Investigating the Concept and Origin of Viruses

Arshan Nasir, Ethan Romero-Severson, Jean-Michel Claverie

Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Trends in Microbiology

CellPress
REVIEWS

Opinion

Investigating the Concept and Origin of Viruses

Arshan Nasir,[1,*] Ethan Romero-Severson,[1] and Jean-Michel Claverie[2]

The ongoing COVID-19 pandemic has piqued public interest in the properties, evolution, and emergence of viruses. Here, we discuss how these basic questions have surprisingly remained disputed despite being increasingly within the reach of scientific analysis. We review recent data-driven efforts that shed light into the origin and evolution of viruses and explain factors that resist the widespread acceptance of new views and insights. We propose a new definition of viruses that is not restricted to the presence or absence of any genetic or physical feature, detail a scenario for how viruses likely originated from ancient cells, and explain technical and conceptual biases that limit our understanding of virus evolution. We note that the philosophical aspects of virus evolution also impact the way we might prepare for future outbreaks.

Highlights

The distinctions between virions and viruses and modern and ancient cells are crucial to understand virus origins and evolution.

Viruses can be better defined by their generic features of genome propagation and dissemination rather than physical or biological properties of their virions or hosts.

Virus genomes are characterized by the abundance of virus-specific genes that lack detectable cellular homologs. Despite their abundance, virus-specific genes are rarely discussed in the models of virus origin and evolution.

The alignment-based methods are ill-suited for the origins of life research, especially when the objective is to place fast-evolving organisms or viruses in the tree of life.

Protein structures may provide a better alternative to resolve the very deep branches in the tree of life.

The Need to Redefine Viruses

The COVID-19 pandemic exemplifies the constant threat and pressure exerted by viruses on human health and the global economy. The pandemic has triggered an aggressive international response to contain virus spread, cure the disease, and prevent future infections. In parallel, it has rekindled public curiosity in virus definitions, origins, evolution, and their various modes of emergence. For example, Google search for '*what is a virus*' reached peak popularity in March 2020 coinciding with the global rise in COVID-19 cases. Surprisingly, such fundamental questions have remained unsettled even among evolutionary virologists [1–5] and cause confusion in the media portrayal and public perception. For instance, despite overwhelming scientific evidence supporting a natural zoonotic transmission of SARS-CoV-2 from animals to humans [6], many still suspect that the virus was purposefully engineered in laboratories. Similarly, viruses are generalized as noxious pathogens in common discussions and this focus greatly underestimates the many beneficial roles they play in the biosphere [7,8] and as mutualistic symbionts of many hosts (reviewed in [9,10]). In this article, we revisit fundamental questions about the nature, origins, and evolution of viruses during a time when public interest in virus biology is at its peak. We emphasize the need to rethink viruses in the light of new discoveries [2] and call for broader acceptance of new views that are resisted by (sometimes) century-old concepts established in early virology research (reviewed in [11]).

What Is a Virus?

Defining viruses is surprisingly controversial. This is largely because of the seemingly split nature of the virus reproduction cycle into two distinct stages: (i) an intracellular stage during which the virus reprograms the infected cell to produce viral particles or **virions** (see Glossary), and (ii) an extracellular stage during which virions escape the infected cells and persist in the external environment (similar to plant seeds [12]).[1] Both stages, when considered separately, provide dramatically contrasting views about the nature and roles of viruses. For example, virions are metabolically

[1]Theoretical Biology and Biophysics (T-6), Los Alamos National Laboratory, Los Alamos, NM, USA
[2]Aix Marseille University, CNRS, IGS, Structural and Genomic Information Laboratory (UMR7256), Mediterranean Institute of Microbiology (FR3479), Marseille, France

*Correspondence:
anasir@lanl.gov (A. Nasir).

[1] In addition, a third stage may exist if the virus genome either integrates into the host DNA or becomes part of the host cytoplasm. Such examples may not lead to virion production or diseases. Because classical signs of virus infection (e.g., virion production, cell rupture) may not be obvious, it is possible that we have massively underestimated nonharmful virus–cell interactions involving virus genome endogenization and domestication by cells [83].

inert infectious particles that do not meet any of the criteria we may use to define 'life' or living organisms [2]. However, since they can be purified, counted, and visualized under the microscope, their physical and biochemical properties (e.g., size, shape, metabolic capabilities, capsid) along with host/tissue specificity have become popular in the description, illustration, and naming of viruses (e.g., human immunodeficiency virus). These, in turn, have shaped our perceptions about viruses as nonliving inanimate biological objects that are, paradoxically, infectious.

Treating virions as viruses is a conceptual mistake [2,12–16] that overlooks the dramatic changes viruses introduce inside infected cells. A virus-infected cell can effectively be transformed into a 'hot spot' for virion production [17] and can practically lose its identity (i.e., it now produces virions rather than two daughter cells) [18]. In some viral infections, large cell-like '**virion factories**' are clearly visible [19]. This remarkable transformation is due to the virus-mediated manipulation and alteration of host metabolism and defenses [7]. The intracellular stage therefore involves substantial viral activity and is often the target of antiviral drugs to combat virus infection (e.g., antivirals that target HIV polymerase). Despite its immense role in establishing virus infection and existence inside the infected cells, it has unfortunately been referred to as the 'eclipse' or 'vegetative' phase [20,21] to indicate lack of hallmark signs of virus infection (e.g., virion production, plaques, and cell rupture) and ignored in the definitions and descriptions of viruses. As suggested by Jean-Michel Claverie, virion factory better represents the 'virus self' and virions are simply means to disseminate genetic information much like human gametes and plant seeds [12]. In other words, we should depart from the established usage of the word 'virus' as being synonymous to 'virion'. The term 'virus' should refer to the process encompassing all phases of the virus infection cycle [3]. In this context, questioning the origin of 'viruses' takes a completely different and much broader meaning than simply questioning the origin of the virus particles [2,11,13,16].

Avoid the Presence/Absence Criteria to Define Viruses

The virion- and host-centric virus definitions can cause ambiguities in distinguishing different viral lineages and even viruses from cellular organisms. For example, Forterre recently proposed to redefine viruses as 'capsid-encoding organisms' [22] and later as 'virion-encoding organisms' [2]. Both definitions recognize viruses as 'organisms' that produce capsids/virions and rightly put emphasis back on the intracellular stage of virus infection cycle. However, these views suffer from our 'human' habit of classifying biological entities based on the presence/absence or contrast of physical and genetic features. As we discuss later, such definitions rarely withstand the test of time and are vulnerable to change with new discoveries. For example, viruses were long considered tiny and submicroscopic biological entities (properties that describe virions not viruses!) before the discovery of 'giant viruses' with genomes and virions bigger than the genomes and sizes of many parasitic cells [23–25]. In fact, holding onto the century-old size/shape virion-centric definitions delayed the discovery of giant viruses by more than a decade.[2] Similarly, some scientists consider viruses 'non-living' because they do not encode metabolism-related genes [1]. However, this feature is neither unique nor common to all viruses. Many **endosymbiotic** cellular organisms are also characterized by extremely reduced metabolic and translational machineries [26–28], and recent metagenomic surveys have verified the existence of several, and likely very ancient, metabolic genes in the genomes of giant viruses [7]. These genes likely help reconfigure the metabolism of infected cells during virus infection [7].

Glossary

Endosymbiosis: the intimate existence of organisms inside the cells or body of other organisms. Notable examples include endosymbiosis of the ancestors of mitochondria and chloroplasts by proto-eukaryotes or the ancestors of eukaryotes.
Last universal common ancestor (LUCA): the common ancestor of modern cells, Archaea, Bacteria, and Eukarya. LUCA was not the first cell. It was the last population of cells that diversified into modern cells.
Orthologous: refers to genes that diverged from the common ancestor as a result of speciation.
Tree of life: a diagram that describes the evolutionary history among modern species using the metaphors of branching patterns, roots, and leaves to represent evolutionary relationships, ancestors, and modern species, respectively. The topology of the tree of life and the place of viruses in the tree are hotly debated topics.
Virion: virus particle that can be purified and visualized. The core of a virion comprises the virus nucleic acid (DNA or RNA) enclosed inside a protein shell called a capsid.
Virion factories: intracellular compartments, formed inside virus-infected cells, that increase virus replication.

[2] The first giant virus, *Acanthamoeba polyphaga mimivirusi,* was initially mistaken for a Gram-positive bacterium. It was first discovered in 1992 during a pneumonia outbreak in Bradford, UK and the large size of its virion misled scientists to believe that it must be a bacterium (called 'Bradfordcoccus'). Its virus nature was finally revealed in 2003 [84] and the virus was aptly named 'mimivirus' for 'bacteria-mimicking virus'. This is a famous example where adhering to century-old virion and size-based virus definitions delayed a significant discovery.

Using virion or capsid to distinguish viral lineages and viruses from cells can generate similar confusions. For example, it can complicate classifications for virus-like genetic elements and viruses that either lack virions (e.g., plasmids, viroids [29]) or encode only part of the virion (e.g., polydnaviruses). For example, the genome of polydnaviruses is dispersed within the genome of parasitoid wasps. The polydnavirus-associated wasps encode the virion packaging system and utilize virions as gene delivery vectors to infect caterpillars [30]. Since the virion is encoded by the wasp genome, polydnavirus-associated wasps may better resemble virion-encoding organisms under the virion-centric definition [31]. Similarly, virus-infected cells can excrete vesicles containing the virus nucleic acid, [32] and healthy cells routinely utilize extracellular vesicles for genetic communication [33]. These examples generalize the concept and morphology of a 'virion'. Similarly, capsid-like compartments have been detected in cellular organisms where they perform functions such as storage of enzymes [34], and many viral capsid proteins either evolved directly from cellular proteins [35] or have distant homologs in cellular genomes [36]. These examples blur the separation of viruses from cells (and other parasitic genetic elements) based on the presence/absence of physical or genetic descriptors.

In sum, we discourage the use of any virus definition based on the presence or absence of any subset of genes or physical features (e.g., size, morphology, capsid proteins) because such definitions are often ambiguous, not broadly applicable, and more importantly prone to change with new discoveries. We assert that viruses can be better defined by their generic properties of genome dissemination and propagation [11]. Viruses replicate using the macromolecular machinery of other biological entities. This prong establishes absolute parasitism, which is a hallmark of viruses and virus-like genetic elements. Another feature of viruses is the ability to encapsulate and disseminate genomes in metabolically inert structures. This prong can also be generalized, and such structures could be any type of infectious particle without constraints of size, shape, or biochemical composition (e.g., vesicles) [37]. This definition encompasses both the encapsulated and non-encapsulated genomes (e.g., plasmids) and emphasizes the generic feature of how viruses propagate in cells rather than being dependent on the presence/absence of specific biomarkers [11].

Origins of Viruses: Which Hypothesis Is Biologically Plausible?

Under our generic definition, virus origin must mean the origin of parasitism and the subsequent ability of those parasitic entities to propagate via the production of metabolically inert structures. Since all modern-day viruses strictly parasitize cells (with the exception of virophages that parasitize the viral factory of other viruses) [38,39], we can assume that virus-mediated parasitism and propagation originated only after cells appeared in evolution as cells would provide both the resources to parasitize upon and the means for genome dissemination (e.g., capsids/vesicles). We therefore rule out virus existence in a 'pre-cellular' world as it would be incompatible with the proposed virus definition (Figure 1 for comparative scenarios).

The next logical questions are the timings and mechanisms of when and how the first viruses appeared. The former question is relatively straightforward. In our view, viruses originated from 'ancient' cells that existed before the **last universal common ancestor** (**LUCA**) diversified into modern cells (i.e., the three superkingdoms, Archaea, Bacteria, and Eukarya) [40].[3] There are multiple lines of evidence supporting this timing. For example, the genomes of archaeoviruses, bacterioviruses, and eukaryoviruses, are characterized by the abundance of

[3] In the 1970s, Carl Woese pioneered the method of using molecular sequences to study evolution. His work led to the recognition of Archaea [85], then called the 'third domain' of life [86]. Archaea have recently taken center stage in evolutionary debates regarding the origin of eukaryotes. There is great controversy on whether Archaea were the first group of diversified organisms on Earth [87], are a sister group to eukaryotes [88,89], or are our ancestors [90,91].

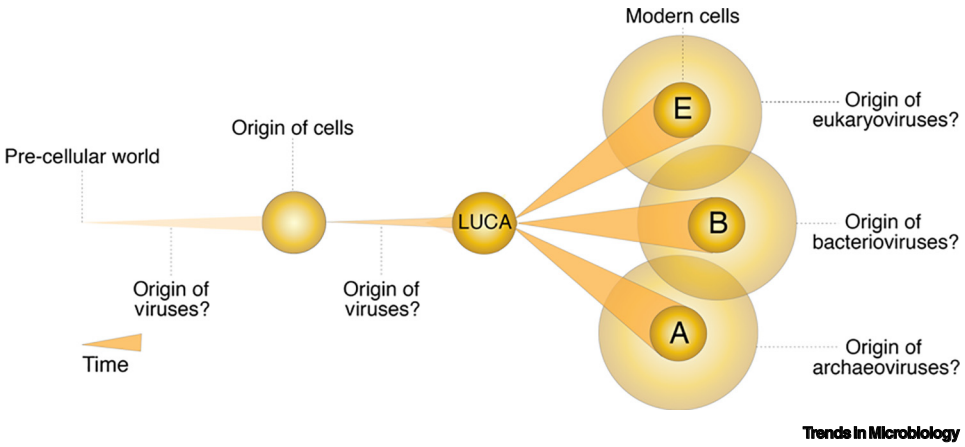

Figure 1. Different Scenarios for the Origin of Viruses. Viruses originated either prior to or from cells. A pre-cellular scenario is incompatible with the proposed generic definition of virus propagation inside cells. In turn, the origin of archaeoviruses from Archaea, bacterioviruses from Bacteria, and eukaryoviruses from Eukarya also seems less likely as these viruses share several conserved protein folds involved in virion synthesis and other functions, indicating that they may have evolved prior to the diversification of LUCA into modern cells. These considerations support an intermediate timing for the origin of viruses, that is, from ancient cells that existed prior to LUCA. Modified from [82]. Abbreviation: LUCA, last universal common ancestor.

virus-specific genes that lack detectable homologs in cellular genomes [41]. While these genes can be strain-specific with a recent *de novo* origin [42,43], their abundance and existence in diverse virus groups suggests their accumulation likely started very early in evolution. Similarly, viral lineages that infect distantly related hosts from all three superkingdoms share several conserved three-dimensional (3D) protein structural folds that also indicate that these lineages likely existed prior to LUCA diversification [44]. New viruses would then evolve from existing viruses via natural processes such as recombination and in response to new and emerging hosts.

The mechanisms of how ancient cells evolved into viruses are relatively less clear. Krupovic *et al*. recently proposed a hybrid model to answer this question [45]. According to their model, viral nucleic acids evolved in the pre-cellular world and virus propagation mechanisms evolved via the modification of cellular proteins to function as virus capsids once cells appeared in evolution. In their view, giant viruses such as pandoraviruses and Mimiviruses gradually became bigger due to frequent gene capture from host cells [46]. Their model thus explains the massive genetic diversity seen in virus replicons and proposes mechanisms for the origin of virus capsids and giant viruses. We disagree with the model on two major points.

First, it is unnecessary to invoke a pre-cellular world to explain the observed replicon and genetic diversity among modern-day viruses. This diversity can simply unfold in the pool of ancient cells that existed prior to LUCA. Second, the proposed incremental growth of viral genomes, especially large DNA viruses, via gene gain from hosts is incompatible with our knowledge of how endosymbiotic/parasitic cells evolve. Cells committed to obligate parasitism are characterized by extreme genomic and physical reduction as they increase dependency on their hosts [26,47]. It makes sense to think that viruses, which are the ultimate examples of parasitism, would also evolve similarly. This ‘reduction’ scenario is more parsimonious when one considers the gigantic genome sizes of pandoraviruses (~2500 genes). There is no clear incentive as to why a small-sized virus genome (~5 genes in papillomaviruses) would adopt a

pathway towards gigantism, when it is already a well-established parasite. Moreover, regular and recent gene uptake from cells is expected to leave detectable similarity traces in the viral genomes. However, >90% of *Pandoravirus* genes show no similarity to cellular genes [24] (a feature conserved in many viruses [41], see also [48]) and viruses encode several protein fold structures that have never been detected in cells [41]. It is strange to think that regularly and recently captured viral genes are no longer recognizable whereas presumably ancient 'core' genes used to build virus phylogenies are readily recognizable.

Surprisingly, the existence and abundance of virus-specific genes (i.e., genes or protein folds detected only in viral genomes) are rarely discussed in the models of virus origin and evolution. Instead, homology of a subset of 'core' virus genes to their cellular counterparts is used to generalize the notion that viruses evolve by acquiring cellular genes [49]. This practice yields an incomplete view of the composition and evolution of virus genomes and ignores the significant *de novo* gene creation abilities in viruses, especially in pandoraviruses [42]. Moreover, 'core' genes describe the evolutionary histories of individual genes and not the whole organisms. They can be patchily distributed, similar to virus-specific genes, and the number of available core genes for phylogenetic studies is strongly dependent on the number of sampled genomes and their taxonomic range [50]. For example, no single gene or protein fold is conserved across all RNA and DNA viruses and very few are conserved across diverse virus families (e.g., DNA polymerase is conserved in many DNA viruses but not in papillomaviruses, and RNA polymerase is conserved in many RNA viruses but not in satellite viruses). The patchy distribution of both the core and virus-specific genes is expected from random reductive evolution where lost and conserved genes were randomly selected from ancient cells. We therefore propose that viruses, especially DNA viruses, evolved from one or multiple ancient cells via reduction [11,41,51,52]. This scenario better aligns with the evolutionary biology of endosymbiotic and parasitic cellular organisms and is more plausible considering the unique composition of virus genomes.

The proposed reduction model is based on the generic definition of viruses and does not suggest that ancient cells reduced into virions. That would be mistaking viruses for their virions, a classical mistake that we have just criticized. Instead, we simply propose that ancient cells were the first to discover the benefits of parasitization and propagation via released particles (e.g., vesicles) [37]. Gradually, the ancient cells devolved as the released particles became fully capable of repeating the cycle of invasion and escape in coinhabiting cellular lineages. While the concept of a 'virus-like cell' or a 'cellular ancestor of virus' may be difficult to imagine, we already know several examples of viral genome endogenization into host DNA [53] and viral factories that alter the nature of the infected cells [12,18]. These modern-day events transiently restore the ancient 'cellular self' that may have been a more permanent feature in the past. In fact, cytoplasmic virion factories behave like a pseudo-nucleus where virus genome replication and translation are separated from host cytoplasm [54,55]. Some authors suggest that the eukaryotic nucleus likely evolved directly from an ancient viral factory [56,57].

Pathways to DNA Cells and Viruses

The ancient pre-LUCA cells likely harbored segmented RNA genomes [41,58,59]. It is therefore logical to think that RNA viruses evolved first from RNA cells, and that later, DNA viruses evolved directly from RNA viruses, and in parallel, DNA cells evolved from RNA cells (Figure 2A). This scenario implies that RNA viruses are the ancestors of DNA viruses and was supported in a recent phylogenomic analysis [41]. A second alternative could be that RNA viruses evolved directly from RNA cells, and DNA viruses evolved directly from DNA cells. Thus, both groups

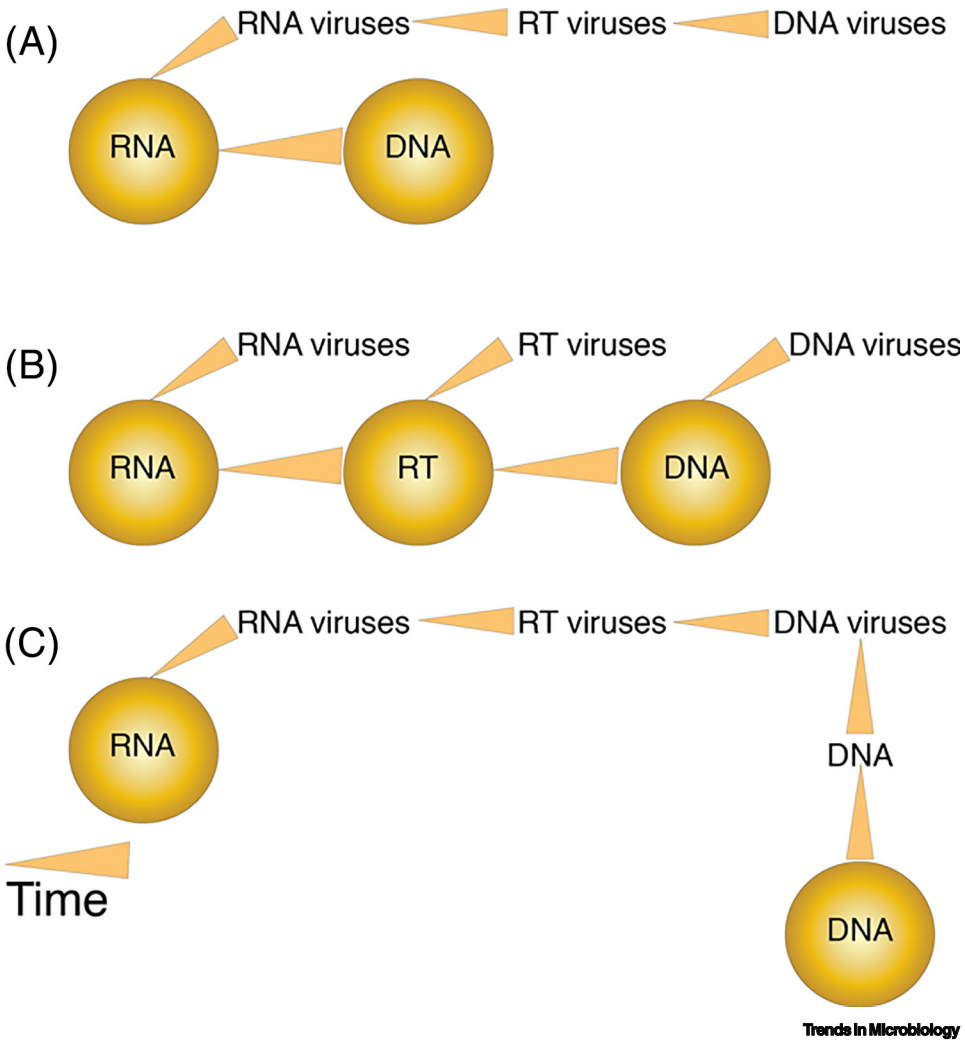

Figure 2. Different Scenarios for the Evolution of Different Virus Replicon Groups. (A) RNA viruses evolved from RNA cells and later evolved into retrotranscribing (RT) and DNA viruses. In parallel, RNA cells evolved into DNA cells. (B) The evolution of RNA, RT, and DNA viruses followed the emergence of RNA, RT, and DNA cells, respectively. (C) RNA viruses evolved from RNA cells and later evolved into RT and DNA viruses. RNA cells evolved into DNA cells once DNA was invented by viruses [60].

evolved independently from different cellular ancestors and possibly via different mechanisms (Figure 2B). Finally, a third alternative could be the evolution of RNA viruses from RNA cells. RNA viruses later invented DNA to escape the defenses of RNA cells. The invention of DNA was later picked up by RNA cells to become DNA cells [60,61] (Figure 2C). Testing these alternatives is challenging since molecular data are limited in their ability to resolve deep evolutionary events.

Existing Methods Are Ill-Suited to Study Virus Origins

In standard phylogenetic analyses, gene and protein sequences are aligned to elucidate the phylogenetic history of a group of organisms. This alignment is used to infer a phylogenetic tree using various methods [62]. While the alignment-dependent methods work very well in resolving the evolutionary relationships among closely-related (micro)organisms and have significant other

applications, they are probably not suited for the origins or '**tree of life**' research [63]. This is especially true when the objective is to place fast-evolving organisms and viruses in the tree of life [64]. First, the subset of virus genes for which reliable homologs can be found is extremely small [48]. This fact greatly limits the choice and the number of available **orthologous** genes to be used in phylogeny reconstruction. This sometimes leads authors to resort to subjective, nonstatistically supported approaches to suggest distant homology relationships [65]. Another problem is the recovery of a reliable alignment of homologous genes from a diverse set of genomes. In general, statistically detectable sequence similarity fades over evolutionary time, sometimes leading to complete loss of evolutionary signal due to mutation saturation [66,67]. In addition, protein domain (i.e., structural and functional units within proteins) gains, losses, rearrangements, duplications, and transfers are frequent events in the evolution of genes and genomes [68,69]. These events can happen at different rates in different lineages and thus add many unaligned or poorly aligned regions in sequence alignments [70]. Recovery of a reliable alignment therefore often requires significant manual curation (e.g., removal of a large proportion of poorly aligned sites), which impacts reproducibility by introducing subjectivity [71], and may even be impossible for diverse RNA virus groups [64]. Indeed, the genomes of RNA and retrotranscribing viruses exhibit very high mutation rates [72]. HIV lineages evolving within the same host can differ by 5–10% whereas intrahuman genetic variation could be <0.1% even after ~2.5 million years [73]. These facts greatly limit our ability to reconstruct past evolutionary events using molecular sequence information alone and prompt us to evaluate the potential of alternative, more conserved, molecular characters such as protein structures [74] (Box 1).

Concluding Remarks

Viruses can be better defined based on the generic features of genome dissemination rather than specific virion-associated or physical (size) properties. The defining feature of virus genomes is the existence and abundance of virus-specific genes and protein folds that have no homologs in the cellular world. These genes are rarely discussed in the models of virus origin and evolution, and instead most evolutionary studies rely on a very small subset of viral genes for which we can find reliable cellular homologs. Often such homologies are interpreted as gene uptake from cells by viruses, which is an oversimplified notion for the evolution of virus genomes. Moreover, the fast mutation rates of RNA and retrotranscribing viruses almost make it impossible to recover a reliable alignment for deep virus evolutionary studies. In this regard, focusing on alternative molecular characters that are better conserved in evolution (e.g., protein structures) can possibly provide better solutions. Viruses are likely very old and originated from ancient RNA cells that predated LUCA. They continue to play important roles in the evolution of cells and exert enormous pressure on human health and the global economy. Updating our views on the origins and definitions of viruses (e.g., the distinction between virion and virus) may also help to clarify our thinking about the risk of emergence and spread of new viral diseases (see Outstanding Questions).

Box 1. Protein Structures Can Improve Deep Evolutionary Inferences

Advancements in structural biology allow us to explore and utilize new sets of molecular characters to study deep evolution. Protein function is usually determined directly by the 3D shape of the protein. This fact constrains the preservation of protein structure over longer periods of time, as tampering with the structure could lead to loss of function and could be quite damaging [63,75,76]. As of 26 May 2020, there are ~160 000 protein structural entries in the RCSB Protein Data Bank [77]. These structures correspond to ~1400 protein folds [78], indicating that protein structure space is relatively well sampled and possibly finite [79]. Illergård *et al*. showed that protein structures evolve at least three to ten times slower than protein sequences [74]. Protein folds are thus (apparently) advantageous as they are remarkably conserved across all species, (and even) viruses, as revealed by their use in recent studies [80,81]. It is possible that a large number of protein folds we see today are very ancient, and even predated LUCA [37,41]. Their use could thus be extremely powerful if subjected to careful phylogenomic and comparative analyses. However, much like the resistance to accepting emerging viewpoints on viruses, protein structures have been rarely utilized in evolutionary studies. This remains another major roadblock in our understanding of virus origins and evolution.

Outstanding Questions

How to visualize and illustrate viruses (virus factories) if virions cannot and should not be used to describe viruses?

How are RNA and DNA viruses evolutionarily related?

Do well defined boundaries exist between cellular and viral lineages?

How do we explain the origin and abundance of virus-specific genes in viral genomes?

Can protein structure-based methods improve the evolutionary studies of viruses?

Acknowledgments

The authors would like to thank Chantal Abergel for continuous discussions and advice. A.N. is supported by the US Department of Energy Laboratory Directed Research and Development (LDRD) program at the Los Alamos National Laboratory (20180751PRD3). E.R.-S. is supported by the LDRD program under project number 20180612ECR and by NIH/NIAID grant R01AI087520 to Thomas Leitner.

References

1. Moreira, D. and Lopez-Garcia, P. (2009) Ten reasons to exclude viruses from the tree of life. *Nat. Rev.* 7, 306–311
2. Forterre, P. (2016) To be or not to be alive: How recent discoveries challenge the traditional definitions of viruses and life. *Stud. Hist. Phil. Biol. Biomed. Sci.* 59, 100–108
3. Dupré, J. and Guttinger, S. (2016) Viruses as living processes. *Stud. Hist. Phil. Biol. Biomed. Sci.* 59, 109–116
4. Claverie, J.M. and Abergel, C. (2013) Open questions about giant viruses. *Adv. Virus Res.* 85, 25–56
5. Van Regenmortel, M.H. (2016) The metaphor that viruses are living is alive and well, but it is no more than a metaphor. *Stud. Hist. Phil. Biol. Biomed. Sci.* 59, 117–124
6. Andersen, K.G. *et al.* (2020) The proximal origin of SARS-CoV-2. *Nat. Med.* 26, 450–452
7. Moniruzzaman, M. *et al.* (2020) Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat. Commun.* 11, 1–11
8. Schulz, F. *et al.* (2020) Giant virus diversity and host interactions through global metagenomics. *Nature* 578, 432–436
9. Roossinck, M.J. (2011) The good viruses: viral mutualistic symbioses. *Nat. Rev. Microbiol.* 9, 99–108
10. Roossinck, M.J. (2015) Move over, bacteria! viruses make their mark as mutualistic microbial symbionts. *J. Virol.* 89, 6532–6535
11. Claverie, J.M. and Abergel, C. (2016) Giant viruses: The difficult breaking of multiple epistemological barriers. *Stud. Hist. Phil. Biol. Biomed. Sci.* 59, 89–99
12. Claverie, J.M. (2006) Viruses take center stage in cellular evolution. *Genome Biol.* 7, 110
13. Bandea, C.I. (1983) A new theory on the origin and the nature of viruses. *J. Theor. Biol.* 105, 591–602
14. Van Regenmortel, M.H. (2009) Logical puzzles and scientific controversies: The nature of species, viruses and living organisms. *Syst. Appl. Microbiol.* 33, 1–6
15. Boycott, A.E. (1928) The transition from live to dead: the nature of filtrable viruses. *Proc. R. Soc. Med.* 22, 55–69
16. Bandea, C.I. (2009) The origin and evolution of viruses as molecular organisms. *Nat. Prec.* Published October 23, 2009. https://doi.org/10.1038/npre.2009.3886.1
17. Lwoff, A. (1967) Principles of classification and nomenclature of viruses. *Nature* 215, 13–14
18. Forterre, P. (2011) Manipulation of cellular syntheses and the nature of viruses: The virocell concept. *C. R. Chim.* 14, 392–399
19. Suzan-Monti, M. *et al.* (2007) Ultrastructural characterization of the giant volcano-like virus factory of *Acanthamoeba polyphaga* Mimivirus. *PLoS One* 2, e328
20. Lwoff, A. (1957) The concept of virus. *J. Gen. Microbiol.* 17, 239–253
21. Jacob, F. and Wollman, E. (1961) Viruses and genes. *Sci. Am.* 204, 93–107
22. Raoult, D. and Forterre, P. (2008) Redefining viruses: lessons from Mimivirus. *Nat. Rev.* 6, 315–319
23. Raoult, D. *et al.* (2004) The 1.2-megabase genome sequence of Mimivirus. *Science* 306, 1344–1350
24. Philippe, N. *et al.* (2013) Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* 341, 281–286
25. Legendre, M. *et al.* (2015) In-depth study of *Mollivirus sibericum*, a new 30,000-y-old giant virus infecting *Acanthamoeba*. *Proc. Natl. Acad. Sci. U. S. A.* 112, E5327–E5335
26. López-Madrigal, S. *et al.* (2011) Complete genome sequence of '*Candidatus* Tremblaya princeps' strain PCVAL, an intriguing translational machine below the living-cell status. *J. Bacteriol.* 193, 5587–5588
27. Luef, B. *et al.* (2015) Diverse uncultivated ultra-small bacterial cells in groundwater. *Nat. Commun.* 6, 6372
28. Omsland, A. *et al.* (2014) Chlamydial metabolism revisited: Interspecies metabolic variability and developmental stage-specific physiologic activities. *FEMS Microbiol. Rev.* 38, 779–801
29. Sato, Y. *et al.* (2020) Hadaka virus 1: a capsidless eleven-segmented positive-sense single-stranded RNA virus from a phytopathogenic fungus, *Fusarium oxysporum*. *mBio* 11, e00450-20
30. Burke, G.R. and Strand, M.R. (2012) Polydnaviruses of parasitic wasps: domestication of viruses to act as gene delivery vectors. *Insects* 3, 91–119
31. Desjardins, C.A. (2012) Unusual viral genomes: Mimivirus and the polydnaviruses. In *Parasitoid Viruses: Symbionts and Pathogens* (Beckage, N.E. and Drezen, J.M., eds), pp. 115–125, Elsevier
32. Gill, S. *et al.* (2019) Extracellular membrane vesicles in the three domains of life and beyond. *FEMS Microbiol. Rev.* 43, 273–303
33. Soler, N. and Forterre, P. (2020) Vesiduction: the fourth way of HGT. *Environ. Microbiol.* 22, 2457–2460
34. Sutter, M. *et al.* (2008) Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nat. Struct. Mol. Biol.* 15, 939–947
35. 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
36. Nasir, A. and Caetano-Anollés, G. (2017) Identification of capsid/coat related protein folds and their utility for virus classification. *Front. Microbiol.* 8, 380
37. Nasir, A. *et al.* (2015) Untangling the origin of viruses and their impact on cellular evolution. *Ann. N. Y. Acad. Sci.* 1341, 61–74
38. La Scola, B. *et al.* (2008) The virophage as a unique parasite of the giant mimivirus. *Nature* 455, 100–104
39. Claverie, J.M. and Abergel, C. (2009) Mimivirus and its virophage. *Annu. Rev. Genet.* 43, 49–66
40. Forterre, P. (2005) The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. *Biochimie* 87, 793–803
41. Nasir, A. and Caetano-Anollés, G. (2015) A phylogenomic data-driven exploration of viral origins and evolution. *Sci. Adv.* 1, e1500527
42. Legendre, M. *et al.* (2019) *Pandoravirus celtis illustrates the microevolution processes at work in the giant Pandoraviridae genomes*. 10 pp. 1–11
43. Legendre, M. *et al.* (2018) Diversity and evolution of the emerging Pandoraviridae family. *Nat. Commun.* 9, 2285
44. Abrescia, N.G.A. *et al.* (2012) Structure unifies the viral universe. *Annu. Rev. Biochem.* 81, 795–822
45. Krupovic, M. *et al.* (2019) Origin of viruses: primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.* 17, 449–458
46. Koonin, E.V. and Yutin, N. (2019) Evolution of the large nucleocytoplasmic DNA viruses of eukaryotes and convergent origins of viral gigantism. *Adv. Virus Res.* 103, 167–202
47. McCutcheon, J.P. and Moran, N.A. (2011) Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* 10, 13–26
48. Boratto, P.V.M. *et al.* (2020) Yaravirus: A novel 80-nm virus infecting *Acanthamoeba castellanii*. *Proc. Natl. Acad. Sci. U. S. A.* 117, 16579–16586
49. Moreira, D. and López-García, P. (2015) Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes? *Philos. Trans. R. Soc. B Biol. Sci.* 370, 20140327
50. Dagan, T. and Martin, W. (2006) The tree of one percent. *Genome Biol.* 7, 118
51. Nasir, A. *et al.* (2012) Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. *BMC Evol. Biol.* 12, 156

52. Claverie, J.M. *et al.* (2006) Mimivirus and the emerging concept of 'giant' virus. *Virus Res.* 117, 133–144
53. Feschotte, C. and Gilbert, C. (2012) Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* 13, 283–296
54. Forterre, P. (2010) Defining life: The virus viewpoint. *Orig. Life Evol. Biosph.* 40, 151–160
55. Chaikeeratisak, V. *et al.* (2017) Assembly of a nucleus-like structure during viral replication in bacteria. *Science* 355, 194–197
56. Forterre, P. and Gaïa, M. (2016) Giant viruses and the origin of modern eukaryotes. *Curr. Opin. Microbiol.* 31, 44–49
57. Bell, P.J.L. (2001) Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus? *J. Mol. Evol.* 53, 251–256
58. Woese, C.R. (1983) The primary lines of descent and the universal ancestor. In *Evolution from Molecules to Men* (Bendall, D.S., ed.), pp. 209–233, Cambridge University Press
59. Poole, A.M. *et al.* (1998) The path from the RNA world. *J. Mol. Evol.* 46, 1–17
60. Forterre, P. (2006) Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: A hypothesis for the origin of cellular domain. *Proc. Natl. Acad. Sci. U. S. A.* 103, 3669–3674
61. Claverie, J.M. and Abergel, C. (2010) Mimivirus: The emerging paradox of quasi-autonomous viruses. *Trends Genet.* 26, 431–437
62. Kapli, P. *et al.* (2020) Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* 21, 428–444
63. Caetano-Anollés, G. and Nasir, A. (2012) Benefits of using molecular structure and abundance in phylogenomic analysis. *Front. Genet.* 3, 172
64. Holmes, E.C. and Duchêne, S. (2019) Can sequence phylogenies safely infer the origin of the global virome? *mBio* 10, e00289-19
65. Krupovic, M. *et al.* (2020) Evolution of a major virion protein of the giant pandoraviruses from an inactivated bacterial glycoside hydrolase. *Virus Evol.*, veaa059
66. Sober, E. and Steel, M. (2002) Testing the hypothesis of common ancestry. *J. Theor. Biol.* 218, 395–408
67. Nasir, A. *et al.* (2016) Arguments reinforcing the three-domain view of diversified cellular life. *Archaea* 2016, 1851865
68. Nasir, A. *et al.* (2014) Global patterns of protein domain gain and loss in superkingdoms. *PLoS Comput. Biol.* 10, e1003452
69. Moore, A.D. and Bornberg-Bauer, E. (2012) The dynamics and evolutionary potential of domain loss and emergence. *Mol. Biol. Evol.* 29, 787–796
70. Caetano-Anollés, G. *et al.* (2018) Rooting phylogenies and the tree of life while minimizing ad hoc and auxiliary assumptions. *Evol. Bioinform.* 14 1176934318805101
71. Philippe, H. *et al.* (2017) Pitfalls in supermatrix phylogenomics. *Eur. J. Taxon.* 283, 1–25
72. Sanjuán, R. *et al.* (2010) Viral mutation rates. *J. Virol.* 84, 9733–9748
73. Leitner, T. (2018) The puzzle of hiv neutral and selective evolution. *Mol. Biol. Evol.* 35, 1355–1358
74. Illergård, K. *et al.* (2009) Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins* 77, 499–508
75. Chothia, C. *et al.* (2003) Evolution of the protein repertoire. *Science* 300, 1701–1703
76. Gerstein, M. and Hegyi, H. (1998) Comparing genomes in terms of protein structure: Surveys of a finite parts list. *FEMS Microbiol. Rev.* 22, 277–304
77. Burley, S.K. *et al.* (2019) RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* 47, D464–D474
78. Andreeva, A. *et al.* (2020) The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.* 48, D376–D382
79. Caetano-Anollés, G. *et al.* (2009) The origin, evolution and structure of the protein world. *Biochem. J.* 417, 621–637
80. Mughal, F. *et al.* (2020) The origin and evolution of viruses inferred from fold family structure. *Arch. Virol.* 1–15
81. Bokhari, R.H. *et al.* (2020) Bacterial origin and reductive evolution of the CPR group. *Genome Biol. Evol.* 12, 103–121
82. Nasir, A. *et al.* (2012) Viral evolution Primordial cellular origins and late adaptation to parasitism. *Mob. Genet. Elements* 2, 247–252
83. Nasir, A. *et al.* (2017) Long-term evolution of viruses: A Janus-faced balance. *BioEssays* 39, e201700026
84. La Scola, B. *et al.* (2003) A giant virus in amoebae. *Science* 299, 2033
85. Woese, C.R. and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5088–5090
86. Woese, C.R. *et al.* (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* 87, 4576–4579
87. Staley, J.T. and Caetano-Anollés, G. (2018) Archaea-First and the co-evolutionary diversification of domains of life. *BioEssays* 40, 1800036
88. Da Cunha, V. *et al.* (2017) Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.* 13, e1006810
89. Da Cunha, V. *et al.* (2018) Asgard archaea do not close the debate about the universal tree of life topology. *PLoS Genet.* 14, e1007215
90. Zaremba-Niedzwiedzka, K. *et al.* (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541, 353–358
91. Williams, T.A. *et al.* (2020) Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.* 4, 138–147