The LUCA and its complex virome

Mart Krupovic, Valerian Dolja, Eugene Koonin

The LUCA and its complex virome

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

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

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

Abstract

The last universal cellular ancestor (LUCA) is the most recent population of organisms from which all cellular life on Earth descends. Reconstruction of the genome and phenotype of the LUCA is a major challenge in evolutionary biology. Given that all life forms are associated with viruses and/or other mobile genetic elements, there is no doubt that the LUCA was a host to viruses. Here, by projecting back in time using the extant distribution of viruses across the two primary domains of life, bacteria and archaea, and tracing the evolutionary histories of some key virus genes, we attempt a reconstruction of the LUCA virome. Even a conservative version of this reconstruction suggests a remarkably complex virome that already included the main groups of extant viruses of bacteria and archaea. We further present evidence of extensive virus evolution antedating the LUCA. The presence of a highly complex virome implies substantial genomic and pan-genomic complexity of the LUCA itself.

[H1] Introduction

Viruses and other mobile genetic elements (MGE) are involved in parasitic or symbiotic relationships with all cellular life forms[1-5], and theoretical models indicate that emergence of such selfish elements is an intrinsic feature of replicator systems[6-11]. Thus, genetic parasites must have been inalienable components of life from its very beginnings. Unlike cellular life forms, viruses employ all existing types of nucleic acids as replicating genomes that are packaged into virions. This diversity of the replication and expression strategies has been captured in a systematic form in the 'Baltimore classification' of viruses[12]. Recently, we undertook a comprehensive reappraisal of the findings of virus phylogenomics to assess the evolutionary status of each of the Baltimore classification groups[13]. This synthesis culminated in the identification of four realms (the highest rank in virus taxonomy) of viruses that are monophyletic with respect to their core gene sets and partially overlap with the Baltimore classification: *Riboviria*, *Monodnaviria*, *Duplodnaviria* and *Varidnaviria*. *Riboviria* includes viruses with positive-sense, negative-sense and double-stranded RNA (dsRNA) genomes as well as reverse-transcribing viruses with RNA and DNA genomes. Members of this realm are unified by the homologous RNA-dependent RNA polymerases (RdRPs) and reverse transcriptases (RTs). *Monodnaviria* includes single-stranded DNA (ssDNA) viruses together with small dsDNA viruses (papillomaviruses and polyomaviruses) that are unified by the distinct endonuclease (or its inactivated derivative) involved in the initiation of the genome replication. *Duplodnaviria* include tailed dsDNA bacteriophages and archaeal viruses along with animal herpesviruses that are unified by the distinct morphogenetic module consisting of HK97-fold major capsid proteins (MCPs), homologous genome packaging ATPases-nucleases (terminases), portal proteins and capsid maturation proteases. *Varidnaviria* is an enormously diverse assemblage of viruses infecting bacteria, archaea and eukaryotes, that are unified by the vertical jelly-roll MCPs (most groups possess double jelly-roll (DJR) MCPs but some have a single vertical jelly-roll domain that is the likely an ancestral form) along with a distinct type of genome packaging ATPase present in most constituent groups. This megataxonomy of viruses has been recently formally adopted by the International Committee for the Taxonomy of Viruses (ICTV)[14,15]. Apart from the four monophyletic realms, several groups of viruses remain unaffiliated in the emergent megataxonomy, most notably, the diverse dsDNA viruses of hyperthemophilic archaea that form several distinct, seemingly, unrelated groups[16-18].

In another recent synthesis, we examined the origins of the replication and structural modules of viruses, and posited a 'chimeric' scenario of virus evolution[19]. Under this model, the replication machineries of each of the four realms derive from the primordial pool of genetic elements, whereas the major virion structural proteins were acquired from cellular hosts at different stages of evolution giving rise to bona fide viruses.

In this Perspective article, we combine this recent work with observations on the host ranges of viruses in each of the four realms, along with deeper reconstructions of virus evolution, to tentatively infer the composition of the virome of the LUCA.

[H1] The LUCA

Evidently, in order to make any meaningful inferences regarding the viruses that infected the LUCA, we must have at least a general notion of the characteristics of this ancestral life form. Considerable efforts have been undertaken over the years to deduce the gene composition and biological features of the

LUCA from comparative genome analyses combined with biological reasoning[20-22]. These inferences are challenged by the complex evolutionary histories of most genes (with partial exception for the core components of the translation and transcription systems) that involved extensive horizontal transfer and non-orthologous gene displacement[23-25]. Nevertheless, on the strength of combined evidence, it appears likely that the LUCA was a prokaryote-like organism (that is, like bacteria or archaea) of considerable genomic and organizational complexity[20,26-28]. Formal reconstructions of the ancestral gene repertoires based on maximum parsimony and maximum likelihood approaches assign several hundred genes to the LUCA that are responsible for most of the core processes characteristic of prokaryotic cells[22,27,29], perhaps, making it comparable to the simplest extant free-living bacteria and archaea (~1,000 genes or even more complex given that the accessory gene repertoire is not amenable to a straightforward reconstruction). However, the nature of the replication and membrane machineries of LUCA remains unclear due to the drastic differences between the respective systems of bacteria and archaea, the two primary domains of life[30-33].

The fact that the replicative DNA polymerases (DNAPs) of bacteria, archaea and eukaryotes are not homologous has prompted ideas of an RNA-based LUCA[31,34,35]. However, the recent discovery of the structural similarity between the catalytic cores of the archaeal replicative family D DNA polymerase, PolD, and the universal DNA-directed RNA polymerase (RNAP)[36,37] implies a common origin of replication and transcription, and suggests an 'archaeal-like' replication machinery in LUCA, with PolD serving as the replicative DNAP[38]. Evolutionary reconstructions point to a fairly complex replication apparatus in the LUCA, with processive replication aided by the sliding clamp (proliferating cell nuclear antigen; PCNA), clamp loader, replicative helicase and the ssDNA-binding protein. Similarly, the transcription system of the LUCA can be inferred to have already included a multisubunit RNAP with the duplicated large subunits, some smaller subunits and multiple transcription regulators[39].

The membranes of archaea and bacteria consist of different types of phospholipids, namely, isoprenoid ethers and fatty acid esters, respectively, with different chiralities of the glycerol-phosphate moiety[33]. Although the possibility of a membrane-less LUCA has been brought up[40,41], the general considerations on the essentiality of compartmentalization and the universal conservation of certain key membrane-associated components, such as the signal recognition particle, leave little doubt that LUCA had membrane-bounded cells[42]. The nature of the membrane in the LUCA remains uncertain, however. Phylogenomic analyses indicate that LUCA encoded the biosynthetic pathways for both bacterial and archaeal phospholipids, implying an ancestral mixed membrane, with subsequent differentiation[30-33]. Notably, preliminary data suggest that bacteria of the Fibrobacteres–Chlorobi–Bacteroidetes (FCB) group superphylum and related candidate phyla encode a complete pathway for archaeal membrane lipid biosynthesis, in addition to the bacterial fatty acid membrane pathway, suggesting that certain contemporary bacterial lineages have mixed heterochiral membranes[43]. Such a possibility is consistent with the results of recent experiments demonstrating the viability of bacteria with an engineered mixed, archaeal–bacterial membrane[44]. The same considerations apply to the cell wall that is represented by the peptidoglycan in most bacteria[45,46] and the proteinaceous S-layer in archaea and some bacteria[47]. Importantly, however, in this case, the possibility of a wall-less LUCA cannot be dismissed.

The genetic composition of modern prokaryotes is best described in terms of the pangenome, that is, the entirety of the genes that are found in organisms with closely related core genomes that are traditionally considered to constitute a species[48-50]. The accessory genes that are present in each strain in addition to the core genome, and collectively account for the bulk of the pangenome, include diverse

anti-parasite defence systems, genes involved in inter-microbial conflicts, such as antibiotic production and resistance, and integrated MGE. Given that genetic parasites are intrinsic components of any replicator system, this pangenome structure necessarily should have been established at the earliest stages of cellular evolution. Thus, although important features of the LUCA remain to be clarified, we can conclude with reasonable confidence that it was a prokaryotic population with a pangenomic complexity comparable to that of the extant archaea and bacteria. The attempt on the reconstruction of the LUCA virome that we undertake here provides some insights into the pangenome of the LUCA that we discuss in the final section.

[H1] LUCA and the four viral realms

Examination of the host ranges of viruses in each of the four realms[13], and in particular, assessment of the relationships between bacterial and archaeal members allows us to make inferences on the composition of the LUCA virome. Widespread groups with a clear dichotomy between archaeal and bacterial viruses are the best candidates for components of the virome of the LUCA. We start by mapping the major groups of viruses to the evolutionary trees of bacterial and archaeal hosts[51] (omitting eukaryotes as a derived domain of life that emerged at a later stage of evolution and hence is irrelevant as far as the LUCA is concerned[52]) and thereby inferring their likely presence or absence in the LUCA virome. The results of this reconstruction (Figures 1 and 2) suggest that the LUCA virome was dominated by dsDNA viruses. More specifically, several groups of tailed dsDNA viruses (*Duplodnaviria*) were assigned to the LUCA virome indicating that, at least, this realm of viruses already reached considerable diversity prior to the radiation of archaea and bacteria (Figure 3). All viruses of this realm share homologous MCPs (HK97-fold), large and small terminase subunits, prohead maturation proteases and portal proteins indicating that their morphogenetic modules are monophyletic[53-58]. Duplodnaviruses are broadly distributed among both bacteria and archaea (Figures 1 and 2), and crucially, comparative genomic analyses suggest that the archaeal and bacterial viruses within *Duplodnaviria*, on a broad scale, have coevolved with their respective hosts[59] (see discussion below).

Tailed bacteriophages are nearly universal among bacteria[60]. In archaea, duplodnaviruses or related proviruses (virus genome integrated into the cellular chromosome) have been detected in many mesophilic as well as extremophilic lineages of the phyla Euryarchaeota and Thaumarchaeota[56,61]. Furthermore, HK97-fold MCPs were identified in uncultivated archaea of the proposed phyla Aenigmarchaeota, Altiarchaeota, Nanoarchaeota, Micrarchaeota, Iainarchaeota and Asgardarchaeota (Fig. 2). However, given the potential artifacts associated with binning of contigs from environmental genomics projects, the host assignment for these (pro)viruses should be considered with utmost caution. Nevertheless, the distribution of tailed archaeal duplodnaviruses appears to encompass highly diverse environments, mirroring the situation of their bacterial relatives, and consistent with the presence of this group in the LUCA virome. Although it is difficult to precisely map specific groups of duplodnaviruses to the LUCA virome, the presence of viruses with short tails (podovirus morphology), long non-contractile tails (siphovirus morphology) and contractile tails (myovirus morphology) in both bacteria and archaea implies that all these morphologies were already represented in the LUCA virome (Figure 3). The alternative possibility, namely, that all three major groups of duplodnaviruses (that is, siphoviruses, myoviruses and podoviruses) were transferred between bacteria and archaea at later stages of evolution cannot be formally excluded but appears less parsimonious. Furthermore, the

observations that many bacterial members of the *Duplodanviria* encode archaeal-like genome replication modules[62], which are not homologous to the bacterial functional counterparts, also argues in favor of the origin of this virus group antedating the archaeal–bacterial divide.

The second realm of dsDNA viruses, *Varidnaviria*, is represented in prokaryotes by four families of bacterial viruses (*Tectiviridae*, *Corticoviridae*, *Autolykiviridae* and *Finnlakeviridae*), one family of archaeal viruses (*Turriviridae*), and the family *Sphaerolipoviridae*, in which different genera include viruses infecting either bacteria or archaea. However, mining metagenomic data for homologs of the DJR MCP using sensitive computational methods resulted in the discovery of a vast diversity of previously unknown viruses of this realm that, in all likelihood, infect prokaryotes[63,64]. Actual host assignments await but some of these virus genomes were found in geothermal habitats, strongly suggesting archaeal hosts[63,64]. Perhaps, even more informative has been the analysis of bacterial and archaeal genomes for the presence of proviruses encoding DJR MCPs, which has substantially expanded the reach of *Varidnaviria* in both prokaryotic domains[65-67]. Phylogenetic analysis of the concatenated DJR MCP and genome packaging ATPases of archaeal varidnaviruses suggested coevolution of this group of viruses with the major archaeal lineages, rather than recent horizontal transfer from bacteria[66]. Thus, most likely, the LUCA virome also included multiple groups of dsDNA viruses with vertical (both single and double) jelly-roll MCPs (Fig. 3). Furthermore, reconstruction of the DJR MCP evolution sheds light on the pre-LUCA stages of virus evolution as discussed in the next section.

Among the ssDNA viruses (realm *Monodnaviria*), only members of a single order, *Tubulavirales* (until recently known as the family *Inoviridae*), consisting of filamentous or rod-shaped viruses, appear to be hosted by both bacteria and archaea. However, whereas tubulaviruses are ubiquitous in bacteria, their association with archaea was inferred from putative proviruses present in several archaeal lineages, namely, methanogens and aenigmarchaea[68]. Such distribution has been judged best compatible with horizontal virus transfer from bacteria to archaea[68]. Given their ubiquity in bacteria, the origin of filamentous bacteriophages concomitantly or soon after the emergence of the last bacterial common ancestor (LBCA) appears likely whereas their presence in LUCA cannot be ruled out either (Fig. 3). Similarly, microviruses with icosahedral capsids and circular ssDNA genomes are nearly ubiquitous in the environment and are genetically highly diverse[69-71]. Although for the vast majority of these viruses the hosts are unknown, the few known isolates infect broadly diverse bacteria from five different phyla (Fig. 1). It is thus likely that microviruses have a long-standing evolutionary history in bacteria, which probably dates back at least to the LBCA (Fig. 3).

In the extant biosphere, RNA viruses dominate the eukaryotic virome but are rare in bacteria, compared to DNA viruses, and unknown in archaea[72]. Bacterial RNA viruses are represented by two families, the positive-sense RNA *Leviviridae* and dsRNA *Cystoviridae*[60,73]. The host range of the experimentally identified members of both families is limited to a narrow range of bacteria (almost exclusively, Proteobacteria). However, recent metagenomics efforts have drastically expanded the known diversity of leviviruses indicating that their share in the prokaryotic virome had been substantially under-appreciated[74,75].

Reverse-transcribing viruses are conspicuously confined to eukaryotes although prokaryotes carry a substantial diversity of non-packaging (that is, non-viral) retroelements, for example, group II introns[76,77]. The extant distribution of the viruses of the realm *Riboviria*, with its drastic display of eukaryotic over prokaryotic host ranges, might appear paradoxical, given the broadly accepted RNA

world concept of the origin of life[78-80], implying the early origin of RdRP and RT, and as a consequence, the primordial status of RNA viruses. The origin of leviviruses within bacteria is best compatible with their currently characterized distribution (Fig. 1) and is a distinct possibility. However, given the lack of obvious direct ancestors of the RdRP among RTs of bacterial retroelements, and the ever expanding diversity of leviviruses through metagenomics[74,75], we consider that origin of levivirus ancestors at the pre-LUCA stage of evolution and their presence in the virome of the LUCA cannot be ruled out, even if not supported by the currently available data. Conceivably, at the LUCA stage and later, primordial RNA viruses were losing the evolutionary competition with the more efficient dsDNA viruses and went extinct in many lines of descent, including archaea. Under this scenario, the renaissance of the RNA viruses occurred only in eukaryotes, arguably, thanks to the combination of barriers for DNA virus replication created by the nucleus and the emergence of the cytosolic endomembrane system that became a niche favorable to RNA virus reproduction[72].

Furthermore, unlike the LUCA, for which most evolutionary reconstructions suggest a mesophilic or a moderate thermophilic lifestyle[81,82], the last common ancestors of bacteria and archaea are inferred to have been thermophiles or hyperthermophiles[83,84]. Extreme high temperatures might be restrictive for propagation of RNA viruses and thus could represent a bottleneck associated with the demise of the ancestral RNA virome (and potentially explain why RNA viruses are unknown in archaea)[85]. The family *Cystoviridae* that includes dsRNA viruses has an even narrower host range than the leviviruses, suggesting a later origin. Thus, of the realm *Riboviria*, positive-sense RNA viruses are a putative component of the LUCA virome, whereas dsRNA viruses, negative-sense RNA viruses and all reverse-transcribing viruses appear to be subsequent additions to the virus world, the latter two taxa emerging only in eukaryotes.

The ancestral status of many archaea-specific virus groups is difficult to ascertain. However, some monophyletic virus assemblages, such as those with spindle-shaped virions[16,86], infect hosts from all major archaeal lineages (Figure 2) and thus can be traced to the last archaeal common ancestor (LACA). Therefore, their presence in the LUCA virome, with subsequent loss in the bacterial lineage, cannot be ruled out either.

[H1] Virus evolution before the LUCA

Likely cellular ancestors are identifiable for many major virion proteins, on the basis of phylogenomic analyses of the corresponding protein families[87]. Reconstruction of the evolutionary paths from ancestral host proteins to viral capsids sheds light on the early stages of evolution of both realms of dsDNA viruses (Figure 4). The DJR MCP of the *Varidnaviria* appears to be a unique virus feature, with no potential cellular ancestors detected. By contrast, the single jelly-roll (SJR) MCP of numerous RNA viruses that was also acquired by ssDNA viruses through recombination can be traced to ancestral cellular carbohydrate-binding proteins, with several probable points of entry into the virus world[87]. Thus, the DJR MCP, in all likelihood, evolved from the SJR MCP early in the evolution of viruses. Remarkably, apparent evolutionary intermediates are detectable in two virus families. Viruses in the family *Sphaerolipoviridae* encode two 'vertically' oriented SJR MCPs that are likely to represent the ancestral duplication preceding the fusion that gave rise to the DJR MCP[88-90]. The recently discovered archaeal dsDNA viruses in the family *Portogloboviridae*[91] contain one SJR MCP[92] and thus appear to represent an even earlier evolutionary intermediate (Figure 4). Indeed, structural comparisons of the SJR

MCPs from RNA and DNA viruses show that the portoglobovirus MCP is most closely related to the MCPs of sphaerolipoviruses[92]. Combined with the inferred presence in the LUCA virome, of multiple groups of *Varidnaviria*, the discovery of the intermediate MCP forms in capsids of extant viruses implies extensive evolution of varidnaviruses predating the LUCA. The families *Portogloboviridae* and *Sphaerolipoviridae* appear to be relics of the pre-LUCA evolution of varidnaviruses and, accordingly, must have been part of the LUCA virome.

For the members of the second realm of dsDNA viruses, *Duplodnaviria*, no cellular ancestor was detected in the dedicated comparative analyses of the sequences and structures of virion proteins[87]. However, a recent structural comparison has shown that the main scaffold of the HK97-like MCP belongs to the strand-helix-strand-strand (SHS2) fold (with the insertion of an additional, uncharacterized domain of the DUF1884 (PF08967) family[93]) and appears to be specifically related to the dodecin family of the SHS2-fold proteins[94]. Dodecins are widespread proteins in bacteria and archaea that form dodecameric compartments involved in flavin sequestration and storage[95], and thus are plausible ancestors for the HK97-fold MCP. Although, in this case, there are no detectable evolutionary intermediates among viruses, the inferred presence of multiple groups of duplodnaviruses in the LUCA virome implies that the recruitment of dodecin and the insertion of DUF1884 are ancient events. Consistently, viruses with short tails (podovirus morphology), long non-contractile tails (siphovirus morphology) and long contractile tails (myovirus morphology) are all found in both bacteria and archaea, indicating that the morphogenetic toolkit of viruses with HK97-fold MCPs attained considerable versatility in the pre-LUCA era.

[H1] Virus replication modules

Each virus genome includes two major functional modules, one for virion formation (morphogenetic module), and the other one for genome replication[96]. The two modules rarely display congruent histories over long evolutionary spans and instead are exchanged horizontally between different groups of viruses through recombination, continuously producing new virus lineages. In the previous sections, we show that the morphogenetic modules including the vertical jelly-roll and HK97-fold MCPs can be traced to the LUCA virome.

What about the replication modules? One of the most widespread replication modules in the virosphere is the rolling circle replication endonuclease (RCRE) of the HUH superfamily[97]. Homologous RCREs are encoded by viruses with single and double jelly-roll MCPs, HK97-like MCPs and morphologically diverse ssDNA viruses, and are also found in many families of bacterial and archaeal plasmids and transposons[98]. Thus, RCRE can be confidently assigned to the LUCA virome or mobilome (that is, all the MGEs of the LUCA).

Protein-primed family B DNA polymerases (pPolB) represent another replication module with a broad distribution spanning several families of viruses and non-viral MGEs[62]. pPolB is present in bacteria-infecting members of the realms *Duplodnaviria* (phi29-like podoviruses) and *Varidnaviria* (*Tectiviridae*, *Autolykiviridae* and diverse varidnavirus genomes identified in metagenomic data) as well as in several families of archaeal viruses (*Halspiviridae*, *Thaspiviridae, Ovaliviridae* and *Pleolipoviridae*). In phylogenetic analyses, pPolBs split into two separate clades corresponding to bacterial and archaeal

viruses[99,100], strongly suggesting that they have coevolved with bacterial and archaeal lineages ever since their divergence from LUCA.

Two other key replication proteins that are among the most common in bacterial and archaeal viruses and MGEs are primases of the archaeo-eukaryotic primase (AEP) superfamily[101,102] and superfamily 3 helicases (S3H)[62]. Whereas S3H are exclusive to viruses and MGEs, the viral AEPs form specific families which are not closely related to the cellular homologs. Notably, bacteria do not employ AEP for primer synthesis and so bacterial viruses could not have recruited this protein from their hosts. Thus, AEPs and S3H, along with RCRE and pPolB, appear to represent major components of the replication modules of the LUCA virome.

More generally, contemporary duplodnaviruses display a remarkable diversity of genome replication modules, from minimalist initiators that recruit cellular DNA replisomes for viral genome replication to near-complete virus-encoded DNA replication machineries[62,103]. In many cases, these DNA replication proteins do not have close cellular homologs, suggesting long evolutionary history within the virus world. Notably, some of the phage proteins, such as helicase loaders, have replaced their cellular counterparts at the onset of certain bacterial lineages for replication of cellular chromosomes[104]. Although some bacterial tailed dsDNA viruses encode replication factors of apparent bacterial origin, in archaeal duplodnaviruses)[57], the proteins involved in informational processes, including components of genome replication machinery, DNA repair and RNA metabolism, are of archaeal type, with none of the known archaeal viruses encoding components of bacterial-type replication machinery[61,62]. Finally, tailed archaeal viruses carry archaeal or eukaryotic-like promoters[105,106], consistent with the fact that none of the known archaeal viruses encode RNA polymerases[16,107], further pointing to long-term coevolution with the hosts. These considerations argue against (recent) horizontal transfers of duplodnaviruses between bacteria and archaea accounting for the observed distribution of these viruses, even though some such transfers might have occurred. Thus, analyses of duplodnavirus and varidnavirus genome replication modules complement those of the morphogenetic modules and suggest extensive divergence of both groups of viruses in the pre-LUCA era.

[H1] Conclusions

The informal reconstructions attempted here suggest a remarkably diverse, complex LUCA virome. This ancestral virome was likely dominated by dsDNA viruses from the realms *Duplodnaviria* and *Varidnaviria.* In addition, two groups of ssDNA viruses (realm *Monodnaviria*), namely, *Microviridae* and *Tubulavirales*, can be traced to LBCA, whereas spindle-shaped viruses, most likely, infected LACA. The possibility that these virus groups were present in the LUCA virome but were subsequently lost in one of the two primary domains cannot be dismissed. The point of origin of the extant bacterial positive-sense RNA viruses (realm *Riboviria*) remains uncertain, with both bacterial and primordial origins remaining viable scenarios. Further virus prospecting efforts could shed light on the history of these viruses. Although the inferred LUCA virome, in all likelihood, did not include members of many extant groups of viruses of prokaryotes, its apparent complexity seems to exceed the typical complexity of well-characterized viromes of a bacterial or archaeal species. These observations imply that the LUCA was not a homogenous microbial population, but rather, a community of diverse microorganisms, with a shared gene core that was inherited by all descendant life-forms and a diversified pangenome that included various genes involved in virus–host interactions, in particular, multiple defence systems.

According to the 'chimeric' scenario of virus origins, different groups of viruses evolved through recruitment of cellular proteins as virion components[19]. Here, we present evidence that — contingent on our mapping of both duplodnaviruses and varidnaviruses to the virome of LUCA — several such events occurred already in the earliest phase of life's evolution, from the primordial pool of replicators to the LUCA. Moreover, virus evolution during that early era went through multiple, distinct stages as demonstrated by the reconstructed histories of the capsid proteins of the two realms of dsDNA viruses. The cellular ancestors of the varidnavirus DJR MCPs, the SJR-containing carbohydrate-binding or nucleoplasmin-like proteins, and the dodecins, the ancestors of the duplodnavirus MCPs, belong to expansive protein families that have already undergone substantial diversifying evolution prior to the origins of the two realms of viruses. The respective protein families do not belong to the universal core of cellular life, so their apparent pre-LUCA diversification further emphasizes the substantial pangenomic, organizational and functional complexity of the LUCA. This conclusion is indeed compatible with the previous inferences on LUCA made from the analysis of coalescence in different families of ancient genes, namely, that a common ancestor containing all the genes shared by the three domains of life has never existed[108].

Straightforward thinking on the LUCA virome might have envisaged it as a domain of RNA viruses descending from the primordial RNA world. The reconstructions, however, suggest otherwise, indicating that the LUCA was similar to the extant prokaryotes with respect to the repertoire of viruses it hosted. These findings do not defy the RNA world scenario but mesh well with the conclusion that DNA viruses have evolved and diversified extensively already in the pre-LUCA era. The RNA viruses, after all, might have been the first to emerge but, by the time LUCA lived, they have already been largely supplanted by the more efficient DNA virosphere.

Author contributions

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

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. M.K. was supported by l'Agence Nationale de la Recherche grant ANR-17-CE15-0005-01.

Peer review information

Nature Reviews Microbiology thanks J. P. Gogarten, P. López-Garcìa and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note

Supplementary information

Supplementary information is available for this paper at https://doi.org/10.1038/s415XX-XXX-XXXX-X

References

1 Forterre, P. & Prangishvili, D. The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. *Ann N Y Acad Sci* **1178**, 65-77 (2009).

2 Koonin, E. V., Senkevich, T. G. & Dolja, V. V. The ancient Virus World and evolution of cells. *Biol Direct* **1**, 29 (2006).

3 Moreira, D. & Lopez-Garcia, P. Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol* **7**, 306-11 (2009).

4 Koonin, E. V. & Dolja, V. V. A virocentric perspective on the evolution of life. *Curr Opin Virol* **3**, 546-57 (2013).

5 Raoult, D. & Forterre, P. Redefining viruses: lessons from Mimivirus. *Nat Rev Microbiol* **6**, 315-9 (2008).

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

7 Szathmary, E. & Demeter, L. Group selection of early replicators and the origin of life. *J Theor Biol* **128**, 463-86 (1987).

8 Takeuchi, N. & Hogeweg, P. The role of complex formation and deleterious mutations for the stability of RNA-like replicator systems. *J Mol Evol* **65**, 668-86 (2007).

9 Takeuchi, N. & Hogeweg, P. Evolutionary dynamics of RNA-like replicator systems: A bioinformatic approach to the origin of life. *Phys Life Rev* **9**, 219-63 (2012).

10 Eigen, M. The origin of genetic information: viruses as models. *Gene* **135**, 37-47 (1993).

11 Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465-523 (1971).

12 Baltimore, D. Expression of animal virus genomes. *Bacteriol Rev* **35**, 235-41 (1971).

13 Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol Mol Biol Rev* **84**, e00061-19 (2020).

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

15 Siddell, S. G. *et al.* Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018). *Arch Virol* **164**, 943-946 (2019).

16 Prangishvili, D. *et al.* The enigmatic archaeal virosphere. *Nat Rev Microbiol* **15**, 724-739 (2017).

17 Munson-McGee, J. H., Snyder, J. C. & Young, M. J. Archaeal Viruses from High-Temperature Environments. *Genes (Basel)* **9**, E128 (2018).

18 Iranzo, J., Koonin, E. V., Prangishvili, D. & Krupovic, M. Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. *J Virol* **90**, 11043-11055 (2016).

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

20 Koonin, E. V. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* **1**, 127-36 (2003).

21 Glansdorff, N., Xu, Y. & Labedan, B. The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol Direct* **3**, 29 (2008).

22 Weiss, M. C. *et al.* The physiology and habitat of the last universal common ancestor. *Nat Microbiol* **1**, 16116 (2016).

23 Doolittle, W. F. Uprooting the tree of life. *Sci Am* **282**, 90-5 (2000).

24 Puigbo, P., Wolf, Y. I. & Koonin, E. V. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J Biol* **8**, 59 (2009).

25 Berkemer, S. J. & McGlynn, S. E. A new analysis of archaea-bacteria domain separation: variable phylogenetic distance and the tempo of early evolution. *Mol Biol Evol* (2020).

26 Fournier, G. P., Andam, C. P. & Gogarten, J. P. Ancient horizontal gene transfer and the last common ancestors. *BMC Evol Biol* **15**, 70 (2015).

27 Ouzounis, C. A., Kunin, V., Darzentas, N. & Goldovsky, L. A minimal estimate for the gene content of the last universal common ancestor--exobiology from a terrestrial perspective. *Res Microbiol* **157**, 57-68 (2006).

28 Gogarten, J. P. & Taiz, L. Evolution of proton pumping ATPases: Rooting the tree of life. *Photosynth Res* **33**, 137-46 (1992).

29 Mirkin, B. G., Fenner, T. I., Galperin, M. Y. & Koonin, E. V. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* **3**, 2 (2003).

30 Brown, J. R. & Doolittle, W. F. Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev* **61**, 456-502 (1997).

31 Leipe, D. D., Aravind, L. & Koonin, E. V. Did DNA replication evolve twice independently? *Nucleic Acids Res* **27**, 3389-401 (1999).

32 Pereto, J., Lopez-Garcia, P. & Moreira, D. Ancestral lipid biosynthesis and early membrane evolution. *Trends Biochem Sci* **29**, 469-77 (2004).

33 Lombard, J., Lopez-Garcia, P. & Moreira, D. The early evolution of lipid membranes and the three domains of life. *Nat Rev Microbiol* **10**, 507-15 (2012).

34 Mushegian, A. R. & Koonin, E. V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* **93**, 10268-73 (1996).

35 Forterre, P. 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**, 3669-74 (2006).

36 Raia, P. *et al.* Structure of the DP1-DP2 PolD complex bound with DNA and its implications for the evolutionary history of DNA and RNA polymerases. *PLoS Biol* **17**, e3000122 (2019).

37 Sauguet, L. The Extended "Two-Barrel" Polymerases Superfamily: Structure, Function and Evolution. *J Mol Biol* **431**, 4167-4183 (2019).

38 Koonin, E. V., Krupovic, M., Ishino, S. & Ishino, Y. The replication machinery of LUCA: Common origin of DNA replication and transcription *BMC Biology* **18**, 61 (2020).

39 Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nat Rev Microbiol* **9**, 85-98 (2011).

40 Martin, W. & Russell, M. J. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philos Trans R Soc Lond B Biol Sci* **358**, 59-83; discussion 83-5 (2003).

41 Koonin, E. V. & Martin, W. On the origin of genomes and cells within inorganic compartments. *Trends Genet* **21**, 647-54 (2005).

42 Mulkidjanian, A. Y., Galperin, M. Y. & Koonin, E. V. Co-evolution of primordial membranes and membrane proteins. *Trends Biochem Sci* **34**, 206-15 (2009).

43 Villanueva, L. *et al.* Bridging the divide: bacteria synthesizing archaeal membrane lipids. *bioRxiv* **doi: https://doi.org/10.1101/448035** (2018).

44 Caforio, A. *et al.* Converting Escherichia coli into an archaebacterium with a hybrid heterochiral membrane. *Proc Natl Acad Sci U S A* **115**, 3704-3709 (2018).

45 Rajagopal, M. & Walker, S. Envelope Structures of Gram-Positive Bacteria. *Curr Top Microbiol Immunol* **404**, 1-44 (2017).
46 Auer, G. K. & Weibel, D. B. Bacterial Cell Mechanics. *Biochemistry* **56**, 3710-3724 (2017).
47 Sleytr, U. B., Schuster, B., Egelseer, E. M. & Pum, D. S-layers: principles and applications. *FEMS Microbiol Rev* **38**, 823-64 (2014).
48 Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Curr Opin Microbiol* **23**, 148-54 (2015).
49 Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr Opin Genet Dev* **15**, 589-94 (2005).
50 Puigbo, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y. I. & Koonin, E. V. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol* **12**, 66 (2014).
51 Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**, 996-1004 (2018).
52 Embley, T. M. & Martin, W. Eukaryotic evolution, changes and challenges. *Nature* **440**, 623-30 (2006).
53 Kristensen, D. M., Cai, X. & Mushegian, A. Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. *J Bacteriol* **193**, 1806-14 (2011).
54 Iranzo, J., Krupovic, M. & Koonin, E. V. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *MBio* **7**, e00978-16 (2016).
55 Duda, R. L. & Teschke, C. M. The amazing HK97 fold: versatile results of modest differences. *Curr Opin Virol* **36**, 9-16 (2019).
56 Krupovic, M., Forterre, P. & Bamford, D. H. Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. *J Mol Biol* **397**, 144-60 (2010).
57 Sencilo, A. *et al.* Snapshot of haloarchaeal tailed virus genomes. *RNA Biol* **10**, 803-16 (2013).
58 Cheng, H., Shen, N., Pei, J. & Grishin, N. V. Double-stranded DNA bacteriophage prohead protease is homologous to herpesvirus protease. *Protein Sci* **13**, 2260-9 (2004).
59 Low, S. J., Dzunkova, M., Chaumeil, P. A., Parks, D. H. & Hugenholtz, P. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nat Microbiol* **4**, 1306-1315 (2019).
60 Dion, M. B., Oechslin, F. & Moineau, S. Phage diversity, genomics and phylogeny. *Nat Rev Microbiol* **18**, 125-138 (2020).
61 Philosof, A. *et al.* Novel Abundant Oceanic Viruses of Uncultured Marine Group II Euryarchaeota. *Curr Biol* **27**, 1362-1368 (2017).
62 Kazlauskas, D., Krupovic, M. & Venclovas, C. The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Res* **44**, 4551-64 (2016).
63 Yutin, N., Bäckström, D., Ettema, T. J. G., Krupovic, M. & Koonin, E. V. Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis. *Virol J* **15**, 67 (2018).
64 Kauffman, K. M. *et al.* A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature* **554**, 118-122 (2018).
65 Krupovic, M. & Bamford, D. H. Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. *Virology* **375**, 292-300 (2008).
66 Krupovic, M. *et al.* Integrated mobile genetic elements in Thaumarchaeota. *Environ Microbiol* **21**, 2056-2078 (2019).
67 Jalasvuori, M. & Koskinen, K. Extending the hosts of Tectiviridae into four additional genera of Gram-positive bacteria and more diverse Bacillus species. *Virology* **518**, 136-142 (2018).

68 Roux, S. *et al.* Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat Microbiol* **4**, 1895-1906 (2019).
69 Roux, S., Krupovic, M., Poulet, A., Debroas, D. & Enault, F. Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS One* **7**, e40418 (2012).
70 Creasy, A., Rosario, K., Leigh, B. A., Dishaw, L. J. & Breitbart, M. Unprecedented Diversity of ssDNA Phages from the Family Microviridae Detected within the Gut of a Protochordate Model Organism (Ciona robusta). *Viruses* **10** (2018).
71 Tisza, M. J. *et al.* Discovery of several thousand highly diverse circular DNA viruses. *Elife* **9**, e51971 (2020).
72 Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479-480**, 2-25 (2015).
73 Wolf, Y. I. *et al.* Origins and Evolution of the Global RNA Virome. *MBio* **9**, e02329-18 (2018).
74 Krishnamurthy, S. R., Janowski, A. B., Zhao, G., Barouch, D. & Wang, D. Hyperexpansion of RNA Bacteriophage Diversity. *PLoS Biol* **14**, e1002409 (2016).
75 Callanan, J. *et al.* Expansion of known ssRNA phage genomes: From tens to over a thousand. *Sci Adv* **6**, eaay5981 (2020).
76 Gladyshev, E. A. & Arkhipova, I. R. A widespread class of reverse transcriptase-related cellular genes. *Proc Natl Acad Sci U S A* **108**, 20311-6 (2011).
77 Zimmerly, S. & Wu, L. An Unexplored Diversity of Reverse Transcriptases in Bacteria. *Microbiol Spectr* **3**, MDNA3-0058-2014 (2015).
78 Gilbert, W. The RNA World. *Nature* **319**, 618 (1986).
79 Joyce, G. F. The antiquity of RNA-based evolution. *Nature* **418**, 214-21 (2002).
80 Bernhardt, H. S. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)(a). *Biol Direct* **7**, 23 (2012).
81 Catchpole, R. J. & Forterre, P. The Evolution of Reverse Gyrase Suggests a Nonhyperthermophilic Last Universal Common Ancestor. *Mol Biol Evol* **36**, 2737-2747 (2019).
82 Cantine, M. D. & Fournier, G. P. Environmental Adaptation from the Origin of Life to the Last Universal Common Ancestor. *Orig Life Evol Biosph* **48**, 35-54 (2018).
83 Boussau, B., Blanquart, S., Necsulea, A., Lartillot, N. & Gouy, M. Parallel adaptations to high temperatures in the Archaean eon. *Nature* **456**, 942-5 (2008).
84 Groussin, M. & Gouy, M. Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in archaea. *Mol Biol Evol* **28**, 2661-74 (2011).
85 Forterre, P. The common ancestor of archaea and eukarya was not an archaeon. *Archaea* **2013**, 372396 (2013).
86 Krupovic, M., Quemin, E. R., Bamford, D. H., Forterre, P. & Prangishvili, D. Unification of the globally distributed spindle-shaped viruses of the Archaea. *J Virol* **88**, 2354-8 (2014).
87 Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc Natl Acad Sci U S A* **114**, E2401-E2410 (2017).
88 Santos-Perez, I. *et al.* Structural basis for assembly of vertical single beta-barrel viruses. *Nat Commun* **10**, 1184 (2019).
89 De Colibus, L. *et al.* Assembly of complex viruses exemplified by a halophilic euryarchaeal virus. *Nat Commun* **10**, 1456 (2019).
90 Rissanen, I. *et al.* Bacteriophage P23-77 capsid protein structures reveal the archetype of an ancient branch from a major virus lineage. *Structure* **21**, 718-26 (2013).
91 Liu, Y. *et al.* A Novel Type of Polyhedral Viruses Infecting Hyperthermophilic Archaea. *J Virol* **91**, e00589-17 (2017).

92 Wang, F. *et al.* A packing for A-form DNA in an icosahedral virus. *Proc Natl Acad Sci U S A* **116**, 22591-22597 (2019).
93 Kelley, L. L. *et al.* Structure of the hypothetical protein PF0899 from Pyrococcus furiosus at 1.85 A resolution. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **63**, 549-52 (2007).
94 Holm, L. DALI and the persistence of protein shape. *Protein Sci* **29**, 128-140 (2020).
95 Grininger, M., Zeth, K. & Oesterhelt, D. Dodecins: a family of lumichrome binding proteins. *J Mol Biol* **357**, 842-57 (2006).
96 Krupovic, M. & Bamford, D. H. Order to the viral universe. *J Virol* **84**, 12476-9 (2010).
97 Ilyina, T. V. & Koonin, E. V. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Res* **20**, 3279-85 (1992).
98 Kazlauskas, D., Varsani, A., Koonin, E. V. & Krupovic, M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nat Commun* **10**, 3425 (2019).
99 Redrejo-Rodríguez, M. *et al.* Primer-Independent DNA Synthesis by a Family B DNA Polymerase from Self-Replicating Mobile Genetic Elements. *Cell Rep* **21**, 1574-1587 (2017).
100 Kim, J. G. *et al.* Spindle-shaped viruses infect marine ammonia-oxidizing thaumarchaea. *Proc Natl Acad Sci U S A* **116**, 15645-15650 (2019).
101 Iyer, L. M., Koonin, E. V., Leipe, D. D. & Aravind, L. Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. *Nucleic Acids Res* **33**, 3875-96 (2005).
102 Kazlauskas, D. *et al.* Novel Families of Archaeo-Eukaryotic Primases Associated with Mobile Genetic Elements of Bacteria and Archaea. *J Mol Biol* **430**, 737-750 (2018).
103 Weigel, C. & Seitz, H. Bacteriophage replication modules. *FEMS Microbiol Rev* **30**, 321-81 (2006).
104 Brezellec, P. *et al.* Domestication of Lambda Phage Genes into a Putative Third Type of Replicative Helicase Matchmaker. *Genome Biol Evol* **9**, 1561-1566 (2017).
105 Pfister, P., Wasserfallen, A., Stettler, R. & Leisinger, T. Molecular analysis of Methanobacterium phage psiM2. *Mol Microbiol* **30**, 233-44 (1998).
106 Iro, M. *et al.* The lysogenic region of virus phiCh1: identification of a repressor-operator system and determination of its activity in halophilic Archaea. *Extremophiles* **11**, 383-96 (2007).
107 Dellas, N., Snyder, J. C., Bolduc, B. & Young, M. J. Archaeal Viruses: Diversity, Replication, and Structure. *Annu Rev Virol* **1**, 399-426 (2014).
108 Zhaxybayeva, O. & Gogarten, J. P. Cladogenesis, coalescence and the evolution of the three domains of life. *Trends Genet* **20**, 182-7 (2004).
109 Mendler, K. *et al.* AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res* **47**, 4442-4448 (2019).
110 Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* **4**, e03125 (2015).
111 Krupovic, M. & Bamford, D. H. Putative prophages related to lytic tailless marine dsDNA phage PM2 are widespread in the genomes of aquatic bacteria. *BMC Genomics* **8**, 236 (2007).
112 Pawlowski, A., Rissanen, I., Bamford, J. K., Krupovic, M. & Jalasvuori, M. Gammasphaerolipovirus, a newly proposed bacteriophage genus, unifies viruses of halophilic archaea and thermophilic bacteria within the novel family Sphaerolipoviridae. *Arch Virol* **159**, 1541-54 (2014).
113 Wang, F. *et al.* Structure of a filamentous virus uncovers familial ties within the archaeal virosphere. *Virus Evol* **6**, veaa023 (2020).
114 Weidenbach, K. *et al.* Methanosarcina Spherical Virus, a Novel Archaeal Lytic Virus Targeting Methanosarcina Strains. *J Virol* **91**, e00955-17 (2017).

115 Liu, Y. *et al.* Identification and characterization of SNJ2, the first temperate pleolipovirus integrating into the genome of the SNJ1-lysogenic archaeal strain. *Mol Microbiol* **98**, 1002-20 (2015).
116 Wang, J. *et al.* A novel family of tyrosine integrases encoded by the temperate pleolipovirus SNJ2. *Nucleic Acids Res* **46**, 2521-2536 (2018).
117 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402 (1997).

Figures and legends

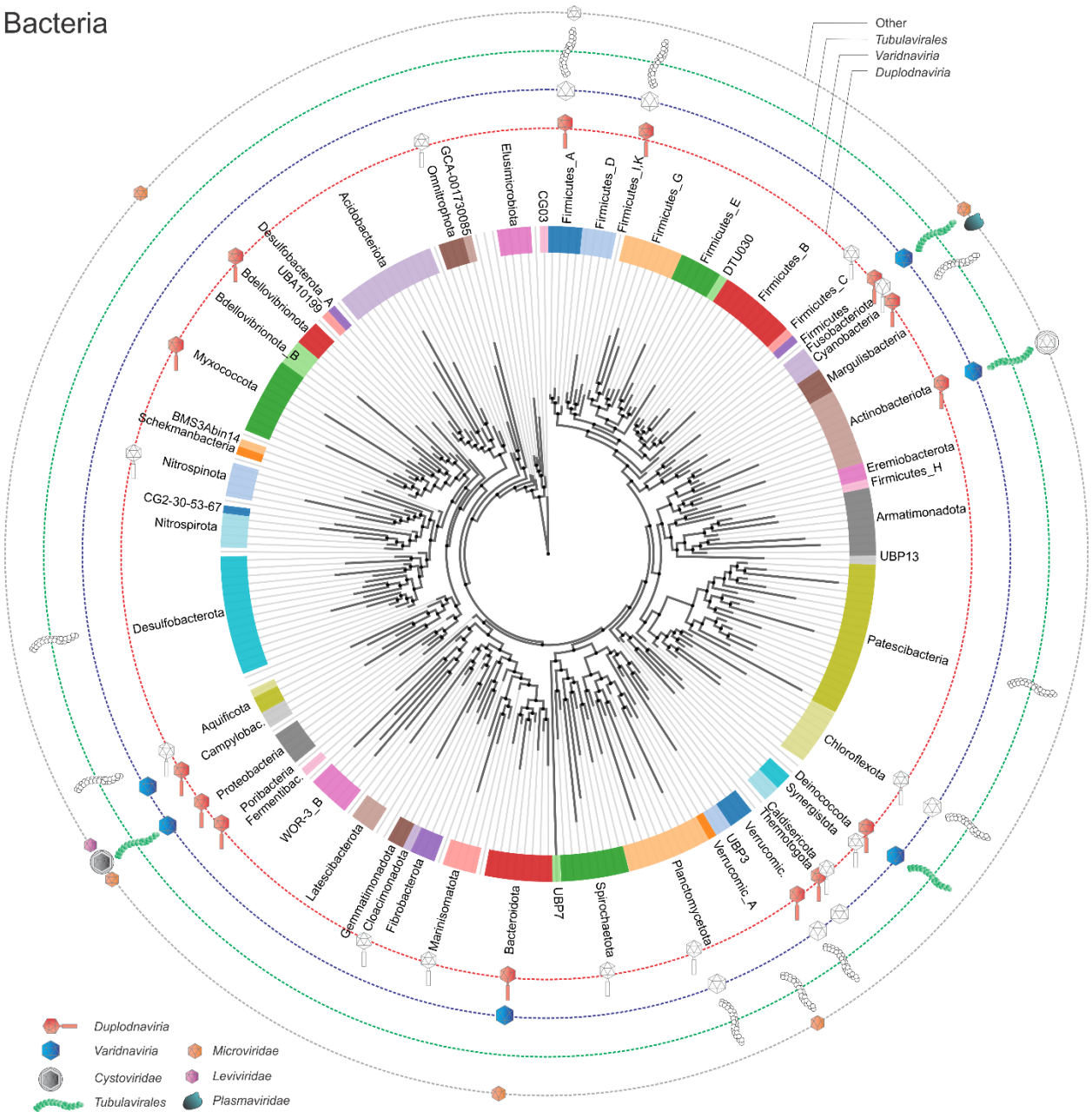

Figure 1. Distribution of known viruses across the evolutionary tree of bacteria.

The figure shows the latest phylogenetic tree of bacteria, with all phyla indicated, and the major groups of viruses known to infect members of these phyla. Virus groups are represented by symbols depicting the corresponding virions. Coloured and open symbols represent virus isolates and virus genomes or putative prophages, respectively. The symbols are arranged on 4 concentric rings: the innermost ring depicts the distribution of members of the realm *Duplodnaviria* (families *Siphoviridae*, *Podoviridae*, *Myoviridae*, *Ackermannviridae* and *Herelleviridae*); the second ring shows members of the realm *Varidnaviria* (families *Tectiviridae*, *Corticoviridae*, *Finnlakeviridae*, *Sphaerolipoviridae* and *Autolykiviridae*); the third ring shows members of the order *Tubulavirales* (families *Inoviridae* and *Plectroviridae*); the fourth ring includes all other virus groups, namely, *Microviridae*, *Leviviridae*,

Cystoviridae and *Plasmaviridae*. The phylogeny and taxonomic nomenclature are from the Genome Taxonomy Database[51] and were visualized with AnnoTree[109]. The information on virus distribution for virus isolates with completely sequenced genomes is from GenBank. The provirus distribution is from previously published work on *Duplodnaviria*[110], *Tubulavirales*[68], *Varidnaviria*[63,64,111,112]. Supplementary Data 1 shows the known virus-host associations across the domains Bacteria and Archaea. In the spreadsheet, a genus name is indicated if a virus is known to infect (or be associated as a provirus with) any member of the phylum or class of Bacteria and Archaea, respectively.

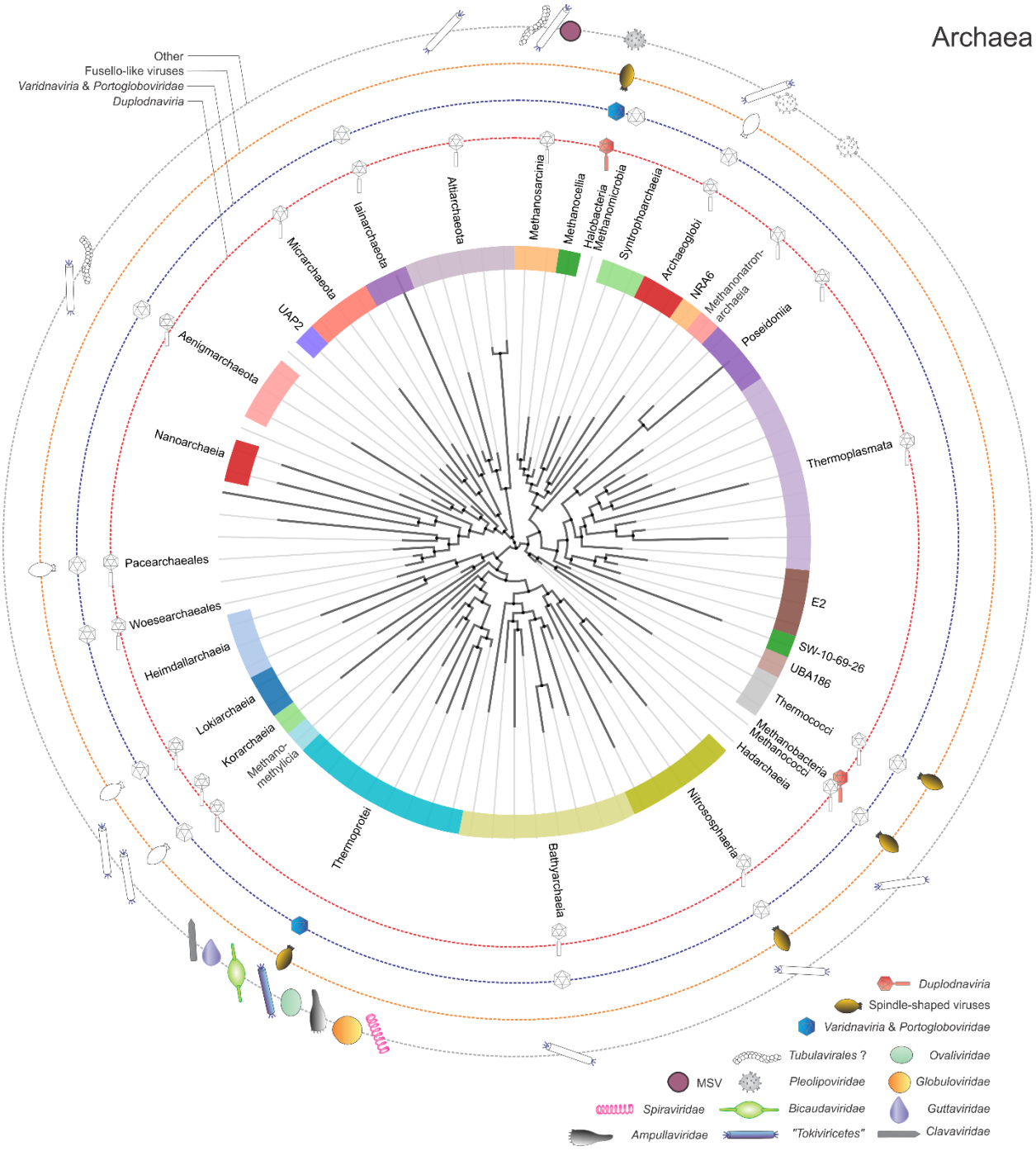

Figure 2. Distribution of known viruses across the evolutionary tree of archaea.

The figure shows the latest phylogenetic tree of archaea, with all phyla indicated, and the major groups of viruses known to infect members of these phyla. Virus groups are represented by symbols depicting the corresponding virions. Coloured and open symbols represent virus isolates and virus genomes or putative proviruses, respectively. The symbols are arranged on 4 concentric rings: the innermost ring depicts the distribution of members of the realm *Duplodnaviria*; the second ring shows members of the realm *Varidnaviria* (families *Turriviridae* and *Sphaerolipoviridae*) and family *Portogloboviridae*; the third

ring shows spindle-shaped viruses (families *Fuselloviridae*, *Halspiviridae* and *Thaspiviridae* and unclassified viruses of Thermococcales); the fourth ring consists of all other virus groups, including the candidate class *"Tokiviricetes"* (*Rudiviridae, Lipothrixviridae* and *Tristromaviridae*)[113] and families *Ampullaviridae, Ovaliviridae, Bicaudaviridae, Guttaviridae, Globuloviridae, Clavaviridae, Spiraviridae* and *Pleolipoviridae,* and unclassified Methanosarcina spherical virus (MSV)[114]. Putative proviruses related to bacterial members of the *Tubulavirales* were identified in some archaeal genomes[68] and are indicated with open symbols. The phylogeny and taxonomic nomenclature are from the Genome Taxonomy Database[51] and were visualized with AnnoTree[109]. The information on virus distribution for virus isolates with completely sequenced genomes is from GenBank. The provirus distribution is from previously published work on *Duplodnaviria*[56,66], *Varidnaviria*[63,66], *Pleolipoviridae*[115,116]. Additional information was obtained by performing BLASTP searches[117] queried with the major capsid proteins of the corresponding viruses against the archaeal genome database at the NCBI. Supplementary Data 1 shows the known virus-host associations across the domains Bacteria and Archaea. In the spreadsheet, a genus name is indicated if a virus is known to infect (or be associated as a provirus with) any member of the phylum or class of Bacteria and Archaea, respectively.

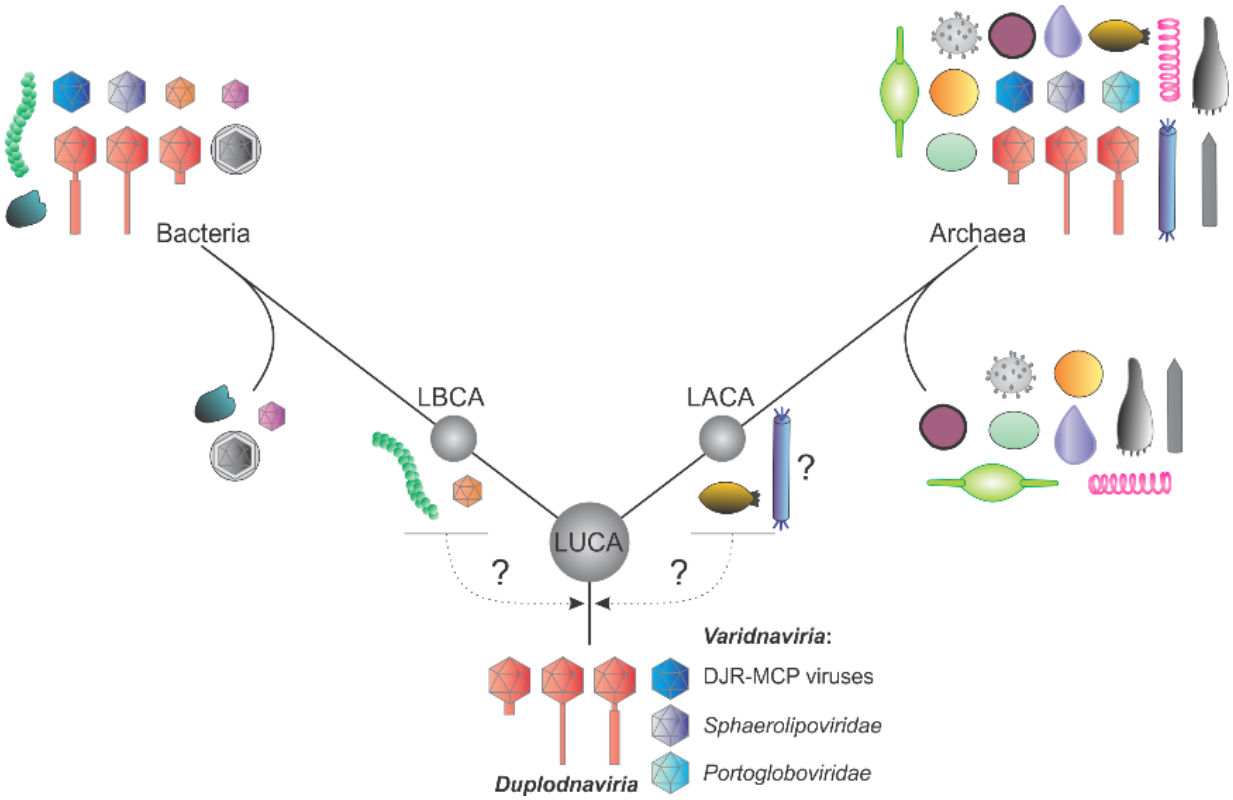

Figure 3. Reconstruction of the LUCA virome from the divergence of the bacterial and archaeal viromes.

This figure shows the hypothetical complex virome of the last universal cellular ancestor (LUCA) as reconstructed from the distribution of viruses among extant phyla of bacteria and archaea. Also schematically depicted are the split of the LUCA virome into the viromes of the last bacterial common ancestor (LBCA) and the last archaeal common ancestor (LACA), as well as the subsequent diversification that resulted in the extant viromes. Divergence of bacteria and archaea from the LUCA is depicted as a bifurcation. Viruses predicted to be associated with the LBCA and the LACA are indicated next to the corresponding grey spheres. Dotted arrows indicate the possibility that the respective viruses might have been represented in the LUCA virome.

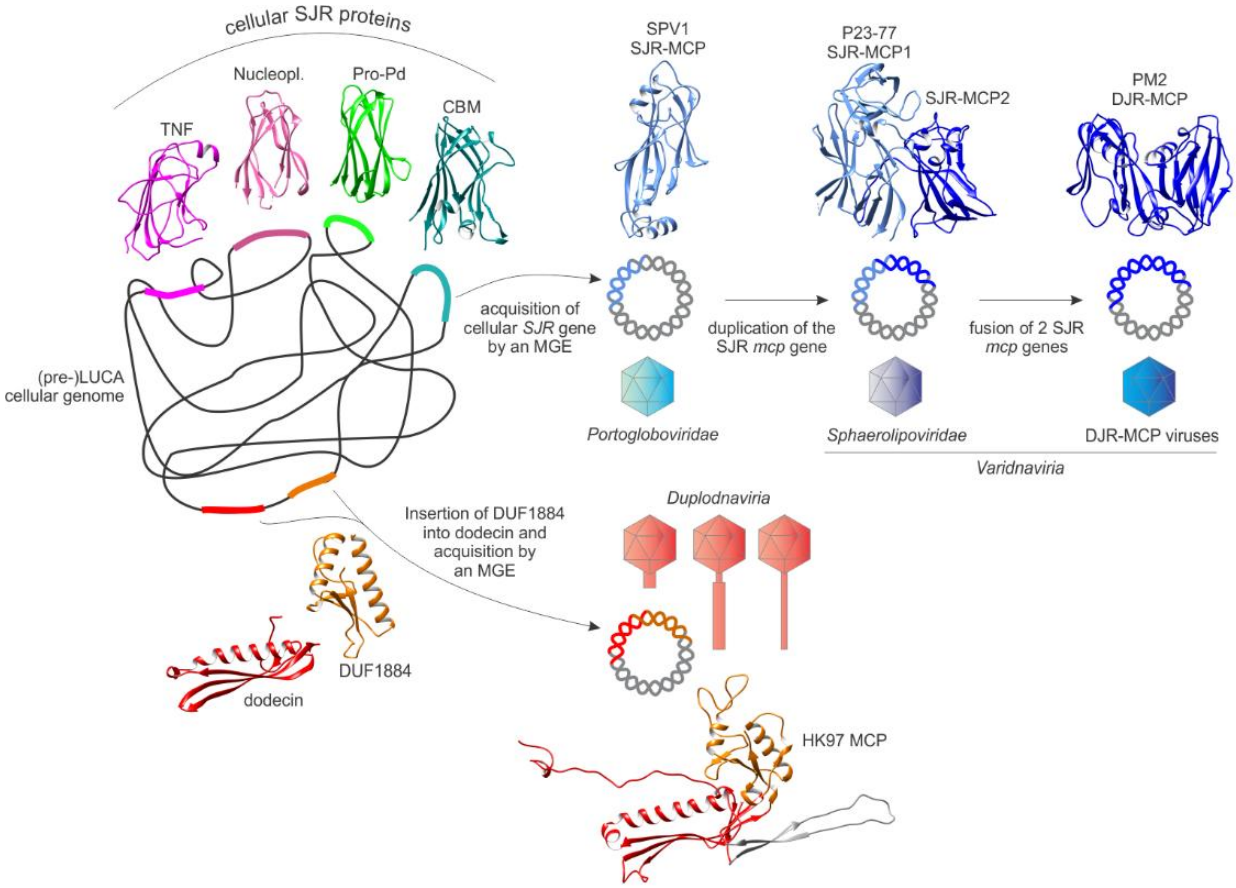

Figure 4. Evolution of double-stranded DNA viruses antedating the LUCA.

The figure shows the origin of single and double jelly-roll (SJR and DJR, respectively), and HK97-fold major capsid proteins (MCP) from cellular ancestors. The capture of the vertical SJR MCP precipitated the emergence of the virus realm *Varidnaviria*, whereas the acquisition of the HK97 MCP gave rise to the realm *Duplodnaviria*. Major evolutionary events are described next to the corresponding arrows. The likely cellular ancestors of the MCPs are shown with thick coloured lines, and the structures of the similarly coloured corresponding proteins are shown next to them: TNF, tumor necrosis factor superfamily protein (PDB ID: 2hey); nucleopl, nucleoplasmin (PDB ID: 1nlq); Pro-Pd, P domain of a subtilisin-like protease (PDB ID: 3afg); CBM, carbohydrate-binding module (PDB ID: 4d3l); dodecin family protein (PDB ID: 3qkb); DUF1884 family protein (PDB ID: 2pk8). The SJR and DJR MCPs and the corresponding virion symbols of members of the realm *Varidnaviria* are colored with different shades of blue. MCPs of duplodnaviruses are represented by gp5 protein of bacteriophage HK97 (PDB ID: 1ohg); *Portogloboviridae* is represented by VP4 of Sulfolobus polyhedral virus 1 (SPV1; PDB ID: 6oj0); *Sphaerolipoviridae* is represented by a heterodimer of MCPs, VP16 and VP17, of *Thermus* bacteriophage P23-77 (PDB ID: 3zn6); DJR MCP viruses are represented by P2 protein of *Pseudoalteromonas* bacteriophages PM2 (family *Corticoviridae*; PDB ID: 2vvf).