The enzymes for genome size increase and maintenance of large (+)RNA viruses

Francois Ferron, Bhawna Sama, Etienne Decroly, Bruno Canard

The enzymes for genome size increase and maintenance of large (+)RNA viruses

François Ferron, [1,2] Bhawna Sama, [1] Etienne Decroly, [1] and Bruno Canard [1,2,*]

With sizes <50 kb, viral RNA genomes are at the crossroads of genetic, biophysical, and biochemical stability in their host cell. Here, we analyze the enzymatic assets accompanying large RNA genome viruses, mostly based on recent scientific advances in Coronaviridae. We argue that, in addition to the presence of an RNA exonuclease (ExoN), two markers for the large size of viral RNA genomes are (i) the presence of one or more RNA methyltransferases (MTases) and (ii) a specific architecture of the RNA-dependent RNA polymerase active site. We propose that RNA genome expansion and maintenance are driven by an evolutionary ménage-à-trois made of fast and processive RNA polymerases, RNA repair ExoNs, and RNA MTases that relates to the transition between RNA- to DNA-based life.

Highlights

Gene expression originated from early protoviruses ('replicators') resembling (+)RNA viruses.

RNA genome size expansion corelates with specific RNA-dependent RNA polymerase signature sequences.

RNA genome size expansion and maintenance requires more than an increasing replication accuracy provided by the acquisition of repair exonucleases (ExoNs).

The presence of RNA methyltransferase (MTase) genes corelates with increasing RNA genome size.

Large RNA genomes result from the coexistence of at least a processive RNA-dependent RNA polymerase, an RNA repair ExoN, and at least one RNA MTase.

The acquisition of enzymes for genome size expansion and stability

Viruses survive and adapt to environmental changes while maintaining major functional and structural elements [1]. This inner strength relies on an evolutionary paradigm involving seemingly antagonistic properties of genome stability and plasticity simultaneously. In a changing environment, viral loss of fitness (see Glossary) is buffered by genome evolution involving nucleic acid sequence adaptation (mutations, insertions, and deletions) and rearrangements (RNA recombination, gene acquisition/loss). In RNA viruses, this empirical exploration along fitness landscapes remains constrained by physical properties of the RNA molecule. To allow expansion and overcome the RNA biophysical limitation as a genetic material, physical protection, sequence stability, and specialized enzymes are required.

In light of recently discovered large (+)RNA virus genomes, we argue that they have achieved genome stability and maintenance through the acquisition and synergy of high-fidelity RNA-dependent RNA polymerase (RdRp) and/or an RNA repair exonuclease (ExoN) and RNA methyltransferases (MTases), respectively. Thus, this opinion article focuses on these enzymes and their signature sequence accompanying large (+)RNA viruses.

Prebiotic RNAs and their evolution toward RNA and DNA genomes

It is generally accepted that RNA appeared before DNA during evolution, supporting the existence of an ancestral RNA world [2]; together with primary peptides, short oligoribonucleotides acquired catalytic activity (ribozymes, ribonucleoproteins, and ribosome-like factories) and later peptide-coding capacity stored in their nucleic acid sequence. The appearance of membranes provided these primordial genomes (or 'replicators') new tools adapted to thrive in their environment. The RNA replication machinery of these replicators, or 'proto-RNA viruses,' arose from the primordial pool of genetic elements, while capsid proteins were captured from hosts and appeared later at different stages of evolution [3]. Replicators have expanded in size up to a last universal common RNA ancestor (LUCRA), after which DNA genomes appeared. Due to its improved biophysical properties compared with RNA, DNA successfully persisted: DNA genome

[1]Centre National de la Recherche Scientifique, Aix-Marseille Université, CNRS UMR 7257, AFMB, Case 925, 163, Avenue de Luminy, 13009 Marseille, France
[2]European Virus Bioinformatics Center, Leutragraben 1, 07743 Jena, Germany

*Correspondence:
bruno.canard@univ-amu.fr (B. Canard).

sizes range from a few kilobases (kb) for DNA viruses to multimegabases for higher-order eukaryotic organisms. The 'virus early' hypothesis [3] is conceptually compatible with this scenario: after this transition, proto-RNA viruses might have acquired and evolved their own polymerase, the RdRp, in parallel to the appearance of reverse transcriptase (RT), one of the most ancient DNA polymerase folds known. This is supported by the fact that DNA and RNA polymerases share a double psi-beta barrel fold, which is as ancient as the RNA recognition motif-containing fold or the palm domain fold found in viral RdRp and RT [4]. Following this event that enabled RNA genome replication, tracking individual RNA genome size expansion and shrinkage at the level of individual species becomes a daunting task. Nevertheless, RNA viruses evolved together with their hosts through a myriad of events and mechanisms [5].

In the tree of life, RNA viruses are the only group of organisms known to have an RNA genome. RNA genome sizes range from a few kb up to ~41 kb in length [6,7] in the case of (+)RNA viruses (Figure 1A,B). In this article, 'RNA virus' will refers to (+)RNA viruses unless specified otherwise.

Are there enzyme-specific markers of large RNA genomes?

The occurrence and distribution of enzyme-encoding genes according to virus genome size has been proposed to be an evolutionary marker in these large RNA genomes lying at this genome size limit [6]; indeed, a positive correlation exists between RNA virus genome sizes and the presence of RNA helicase [8] and RNA ExoN domains [6]. Although the presence of the helicase seems to be an ancestral acquisition dictated by a functional requirement for RNA viral genomes >6 kb in size [9], the ExoN gene was acquired through evolution to provide an RNA synthesis proofreading system [10–13] essential for large RNA genome stability (discussed later). In some large Nidovirales, ExoN is bound to a processive replicative RdRp and corrects mismatched bases appearing during viral RNA synthesis [12,13]. Over the past decade, a view of the overall repair process has been refined in coronaviruses [14,15]. Beyond these acquisitions, large genome nidoviruses and nidovirus-like [6,16] viruses reside at the upper boundary of the largest RNA genome sizes, and their members code for an unusually large number of RNA modification enzymes carrying functions matching that of small DNA viruses, such as DNA bacteriophages [17,18]. Thus, markers of genome expansion can be tracked by following the acquisition of new genes (proteins) directly involved in RNA 'housekeeping.'

Is RNA synthesis fidelity increasing with genome size?

Most RNA viruses code for their own RdRps, generally quoted as being of 'low fidelity.' In DNA-based organisms, the mutation rate (μ_g) stays rather constant (~0.003 mutations per genome per replication) for genomes differing by several orders of magnitude [19,20], and sometimes strikingly lower (~10^{-9} to 10^{-12}), as reported in more recent studies [21,22]. By contrast, a μ_g has been measured between ~0.8 and 6.5 mutations per genome per replication for (+)RNA viruses devoid of RNA repair system [23,24] and of ~0.075 for coronaviruses [23]. Thus, for large RNA genomes such as the latter, fidelity is higher by about one order of magnitude. Strikingly, the mutation rates per nucleotide per replication cycle (i.e., weighting with genome size) for small RNA genomes do exhibit a lower fidelity of replication [25]. One can thus ask if structural markers of RNA fidelity synthesis can be detected in RdRp sequences and structures and if they correlate with genome size.

The maximum possible intrinsic fidelity of viral RdRps

Fitzsimmons and colleagues suggested that viruses could have high mutation rates because it is hard to be simultaneously fast and accurate while facilitating adaptation [26]. Viral replication needs to be fast for survival purposes over being accurate, thus leading RNA viruses to be

Glossary

Epitranscriptomics: the study of post-transcriptional biochemical modifications occurring on RNA molecules.

Fidelity: a measurement of the adherence to Watson-Crick rules during nucleic acid synthesis. Fidelity is a property of the polymerase and is expressed as the frequency of deviation (i.e., misinsertion) from Watson-Crick base-pairing rules. It equals the mutation rate when expressed per nucleotide inserted per template site. It is often blurred by natural selection of resulting mutated genomes, which thus represent only a part of the generated genetic diversity.

Fitness: for viruses, the capacity to produce infectious progeny in a given environment. The definition is extended to the adaptation to their surrounding environment to survive, thrive, and reproduce for passing on the genetic information to the next generations.

Mutation rate: the frequency of appearance of nucleotide changes. It can be expressed per nucleotide per template site (i.e., fidelity, see earlier), or per nucleotide per replicated genome, or per nucleotide per replicated cell, or per cell passage. When the number of replicated genomes in a cell or organism is unknown, it is impossible to estimate how many viral genomes have disappeared as a result of counterselection in the cell host.

Nsp14: a bifunctional enzyme carrying both an RNA MTase in its C-terminus and a 3′ to 5′ exonuclease domain (ExoN) in its N-terminus, which excises mispaired bases ('errors') occurring during RNA synthesis and thus restores fidelity to levels compatible with a required genetic stability. Nsp14 ExoN is activated by nsp10, and both are products of expression of the coronavirus Orf1ab, further processed into nsp1–nsp16.

Processive: a nucleic acid polymerase is processive when it catalyzes nucleotide incorporation continuously without dissociation from its template. Processivity is defined as a number, which represents the number of nucleotides added to a growing strand per encounter of the polymerase with its nucleic acid substrate.

RNA viruses: viruses that have RNA as a genetic material. This RNA can be double-stranded (ds), such as reoviruses, or single-stranded (ss).

subject to other trade-offs of evolutionary significance. Thus, the question of RdRp fidelity in relation to genome size began to be indirectly addressed recently with the highly processive severe acute respiratory syndrome coronavirus (SARS-CoV) main RdRp complex made of nonstructural protein 12 (nsp12; RdRp) and nsp7/nsp8 (processivity factor). This complex, surprisingly, was found to exhibit a low fidelity of RNA synthesis in in vitro enzyme assays relative to other known RdRps belonging to smaller genome viruses [10]. This counterintuitive finding can be resolved as follows.In large Nidovirales, the RdRp complex is able to co-opt an RNA repair ExoN (carried in coronavirus nsp14 [13,14]). The current view, demonstrated in CoV-infected cells [10,11,27] and using purified enzyme assays [13,14,28], is that the SARS-CoV nsp12 enzyme has 'relaxed' its fidelity of nucleotide selection.When nucleotide misincorporation occurs, the nsp12-associated nsp14 ExoN, which is regulated by nsp10, restores accurate RNA synthesis through mismatch excision.

ssRNA viruses can have positive (+) sense ssRNA [abbreviated herein as (+) RNA viruses], whose functional mRNA directly translates into proteins (e.g., coronaviruses), as opposed to negative (−) sense ssRNA [reported as (−)RNA], which first produces mRNA and then translates into proteins (e.g., orthomyxoviruses). Unless specified otherwise, RNA viruses discussed here are (+)RNA viruses due to their immediate evolutionary connection of proto-RNA viruses (see 'The acquisition of enzymes for genome size expansion and stability' in the main text).

SAM-dependent RNA methyltransferases (RMTases): enzymes that use S-adenosyl-methionine (SAM) as the cofactor to methylate the RNA acceptor molecule, generating S-adenosyl-homocysteine (SAH) as a by-product.

Structural markers of fidelity corelate with large RNA virus genome size

Besides the co-optation of an exonuclease by the viral RdRp, structural determinants of RNA synthesis fidelity at the polymerase active site have only been investigated recently for large RNA genomes. However, structure- and sequence-based alignments of viral RdRps have identified seven conserved motifs, A to G, which constitute the core of the RdRp RNA synthesis machinery, and fidelity markers have been mapped mainly to structural motifs A, B, C, and F [29,30]. More specifically, Shannon et al. [31] have recently determined that, in parallel to its infidelity in RNA synthesis, the SARS-CoV RNA synthesis complex has an approximately tenfold faster nucleotide incorporation rate than any other known viral RdRp. This is a logical evolution to cope with synthesis of its large (~30 kb) RNA genome: the faster RNA synthesis rate, the lesser detection in a given time window by cellular innate immunity guardians.

Beyond this example, we argue that through analysis of sequence alignments and structure-based studies, markers of RdRp active site fidelity can be determined (Figure 2). At least two specific structural features are apparent in the RdRp's catalytic site that we argue corelate with genome size. These structural features are located in motif F and motif C. Small genome (+)RNA viruses, exemplified here by Picornaviridae and Flaviviridae, possess a long side-chain amino acid (E or Q) at the second amino acid position after the absolutely conserved first Lys of motif F. Arteriviridae have a Gln at this position, which is corelated with the presence of the large genome (+)RNA Nidovirales signature sequence SDD at motif C. We feel it is important to note that none of these viruses have an ExoN, and thus their RNA synthesis fidelity must be more strictly controlled at the nucleotide selection/insertion step than that of viruses having an ExoN. Upon increase in genome size, slightly below 20 kb in size (starting with CASV; Figure 2A), a short chain amino acid Ala (or Ser) appears in motif F, with a correlation with the presence of SDD in motif C and the presence of ExoN. The presence of this motif F Ala is a marker of RNA synthesis infidelity, as its short side chain leaves a wider space for the base of the incoming nucleotide [32,33]. In SARS-CoV, and together with an additional hydrogen bond provided by the Ser of the SDD motif, its presence has been suggested to be involved in the unusual nucleotide polymerization rate and infidelity measured during RNA synthesis in the absence of ExoN [31]. It is thus expected that large flaviviruses, which do not have ExoN, have expanded their genome size together with GDD, without evolving an Ala in motif F so far (Figure 2A,B). We therefore propose that large, Flavi-like genomes represent, up to now, the maximum achievable RNA synthesis fidelity without the help of an ExoN-mediated proofreading system. In other words, without presuming what actually drove genome expansion, one structural molecular signature for fidelity in the RdRp is located in motif C (GDD).

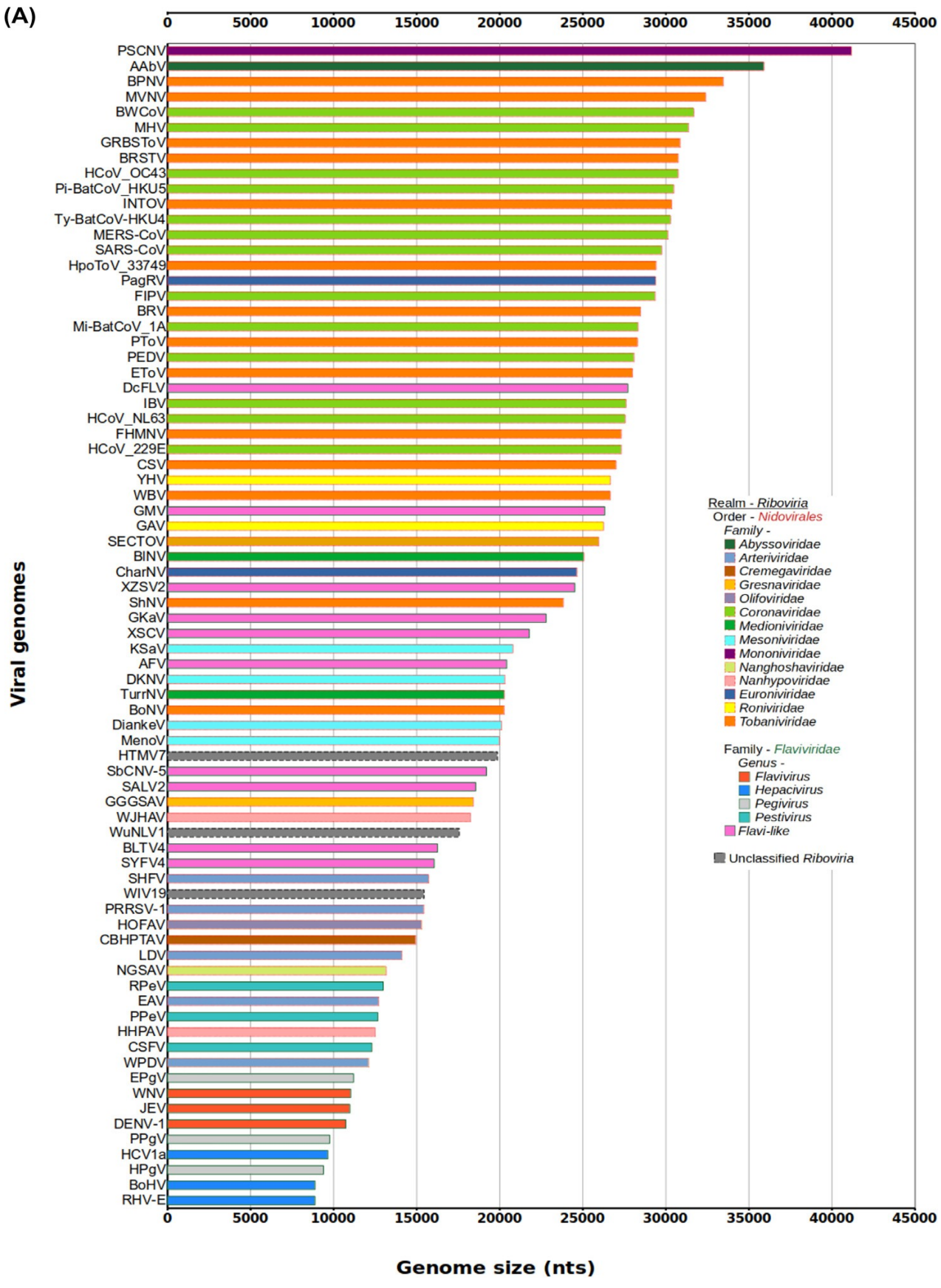


Figure 1. Large (+)RNA virus genomes sorted according to genome size, with their respective color-coded family or genus. (A) Nidovirales, Flaviviridae, and Flavi-like and some unclassified Riboviria having a genome >8000 nucleotides (nt) in length. Table S1. (B) Simplified view of (A), showing representative order, family, and genus. Created with BioRender. The abbreviated virus names used throughout the text and figures are given in Table S1 in the supplementary information online.

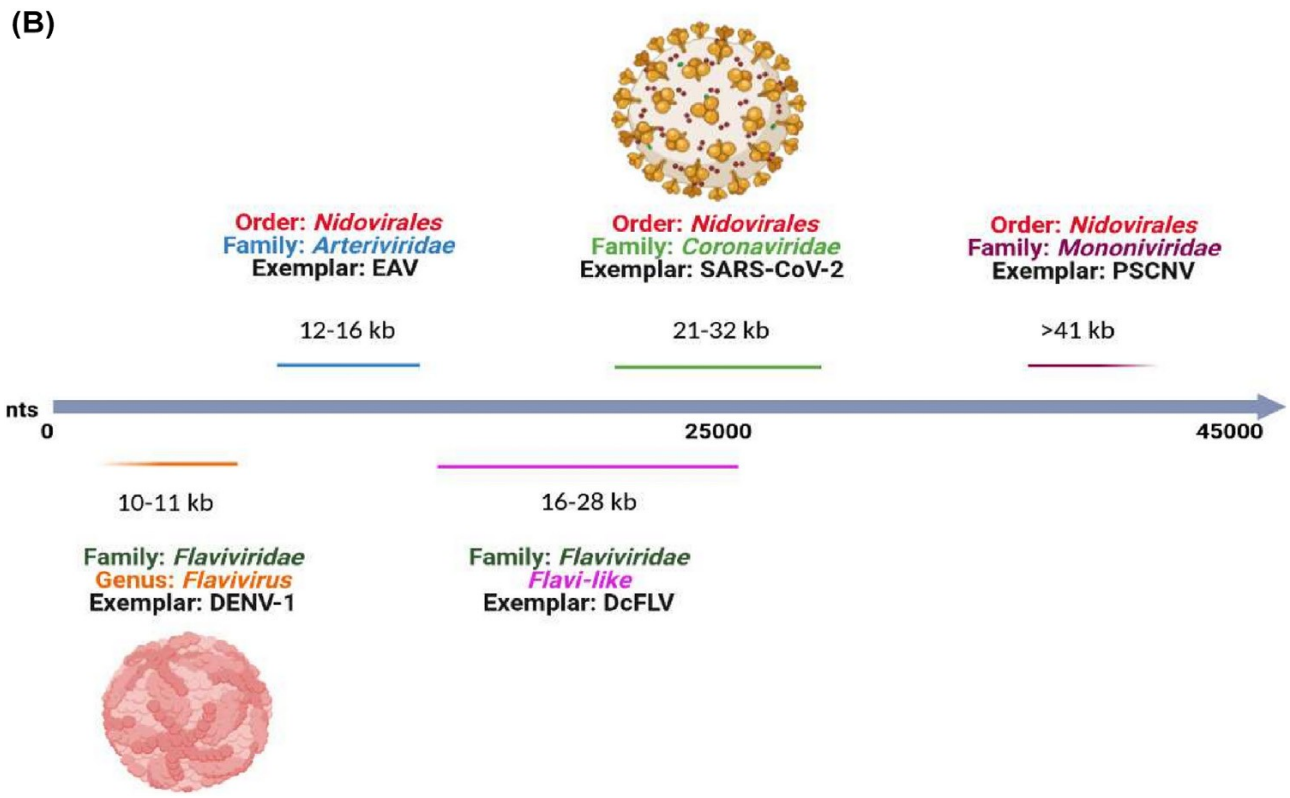


Figure 1 (continued).

Ongoing discoveries of large RNA virus genomes

Remarkably, the association of ExoN to Nidovirales RdRp complex validates a prediction made by Drake in 1993 [19]: 'RNA viruses would have to acquire several host genes and adapt them to RNA substrates to achieve a major reduction in spontaneous mutation rate. The result would be a substantial increase in genome size.' This prediction has been put in perspective with the recent discovery of the 41.2-kb RNA genome planarian secretory cell nidovirus (PSCNV) [7], which carries an ExoN signature sequence, but it is not known if the fidelity of its RdRp is increased relative to that of coronaviruses. Above 35 kb in genome size, only two genome sequences are known to have a GDD in motif C. For the only one >40 kb in size (PSCNV), a single genome sequence is available with a different motif F. Because no structural data are available for the RdRps of these two extremely long RNA genomes, we are left to admit that another yet unknown factor or feature of possible active site geometry may exist to promote additional RNA synthesis fidelity at these RdRp active sites.

Perhaps PSCNV RdRp increased RNA synthesis accuracy through evolution relative to coronaviruses while conserving the ExoN-mediated RNA repair capability. The discovery of these largest PSCNV RNA genomes suggests that other large RNA viruses will be discovered, carrying both a high-fidelity RdRp and matching ExoN activity. More sequences of >35 kb RNA genomes, as well as RdRp structures at atomic resolution, are needed to discover if specific structural marker(s) appear to address a putative need for highest RNA synthesis fidelity.

Novel large RNA virus genomes challenge the ExoN paradigm

Unless postreplicative mismatch repair or any other molecular system increasing RNA synthesis fidelity is discovered, we are at the dawn of discovering very large (>40 kb) RNA genomes generated and maintained by 'super'-accurate RdRps complemented by highly active RNA repair 3′ to 5′ exonucleases. Recent discoveries in the invertebrate virosphere challenge

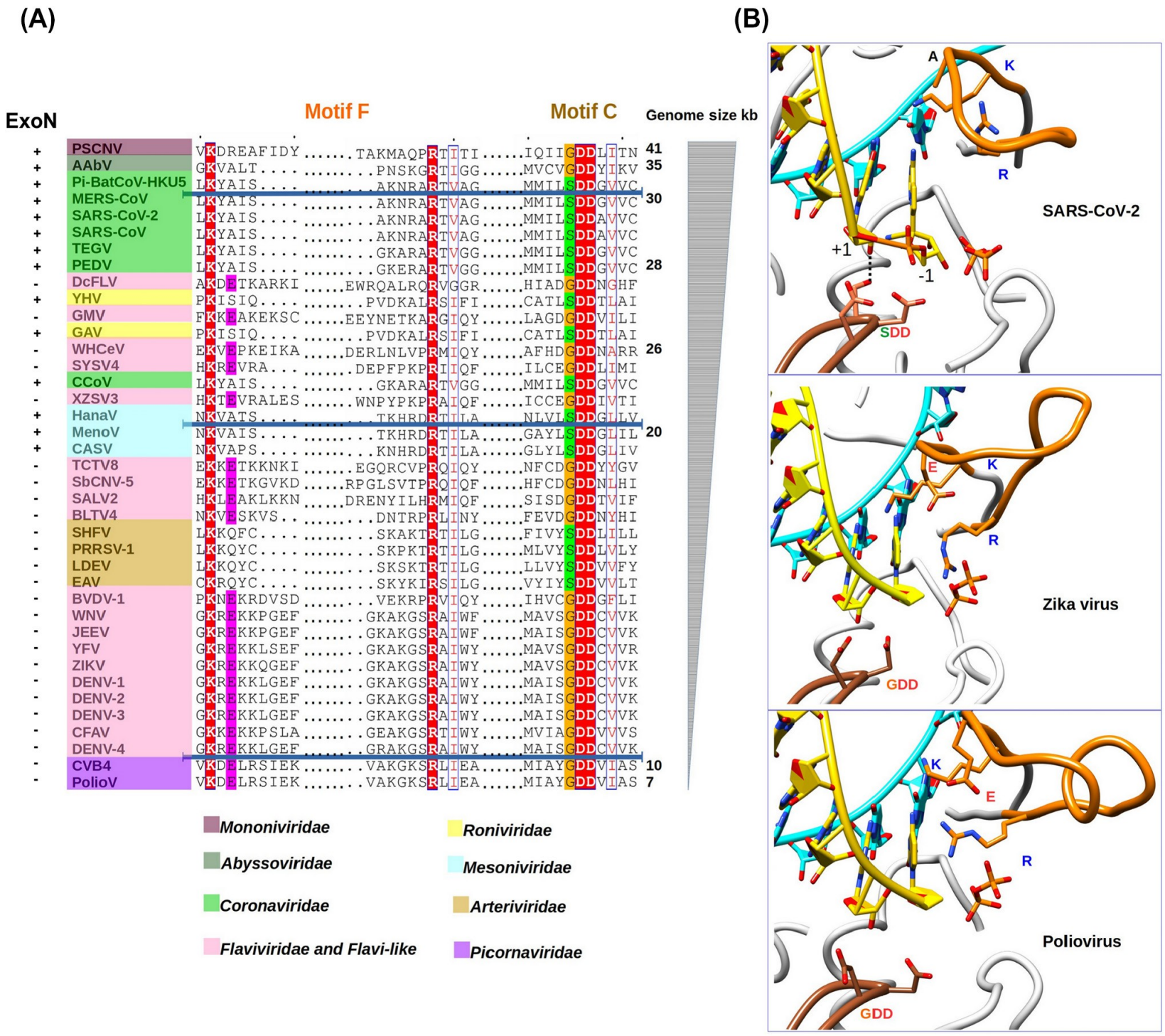


Figure 2. Sequences and structural features comparison of (+)RNA virus RNA-dependent RNA polymerase (RdRp) active sites order by genome size.

(A) Motifs F (orange) and C (brown) alignment of a selection of representative viral RdRp sequences sorted by genome length. The presence (+) or absence (−) of an exonuclease (ExoN) signature sequence is indicated on the left. Genome size is indicated and sorted by decreasing order on the right side of the table. Virus names are abbreviated according to the nomenclature (see Table S1 in the supplementary information online), and they are boxed in color according to their viral family. The catalytic motif C, which comes downstream of motif F, is either GDD or SDD. In motif F, the conserved lysine and arginine bordering the motif de fine the two NTP-interacting residues. The overall length of the motif F loop is variable from one family to the other and is illustrated by dots. The presence of a glutamate (E, purple box) in motif F is mainly corelated to the glycine (orange box) in motif C, else replaced by alanine, serine, or glutamine when the serine (green box) is observed in motif C. Alanine, conserved in Coronaviridae, is a marker of infidelity (see main text). (B) Conserved motifs F (orange) and C (brown) in RdRp structures of (top to bottom): severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)nonstructural protein 12 (Protein Data Bank accession no. 7BV2), Zika virus (PDB accession no. 5TFR), and poliovirus (PDB accession no. 3OI9). The RNA and pyrophosphate position in Zika virus were deduced from superimposition with the poliovirus RdRp structure. In motif F, the CoV polymerase carries an alanine (A547) instead of a glutamate residue (E161), removing a highly conserved interaction that positions the motif F arginine for interactions with the NTP and pyrophosphate. The serine (S759) of motif C (SDD) is able to interact with 2′-OH ribose in position −1 of the primer (broken line). Abbreviations: ExoN, exonuclease; RdRp, RNA-dependent RNA polymerase.

what is actually promoting genome size increase or simply assisting large genome size maintenance. In particular, the discovery of large Flavi/Pesti-like viruses has uncovered 27.7-kb RNA genomes devoid of any obvious ExoN signature sequence [34]. ExoN domains, however, are present in other viruses showing innate immunity suppression and indirectly connected to replication [e.g., arenavirus nucleoproteins (NPs) [35,36]]. In this latter case, genomes are segmented, though, and much smaller than those of large nidoviruses; their RNAs thus may not withstand the same constraints as large unsegmented (+)RNA genomes. It is thus logical to propose that other factors promote genetic stability of large RNA genomes: genetic stability should match chemical stability.

Large RNA viruses first solved a chemical RNA stability issue

Genetic stability cannot be achieved on a biophysically unstable genetic material: the longer an RNA molecule, the greater the probability that it would lose its coding functionalities. Bases could lose their coding capacity through modification or elimination, or, more simply, the RNA could be cleaved. Cleavage may happen through host-mediated enzymatic nucleolytic cleavage [37] or internal, cis-mediated self-cleavage through ribozyme-like reactions [38]. To overcome the poor RNA stability, acquiring enzymes enhancing the chemical RNA stability or protective proteins seems a logical evolutionary path. Accordingly, it is now well known that RNAs can carry >100 epitranscriptomic RNA modifications [39], and viral RNAs are being increasingly reported as carriers of these epitranscriptomic marks (reviewed in [40]).

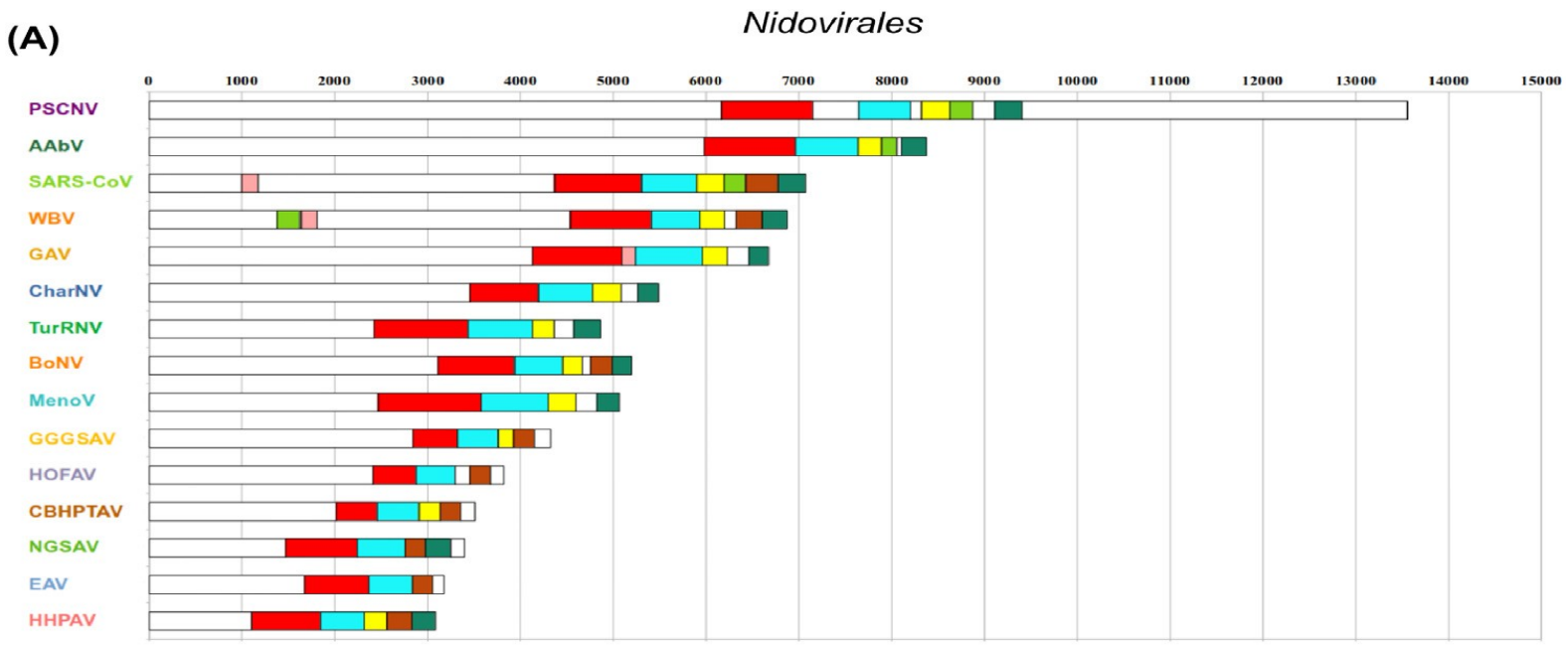
The 2′-hydroxy group stands out as a stability provider: it is a nucleophile promoting non-enzymatic alkaline internal RNA hydrolysis through a chemical mechanism shared by all of the naturally occurring, small, self-cleaving RNAs, by ribonuclease A and other ribonucleases [41], and either its 2′-O-methylation or its disappearance (in DNA) significantly increases nucleic acid stability. Likewise, SAM-dependent RNA methyltransferases (RMTases) are ancient protein folds susceptible to having been present early in evolution concomitantly with the RNA world [42]. Thus, RNA 2′-O-MTases are excellent candidates to be ancillary stability factors. In modifying RNA, RMTases may dampen the nucleophilic potential of chemical groups present in nucleobases, ribose, and phosphate that could promote intra- or intermolecular RNA cleavage.

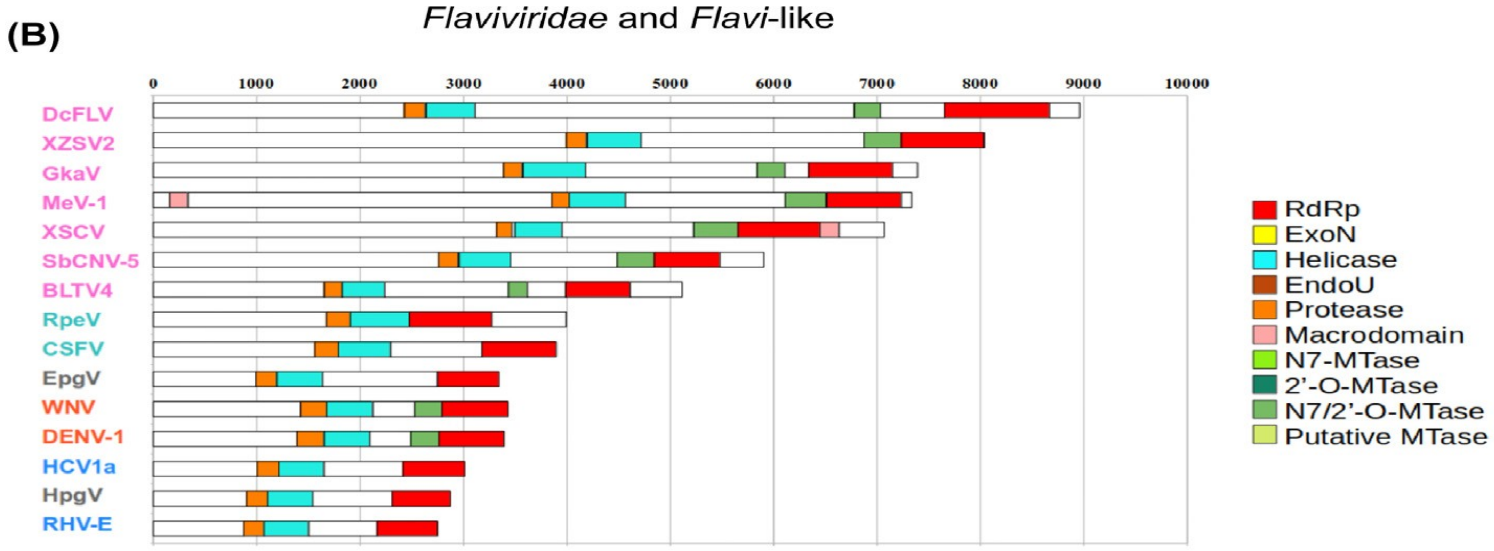
Large RNA genomes are generally associated with RNA MTases

The analysis of large viral RNA genomes [16,43], such as genomes of nidovirus (~12–41 kb), flavivirus (~8.8–13 kb), flavivirus-like (~16–27.7 kb), and some unclassified Riboviria resembling mamastroviruses (up to ~20 kb), indicates that above ~17 kb of contiguous RNA, all viral RNA genomes contain at least one detectable RNA MTase signature sequence (Figure 3). Currently known RNA genomes >27 kb always carry both one ExoN and at least one MTase signature sequence. Unlike flavivirus-like genomes, which are so far limited to <30 kb in size, acquiring both ExoN and RMTase is linked to nidovirus-like genome expansion and maintenance.

Viral RNA MTases are not solely involved in RNA capping

Two main MTase activities (RNA-cap N7-guanine and 2′-hydroxyl methylations) carried by viral enzymes participate in viral RNA capping [44]. However, among the three virus taxa mentioned earlier, two do not rely on conventional RNA capping: Mamastroviruses use a protein cap (VPg) at the 5′-end of viral RNA [45], pestiviruses (Flaviviridae) have no cap at all but rather an internal ribosome entry site (IRES) whose structure promotes protection and efficient translation [46], and the situation is far from being clear for Nidovirales regarding the presence of a canonical RNA cap. Furthermore, arterivirus members do not code for any RMTase nor capping enzyme, the presence and structure of an RNA cap has been demonstrated only for one virus of the





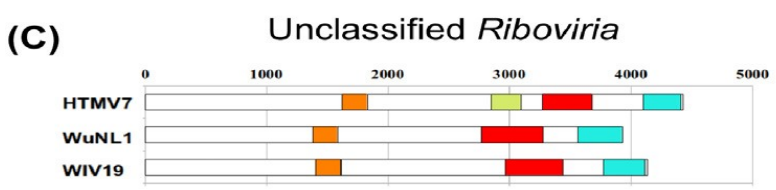


Figure 3. Polyproteins of the three representative taxa,sorted by size in amino acids. Enzyme domains were identified using the HHblits and HHpred tools of the Bioinformatics toolkit [59]. for highlighting the markers of large RNA genomes – ExoN, MTase(s). (A) Graph depicts the domain polyprotein 1ab (pp1ab) organization for representative virus (es) of the 14 families belonging to Nidovirales (genome size ~12–41 kb). When available, authentic cleavage sites were used to predict protein gene products of the Orf1ab, Orf1a, and Orf1b polyproteins (pp). The boundaries were otherwise approximately ±10 amino acids determined using structural homologies detected using HHPred, except for the N-terminal boundary of the Orf1b gene product. In Nidovirales, the absence of any structural data or homology (outside the order) on the N-terminus of the RNA-dependent RNA polymerase (RdRp) gene [nonstructural protein 9 (nsp9)Arteriviridae, nsp12 in Coronaviridae], which was used for phylogenic analysis, precludes a precise sequence homology search in this limited area between the nsp10 and nsp12 proteins (coronavirus gene product naming). (B) Graph depicts the pp organization for different enzyme domains of Flaviviridae viruses (genome size 8.8–27.7kb). To represent the diversity, few members of each of the genus – Hepacivirus, Pegivirus, Flavivirus, Pestivirus, and the longer flavivirus-like have been shown. (C) Graph depicts the pp organization for different enzyme domains of three of the unclassified Riboviria members (up to ~20 kb). Abbreviations: RdRp, RNA dependent RNA polymerase; ExoN, exonuclease; EndoU, endoribonuclease; MTase, methyltransferase. The abbreviated virus names are given in Table S1.

Torovirus [47], and an RNA cap has only been inferred in other members of Nidovirales RNAs due to the presence of RMTase sequences [44].

In parallel, it is interesting to note that several recent reports have shown that viral 2′-O-MTase specificities are not limited to RNA caps. The Ebolavirus MTase domain of the L protein methylates internal adenosine residues in RNA [48]. Similar observations had also been reported for Zika and Dengue viruses [49,50]. It is thus tempting to speculate that these Nidovirales and Flavi-like MTase substrate specificities are much wider beyond that of a viral RNA cap. They could indeed be methylating viral RNA internally and provide increased stability. Also, because one could argue that not all virus genomes of 10–11 kb in size encode an RMTase, we propose, using the same logic as for the helicase, that it provides a selective advantage, as demonstrated and mentioned earlier for Zika and Dengue viruses.

A 'battle' of RNA MTases in the infected cell

Recent reports mention host cellular MTases acting on the invading viral RNA [40]. For many viruses, these epitranscriptomic marks on viral RNA were described to play key roles in regulating several viral functions, including gene expression. In cellular methylation, the N^6-methyladenosine is an abundant RNA modification found in viral RNAs (for review, see [51]). These modifications induce pleiotropic function, which is regulated during the virus replication steps. For example, in influenza, N^6-adenosine MTases have been shown to regulate splicing and export of viral RNA, genome packaging, and positively or negatively impact viral gene expression [52]. In flaviviruses and others (reviewed in [50]), N^6-methyladenosine dampens viral expression (antiviral effect). Conversely, during an HIV infection, FTSJ3, a cellular RNA 2′-O-MTase, is recruited by transactivation response element RNA-binding protein (TRBP) to methylate HIV RNA at 17 positions [53]. These methylations seem to have a proviral effect because they have been reported to favor viral escape from detection by MDA5 and, in turn, the secretion of type I interferon. Thus, cellular and viral RNA MTases may well compete for RNA methylation in order to establish their respective phenotypic outcomes.

RNA processing enzymes, LUCRA, and a possible bacteriophage connection

Could genomes of large (+)RNA viruses share vestigial features with the LUCRA? And would there be any clues of the RNA-to-DNA transition remaining in the largest (+)RNA viral genomes? Or did the RNA-to-DNA transition occur in small protoviruses, thus suggesting that large genome RNA viruses represent an evolutionary dead end after which size increase was no longer possible?

The appearance of DNA from RNA or precursors requires two essential enzymes: the thymidine synthase (TS), and the ribonucleotide reductase (RR). The RR is a metalloenzyme proceeding through free radical chemistry and metal-sulfur clusters [54], the latter metal clusters being of unusual abundance in large Nidovirales genomes [55]. The TS is an unusual MTase, catalyzing radical-based synthesis of dTMP from dUMP, by means of 5,10-methylenetetrahydrofolate (not SAM) as the methyl donor [56]. Both enzymes are widely distributed in all kingdoms of life, including DNA bacteriophages. Interestingly, the large genome coronavirus 229E carries a significant amount of 5-methylcytosine, a precursor of thymidine [57], bearing an amino group in position 4. This indicates either that CoV RdRp is insensitive to base methylation of CTP (and potentially other NTP substrates) or that coronavirus RNA genomes show chemical proximity to DNA without apparent harm. Large Nidovirales RNA genomes share various features (i.e., activity embedded in the same protein folding) with DNA bacteriophages: endonucleases, exonucleases, primase/processivity factors, SAM-dependent MTases, helicases, polymerase I–type polymerases, and the oligonucleotide/oligosaccharide-binding fold single-strand binding protein (SSB) nsp9 of large coronaviruses [17,18]. The latter nsp9 has no homologue

in the RNA virus world but is structurally similar to,for example, the Escherichia coli T7 and T4 bacteriophage SSB gene 2.5 and gene 32, respectively [17].

Concluding remarks

Due to their possible connection to the primordial RNA world, (+)RNA viruses could be considered as a special evolutionary case in the virosphere. Evolutionary forces, however, may have dictated to large proto-RNA viruses other solutions than embracing the DNA world: For (−)RNA viruses, genome segmentation (arena-, bunya-, and orthomyxoviruses) could be one, preventing a putatively costly maintenance of stable long RNA stretches while also providing an evolutionary benefit through genome reassortment [58]; the use of NP protecting the constantly attacked genomic RNA could be another means to offer some protection. Be it related or not, large Nidovirales are the only (+)RNA virus genomes encoding their own NP for genome encapsidation.

We argue that size expansion and maintenance of large replicating RNA molecules resembling (+)RNA viruses have been promoted through both evolving an accurate RdRp enhanced by ExoN and RNA MTase–mediated stabilization (see Outstanding questions). Only a concerted improvement of RNA stability and RNA synthesis fidelity through these three essential enzymes might have been responsible, at least in part, for genome expansion and maintenance (Figure 4). Future studies on large (+)RNA genomes and their enzymes represent an exciting set of challenges ahead

Outstanding questions

What is the workload of the ExoN involved in RNA synthesis proofreading?

How accurate can a viral RdRp core be?

What are the substrate specificities of RMTases from large RNA viruses?

Does the highest RdRp fidelity correspond to the largest RNA genome?

Are larger RNA virus genomes to be discovered?

What would be further gene markers in LUCRA?

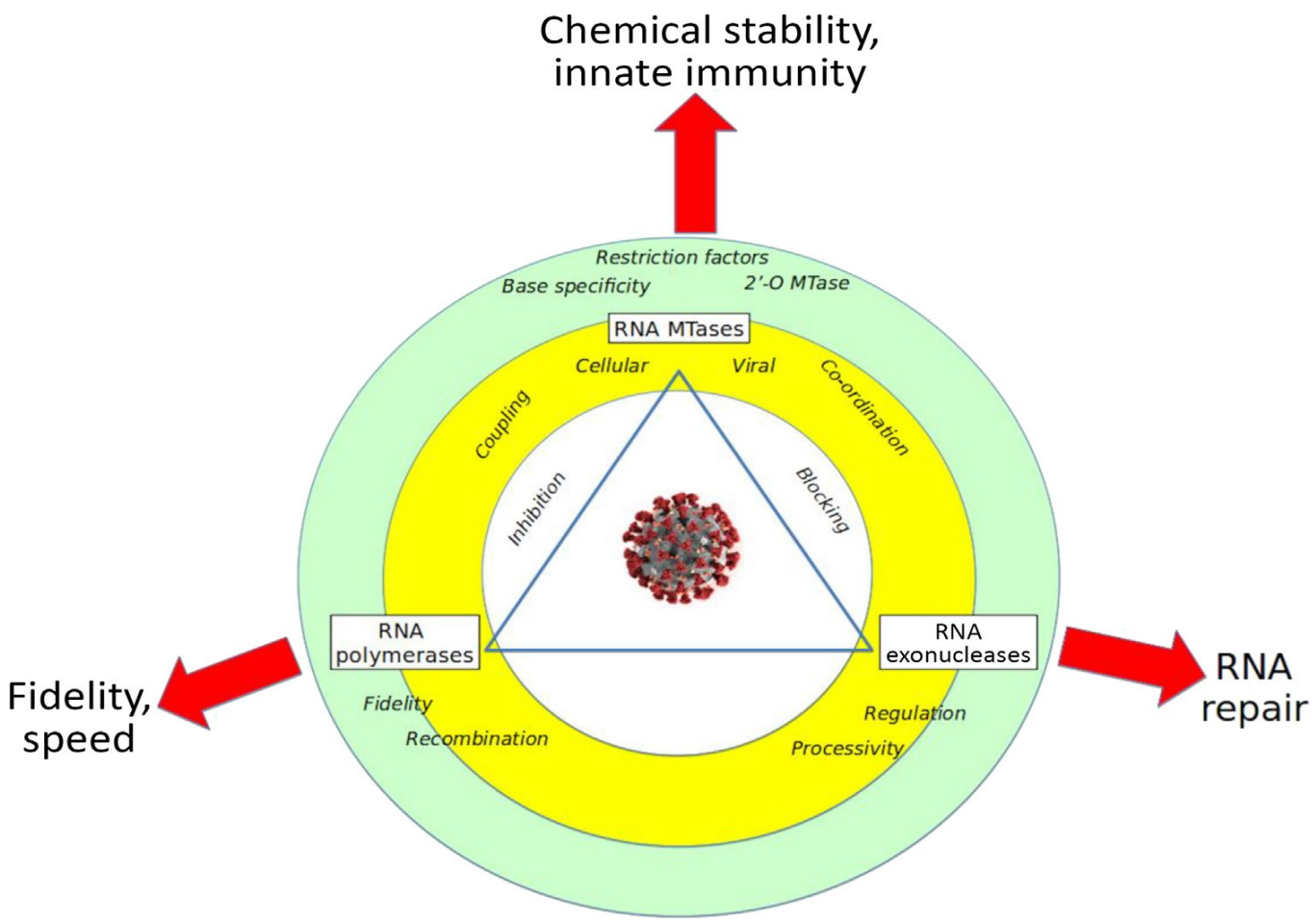


Figure 4. A diagram showing evolutionary forces at work through synergy of three enzymes [RNA-dependent RNA polymerase (RdRp), exonuclease, and RNA methyltransferase (MTase)] to promote genome size increase and maintenance in (+)RNA viruses. Red arrows represent three positive forces for successful genome expansion through evolution. The respective measurable enzyme properties are represented in the inner circles, together with their resulting effects on viral RNA genomes. We argue that large (+)RNA genomes result from the presence, action, and synergy of these three enzymatic activities.

to fully understand the astonishing genome plasticity and evolutionary capability of RNA viruses. These studies will certainly serve as science-based control measures against pathogenic RNA viruses, such as predicting the pandemic potential of emerging viruses, vaccine design, and drug resistance. In the current context of the SARS-CoV-2 pandemic and the recent shining success of mRNA-based vaccines, it seems worth investigating RNA modification enzymes promoting RNA stability, even though they may originate from exotic, nonpathogenic RNA viruses.

Acknowledgments

We thank Professor Olve Peersen and Dr Ashleigh Shannon for their thoughtful input and discussion of viral RNA-dependent RNA polymerase active sites and the viral RNA replication/transcription process in general. This work was supported by the Fondation pour la Recherche Médicale (FRM; Aide aux Équipes 2019-2022) and the Fondation Mediterranée Infection (Infectiopole Sud).

Declaration of interests

The authors have no interests to declare.

Supplemental information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.tibs.2021.05.006.

References

1. Agol, V.I. and Gmyl, A.P. (2018) Emergency services of viral RNAs: repair and remodeling. Microbiol. Mol. Biol. Rev. 82, e00067-17
2. Cech, T.R. (2012) The RNA worlds in context. Cold Spring Harb. Perspect. Biol. 4, a006742
3. Krupovic, M. et al. (2019) Origin of viruses: primordial replicators recruiting capsids from hosts. Nat. Rev. Microbiol. 17, 449–458
4. Koonin, E.V. et al. (2020) The replication machinery of LUCA: common origin of DNA replication and transcription. BMC Biol. 18, 61
5. Wolf, Y.I. et al. (2018) Origins and evolution of the global RNA virome. mBio 9, e02329-18
6. Lauber, C. et al. (2013) The footprint of genome architecture in the largest genome expansion in RNA viruses. PLoS Pathog. 9, e1003500
7. Saberi, A. et al. (2018) A planarian nidovirus expands the limits of RNA genome size. PLoS Pathog. 14, e1007314
8. Koonin, E.V. et al. (2015) Origins and evolution of viruses of eukaryotes: the ultimate modularity. Virology 479–480, 2–25
9. Gorbalenya, A.E. and Koonin, E.V. (1989) Viral proteins containing the purine NTP-binding sequence pattern. Nucleic Acids Res. 17, 8413–8440
10. Eckerle, L.D. et al. (2007) High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. J. Virol. 81, 12135–12144
11. Denison, M.R. et al. (2011) Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA Biol. 8, 270–279
12. Bouvet, M. et al. (2012) RNA 3′-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. Proc. Natl. Acad. Sci. U. S. A. 109, 9372–9377
13. Ferron, F. et al. (2018) Structural and molecular basis of mismatch correction and ribavirin excision from coronavirus RNA. Proc. Natl. Acad. Sci. U. S. A. 115, E162–E171
14. Subissi, L. et al. (2014) One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. Proc. Natl. Acad. Sci. U. S. A. 111, E3900–E3909
15. Smith, E.C. et al. (2014) Thinking outside the triangle: replication fidelity of the largest RNA viruses. Annu. Rev. Virol. 1, 111–132
16. Shi, M. et al. (2016) Redefining the invertebrate RNA virosphere. Nature 540, 539–543
17. Egloff, M.-P. et al. (2004) The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world. Proc. Natl. Acad. Sci. U. S. A. 101, 3792–3796
18. Smith, E.C. and Denison, M.R. (2013) Coronaviruses as DNA wannabes: a new model for the regulation of RNA virus replication fidelity. PLoS Pathog. 9, e1003760
19. Drake, J.W. (1993) Rates of spontaneous mutation among RNA viruses. Proc. Natl. Acad. Sci. U. S. A. 90, 4171–4175
20. Drake, J.W. (1991) A constant rate of spontaneous mutation in DNA-based microbes. Proc. Natl. Acad. Sci. U. S. A. 88, 7160–7164
21. Sung, W. et al. (2012) Drift-barrier hypothesis and mutation-rate evolution. Proc. Natl. Acad. Sci. U. S. A. 109, 18488–18492
22. Ossowski, S. et al. (2010) The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science 327, 92–94
23. Sanjuán, R. et al. (2010) Viral mutation rates. J. Virol. 84, 9733–9748
24. Drake, J.W. and Holland, J.J. (1999) Mutation rates among RNA viruses. Proc. Natl. Acad. Sci. U. S. A. 96, 13910–13913
25. Bradwell, K. et al. (2013) Correlation between mutation rate and genome size in riboviruses: mutation rate of bacteriophage Qβ. Genetics 195, 243–251
26. Fitzsimmons, W.J. et al. (2018) A speed-fidelity trade-off determines the mutation rate and virulence of an RNA virus. PLoS Biol. 16, e2006459
27. Eckerle, L.D. et al. (2010) Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS Pathog. 6, e1000896
28. Bouvet, M. et al. (2010) In vitro reconstitution of SARS-coronavirus mRNA cap methylation. PLoS Pathog. 6, e1000863
29. Peersen, O.B. (2017) Picornaviral polymerase structure, function, and fidelity modulation. Virus Res. 234, 4–20
30. Selisko, B. et al. (2018) Structural and functional basis of the fidelity of nucleotide selection by flavivirus RNA-dependent RNA polymerases. Viruses 10, 59
31. Shannon, A. et al. (2020) Rapid incorporation of favipiravir by the fast and permissive viral RNA polymerase complex results in SARS-CoV-2 lethal mutagenesis. Nat. Commun. 11, 4682
32. Gong, P. and Peersen, O.B. (2010) Structural basis for active site closure by the poliovirus RNA-dependent RNA polymerase. Proc. Natl. Acad. Sci. U. S. A. 107, 22505–22510

33. Peersen, O.B. (2019) A comprehensive superposition of viral polymerase structures. Viruses 11, 745
34. Matsumura, E.E. et al. (2016) Complete genome sequence of the largest known flavi-like virus, Diaphorina citri flavi-like virus, a novel virus of the Asian citrus psyllid, Diaphorina citri. Genome Announc. 4, e00946-16
35. Hastie, K.M. et al. (2011) Structure of the Lassa virus nucleoprotein reveals a dsRNA-specific 3′ to 5′ exonuclease activity essential for immune suppression. Proc. Natl. Acad. Sci. U. S. A. 108, 2396–2401
36. Papageorgiou, N. et al. (2020) Brothers in arms: structure, assembly and function of Arenaviridae nucleoprotein. Viruses 12, 772
37. Hartmann, G. (2017) Nucleic acid immunity. Adv.Immunol. 133, 121–169
38. Jimenez, R.M. et al. (2015) Chemistry and biology of self-cleaving ribozymes. Trends Biochem. Sci. 40, 648–661
39. Helm, M. and Motorin, Y. (2017) Detecting RNA modifications in the epitranscriptome: predict and validate. Nat. Rev. Genet. 18, 275–291
40. Netzband, R. and Pager, C.T. (2020) Epitranscriptomic marks: emerging modulators of RNA virus gene expression. Wiley Interdiscip. Rev. RNA 11, e1576
41. Fedor, M.J. (2000) Structure and function of the hairpin ribozyme. J. Mol. Biol. 297, 269–291
42. Rana, A.K. and Ankri, S. (2016) Reviving the RNA world: an insight into the appearance of RNA methyltransferases. Front. Genet. 7, 99
43. Shi, M. et al. (2016) Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the Flaviviridae and related viruses. J. Virol. 90, 659–669
44. Decroly, E. et al. (2011) Conventional and unconventional mechanisms for capping viral mRNA. Nat. Rev. Microbiol. 10, 51–65
45. Mendez, E. and AAstroviruses. In Fields Virology (Vol. 1, 5th edn) (Knipe, D.M. and Howley, P.M., eds), pp. 981–1000, Lippincott Williams & WilkinsLippincott Williams & Wilkins, pp. 981–1000Astroviruses. In Fields Virology (Vol. 1, 5th edn) (Knipe, D.M. and Howley, P.M., eds), pp. 981–1000, Lippincott Williams & Wilkins
46. Martínez-Salas, E. et al. (2008) New insights into internal ribosome entry site elements relevant for viral gene expression. J. Gen. Virol. 89, 611–626
47. van Vliet, A.L.W. et al. (2002) Discontinuous and non-discontinuous subgenomic RNA transcription in a nidovirus. EMBO J. 21, 6571–6580
48. Martin, B. et al. (2018) The methyltransferase domain of the Sudan ebolavirus L protein specifically targets internal adenosines of RNA substrates, in addition to the cap structure. Nucleic Acids Res. 46, 7902–7912
49. Dong, H. et al. (2012) 2′-O methylation of internal adenosine by flavivirus NS5 methyltransferase. PLoS Pathog. 8, e1002642
50. McIntyre, W. et al. (2018) Positive-sense RNA viruses reveal the complexity and dynamics of the cellular and viral epitranscriptomes during infection. Nucleic Acids Res. 46, 5776–5791
51. Imam, H. et al. (2020) Epitranscriptomic(N6-methyladenosine) modification of viral RNA and virus-host interactions. Front. Cell. Infect. Microbiol. 10, 584283
52. Courtney, D.G. et al. (2017) Epitranscriptomic enhancement of influenza A virus gene expression and replication. Cell Host Microbe 22, 377–386.e5
53. Ringeard, M. et al. (2019) FTSJ3 is an RNA 2′-O-Methyltransferase recruited by HIV to avoid innate immunity sensing. Nature 565, 500–504
54. Nordlund, P. and Reichard, P. (2006) Ribonucleotide reductases. Annu. Rev. Biochem. 75, 681–706
55. Snijder, E.J. et al. (2003) 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
56. Costi, M.P. et al. (2005) Thymidylate synthase structure, function and implication in drug discovery. Curr. Med. Chem. 12, 2241–2258
57. Viehweger, A. et al. (2019) Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome Res. 29, 1545–1554
58. Ojosnegros, S. et al. (2011) Viral genome segmentation can result from a trade-off between genetic content and particle stability. PLoS Genet. 7, e1001344
59. Zimmermann, L. et al. (2018) A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J. Mol. Biol. 430, 2237–2243