Giant RNA genomes: Roles of host, translation elongation, genome architecture, and proteome in nidoviruses

Benjamin W Neuman, Alexandria Smart, Orian Gilmer, Redmond Smyth, Josef Vaas, Nicolai Böker, Dmitry Samborskiy, Ralf Bartenschlager, Stefan Seitz, Alexander E Gorbalenya, et al.

PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Giant RNA genomes: Roles of host, translation elongation, genome architecture, and proteome in nidoviruses

Benjamin W. Neuman[a,1], Alexandria Smart[b,1], Orian Gilmer[b,1], Redmond P. Smyth[b,c,1], Josef Vaas[d,e], Nicolai Böker[f,g], Dmitry V. Samborskiy[h], Ralf Bartenschlager[d,e], Stefan Seitz[d,e], Alexander E. Gorbalenya[i,2], Neva Caliskan[b,j,2], and Chris Lauber[f,g,2]

Affiliations are included on p. 11.

Edited by Eugene Koonin, NIH, Bethesda, MD; received July 25, 2024; accepted January 9, 2025

Positive-strand RNA viruses of the order *Nidovirales* have the largest known RNA genomes of vertebrate and invertebrate viruses with 36.7 and 41.1 kb, respectively. The acquisition of a proofreading exoribonuclease (ExoN) by an ancestral nidovirus enabled crossing of the 20 kb barrier. Other factors constraining genome size variations in nidoviruses remain poorly defined. We assemble 76 genome sequences of invertebrate nidoviruses from >500.000 published transcriptome experiments and triple the number of known nidoviruses with >36 kb genomes, including a 64 kb RNA genome. Many of the identified viral lineages acquired putative enzymatic and other protein domains linked to genome size, host phyla, or virus families. The inserted domains may regulate viral replication and virion formation, or modulate infection otherwise. We classify ExoN-encoding nidoviruses into seven groups and four subgroups, according to canonical and noncanonical modes of viral replicase expression by ribosomes and genomic organization (reModes). The most-represented group employing the canonical reMode comprises invertebrate and vertebrate nidoviruses, including coronaviruses. Six groups with noncanonical reModes include invertebrate nidoviruses with 31-to-64 kb genomes. Among them are viruses with segmented genomes and viruses utilizing dual ribosomal frameshifting that we validate experimentally. Moreover, largest polyprotein length and genome size in nidoviruses show reMode- and host phylum-dependent relationships. We hypothesize that the polyprotein length increase in nidoviruses may be limited by the host-inherent translation fidelity, ultimately setting a nidovirus genome size limit. Thus, expansion of ExoN-encoding RNA virus genomes, the vertebrate/invertebrate host division, the control of viral replicase expression, and translation fidelity are interconnected.

data-driven virus discovery | nidovirus evolution | RNA genome expansion | programmed ribosomal frameshifting | coronavirus genomics

Significance

Nidoviruses, to which the coronaviruses belong, have the largest known RNA genomes. The expansion of nidovirus genomes during evolution must be compatible with their roles in replication, transcription, translation, and encapsidation. By mining raw transcriptomic data of >500.000 public projects of diverse animals, we describe 76 invertebrate nidovirus genomes, including bi- and trisegmented genomes and giant RNA genomes up to 64 kb. The updated proteome of nidoviruses includes putative enzymes and undescribed cofactors of viral infection or modulators of virus–host interaction. We define modes of nidovirus replicase expression, which together with translation fidelity and host dependencies may set limits to RNA genome size expansions. The reported advancements lay the foundation for research of viruses with giant RNA genomes.

Diverse RNA viruses with no DNA stage in their life cycle have been detected in all eukaryotic and bacterial life forms. Based on genome type, they belong to either positive-strand (ssRNA+), negative-strand (ssRNA–), or double-strand RNA (dsRNA) viruses (1). Their average genome size is around 10-11 kb and only few lineages have evolved RNA genomes exceeding 20 kb, according to our analysis of the curated RefSeq database (2, 3) and several recent large-scale metagenomics studies that define the current sampling of the RNA virosphere (4–7). Low fidelity of replication was implicated in limiting genome size in RNA viruses, compared to DNA-based entities (8–10). This limitation may have been alleviated in viruses with >20 kb genomes by either genome segmentation such as in the dsRNA order *Reovirales* (11, 12) or by acquisition of a proofreading DEDDh subfamily exoribonuclease (ExoN) in the single-segment ssRNA+ viruses of the order *Nidovirales* (13–15). This virus order represents a monophyletic group of animal viruses with unsegmented or bisegmented (5, 16) genomes and an extraordinarily wide range of RNA genome sizes from 12 up to 41 kb (17–19), although larger nido-like genomes may exist (20, 21).

Improving our understanding of how nidoviruses have evolved and use such exceptionally large RNA genomes provides a unique window into the biology of primordial life (22). Also, it informs applied research on many pathogens including severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which belongs to the nidovirus family *Coronaviridae* (23) that together with the nidovirus family *Tobaniviridae* include three viruses with the largest known RNA genomes of vertebrate hosts (36.1 to 36.7 kb) (16, 24, 25). Two of these viruses have bisegmented genomes and infect aquatic hosts (5, 16, 24). Based on protein content, the nidovirus genome can be divided, from the 5′- to 3′-end, into three functional regions that predominantly control genome expression, genome replication, and virus dissemination, respectively (26). This genomic organization and the associated division of labor may facilitate

Author contributions: C.L., S.S., N.C., and A.E.G. conceptualized the study approach; C.L., N.C., A.E.G., and R.B. supervised the research; C.L., R.P.S., J.V., N.B., B.W.N., A.E.G., and D.V.S. performed computational research; A.S. performed wet lab experiments; C.L., R.P.S., A.S., O.G., S.S., B.W.N., and A.E.G. analyzed the data; C.L., N.C., A.S., N.B., B.W.N., and A.E.G. created data visualizations; and C.L., A.E.G., and N.C. wrote the paper with inputs from all authors.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]B.W.N., A.S., O.G., and R.P.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: a.e.gorbalenya@lumc.nl, neva.caliskan@ur.de, or chris.lauber@twincore.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2413675122/-/DCSupplemental.

Published February 10, 2025.

and constrain genome size change. For nidoviruses with a canonical single-segment genome and multiple open reading frames (ORFs), the three regions correspond to overlapping large ORF1a and ORF1b, and an array of multiple ORFs at their 3′ (3′ORFs), respectively. Upon translation of the genomic RNA by the ribosome, two polyproteins are synthesized: pp1a, encoded in ORF1a, and pp1ab that is produced by extending translation into ORF1b by –1 programmed ribosomal frameshifting (–1PRF) (27–30). Proteolytic autoprocessing of pp1a/pp1ab produces subunits of the replication–transcription complex (RTC) (31–33). In contrast, 3′ORFs encode structural and accessory proteins that are expressed from a nested set of 5′-coterminal subgenomic mRNAs whose RTC-mediated synthesis on the antigenomic template (transcription) is controlled by discontinuous leader and body transcription-regulating sequences (TRS) in most characterized nidoviruses (34–38). Other nidoviruses may use leaderless transcription (39, 40).

Nidoviruses are distinguished by an array of five universally conserved domains, including four key enzymes of replication: main 3C-like protease (Mpro or 3CLpro) responsible for pp1a/pp1ab autoprocessing downstream of its location; core RTC components including nucleotidyltransferase (NiRAN) that controls the 5′-terminal structure of viral RNAs and is the nidovirus marker domain; RNA-dependent RNA polymerase (RdRp); and Zn-binding domain (ZBD) fused to superfamily 1 helicase (HEL1) (41). In the RTC, they are assisted by many other viral nonstructural protein products of diverse functions and variable conservation. They include ExoN, SAM-dependent N7-methyltransferase (NMT), and SAM-dependent 2′-O-methyltransferase (OMT) encoded downstream of HEL1 and most common in nidoviruses with genomes >20 kb (15, 42). Experimental characterization of SARS-CoV-2 and a few other nidoviruses defines our current understanding of the RTC function and structure that may also involve modulation of innate immunity (31–33). Comparative genomics led to the uncovering of most replicase functions and it places the obtained experimental results within an evolutionary framework. This includes mutation patterns and lineage- and host-specific variations in the RTC subunit composition (41).

Unlike 128 known nidovirus species expressing replicase proteins from two overlapping ORFs, two known nidoviruses with extra large RNA genomes have in-frame ORF1a and ORF1b, either as part of a single genome-wide ORF [41.2 kb genome of planarian secretory cell nidovirus (PSCNV)] (17) or separated by a termination codon [35.9 kb genome of Aplysia abyssovirus 1 (AAbV)] (18). PSCNV and AAbV may use either –1PRF in a noncanonical way or a readthrough mechanism, respectively, to attenuate genomic RNA translation at a position similar to that of the ORF1a/ORF1b junction. This raises the question whether expansion of RNA genomes to very large sizes in nidoviruses is not supported by canonical replicase ORF organization and expression.

To learn about extra large RNA viruses, we seek to expand the genomic characterization of nidoviruses. We apply a Data-Driven Virus Discovery (DDVD) approach (43) to mine raw sequencing data from the Sequence Read Archive (SRA) (44) for divergent RNA viral sequences across a wide eukaryotic host spectrum. We analyze 581,629 eukaryotic SRA transcriptome experiments in a sensitive and highly parallelized fashion, as described previously (16, 45–47). Among the identified set of RdRp-encoding sequences are 116 nidovirus sequences found exclusively in datasets from vertebrate and invertebrate animals. Each of the new viruses with coding-complete genomes reported here encodes the five most conserved replicase domains that distinguish nidoviruses from other viruses. We will refer to nidoviruses in the genome size ranges of 20 to 36 kb and 36 to 64 kb, as large and giant respectively, following a classification framework introduced previously (48) and using two apparent size barriers during genome expansion (15, 26) (and this study). Here, we describe the identification, classification and functional characterization of large and giant nidoviruses in invertebrates. The subset of *vertebrate*-associated nidoviruses is described in another study (16).

Results and Discussion

Discovery of Giant Invertebrate Nidoviruses. We report 76 new nidovirus genomes associated with 72 different *invertebrate* host species (Dataset S1). In this study, invertebrates are defined as all Metazoa not belonging to the clade Vertebrata (phylum Chordata). All new genomes passed a stringent assembly quality control (*SI Appendix*, Fig. S1). Forty of the genomes are coding-complete (and we consider four additional genomes to be nearly coding-complete) with sizes between 18.1 to 64.3 kb (*SI Appendix*, Fig. S2 and Dataset S1). Notably, only six out of the 76 discovered nidovirus genome sequences show >90% nucleotide sequence similarity to five partial and one coding-complete nidovirus genomes found in another DDVD study in which SRA datasets were analyzed (5), and only three similar sequence fragments were reported in a large-scale RNA virus discovery study of metatranscriptomes (6) (*SI Appendix*, *Extended Materials and Methods* and Dataset S2). These comparisons show that the data analysis pipeline used in our study has excellent sensitivity regarding the discovery of highly divergent nidoviruses.

The nidoviruses reported here outnumber three recent RNA virus discovery studies reporting >10^5 viruses with respect to sampling at the upper range of genome size (Fig. 1*A*). They account for over 50% of approved and tentative RNA virus species with genomes larger than 32 kb and 100% of genomes larger than 41.2 kb (Fig. 1*B*). The newly discovered nidoviruses include viruses with either unsegmented or segmented genomes (*SI Appendix*, Fig. S2). The genomes of Crassostrea gigas nidovirus (CGNV) from an oyster, Pomacea canaliculata nidovirus (PCNV) from a snail, and Poecilobdella javanica nidovirus (PJNV) from a leech have lengths of, respectively, 64.3, 54.1, and 45.8 kb (Fig. 1*C*). Together with a 47.3 kb nido-like viral genome reported recently (20), they are the largest known RNA viruses, exceeding the median RNA virus genome size of 10 to 11 kb four- to six-fold (Fig. 1*B*). They highlight the extraordinary capacity of ExoN-encoding nidoviruses for genome expansion, notwithstanding a recently reported 39.8 kb flavi-like virus that is thought to have alleviated genome size constraints by an ExoN-independent mechanism of RNA synthesis involving homologs to cellular domains associated with nucleic acid metabolism (49).

Functional Annotation of Nidovirus Genomes. All identified nidovirus genomes with completely assembled ORF1b or its equivalent encode the replicative protein domains NiRAN, RdRp, ZBD, HEL1, ExoN, NMT, and OMT in the expected order (Fig. 2), with a single exception of Longidorus elongatus nidovirus (LENV) (see below). However, OMTs in several viruses have no complete counterpart to the catalytic K-D-K-E tetrad, with the SAM-binding site being mutated. Like the previously described PSCNV (17), several new nidoviruses have replacements of otherwise conserved two Cys and two His residues coordinating a Zn^{2+} in the active site of ExoN. In addition to the ORF1b-encoded protein domains, the invertebrate nidovirus genomes with a complete ORF1a homologous region or major parts of it also encode a 3CLpro flanked by transmembrane domains upstream and downstream of it (Fig. 2 and *SI Appendix*, Fig. S2). The exact boundaries of these and other protein domains remain uncertain, as well as 3CLpro cleavage sites, which may deviate from the canonical sites in coronaviruses (17, 50).

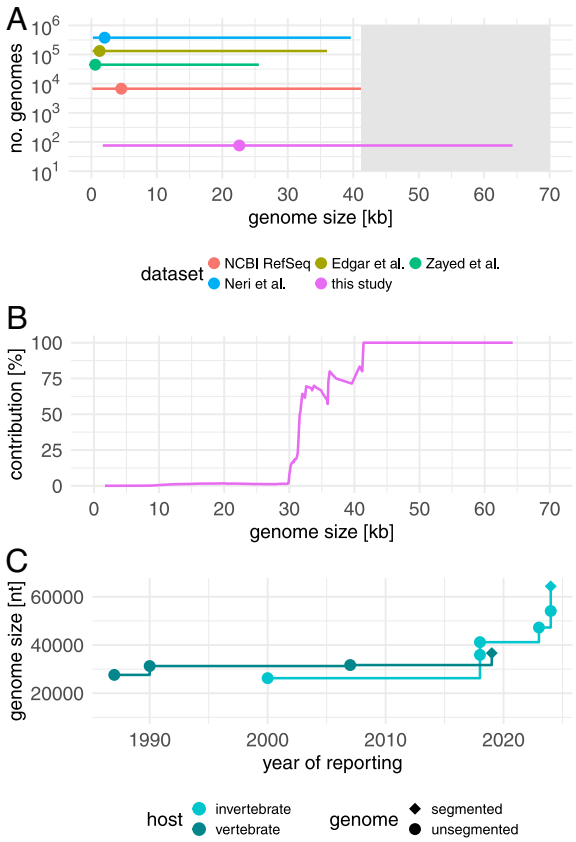

Fig. 1. Genomes sizes of nidoviruses and other RNA viruses. (*A*) Genome size ranges across RNA viruses in NCBI RefSeq and three recent large-scale virus discovery studies (5–7) and nidovirus genome size range reported in this study; circles indicate median values. (*B*) Percentage of RNA viruses contributed by this study within particular genome size ranges. (*C*) Size increase of the largest known genome of vertebrate and invertebrate nidoviruses from 1987 until 2024.

Also, we identified tentative structural and various putative accessory proteins encoded by the identified nidoviruses using hidden Markov models for scanning various sequence and structural databases (*SI Appendix*, *Extended Materials and Methods* and Dataset S3). In the 3′-part of the genome (downstream of ORF1b or on the second segment) we detected similarity with viral glycoproteins of different origins, E2 of alphaviruses (nine viruses), and G2 of nairo-/hantaviruses (six viruses; Fig. 2 and *SI Appendix*, Figs. S2 and S3 and Dataset S3). Moreover, 11 newly found viruses, including CGNV, PCNV, PJNV, and Actinia tenebrosa nidovirus (ATNV) encode a serine proteinase domain with two-barrel chymotrypsin-like fold in the structural module (CPPro; *SI Appendix*, Fig. S3 and Dataset S3). In alphaviruses, CPPro initiates autocatalytic processing of the structural polyprotein, containing E1 and E2 glycoproteins, and is used as nucleocapsid (51, 52). Seven nidoviruses that encode CPPro upstream of the E2 homolog might employ a similar mechanism, although no particular sequence affinity was found between alphavirus and nidovirus CPPros. Also, CPPros may have other functions, particularly in four nidoviruses that do not encode E2 homologs.

We identified several notable insertions of protein domains with putative enzymatic functions in pp1a/pp1ab of separate viral lineages or individual viruses (*SI Appendix*, Fig. S3). This included homologs of ADP-ribose-1″-phosphatases (ADRP, known also as macrodomain) (53), NAD- and ADP-ribose domain (NADAR) (53), Alkylated DNA repair dioxygenase (AlkB) (54), ribonuclease T2 (55) and endonuclease HNH (56) (RNase T2 and RNase HNH), methyltransferases (MT) including OMT homologs, a protein kinase, RNA ligase 2′–5′ (RLigase2p5) (53), RNA 2′-phosphotransferase (PTP1) (57), exoribonuclease 1 (ERI1), domain 3 of NAD+-dependent DNA ligase (NA3Li3D) (58), papain-like protease (48), ZBD paralog, and a new family of proteins containing nucleotide-binding Walker-box A and B motifs (59) named REXA (see below), among others. Many belong to protein (super)families that are part of the proteomes of previously characterized nidoviruses or other RNA viruses. RNase HNH, PTP1, ERI1, RLigase2p5, NA3Li3D, and REXA expand the known repertoire. These domains are often ancient ADP/NAD+-dependent and cooperate in diverse RNA-based processes by defining direction and pace of reactions. Several nidoviruses encode two divergent copies of the same domain, such as RNase HNH, RNase T2, AlkB, NADAR, ADRP, or G2, that may have been acquired independently or through duplication (*SI Appendix*, Fig. S3). Several of the inserted domains appeared to be truncated in comparison with homologs from reference databases or have unusual substitutions in putative functional sites, informing future studies.

Coding regions for the domains were inserted at various sites in the genome, with three apparent hotspots: upstream of 3CLpro in ORF1a, between the RdRp and ZBD in ORF1b, and between the HEL1 and ExoN in ORF1b (*SI Appendix*, Fig. S3). Relative to coronaviruses, 21 different annotated domain varieties were inserted in the most conserved replicase area flanked by NiRAN and OMT. The largest number of nine varieties was observed for ADRP with six of them being located at different positions in ORF1b of the individual recipient genomes. Insertion of an ADRP domain in ORF1b was previously reported only for a vertebrate coronavirus from Pacific salmons, between the uridylate-specific endoribonuclease (NendoU) and OMT (Fig. 2) (16). The identified ADRPs diverged profoundly and could play diverse ADP-dependent roles documented for these proteins (53). Other frequent insertions observed at different genomic sites included NADAR, RNase T2 and RNase HNH, and MTs with different specificities (*SI Appendix*, Fig. S3).

The domain insertions may be specific to either individual viruses or viral clades of variable diversity. Several suggest linkages to virus family, genomic region or genome size. A 3-domain combination of the adjacent REXA, AlkB, and MT was inserted between the RdRp and the ZBD of all 15 known members of an undescribed viral clade with putative spider hosts (*SI Appendix*, Fig. S3). Also, these viruses with genomes in the size range of 29.2 to 32.1 kb may encode PLpro, DUF5652, and AlkB in ORF1a as well as a RNase T2 homolog in structural module. These seven domains may work in a host-dependent manner starting in an ancestor of these viruses (*SI Appendix*, Fig. S3).

REXA homologs were identified in four locations adjacent to NiRAN, RdRp, HEL1, and ExoN of 18 nidoviruses, all found in Arthropoda (see below). They range in size from 120 to 180 aa and some have substitutions in the nucleotide-binding site residues, indicating that they may be enzymatically compromised. We identified their most closely related homolog inserted into the putative RdRp of a flavi-like virus with a 22 kb RNA genome (60). Due to the strong genetic linkage of this domain with key enzymes of replication in RNA viruses with large genomes, we named this domain RNA genome expansion and replicase associated (REXA) for brevity.

Unlike other known nidoviruses, Macrobrachium nipponense nidovirus encodes two distantly related 5′-3′exoribonucleases, proofreading ExoN and ERI1, involved in divergent RNA metabolic pathways (61) (*SI Appendix*, Fig. S3). The later segregates with RdRp in the 35.1 kb genome, that is considerably larger compared to the 26.3-to-32.0 kb size range of other roniviruses, indicating a possible link to genome expansion in the family *Roniviridae*. Likewise, a combination of NADAR_3 and ADRP_3 domain varieties was shared between the 40.8 kb genome of Penaeus monodon nidovirus and the distantly related, 36.6 and 45.8 kb genomes of WPNV and PCJV (*SI Appendix*, Fig. S3),

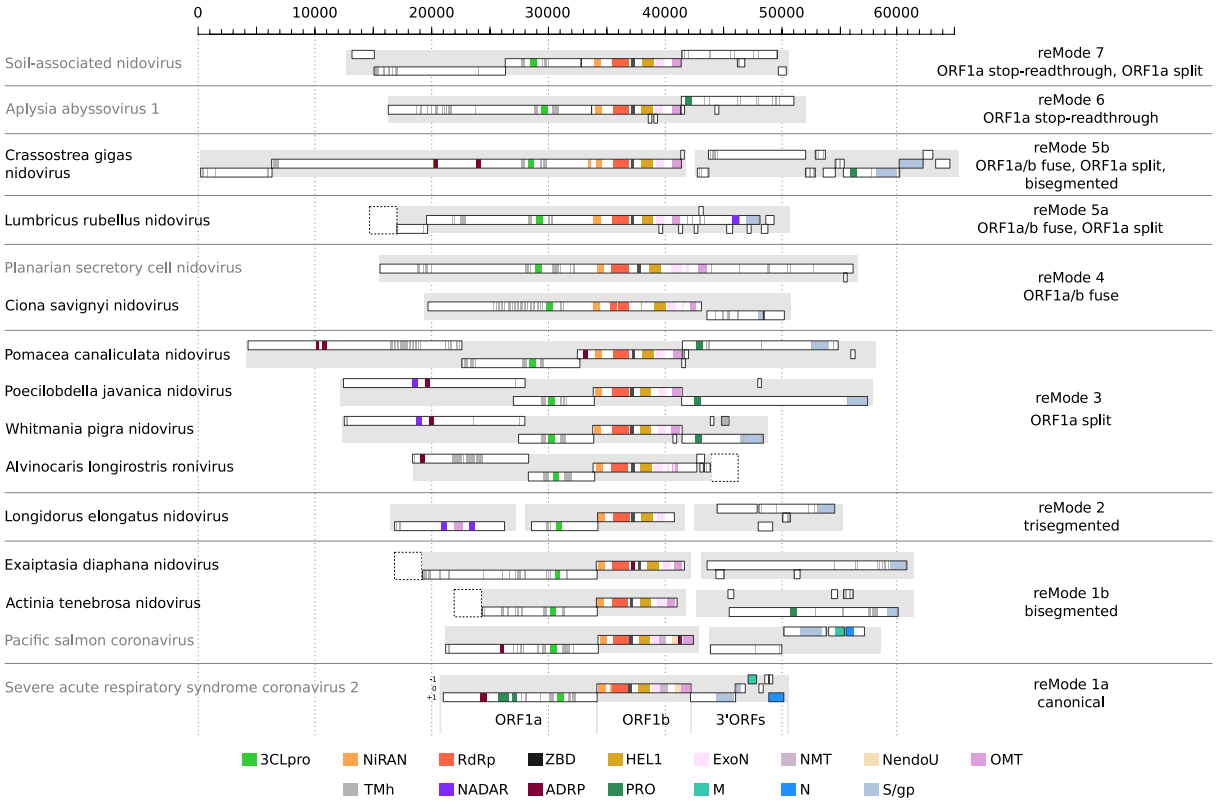

Fig. 2. Genomic organization, functional annotation, and mechanisms of regulating translation of replicative proteins in nidoviruses. Genomic organizations of selected nidovirus genomes drawn to scale in units of nucleotides. Newly identified nidoviruses and reference viruses have black and gray names, respectively. White rectangles indicate predicted ORFs of at least 300 nt in length delimited by termination codons at both ends. Selected nidovirus proteins are in color; for a more detailed domain annotation see *SI Appendix*, Fig. S3. Rectangles with dashed borders indicated sequence of unknown length predicted to be missing in incomplete assemblies. Seven main reModes and four submodes are indicated to the right; ORF1a/b fuse, ORF1a and ORF1b equivalents fused in-frame; ORF1a split, ORF1a equivalent split into overlapping ORF1a1 and ORF1a2. The reference viruses are based on the following NCBI Genbank/RefSeq entries: Soil-associated nidovirus 2 (BK066825.1), AAbV (NC_040711.1), PSCNV (NC_040361.1), Pacific salmon coronavirus (MK611985.1), SARS-CoV-2 (NC_045512.2). Domain abbreviations: TMh (transmembrane helix), ADP-ribose-1″-phosphatase (ADRP), 3C-like protease (3CLpro), nucleotidyltransferase (NiRAN), RNA-dependent RNA polymerase (RdRp), Zinc-binding domain (ZBD), superfamily HEL1, exonuclease (ExoN), N-methyltransferase (NMT), uridylate-specific endoribonuclease (NendoU), O-methyltransferase (OMT), NAD- and ADP-ribose domain (NADAR), protease (PRO), matrix protein (M), nucleocapsid protein (N), viral glycoprotein (S/gp).

suggesting a possible role of these two domains in the evolution of giant nidovirus genomes.

The genetic segregation of the inserted protein domains with key replicative enzymes suggests that they are incorporated in the RTC and modulate genome replication and/or transcription in nidoviruses. Surprisingly, three domains were inserted *inside* a main replicase protein in loop regions connecting conserved motifs. We identified a putative RNA MT inside the NiRAN of CGNV, a REXA domain inside the linker domain connecting NiRAN and RdRp of Episyrphus balteatus mesonivirus (EBMV), and another REXA domain inside the ExoN of Nilaparvata lugens nidovirus. These domains may be expressed along with their recipient domains in the same protein indicating hitherto unknown dependencies of key enzymes controlling initiation, elongation, and proofreading of RNA synthesis. Indeed, the 3D model of the REXA domain insertion in EBMV was structurally compatible with the NiRAN-RdRp domain model of this virus (*SI Appendix, Extended Materials and Methods* and Fig. S4). It received several hits just above the significance threshold in a Dali server search of the PDB (top hit: vacuolar protein sorting-associated protein 72, chain: 6gej-M, Z = 4.7, RMSD = 4.8, 10% identity), indicating that orthologs of this domain are yet to be structurally characterized.

Phylogenetic Analysis, Sequence-Based Classification, and Host Association of Discovered Nidoviruses. All 40 coding-complete and 12 partial genomes encode the nidovirus-specific array of five conserved domains (3CLpro-NiRAN-RdRp-ZBD-HEL1) (*SI Appendix*, Fig. S2 and Dataset S1). Based on RdRp sequence similarity, the discovered viruses form numerous divergent phylogenetic lineages within the *Nidovirales* (Fig. 3). Notably, this includes three viruses (Eurytemora affinis nidovirus, Frankliniella intonsa nidovirus, Megalurothrips usitatus nidovirus) that form the two most basal lineages in the nidovirus phylogeny rooted with the order *Picornavirales* (Fig. 3 and *SI Appendix*, *Extended Materials and Methods* and Dataset S1). The remaining 24 nidoviruses with an incomplete 5-domain array (in the most extreme case we only retrieved the RdRp coding region) clustered with viruses with a complete domain array (Fig. 3 and Dataset S1). Together, these results provide strong support for all newly identified viruses being members of the *Nidovirales*, although this assignment remains provisional for the viruses with coding-incomplete genomes.

We tentatively classify the newly identified nidoviruses into 18 virus family-like operational taxonomic units (fOTUs), of which 11 are new, using phylogenetic analysis and DEmARC, a quantitative, sequence-based classification approach applied here to the RdRp (62, 63) (*SI Appendix*, *Extended Materials and Methods* and Fig. S5 and Dataset S4). The deep interfamily relationships showed no strong support in the rooted tree, and this classification must be refined in future for it to become part of virus taxonomy. Notwithstanding that, the viruses described here may triple the number of recognized invertebrate nidovirus families (from 6 currently, to 17) and almost double the total number of nidovirus families (from 14 currently, to 25). Many of the newly delineated nidoviral taxa have extraordinary features of

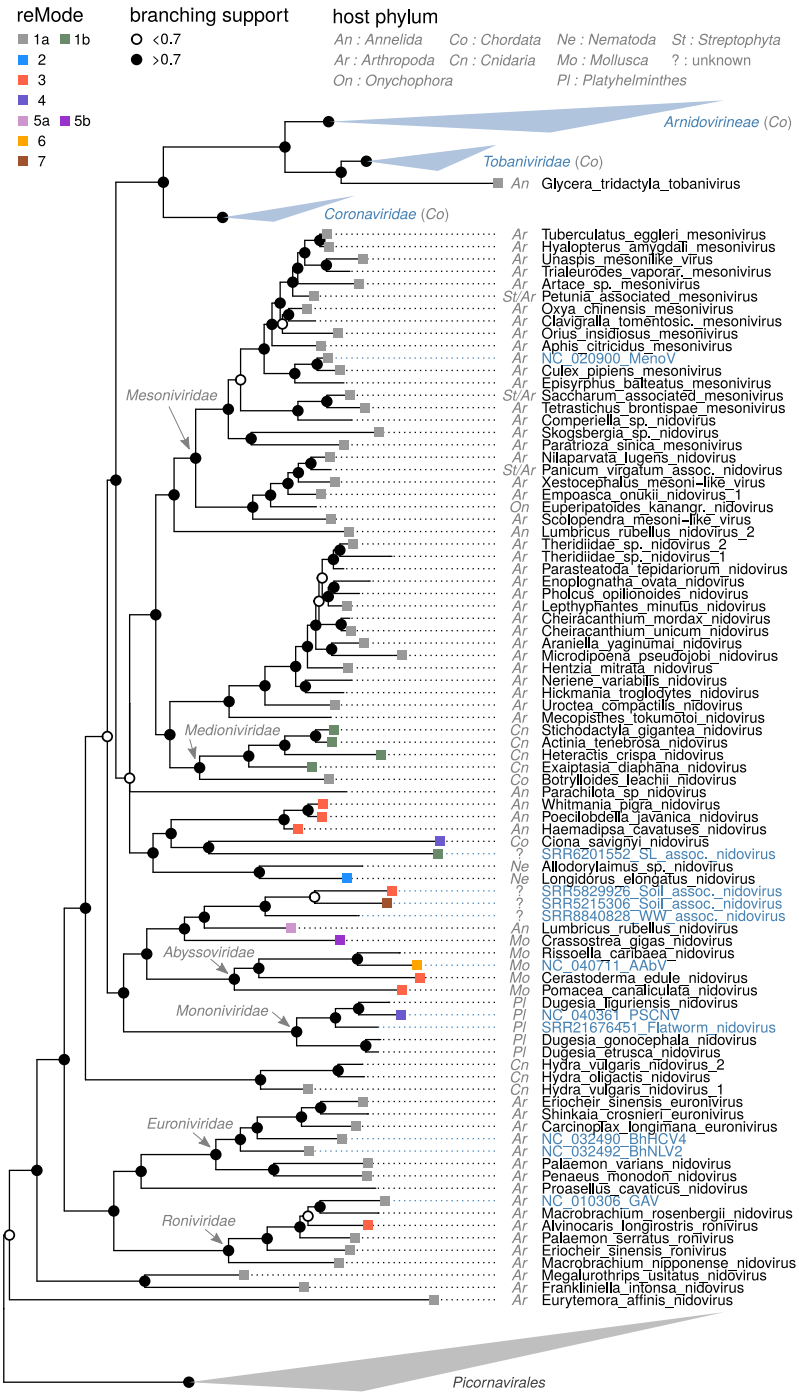

Fig. 3. Phylogeny of newly discovered nidoviruses. RdRp phylogeny reconstructed using PhyML with SH-like branching support. *Picornavirales* reference sequences are used as an outgroup and were collapsed (gray triangle). Names of novel nidoviruses reported in this study are in black and those of selected reference viruses are in blue. White and black circles at internal nodes indicate SH-like branching support smaller and larger than 0.7, respectively. The host phylum, either known for reference viruses or inferred from the respective sequencing experiment for newly discovered viruses, is shown in gray next to virus/taxa names. Nidoviruses identified in land plants (*Streptophyta*) experiments are assumed to have arthropod hosts (St/Ar). Colored squares at tips indicate the replicase expression mode (reMode). Major vertebrate nidovirus taxa were collapsed (blue triangles). Virus abbreviations: MenoV (Meno virus), SL assoc. nidovirus (saline lake-associated nidovirus), Soil assoc. nidovirus (soil-associated nidovirus), WW assoc. nidovirus (wastewater-associated nidovirus), AAbV, PSCNV, BhHC4 (Beihai hermit crab virus 4), BhNLV2 (Beihai Nido-like virus 2), GAV (Gill-associated virus).

functional and evolutionary significance, which are evident in the virus proteome (see above, *SI Appendix*, Fig. S3) and are also described below.

Most of the reported nidoviruses were identified in Arthropoda hosts (55 nidoviruses) and, depending on tree partitioning, they form four or five suprafamily clusters in the rooted RdRp-tree, including several basal clusters (Fig. 3). In all likelihood, three nidoviruses from plant experiments are of arthropod host origin according to their position in the tree (Fig. 3). Other notable host-related clusters include nidoviruses that infect phyla Annelida (three clusters), Cnidaria (two), Mollusca (one), Nematoda (one), and Platyhelminthes (one). These virus–host associations indicate a host barrier for nidovirus dissemination at the host phylum level. This barrier might have been crossed on several occasions, since the sole known nidovirus infecting a host from phylum Onychophora intertwines with nidoviruses isolated from Arthropoda while one of the Annelida clusters formed by Lumbricus rubellus nidovirus 2 intertwines with Arthropoda clusters (Fig. 3). Moreover, we found two nidoviruses isolated from Tunicata hosts (phylum Chordata), Ciona savignyi nidovirus (CSNV) and Botrylloides leachii nidovirus, to be separated and cluster with different invertebrate nidoviruses far from families *Coronaviridae* and *Tobaniviridae* infecting Vertebrata hosts (phylum Chordata) (Fig. 3). No nidoviruses have yet been identified in several dozen other invertebrate host phyla.

Classification of Nidoviruses Based on Genomic Organization and Translation Regulation. The 76 discovered viruses have diverse genome coding organizations that implicate different mechanisms for the expression of the RdRp and other RTC components. Based on the number of genome segments and organization of ORF(s) encoding

the replicase proteins, we recognized seven (plausible) modes of replicase expression by translation (denoted "reModes" from 1 to 7) and four submodes of two reModes that are denoted with suffix "a" or "b" after the respective reMode numeral (Fig. 2). This classification is defined by variations at three genome regions listed from 5′-end to 3′-end and including key signals of genome expression: upstream of 3CLpro (the N-terminal most cognate 3CLpro cleavage site), upstream of NiRAN (ORF1a/b -1PRF), and downstream of OMT (ORF1b terminal codon) in coronaviruses (26). They demark the beginning of 3CLpro control over replicase pp1a/pp1ab processing, separate loci encoding accessory subunits and key replicase enzymes of the RTC, and separate replicase from structural genes, respectively. The first two sites (controlling replicase expression) define the main reMode while the third site (concerning expression of structural proteins) is used to discriminate submodes. The reModes 1, 4, and 6 accommodate already described mechanisms.

The reMode 1 includes two submodes, the canonical ORF1a/b –1PRF-based mechanism of unsegmented genomes (reMode 1a; n = 40 newly discovered virus species) and bisegmented genomes that encode replicase and structural genes on separate segments (reMode 1b; n = 6). The two others are employed by the single-ORF (ORF1a/b in-frame fuse) PSCNV (reMode 4; n = 4) and AAbV whose ORF1a and ORF1b are connected by a readthrough stop codon (reMode 6; n = 1) (Fig. 2 and Dataset S1). Another mode (reMode 7; n=1) is used by a soil-associated nidovirus that is reported independently in two studies (20, 64). This 38 kb viral genome has the 3′-end of ORF1a and the 5′-end of ORF1b connected by a readthrough stop codon as described for AAbV (18) but encodes two additional ORFs upstream of the 3CLpro (Fig. 2).

Viruses with the remaining three reModes (2, 3, and 5) including one represented by two sub-reModes (5a and 5b) have not been described previously (Fig. 2 and Dataset S1). Two viruses have trisegmented genomes with segmentation affecting replicase coding and expression (reMode 2; n = 2). Six nidoviruses have replicase expression regulated by two –1PRFs (one of which is upstream of the 3CLpro) that extend translation of the viral genome RNA at junctions in two pairs of overlapping ORFs (reMode 3; n = 6). Two nidoviruses also have two –1PRFs of which the first is located in an overlapping ORF region upstream of the 3CLpro and results in translation extension into a new ORF encoding RTC subunits, while the second redirects a fraction of ribosomes to terminate translation in a manner described for PSCNV (reMode 4). This reMode 5 is employed by unsegmented (reMode 5a; n = 1) and bisegmented (reMode 5b; n = 1) viruses.

Nidoviruses of invertebrate hosts using canonical reMode 1a outnumber by four times all others combined at the species(-like) rank (78 vs. 19 taxa), while remain comparable at the family(-like) rank (10 vs. 9) (Dataset S4), indicating that viruses of this submode are the best albeit unevenly sampled (Fig. 4*A*). Thus, invertebrate nidoviruses with noncanonical reModes (1b to 7) may be much more common than suggested by their current and relatively low sampling. Indeed, the majority of virus species employing reMode 1a cluster separately from others in few family-restricted monophyletic lineages (Fig. 3). Likewise, group-specific phylogenetic clustering is also observed for viruses of reModes 1b, 3, 4, and 6, which include two or more viruses (Fig. 3 and Dataset S4). It implies that once emerged, a reMode-specific genomic make-up persists in evolution. Furthermore, each of reModes 1b, 3, and 4 are observed in two very distantly related lineages, indicating that they may have emerged independently at least twice during the evolution of nidoviruses.

Many of the viruses that form most basal lineages in the nidovirus phylogeny have reMode 1a, which thus may represent the ancestral state, if the rooting is confirmed. For instance, two groups of reMode 1b viruses as well as shrimp Alvinocaris longirostris ronivirus of reMode 3 interleave with numerous viruses of reMode 1a representing the family *Roniviridae*, indicating that they may have emerged from reMode 1a viruses. There are exceptions from the phylogenetic clustering of reModes, including nidoviruses of reModes 3, 5, 6, and 7 derived from mollusk, annelid, and soil projects, which are sister to each other and are deeply rooted in the nidovirus phylogeny (Fig. 3).

Association of Replicase Expression Modes With Genome Expansion Trajectories. The wide genome size range of 20 to 64 kb suggests that multiple factors may determine genome size in ExoN-encoding nidoviruses. Nidovirus genomes may be frozen at different stages of the wavelike expansion trajectory shaped by a hierarchical division of labor between three genome regions responsible for genome replication and expression, and virion dissemination (26). ExoN-encoding viruses of reMode 1a and 1b, which employ canonical replicase expression, include vertebrate and invertebrate viruses with genomes in the size range from 20 to 41 kb (Fig. 4*A*), with only three having genomes larger than 36 kb: the 36.1 kb Veiled chameleon serpentovirus (25), the 36.1 kb Hypomesus transpacificus coronavirus (16), and the 40.8 kb Penaeus monodon nidovirus from this study (Dataset S1). For vertebrate nidoviruses, a genome size of about 36 kb seems to be close to the upper limit, given how modestly (9 kb) the known largest genome size of vertebrate-associated nidoviruses increased over 38 y since the first coronavirus genome was reported (Fig. 1*C*). In contrast, viruses with reModes 2 to 7 belong to seven fOTUs of invertebrate nidoviruses with 31-to-64 kb genomes, indicating a linkage between reMode, the vertebrate/invertebrate host divide and genome expansion in nidoviruses.

Genome Segmentation May Facilitate Nidovirus Genome Expansions. For both vertebrate and invertebrate nidoviruses, the largest known genome is segmented (Figs. 1*C* and 4*A*), implicating genome segmentation into the evolution of very large RNA genomes [similar to the multisegment genomes in the size range from 18 to 30 kb in the order *Reovirales* (11, 12)]. Interestingly, none of the segmented genomes of either vertebrate or invertebrate hosts are smaller than 30 kb, indicative of a lower genome-size constraint on genome segmentation in nidoviruses.

Segmentation decreases the size of single molecules to be copied faithfully, and that gain alone may be sufficient to produce relatively large segmented genomes. Indeed, the size of the largest segment in the multisegmented nidoviruses other than CGNV is in the 14 to 24 kb range, which is at the low end of genome size range of nidoviruses (Fig. 4 *A* and *B*) across major lineages in different hosts (Fig. 3). In the 64 kb CGNV, the largest segment is of 41 kb, which is just above the 36 kb size threshold separating large and giant nidoviruses, including (with very few exceptions) viruses with canonical reMode 1a and noncanonical reModes, respectively. Thus, segmentation in this virus may be linked to the use of reMode 5b in addition to its host and unique proteome (see also below).

Theoretically, the largest gain of genome size increase by segmentation may be achieved in viruses that split the genome into segments of equal size (Fig. 4*B*). Based on their segment and genome sizes, segmented invertebrate nidoviruses in the genome size range of 31 to 36 kb are close to the theoretical genome size limit, indicating that they use segmentation most efficiently for the purpose of genome expansion. In contrast, the giant 64 kb genome of CGNV is split unevenly between 41 kb and 23 kb segments; if both segments were as large as its largest, the virus

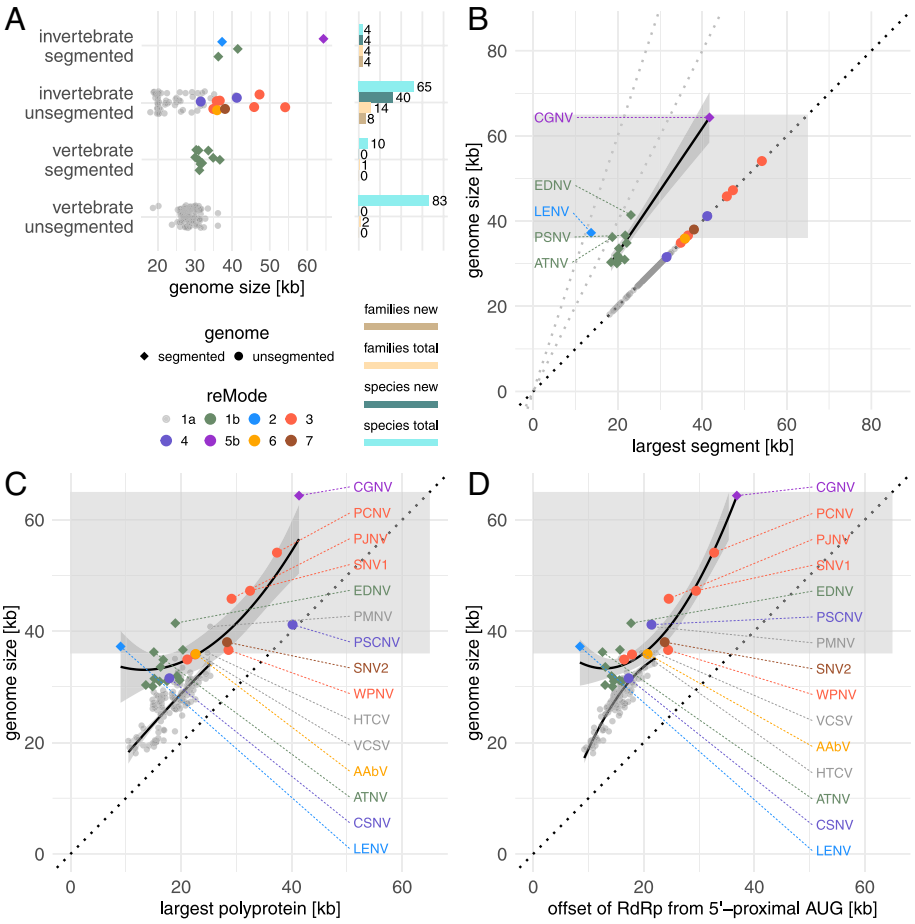

Fig. 4. Nidovirus genome size expansion and its constraints. (*A*) Genome sizes of ExoN-encoding vertebrate and invertebrate nidoviruses with segmented and unsegmented genomes. Only newly discovered nidoviruses with coding-complete genomes are included in addition to 156 coding-complete nidovirus genomes from NCBI RefSeq. Viruses with segmented and unsegmented genomes are shown as diamonds and circles, respectively. Symbol color indicates reMode. The number of total and new virus species and virus families per group is indicated at the right. For the same viruses, the relation of genome size to the size of the largest segment (*B*), the length of the largest polyprotein (*C*), and the position of the RdRp in the pp1ab polyprotein (*D*) is shown. The black dotted line is the main diagonal. The black solid curves are local polynomial regression (LOESS) fits with confidence intervals shown as dark-gray background. LOESS is applied to the bisegmented viruses in *B* and separately to reMode 1a and non-reMode 1a viruses in *C* and *D*. The gray dotted lines in *B* show the theoretical genome size limit for bi- and trisegmented viruses. The light-gray rectangle in the background highlights the genome size region above 36 kb. The following viruses are labeled: LENV, ATNV, Exaiptasia diaphana nidovirus (EDNV), Whitmania pigra nidovirus (WPNV), PCNV, PJNV, Haemadipsa cavatuses nidovirus (HCNV), CSNV, CGNV, Penaeus monodon nidovirus (PMNV), Veiled chameleon serpentovirus (VCSV, NC_076911.1), PSCNV (NC_040361.1), AAbV (NC_040711.1), Pacific salmon nidovirus (PSNV), Hydra vulgaris nidovirus 2 (HVNV2), Soil-associated nidovirus 1 (SNV1) (20), Soil-associated nidovirus 2 (SNV2, BK066825.1) (20, 64).

genome size would be 82 kb. To accommodate multiple segments of a genome, the nidovirus life cycle may be further elaborated, with segment packaging into virus particles being one of the affected stages, as studies of other segmented viruses found (65, 66). Evolving respective mechanisms and toolkits must engage additional constraints and this may limit evolutionary space and host range of multisegmented nidoviruses, which is in line with the relative minority of these viruses in the available sampling.

Translation Fidelity May Limit Nidovirus Polyprotein Length and Genome Size. Nidovirus genomes can encode extraordinarily long polyproteins up to 13.8 thousand aa (Fig. 2), whose size may be comparable to the largest known host proteins and exceed the average of 0.3 thousand aa by multifold. This size is in line with an estimated rate of amino acid misincorporation of 10^{-3} to 10^{-4} (67, 68) that may be modulated by external factors, including mRNA turnover (69, 70). The fidelity of translation is therefore another factor that can constrain the expansion of positive-stranded RNA virus genomes. In contrast to RdRp-mediated fidelity of genome replication that is virus-specific, translation fidelity is host-determined.

We hypothesized that, if translation fidelity constrains nidovirus genome size, there would be an upper limit on the length of a polyprotein synthesized from the viral genome. Such a length limit would keep the number of amino acids that are misincorporated on average by the ribosome during protein synthesis to remain compatible with protein function. To test this hypothesis, we compared genome size with the longest in silico translated ORF (hereafter largest polyprotein) in nidoviruses. Although nidoviruses may encode several large proteins, the largest polyprotein always includes the RdRp and other universally conserved replicase domains. Since this polyprotein is the first synthesized from the incoming genomic RNA or its largest segment in segmented viruses, it plays a critical role in the nidovirus life cycle by forming the RTC core. We observed that genome size and largest polyprotein length show a near linear-like relation when jointly considering all viruses or viruses from well-sampled reModes 1a, 1b, and 3, separately (*SI Appendix*, Fig. S6*A*). Yet, several viruses deviate considerably from the regression. To gain insight into this variation, we compared the polyprotein-vs-genome-size relation for nidoviruses with canonical reMode 1a and noncanonical reModes (Fig. 4*C*). Only a single, coding-incomplete genome is available for reMode 5a. This reMode is therefore excluded from analyses of genome expansion constraints that require coding-complete genomes.

For reMode 1a viruses, the regression was linear-like, showing that genome and polyprotein lengths are strongly linked during genome size change (Fig. 4*C*). In contrast, for the second group of viruses with noncanonical reModes, we observed a nonlinear, power-like relation (albeit with a relatively large regression confidence interval) (Fig. 4*C*). Accordingly, the noncanonical reMode viruses in the genome size ranges of 31-to-41 kb and 41-to-64 kb have different genome size shares for encoding the longest polyprotein: from 24-to-78% and 46-to-98%, respectively. At the extremes, these two percent ranges include the smallest and largest shares observed in nidoviruses (Fig. 4*C*). The latter subgroup includes five viruses among which are the two largest nidoviruses CGNV and PCNV, as well as PSCNV. These three viruses have very different genome organizations but form a group encoding the largest replicase polyproteins of comparable lengths (from 37.3 to 41.3 kb, or 12.4 to 13.8 thousand amino acids). Apparently, this polyprotein length may be close to its upper limit and may thus reveal a constraint on genome expansion.

While all encoded proteins matter for virus fitness, viruses must be most sensitive to faithful synthesis of RTC components. Within this framework, the offset of the RdRp from the initiator AUG codon may be considered as a proxy for the RTC synthesis in the analysis of protein size limits. We used this characteristic instead of the largest polyprotein in comparisons with genome size and observed improved confidence for the regressions involving reMode1a and noncanonical reModes, with the single-ORF PSCNV moved closer to the regression line (compare Fig. 4 *C* and *D*). These results connect translation elongation of the RNA genome to RTC synthesis, offering a biological interpretation of the observed size relationships.

Given translation is host controlled, we also partitioned nidoviruses according to the phylum of their host for the purpose of this analysis (*SI Appendix*, Fig. S6*B*). Remarkably, the linear-like relation described above for reModes 1a, 1b, and 3 were largely reproduced for nidoviruses infecting hosts of the phyla Arthropoda, Chordata, and Annelida due to the respective strong reMode-host associations. Notably, this linear relation of genome size and polyprotein length also applied to viruses with hosts from the phylum Mollusca, which includes viruses of three reModes with two having the largest known nidovirus genomes. These observations show that translation factors at the host phylum level may shape nidovirus genome expansions.

Dual Ribosomal Frameshifting in a Subset of Nidoviruses with Giant Genomes. In reMode 3 viruses (with genome sizes in the range from 35 to 54 kb), which include PCNV, PJNV, and four other invertebrate viruses in three distinct clades in the nidovirus tree (Fig. 3), the exceptionally large ORF1a region is represented by two overlapping ORFs, denoted ORF1a1 and ORF1a2 from here on, with the overlap locating upstream of 3CLpro (Fig. 2). Otherwise, viruses of reMode 3 resemble reMode 1a in replicase ORF organization. Notably, this genomic organization implies that translation of the viral RdRp locus in these viruses from the incoming genomic RNA must involve two instead of one −1PRF event observed in reMode 1a. In addition, such mechanism would constitute an additional layer of controlling the stoichiometries of viral replicase proteins encoded in ORF1a1 relative to those in ORF1a2 and ORF1b, like observed in arteriviruses (71, 72). In coronaviruses, a positionally equivalent genomic region encodes multiple functions including those involved in the formation of a molecular pore that spans the double membrane of the coronavirus replication organelle (73) and a factor (nsp1) that is involved in the shutdown of host mRNA translation (74, 75).

The presence of bioinformatically predicted frameshift signals (*SI Appendix, Extended Materials and Methods*) in and immediately downstream of the ORF1a1-ORF1a2 and the ORF1a2-ORF1b overlapping region in the PCNV genome, including predicted RNA secondary structures and two identical slippery sequences (UUUAAAC), strongly support this reMode 3 expression mechanism. Similar genetic elements with PRF signals were identified in the ORF-overlapping regions of most of the other reMode 3 viruses (*SI Appendix*, Fig. S7).

Chemical Probing of the Structural Ensembles Associated with Dual Ribosomal Frameshifting. To investigate the structural folds of the predicted nidoviral frameshift RNAs, we next probed the structures of the RNAs using the chemical dimethyl sulfate (DMS), which reacts with unpaired A and C residues (*SI Appendix, Extended Materials and Methods*) (76). To account for possible effects of temperature on RNA structure, we performed folding and DMS treatment at 22 and 37 °C. We did not observe substantial differences in RNA reactivities, indicating that the viral RNA folds are likely to be the same at both temperatures.

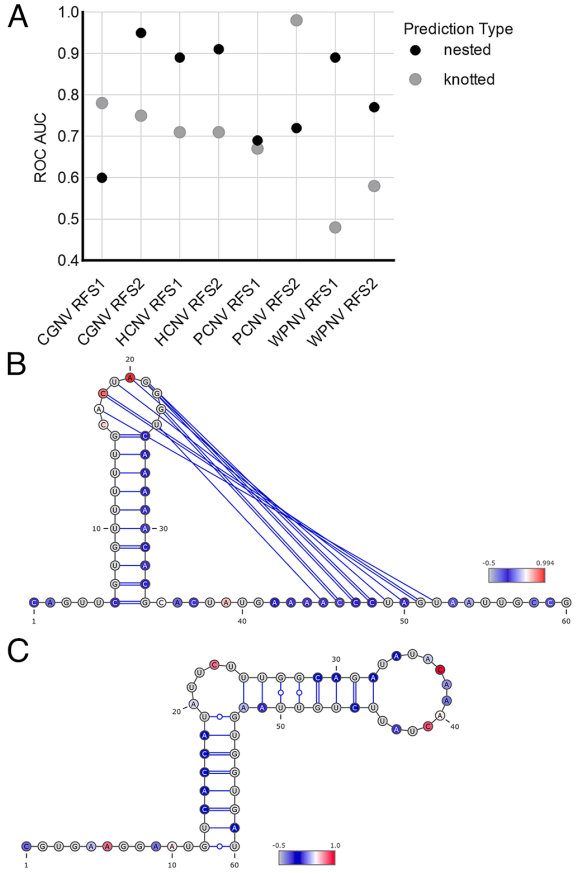

Fig. 5. RNA structures associated with dual ribosomal frameshifting in four nidoviruses. (*A*) ROC-AUC analysis for the RNA variants of four nidoviruses. ROC score was calculated for the nested (black) and knotted (gray) RNA secondary structure models (*SI Appendix*, Fig. S4). (*B* and *C*) RNA secondary structure models of CGNV RFS1 and RFS2. Models were obtained using DMS reactivities as constraints for in silico folding in RNAStructure 3. Blue lines represent PK interactions supported by DMS-MaP. Red and blue nucleotides: red denotes highly reactive *A*, *C* residues, whereas blue indicates unreactive residues that are likely paired.

We then measured the match between DMS reactivities and predicted pseudoknot or nested RNA structures, using a receiver operating characteristic area under the curve (ROC-AUC) analysis (*SI Appendix, Extended Materials and Methods*) (Fig. 5*A*). Overall, the ROC-AUC analysis favored either the nested model or the pseudoknot-containing model for each of the analyzed RNAs, except for PCNV RFS1, which showed a score close to 0.7 for both models. In CGNV RFS1, the RNA fold was consistent with a pseudoknot, whereas the most likely conformation of CGNV RFS2 was a stem loop with a bulge (Fig. 5 *B* and *C*). Similarly, HCNV RFS1 and RFS2 folded into stem loops with a bulge (*SI Appendix*, Fig. S8). In PCNV RFS1, our analysis predicted a pseudoknot, in which the loop of the hairpin base paired with 5′ upstream nucleotides. PCNV RFS2 on the other hand likely existed as a pseudoknot or formed a simple hairpin (*SI Appendix*, Fig. S9). Last, both WPNV RFS1 and RFS2 folded into long stem loops with a bulge (*SI Appendix*, Fig. S10).

Dual-Fluorescent Reporter Probing of Dual Ribosomal Frameshifting Sites. To provide experimental evidence for −1 ribosomal frameshifting in nidoviruses with two −1PRF sites and to test whether the putative frameshift sites are functional in cells, we have employed the dual fluorescence reporter assay that we developed previously (*SI Appendix, Extended Materials and Methods*) (76, 77). To generate dual-fluorescence reporter constructs, approximately

150 nt fragments of each nidoviral genome spanning the putative programmed ribosome frameshift sites and secondary structural elements were placed between the coding sequence of EGFP and mCherry (see *Methods* for details, Fig. 6 *A* and *B* and Dataset S5). Accordingly, we calculated the site efficiency of frameshifting in viruses of reModes 3 (WPNV, PCNV, and HCNV) and 5b (CGNV), and against several controls that worked as expected (Fig. 6*C*). Two sites in CGNV, RFS1 and RFS2, led to 0.03% and 2.3% frameshift, respectively. The observed low and biologically implausible efficiency of frameshifting at RFS1 may be due to virus or host factors not available in the tested systems or regulatory signals located outside the genome region studied. Likewise, these factors might upregulate the low frameshifting of 2.3% at RFS2 of CGNV, although this efficiency would read as 97.8% efficiency for translations of an ORF1b equivalent. In the reMode 3 viral sequences significant levels of frameshift were observed at least in one of the frameshift sites (Fig. 6*D*). PCNV RFS1 and RFS2 had modest frameshift levels of 23.1% and 11.5%, respectively. Phylogenetically sister WPNV and HCNV were similar in showing little frameshifting at RFS1, 6.6% and 1.53%, respectively, and high frameshifting at RFS2, 39.9% and 83%, respectively. The latter is among the highest in vivo measured frameshift efficiencies reported so far in viral sequences (Fig. 6*D*) (78, 79). Of note, the amount of frameshift varied widely among the different viral mRNAs. This aligns with the notion that the RNA structure directly flanking the slippery site (Fig. 6*A*) not only contributes to increased levels of frameshifting, but plays a direct role in fine-tuning the exact amount of –1PRF needed by the virus to regulate the production of viral proteins at the correct stoichiometry from each reading frame.

The Nidovirus with the Largest RNA Genome of 64 kb. CGNV has the largest known RNA virus genome and identification of its segment 2 was most challenging. To connect segment 2 to the replicase-encoding segment 1 of CGNV (*SI Appendix*, *Extended Materials and Methods*), we obtained several lines of evidence. First, we detected significant sequence similarity of a coding region on CGNV segment 2 with viral glycoproteins (Fig. 2 and Dataset S6). Second, no RNA or DNA polymerase is encoded on segment 2, suggesting that it is part of a segmented genome and not the unsegmented genome of another virus that had infected the same animal. Third, an analysis of transcription signals and read depth (see below) is in full agreement with the expression of segment 2-encoded proteins via subgenomic RNAs. And fourthly, the size of segment 2 matches the size that is expected from a regression analysis of nidovirus genome size vs. the size of the subgenomic module (part of the genome that is expressed from subgenomic RNA) (*SI Appendix*, Fig. S6*C*).

The CGNV proteome has features not found in other nidoviruses. The unique RNA MT insertion in the NiRAN domain (*SI Appendix*, Fig. S3) may affect regulation of RNA synthesis. Moreover, the CGNV genome encodes homologs of MIEAP, LRR, and SH3, the last two of which are frequently involved in protein–protein interactions to mediate signaling in cellular processes including innate immunity (80, 81). Segment-2-encoded proteins include CPPro followed by two divergent copies of nairo/hantavirus G2 homologs. Together, these features may be linked to promoting genome expansion.

The dual frameshifting mechanism in the reMode 3 viruses considered above results in elongation of translation of the viral genomic RNA upon –1PRF at both sites to express the RdRp and the other replicase proteins. A similar but distinct two-site frameshifting mechanism is employed by the sole member of reMode 5b viruses, CGNV. Here, –1PRF at the first site located within the overlap of the two ORFs encoded by segment 1 results in translation progression, while –1PRF at the second site located in the N-terminal NiRAN coding region terminates translation of a ribosome fraction (Fig. 2 and *SI Appendix*, Fig. S7). This shifting of ribosomes out of the coding reading frame at the second CGNV site to regulate the stoichiometries of replicase and other nonstructural proteins resembles the mechanism described for PSCNV of reMode 4 (17). Compared to all other nidoviruses, both CGNV and PSCNV also

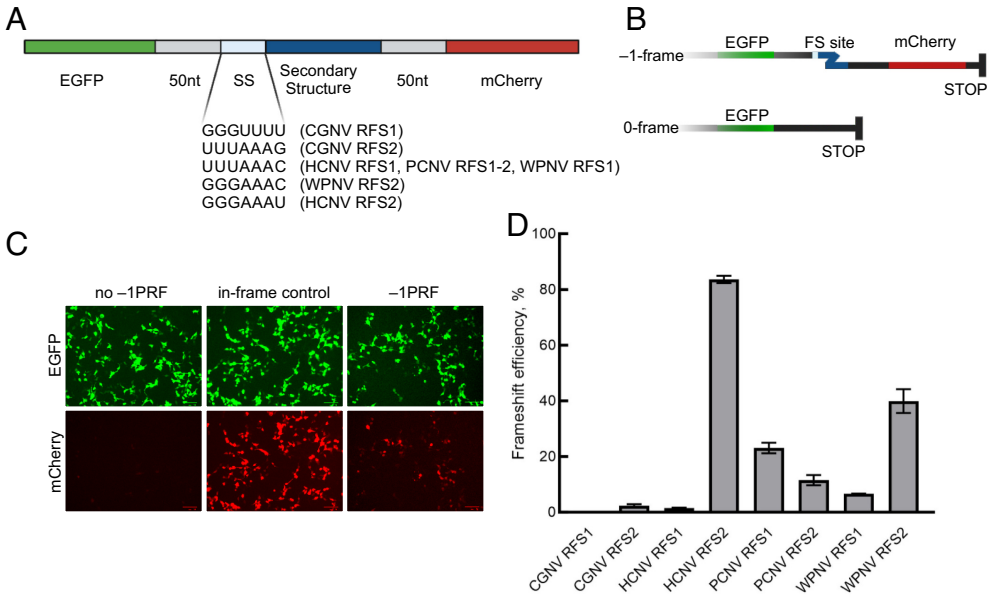

Fig. 6. Dual ribosomal frameshifting efficiency in four nidoviruses. (*A*) Schematic representation of the dual-fluorescence frameshift reporters. Sequences coding for the EGFP and mCherry are linked by a fragment of the respective frameshift site. Frameshift sequences are composed of the SS (slippery site), secondary structure, and included 50 nucleotides upstream and downstream of the putative nidovirus frameshift site. (*B*) EGFP and mCherry are separated by a self-cleaving 2A peptide as well as by a stop codon in-frame with EGFP. As a result, 0-frame translation would produce only EGFP, whereas –1PRF would produce both EGFP and mCherry. The ratio of mCherry to EGFP fluorescence is used for quantification. (*C*) Confocal microscopy images of cells transfected with the EGFP-mCherry control construct lacking frameshift signal (no –1PRF), an in-frame control, and –1PRF constructs. (*D*) Comparison of relative frameshift efficiency of different nidovirus frameshift sites. Data points represent the mean ± SD (n = 3 independent experiments).

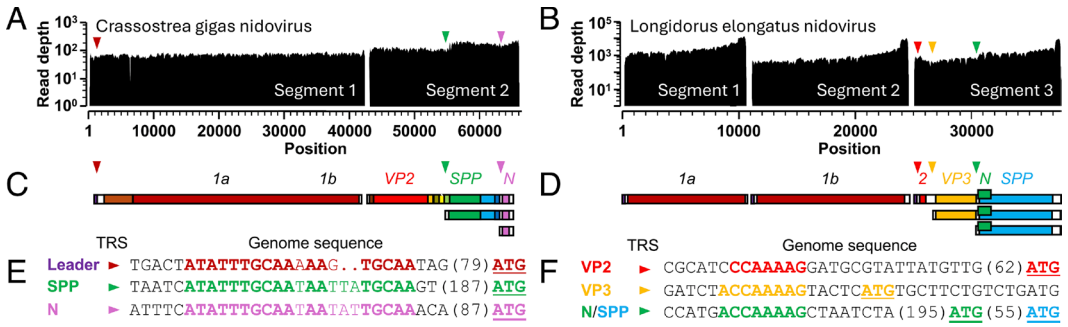

Fig. 7. Evidence for subgenomic RNA production in two segmented nidoviruses. Sequence read depth maps (*A* and *B*), maps of genome coding capacity (*C* and *D*), and putative transcription-regulatory sequences (*E* and *F*) are shown for CGNV (reMode 5b; panels *A*, *C*, and *E*) and LENV (reMode 2; panels *B*, *D*, and *F*). Inferred start codons are underlined and color-coded to match the genome maps. Abbreviations: 1a, approximate position of coding region equivalent to ORF1a; 1b, approximate position of coding region equivalent to ORF1b; 2, ORF2-encoded protein; VP2, virus protein 2; VP3, virus protein 3; SPP, structural polyprotein with similarity to that of alphaviruses; N, nucleoprotein-like basic soluble protein.

have the RdRp (and other key replicase enzymes) positioned most distant in the synthesized polyprotein at more than 12,000 amino acids C-terminal to the polyprotein N terminus (Fig. 4 *C* and *D*).

A Nidovirus with a Trisegmented Genome. The sole member of reMode 2 viruses, LENV from a nematode, has a trisegmented 37.3 kb genome which encodes ORF1a1, ORF1a2/ORF1b, and the structural ORFs on different segments (Fig. 2). High read coverage depth values (see below) and conserved sequence termini that are shared among the three segments and which end with expected poly(A) tails in case of the 3′-ends show that the trisegmentation of the genome is authentic (*SI Appendix*, Fig. S11), making LENV currently the sole known nidovirus with three genome segments. This trisegmented genomic organization enables differential regulation of viral protein amounts via segment-specific copy numbers.

The LENV RdRp-encoding polyprotein of segment 2 is the shortest among the ExoN-encoding invertebrate nidoviruses, and its gene size of 9.5 kb is comparable to those of nidoviruses with two-fold smaller single-segment genomes (37 kb vs. 19 kb; see Fig. 4 *A* and *C*). This size parallel may be of functional significance, since LENV and the 20.3 kb genome bovine nidovirus of the family *Tobaniviridae* (82) are the only nidoviruses lacking an otherwise conserved OMT domain at the C-terminus of this polyprotein. Strikingly, LENV encodes an OMT homolog in the single large ORF on segment 1 (Fig. 2), in apparent support of segments 1 and 2 belonging to the same virus. This OMT is flanked by two NADARs, and OMT-NADAR linkage is found in three other nidoviruses (*SI Appendix*, Fig. S3).

Moreover, the LENV ZBD has replacements of several Cys/His residues that are otherwise conserved across nidoviruses. For the closest known relative of LENV, Allodorylaimus sp. nidovirus (Fig. 3), we only reconstructed a short genome fragment that does not include the ZBD (*SI Appendix*, Fig. S2). A wider nidovirus sampling in nematodes is therefore needed to study whether and how these Cys/His substitutions in the ZBD may affect its Zinc-binding and helicase regulatory function (83).

Transcriptional Analysis of Noncanonical reMode Viruses. SRA datasets containing the identified nidovirus genomic sequences were further examined for three types of RNA data commonly associated with subgenomic transcription in nidoviruses: a) presence of putative leader and body TRS upstream of important structural genes, b) stepwise increases in sequence read depth associated with production of one or more 3′-coterminal sgRNA species, and c) reads joining the genomic leader sequence to a downstream body sequence at a TRS (*SI Appendix*, *Extended Materials and Methods*). WPNV (reMode 3) and CSNV (reMode 4) both show polar read depth increases as well as presence of putative leader-body-junction reads consistent with canonical nidovirus subgenomic RNA production (*SI Appendix*, Fig. S12). In contrast, CGNV (reMode 5b) shows stepwise read depth increases at two putative TRS, but does not contain any sequences that would be consistent with leader addition (Fig. 7). CGNV may therefore produce leaderless sgRNA species, as reported for two other previously characterized invertebrate nidoviruses, GAV (84) and AAbV (39), but with leader TRS at the 5′-end of segment 1, and body TRS on segment 2 (Fig. 7). LENV (reMode 2) does not show marked read depth increases at putative TRS upstream of three inferred structural genes on segment 3. Notably, and uniquely among all known multisegment nidoviruses, all three segments of LENV contained a near-identical 206 nt sequence at the 5′ termini (*SI Appendix*, Fig. S11), suggestive of a new expression strategy that may be related to canonical nidovirus discontinuous transcription, leader-switching (84–86), or both (Fig. 7). Taken together with results for reMode 1 viruses (87, 88) and reMode 6 (39), there is evidence for production of 3′-coterminal sgRNA species, a hallmark of canonical nidovirus replication, across most nidovirus reModes.

Conclusions

Data-Driven Virus Discovery reveals several novel evolutionary trajectories taken by different nidoviruses during evolution of the largest and most complex RNA genomes known. Nidoviruses therefore provide a unique window into the evolution and function of giant RNA genomes of sizes that seem inaccessible to other viruses. In all known nidoviruses using proofreading ExoN, we document restricted RNA genome size ranges that are associated with genome segmentation, genomic organization, translation elongation, proteome content, or infected animal host in addition to phylogenetic family-based constraints established before (3). The genome size range of invertebrate nidoviruses is more than twice as large as that of vertebrate nidoviruses, which is correlated with the respective diversities and time scales of host evolution.

In the current dataset, invertebrates are represented by only a subset of known taxa. The total number of described (invertebrate) host taxa may therefore be a more informative denominator for estimating the global nidovirus diversity. Our and previous large-scale studies did not discover genuine non-metazoan-infecting nidoviruses, suggesting that contemporary nidoviruses are restricted to infect animals.

The nidovirus proteome is larger and more diverse than previously thought, which includes protein domains that currently cannot be functionally annotated with sequence- or structure-based

methods due to a lack of any appreciable similarity to known proteins. The insertions of putative nucleotide-dependent and RNA-processing domains, including REXA, in ORF1a and ORF1b in many of the nidovirus genomes and their genomic segregation with key replicative proteins suggest that they are cofactors of viral replication. Host homologs of these domains may be probed as potential transient cofactors of the RTC for other nidoviruses that do not encode these domains in the genome, including coronaviruses. Other and/or additional roles of the acquired proteins are conceivable, such as interference with the innate immune system of the host, especially for domains inserted upstream of 3CLpro in pp1a/pp1ab.

Multiple nidovirus lineages have independently reached states of high genetic complexity by evolving exceptionally large genomes and sophisticated translational control mechanisms such as dual frameshifting. The function of dual frameshifting in the nidovirus life cycle, the influence of body temperature of invertebrate hosts on frameshifting efficiency, how cis-acting RNA elements lead to such varying levels of frameshifting and whether these sites are additionally regulated by host and viral trans-acting factors will remain to be explored.

Methods

SRA-Based Virus Discovery. Our computational virus discovery workflow and its application to SRA data are described in detail in a previous study (16). It is highly parallelized and was run on a high-performance computing cluster. The workflow involves sensitive detection of raw viral sequencing reads via profile hidden Markov model–based sequence homology searches followed by targeted (seed-based) viral genome assembly. The assembly quality is assess by two metrics that quantify the per-base (minimal coverage) and contig-wide (mean alignment score) accuracy (*SI Appendix, Extended Materials and Methods*) (16).

Functional Annotation of Nidovirus Genomes. We conducted several complementary analyses to functionally characterize protein domains encoded by the newly discovered nidovirus genomes, which included: i) multiple sequence alignment to in-house reference nidovirus genomes and their annotated proteins, ii) profile HMM searches against a custom database of conserved protein domains of nidovirus and other RNA viruses, iii) profile HMM searches using the MPI Bioinformatics Toolkit (https://toolkit.tuebingen.mpg.de), and iv) predicted protein 3D structure-based comparisons (*SI Appendix, Extended Materials and Methods*).

Phylogenetic Analysis. Multiple RdRp protein sequence alignments were built separately for ingroup (nidovirus) and outgroup (picornavirus) sequences and then combined using profile-based alignment (*SI Appendix, Extended Materials and Methods*). Phylogenetic reconstruction was performed using maximum likelihood and Bayesian approaches (*SI Appendix, Extended Materials and Methods*). The applied tools generated trees with very similar topologies and the maximum likelihood version was used for subsequent analyses and visualization.

Sequence-Based Virus Classification. We used a pairwise distance-based approach, DEmARC (62, 63) v1.4, to classify viruses into operational taxonomic units at the family level (*SI Appendix, Extended Materials and Methods*). Since most newly described nidoviruses prototype new species, the reported demarcation at this rank for new nidoviruses is trivial and accurate. For other nidoviruses, we used formally recognized species. We used species taxa (89) in our comparative analyses to correct for highly uneven virus sampling.

Dimethyl Sulfate Mutational Profiling (DMS-MaP) of RNA Structures. We performed in vitro DMS modification of the frameshift RNA elements as described previously (90) and in *SI Appendix, Extended Materials and Methods*, followed by sequencing on an Illumina NovaSeq instrument using paired-end PE250 sequencing (Novogene).To measure the agreement between the predicted base pairing in a model and the experimentally determined DMS reactivity, ROC-AUC (receiver operator characteristic – area under the curve) analysis was carried out.

Assessment of Ribosomal Frameshifting in Cells Using Dual Fluorescence Reporters. To generate dual-fluorescence reporter constructs to assess frameshift efficiencies, -1 programmed frameshift sites of each nidovirus were placed between the coding sequence of EGFP and mCherry such that EGFP would be produced in 0-frame and mCherry in –1-frame (76, 77). An in-frame 100% mCherry control for each vector was generated to normalize the mCherry fluorescence signal. HEK293 cells were transfected and harvested at 24 h posttransfection (*SI Appendix, Extended Materials and Methods*). Flow cytometry was performed on a NovoCyte Quanteon (ACEA) instrument, and frameshifting efficiency was quantified (*SI Appendix, Extended Materials and Methods*).

Data, Materials, and Software Availability. Transcriptomic data used in this study are publicly available through the Sequence Read Archive repository (44). Viral genome sequences assembled in this study will be made available through the NCBI Genbank database upon publication. The DMS sequencing data produced in this study is available via GEO Accession No.: GSE283188 (91). The Virushunter and Virusgatherer tools are available on GitHub: https://github.com/lauberlab/VirusHunterGatherer (92). All other data are included in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. C.L. presented the virus discovery part and initial data on dual ribosomal frameshifting reported in this paper at the international Nido2023 symposium (Montreux, Switzerland, May 2023). C.L. is supported by the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy - EXC 2155–Project No. 390874280. C.L., S.S., and R.B. acknowledge support by KA1-Co-02 "CoViPa", a grant from the Helmholtz Association's Initiative and Network Fund. N.C. is funded by the Helmholtz Association and European Research Council–Project No. 948636. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank all colleagues in the scientific community who make their sequencing data publicly accessible. We acknowledge the National Center for Biotechnology Information for providing an elaborate platform to exchange sequencing data. The authors gratefully acknowledge the computing time made available to them on the high-performance computers Taurus, Romeo, Julia and Barnard at the NHR (Nationales Hochleistungsrechnen an Hochschulen) Center NHR@TUD of the University of Technology Dresden.

Author affiliations: [a]Department of Biology, Microbial Pathogenesis and Immunity, Texas A&M University, College Station, TX 77840; [b]Helmholtz Institute for RNA-Based Infection Research, Helmholtz Centre for Infection Research, Würzburg 97080, Germany; [c]Institut de Biologie Moléculaire et Cellulaire, Architecture et Réactivité de l'ARN, Université de Strasbourg, Strasbourg 67084, France; [d]Division of Virus-Associated Carcinogenesis (F170), German Cancer Research Center, Heidelberg 69120, Germany; [e]Medical Faculty Heidelberg, Department of Infectious Diseases, Molecular Virology, Heidelberg University, Center for Integrative Infectious Disease Research, Heidelberg 69120, Germany; [f]Institute for Experimental Virology, TWINCORE Centre for Experimental and Clinical Infection Research, a Joint Venture between the Hannover Medical School and the Helmholtz Centre for Infection Research, Hannover 30625, Germany; [g]Cluster of Excellence 2155 RESIST, Hannover 30625, Germany; [h]Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow 119899, Russia; [i]Leiden University Center of Infectious Diseases, Leiden University Medical Center, Leiden 2333 ZA, The Netherlands; and [j]Department of Biochemistry III, University of Regensburg, Regensburg 93053, Germany

1. Y. I. Wolf *et al.*, Origins and evolution of the global RNA virome. *mBio* **9**, e02329-18 (2018).
2. N. A. O'Leary *et al.*, Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
3. C. Lauber, A. E. Gorbalenya, "Taxonomy advancement and genome size change: Two perspectives on RNA virus genetic diversity" in *Virus Evolution: Current Research and Future Directions* (Caister Academic Press, 2016), pp. 215–232.
4. Y. I. Wolf *et al.*, Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat. Microbiol.* **5**, 1262–1270 (2020).
5. R. C. Edgar *et al.*, Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142–147 (2022).
6. U. Neri *et al.*, Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023–4037.e18 (2022).
7. A. A. Zayed *et al.*, Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156–162 (2022).
8. E. Domingo, J. J. Holland, RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.* **51**, 151–178 (1997).

9. M. Combe, R. Sanjuán, Variation in RNA virus mutation rates across host cells. *PLoS Pathog.* **10**, e1003855 (2014).
10. E. C. Holmes, Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol.* **11**, 543–546 (2003).
11. J. Matthijnssens *et al.*, ICTV virus taxonomy profile: Sedoreoviridae 2022: This article is part of the ICTV Virus Taxonomy Profiles collection. *J. Gen. Virol.* **103**, 001782 (2022).
12. J. Matthijnssens *et al.*, ICTV virus taxonomy profile: Spinareoviridae 2022: This article is part of the ICTV Virus Taxonomy Profiles collection. *J. Gen. Virol.* **103**, 001781 (2022).
13. E. J. Snijder *et al.*, Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J. Mol. Biol.* **331**, 991–1004 (2003).
14. E. C. Smith, N. R. Sexton, M. R. Denison, Thinking outside the triangle: Replication fidelity of the largest RNA viruses. *Annu. Rev. Virol.* **1**, 111–132 (2014).
15. P. T. Nga *et al.*, Discovery of the first insect nidovirus, a missing evolutionary link in the emergence of the largest RNA virus genomes. *PLoS Pathog.* **7**, e1002215 (2011).
16. C. Lauber *et al.*, Deep mining of the sequence read archive reveals major genetic innovations in coronaviruses and other nidoviruses of aquatic vertebrates. *PLoS Pathog.* **20**, e1012163 (2024).
17. A. Saberi, A. A. Gulyaeva, J. L. Brubacher, P. A. Newmark, A. E. Gorbalenya, A planarian nidovirus expands the limits of RNA genome size. *PLoS Pathog.* **14**, e1007314 (2018).
18. K. Bukhari *et al.*, Description and initial characterization of metatranscriptomic nidovirus-like genomes from the proposed new family Abyssoviridae, and from a sister group to the Coronavirinae, the proposed genus Alphaletovirus. *Virology* **524**, 160–171 (2018).
19. M. A. Brinton *et al.*, ICTV virus taxonomy profile: Arteriviridae 2021. *J. Gen. Virol.* **102**, 001632 (2021).
20. X. Hou *et al.*, Using artificial intelligence to document the hidden RNA virosphere. *Cell* **187**, 6929–6942 (2024), 10.1016/j.cell.2024.09.027.
21. R. K. French *et al.*, Host phylogeny shapes viral transmission networks in an island ecosystem. *Nat. Ecol. Evol.* **7**, 1834–1843 (2023).
22. M. Krupovic, V. V. Dolja, E. V. Koonin, Origin of viruses: Primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.* **17**, 449–458 (2019).
23. A. E. Gorbalenya *et al.*, The species severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5**, 536–544 (2020).
24. G. J. Mordecai *et al.*, Endangered wild salmon infected by newly discovered viruses. *eLife* **8**, e47615 (2019).
25. L. L. Hoon-Hanks *et al.*, Serpentovirus (Nidovirus) and Orthoreovirus coinfection in captive veiled chameleons (*Chamaeleo calyptratus*) with respiratory disease. *Viruses* **12**, 1329 (2020).
26. C. Lauber *et al.*, The footprint of genome architecture in the largest genome expansion in RNA viruses. *PLoS Pathog.* **9**, e1003500 (2013).
27. I. Brierley *et al.*, An efficient ribosomal frame-shifting signal in the polymerase-encoding region of the coronavirus IBV. *EMBO J.* **6**, 3779–3785 (1987).
28. E. P. Plant *et al.*, A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol.* **3**, e172 (2005).
29. A. E. Firth, I. Brierley, Non-canonical translation in RNA viruses. *J. Gen. Virol.* **93**, 1385–1409 (2012).
30. P. V'kovski, A. Kratzel, S. Steiner, H. Stalder, V. Thiel, Coronavirus biology and replication: Implications for SARS-CoV-2. *Nat. Rev. Microbiol.* **19**, 155–170 (2021).
31. B. Malone, N. Urakova, E. J. Snijder, E. A. Campbell, Structures and functions of coronavirus replication–transcription complexes and their relevance for SARS-CoV-2 drug design. *Nat. Rev. Mol. Cell Biol.* **23**, 21–39 (2022).
32. Z. Lou, Z. Rao, The life of SARS-CoV-2 inside cells: Replication-transcription complex assembly and function. *Annu. Rev. Biochem.* **91**, 381–401 (2022).
33. E. Grellet, I. L'Hôte, A. Goulet, I. Imbert, Replication of the coronavirus genome: A paradox among positive-strand RNA viruses. *J. Biol. Chem.* **298**, 101923 (2022).
34. S. G. Sawicki, D. L. Sawicki, S. G. Siddell, A contemporary view of coronavirus transcription. *J. Virol.* **81**, 20–29 (2007).
35. H. Di *et al.*, Expanded subgenomic mRNA transcriptome and coding capacity of a nidovirus. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E8895–E8904 (2017).
36. R. Madhugiri, M. Fricke, M. Marz, J. Ziebuhr, Coronavirus cis-acting RNA elements. *Adv. Virus Res.* **96**, 127–163 (2016).
37. F. Zirkel *et al.*, Identification and characterization of genetically divergent members of the newly established family Mesoniviridae. *J. Virol.* **87**, 6346–6358 (2013).
38. D. X. Liu, T. S. Fung, K.K.-L. Chong, A. Shukla, R. Hilgenfeld, Accessory proteins of SARS-CoV and other coronaviruses. *Antiviral Res.* **109**, 97–109 (2014).
39. N. S. Kron *et al.*, Expression dynamics of the aplysia abyssovirus. *Virology* **589**, 109890 (2024).
40. N. Sittidilokratna, S. Dangtip, J. A. Cowley, P. J. Walker, RNA transcription analysis and completion of the genome sequence of yellow head nidovirus. *Virus Res.* **136**, 157–165 (2008).
41. A. A. Gulyaeva, A. E. Gorbalenya, A nidovirus perspective on SARS-CoV-2. *Biochem. Biophys. Res. Commun.* **538**, 24–34 (2021).
42. F. Ferron, B. Sama, E. Decroly, B. Canard, The enzymes for genome size increase and maintenance of large (+)RNA viruses *Trends Biochem. Sci.* **46**, 866–877 (2021).
43. C. Lauber, S. Seitz, Opportunities and challenges of data-driven virus discovery. *Biomolecules* **12**, 1073 (2022).
44. R. Leinonen, H. Sugawara, M. Shumway, International Nucleotide sequence database collaboration, The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
45. C. Lauber *et al.*, Deciphering the origin and evolution of hepatitis B viruses by means of a family of non-enveloped fish viruses. *Cell Host Microbe* **22**, 387–399.e6 (2017).
46. C. Lauber, M. Seifert, R. Bartenschlager, S. Seitz, Discovery of highly divergent lineages of plant-associated astro-like viruses sheds light on the emergence of potyviruses. *Virus Res.* **260**, 38–48 (2019).
47. L. C. Chong, C. Lauber, Viroid-like RNA-dependent RNA polymerase-encoding ambiviruses are abundant in complex fungi. *Front. Microbiol.* **14**, 1144003 (2023).
48. A. E. Gorbalenya, L. Enjuanes, J. Ziebuhr, E. J. Snijder, Nidovirales: Evolving the largest RNA virus genome. *Virus Res.* **117**, 17–37 (2006).
49. M. E. Petrone *et al.*, A ~40-kb flavi-like virus does not encode a known error-correcting mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2403805121 (2024).
50. S. Blanck, J. Ziebuhr, Proteolytic processing of mesonivirus replicase polyproteins by the viral 3C-like protease. *J. Gen. Virol.* **97**, 1439–1445 (2016).
51. H. K. Choi *et al.*, Structure of Sindbis virus core protein reveals a chymotrypsin-like serine proteinase and the organization of the virion. *Nature* **354**, 37–43 (1991).
52. P. Melancon, H. Garoff, Processing of the Semliki Forest virus structural polyprotein: Role of the capsid protease. *J. Virol.* **61**, 1301–1309 (1987).
53. L. M. Iyer, A. M. Burroughs, V. Anantharaman, L. Aravind, Apprehending the NAD+–ADPr-dependent systems in the virus world. *Viruses* **14**, 1977 (2022).
54. E. Van Den Born *et al.*, Viral AlkB proteins repair RNA damage by oxidative demethylation. *Nucleic Acids Res.* **36**, 5451–5461 (2008).
55. N. Luhtala, R. Parker, T2 family ribonucleases: Ancient enzymes with diverse roles. *Trends Biochem. Sci.* **35**, 253–259 (2010).
56. B. L. Stoddard, Homing endonuclease structure and function. *Q. Rev. Biophys.* **38**, 49 (2006).
57. A. Munir, A. Banerjee, S. Shuman, NAD+-dependent synthesis of a 5′-phospho-ADP-ribosylated RNA/DNA cap by RNA 2′-phosphotransferase Tpt1. *Nucleic Acids Res.* **46**, 9617–9624 (2018).
58. J. Y. Lee *et al.*, Crystal structure of NAD+-dependent DNA ligase: Modular architecture and functional implications. *EMBO J* **19**, 1119–1128 (2000).
59. J. E. Walker, M. Saraste, M. J. Runswick, N. J. Gay, Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J* **1**, 945–951 (1982).
60. M. Shi *et al.*, Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the Flaviviridae and related viruses. *J. Virol.* **90**, 659–669 (2016).
61. M. F. Thomas, N. D. L'Etoile, K. M. Ansel, Eri1: A conserved enzyme at the crossroads of multiple RNA-processing pathways. *Trends Genet.* **30**, 298–307 (2014).
62. C. Lauber, A. E. Gorbalenya, Partitioning the genetic diversity of a virus family: Approach and evaluation through a case study of picornaviruses. *J. Virol.* **86**, 3890–3904 (2012).
63. C. Lauber, A. E. Gorbalenya, Toward genetics-based virus taxonomy: Comparative analysis of a genetics-based classification and the taxonomy of picornaviruses. *J. Virol.* **86**, 3905–3915 (2012).
64. C. B. Buck *et al.*, Widespread horizontal gene transfer among animal viruses. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.03.25.586562 (Accessed 7 May 2024).
65. T. Noda, Selective genome packaging mechanisms of influenza A viruses. *Cold Spring Harb. Perspect. Med.* **11**, a038497 (2021).
66. A. Borodavka, U. Desselberger, J. T. Patton, Genome packaging in multi-segmented dsRNA viruses: Distinct mechanisms with similar outcomes. *Curr. Opin. Virol.* **33**, 106–112 (2018).
67. E. B. Kramer, P. J. Farabaugh, The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA* **13**, 87–96 (2007).
68. H. S. Zaher, R. Green, Fidelity at the molecular level: Lessons from protein synthesis. *Cell* **136**, 746–762 (2009).
69. R. M. Voorhees, V. Ramakrishnan, Structural basis of the translational elongation cycle. *Annu. Rev. Biochem.* **82**, 203–236 (2013).
70. K. Mohler, M. Ibba, Translational fidelity and mistranslation in the cellular response to stress. *Nat. Microbiol.* **2**, 17117 (2017).
71. Y. Li *et al.*, Transactivation of programmed ribosomal frameshifting by a viral protein. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E2172–E2181 (2014).
72. Y. Fang *et al.*, Efficient −2 frameshifting by mammalian ribosomes to synthesize an additional arterivirus protein. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E2920–E2928 (2012).
73. G. Wolff *et al.*, A molecular pore spans the double membrane of the coronavirus replication organelle. *Science* **369**, 1395–1398 (2020).
74. T. Fisher *et al.*, Parsing the role of NSP1 in SARS-CoV-2 infection. *Cell Rep.* **39**, 110954 (2022).
75. K. Schubert *et al.*, Universal features of Nsp1-mediated translational shutdown by coronaviruses. *Mol. Cell* **83**, 3546–3557.e8 (2023).
76. P. Bohn, A.-S. Gribling-Burrer, U. B. Ambi, R. P. Smyth, Nano-DMS-MaP allows isoform-specific RNA structure determination. *Nat. Methods* **20**, 849–859 (2023).
77. M. M. Zimmer *et al.*, The short isoform of the host antiviral protein ZAP acts as an inhibitor of SARS-CoV-2 programmed ribosomal frameshifting. *Nat. Commun.* **12**, 7193 (2021).
78. R. J. Riegger, N. Caliskan, Thinking outside the frame: Impacting genomes capacity by programmed ribosomal frameshifting. *Front. Mol. Biosci.* **9**, 842261 (2022).
79. L. K. Finch *et al.*, Characterization of ribosomal frameshifting in Theiler's murine encephalomyelitis virus. *J. Virol.* **89**, 8580–8589 (2015).
80. F. M. Ausubel, Are innate immune signaling pathways in plants and animals conserved? *Nat. Immunol.* **6**, 973–979 (2005).
81. J. D. Scott, T. Pawson, Cell signaling in space and time: Where proteins come together and when they're apart. *Science* **326**, 1220–1224 (2009).
82. R. Tokarz *et al.*, Discovery of a novel nidovirus in cattle with respiratory disease. *J. Gen. Virol.* **96**, 2188–2193 (2015).
83. J. Chen *et al.*, Ensemble cryo-EM reveals conformational states of the nsp13 helicase in the SARS-CoV-2 helicase replication–transcription complex. *Nat. Struct. Mol. Biol.* **29**, 250–260 (2022).
84. J. A. Cowley, C. M. Dimmock, P. J. Walker, Gill-associated nidovirus of Penaeus monodon prawns transcribes 3′-coterminal subgenomic mRNAs that do not possess 5′-leader sequences. *J. Gen. Virol.* **83**, 927–935 (2002).
85. S. Makino, S. A. Stohlman, M. M. Lai, Leader sequences of murine coronavirus mRNAs can be freely reassorted: Evidence for the role of free leader RNA in transcription. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 4204–4208 (1986).
86. S. Makino, M. M. Lai, High-frequency leader sequence switching during coronavirus defective interfering RNA replication. *J. Virol.* **63**, 5285–5292 (1989).
87. K. Stirrups *et al.*, Leader switching occurs during the rescue of defective RNAs by heterologous strains of the coronavirus infectious bronchitis virus. *J. Gen. Virol.* **81**, 791–801 (2000).
88. D. Kim *et al.*, The architecture of SARS-CoV-2 transcriptome. *Cell* **181**, 914–921.e10 (2020).
89. A. E. Gorbalenya, S. G. Siddell, Recognizing species as a new focus of virus research. *PLoS Pathog.* **17**, e1009318 (2021).
90. L. Pekarek *et al.*, Cis-mediated interactions of the SARS-CoV-2 frameshift RNA alter its conformations and affect function. *Nucleic Acids Res.* **51**, 728–743 (2023).
91. R. P. Smyth, DMS sequencing data from "Giant RNA genomes: Roles of host, translation elongation, genome architecture, and proteome in nidoviruses". Gene Expression Omnibus (GEO). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE283188. Deposited 29 November 2024.
92. L. Chuprikova, L. C. Chong, S. Ruff, S. Seitz, C. Lauber, Virushunter and virusgatherer. GitHub. https://github.com/lauberlab/VirusHunterGatherer. Deposited 8 September 2020.