The global virome: How much diversity and how many independent origins?

Eugene V Koonin, Mart Krupovic, Valerian Dolja

The global virome: how much diversity and how many independent origins?

Eugene V. Koonin[1†], Mart Krupovic[2,†] and Valerian V. Dolja[3]

[1]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA.
[2]Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, Paris, France.
[3]Department of Botany and Plant Pathology, Oregon State University, OR, USA.

[†]e-mail: koonin@ncbi.nlm.nih.gov; mart.krupovic@pasteur.fr

Summary

Viruses are considered to be the most abundant biological entities on earth. They also display striking genetic diversity as emphatically demonstrated by the recent advances of metagenomics and metatranscriptomics. But what are the limits of this diversity, that is, how many virus species in the earth virome? By combining the available estimates of the number of prokaryote species with those of the virome size, we obtain back-of-the-envelope estimates of the total number of distinct virus species, which come out astronomically large, from about 10^7 to about 10^9. The route of virus origins apparently involved non-viral replicators capturing and exapting various cellular proteins to become virus capsid subunits. How many times in the history of life has this happened? In other words, how many realms of viruses, the highest rank taxa that are supposed to be monophyletic, comprise the global virome? We argue that viruses emerged on a number (even if far from astronomical) independent occasions, so the number of realms will considerably increase from the current 6, by splitting some of the current realms, giving the realm status to some of the currently unclassified groups of viruses and discovery of new distinct groups.

Viruses are often considered to be the most numerous entities in the global biosphere. The most common estimates suggest that there are on the order of 10^{31} virus particle on the planet at any given moment, about an order of magnitude greater than the total number of cells (Mushegian, 2020). To the best of our current understanding, all organisms on earth are hosts to multiple viruses, with the possible exception of some endosymbiotic bacteria (Koonin et al., 2020). Empirical observations on the ubiquity of viruses are buttressed by theoretical argument on the inevitable emergence of genetic parasites in any replicator system (Koonin et al., 2017). Virus genomes are also extremely diverse, and the ongoing metagenomic-metatranscriptomic revolution reveals the vast scale of that diversity (Dutilh, 2014; Dolja and Koonin, 2018). The case of RNA viruses can serve as an apt illustration. Astonishingly, analysis of a single metatranscriptome, apparently coming from an environment rich in unicellular eukaryotes hosting RNA viruses, resulted in a twofold expansion of the known RNA virome (Wolf et al., 2020). Three independent subsequent studies exploring thousands of metatranscriptomes from diverse environments (Edgar et al., 2022; Neri et al., 2022; Zayed et al., 2022) each led to further, several fold increase in the number of known distinct RNA viruses (distinct, in this case, means not too closely related to each other, more specifically, clusters of genomes with similar sequences that roughly correspond to a virus species level), which combined, would amount to a more than an order

of magnitude expansion. Rarefaction analysis shows that saturation of the RNA virus diversity is not yet in sight (Neri et al., 2022). Metagenomic studies indicate that the case of DNA viruses is similar, and expansion of some groups, for instance, tailless bacteriophages (Kauffman et al., 2018; Yutin et al., 2018a), or tailed phages of the expansive order *Crassvirales* (Guerin et al., 2018; Yutin et al., 2018b; Yutin et al., 2021) has been even more dramatic.

So how many distinct viruses, or virus species, are there in the global virome altogether? Given that metagenomic and metatranscriptomic analyses (below we refer to these collectively as metaviromics insofar as applied to virus discovery) are not yet approaching saturation, this number cannot be inferred by extrapolation from available data. However, to obtain a rough, back of the envelope estimate, we can take a different approach modeled over that employed previously to estimate the number of unique microbial genes (Wolf et al., 2016). The great majority of viruses on earth are tailed and tailless phages infecting bacteria; viruses of archaea and eukaryotes are only relatively small additions. Let us conservatively assume that there are 10^6 to 10^7 bacterial species on earth (Curtis et al., 2002) (some estimates are orders of magnitude higher (Locey and Lennon, 2016)). Most if not all bacteria are hosts to multiple viruses. For *Escherichia coli* alone, about a hundred bacteriophages have been identified (Maffei et al., 2021), whereas for *Mycobacterium smegmatis* mc2155, more than 10,000 individual mycobacteriophages have been isolated, although only 2,100 of these have been sequenced and thus it remains to be determined how many different virus species they represent (Hatfull, 2022). Furthermore, analysis of CRISPR spacers, the majority of which appear to be virus-derived but do not match known viruses, implies large, host species-specific viromes (Shmakov et al., 2020). Let us assume 10 to 100 virus species per host species as a conservative estimate. Then, the size of the global virome can be crudely estimated at 10^7 to 10^9 distinct virus species – obviously, even the low bound in this range, probably, a vast underestimate, is a huge number. The upper bound appears more realistic, so there is likely to be about a billion virus species if not more on earth – evidently, a long way to go from the currently recognized 10^4 species (Walker et al., 2022) until we know them all.

As a bonus, we can also roughly estimate the size of the virus pangenome, in other words, the total number of genes in the virosphere. Large viruses encompass many poorly conserved, species-specific genes that obviously represent the bulk of the virus pangenome. Assuming 10 such unique genes per virus species, there would be 10^8 to 10^{10} unique virus genes altogether, a vast gene repertoire, to put it modestly.

Having obtained some striking, even if rough numbers characterizing the vastness of the virosphere, we now ask a different question that is likely to appear burning to anyone interested in virus diversity and evolution: how many times have viruses evolved independently? Recent analyses of the provenance of virus hallmark genes (VHG) encoding the key proteins involved in virus replication and virion formation suggest the principal route of virus origin: non-viral selfish replicators capture of host proteins that are exapted (repurposed) to become virion components, in particular, capsid proteins (CP) (Krupovic and Koonin, 2017; Krupovic et al., 2019; Koonin et al., 2022). This chimeric origin scenario, most likely, applies to the ultimate origin of viruses concomitant with the earliest stages of the evolution of cells and was recapitulated at least several times at later stages. But how many times did it happen, actually? Is origin of viruses by this chimeric route a common, routine or an extremely rare event? A single common origin of viruses appears to be out of the question, simply, because viruses do not share a single universally conserved gene (Koonin et al., 2020). This is a sharp contrast with cellular life forms which all share about 100 universal genes and are confidently traced back to the Last Universal Cellular Ancestor (LUCA) (Koonin, 2003). Clearly, there has been no last universal virus ancestor and viruses were already quite diverse at the time of LUCA (Krupovic et al., 2020). But then, how many independent ancestors for the contemporary global virome? In the comprehensive hierarchical virus taxonomy recently adopted by the International Committee

on Virus Taxonomy (ICTV), there are 6 taxa at the highest level, realms, each of which is thought to be monophyletic (International Committee on Taxonomy of Viruses Executive Committee, 2020; Koonin et al., 2020). In this respect, each realm is equivalent to the entirety of cellular life forms: there is a conserved core of VHG that defines the identity of the respective realm, even if it consists of a single gene.

There are four expansive realms of viruses: *Riboviria*, *Monodnaviria*, *Duplodnaviria* and *Varidnaviria*, each including a huge diversity of viruses that is rapidly growing through large scale metaviromics studies (Koonin et al., 2020). The realm *Riboviria* includes viruses with positive-sense, negative-sense and double-strand (ds) RNA genomes, together with the reverse-transcribing viruses that possess either RNA or DNA genomes. The viruses in this vast realm are unified by the homologous RNA-dependent RNA polymerases (RdRPs) and reverse transcriptases (RTs). The realm *Monodnaviria* consists of single-stranded (ss) DNA viruses and small dsDNA viruses (papillomaviruses and polyomaviruses), largely, with circular genomes. The single hallmark gene that holds this realm together encodes a distinct endonuclease (or its inactivated derivative) that is involved in the initiation of the genome replication, mostly, via the rolling circle replication mechanism. The realm *Duplodnaviria* includes tailed bacterial and archaeal viruses with dsDNA genomes, along with animal herpesviruses and the recently discovered mirusviruses (Gaïa et al., 2022). All these viruses share a distinct structural gene module that encodes the HK97-fold major capsid protein (MCP), genome packaging ATPase-nuclease (terminase), portal protein and capsid maturation protease. The second realm of viruses with dsDNA genomes, *Varidnaviria*, consists of diverse viruses infecting bacteria, archaea and eukaryotes that are unified by the vertical jelly-roll MCPs. Most viruses in this realm belong to the kingdom *Bamfordvirae* and possess double jelly-roll (DJR) MCP, but viruses in the smaller kingdom *Helvetiavirae* feature two MCPs each consisting of a single vertical jelly-roll domain. In addition, two small virus realms, *Adnaviria* and *Ribozyviria*, were recently formally recognized by the ICTV. The realm *Adnaviria* consists of filamentous and rod-shaped viruses that infect hyperthermophilic archaea of the phylum *Thermoproteota* and encapsidate linear dsDNA in the A-form (Krupovic et al., 2021). The realm *Ribozyviria* includes human hepatitis delta virus and its relatives infecting other animals (Hepojoki et al., 2021). The genomes of the ribozyviruses are viroid-like circular RNAs encoding a nucleocapsid protein (Delta antigen).

The virus realms are supposed to be monophyletic, that is, to have evolved from a single common ancestor virus. However, even apart from the fact that the monophyly hinges only on a few or even one VHG, this is not actually the case for at least three of the four current major realms. In particular, although the two kingdoms of the realm *Riboviria* share a homologous replicative enzyme, the origins of the viruses themselves, resulting from the recruitment of distinct capsid proteins, were clearly independent (Figure 1). Thus, under the premise that taxa should be strictly monophyletic, the realm *Riboviria* ought to be split into two or even three separate realms, given that two groups of *Pararnavirae*, namely, *Ortervirales* and *Blubervirales*, apparently evolved from distinct retrotransposons (Gong and Han, 2018; Krupovic et al., 2018). The case of the realm *Monodnaviria* is even more striking: viruses that currently comprise this realm evolved on at least four independent occasions via parallel routes, namely, capture, by small rolling-circle plasmids, of either a virus gene encoding a CP or cellular genes repurposed as CPs (Kazlauskas et al., 2019) (Figure 1). In the realm *Varidnaviria*, the two single jelly-roll vertical CPs of helvetiaviruases initially were presumed to be an ancestral form of the MCP, giving rise to the DJR MCP through gene fusion. However, recent analysis of the evolution of the vertical jelly-roll MCPs from cellular ancestors that revealed independent capsid origins in *Bamfordvirae* and *Helvetiavirae*, strongly suggests a split of this realm (Krupovic et al., 2022). The monophyly of the remaining three virus realms, *Duplodnaviria*, *Adnaviria* and *Ribozyviria*, seems to hold for the time being. However, there are already a number of viruses that appear to be unrelated to the established realms and hence probably evolved independently, thus being

candidates for new, small realms. These distinct enigmatic viruses include several virus families infecting archaea, in particular, viruses with lemon-shaped virions (Krupovic et al., 2014; Wang et al., 2022); animal anelloviruses that have small ssDNA genomes but appear to lack the hallmark rolling-circle endonuclease which unifies the members of *Monodnaviria* (Taylor et al., 2022); large dsDNA viruses of the class *Naldaviricetes* (baculo-like viruses) infecting diverse arthropods (Petersen et al., 2022); and a variety of viruses with circular RNA genomes, discovered mostly by metatranscriptome mining, that resemble viroids but encode proteins unrelated to the nucleocapsid protein of ribozyviruses (Linnakoski et al., 2021; Lee et al., 2022).

Thus, clearly, although new viruses do not emerge "every other day", there have been many points of virus origin. It is not inconceivable that, before we are done with the global virome, the number of distinct, independently evolving groups of viruses that, at least in principle, have to be classified as realms will far exceed the current six. Virus taxonomists, then, will face the burning question: classify viruses into realms in a strict adherence to the monophyly criterion or lump together some groups that share core VHG but evolved independently, for the sake of tradition and convenience? In any case, the question on the diversity and evolution of the virosphere remain burning, but thanks to the remarkable advances of metaviromics and the concerted efforts in comparative genomics, we are starting to glean interesting answers.

References

Curtis, T.P., Sloan, W.T., and Scannell, J.W. (2002) Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci U S A* **99**: 10494-10499.

Dolja, V.V., and Koonin, E.V. (2018) Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Res* **244**: 36-52.

Dutilh, B.E. (2014) Metagenomic ventures into outer sequence space. *Bacteriophage* **4**: e979664.

Edgar, R.C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D. et al. (2022) Petabase-scale sequence alignment catalyses viral discovery *Nature* **602**: 142-147.

Gaïa, M., Meng, L., Pelletier, E., Forterre, P., Vanni, C., Fernandez-Guerra, A. et al. (2022) Plankton-infecting relatives of herpesviruses clarify the evolutionary trajectory of giant viruses. *bioRxiv* **https://www.biorxiv.org/content/10.1101/2021.12.27.474232v2**.

Gong, Z., and Han, G.Z. (2018) Insect Retroelements Provide Novel Insights into the Origin of Hepatitis B Viruses. *Mol Biol Evol* **35**: 2254-2259.

Guerin, E., Shkoporov, A., Stockdale, S.R., Clooney, A.G., Ryan, F.J., Sutton, T.D.S. et al. (2018) Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* **24**: 653-664 e656.

Hatfull, G.F. (2022) Mycobacteriophages: From Petri dish to patient. *PLoS Pathog* **18**: e1010602.

Hepojoki, J., Hetzel, U., Paraskevopoulou, S., Drosten, C., Harrach, B., Zerbini, F.M. et al. (2021) Create one new realm (Ribozyviria) including one new family (Kolmioviridae) including genus Deltavirus and seven new genera for a total of 15 species. *International Committee on Taxonomy of Viruses Taxonomic Proposal*.

International Committee on Taxonomy of Viruses Executive Committee (2020) The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat Microbiol* **5**: 668-674.

Kauffman, K.M., Hussain, F.A., Yang, J., Arevalo, P., Brown, J.M., Chang, W.K. et al. (2018) A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature* **554**: 118-122.

Kazlauskas, D., Varsani, A., Koonin, E.V., and Krupovic, M. (2019) Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nat Commun* **10**: 3425.

Koonin, E.V. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* **1**: 127-136.

Koonin, E.V., Wolf, Y.I., and Katsnelson, M.I. (2017) Inevitability of the emergence and persistence of genetic parasites caused by evolutionary instability of parasite-free states. *Biol Direct* **12**: 31.

Koonin, E.V., Dolja, V.V., and Krupovic, M. (2022) The logic of virus evolution. *Cell Host Microbe* **30**: 917-929.

Koonin, E.V., Dolja, V.V., Krupovic, M., Varsani, A., Wolf, Y.I., Yutin, N. et al. (2020) Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol Mol Biol Rev* **84**: e00061-00019.

Krupovic, M., and Koonin, E.V. (2017) Multiple origins of viral capsid proteins from cellular ancestors. *Proc Natl Acad Sci U S A* **114**: E2401-E2410.

Krupovic, M., Dolja, V.V., and Koonin, E.V. (2019) Origin of viruses: primordial replicators recruiting capsids from hosts. *Nat Rev Microbiol* **17**: 449-458.

Krupovic, M., Dolja, V.V., and Koonin, E.V. (2020) The LUCA and its complex virome. *Nat Rev Microbiol* **18**: 661-670.

Krupovic, M., Makarova, K.S., and Koonin, E.V. (2022) Cellular homologs of the double jelly-roll major capsid proteins clarify the origins of an ancient virus kingdom. *Proc Natl Acad Sci U S A* **119**: e2120620119.

Krupovic, M., Quemin, E.R., Bamford, D.H., Forterre, P., and Prangishvili, D. (2014) Unification of the globally distributed spindle-shaped viruses of the Archaea. *J Virol* **88**: 2354-2358.

Krupovic, M., Kuhn, J.H., Wang, F., Baquero, D.P., Dolja, V.V., Egelman, E.H. et al. (2021) *Adnaviria*: a new realm for archaeal filamentous viruses with linear A-form double-stranded DNA genomes. *J Virol* **95**: e0067321.

Krupovic, M., Blomberg, J., Coffin, J.M., Dasgupta, I., Fan, H., Geering, A.D. et al. (2018) *Ortervirales*: New virus order unifying five families of reverse-transcribing viruses. *J Virol* **92**: e00515-00518.

Lee, B.D., Neri, U., Roux, S., Wolf, Y.I., Camargo, A.P., Krupovic, M. et al. (2022) A vast world of viroid-like circular RNAs revealed by mining metatranscriptomes. *bioRxiv* **https://www.biorxiv.org/content/10.1101/2022.07.19.500677v1**.

Linnakoski, R., Sutela, S., Coetzee, M.P.A., Duong, T.A., Pavlov, I.N., Litovka, Y.A. et al. (2021) Armillaria root rot fungi host single-stranded RNA viruses. *Sci Rep* **11**: 7336.

Locey, K.J., and Lennon, J.T. (2016) Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A* **113**: 5970-5975.

Maffei, E., Shaidullina, A., Burkolter, M., Heyer, Y., Estermann, F., Druelle, V. et al. (2021) Systematic exploration of Escherichia coli phage-host interactions with the BASEL phage collection. *PLoS Biol* **19**: e3001424.

Mushegian, A.R. (2020) Are There 10(31) Virus Particles on Earth, or More, or Fewer? *J Bacteriol* **202**: e00052-00020.

Neri, U., Wolf, Y.I., Roux, S., Camargo, A.P., Lee, B., Kazlauskas, D. et al. (2022) Expansion of the global RNA virome reveals new clades of bacteriophages. *Cell* **In press**. https://doi.org/10.1016/j.cell.2022.08.023

Petersen, J.M., Bezier, A., Drezen, J.M., and van Oers, M.M. (2022) The naked truth: An updated review on nudiviruses and their relationship to bracoviruses and baculoviruses. *J Invertebr Pathol* **189**: 107718.

Shmakov, S.A., Wolf, Y.I., Savitskaya, E., Severinov, K.V., and Koonin, E.V. (2020) Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. *Commun Biol* **3**: 321.

Taylor, L.J., Keeler, E.L., Bushman, F.D., and Collman, R.G. (2022) The enigmatic roles of Anelloviridae and Redondoviridae in humans. *Curr Opin Virol* **55**: 101248.

Walker, P.J., Siddell, S.G., Lefkowitz, E.J., Mushegian, A.R., Adriaenssens, E.M., Alfenas-Zerbini, P. et al. (2022) Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Arch Virol* **doi: 10.1007/s00705-022-05516-5**.

Wang, F., Cvirkaite-Krupovic, V., Vos, M., Beltran, L.C., Kreutzberger, M.A.B., Winter, J.M. et al. (2022) Spindle-shaped archaeal viruses evolved from rod-shaped ancestors to package a larger genome. *Cell* **185**: 1297-1307 e1211.

Wolf, Y.I., Makarova, K.S., Lobkovsky, A.E., and Koonin, E.V. (2016) Two fundamentally different classes of microbial genes. *Nat Microbiol* **2**: 16208.

Wolf, Y.I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D. et al. (2020) Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat Microbiol* **5**: 1262-1270.

Yutin, N., Backstrom, D., Ettema, T.J.G., Krupovic, M., and Koonin, E.V. (2018a) Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis. *Virol J* **15**: 67.

Yutin, N., Makarova, K.S., Gussow, A.B., Krupovic, M., Segall, A., Edwards, R.A., and Koonin, E.V. (2018b) Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol* **3**: 38-46.

Yutin, N., Benler, S., Shmakov, S.A., Wolf, Y.I., Tolstoy, I., Rayko, M. et al. (2021) Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features *Nat Commun* **12**: 1044.

Zayed, A.A., Wainaina, J.M., Dominguez-Huerta, G., Pelletier, E., Guo, J., Mohssen, M. et al. (2022) Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**: 156-162.

Author contributions

M.K., V.V.D and E.V.K. researched data for article, substantially contributed to the discussion of the content, and wrote the manuscript.

Competing interests

The authors report no competing interests.

Acknowledgements

E.V.K. is supported by the funds of the Intramural Research Program of the National Institutes of Health of the USA. V.V.D. was partially supported by NIH/NLM/NCBI Visiting Scientist Fellowship.

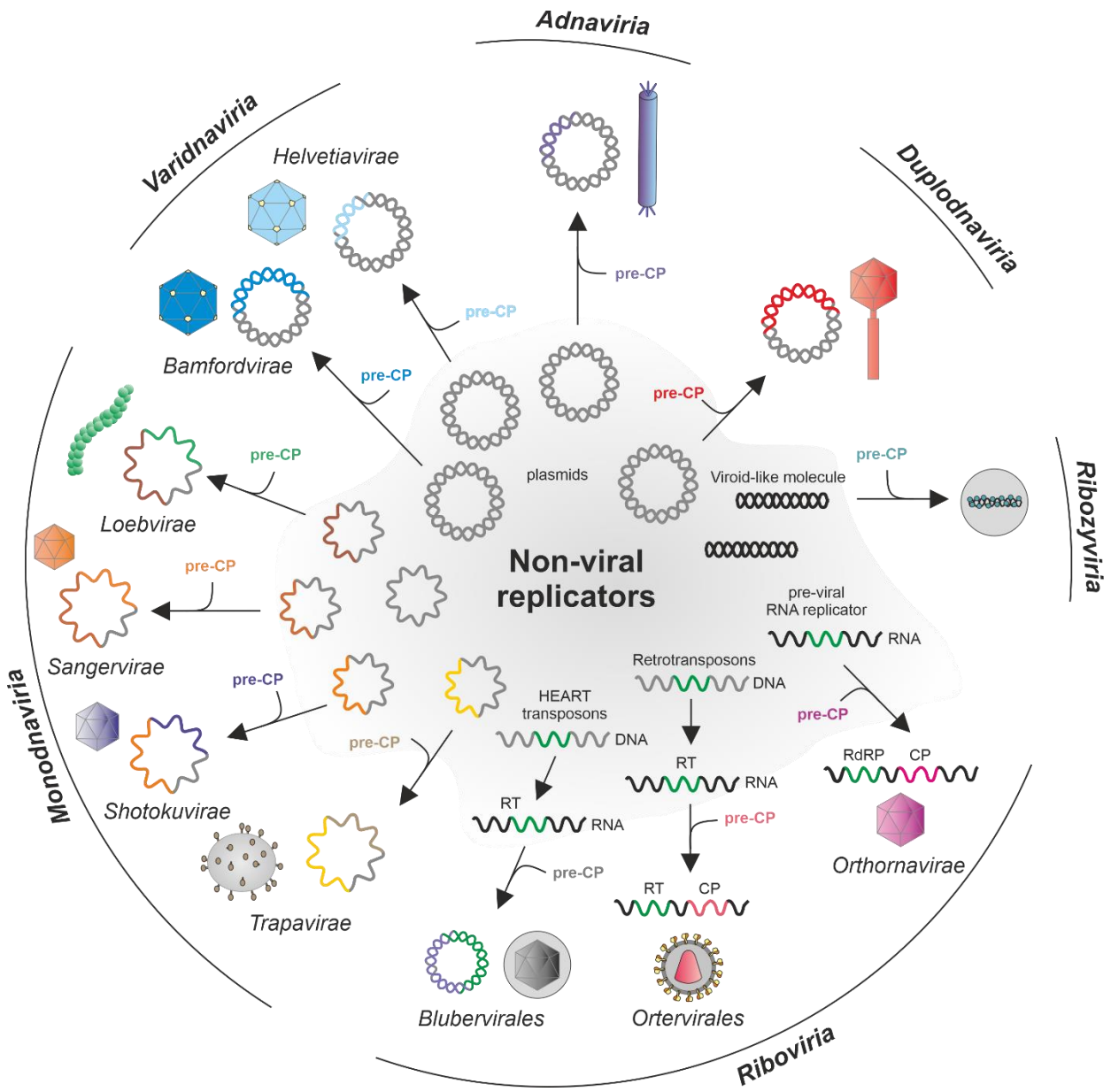

Figure 1. Independent emergence of major groups of viruses: the realms and their possible split. Virogenesis involves acquisition by non-viral replicators, such as plasmids and transposons, of cellular genes which are exapted to form viral capsids. Potential independent routes of virogenesis from non-viral replicators are shown for three of the six realms, with the highest monophyletic taxa indicated. Non-viral DNA and RNA replicators are depicted in grey and black, respectively. Different colors show independent ancestors of capsid proteins. Abbreviations: CP, capsid protein; RdRP, RNA-dependent RNA polymerase; RT, reverse transcriptase.