The LUCA and its complex virome

Mart Krupovic, Valerian V. Dolja and Eugene V. Koonin

Abstract | The last universal cellular ancestor (LUCA) is the most recent population of organisms from which all cellular life on Earth descends. The 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 the substantial genomic and pan-genomic complexity of the LUCA itself.

Viruses and other mobile genetic elements (MGEs) are involved in parasitic or symbiotic relationships with all cellular life forms1–5, and theoretical models indicate that the emergence of such selfish elements is an intrinsic feature of replicator systems6–11. 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 packaged into virions. This diversity of the replication and expression strategies includes and partially overlap with the Baltimore classification12. 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 double-stranded DNA (dsDNA) viruses (papillomaviruses and polyomaviruses) that are unified by the distinct endonuclease (or its inactivated derivative) involved in the initiation of 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 jelly-roll (SIR) domain that is the likely 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 Viruses14,15. Apart from the four monophyletic realms, several groups of viruses remain unaffiliated in the emergent megataxonomy, most notably the diverse dsDNA viruses of hyperthermophilic archaea that form several distinct, seemingly unrelated groups16–18.

In another recent synthesis, we examined the origins of the replication and structural modules of viruses and posited a ‘chimeric’ scenario of virus evolution19. 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 last universal cellular ancestor (LUCA; also referred to as the last universal common ancestor of cellular organisms).

The LUCA

Evidently, 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 genetic composition and biological features of the LUCA from comparative genome analyses combined with biological reasoning20–22. 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 displacement23–25. 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 complexity26–28. 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 cells29,30, 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 owing to the drastic differences between the respective systems of bacteria and archaea, the two primary domains of life.  

The fact that the replicative DNA polymerases of bacteria, archaea and eukaryotes are not homologous has prompted ideas of an RNA-based LUCA. However, the recent discovery of the structural similarity between the catalytic cores of the archaeal replicative family D DNA polymerase (PolD) and the universal RNA-directed DNA polymerase implies a common origin of replication and transcription and suggests an ‘archaeal-like’ replication machinery in LUCA, with PolD serving as the replicative DNA polymerases. Evolutionary reconstructions point to a fairly complex replication apparatus in the LUCA, with PolD serving as the sliding clamp (proliferating cell nuclear antigen), 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 RNA polymerase with the duplicated large subunits, some smaller subunits and multiple transcription regulators.  

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. Although the possibility of a membrane-less LUCA has been discussed, 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 the LUCA had membrane-bound cells; however, the nature of the membrane in the LUCA remains uncertain. Phylogenetic analyses indicate that the LUCA encoded the biosynthetic pathways for both bacterial and archaeal phospholipids, implying an ancestral mixed membrane, with subsequent differentiation. Notably, preliminary data suggest that the major groups of viruses to the evolutionary trees of bacterial and archaeal hosts (omitting eukaryotes as a derived domain of life that emerged at a later stage of evolution and is hence irrelevant as far as the LUCA is concerned), thereby inferring their likely presence or absence in the LUCA virome. The results of this reconstruction 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 had already reached considerable diversity prior to the radiation of archaea and bacteria. 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. Duplodnaviruses are broadly distributed among both bacteria and archaea, indicating that the archaeal and bacterial viruses within Duplodnaviria, on a broad scale, have coevolved with their respective hosts (see discussion below). Tailed bacteriophages are nearly universal among bacteria. 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. Furthermore, HK97-fold MCPs were identified in uncultivated archaea of the proposed phyla Aenigmarchaeota, Altiarchaeota, Nanoarchaeota, Micarchaeota, Iainarchaeota and Asgardarchaeota. However, given the potential artefacts associated with the binning of viral modules62, which are not homologous to the bacterial functional counterparts, also argues in favour of the origin of this virus
Fig. 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 four 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 Variddaviria (families Tectiviridae, Corticoviridae, Finnlakeviridae, Sphaerolipoviridae and Autolykiviridae); the third ring shows members of the order Tubulavirales (families Inoviridae and Plectroviridae); and the fourth ring includes all other virus groups, namely Microviridae, Leviridae, Cystoviridae and Plasmaviridae. The phylogeny and taxonomic nomenclature were retrieved from the Genome Taxonomy Database\textsuperscript{51} and visualized with AnnoTree\textsuperscript{109}. The information on virus distribution for virus isolates with completely sequenced genomes was obtained from GenBank. The provirus distribution was retrieved from previously published work on Duplodnaviria\textsuperscript{110}, Tubulavirales\textsuperscript{68} and Varidnaviria\textsuperscript{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.
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 homologues 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 prokaryotes63,64. Actual host assignments await but some of these virus genomes were found in geothermal
hypothesis was tested by searching for 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.

While the vast majority of these viruses the hosts are unknown, the few known isolates infect broadly diverse bacteria from five different phyla (Fig. 2). It is likely that microviruses have a long-standing evolutionary history in bacteria, which probably dates back at least to the LUCA (Fig. 3).

In the extant biosphere, RNA viruses dominate the eukaryotic virome but are rare in bacteria (compared with DNA viruses) and unknown in archaea. Bacterial RNA viruses are represented by two families, the positive-sense RNA Leviridae and dsRNA Cystoviridae. The host range of 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 leviruses, indicating that their share in the prokaryotic virome had been substantially under-appreciated.

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. 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, implying the early origin of RdRP and RT and, as a consequence, the primordial status of RNA viruses. The origin of leviruses within bacteria is best compatible with their currently characterized distribution 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 leviruses through metagenomics, we consider that the 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 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 due 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 favourable to RNA virus reproduction.

Furthermore, unlike the LUCA, for which most evolutionary reconstructions suggest a mesophilic or a moderate thermophilic lifestyle, the last common ancestors of bacteria and archaea are inferred to have been thermophiles or hyperthermophiles. Extremely high temperatures might be restrictive for the 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). The family Cystoviridae, which includes dsRNA viruses, has an even narrower host range than the leviruses, suggesting a later origin. Thus, of the realm Riboviria, positive-sense RNA viruses are a putative component of the LUCA virome, whereas dsDNA 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 spindleshaped virions, infect hosts from all major archaean lineages and can be traced to the last archaean common ancestor. Therefore, their presence in the LUCA virome, with subsequent loss in the bacterial lineage, cannot be ruled out either.

**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. The 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 (Fig. 4). The DJR MCP of the Varidnaviria appears to be a unique virus feature, with no potential cellular ancestors detected. By contrast, the 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 world94. 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 MCP88–90. The recently discovered archaeal dsDNA viruses in the family Portogloboviridae91 contain one SJR MCP88 and thus appear to represent an even earlier evolutionary intermediate (Fig. 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 sphaerolipoviruses92. 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 proteins87. 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) family93) and appears to be specifically related to the dodecacin family of the SHS2-fold proteins94. Dodecaines are widespread proteins in bacteria and archaea that form dodecameric compartments involved in flavin sequestration and storage95 and are thus 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 dodecacin 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.

**Virus replication modules**

Each virus genome includes two major functional modules, one for virion formation (morphogenetic module) and one for genome replication96. The two modules rarely display congruent histories over long evolutionary spans and are instead exchanged horizontally between different groups of viruses through recombination, continuously producing new virus lineages.

**Key**

- Tubulivirales
- DJR MCP viruses
- Sphaerolipoviridae
- Microviridae
- Leviviridae
- Duplodnaviria
- Plasmaviridae
- Cystoviridae
- Bicaudaviridae
- Pleolipoviridae
- Globuloviridae
- Ovalliviridae
- MSV
- Guttaviridae
- Spindle-shaped viruses
- Portogloboviridae
- Spiraviridae
- Ampullaviridae
- ‘Tokiviricetes’
- Clavaviridae

---

**Fig. 3** | Reconstruction of the LUCA virome from the divergence of the bacterial and archaean viromes. This figure shows the hypothetical complex virome of the last universal cellular ancestor (LUA) as reconstructed from the distribution of viruses among extant phyla of bacteria and archaea. Also schematically depicted are the split of the LUA virome into the viromes of the last bacterial common ancestor (LBCA) and the last archaean common ancestor (LACA) as well as the subsequent diversification that resulted in the extant viromes. The divergence of bacteria and archaea from the LUA 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 LUA virome. DJR, double jelly-roll; MCP, major capsid proteins; MSV, Methanosarcina spherical virus.
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.

One of the most widespread replication modules in the virosphere is the rolling circle replication endonuclease (RCRE). 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 superfamily protein (Pro-Pd) (PDB ID: 3afg); P domain of a subtilisin-like protease (Pro-Pd) (PDB ID: 2hey); nucleoplasmin (PDB ID: 1nlq); P domain of a subtilisin-like protease (Pro-Pd) (PDB ID: 3afg); and DUF1884 family protein (PDB ID: 2pk8). The SJR and DJR MCPs and the corresponding virion symbols of members of the realms Varidnaviria and Duplodnaviria are coloured with different shades of blue. MCPs of duplodnaviruses are represented by the 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); and DJR MCP viruses are represented by the P2 protein of bacteriophage HK97 (PDB ID: 3qkb); and DUF1884 family protein (PDB ID: 2pk8). The SJR and DJR MCP viruses are represented by the 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); and DJR MCP viruses are represented by the P2 protein of bacteriophage HK97 (PDB ID: 3qkb).

In many cases, these DNA replication proteins do not have close cellular homologues, suggesting a 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 the replication of cellular chromosomes. Although some tailed bacterial dsDNA viruses encode replication factors of apparent bacterial origin, in archaeal duplodnaviruses, the proteins involved in informational processes, including components of the genome replication machinery, DNA repair and RNA metabolism, are of archaeal type, with none of their homologues found in contemporary bacterial viruses.
of the known archaeal viruses encoding components of the bacterial-type replication machinery\textsuperscript{19,20}. Finally, tailed archaeal viruses carry archael or eukaryotic-like promoters\textsuperscript{15,21}, consistent with the fact that none of the known archaeal viruses encode RNA polymerases\textsuperscript{5,16,20}, 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.

**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 Tubulaviridae, can be traced to the LBCA, whereas spindle-shaped viruses, most likely infected the last archael common ancestor. 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 extent 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 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 the 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\textsuperscript{10}.

Straightforward thinking on the LUCA virome might have envisaged it as a domain of RNA viruses descending from the primordial RNA world. However, the reconstructions 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 the LUCA lived, they had already been largely supplanted by the more efficient DNA virosphere.

*Marc Kruopovic\textsuperscript{11}*, Valerian V. Dolja\textsuperscript{12} and Eugene V. Koonin\textsuperscript{13}

\textsuperscript{11}Archaeal Virology Unit, Institut Pasteur, Paris, France.

\textsuperscript{12}Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA.

\textsuperscript{13}National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA.

\textsuperscript{14}Email: mail.kruopovic@pasteur.fr

https://doi.org/10.1038/s41579-020-0408-x

**Published online:** 14 July 2020


29. Mirkin, B. C., Fenn, T. L., 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


**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.

**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 declare no competing interests.

**Peer review information**

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

**Publisher’s note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Supplementary information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41579-020-0408-x.

© Springer Nature Limited 2020