CHAPTER 6

The footprint of genome architecture in the largest genome expansion in RNA viruses

Chris Lauber
Jelle J. Goeman
Maria del Carmen Parquet
Phan Thi Nga
Eric J. Snijder
Kouichi Morita
Alexander E. Gorbalenya

manuscript in preparation

Abstract

Small genome sizes of RNA viruses (2 to 32kb) have been linked to the high mutation rate during RNA replication that is thought to lack proof-reading. This paradigm is now being reviewed owing to the discovery of a 3'-to-5'exoribonuclease (ExoN) in nidoviruses, a monophyletic group of viruses with non-segmented, single-stranded RNA genomes of positive polarity and conserved genome architecture. The ExoN, homolog of a canonical DNA proof-reading enzyme, is exclusively encoded by nidoviruses with genomes larger than 20 kb. All other known non-segmented RNA viruses employ smaller genomes. Here we use evolutionary analyses to show that the two- to three-fold expansion of the nidovirus genome was accompanied by a vast amount of replacements in conserved proteins at the scale observed in the Tree of life. To unravel common patterns of such genetically diverse viruses, we exploited functional conservation of the nidovirus genome architecture. This conservation allowed us to partition each genome into five spatially collinear regions in an alignment-free manner. Each genomic region was analyzed for its contribution to genome size change under both linear and non-linear conditions. The non-linear model statistically outperformed the linear one and captured >92% of data variation. Accordingly, individual nidoviruses were found to have reached different points on a common expansion trajectory dominated by three consecutive, region-specific size increases. Our findings indicate a hierarchical relation between the three involved genome regions that are distinguished by expression mechanism. In the order of size increase these regions predominantly control genome replication, genome expression, and virus dissemination, respectively. In contrast to the observed directionality in the evolutionary dimension these fundamental biological processes cooperate bi-directionally on a functional level in the virus life cycle. Collectively, our findings suggest that genome architecture and the associated division of labor control genome size and may set its limits in RNA viruses.

Author Summary

RNA viruses include many major pathogens. Virus adaptation to their hosts is facilitated by fast mutation and constrained by small genome sizes, which are both due to the extremely high error rate of viral polymerases. Using an innovative computational approach we now provide evidence for additional forces that may control genome size and, consequently, affect virus adaptation to the host. We analyzed nidoviruses, a monophyletic group of viruses that populate the upper ~60% of the RNA virus genome size scale, evolved a conserved genomic architecture, and infect vertebrate and inverterbrate species. They include viruses with the largest known RNA genomes that exclusively encode a 3'-to-5'exoribonuclease, homolog of a canonical DNA proof-reading enzyme, which improves the replication fidelity. We show that the evolutionary space explored by these viruses exceeds that of the Tree of life for comparable protein datasets, although the time-scale of nidovirus evolution remains unknown. Extant nidoviruses with different genome sizes reached particular points on a common non-linear genome expansion trajectory. This trajectory may be shaped by the division of labor between open reading frames that predominantly control genome replication, genome expression, and virus dissemination, respectively. Ultimately, genomic architecture may determine the observed limit of genome size in contemporary RNA viruses.

Introduction

Genome size is a net result of evolution driven by the environment, mutation, and the genetics of the organism[308,442]. Particularly, mutation rate is a powerful evolutionary factor[116]. The relation between mutation rate and genome size is inversely proportional for a range of life forms from viroids to viruses to bacteria, and it is slightly positive for eukaryotes, suggestive a causative link[155,307,431]. The genome size of RNA viruses is restricted to a range of ~2-to-32 kb that corresponds to a very narrow band on the genome size scale from 1 kb to 10 Mb at which genome size increase is strongly correlated with mutation rate decrease[404]. This restricted genome size range of RNA viruses is believed to be a consequence of the lack of proof-reading factors resulting in a low fidelity of RNA replication[220,439]. In the above relation, mutation rate and proof-reading serve as a proxy for replication fidelity and genetic complexity, respectively. When combined, replication fidelity, genome size and genetic complexity form the unidirectional triangular relation that was postulated to lock these characteristics in low states in primitive self-replicating molecules[131]. The applicability of this trapping, known as the "Eigen paradox" [276], was also extended to RNA viruses[217]. Recent studies of the order *Nidovirales*, a large group of RNA viruses including those with the largest known genomes, provided strong support for the triangular relation and, unexpectedly, revealed a way of how the Eigen paradox could have been

solved by these viruses[336,432]. These advances established nidoviruses as a prime model for studying genome size evolution in RNA viruses.

The order *Nidovirales* unites viruses with enveloped virions and non-segmented single-stranded RNA genomes of positive polarity (ssRNA+), whose replication is mediated by cognate RNA-depended RNA polymerase (RdRp)[91,360]. The order includes four families - the *Arteriviridae* and *Coronaviridae* (including vertebrate, mostly mammal viruses), and the *Roniviridae* and provisional Mesoniviridae (invertebrate viruses). The unusually broad 12.7-to-31.7 kb genome size range of this monophyletic group of viruses includes the largest known RNA genomes that are employed by viruses from the families *Roniviridae* (~26 kb)[85] and *Coronaviridae* (from 26.3 to 31.7 kb)[90], collectively coined large-sized nidoviruses[174]. Viruses from the *Arteriviridae* (with 12.7-to-15.7 kb genome range)[140] and the recently identified Mesoniviridae (20.2 kb)[284] are considered small-sized and intermediate-sized nidoviruses, respectively. Nidoviruses share a conserved genomic architecture with multiple open reading frames (ORFs) that are flanked by two untranslated regions (UTRs)[49,84,98,336,500]. The two 5'-most ORFs 1a and 1b overlap by a few dozen nucleotides and are translated directly from the genomic RNA to produce polyproteins 1a (pp1a) and pp1ab, the latter involving a -1 ribosomal frameshift (RFS) event[55,366]. The pp1a and pp1ab are autoproteolytically processed to non-structural proteins (nsp), from nsp1 to nsp12 in arteriviruses and from nsp1 to nsp16 in coronaviruses (reviewed in [498]). They encode most components of the membrane-bound replication-transcription complex (RTC)[100,421,462] that mediates genome replication and the synthesis of subgenomic RNAs (known also as transcription)[409,450]. ORF1a encodes proteases for processing of pp1a and pp1ab (reviewed in [498]), trans-membrane domains/proteins (TM1, TM2 and TM3) anchoring the RTC[22,200] and numerous poorly characterized proteins. ORF1b encodes core enzymes of the RTC (see below). Other ORFs, whose number varies considerably among nidoviruses, are located immediately downstream of ORF1b and are expressed from 3'-coterminal subgenomic mRNAs (hereafter collectively referred to as 3'ORFs)[408]. They encode virion and, optionally, so-called "accessory proteins" (reviewed in [53,136,316]).

In addition to the genome architecture, nidoviruses share also an array (synteny) of 6 replicative protein domains. Three domains – an ORF1a-encoded protease with chymotrypsin-like fold (3C-like protease, 3CLpro) [13,27,179], an ORF1b-encoded RdRp[75,179,445] and a superfamily 1 helicase (HEL1)[178,212,417,419] that may form a part or entire protein released from pp1a/pp1ab – represent the most conserved enzymes (reviewed in [169]). For other proteins, a relationship may be established only for some lineages, mostly due to poor sequence similarity. Two tightly correlated properties separate large-sized and intermediate-sized nidoviruses from all other ssRNA+ viruses that form several dozens of families and hundreds species: the genome size exceeding 20 kb and the encoding of a RNA 3'-to-5'exoribonuclease (ExoN)[336]. The latter enzyme is distantly related to a DNA proofreading enzyme, and it is genetically segregated and expressed with RdRp and HEL1[323,432]. Based on these properties ExoN was implicated in improving the fidelity of RNA virus replication.

This hypothesis is strongly supported by an excessive accumulation of mutations in ExoN-defective mutants of two coronaviruses, mouse hepatitis virus[124] and severe acute respiratory syndrome coronavirus (SARS-CoV)[123] (for review see [99]), and the identification of the RNA 3'-end mismatch excision activity in the SARS-CoV nsp10/nsp14 complex[52]. In all likelihood, the on-going characterization of ExoN is expected to reveal the molecular mechanisms that control the fidelity of replication. Regardless of its details, the ExoN acquisition provides the most plausible explanation for the solving of the Eigen paradox with a single evolutionary event that likely liberated the ExoN-encoding nidoviruses for genome expansions beyond the limit observed by other non-segmented ssRNA+ viruses[174,336].

In this study we sought to gain insight into events that led to the emergence of the ExoN-encoding ancestor and for further expansion of the nidovirus genome to sizes threefold the average RNA virus genome size, hereafter referred to as the nidovirus genome expansion (NGE). We show that comparative sequence analysis of nidovirus families are complicated by huge evolutionary distances, at the scale of the Tree of life (ToL), that separate the most conserved proteins. To address this challenge, we exploited functional conservations in the genome architecture that could be established across the nidovirus genome in an alignment-free manner. Consequently we partitioned the genome into five spatially collinear regions. By employing a statistical framework we revealed non-linear, consecutive expansions of the three differentially expressed coding regions (ORF1a, ORF1b, 3'ORFs) that account for 95-99% of the genome. Importantly, these regions predominantly control, respectively, genome replication, genome expression, and virus dissemination, during the virus life cycle. The observed dynamics unveil an evolutionary pathway that accommodated both an enormous accumulation of mutations and virus adaptation to different host species. Our results also indicate that genome architecture and the associated division of labor control the expansion of RNA virus genomes and, contrary to the current paradigm exclusively focusing on replication fidelity, may determine the observed limit on RNA virus genome size.

Results

The scales of per-residue evolutionary change in nidoviruses and the Tree of life are comparable. Nidoviruses have evolved genomes in a size range that accounts for the upper ~60% of the entire RNA virus genome size scale and includes the largest RNA genomes[336]. How much did it take to produce this unprecedented innovation in the RNA virus world? This question could be addressed in two evolutionary dimensions: time and amount of substitutions. Due to both the lack of fossil records and the high viral mutation rate, the time scale of distant relations of RNA viruses remains technically difficult to study. Hence, we sought to estimate the amount of accumulated replacements in conserved nidovirus proteins

and to put it into a biological perspective by comparing it with that accumulated by proteins of cellular species in the ToL.

To this end, we used a rooted phylogeny for a set of 28 nidovirus representatives (Table S1), which is based on a multiple alignment of nidovirus-wide conserved protein regions in the 3CLpro, the RdRp and the HEL1, as described previously[336]. The 28 representatives cover the acknowledged species diversity of nidoviruses with completely sequenced genomes[85,90,140,284] and include two additional viruses. For the arterivirus species Porcine reproductive and respiratory syndrome virus we selected two viruses representing the European and North American types, respectively, because we observed an unusually high divergence of these lineages; for the ronivirus species Gill-associated virus we selected two viruses representing the genotypes gill-associated virus and yellow head virus, respectively, because these viruses showed a genetic distance comparable to that of some coronavirus species (CL & AEG, in preparation). The nidovirus-wide phylogenetic analysis consistently identified the five major lineages: subfamilies *Coronavirinae* and *Torovirinae*, and families *Arteriviridae*, *Roniviridae* and Mesoniviridae. The root was placed at the branch leading to arteriviruses (Fig. 1A) according to outgroup analyses[336]. Accordingly, arteriviruses with genome sizes of 12.7 to 15.7 kb are separated in the tree from other nidoviruses with larger genomes (20.2-31.7 kb).

We compared the evolutionary space explored by nidoviruses, measured in number of substitutions per site in conserved proteins, with that of a single-copy protein dataset representing the ToL[50] (Fig. 1B). Using a common normalized scale of [0,1], comparison of the viral and cellular trees and associated pairwise distance distributions revealed that the distances between cellular proteins (0.05-0.45 range) cover less than half the scale of those separating nidovirus proteins. (Fig. 1C). Unlike cellular species, nidoviruses form few compact clusters, which are very distantly related. The distances between nidovirus proteins are unevenly distributed: intragroup distances between nidoviruses forming major lineages are in the 0.0-0.25 range, while intergroup distances between nidoviruses that belong to different lineages are in the 0.55-1.0 range. The distances separating the intermediate-sized mesonivirus from other nidoviruses tend to be most equidistant, accounting for ~15% of all distances in the 0.55-0.85 range.

The scale of nidovirus genome size change is proportional to the amount of substitutions in the most conserved proteins. To explore the relation of genome size change and the accumulation of substitutions, we plotted pairwise evolutionary distances (PED) separating the most conserved replicative proteins (Y axis) versus genome size difference (X axis) for all pairs of nidoviruses in our dataset (Fig. 2). It should be noted that the observed genome size difference may serve only as a low estimate for the actual genome size change, since it does not account for (expansion or shrinkage) events that happened in parallel between two viruses since their divergence. The obtained 378 values

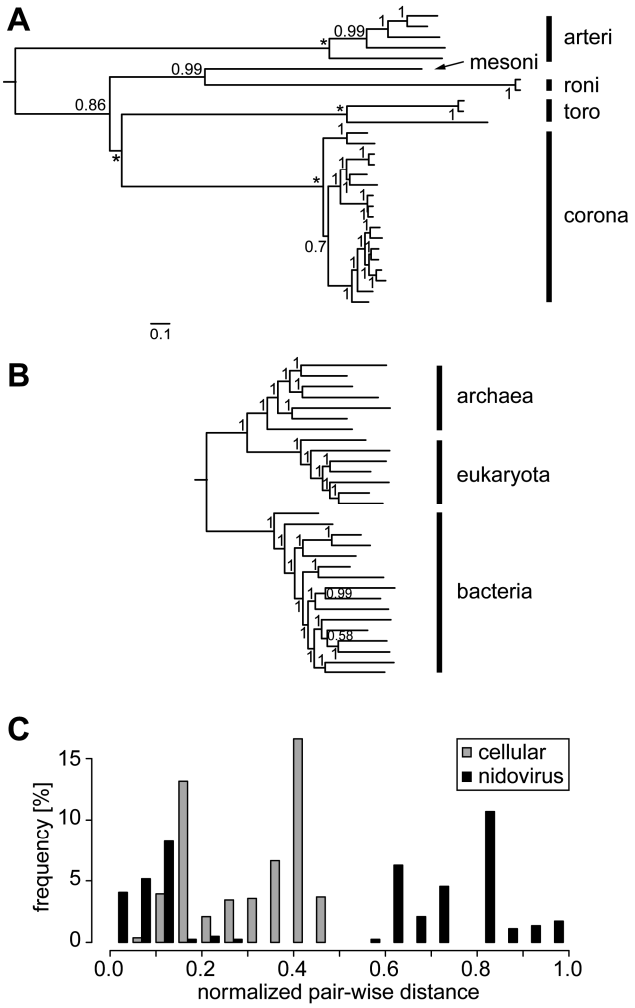

Figure 1. Phylogeny of nidoviruses in comparison to the Tree of life (ToL). Bayesian phylogenies of nidoviruses (A) and ToL (B) are drawn to a common scale of 0.1 amino acid substitutions per position. Major lineages are indicated by vertical bars and names; arteri: *Arteriviridae*, mesoni: Mesoniviridae, roni: *Roniviridae*, toro: *Torovirinae*, corona: *Coronavirinae*. Rooting was according to either (A) domain-specific outgroups[336] or (B) as described[50]. Posterior probability support values and fixed basal branch points (*) are indicated. The nidovirus and ToL alignments include, respectively, three enzymes and 56 single-gene protein families, 604 and 3336 columns, 2.95% and 2.8% gaps. For further details on the nidovirus tree see[336]. (C) Distributions of pair-wise distances for nidovirus and cellular single-copy conserved proteins according to the phylogenies in (A) and (B). The combined set of distances was normalized relative to the largest distance that was set to one.

are distributed highly unevenly, occupying the upper left triangle of the plot. Using phylogenetic considerations, four clusters could be recognized in the plot. Genetic variation within four major virus groups with more than one species (arteri-, corona-, roni-, and toroviruses) is confined to a compact cluster I in the left bottom corner (X range: 0.033-4.521 kb, Y range: 0.051-1.401). Values quantifying genetic divergence between major lineages are partitioned in three clusters taking in account genome sizes: large-sized vs. large-sized nidoviruses (cluster II, X: 0.002-5.433 kb, Y: 3.197-4.292), intermediate-sized vs. other lineages (cluster III, X: 4.475-11.494 kb, Y: 2.896-4.553), and small-sized vs. large-sized nidoviruses (cluster IV, X: 10.536-18.978 kb, Y: 4.159-5.088). Points in the clusters I, III and IV are indicative of a positive proportional relation between genome size change and the accumulation of replacements. The off-diagonal location of the cluster II can be reconciled with this interpretation under a (reasonable) assumption that the three lineages of large-sized nidoviruses expanded their genomes independently and considerably since diverging from the most recent common ancestor (MRCA). This positive relation is also most strongly supported by the lack of points in the bottom-right corner of the plot (large difference in genome size; small genetic divergence). Overall, this analysis indicates that a considerable change in genome size in nidoviruses could have been accomplished only over large evolutionary distances in the most conserved proteins.

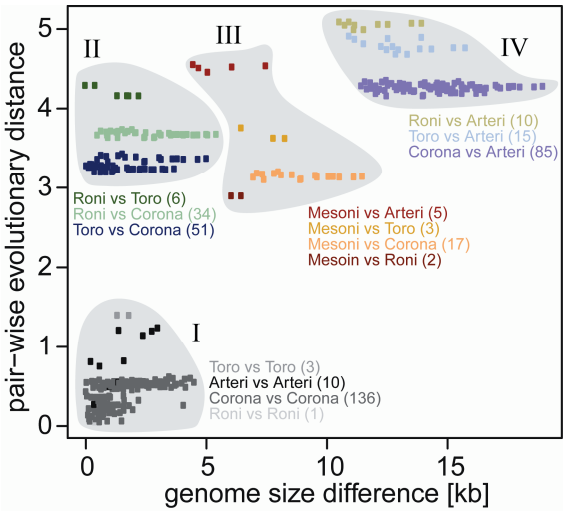

Figure 2. Relationship of evolutionary distance to genome size change in nidoviruses. Evolutionary distance (average number of substitutions per amino acid position in the conserved proteins) in relation to difference in genome size is shown for each pair (n=378) of the 28 nidovirus species. Points are colored according to pairs of major clades shown in Fig. 1A. The number of comparisons for each pair of clades is indicated by numbers in brackets. Points were grouped into clusters I (intra-lineage comparisons), II (large- vs. large-sized inter-lineage comparisons), III (intermediate-sized vs. others) and IV (small- vs. large-sized).

Only a fraction of genome size change may be linked to domain gain and loss. Next, we asked whether genome size change could be linked to domain gain and loss. We analyzed the phylogenetic distribution of protein domains that were found to be conserved in one or more of the five major nidovirus lineages[336]. Ancestral state parsimonious reconstruction was performed for the following proteins: ORF1b-encoded ExoN, N7-methyltransferase (NMT)[73], nidovirus-specific endoribonuclease (NendoU)[232,333], 2'-O-methyltransferase (OMT)[95,96], ronivirus-specific domain (RsD) (this study, see legend to Fig. S1), and ORF1a-encoded ADP-ribose-1"-phosphatase (ADRP)[129,375,402]. This analysis revealed that domain gain and loss have accompanied the NGE (Fig. S1 and Table S2). Particularly, genetically segregated ExoN, OMT and NMT (Fig. 3) were acquired in a yet-to-be determined order in the critical transition from small-sized to intermediate-sized nidovirus genomes. However, the combined size of these domains[336] accounts only for a fraction (49.7%) of the size difference (4,475 nt) between genomes of Nam Dinh virus (NDiV; 20,192 nt) and Simian hemorrhagic fever virus (SHFV), which has the largest known arterivirus genome (15,717 nt). The fraction that could be assigned to these and the three other protein domains is even smaller in other pairs of viruses representing different major nidovirus lineages (CL, AEG unpublished data). This analysis is also complicated by the uncertainty about the genome sizes of nidovirus ancestors that acquired or lost domains.

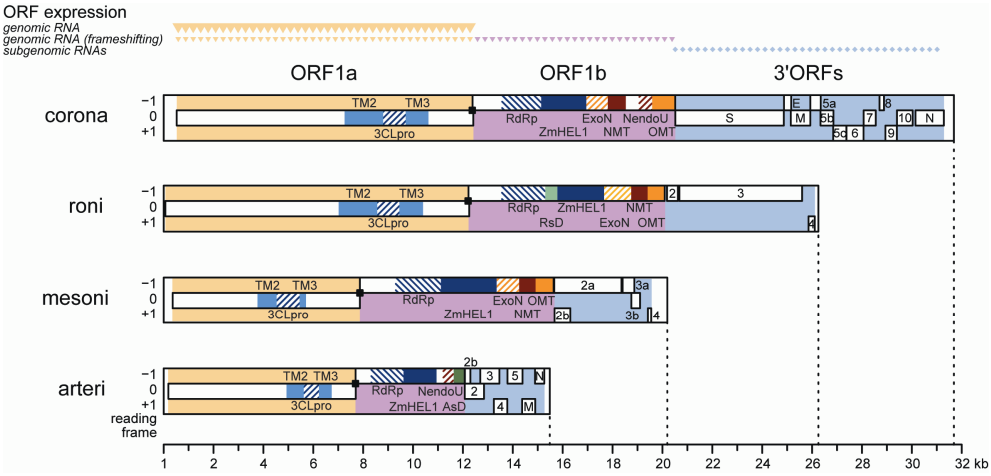

Figure 3. Genomic organization and expression, and key domains of four nidoviruses. The coding regions are partitioned into ORF1a (yellow), ORF1b (violet) and the 3'ORFs (blue), which also differ in expression mechanism as indicated on top. Black squares, ribosomal frameshifting sites. Within ORFs (white rectangles), colored patterns highlight domains identified in: all nidoviruses [TM2, TM3, 3CLpro, RdRp, and Zn-cluster binding domain fused with HEL1 (ZmHEL1)[461] - light and dark blue], large nidoviruses (ExoN, OMT - orange), certain clades (NMT, NendoU - red; ronivirus-specific domain (RsD) - light green; arterivirus-specific domain (AsD)- dark green). Genomic organizations are shown for Beluga whale coronavirus SW1 (corona), gill-associated virus (roni), Nam Dinh virus (mesoni), and porcine respiratory and reproductive syndrome virus North American type (arteri).

The nidovirus genome can be partitioned according to functional conservations in genome architecture. In order to gain further insight in the NGE dynamics, we had to analyze large genome areas in which homology signals are not recoverable in the currently available dataset because of both the extreme divergence of distant nidoviruses and a relatively poor virus sampling (Fig. 1). To address this challenge, we have developed an approach that establishes and exploits relationships between nidovirus genomes on grounds other than sequence homology. To this end, we partitioned the nidovirus genome according to functional conservations in the genome architecture, using results for few characterized nidoviruses and bioinformatics-based analysis for most other viruses (reviewed in [174]). With this approach, the genomes of all nidoviruses can be consistently partitioned in an alignment-free manner into five regions in the order from the 5'- to 3'-end: 5'-UTR, ORF1a, ORF1b, 3'ORFs, and 3'-UTR (Fig. 3). The 5'-UTR and 3'-UTR flank the ORFs area and account for <5% of the genome size in nidoviruses. The borders of the three ORF regions that overlap by few nucleotides in some or all nidoviruses were defined as follows: ORF1a: from ORF1a initiation codon to RFS signal, ORF1b: from RFS signal to ORF1b termination codon, and 3'ORFs: from ORF1b termination codon to the termination codon of the ORF that adjoins the 3'UTR.

It is noteworthy that the three ORF regions are of similar size but differ in expression mechanism (Fig. 3 top). Specifically, ORF1a is the first to be expressed by translation of the incoming virion RNA and, additionally, it encodes 3CLpro that mediates the release of mature proteins from the polyproteins pp1a and pp1ab. The expression of ORF1b, that follows, depends on the ORF1a region in three different ways: (i) the utilization of ribosomes that started translation on the ORF1a initiation codon; (ii) the use of the ORF1a/ORF1b RFS signal located upstream of the ORF1a termination codon; and (iii) the ORF1a-encoded 3CLpro. Finally, the expression of the 3'ORFs depends on products of the ORF1a and ORF1b to form the functional RTC for synthesizing subgenomic mRNAs that are translated to produce 3'ORF-encoded proteins[408]. Thus, ORF1a is the dominant region directly and indirectly controlling the expression of the entire genome.

The nidovirus genome expanded unevenly across three major coding regions. We then asked about how the different regions contributed to the genome expansion. We initially noted that the intermediate position of the mesonivirus between the two other nidovirus groups is observed only in genome but not region-specific size comparisons (Fig. 4). In the latter, the mesonivirus clusters with either small-sized (ORF1a and 3'ORFs) or large-sized (ORF1b) nidoviruses. This non-uniform position of the mesonivirus relative to other nidoviruses is indicative of a non-linear relationship between the size change of the complete genome and its various regions during the NGE. Accordingly, when fitting weighted linear regressions separately to the six datasets formed by nidoviruses with small and large genomes for three regions, support for a linear relationship was found only for the

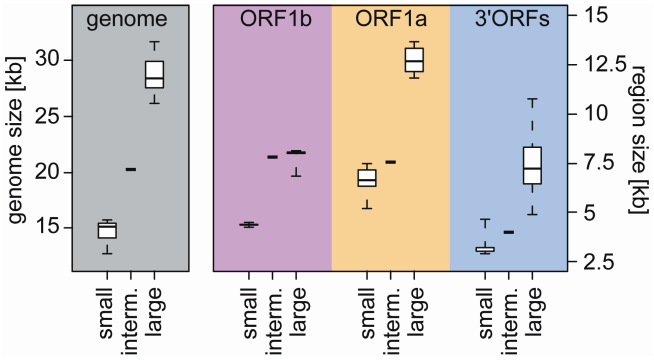

Figure 4. Nidovirus genome and region size differences. Shown are size distributions of genomes (left part) and the three genome coding parts ORF1a, ORF1b and 3'ORFs (right part) for five small-sized arterivirus species (small), 22 large-sized nidovirus species (large) and one intermediate-sized mesonivirus species (interm.). The distributions are represented by box-and-whisker graphs, where the box spans from the first to the third quartile and includes the median (bold line). The whiskers extend (dashed lines) to the extreme values.

3'ORF dataset of large nidoviruses; for all other regions a linear relationship was not statistically significant (Fig. S2). These results prompted us to evaluate linear as well as non-linear regression models applied to a dataset including all known nidovirus species (n=28) (Fig. 5). Two non-linear models were employed: third order monotone splines and a double-logistic regression. In the monotone splines, two parameters – the number and position of knots – determine the regression fit. We identified values for both parameters that result in the best fit (Fig. S3).

Using weighted r2 values, we observed that the splines model captures 92.9-96.1% of the data variation for the three ORF regions. This was a 5-22% gain in the fit compared to the linear model (75.9-90.8%) (Fig. 5). This gain was considered statistically significant (α=0.05) in two F-tests, a specially designed and standard one, as well as in the LV-test for every ORF region (p=0.018 or better) and, particularly, their combination (p=6.2e-5 or better) (Table 1). The splines model also significantly outperforms the double-logistic model (p=0.0011) (Table 1). These results established that the nidovirus genome expanded in a non-linear and region-specific fashion.

The three major coding regions expanded consecutively. Since each region expanded non-linearly during the NGE, so must the entire genome. Revealing its dynamic was our next goal. To this end, we analyzed the contribution of each of the five genomic regions to the overall genome size increase under the three models (Fig. 6 and Fig. S4). The top-ranked splines model (Table 1) predicts a cyclic pattern of overlapping wavelike increases of sizes for the three coding regions (the 5' and 3'UTR account only for a negligibly minor increase that is limited to small nidoviruses). Each of the three coding regions was found to

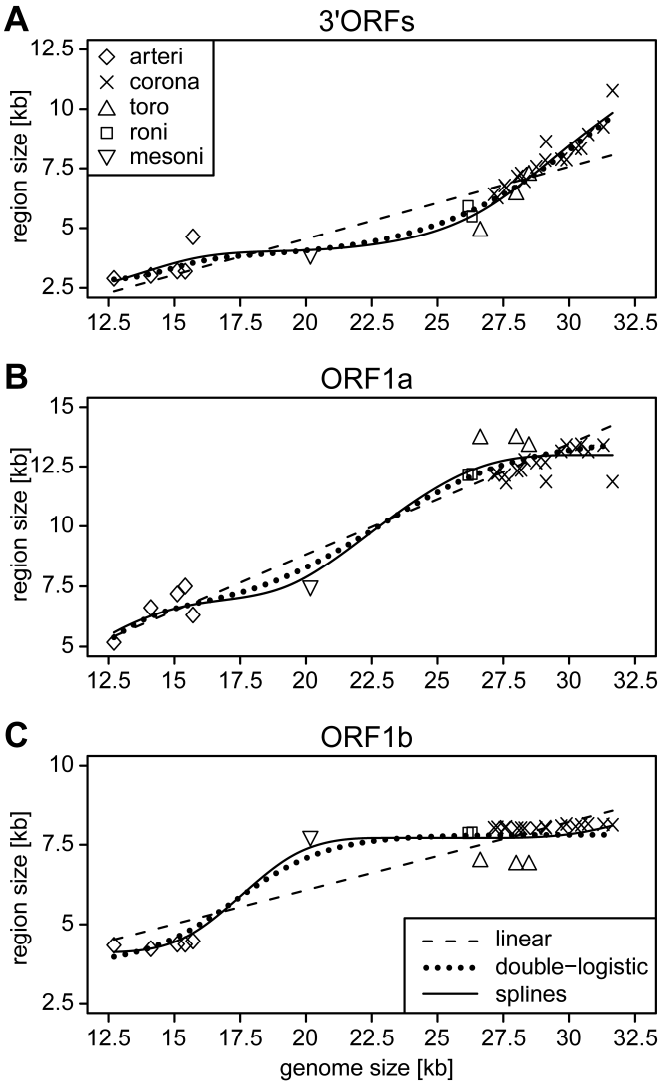

Figure 5. Relationship of sizes of three major coding regions and genome size in the nidovirus evolution. For 28 nidoviruses representing species diversity, absolute sizes of 3'ORFs (A), ORF1a (B), and ORF1b (C) are plotted against the size of the genome. Different symbols were used to group the viruses into five major phylogenetic lineages (see inlet in A). Results of weighted linear, double-logistic and 3rd order monotone splines[380] regression analyses are depicted. The three regression models (see inlet in C) fit the data with weighted r^2 values of 0.908 (linear), 0.948 (double-logistic) and 0.961 (splines) for ORF1a, 0.759, 0.900 and 0.929 for ORF1b, and 0.829, 0.950 and 0.955 for 3'ORFs. For fit comparison of regression models see Table 1.

Table 1. Comparison of regression models.

comparison[a]		test[b]	regression statistics[c]			
model A	model B		ORF1a	ORF1b	3'ORFs	total
linear	splines	F	0.0180*	0.0009*	0.0003*	5.2e-9*
linear	splines	F_{perm}	0.0008*	0.0028*	<1.0e-6*[d]	1.0e-6*
linear	splines	LV	0.0029*	0.0055*	0.0036*	6.2e-6*
linear	dlog	LV	0.0011*	0.0100*	0.0024*	6.5e-6*
dlog	splines	LV	0.0240*	0.0002*	0.20706	1.1e-3*

[a] linear regression model (linear); double-logistic regression model (dlog); 3rd order monotone splines regression model (splines)

[b] standard weighted F test (F); permutation F test (F_{perm}); a weighted version of a test to compare non-nested regression models (LV) as described in [286]

[c] shown is the probability that model A (null hypothesis) fits the data better than model B (alternative hypothesis); asterisks highlight significant values to reject the null in favor of the alternative hypothesis using a confidence level of 0.05; probabilities are calculated separately for ORF1a, ORF1b, 3'ORFs as well as the complete model combining the three coding plus the two UTR regions (total)

[d] non of the 1 million permutations resulted in an F larger than that of the non-permuted dataset

have been increased at different stages during the NGE (Fig. 6). A cycle involves expanding predominantly and consecutively the ORF1b, ORF1a and 3'ORFs region. One complete cycle flanked by two partial cycles are predicted to have occurred during the NGE from small-sized to large-sized nidoviruses. The complete cycle encompasses almost the entire genome size range of nidoviruses, starting from 12.7 kb and ending at 31.7 kb. The dominance of an ORF region in the increase of genome size was characterized by two parameters: a genome size range (X axis in Fig. 6) in which the contribution of a region accounts for a >50% share of the total increase, and by the maximal share it attains in the NGE (Y axis in Fig. 6). For three major regions these numbers are: ORF1b, dominance in the 15.8-19.3 kb range with 72.7% maximal contribution at genome size 17.6 kb; ORF1a, 19.6-25.9 kb and 83.0% at 22.4 kb; 3'ORFs, 26-31.7 kb and 89.8% at 29.4 kb (Fig. 6). Mesonivirus and roniviruses seem to have been "frozen" after the first (ORF1b) and second (ORF1a) wave, respectively. The third wave (3'ORFs) was due to the genome expansion of coronaviruses and, to a lesser extent, toroviruses (compare virus genome sizes on top with wave positions in Fig. 6).

Furthermore, the shapes of the three waves differ. The first one (ORF1b) is most symmetrical and it starts and ends at almost zero contribution to the genome change. This indicates that the ORF1b expansion is exceptionally constrained, which is in line with extremely narrow size ranges of ORF1b in arteri- and coronaviruses (with mean±s.d. of 4362±86 and 8071±50 nt, respectively; Fig. 4 and Fig. 6). The second wave (ORF1a) is tailed at the upper end and is connected to the ORF1a wave from the prior cycle. This ORF seems to have a relatively high baseline contribution (~20%) to the genome size change up to the range of coronaviruses. The third wave (3'ORFs) is most asymmetrical (incomplete),

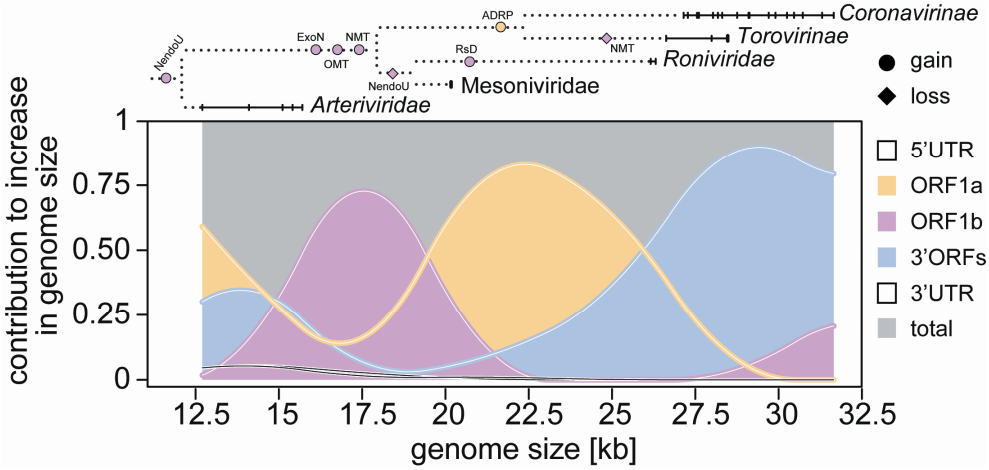

Figure 6. Region-specific, wavelike dynamics of the nidovirus genome expansions. Relative contributions of the genome regions ORF1a, ORF1b, 3'ORFs, 5'UTR and 3'UTR to the increase in genome size are calculated according to the splines regression and plotted on top of each other and against their sum=1. Solid horizontal lines and vertical bars on top: genome size ranges and samplings for nidovirus lineages indicated by names. Dotted lines: topology of major nidovirus branches. Selected domains gained (ExoN, OMT, NMT, RsD and ADRP, circles) and lost (NendoU and NMT, diamonds) are colored according to ORF in which they are encoded. See also Fig. 3, Fig. S1 and text.

as it only slightly decreases from its peak toward the largest nidovirus genome size at which this region remains the dominant contributor (~77%).

One partial cycle, preceding the complete one, is observed inside the genome size range of arteriviruses and involves the consecutive expansions of ORF1a and 3'ORFs, respectively. Also the main, but still very limited contributions of 5'- and 3'-UTRs (<6%) are observed here. The start of another incomplete cycle, involving the expansion of ORF1b and overlapping with the complete cycle, is observed within the upper end of coronavirus genome sizes.

Discussion

In this study we provide, for the first time, a quantitative insight into the large-scale evolutionary dynamics of genome expansion in RNA viruses. We analyzed nidoviruses, a monophyletic group of RNA viruses that populate the upper ~60% of the RNA virus genome size scale and include viruses with the largest known RNA genomes. Nidoviruses infect a broad range of different hosts including vertebrate and inverterbrate species and we now show that the evolutionary space explored by these viruses exceeds that of the ToL for comparable protein datasets. We exploited functional conservation in the genome

architecture in nidoviruses to partition their genomes in five spatially collinear regions. Using a complex statistical framework we reconstructed a non-linear trajectory of region-specific size increase that captured >92% of data variation. This trajectory may be shaped by the division of labor[442] between ORFs that predominantly control genome replication, genome expression, and virus dissemination, respectively. Combined, our results reveal that the genomic architecture severely constrains the NGE. Ultimately, it may determine the observed limit of genome size in contemporary RNA viruses.

Nidoviruses offer the best model for studying the control of RNA genome size. Genome size evolution in RNA viruses, unlike that of DNA-based life forms, has received relatively little attention from the research community. Several reasons may have contributed to this development. The narrow one-order range of small genome sizes that is compatible with the documented extremely high mutation rate[404] might have been perceived as evidence for the lack of meaningful genome size dynamics in RNA viruses. Even if there was any dynamics, its reconstruction could be considered challenging if not impossible to address, since evolutionary signals between distant lineages deteriorate profoundly due to the high mutation rate[220,489]. Consequently, the genome size increase in RNA viruses has so far been associated only with two trends to our knowledge: a concomitant increase of the average size of replicative proteins[33] and a reduction of genome compression measured by gene overlap[34].

In this respect, nidoviruses, which are often regarded an “exception” among RNA viruses[33,217], offer some unique opportunities for studying the evolution of RNA genome size. The genome size of nidoviruses is from ~20-to-200% larger than the “average” 10 kb RNA virus genome. Since nidoviruses form a monophyletic group and show a relatively large protein domain complexity, evolutionary analyses could be pursued.

Our results show that it took a considerable amount of evolutionary work in the most conserved proteins before a noticeable expansion of the nidovirus genome could be detected (Fig. 2). (In other, less conserved proteins the substitution rate is expected to be (much) larger). That relation is in line with an observation that nucleotide substitutions are on average four times more common than insertions/deletions in RNA viruses[404]. Whether this genome size increase also improves virus fitness and could determine the direction of evolution remains to be answered. In this respect we notice that viruses with larger genomes, compared to their small-sized cousins, could be expected to employ a more sophisticated repertoire of proteins for interacting with the host. It is also apparent that large-sized nidoviruses, unlike RNA viruses with smaller genomes, may afford both the acquisition and loss of an ORF as a matter of genome variation. Indeed, SARS-CoV adaptation to human and palm civets was accompanied with a large deletion in the ORF8-ORF10 area[193], and ORF gain/loss was documented in the recent evolution of other coronaviruses[68,302] (for review see [170]). Thus, large genomes could provide nidoviruses with an expanded toolkit to adapt upon crossing species barriers and to explore new niches in established hosts.

Inferring dynamics of genome size expansion in nidoviruses: viruses and protein domains. In our prior studies we already produced an unexpected insight into the control of genome size by identifying the ExoN domain in large-sized nidoviruses[432], a discovery that challenged a major paradigm of RNA virus biology - the universal lack of proof-reading during replication[220,439]. While this paradigm revision is getting support from experimental research[52,123,124,323], the recent discovery of a nidovirus in mosquitos, the mesonivirus NDiV, with a genome size in-between those of small-sized and large-sized nidoviruses, led to the proposal that 20 kb could be the genome size limit for non-segmented RNA viruses lacking ExoN (and proof-reading by implication)[336].

The identification of the 20 kb threshold poses questions about how nidoviruses have arrived at this threshold, crossed it, and expanded their genomes further. For addressing these questions we analyzed the entire ~19 kb genome size variation of nidoviruses (from 12.7 to 31.7 kb). We noted that only the lower ~20% and the upper ~30% of this range was sampled before the NDiV discovery. With the NDiV identification the ~50% non-sampled gap was split roughly in two halves, indicating that this sequence may provide a maximal information gain for analysis of the NGE (see also Fig. S2 in [336]). Indeed, an exceptionally large information value of the mesonivirus to this study is evident in many analyses (Figs. 3-6). On the other hand, the relatively strong impact of this single virus on the results may warrant an additional scrutiny to ensure the validity of conclusions. To this end, we list below other observations, in addition to the strong statistical significance (Table 1), that support the wavelike dynamics of the NGE. First of all, we note that a virus closely related to NDiV (called Cavally virus) was independently identified in a parallel study[500]. Both viruses share all properties that are critical for this study, including the size of genome and ORFs as well as the assignment of protein domains[284]. Second, these two mesoniviruses and the very distant roniviruses with large genomes form a monophyletic group (Fig. 1). This clustering correlates with common (molecular) properties, including the infection of invertebrate hosts and the lack of the NendoU domain, which distinguish mesoni- and roniviruses from other nidoviruses (Fig. S1) and could be expected to apply to other yet-to-be identified viruses of this group as well. Third, even if we restrict our analysis to small- and large-sized nidoviruses, differences between the size range of genomes and the three ORF regions are already apparent (Fig. 4). Particularly striking are the extremely constrained sizes of ORF1b in both arteriviruses and coronaviruses as well as an exceptionally large size range of 3'ORFs in large-sized nidoviruses. These constraints contribute prominently to the first and third wave, respectively, of the major cycle of the NGE (Fig. 6). Thus, the described dynamics of the region-specific genome size increase reflects properties of both mesoniviruses and other nidoviruses, and is expected to sustain upon future updates of virus sampling.

The available poor virus sampling limits the resolution of our reconstruction analysis of domain gain/loss during the NGE. For instance, the critically important acquisition of ExoN seems to be tightly correlated with those of two replicative

methyltransferases, NMT and OMT (Fig. S1). The fact that NMT and ExoN are adjacent domains in a single protein in coronaviruses (nsp14) and OMT resides nearby (nsp16) in pp1ab suggests a link between these domains and indicates that NMT and ExoN might have been acquired in a single event. Furthermore, NMT and OMT were shown to be essential for cap formation at the 5'-end of coronavirus mRNAs[73,95,96], with the OMT-mediated modification being important for the control of innate immunity[503]. These enzymes are yet to be characterized in other large-sized nidoviruses, and this characterization must reconcile the apparent lack of NMT in toroviruses[336] with its essential role in coronaviruses[73].

The ExoN acquisition is a hallmark of the first wave in the NGE because it is expected to have improved the replication fidelity and, thus, made further genome enlargements feasible. In contrast, no domain acquisition with a comparably strong biological rational could be identified for the second wave. Two aspects, both contrasting the first and second wave, are important to notice here. Firstly, while the first wave seems to reflect the genome expansion in a single ancestral lineage that might have given rise to all intermediate- and large-sized nidoviruses (founding event), the second wave is likely to encompass the expansions in several lineages that happened in parallel (Fig. S1b). Secondly, evolutionary relations of proteins in ORF1a (underlying the second wave) are not as extensively documented as those for ORF1b (underlying the first wave), since ORF1a proteins in nidoviruses have diverged far greater. Hence, the domain gain/loss description for the second wave is even less complete than that for the first wave. Most notable is the acquisition of ADRP (formerly X domain [180]) which seems to be part of the second wave in large-sized vertebrate nidoviruses (Fig. 6). This domain belongs to the macrodomain protein family with poorly understood function and a broad phyletic distribution in viruses and cellular organisms[357]. The ADRP was shown to have ADP-ribose-1''-phosphatase activity[375], bind poly-ADP-ribose[129], and its inactivation affected cytokine production in coronavirus-infected cells[137]. It was proposed to regulate RNA replication[432] and coronavirus pathogenesis[137], but its physiological function remains to be established. Unlike the first and second wave, the third one encompasses changes that predominantly happened during the radiation of a subfamily (*Coronavirinae*) rather than several families (Fig. 6); they are being analyzed in a separate study (CL & AEG, in preparation). Improved virus sampling in the future, especially in the genome size range around 20 kb, could be critical for the description of domain gain/loss in ORF1a and its refinement in ORF1b during the NGE (Fig. S1).

Genome architecture and division of labor may control dynamics of genome size expansion in nidoviruses. To analyze the dynamics of the NGE we exploited regional conservation of the expression mechanisms of ORFs in the nidovirus genome. This conservation has no parallel in the cellular world given the enormous accumulation of mutations it accommodated. It was established by combining results of comparative sequence analysis with those obtained by experimental characterization of few selected nidoviruses, mostly representing artriviruses and coronaviruses, the two polar groups in the

genome size dimension. Like with homology, functional considerations – in this case the roles of protein products in the viral life cycle and the order of ORF expression – were invoked to rationalize the observed conservation. Based on the available data, it could be argued that ORF1b, ORF1a, and 3'ORFs play predominant roles in genome replication, genome expression, and virus dissemination, respectively, in all nidoviruses. These three processes are essential for every virus and they form the backbone of the nidovirus life cycle (Fig. 7, bottom)[360]. ORF1b encodes the principal enzymes of RNA synthesis, e.g. RdRp, ORF1a controls the expression of all other ORFs by several mechanisms (see above), and the 3'ORFs encode the components of virus particles that are the principal vehicles of genome dissemination. The regional association of this dominant control of genome replication, genome expression, and virus dissemination may reflect the division of labor between the three non-overlapping coding regions of the genome in the nidovirus life cycle.

The cooperation between products of ORF1b, ORF1a, and 3'-ORFs is bidirectional in the nidovirus life cycle since the functioning of each region is critical for the two other regions. In contrast, the dynamics of genome expansion links these regions in the order ORF1b->ORF1a->3'ORFs (Fig. 7 top). It implies a predominantly unidirectional causative chain of regional expansion during the NGE that suggests a hierarchy of the three underlying biological processes. The association of the first wave of domain acquisitions with ORF1b attests for the universally critical role of replicative enzymes in the NGE beyond the 20 kb threshold that is observed by other ssRNA+ viruses (for discussion see [336]). Regardless in which order the OMT, NMT and ExoN loci were acquired, their products must have been adapted to the RTC whose enzymatic core is believed to be formed by ORF1b-encoded proteins[169,418,445]. Other, less conserved RTC components are encoded in ORF1a[96,200,229,371,409,491]. It is known that proteins encoded in ORF1a and ORF1b interact in coronaviruses[230,352,409] and some of these interactions, e.g. between nsp10 and nsp14 or nsp16, were shown to be essential for functioning of the ORF1b-encoded enzymes involved[51,52,74]. Accordingly, the RTC, already enlarged with the newly acquired ORF1b-encoded subunits, could have triggered and/or sustained expansion of ORF1a. Additionally, it may be prompted by the need to adapt the expression mechanisms for polyproteins 1a and 1ab, which were already increased in size and complexity in the ORF1b-encoded part. The final wave of expansion involving the 3'ORFs may be triggered by the need to adapt virus particles for accommodating the expanded genome[337]. During the NGE, a part of the newly acquired genetic material may have been adapted to facilitate both virus-host interactions[187,224,246,494] and inter-region coordination for the benefit of the processes they control and the life cycle[334]. For instance, in arteriviruses the ORF1a-encoded nsp1 is essential for subgenomic mRNA synthesis and virion biogenesis[332,454,455] and a role in transcription was proposed for an ORF1a-encoded domain of nsp3 in coronaviruses[210]. Thus, factors encoded by ORF1a and ORF1b might constrain the NGE by controlling the expression of the 3'ORFs region and/or the functioning of its products. This would explain why the 3'ORFs expansion could not have been possible before the expansion of ORF1a

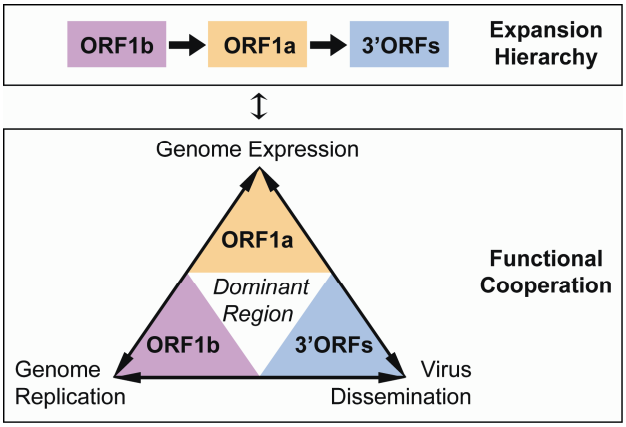

Figure 7. Hierarchy and cooperation in the nidovirus genome expansions. Functional and evolutionary relations between the three major coding regions of the nidovirus genome are depicted. For a brief description on the relationship between these three coding regions and the processes they dominate in the nidovirus life cycle, see text.

and ORF1b. By similar reasoning, an extremely tight control of the ORF1b size (Fig. 4) may set the ultimate size limit to the NGE. Finally, we note that the expansion order of the three coding regions matches their ranking according to sequence conservation, which is evident in the regional distribution of the nidovirus conserved domains (Figs. 2 and 3). This conservation is inversely proportional to the amount of accumulated substitutions, although quantitative characterization of the latter aspect is yet to be systematically documented. Genome changes due to regional-specific expansion and residue substitution may affect each other, and both may contribute to virus adaptation to the host.

Concluding Remarks and Implications. It is broadly acknowledged that extremely high mutation rates and large population sizes allow RNA viruses to explore an enormous evolutionary space and to adapt to their host[33,107]. Yet the low fidelity of replication also confines their evolution within a narrow genome size range that must affect their adaptation. Above, we presented evidence for a new source of constrains of genome expansion in RNA viruses by analyzing nidoviruses which include viruses with improved replication fidelity. In our analysis conserved genome architecture and the associated division of labor emerged as potentially powerful forces for selecting new genes and target genome regions during genome expansion. Importantly, the major diversification of nidoviruses by genome expansion must have started at some early point after the acquisition of ExoN[336]. From that point nidoviruses expanded their genomes in parallel in an increasing number of lineages, each of which may have acquired different domains in a same region. Extant nidoviruses of major lineages have very different genome sizes which we found to correspond to particular

points on the common region-specific genome expansion trajectory. The entire nidovirus (genome size) diversity may serve as a snapshot of different stages of the NGE. For viruses with largest genomes those with smaller genomes represent stages that they have passed in the NGE. For smaller genomes those with the larger ones represent stages that they have not reached in the NGE. It seems that the host may play a role in this process since ExoN-encoding nidoviruses that infect invertebrate are at the low side of genome size. For yet-to-be described nidoviruses, the genome expansion model can predict sizes of three coding regions by knowing only the genome size. The mechanistic basis of this fundamental relation can be probed by comparative structure-function analyses that should also advance the development of nidovirus-based vectors and rational measures of virus control. Thus, the wavelike dynamics model links virus discovery to basic research and its various applications.

This study indicates that genome size in RNA viruses may be restricted by the genome architecture in addition to the low fidelity of replication. Ultimately, these constraints may determine the upper limit of the RNA virus genome size. The reported data point to an important evolutionary asymmetry during genome expansion, which concerns the relation between proteins controlling genome replication, expression, and dissemination, and may be relevant beyond the viruses analyzed here.

Methods

Datasets. A dataset of nidoviruses representing species diversity from the three established and a newly proposed virus family was used (Table S1). A multiple alignment of nidovirus-wide conserved protein domains (28 species, 3 protein families, 604 aa alignment positions, 2.95% gap content) as described previously[336] formed the basis of all phylogenetic analyses. To put the scale of the nidovirus evolution into an independent perspective, we compared it with a cellular dataset previously used to reconstruct the Tree of Life, for which a concatenated alignment of single-copy proteins was used (30 species, 56 protein families, 3336 aa alignment positions, 2.8% gap content)[50]. The proteins used in the nidoviral and cellular datasets are the most conserved in their group and, as such, could be considered roughly equivalent and suitable for the purpose of this comparative analysis.

Phylogenetic analyses. Rooted phylogenetic reconstructions by Bayesian posterior probability trees utilizing BEAST[119] under the WAG amino acid substitution matrix[478] and relaxed molecular clock (lognormal distribution)[118] were performed as described previously[336]. Evolutionary pairwise distances were calculated from the tree branches. A maximum parsimony reconstruction of the ancestral nidovirus protein domain states at internal nodes of the nidovirus tree was conducted using PAML4 [487]. The quality of ancestral reconstructions was assessed by accuracy values provided by PAML4. To correct for non-

independence of the sequences[146] we assigned relative weights to the 28 nidovirus species by using position-based sequence weights[209] that were calculated on the alignment submitted for phylogeny reconstruction. The weights were normalized to sum up to one and were used in regression analyses (see below). The sequence weights varied ~7 fold from 0.017 to 0.116. NDiV, which represents mesoniviruses, showed the largest weight of 0.116 that was distantly followed by those of the bafinivirus White bream virus (WBV; 0.075) and roniviruses (0.06 each); coronaviruses, making up the best-sampled clade, were assigned the lowest weights (0.017 to 0.028 each).

Statistical analysis of genome size change in nidoviruses. The genome of each nidoviruses was consistently partitioned into five genomic regions according to external knowledge (see Results). To model the contribution of each genomic region to the total genome size change, we conducted weighted regression analyses (size of a genomic region on size of the genome) using three models – a linear and two non-linear ones. Position-based sequence weights were used and a confidence level of $\alpha=0.05$ was applied in all analyses. The combined contributions of all genomic regions to the genome size change must obviously sum up to 100%. To satisfy this common constraint, in each analysis, regression functions were fitted simultaneously to sizes of the genomic regions by minimizing the residual sum of squares, thereby constraining the sum of all slopes to be not larger than one. The linear model assumes a constant contribution of each genomic region during evolution which was modeled via linear regions.

In the first non-linear model we applied third order monotone splines with equidistant knots[380]. We chose splines because of their flexibility and generality (we don't rely on a specific regression function). The monotonicity constraint was enforced to avoid overfitting which was observed otherwise, and third order functions were chosen to obtain smooth, second-order derivatives. We explored the dependence of the performance of the splines model on variations in two critical parameters, the number of knots and the start position of the first knot. These two parameters define a knot configuration and determine a partitioning of the data into bins. In the first test we evaluated five different configurations generating from three to seven knots. Configurations using eight or more knots resulted in some bins being empty and were therefore not considered. For each number of knots the position of the first knot and the knot distance were determined as resulting in that configuration for which the data points are distributed most uniformly among the resulting bins. The exception was the 3-knot configuration, in which the position of the second knot was selected as the intermediate position in the observed genome size range (22.2kb). Only configurations with equidistant knots were considered. All probed splines models were evaluated by goodness-of-fit values (weighted version of the coefficient of determination r^2). In the second test we evaluated the model dependence on the position of the first knot by considering all positions that do not result in empty bins for the optimal number of knots determined using the approach described above.

As another non-linear model we used a 7-parameter double-logistic regression function that mimics the splines model and more readily allows for biological interpretations. Since double-logistic regressions did not converge for the 5'- and 3'-UTRs, linear functions were used for these two genome regions instead.

Linear (null hypothesis) and splines (alternative hypothesis) regression models were compared using standard weighted F-statistics and a specially designed permutation test (see below). To exclude overfitting as the cause of support of the more complex models, we utilized a more sophisticated framework (LV-Test) for the comparison of non-nested regression models (linear vs. double-logistic and splines vs. double-logistic) as detailed in [286]. The test was further modified to include weighted residuals according to virus sequence weights that account for sequence dependence.

Since our null hypothesis (linear model) is at the boundaries of the parameter space, we developed a permutation test to further compare the linear and splines models. To this end, genome region sizes were transformed to proportions (region size divided by genome size), randomly permuted relative to genome sizes, and transformed back to absolute values. These transformations are compatible with the constraints of the null hypothesis and the requirement that region sizes have to sum to genome sizes. Weights were not permuted. The linear and splines models were fit to the permuted datasets and F-statistics were calculated as for the original dataset. The p-value of the test is the fraction of F-statistics of permuted datasets that are larger than the F of the original dataset. It was calculated using 1,000,000 permutations that were randomly sampled out of ~10^{29} possible permutations.

Finally, we analyzed the contribution of each genome region to the total change in genome size under the three regression models. The contribution of each region according to a model was calculated as the ratio of change in region size to change in genome size (first derivative of the regression function) along the nidovirus genome size scale. These region-specific contributions were combined in a single plot for visualization purposes.

To conduct all statistical analyses and to visualize the results we used the R package[377].

Accession numbers. Accession numbers of virus genomes utilized in the study are shown in Table S1.

Acknowledgments

We thank Igor Sidorov for discussions and together with Alexander Kravchenko and Dmitry Samborskiy for Viralis management. This research has received funding from the Program of Japan Initiative for Global Research Network on Infectious Diseases (J-GRID), MEXT, Japan, the European Union Seventh Framework Programme (FP7/2007-2013) under the

program SILVER (grant agreement no. 260644), the Netherlands Bioinformatics Centre (BioRange SP3.2.2), the Collaborative Agreement in Bioinformatics between Leiden University Medical Center and Moscow State University (MoBiLe), and Leiden University Fund.

Supporting Information

Table S1. Nidovirus representatives.

virus	virus abbreviation[a]	(sub)family	accession[b]
Nam Dinh virus	NDiV_01-03	Mesoniviridae	DQ458789
Gill-associated virus	GAV_96	*Roniviridae*	AF227196
Yellow head virus	YHV_98	*Roniviridae*	EU487200
White bream virus	WBV-DF24_00	*Torovirinae*	NC_008516
Equine torovirus	EToV-Berne_72	*Torovirinae*	X52374
Bovine torovirus	BToV-Breda1_79	*Torovirinae*	NC_007447
Human coronavirus 229E	HCoV-229E_65	*Coronavirinae*	NC_002645
Human coronavirus NL63	HCoV-NL63_02	*Coronavirinae*	DQ445911
Miniopterus bat coronavirus 1	Mi-BatCoV-1A_05	*Coronavirinae*	NC_010437
Rhinolophus bat coronavirus HKU2	Rh-BatCoV-HKU2_06	*Coronavirinae*	NC_009988
Miniopterus bat coronavirus HKU8	Mi-BatCoV-HKU8_05	*Coronavirinae*	NC_010438
Scotophilus bat coronavirus 512	Sc-BatCoV-512_05	*Coronavirinae*	DQ648858
Porcine epidemic diarrhoea virus	PEDV-CV777_77	*Coronavirinae*	NC_003436
Feline coronavirus	FCoV_79	*Coronavirinae*	NC_007025
SARS coronavirus	SARS-HCoV_03	*Coronavirinae*	AY345988
Tylonycteris bat coronavirus HKU4	Ty-BatCoV-HKU4_04	*Coronavirinae*	EF065505
Pipistrellus bat coronavirus HKU5	Pi-BatCoV-HKU5_04	*Coronavirinae*	EF065509
Rousettus bat coronavirus HKU9	Ro-BatCoV-HKU9_05	*Coronavirinae*	EF065513
Human coronavirus HKU1	HCoV-HKU1_04	*Coronavirinae*	AY884001
Human coronavirus OC43	HCoV-OC43_67	*Coronavirinae*	AY585228
Mouse hepatitis virus	MHV-A59_59	*Coronavirinae*	AY700211
Infectious bronchitis virus	IBV-Beaud_35	*Coronavirinae*	NC_001451
Beluga whale coronavirus SW1	BWCoV-SW1_06	*Coronavirinae*	EU111742
Equine arteritis virus	EAV-CW_96	*Arteriviridae*	AY349167
Simian hemorrhagic fever virus	SHFV_64	*Arteriviridae*	NC_003092
Lactate dehydrogenase-elevating virus	LDV-P_71	*Arteriviridae*	U15146
Porcine respiratory and reproductive syndrome virus, North American type	PRRSV-NA_95	*Arteriviridae*	AF176348
Porcine respiratory and reproductive syndrome virus, European type	PRRSV-LV_91	*Arteriviridae*	M96262

[a] acronym of virus name joined ("_") with sampling year or period for this virus
[b] Genbank/Refseq accession number

Table S2. Nidovirus ancestral protein domain reconstruction.

ancestral node[a]	protein domain[b]											
	NendoU		ExoN		OMT		NMT		ADRP		RsD	
nido (root)	1	1.000	0	0.576	0	0.576	0	0.645	0	1.000	0	1.000
arteri	1	1.000	0	1.000	0	1.000	0	1.000	0	1.000	0	1.000
large nido+mesoni	1	1.000	1	1.000	1	1.000	1	0.836	0	1.000	0	1.000
mesoni+roni	0	1.000	1	1.000	1	1.000	1	1.000	0	1.000	0	1.000
roni	0	1.000	1	1.000	1	1.000	1	1.000	0	1.000	1	1.000
corona+toro	1	1.000	1	1.000	1	1.000	1	0.836	1	1.000	0	1.000
toro	1	1.000	1	1.000	1	1.000	0	1.000	1	1.000	0	1.000
corona	1	1.000	1	1.000	1	1.000	1	1.000	1	1.000	0	1.000

[a] abbreviations: nidoviruses (nido), large and intermediate size nidoviruses (large nido), roniviruses (roni), mesoniviruses (mesoni), toro-/bafiniviruses (toro), coronaviruses (corona), arteriviruses (arteri).

[b] shown are the reconstructed state (presence, 1, or absence, 0) and its accuracy by decimal numbers in the range of [0.500-1.000]) at the respective ancestral node for six domains in a maximum parsimony analysis using PAML.

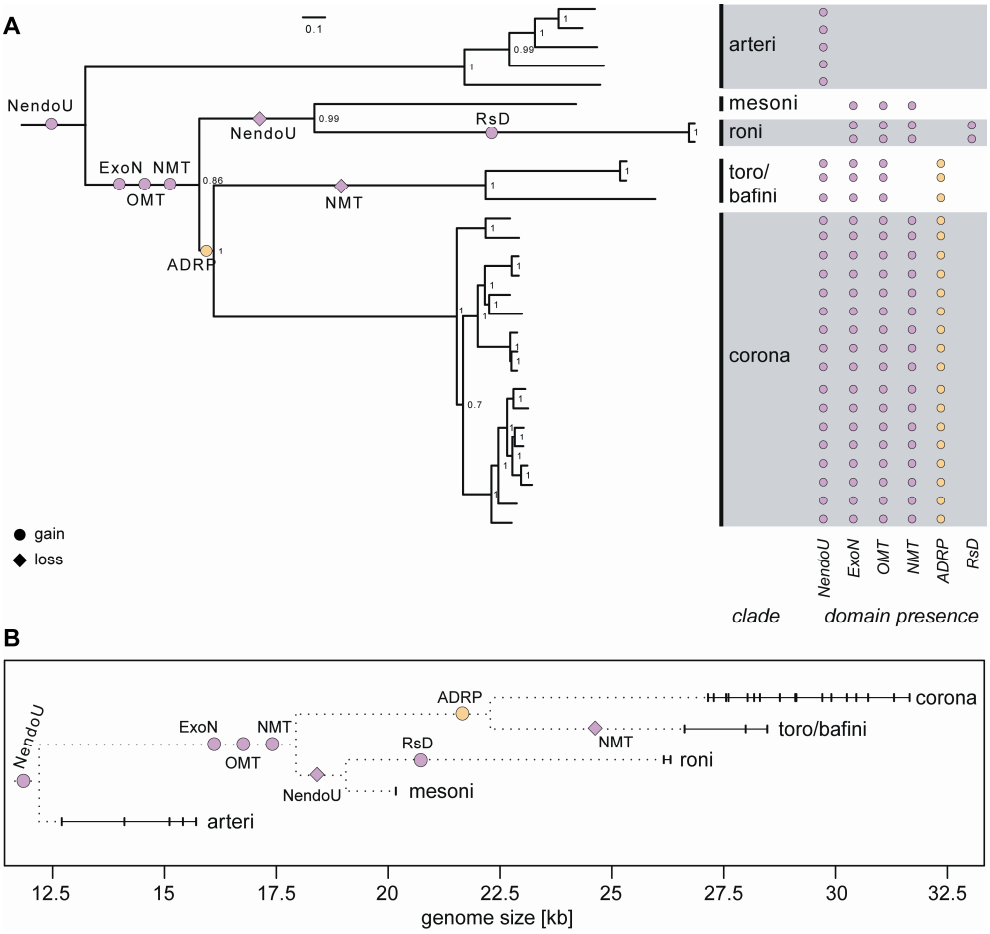

Figure S1. Gain and loss of selected ORF1a/ORF1b domains found in subsets of nidoviruses. (A) Distribution of six selected domains identified in ORF1a (one) and ORF1b (five) conserved in subsets of 28 nidovirus species (right part). One of the ORF1b-encoded domains (RsD) was identified in this study by inspection of the pp1b alignment as a ronivirus-specific insertion (163 aa) that is located between the conserved RdRp and ZmHEL1 domains (see Fig. 3). Colors indicate a domain's ORF location (purple for ORF1b, yellow for ORF1a). The left part shows predicted gain (circles colored according to its ORF location) and loss (colored diamonds) events at internal branches of the nidovirus phylogeny[336]. Nidovirus ancestral domain compositions were reconstructed utilizing a maximum parsimony analysis implemented in PAML4. Support values are shown in Table S2. (B) The nidovirus phylogeny was mapped on the genome size scale (dotted lines). Individual genome sizes of 28 nidovirus species are shown by vertical dashes and the size range within major lineages by horizontal solid lines. Internal nodes in the tree were arbitrarily placed at half the distance of adjacent branching events connecting two lineages while observing the original topology of the phylogeny. Predicted domain gain/loss events are highlighted as in (A).

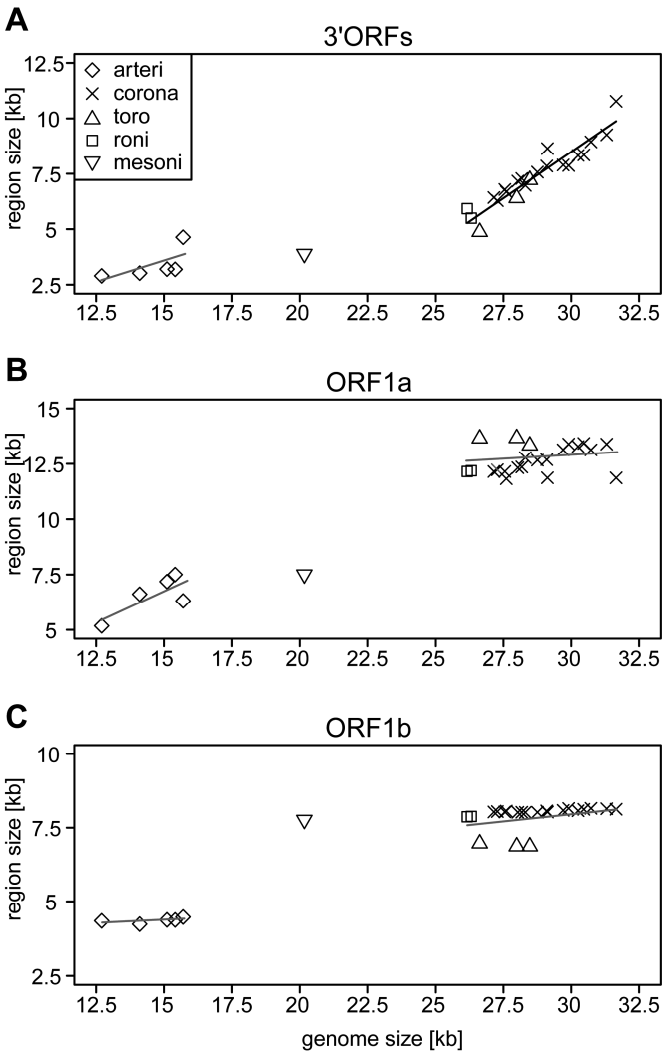

Figure S2. Clade-specific relationship of sizes of three major coding regions and genome size in the nidovirus evolution. For 28 nidoviruses representing species diversity, absolute sizes of 3'ORFs (A), ORF1a (B), and ORF1b (C) are plotted against the size of the genome. Different symbols were used to group the viruses into five major phylogenetic lineages (see inlet in A). Results of weighted linear regression analyses for small-sized (arteri) and large-sized nidoviruses (corona, toro/bafini, roni) are depicted. Regressions with a slope significantly different from zero are shown in black, non-significant ones in grey. The linear regressions fit the data with p=0.11, r^2=0.62 (arteri) and p=0.45, r^2=0.03 (corona, toro/bafini, roni) for ORF1a, p=0.33, r^2=0.31 and p=0.1, r^2=0.13 for ORF1b, and p=0.21, r^2=0.45 and p=6e-11, r^2=0.89 for 3'ORFs. The only significant correlation was observed for 3'ORFs of nidoviruses with large genomes (A) where the regression line showed a slope of 0.84 (±0.07 s.e.).

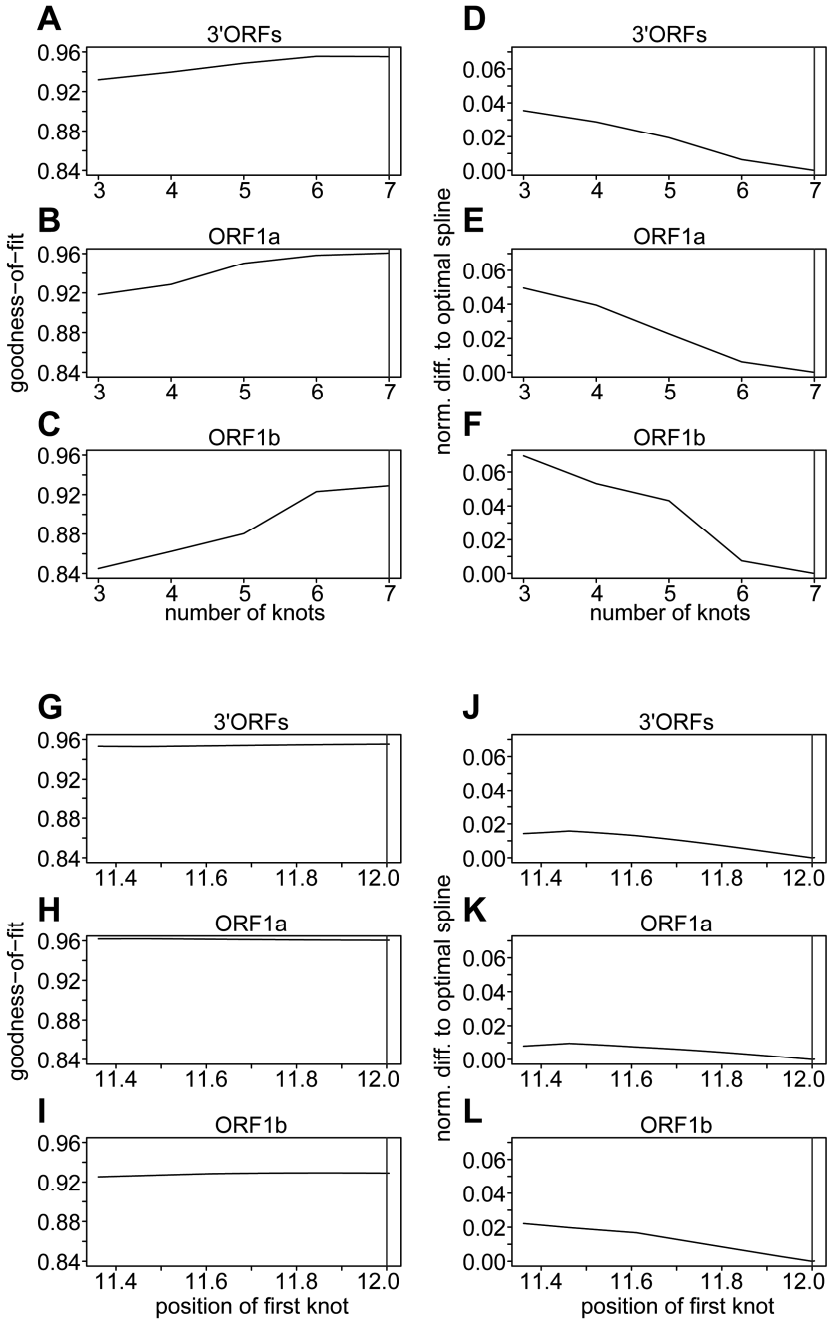
A
3'ORFs
B
ORF1a
C
ORF1b
goodness−of−fit
number of knots
D
3'ORFs
E
ORF1a
F
ORF1b
norm. diff. to optimal spline
number of knots
G
3'ORFs
H
ORF1a
I
ORF1b
goodness−of−fit
position of first knot
J
3'ORFs
K
ORF1a
L
ORF1b
norm. diff. to optimal spline
position of first knot

Figure S3. Sensitivity of the splines regression model to the number of knots and the position of the first knot. Shown are goodness-of-fit in form of weighted r^2 values (A-C, G-I) and sensitivity on the resulting regression curve (D-F, J-L) for different number of knots in the range of 3 to 7 (A-F) and different positions of the first knot (G-L) for the 3'ORFs, ORF1a and ORF1b genome regions. The best fit was obtained for the 7-knot configuration for all three regions (A-C). Hence, the 7-knot configuration was selected as the optimal one. We have also calculated a difference between other splines models compared to the optimal knot number by calculating the absolute difference of the regression curves of two configurations normalized to the size range of observed values (e.g. size ranges of ORF1a, ORF1b or 3'ORFs). This difference was in the range of 1-7% and increased with decreasing knot number in all three regions (D-F); it could be viewed as the loss of fit relative to the 7-knot configuration. Also, we calculated the model dependence on the position of the first knot by evaluating all positions that do not result in empty bins for the 7-knot configuration, which was found to be in the range from 11.4 to 12.0 kb (G-I). There was virtually no dependence of the position of the first knot and the goodness-of-fit (G-L); we selected the position that is closest to the minimal genome size. The knot number (k=7) and position of the first knot (at 12kb resulting in a knot distance of 3.7kb) used in the main calculation are indicated by green vertical lines.

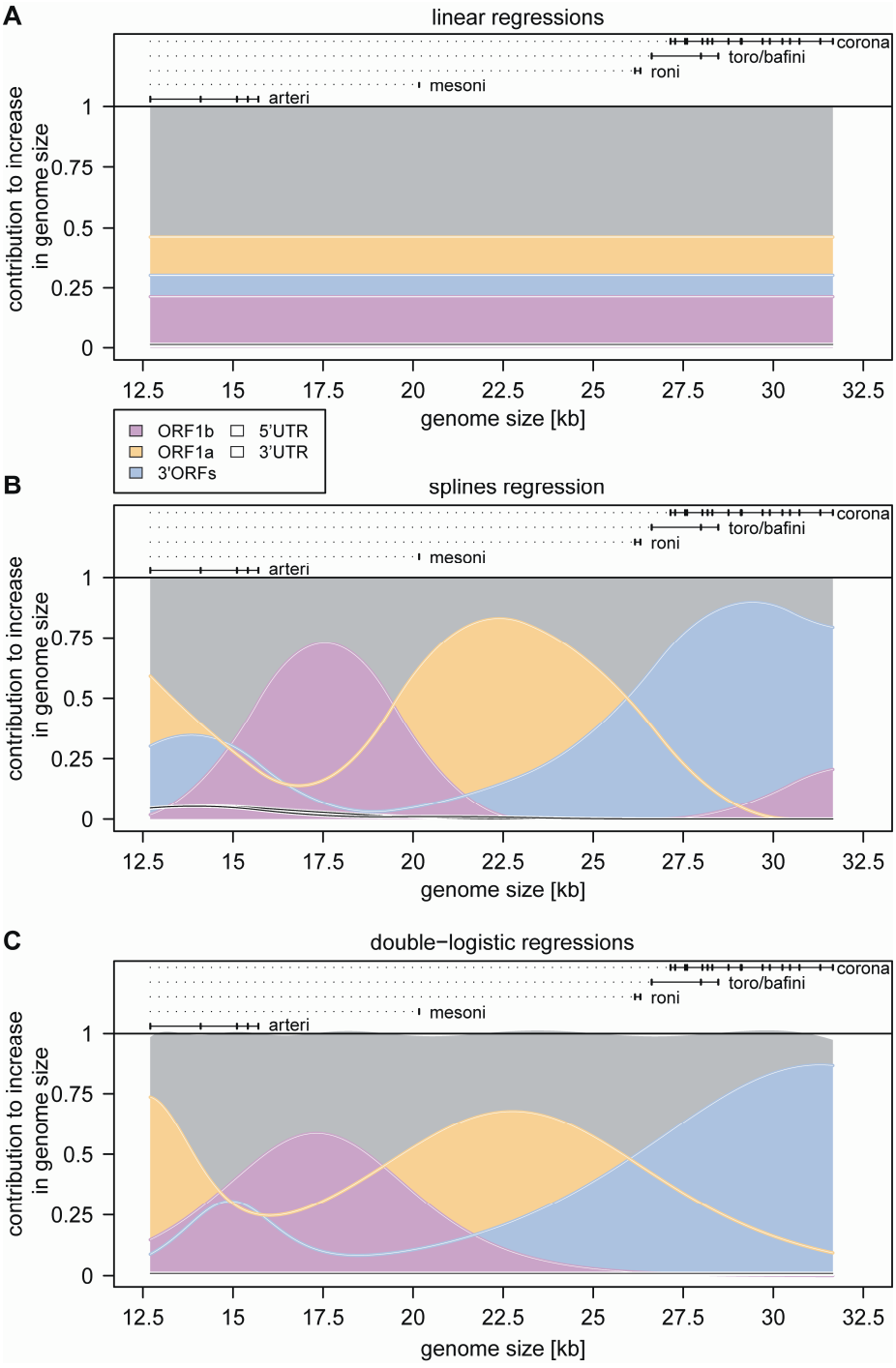
A
linear regressions
corona
toro/bafini
roni
mesoni
arteri
contribution to increase in genome size
1
0.75
0.5
0.25
0
12.5
15
17.5
20
22.5
25
27.5
30
32.5
genome size [kb]
ORF1b
ORF1a
3'ORFs
5'UTR
3'UTR
B
splines regression
C
double−logistic regressions

Figure S4. Modeling contribution of ORF1a, ORF1b, 3'ORFs, 5'UTR and 3'UTR to the nidovirus genome expansion. Relative contributions of ORF1a (yellow), ORF1b (purple), 3'ORFs (blue), and 5' and 3'UTR (black) to the increase in genome size are plotted on top of each other and against their sum=1 (grey) for the linear (A), the splines (B) and the double-logistc (C) regression model. Relative size contributions were calculated based on the regression curves fitted to the five genome parts for a dataset of 28 nidoviruses representing species diversity. Solid horizontal lines and vertical bars on top: genome size ranges and virus samplings for arteri-, corona-, toro-/bafini-, roni- and mesoniviruses. Under the linear model (which was statistically rejected in favor of the non-linear models), the contribution of each region to the genome size change is constant by definition. The ORF1a region accounts for most change (46.3%), followed by 3'ORFs (30.2%), ORF1b (21.3%), 5'UTR (1.3%) and 3'UTR (0.8%). In contrast, the splines and double-logistic models predict a cyclic pattern of overlapping wave-like increases of sizes for the three ORFs regions, with maximal contributions of 72.7%, 83.0% and 89.8% for ORF1b, ORF1a and 3'ORFs, respectively (see also main text). Highly similar cyclic and wave-like patterns of region expansions are predicted by the double-logistic model that mostly differs in the amplitude and range of waves compared to those of the splines model. These similarities suggest that the double-logistic model might be an approximation of the monotone splines model facilitating biologically meaningful interpretations.