Intervirology

Intervirology 2010;53:362–378
DOI: 10.1159/000312921

Published online: June 15, 2010

Giant Viruses: Conflicts in Revisiting the Virus Concept

Patrick Forterre[a, b]

[a]Unité de Biologie du Gène chez les Extrêmophiles, Institut Pasteur, Paris, and [b]Unité de Biologie du Gène chez les Extrêmophiles, Institut de Génétique et Microbiologie, Université Paris-Sud CNRS UMR 8621, Orsay, France

Key Words

Bacteriophage · Horizontal gene transfer · Mimivirus · Nucleo-cytoplasmic large DNA viruses · Tree of life · Virocell · Virus evolution · Virus origin

Abstract

The current paradigm on the nature of viruses is based on early work of the 'phage group' (the pro-phage concept) and molecular biologists working on tumour viruses (the proto-oncogene concept). It posits that viruses evolved from either prokaryotic or eukaryotic cellular genes that became infectious via their association with capsid genes. In this view, after their emergence viruses continued to evolve by stealing cellular genes (the escape model). This paradigm has been challenged recently by scientists who propose that viruses pre-dated modern cells. In particular, the discovery of Mimivirus has stimulated a lot of discussions on the nature of viruses. There are two major schools of thought, those who defend the escape model, suggesting that giant viruses are giant pickpockets (chimera), and those who emphasize their uniqueness and ancient origin. Comparative genomics of Mimivirus and related viruses (nucleo-cytoplasmic large DNA viruses) have produced a lot of data that have been interpreted according to the prejudices of the authors and thus failed until now to generate a consensus. I briefly review here the history of these debates and how they lead to new proposals, such as the definition of viruses as capsid-encoding organisms or else the recognition of their fundamentally cellular nature, the virocell concept.

Introduction

The current paradigm on the nature of viruses was established nearly 50 years ago at the onset of the molecular biology revolution [1–4]. This paradigm rests on two pillars: (1) the assimilation of the virus to the virion (the viral particle), and (2) the assumption that viral genomes originated from fragments of cellular chromosome that became autonomous and infectious (the escape model). The assimilation of viruses with their virions is very pervasive. Hence, whereas Jacob and Wollman correctly noticed in their seminal review 'Viruses and Genes' [5] that 'viruses may exist in three states: the extracellular infectious state, the vegetative state of autonomous replication and finally the proviral state', they describe in the same paper the virus as 'a genetic element enclosed in a protein coat'. The second pillar of the current paradigm on the nature of viruses, the escape model, was originally boosted by the discovery of lysogeny. The finding of viral genomes integrated in cellular genomes led to the 'prophage concept', suggesting that phages originated from pro-phages, identified with portions of cellular genomes. The same idea was later on endorsed by biologists studying eukaryotic viruses after the discovery of proto-oncogenes, supposed to be the precursors of oncogenic viruses (the protovirus hypothesis [4]). These views ended up favouring a model for the origin of viruses in which viruses evolved from either prokaryotic or eukaryotic cellular genes that become infectious through their association with capsid-producing genes. In this scenario, after

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0300–5526/10/0535–0362$26.00/0

Accessible online at:
www.karger.com/int

Patrick Forterre
Unité de Biologie du Gène chez les Extrêmophiles
Institut Pasteur, 25 rue du Docteur Roux
FR–75015 Paris (France)
Tel. +33 1 4568 8791, Fax +33 1 4568 8834, E-Mail forterre@pasteur.fr

their emergence, viruses continue to evolve by stealing cellular genes: the virus pickpocket paradigm [6, 7].

Ironically, the success of the escape model as the second pillar of the current paradigm can be partly explained by the fact that it solves a 'false' problem raised by the first pillar: how can viral genes originate if viruses are inert biological entities? Of course, new virus-specific genes cannot originated in virions but only during the replication of viral genomes during the intracellular stage of virus life. However, surprisingly, this possibility is often practically denied by many evolutionists who reason as if genes could only originate in a cellular chromosome. As a consequence of the current paradigm, viruses are thus often considered as secondary elements of the biosphere, mainly by-products of cellular activity (viruses being 'evolved' by cells, as recently claimed by Moreira and López-García [7]). The confusion between cells as compartments in which genomes (either cellular or viral) indeed evolve, and cells as organisms, members of a particular lineage (Bacteria, Archaea, Eukarya) finally led to the idea that all viral genes should have originated in the genome of an archaeon, a bacterium, or an eukaryote.

For many years the current paradigm remained unchallenged, except by a few authors. Its first pillar was strongly criticized in 1983 by Claudiu Bandea who wrote that 'the living phase of the virus is the intracellular phase of its life cycle' [8]. This allowed her to also challenge the escape model, supporting instead models in which viruses evolved by regression from cellular parasites [8]. My own interest in viruses was awoken at that time by the discovery in Bruce Alberts' laboratory of an unusual type II topoisomerase (Topo II) encoded by the bacteriovirus T4 [9]. In contradiction to the escape model, which predicts that bacteriophages have recruited their proteins from bacteria, it turned out that the T4 Topo II (a heterotrimer) is very different in terms of structure and activity from the bacterial Topo II, DNA gyrase (the only bacterial Topo II then known). I also discovered at that time in the literature that the virus Φ29 infecting *Bacillus subtilis* and the Adenovirus infecting humans encode a similar and unusual protein-primed DNA polymerase with no cellular counterpart [10]. This suggested that viruses might have predated the divergence of eukaryotes and prokaryotes.

The prokaryote/eukaryote paradigm itself was challenged in the 1980s by the work of Carl Woese who demonstrated the existence of three independent evolutionary cellular lineages (formerly urkingdoms, now called domains) from rRNA sequence comparisons [11]. The prokaryote/eukaryote dichotomy, built on structural observation, had to be replaced by a trinity of cellular organisms, Archaea (formerly Archaebacteria), Bacteria (formerly Eubacteria) and Eukarya, based on informational data. Inspired by Woese [12] and Bandea [8], I came to the conclusion that viruses are relics of lost domains, endorsing a new version of the regression theory for the origin of viruses. I suggested that viruses then originated from cellular lineages that have been defeated in the life struggle leading to the last universal common ancestor (LUCA) to all modern cells, but survived by infecting the descendants of LUCA (hitch-hiking with the winning cellular lineage) [13]. However, this proposal went unnoticed, since it was published in a book chapter, by someone with no connection with virologists. Furthermore, the focus of evolutionists between 1980 and 2000 on the rRNA molecules and, more generally, on the translation machinery excluded de facto viruses from historical scenarios for early life evolution.

This situation changed dramatically at the beginning of this century when rapid advances in comparative genomics and structural biology brought viruses once again to the attention of evolutionists. The accumulation of both cellular and viral genome sequences led to the discovery of an increasing number of virus-specific proteins, i.e. viral proteins with only very distantly related cellular homologues, or viral proteins without cellular homologues (except in integrated proviruses). The existence of such proteins was not predicted in the framework of the escape model, and the question of their origin become central to any theory on the origin and nature of viruses. Building on my scenario for the origin of viruses, I suggested that virus-specific proteins originated in ancient cellular lineages that pre-dated LUCA [14, 15], whereas Koonin et al. [16] suggested that proteins without cellular homologues but present in otherwise unrelated viruses (hallmark viral proteins) originated in an ancestral viral world that pre-dated the cellular world.

The recognition that viruses are very ancient opened the possibility of new scenarios for early life evolution in which viruses are major players. In particular, I have postulated that the diversity of virus-specific proteins involved in DNA metabolism testifies to a major role for viruses in the origin of modern DNA genomes [14, 16–18] and possibly in the origin of DNA itself [14]. Meanwhile, Takemura [19] and Bell [20] suggested independently that the eukaryotic nucleus could have originated from a large DNA virus (the viral eukaryogenesis hypothesis). Combining these various scenarios, I even suggested that three distinct viral DNA genomes might have been at the origin of the genomes of the three modern cellular domains [21].

These ideas were in general better received than previous attempts at challenging the current paradigm, because new data from structural analyses have independently started to pave the way for new thinking on the nature of viruses. Hence, unexpectedly, comparative studies made in collaboration between Bamford and Burnett's groups revealed in 1999 that the major capsid proteins of the bacteriophage PRD1 and of the human adenovirus are homologous, i.e. derived from a common ancestral protein [22]. This finding and subsequent work on other viruses led to the conclusion that viruses indeed pre-dated LUCA [23–26]. Several ancient viral lineages were defined by Bamford, who suggests focusing on the structure of capsid proteins to reconstruct the history of viruses, considering that capsids (or more generally virions) are the 'self' of viruses [23]. Finally, the discovery and characterization of unique archaeal viruses at the end of the last century also opens a new window on the virosphere [27]. Viruses from Archaea turned out to be strikingly different from those of Bacteria, adding another nail to the coffin of the prokaryote/eukaryote paradigm. Whereas the findings of Carl Woese and co-workers have shown that the cellular world can be divided into three domains, the work of Wolfram Zillig, David Prangishvili and a few other pioneers show that a distinct viral world corresponds to each cellular domain [for review, see 28]. The problems of the origin of the three cellular domains and the problem of the origin of the three viral worlds thus became intermingled.

At the beginning of the 21st century, the stage was thus set up to revisit the current paradigm on the nature and origin of viruses. The discovery in 2003 of Mimivirus [29, 30] thus occurred at the right time, in an atmosphere that was favourable for the emergence of new ideas. However, the current paradigm also found strong defenders of its cause [6, 7, 31]. In this paper, I will briefly review the history of the debates that have taken place after the discovery of Mimivirus. I will especially discuss the problem of the origin of viral genes and how different views led to different interpretations of the same phylogenies. I will try to show that some of the phylogenies that apparently support horizontal gene transfer (HGT) from cells to viruses can be interpreted the other way around. I will also briefly discuss how the discovery of Mimivirus led to new proposals on the nature of viruses, such as the notion of viruses as viral factories [32], their definition as capsid-encoding organisms [33], or else the idea that viruses are cellular organisms hidden before our eyes, the virocell concept [34].

Mimivirus, a New Branch on the Universal Tree?

The discovery of Mimivirus, a giant DNA virus infecting the amoeba *Entamoeba polyphaga*, and the sequencing of its genome were made by microbiologists and genomists with no strong connection to mainstream virologists [29, 30]. This possibly explains why the authors of this discovery immediately took a strong and challenging position in the debate on the nature of Mimivirus. In the paper describing the Mimivirus genome, they emphasized the presence of genes encoding proteins involved in translation (protein synthesis) that were never previously observed in viral genomes. In opposition to the escape model, they view their data as favouring the regression hypothesis in which viruses derived from intracellular parasitic cellular organisms [30]. In this hypothesis, the translation proteins of Mimivirus were considered as possible remnants of an ancestral complete translation apparatus inherited from its complex cellular ancestor. The discovery of Mimivirus also immediately gave more credit to proposals such as the viral eukaryogenesis hypothesis of Takemura [19] and Bell [20]. Indeed, the size of the Mimivirus genome is half the size of some eukaryotic genomes (2.9 Mb for *Encephalitozoon cuniculi*), while the viral factory of Mimivirus, which is as big as the nucleus of the infected amoeba (fig. 1), and can be seen under the light microscope [35].

Raoult et al. [30] were especially impressed by the presence in the Mimivirus genome of seven genes encoding universal proteins (i.e. proteins present in both Archaea, Bacteria and Eukarya), leading them to build a universal tree of life including Mimivirus. In this tree, Mimivirus branches between Archaea and Eukarya, as a domain on its own. Raoult et al. thus claimed that Mimivirus should be considered as a member of a fourth domain, a view that became widely publicized in the scientific literature and in the media. Unfortunately, the tree published by these authors was plagued with methodological problems, such as the limited number of representatives used for each domain. This opened the way for a counter-attack by proponents of the traditional view on viruses. Hence, Moreira and López-García published a tree for one of the seven universal proteins present in Mimivirus (tyrosyl tRNA synthetase, TyrRS) which, instead of supporting the fourth domain concept, seemed to validate the escape model. Indeed, the Mimivirus TyrRS branches in this tree within Eukarya, close to homologous TyrRS from two species of *Entamoeba* [6] (fig. 2). Moreira and López-García generalized this observation to all Mimivirus proteins and coined at this occasion the term 'giant

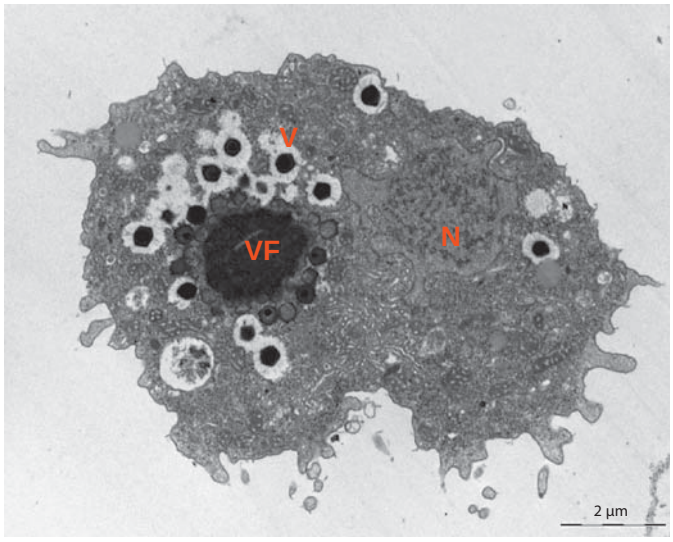

Fig. 1. Electron micrograph of the amoeba *Entamoeba haemolytica* infected by Mimivirus. N = Nucleus; V = virion; VF = viral factory.

pickpocket' to characterize 'giant viruses' [6]. In reply, Ogata et al. [36] argued that TyrRS is the exception that confirms the rule. They noticed that only 87 out of the 1,250 predicted Mimivirus proteins have homologues encoded in the genome of *Entamoeba histolytica,* a close relative of the Mimivirus host. They further reported that only a handful of these 87 proteins are more similar to their *E. histolytica* homologue than to homologues from eukaryotes of other divisions [36]. Finally, they noticed that the TyrRS tree published by Moreira and López-García is inconsistent with the accepted phylogeny of eukaryotes. In this tree, eukaryotic TyrRS (missing in Opistokonts) are divided in two subgroups, one corresponding to Plantae and the other including Mimivirus, Amoebozoa and various protests (fig. 1). In the latter group, Conosa, the clade of Amoebozoa that includes *Dictyostelium discoideum* and *E. histolytica* [37] are not monophyletic, indeed Amoebozoa and Mimivirus group with Giardia, whereas Dictyostelium groups with Plasmodium and Tetrahymena (fig. 2). Ogata et al. [36] thus suggested that the eukaryotic TyrRS might well have been obtained from several giant viruses and not the reverse. Indeed, as shown in figure 3, starting from a four-domains tree including Mimivirus, there are two possibilities to obtain a TyrRS-like three-domains topology, either by removing Mimivirus, if the Mimivirus protein originated from Eukarya, or by removing Eukarya, if eukaryotic proteins originated from relatives of Mimivirus.

More recently, replying to the review paper in which Moreira and López-García propose 'Ten reasons not to include viruses in the universal tree of life' [7], Claverie and Ogata [38] present a new 'four-domains tree', this time based on the universal DNA replication clamp loader protein RFC. They obtained once more a tree in which the Mimivirus protein (RFC) branches between Archaea and Eukarya. However, answering to this reply, López-García and Moreira noticed that Claverie and Ogata [39] have used in their analysis only one of three RFC paralogues that are present in the Mimivirus genomes and in most eukaryotic genomes. When they built a tree including these paralogues, they obtain a phylogeny with three subfamilies of eukaryotic RFC proteins, one for each paralogue. Furthermore, in the phylogenies of the three subfamilies, the Mimivirus RFC branch is close to RFC present in Amoeba, which is in apparent agreement with the pickpocket paradigm (fig. 2, adapted from 39). López-García and Moreira thus concluded once more that Mimivirus (or Mimivirus ancestors) had stolen the RFC genes from its host and that still 'viruses cannot be included in the tree of life' [39]. However, it should be noticed that two of the three Mimivirus RFC (RFC1 and RFC3) and their amoeba relatives branch at the base of their respective eukaryotic RFC subtrees, far from other Conosa *(Dictyostelium)* (fig. 2). This is again at odds with the position of Conosa in the classical eukaryotic phylogeny (illustrated by the third paralogue subtree in which Mimivirus RFC2 and Conoza form a monophyletic group within the eukaryotic tree, suggesting that RFC1 and RFC3 might have been originally transferred to eukaryotes from Mimivirus relatives, as previously discussed in the case of TyrRS, whereas RFC2 is an ancestral eukaryotic RFC which has been transferred from an amoeba to Mimivirus.

The choice of TyrRS or RFC to discuss the origin of Mimivirus raises another question: are these proteins valid phylogenetic markers for this task? This is not obvious, because these proteins are not widely distributed among the nucleo-cytoplasmic large DNA viruses (NCLDV), the viral superfamily that includes Mimivirus. The NCLDV are double-stranded DNA viruses with genomes sizes in the range from 150 kb to 1.2 Mb [40, 41]. They form a superfamily of viruses infecting various eukaryotic hosts (Poxviridae, Asfarviridae, Iridoviridae, Ascoviridae, Mimiviridae, Phycodnaviridae and Marseillevirus) [40, 41]. The TyrRS is only present in Mimivirus, whereas RFC is only present in Mimivirus and one Phycodnavirus (*Ectocarpus siliculosus* virus 1) (the latter always branching late in the eukaryotic tree, suggesting recent transfers from cells to viruses) [39]. This limits the usefulness of

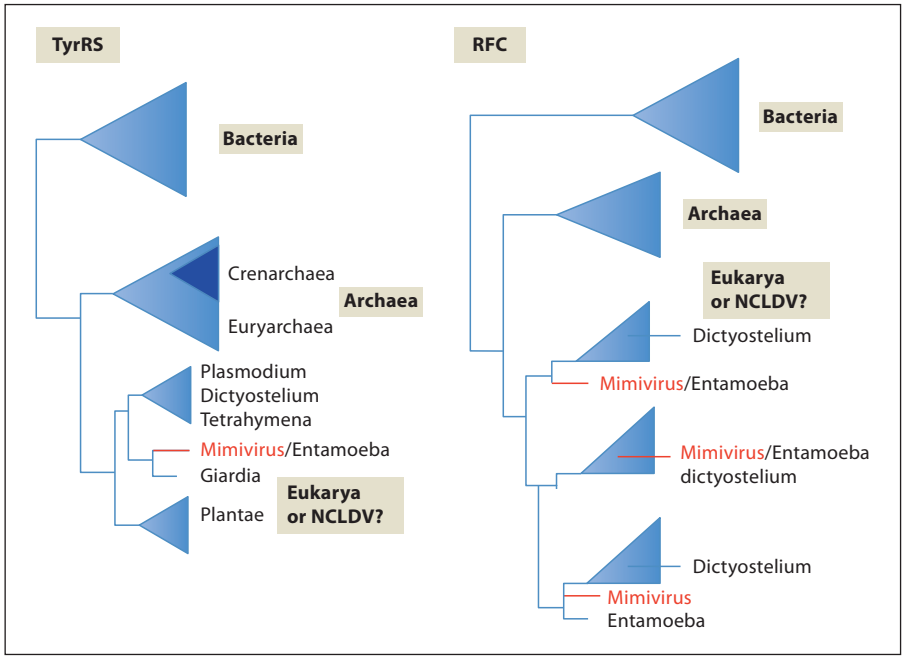

Fig. 2. Schematic representation of the phylogenies of Tyrosyl tRNA synthetases (TyrRS) and replication factor C (RFC). These phylogenies correspond to a three-domains topology. Considering the odd eukaryotic phylogeny of TyrRS and the basal positions of Mimivirus in the RFC tree, one cannot exclude that eukaryotic sequences could be in fine of viral origin. Figure reproduced from [6, 39].

these proteins to discuss the origin of Mimivirus because, in fine, the origin of Mimivirus should be traced to the origin of NCLDV. Indeed, around 50 proteins are found in the majority of NCLDV families (NCLDV hallmark proteins) and 5 in all of them (the NCLDV core genes), suggesting that this group of viruses is monophyletic [40, 41]. The 'fourth domain' concept thus implies that this domain corresponds to the entire NCLDV lineage.

Detailed analyses by Iyer et al. [40] have shown that the divergence of major NCLDV families probably occurred before the divergence of the major eukaryotic lineages. This means that the last common NCLDV ancestor lived before the last eukaryotic common ancestor. This conclusion justifies a priori drawing trees in which Mimivirus (or more correctly the NCLDV lineage) branch indeed between Archaea and Eukarya. If a protein is present in both NCLDV and in the three cellular domains, it is thus not surprising to obtain a four-domains tree (fig. 3). At that point, one should be careful in discussing the origin of such protein. The presence of a protein in the three cellular domains plus NCLDV seems to suggest a priori that this protein was initially a cellular protein present in LUCA that was later on transferred to NCLDV. However, since LUCA coexisted with many viral lineages, another possibility is that this protein first originated in a viral lineage, and was later on transfer to the three cellular domains on one or several occasions. This is precisely the type of scenario predicted by hypoth-

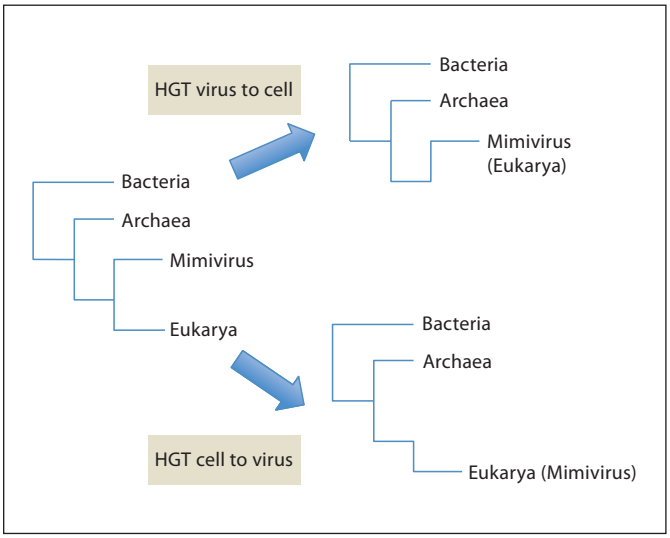

Fig. 3. Two ways to transform a four-domains phylogeny into a three-domains phylogeny. The four-domains tree with Mimivirus is supported by the separation of NCLDV (including Mimivirus) and Eukarya before the divergence of modern eukaryotes [40]. This tree can be transformed into a three-domains topology either if eukaryotic sequences have been recruited from a Mimivirus ancestor (HGT from viruses to cells) or if the Mimivirus sequence is of eukaryotic origin (HGT from cells to viruses).

eses in which DNA replication proteins, such as RFC, first originated in lineages of ancient DNA viruses [14, 15, 18]. Thus, existence of a four-domains tree for a given protein is compatible with either a cellular or viral origin for that protein.

A Chimera at the Node of a Network?

Interestingly, when you scan the literature on Mimivirus, it is clear that the preference of the authors in favour of a particular model for virus origin and evolution has influenced their experimental strategy. Hence, whereas Ogata, Claverie and co-workers have studied in depth Mimivirus genes with no cellular homologues and looked for argument minimizing the importance of HGT in the history of Mimivirus, other authors, such as Koonin, Filée, Chandler, Moreira and Brochier-Armanet, have focused instead on the fraction of Mimivirus proteins that have cellular homologues. Although these proteins probably only represent a small fraction of Mimivirus proteins (between 5 and 15%, see below), all these authors have concluded that Mimivirus is a chimera, because its genome is a mixture of genes from various cellular origins. Indeed, in agreement with the escape model, they have systematically assumed that all Mimivirus genes (with the possible exception of some core NCLDV genes) should have in fine a cellular origin. The notion of viruses as chimera has been further boosted by the discovery of Sputnik, a virus infecting Mimivirus, whose small genome (around 18 kb) encodes several proteins homologous to proteins of viruses with either archaeal, bacterial or eukaryotic hosts [42]. These proteins being respectively assimilated to archaeal, bacterial or eukaryotic proteins in the pickpocket paradigm, Sputnik itself becomes an apparent chimera. The notion of chimera has been highlighted again recently in the description of Marseillevirus, a new NCLDV infecting amoebae, as a 'chimeric microorganism' [43].

With the chimera metaphor, the debates on Mimivirus shifted from the discussion on the nature of viruses to the discussion on the nature and description of the evolutionary process. Focusing on HGT as the major factor in life evolution, some authors have criticized the 'universal tree of life' as 'the tree of 1%', arguing that the evolutionary process should be described as a network, not a tree [44]. Following this line of thought, Didier Raoult recently replied to the paper of Moreira and López-García refuting the possibility of placing viruses in the tree of life by stating that, 'there is no such thing as a tree of life (and of course viruses are out!)' [45], and suggested replacing the tree of life by a rhizome of life [46]. However, a variation introduced by HGT, endosymbiosis or viral integration is not different for natural selection from a variation introduced by mutation (fig. 3). In figure 3, the two organisms A and B that originated by speciation (e.g. two eukaryotic cells) have a clearly defined last common ancestor, even if the speciation of B involved a massive incorporation of genes from another organism C (a bacterium, an archaeon, or a virus). I recommend the recent paper by Gribaldo and Brochier [47] in which these authors clearly explain that most criticisms against the tree as a metaphor to describe the history of life originate from the confusion between the history of species and the history of genomes. Whereas the history of organisms can be in most cases adequately described by a tree, the history of genes is indeed best described in some cases as a network (although these cases might be not as frequent as currently thought). The study of giant viruses is therefore an interesting playground to test various concepts on the nature of viruses and/or the nature of the evolutionary process.

Comparative Genomics of Mimivirus Proteins

Using sensitive BLAST analyses, Aravind, Koonin and co-workers identified 75 and 198 proteins with 'bacterial affinity' and 'eukaryotic affinity', respectively, in the Mimivirus proteome [40]. Based on a similar type of analysis, Filée et al. identified 96 'bacterial-like' proteins and concluded that 78 of them are clearly of bacterial origin [48]. Being more stringent in their criteria (using phylogenetic analysis instead of simple BLAST searches), Moreira and Brochier-Armanet [31] ended up with lower numbers since, starting with 198 Mimivirus proteins originally ascribed to COG families, they ended up with only 126 Mimivirus proteins having clear-cut cellular homologues (97 and 105 having homologues in Bacteria or Eukarya, respectively, and 76 in both). They concluded from their analyses that 29 Mimivirus proteins with cellular homologues are clearly of bacterial origin and 60 of eukaryotic origin.

Strikingly, the authors of all these analyses only identified a very low number of genes of 'viral origin' in the Mimivirus genomes. In particular, Moreira and Brochier-Armanet [31] only identified 4 ORFs among the 198 mimiviral ORFs studied as having a viral origin. For these authors, this low number is a strong argument against the four-domains hypothesis, since the NCLDV lineage itself vanished in their analysis. However, this low number was surprising since the Mimivirus ge-

nome is known to encode at least 25 NCLDV hallmark proteins [49] and to share 23 unique genes with the genomes of *Phycodnaviridae* [40]. The low number of 'viral genes' detected by Moreira and Brochier-Armanet in Mimivirus can be probably explained by the fact that these authors used proteins present in the COG database as the starting point for their analysis. Indeed, this database does not include viral or plasmid genes, such as NCLDV-specific genes.

The underestimation of genes of viral origin in comparative genomic analysis is not surprising since, as previously mentioned, such genes are not supposed to exist in the current paradigm on the nature of viruses. Hence, in the recent analysis of the Marseillevirus genome, genes with homologues encoded by bacterioviruses are included within the category of bacterial genes [see fig. 2 in 43], assuming that genes present in the genomes of bacterioviruses originated from Bacteria. However, this is misleading, since most genes encoded by bacterioviruses have no bacterial homologues, except in pro-viruses integrated in bacterial genomes.

In all comparative genomic analyses performed on Mimivirus proteome, a striking observation is the low number of Mimivirus proteins with cellular homologues (between 116 and 273, depending of the criteria stringency used for their detection) compared to the number of putative Mimivirus proteins (911) [30, 31, 40, 48]. This dramatically emphasizes that the origin of more than 600 Mimivirus proteins remains a priori mysterious. The number of Mimivirus proteins without cellular homologues might be in fact even higher, since Raoult et al. [30] only considered ORFs encoding proteins of more than 100 amino acids in their analysis, although viral genomes always encode many small proteins. For instance, T4, whose genome has been very well annotated from genetic and biochemical studies, encodes 298 proteins for a genome of 169 kb, 103 of them comprising less than 100 amino acids and 16 less than 50 amino acids [for review, see 50]. By extrapolation, the genome of Mimivirus (1,180 kb) should encode many more than 911 proteins, possibly even more than the 1,262 ORFs detected by Raoult et al. [30] in the Mimivirus genome. I suspect that many ORFs removed from the list of protein-coding genes based on their size, anomalous nucleotide composition, overlapping with other genes or absence of viral homologues [30, 40] are in fact bona fide viral proteins.

Although the number of Mimivirus proteins with cellular homologues is already low, the number of the proteins that really testify for an HGT from cells to viruses might be even lower. Indeed, whereas in the current paradigm all HGTs have occurred from cells to viruses, in reality, many HGTs should have occurred the other way around. Of course, the relative importance of these two types of HGT probably varies depending of gene category. For instance, there is no doubt that many bacterial-like proteins of Mimivirus are indeed of bacterial origin (see below for possible exceptions). Some families of NCLDV that infect eukaryotes also grazing on bacteria, such as *Phycodnaviridae* and the Marseillevirus, also encode between 5 and 10% of bacterial-like genes [43, 48, 51]. These genes are not located randomly but form islands that tend to be localized toward the ends of linear genomes in Mimivirus and *Phycodnaviridae* [48]. These islands contain numerous mobile genetic elements normally found in Bacteria, as well as relatively large numbers of bacterial-like homing endonucleases and inteins. Interestingly, the genomes of T4 also contain up to 10% of bacterial-like genes [48, 51]. Filée et al. [48] thus suggested that foreign DNA can be introduced at the extremities of linear NCLDV genomes by a splice or patch process dependent on a replication mechanism similar to T4 recombination-primed DNA replication. Note that this explanation implies that the circular genomes of Marseillevirus originated from NCLDV with linear genomes.

Bacterial-like proteins present in Mimivirus teach us a lot about the interaction between bacteria and viruses in their eukaryotic hosts. However, they are clearly not appropriate to discuss the origin of NCLDV. Indeed, most of them have been only recently and independently acquired in Mimivirus and *Phycodnaviridae* [48, 51]. Eukaryotic-like proteins in Mimivirus are a priori more directly connected to the problem of NCLDV origin. In the pickpocket paradigm, these proteins are supposed to derive from proteins that were picked by ancestors of Mimivirus from their eukaryotic hosts. This process being a priori still ongoing, one would expect in that model to find in Mimivirus a large proportion of proteins recently recruited from *Entamoeba* or close relatives (Conosa). However, as initially observed by Ogata et al. [36], the number of Mimivirus proteins that are specifically related to Conosa is low. Hence, out of 60 Mimivirus eukaryotic-like proteins detected in their analysis, Moreira and Brochier-Armanet [31] identified only 13 proteins more closely related to Conosa than to homologues in other eukaryotic groups. To explain such observation, which clearly challenges the pickpocket theory, Moreira and López-García suggested that most Mimivirus proteins either originated from cellular genes which are not yet present in databases (when they have no cellular homologues) or testify to the fact that Mimivirus can infect hosts belonging to eukaryotic

lineages other than Conosa [31]. In fact, they systematically favoured an interpretation that fits with the virus pickpocket paradigm, although alternative possibilities exist. For instance, they noticed that 'certain mimiviral ORFs are exclusively shared by this virus and its amoebal host and they represent probable additional host-to-virus HGT events' [31]. However, in such situation, it is not possible to decide a priori if the HGT has occurred from host to virus or from virus to host.

I have argued elsewhere that many viral genes with cellular homologues testify in fact for HGT from viruses to cells because viral genes are much more abundant than cellular genes in the biosphere [53], making the probability of transfers higher from viruses to cells than the opposite [52]. Indeed, whereas genes with clear recent cellular origin are rare in viral genomes, genes with recent viral/plasmid origin are abundant in cellular genomes [54, 55]. It is therefore likely that many Mimivirus genes with eukaryotic homologues were originally transferred from ancestors of Mimivirus to ancestors of Conosa at various stages of eukaryote evolution. We will discuss later on how one can try to distinguish this category of proteins from those that have been indeed transferred from viruses to cells.

To sum up, (1) most Mimivirus proteins (around 80–90%) have no cellular homologues (a situation in fact typical for viruses), (2) Mimivirus proteins of bacterial origin have been recruited recently, and (3) the origin of those with eukaryotic homologues can be disputed. However, despite these observations, giant viruses have been systematically called 'chimera' by all authors who have focused their analyses on Mimivirus genes with cellular homologues. Moreira and Brochier-Armanet [31] introduced the term chimera in the title of their paper 'Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes', writing in their abstract, 'our phylogenetic trees strongly suggest that Mimivirus acquired most of these genes by HGT either from its amoebal hosts or from bacteria that parasitise the same hosts.' Filée et al. [56] concluded that their analyses 'further strengthen the notion that giant viruses are chimeras of genes from disparate sources.' They refer to Moreira and Brochier-Armanet to support this view, writing 'a phylogenomic study of Moreira and Brochier of the Mimivirus genome has confirmed our conclusion of massive acquisition of cellular genes from bacteria and from the host'. The term massive is surprising, referring to such a small fraction (5–10%) of the Mimivirus proteome. This fraction is lower than that of viral/plasmid genes which have been recently transferred into archaeal or bacterial genomes [54] and much lower than the amount of viral genes which can be detected in the human genome [55]. If we describe giant viruses as chimeric microbes, we should thus describe ourselves as chimeric macrobes. However, nobody has previously suggested defining eukaryotes as fundamentally giant pickpockets of retro-viral genes.

The chimera metaphor is, in my opinion, misleading for living organisms since there is a long-term continuity between organisms (from mother cells to daughter cells or from parents to offspring) in evolutionary lineage (fig. 4). It is not because the germline of your son has been infected by a new retrovirus that the continuity of your lineage has been broken! In its original sense, the term chimera refers to a monstrous organism born from the fusion of several 'normal' organisms (animals in the mythology). This is not the case for NCLDV, which are bona fide organisms with their own evolutionary history (see below) and did not originate from the fusion of Bacteria and Eukarya (despite the presence in their genomes of bacterial and eukaryal-like genes). In fine, the use of the chimera metaphor reminds us that for many biologists, Archaea, Bacteria and Eukarya are 'normal organisms' (because they are cellular), whereas viruses are not considered as 'organisms', but as monstrous.

You See What You Want

Practically, it is difficult to determine the direction of an ancient HGT between cells and viruses from phylogenetic analysis (and impossible from simple BLAST analysis). In only a few cases is the interpretation straightforward. For instance, there is little doubt that a viral protein has been borrowed from its eukaryotic host if homologous cellular proteins are widespread in many eukaryotic divisions, producing a reasonable eukaryotic phylogeny, and if the viral protein branches as a sister group or within the group containing related proteins from virus hosts (see the example of Mimivirus in fig. 5a). Similarly, it is most likely that a viral protein has a bacterial origin if its bacterial homologues are widespread in many bacterial phyla and if the viral protein branches within one of these phyla (fig. 6a). However, the majority of phylogenies obtained with Mimivirus proteins do not fit with these simple schemes, being more complex and difficult to interpret. This is typical for phylogenies grouping viral and cellular homologues. I previously observed such a situation for proteins involved in DNA metabolism [57–59]. In these cases, it is difficult to determine the direction of ancient HGT. In fact, since viruses or cells could have recruited their genes from cellular or viral lineages now

extinct, it is impossible to decide the direction of the transfer rigorously. The decision will finally rest on the investigator's subjective assumption that will be based on his/her prejudices. For instance, the impact of the pick-pocket paradigm is well illustrated by the interpretation that Filée et al. gave of their phylogenies of six core NCLDV genes [see fig. 3 in 56]. These authors wrote that their phylogenies showed that all HGTs 'are the result of transfer from the host' and concluded that: 'these six core genes are therefore acquired by polarized lateral gene transfer from the host'. However, when I looked at their phylogenies, I remarked that in five of them, some NCLDV proteins have a basal position in the eukaryotic tree (as schematically depicted in fig. 4b). This indicated that these five proteins might also have a viral origin, as previously discussed in the cases of RFC (see the schemes of fig. 3). In contrast, Filée et al. have reached their conclusion (all HGTs from cells to viruses) by focusing on 12 cases in which other NCLDV viruses group with their hosts. For me, these cases correspond to secondary transfers in which a protein initially of viral origin is re-introduced in viruses after an historical passage in cells.

To quantify (and qualify) the impact of opposite prejudices on the interpretation of Mimivirus phylogenies, I have reviewed all phylogenies published in the supplementary material of the Moreira and Brochier-Armanet paper [31], in order to compare their interpretation with mine. Most of these phylogenies are complex, with intermixing of cellular groups that are otherwise evolutionarily unrelated. In particular, Conosa often do not form a monophyletic group included in Amoebozoa, different groups of Conosa being dispersed between various eukaryotic divisions (as in fig. 5b). As a consequence, even if the Mimivirus protein is closely related to a protein of its host, it is not easy to determine if the latter is a bona fide Conosa protein (i.e. which originated and evolved in the eukaryotic domain) or if it has been introduced in Conosa by Mimivirus or by other NCLDV. Importantly, if a protein that originated first in a NCLDV lineage has been introduced in ancient proto-eukaryotes before the separation of the various modern eukaryotic divisions, the Mimivirus protein should be located at the base of the eukaryotic tree. This situation is frequently observed, as in the case of RFC. In many trees, Conosa proteins are grouped with the Mimivirus proteins at the base of the eukaryotic tree, far from other Amoebozoa, suggesting a recent secondary transfer from Mimivirus to Conosa, as in the case of the phylogenies of core proteins discussed in the paper of Filée et al. [56]. It is therefore likely in these cases that the Conosa proteins are of viral origin and not the reverse.

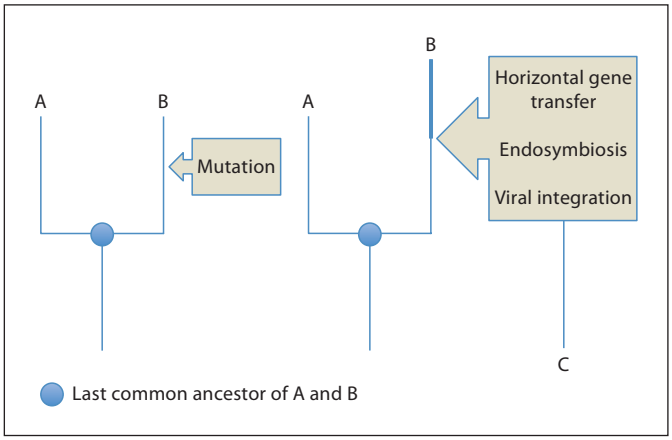

Fig. 4. The nature of variation does not modify the general pattern of descent with modification. In the left tree, divergence between A and B originated from a single point mutation in the lineage leading to B. In the right tree, divergence between A and B originated from the acquisition of new genetic material from the organism C (a cell or a virus) by lateral gene transfer, endosymbiosis or viral integration. This produces a more drastic phenotype modification and an acceleration of the rate of evolution in the lineage leading to B (symbolized by a long branch and bold line). However, in both cases, one can define a tree-like pattern and a similar common ancestor to A and B. The organism C does not belong to the same evolutionary lineage as A and B and cannot be considered as a bona fide ancestor of the organism B, even if it has contributed to its genetic makeup.

In several cases, the Mimivirus protein has homologues only in one eukaryotic group. This situation is observed in particular with animals and fungi alone or with Mimivirus at the base of opistokonts (fig. 5c). In the pick-pocket paradigm, this is explained by the transfer into Mimivirus of a protein that appears recently in animals, fungi or opistokonts. However, the transfer of an animal protein to Mimivirus appears highly unlikely since the ancestor of Mimivirus had probably never infected an animal cell before becoming a parasite of amoeba. It is more likely that an animal ancestor has received this gene from a NCLDV that shared it with Mimivirus.

In the case of trees supporting a bacterial origin for the Mimivirus proteins, one should distinguish again those in which the Mimivirus protein branches clearly within the bacterial tree (clear case of transfer; see fig. 6a) and those in which the Mimivirus protein is basal or nearly basal to the bacterial tree (sometimes itself with few representatives; see fig. 6b). In the latter case, it is unclear if the Mimivirus is really a 'bacterial protein' or if the root of the tree should be located between the Mimivirus protein and the bacterial domain. In the latter case, the bacterial-like pro-

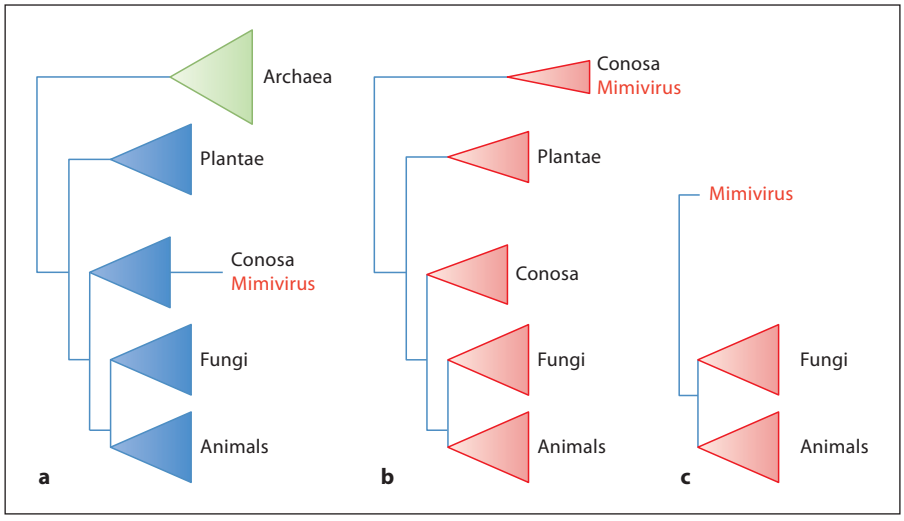

Fig. 5. Phylogenetic analysis and gene transfer between Mimivirus and eukaryotes. **a** Schematic representation of a theoretical phylogeny supporting an eukaryotic origin for a Mimivius protein. **b** Schematic representation of a theoretical phylogeny supporting a viral origin for Conosa proteins related to Mimivirus. **c** Schematic representation of a theoretical phylogeny suggesting a possible viral origin for a eukaryotic protein present in a limited number of eukaryotic divisions (in the illustrated case, opistokonts).

tein might be in fact a viral protein that has been introduced in Bacteria and not a bona fide bacterial protein.

Moreira and Brochier-Armanet concluded from their analyses that 29, 60, 1 and 4 Mimivirus proteins have bacterial, eukaryal, archaeal and viral origin, respectively [31]. In contrast, based on the above considerations, I found 32, 34 and 21 Mimivirus proteins of bacterial, eukaryal and viral origin, respectively. The main discrepancy is therefore the number of Mimivirus proteins of 'viral origin' detected in both analyses (21 vs. 4). I suppose that Mimivirus proteins with both eukaryotic and NCLDV homologues have been systematically dubbed 'proteins of eukaryotic origin' by Moreira and Brochier-Armanet. In contrast, I have considered that many of them were ancestral NCLDV proteins (pre-dating the divergence of eukaryotic divisions) or proteins that originated in various NCLDV lineages and were introduced later on in eukaryotes.

It is interesting to consider how Moreira and López-García used the conclusion of the Moreira and Brochier-Armanet analyses to support their paper 'Ten reasons to exclude viruses from the tree of life' [7]. In figure 2 of that paper, they symbolize the linear genome of Mimivirus by a ring only figuring genes with cellular homologues. The ring is superimposed on a virion, as expected from the classical view of viruses. This gives the impression that the Mimivirus genome is indeed a chimera of genes of cellular origin (mainly eukaryotic genes). I have translated the figure of Moreira and López-García into a column (linear genome) and compared their data with those obtained in my analysis (fig. 7) and superimposed my column onto a Mimivirus viral factory (and a virocell, see below). Importantly, I considered all Mimivirus genes

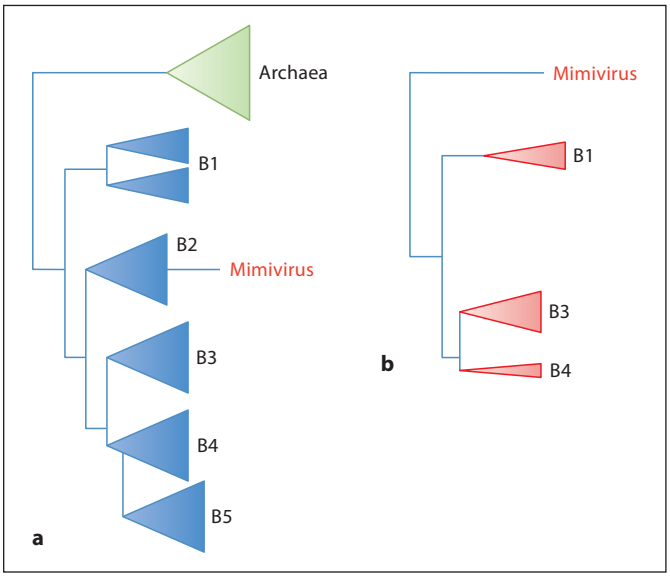

Fig. 6. Phylogenetic analyses and gene transfer between Mimivirus and Bacteria. **a** Schematic representation of a theoretical phylogeny supporting a bacterial origin for a Mimivius protein. **b** Schematic representation of a theoretical phylogeny supporting a possible viral origin for bacterial proteins.

(with or without cellular homologues). The difference between the two columns is striking. It illustrates how two different views of 'what is a virus' produce two different interpretations (and illustrations) of the same data. In one view, the Mimivirus genome appears indeed mainly as a chimera of genes of cellular origin, whereas in the other, it appears as an engine to produce new genes, surround-

ing a core of ancient viral genes and a sizable but limited portion of stolen cellular genes.

Many new giant virus genomes will be probably available in the near future. One can hope that their analyses will be performed taking into consideration that genes can have a viral origin and not only a cellular one. The most important point will be to focus on NCLDV hallmark genes that can really teach us more than the others on the origin of NCLDV. In fact, despite their adhesion to the chimera metaphor, Koonin, Filée and co-workers have already published important analyses on the NCLDV core proteins [40, 56, 60] and proposed various hypotheses on the origin of this lineage and subsequent evolution of the various NCLDV families. I will briefly review them now as a starting point for discussions on the nature of the NCLDV ancestor, the mode of evolution of NCLDV and, in fine, their possible origin.

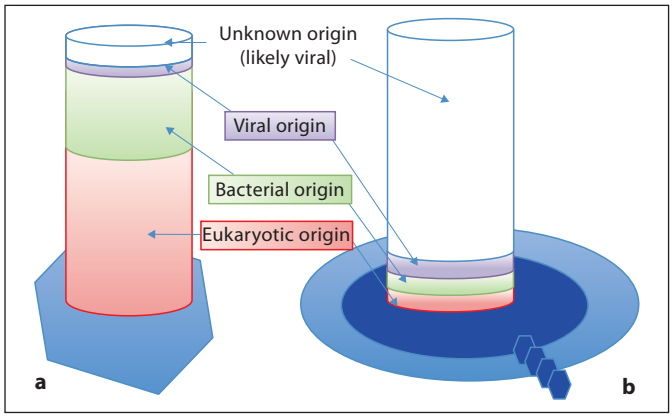

Fig. 7. Two visions of the same virus. **a** Column representing the percentage of Mimivirus genes with cellular homologues coloured according to their putative origin (adapted from 7). **b** Column representing the percentage of all Mimivirus genes coloured according to their putative origin according to the analysis performed for this work. In **a**, Mimivirus is represented by its virion, as in [7]; in **b**, Mimivirus is represented by the infected cell, according to the virocell concept [34].

The NCLDV Ancestor

The NCLDV core proteins can be used to reconstruct the history of NCLDV, much like core proteins conserved in all Archaea (ribosomal proteins, or else RNA polymerase subunits) have been used to reconstruct the history of Archaea [61]. In the most recently published trees, the Mimivirus lineage forms a clade with Phycodnaviridae, whereas the Marseillevirus forms a clade with Iridoviridae and Ascoviridae [41, 43]. An important task to determine the origin of NCLDV is thus to reconstruct their last common ancestor. This is not easy since the ancestral set of NCLDV proteins has been certainly affected by gene loss and non-orthologous or orthologous displacements [41, 56]. For instance, the complex RNA polymerase encoded by some NCLDV viruses was most likely present in the NCLDV ancestor and lost several times in various lineages [40]. Furthermore, we do not know where the root of the NCLDV tree is located. Accordingly, it is not possible to identify the ancestral NCLDV proteins based on a rigorous cladistic analysis. One can only infer a minimal set of NCLDV ancestral proteins using parsimony principles. Using this strategy, this minimal set has been first estimated to be 28 [62], then to be 41 [40] and more recently to be 47 [41], increasing with the number of NCLDV sequences available. Assuming that these proteins were possibly the only ones present in the NCLDV ancestor, Filée et al. [56] suggested that NCLDV 'have evolved by significant growth of a simple DNA virus by gene acquisition from cellular sources'. This view is again based on the idea that most viral genes should have in fine a cellular origin. To support their hypothesis, these authors noticed that the number of genes acquired by HGT in NCLDV correlates with their genome size (bigger genome, more acquired genes) [56]. However, there is also a clear correlation between genome size and the number of genes acquired by HGT in Bacteria or Archaea. This does not prevent most evolutionists from believing that the genome size of the ancestral archaeon or bacterium was not very different from those of their modern descendants (again being treated differently from viruses). Similarly, most evolutionists usually agree that the genome of the LUCA should have encoded many more proteins than those present in the set of universal proteins. For instance, Koonin [63], who identified about 60 universal proteins, has estimated that LUCA might have encoded around 500–600 proteins. Accordingly, with a minimal set of 40–50 proteins, the NCLDV ancestor might have encoded around 400–500 proteins, much in the range of modern NCLDV. Indeed, in their more recent analysis, Yutin et al. [41] conclude that, 'the common ancestor of an extant NCLDV even might have been a giant virus' (fig. 8).

Major Mechanism of NCLDV Evolution

It is unlikely that NCLDV evolved mainly by 'gene acquisition from cellular sources', as suggested by the pick-pocket paradigm, since the genomes of modern NCLDV

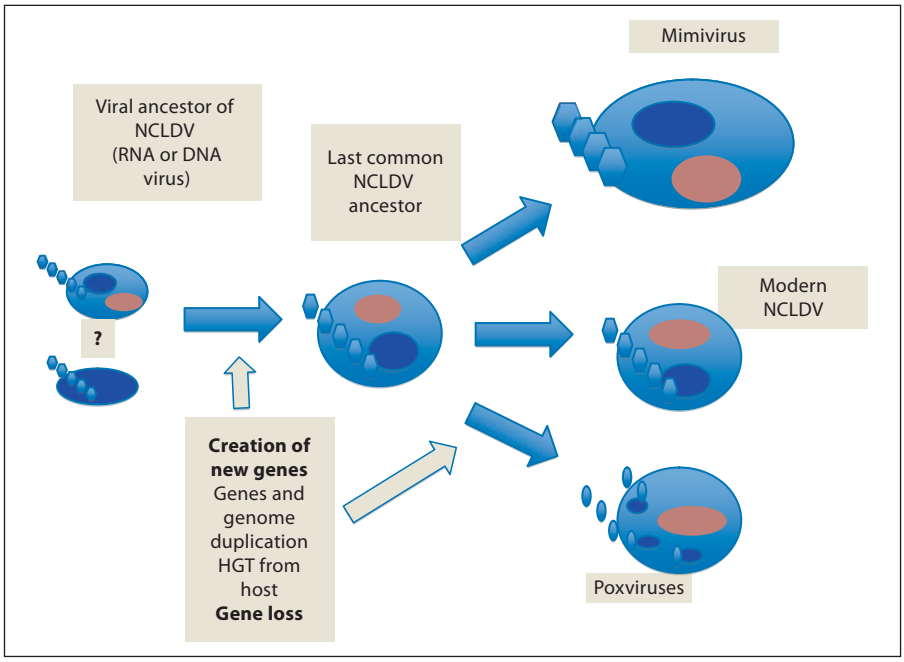

Fig. 8. Schematic representation of NCLDV evolution in the viral origin scenario [18, 56]. NCLDV are represented by infected cells, according to the virocell concept [34]. The viral factories are represented by dark blue circles and the decaying nucleus of the infected cell by pink circles surrounded by dotted lines.

contain relatively few genes testifying to acquisition from their hosts. NCLDV thus should have evolved by more complex and diverse processes (fig. 8). The especially large size of Mimivirus seems to be the result of multiple specific lineage duplications as well as genome duplications [64]. Similar mechanisms might have operated to a lesser extent in other NCLDV [56]. However, in my opinion, the major mechanism of NCLDV evolution has been the creation of new genes. Indeed, all NCLDV harbour numerous ORFs encoding putative proteins with neither cellular nor viral homologues, except sometimes in close members of the same NCLDV family (viral lineage-specific genes). Analysis of lineage-specific NCLDV proteins by Ogata and Claverie [65] has shown that they exhibit the same position-dependent nucleotide statistics as the rest of the genome, suggesting that most of them are truly viral genes and do not correspond to recent HGT. These proteins have probably originated at different epochs in NCLDV lineages. The creation of new viral genes is indeed continuously driven by the selection of novel functions to counteract the defences of the host and to manipulate its biology. This led in particular to the production of small proteins that interact with cellular proteins to modify their function, explaining the plethora of small ORFs and genes in viral genomes [66]. Family-specific NCLDV proteins are indeed shorter on average than NCLDV proteins with cellular homologues. Interestingly, Ogata and Claverie [65] noticed that NCLDV family-specific proteins include a larger fraction of predicted transmembrane proteins compared to NCLDV proteins with cellular homologues. This resembles the case of large bacterioviruses like T4 that also encode many proteins that integrate into the membrane of the infected cell [49]. In the case of NCLDV, these proteins should be also essential for manipulation of the intracellular membrane system of eukaryotic cells. Finally, Ogata and Claverie [65] noticed that NCLDV family-specific proteins are rich in low-complexity sequences, resembling eukaryotic proteins involved in nucleic acid interactions and in the architecture of the cytoskeleton. Low sequence complexity is also a characteristic of new genes produced by overprinting of ancient genes, a frequent and well-studied phenomenon in RNA viruses [67]. They can also be more frequent in viruses because mistakes such as replication slippage or abnormal recombination may occur more frequently in viral genome replication. If low complexity is indeed a frequent property of viral proteins, it is tempting to speculate that many eukaryotic proteins involved in interactions with nucleic acids or in cytoskeleton architecture that exhibit this property first originated in viral lineages.

Ogata and Claverie [65] concluded from their study that 'a large viral genome may act as an invention factory for new genes'. This is true indeed for any kind of viral genome, RNA or DNA, small or big. Of course, larger genomes are also larger reservoirs for the creation of new

genes. In that context, it is obvious that, as previously suggested here, many of the new genes created in NCLDV lineages should have been regularly transferred (and sometimes fixed) into the genomes of eukaryotic cells following NCLDV infection.

Origin of the NCLDV Lineage

Several hypotheses have been proposed to explain the origin of NCLDV. All these hypotheses are strongly dependent on the favourite scenario of their authors for early life evolution. In the framework of regression hypotheses, Claverie has suggested that NCLDV originated from cells of a proto-eukaryotic lineage [32]. Elaborating on the viral eukaryogenesis hypothesis, Claverie proposed that transition between ancient giant viruses and nuclei of proto-eukaryotic cells occurred several times in both directions during proto-eukaryotic evolution, opening the possibility that NCLDV originated from a primitive nucleus (the nuclear viriogenesis hypothesis) instead of the reverse (viral eukaryogenesis). In that hypothesis, the chromosome of a proto-eukaryotic cell has captured the genes encoding the virion of a small DNA virus to become the NCLDV ancestor (after the loss of its ribosome-encoding genes). The capsid protein of NCLDV contains the double jelly-roll fold that defines a lineage of double-stranded DNA viruses that spans the three cellular domains of life [26]. Interestingly, the major capsid protein of Sputnik also seems to harbour a double jelly-roll fold [68]. In the nuclear virogenesis hypothesis, it is therefore tempting to suggest that NCLDV recruited their capsid protein from an ancestor of Sputnik. However, a weak point of this hypothesis is that the capsid protein of a small virus should abruptly become able to form a giant structure in order to enclose a now large chromosome.

The alternative to the cellular regression hypothesis is that the NCLDV ancestors were always viral. Koonin and co-workers suggested that NCLDV originated from the fusion of both large and small bacterial and archaeal DNA viruses that infected the bacterium and the archaeon that mixed (by endosymbiosis, fusion or any other mechanism) to produce the first eukaryotic cell [40; see fig. 8]. Scenarios mixing Archaea and Bacteria for the origin of eukaryotes are presently very popular. However, in my opinion, they are all very unlikely, because they suppose a dramatic acceleration of the evolutionary rate of the mixed system (creation of hundreds of new protein domains and complex molecular mechanisms, acceleration of the rate of protein evolution) to an extant which has never been observed in actual mixed systems, such as the endosymbiosis of cyanobacteria and eukaryotes to give Plantae (a eukaryotic division, not a new domain) [for critical discussions of these scenarios, see 69, 70]. In the case of the origin of eukaryotic viruses, mixed scenarios imply a rapid and complete transformation of a small bacterial or a small archaeal virus of the double jelly-roll lineage into a giant NCLDV, with complete modification of their replication proteins. In fact, the dramatic differences between viruses infecting the three domains of life (with few exceptions) and, in particular, the uniqueness of most eukaryotic viruses, appear to me another strong argument against mixed scenarios for the origin of eukaryotes.

To explain why viruses infecting different domains are so different both in terms of virion morphotypes and genome contents, it has been suggested that the three cellular lineages at the origin of modern domains selected independently three distinct parts of the ancestral virosphere [28]. In that hypothesis, NCLDV should have originated from large DNA viruses that were already part of this ancestral virosphere and infected cells from the lineage that gives rise to proto-eukaryotes. Several lineages of large DNA viruses (including the NCLDV ancestor) probably emerged from smaller DNA viruses, themselves descendants of RNA viruses, during the late RNA world or during the early DNA world, i.e. after the transition from RNA cells to DNA cells (fig. 8) [18, 21, 56]. They infected either large late RNA cells or large early DNA cells, probably related to the proto-eukaryotic lineage.

An important point would be to determine if the primordial virus ancestors of NCLDV produced viral factories (upper drawing in fig. 8) as many modern eukaryotic viruses, or if they transformed the primitive cells into viral factory (lower drawing in fig. 7) as modern bacterioviruses and archaeoviruses do (see below). The traditional view is that prokaryotes preceded proto-eukaryotes, but this is only based on a strong prejudice. Endomembrane formation, a characteristic feature of eukaryotes, has now been detected in some bacteria [71], and one cannot exclude the possibility that LUCA and its ancestors already had internal membrane systems that have been lost later on in all Archaea and in most Bacteria.

There are many controversial hypotheses on the nature of LUCA and the topology of the universal tree of life. Usually, viruses are forgotten from these scenarios. A few years ago, I suggested that viruses might have 'invented' DNA and that DNA was later on transferred from viruses to cells [14, 15]. In a specific version of this hypothesis – 'the three RNA cells, three DNA viruses'

hypothesis – this virus-induced RNA to DNA transition in cellular genomes occurred three times independently, at the onset of each domain [21]. Three founder viruses were then at the origin of DNA and DNA replication mechanisms in modern Archaea, Bacteria and Eukarya, respectively. An ancient NCLDV is a priori a good candidate to be the founder virus for eukaryotes because of its large size, linear genome and the ability of its modern descendants to manipulate intracellular membranes. However, since most NCLDV proteins involved in DNA and RNA metabolisms (RNA and DNA polymerases) are only distantly related to those involved in eukaryotic DNA replication whereas others are clearly distinct (such as the DNA primase), it is likely that the founder virus of the eukaryotic domain (if it existed) was not a bona fide NCLDV but a member of another lineage of large DNA viruses, only distantly related to NCLDV. In my present opinion, the most likely scenario is that several lineages of large DNA viruses (plasmids) which co-evolved with proto-eukaryotic lineages contributed in fact to the formation of modern eukaryotes. This could explain the presence of multiple DNA polymerases and RNA polymerases in eukaryotes. Instead of being paralogues, these proteins might have been introduced in eukaryotic cells by different DNA viruses from different extinct lineages.

Mimivirus and the Nature of Viruses

The current ideas of the scientific community on the nature of viruses have been deeply influenced by extensive studies of bacterioviruses by early molecular biologists. Although one of the earliest and most studied viruses, T4, has a quite large genome, later work has mainly focused on viruses with smaller genomes, such as lambda (48.5 kb), M13 (6.4 kb), Qβ (4.1 kb) or SV40 (5 kb). This imprints a vision of viruses as simple biological entities that were more akin to molecular machines than to living organisms. The discovery of Mimivirus shakes these views by breaking the frontier between viruses and cells. The genome of Mimivirus turned out to be very similar in size, gene density and gene contents to those of small bacteria or archaea. The only characteristic that distinguishes the Mimivirus genome from a cellular genome is the absence of genes encoding ribosomal proteins. Didier Raoult and myself thus proposed to define viruses as 'capsid-encoding organisms', versus cells being defined as 'ribosome-encoding organisms' [33]. This raised the risk of strengthening the confusion between virus and virions if one focuses on the first word of the definition – capsid – instead of the third – organism. Fortunately, the discovery of the giant viral factory of Mimivirus also helps to realize that the virion is not the virus, as previously stressed by Bandea [8]. This was clearly stated by Claverie [32], who suggests considering that the real viral organism is the viral factory. The viral factory indeed looks like an intracellular parasite with defined borders. The concept of viruses as viral factories can be easily extended to all NCLDV and other viruses (including RNA viruses) that replicate in the cytoplasm of their eukaryotic host. These viruses indeed manipulate the intracellular membrane network to produce structures similar to Mimivirus viral factories [72, 73]. However, the identification of a viral factory is not so easy in the case of viruses whose genomes replicate in the nucleus of eukaryotic cells. Furthermore, the viral infections of Archaea and Bacteria do not produce visible viral factories. In one of the best-documented cases, T4 infection, small viral-like factories (black rounded structures) that can been seen in thin sections by electron microscopy correspond to aggregates of capsid proteins that are stocked as precursors of the T4 head, barely living organisms [74].

Interestingly, Lwoff [1] wrote in 1957 that: 'viruses transform the cell into a viral factory'. Taking into account that infected cells produce virions and not viruses, this sentence can be translated as 'viruses transform the cell into a virion factory', i.e. a virus (if we identify viruses and virion factories). Following this logic, one can consider that the living form of the virus corresponds to the infected cell itself [75]. In that case, viruses can be described as cellular organisms (thus definitely alive; see fig. 7 and 8 for illustration of this concept). I defined previously organisms as an integrated collection of organs – molecular and/or cellular – evolving by natural selection [76], and Bandea [77] has suggested defining viruses as 'molecular organisms'. However, I think now that all living organisms should be also cellular by definition because the integration of organs (either molecular or cellular) can only occur in a cellular context (within a cell in a unicellular organism or between cells in a multicellular organism). In this view, viruses are indeed cellular organisms that comprise a collection of molecular organs (virion, replicon, virion factories) whose integration occurs in the infected cell. I recently proposed the terms 'virocell' to characterize the cellular form of the virus, and ribocells to designate cells of ribosome-encoding organisms [34]. In my opinion, the virocell concept, by focusing the attention on the cellular state of the virus life

cycle, helps to understand how viruses can create new genes, i.e. virus-specific proteins. These genes appear in the course of viral genome replication (in the cellular setting) much like new cellular genes originate during the replication of cellular genomes (by point mutation, insertion or deletion, extensive recombination, slippage or frameshift). Note that in the case of an amoeba infected by a Mamavirus and Sputnik, we have now a virocell with two viral organisms. These examples help us to understand that the virocell (or ribocell) concept should not be confused with the concept of organism (fig. 9).

The concept of virocell raises interesting issues. For instance, if the infected amoeba is the virocell of Mimivirus, it suggests that the viral factory of Mimivirus is the nucleus of the virocell. In that framework, the evolutionary transition from a large DNA virus to the nucleus of a proto-eukaryotic cell (or vice versa) becomes the transition between the nucleus of a ribocell and the nucleus of a virocell. The viral eukaryogenesis or nuclear virogenesis hypotheses are much easier to defend in this context. These hypotheses were usually rejected by scientists who, still confusing viruses and virions, used to ask: how a virion can become a cell nucleus? It becomes much more credible if the virus is assimilated to the virocell and the viral factory to its nucleus.

The lysogenic state also asks for second thought. In a lysogenic situation, we have at least two organisms, a ribosome-encoding organism and a capsid-encoding organism, that cohabit in the same cell, a ribovirocell [34]. A ribovirocell includes in the same cellular structure organisms from different evolutionary lineages, defined by genetic continuity. In the case of an amoeba infected by a bacterium and a Mimivirus, itself infected by a Sputnik, we have now four organisms that coexist temporarily in the same cell (fig. 9). This cell is originally provided by the amoeba that can be rightly described as a 'melting pot' of organisms from various sources, and the complex ribovirocell that emerges from this association can be probably dubbed a chimera in that case. In the infected amoeba, the four organisms present can exchange genes extensively. However, each of them conserves its integrity (genetic continuity) and tree-like evolution pattern, i.e. the amoeba (if it survives) remains member of a particular eukaryotic division, and Mimivirus remains an NCLDV. An organism is an historical product and integrates the memory of all previous interactions of its lineages with others that had co-existed with it during life evolution, but it does not lose its identity through this process.

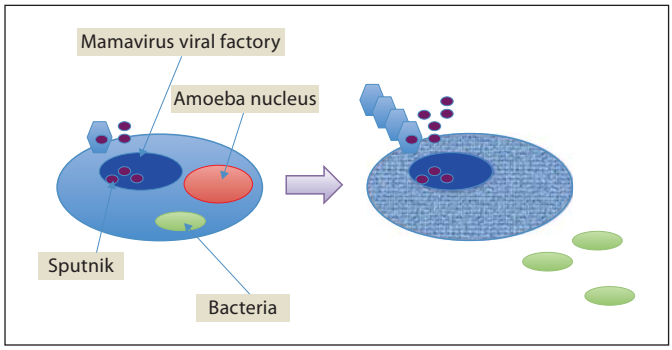

Fig. 9. Left panel: an amoeba infected with a bacterium is attacked by a Mamavirus associated to a Sputnik virus, four organisms coexist in the same cell. Right panel: the amoeba nucleus has been transformed into a virocell formed by two viral organisms, Mamavirus viral factory and Sputnik.

Viruses and/in the Tree of Life

The tree of life can be seen as a metaphor to represent the history of life. In that definition, all living organisms should be present in this tree, including viruses. However, in that case, the tree of life cannot be a tree in a strict sense of the term because viral evolution only partly occurs in a tree-like fashion. Whereas tree-like evolution seems to be the rule for large viruses, such as T4-like bacterioviruses or NCLDV [41, 78], many events of fusion have probably occurred between different lineages of viruses with small genomes, or between such viruses and plasmids [79]. The tree of capsid-encoding genes therefore does not coincide with the tree of replicons. Recombination between viral and/or plasmid genomes can lead to the formation of completely new lineages, as exemplified by the origin of geminiviruses from the fusion of a single-stranded DNA virus and a bacterial plasmid [80]. As a consequence, a new organism is created, starting a new evolutionary lineage. Furthermore, although plasmids and viruses mainly co-evolve with their hosts, as recently shown in the case of plasmids from Euryarchaea [81], they can sometimes jump from one cellular lineage to another, as in the case of retroviruses shifting between different vertebrate hosts. One is thus faced with a dilemma, either we consider the tree of life as a metaphor to represent the history of life, which cannot be viewed as a classical tree, or we adhere to a strict phylogenetic definition of the tree of life, which (as it excludes viruses) is not universal. A possibility would be to distinguish between a universal network of life (including both viruses and cells) and a universal tree of life (including only ribocells). However, the network metaphor would be again

partly misleading since virus/plasmid mainly co-evolve with their cellular hosts and jumping of viruses between lineages is usually restricted to closely related lineages.

As recently quoted by Bandea, 'the trunk and many of the branches of tree of life are embedded in a viral shell' [82]. I basically agree with this view and suggest maintaining the tree of life metaphor as a useful descriptor of life history, keeping in mind that this tree encompasses various evolutionary processes, depending of the type of organism. Schematically, the tree of life is a combination of two components, a tree-like component corresponding to the history of modern ribocells, and a co-evolving shell (the word of plasmids and virocells). In the case of large and complex eukaryotic DNA viruses, such as NCLDV, the tree of ribocells and virocells can be probably superimposed to a great extent. An exciting research program would be to screen extensively all eukaryotic divisions to discover new NCLDV families and even completely new families of large DNA viruses. This would help us to reconstruct with more confidence the history of eukaryotes, of their viruses and of the creative co-evolution of these different but co-dependent organisms.

Acknowledgments

I am grateful to David Prangishvili, Simonetta Gribaldo and Mart Krupovic for critical reading of the manuscript.

References

1 Lwoff A: The concept of virus. J Gen Microbiol 1957;17:239–253.
2 Lwoff A: Principles of classification and nomenclature of viruses. Nature 1967;215:13–14.
3 Lwoff A: Interaction among virus, cell, and organism. Science 1966;152:1216–1220.
4 Temin HM: The DNA provirus hypothesis. Science 1976;192:1075–1080.
5 Jacob F, Wollman E: Viruses and genes. Sci Am 1961;204:93–107.
6 Moreira D, López-García P: Comment on 'The 1.2-megabase genome sequence of Mimivirus'. Science 2005;308:1114.
7 Moreira D, López-García P: Ten reasons to exclude viruses from the tree of life. Nat Rev Microbiol 2009;7:306–311.
8 Bandea C: A new theory on the origin and the nature of viruses. J Theor Biol 1983;105:591–602.
9 Liu LF, Liu CC, Alberts BM: T4 DNA topoisomerase: a new ATP-dependent enzyme essential for initiation of T4 bacteriophage DNA replication. Nature 1979;281:456–461.
10 Bernad A, Zaballos A, Salas M, Blanco L: Structural and functional relationships between prokaryotic and eukaryotic DNA polymerases. EMBO J 1987;6:4219–4225.
11 Woese CR, Fox GE: Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 1977;74:5088–5090.
12 Woese CR: Archaebacteria. Sci Am 1981;244:98.
13 Forterre P: New hypotheses on the origin of prokaryotes, eukaryotes and viruses; in Trân Thanh Vân JK, Mounolou JC, Schneider J, McKay C (eds): Frontiers of Life. Gif sur Yvette, Editions Frontières, 1991, pp 221–233.
14 Forterre P: The origin of DNA genomes and DNA replication. Curr Opin Microbiol 2002;5:525–532.
15 Forterre P: The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. Biochimie 2005;87:793–803.
16 Koonin EV, Senkevich TG, Dolja VV: The ancient virus world and evolution of cells. Biol Direct 2006;9:1–29.
17 Forterre P: Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. Mol Microbiol 1999;33:457–465.
18 Forterre P: The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 2006;117:5–16.
19 Takemura M: Poxviruses and the origin of the eukaryotic nucleus. J Mol Evol 2001;52:419–425.
20 Bell PJ: Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus? J Mol Evol 2001;53:251–256.
21 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 2006;103:3669–3674.
22 Benson SD, Bamford JK, Bamford DH, Burnett RM: Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. Cell 1999;98:825–833.
23 Bamford DH: Do viruses form lineages across different domains of life? Res Microbiol 2003;154:231–236.
24 Baker ML, Jiang W, Rixon FJ, Chiu W: Common ancestry of herpes viruses and tailed DNA bacteriophages. J Virol 2005;79:14967–14970.
25 Bamford DH, Grimes JM, Stuart DI: What does structure tell us about virus evolution? Curr Opin Struct Biol 2006;15:655–663.
26 Krupovic M, Bamford DH: Virus evolution: how far does the double beta-barrel viral lineage extend? Nat Rev Microbiol 2008;6:941–948.
27 Martin A, Yeats S, Janekovic D, Reiter WD, Aicher W, Zillig W: SAV 1, a temperate u.v.-inducible DNA virus-like particle from the archaebacterium *Sulfolobus acidocaldarius* isolate B12. EMBO J 1984;3:2165–2168.
28 Prangishvili D, Forterre P, Garrett RA: Viruses of the Archaea: a unifying view. Nat Rev Microbiol 2006;4:837–848.
29 La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie JM, Raoult D: A giant virus in amoebae. Science 2003;299:2033.
30 Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM: The 1.2-megabase genome sequence of Mimivirus. Science 2004;306:1344–1350.
31 Moreira D, Brochier-Armanet C: Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. BMC Evol Biol 2008;8:12.
32 Claverie JM: Viruses take center stage in cellular evolution. Genome Biol 2006;7:110.
33 Raoult D, Forterre P: Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol 2008;6:315–319.
34 Forterre P: Manipulation of cellular syntheses and the nature of viruses: the virocell concept. Comptes Rendus Acad Sci, in press.
35 Suzan-Monti M, La Scola B, Barrassi L, Espinosa L, Raoult D: Ultrastructural characterization of the giant volcano-like virus factory of Acanthamoeba polyphaga Mimivirus. PLoS ONE 2007;2:e328.

36 Ogata H, Abergel C, Raoult D, Claverie JM: Response to comment on the 1.2-Megabase genome sequence of Mimivirus. Science 2005;308:1114–1115.
37 Cavalier-Smith T: Megaphylogeny, cell body plans, adaptive zones: causes and timing of eukaryote basal radiations. J Eukaryot Microbiol 2009;56:26–33.
38 Claverie JM, Ogata H: Ten good reasons not to exclude giruses from the evolutionary picture. Nat Rev Microbiol 2009;7:615.
39 López-García P, Moreira D: Yet viruses cannot be included in the tree of life. Nat Rev Microbiol 2009;7:615.
40 Iyer LM, Balaji S, Koonin EV, Aravind L: Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res 2006;117:156–184.
41 Yutin N, Wolf YI, Raoult D, Koonin EV: Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 2009;6:223.
42 La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Merchat M, Suzan-Monti M, Forterre P, Koonin E, Raoult D: The virophage as a unique parasite of the giant Mimivirus. Nature 2008;455:100–104.
43 Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, Robert C, Azza S, Sun S, Rossmann MG, Suzan-Monti M, La Scola B, Koonin EV, Raoult D: Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc Natl Acad Sci USA 2009;106:21848–21853.
44 Dagan T, Martin W: The tree of one percent. Genome Biol 2006;7:118.
45 Raoult D: There is no such thing as a tree of life (and of course viruses are out!). Nat Rev Microbiol 2009;7:615.
46 Raoult D: The post-Darwinist rhizome of life. Lancet 2010;375:104–105.
47 Gribaldo S, Brochier C: Phylogeny of prokaryotes: does it exist and why should we care? Res Microbiol 2009;160:513–521.
48 Filée J, Siguier P, Chandler M: I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends Genet 2007;23:10–15.
49 Koonin EV: Virology: Gulliver among the Lilliputians. Curr Biol 2005;15:R167–R169.
50 Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Rüger W: Bacteriophage T4 genome. Microbiol Mol Biol Rev 2003;67:86–156.
51 Filée J, Chandler M: Convergent mechanisms of genome evolution of large giant DNA viruses. Res Microbiol 2008;159:325–331.
52 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 NY Acad Sci 2009;1178:65–77.
53 Kristensen DM, Mushegian A, Dolja VV, Koonin EV: New dimensions of the virus world discoverde through metagenomics. Trends Microbiol 2010;18:11–19.
54 Cortez D, Forterre P, Gribaldo S: A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. Genome Biol 2009;10:R65.
55 De Parseval N, Heidmann T: Human endogenous retroviruses: from infectious elements to human genes. Cytogenet Genome Res 2005;110:318–332.
56 Filée J, Pouget N, Chandler M: Phylogenetic evidence for extensive lateral acquisition of cellular genes by nucleocytoplasmic large DNA viruses. BMC Evol Biol 2008;8:320.
57 Filée J, Forterre P, Sen-Lin T, Laurent J: Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. J Mol Evol 2002;54:763–773.
58 Filée J, Forterre P, Laurent J: The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. Res Microbiol 2003;154:237–243.
59 Forterre P, Gribaldo S, Gadelle D, Serre MC: Origin and evolution of DNA topoisomerases. Biochimie 2007;89:427–446.
60 Yutin N, Koonin EV: Evolution of DNA ligases of nucleo-cytoplasmic large DNA viruses of eukaryotes: a case of hidden complexity. Biol Direct 2009;4:51.
61 Brochier C, Forterre P, Gribaldo S: An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. BMC Evol Biol 2005;5:36.
62 Iyer LM, Aravind L, Koonin EV: Common origin of four diverse families of large eukaryotic DNA viruses. J Virol 2001;75:11720–11734.
63 Koonin EV: Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 2003;1:127–136.
64 Suhre K: Gene and genome duplication in Acanthamoeba polyphaga Mimivirus. J Virol 2005;79:14095–14101. Erratum in: J Virol 2005;79:15591.
65 Ogata H, Claverie JM: Unique genes in giant viruses: regular substitution pattern and anomalously short size. Genome Res 2007;17:1353–1361.
66 Comeau AM, Krisch HM: War is peace: dispatches from the bacterial and phage killing fields. Curr Opin Microbiol 2005;8:488–494.
67 Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D: Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. J Virol 2009;83:10719–10736.
68 Sun S, La Scola B, Bowman VD, Ryan CM, Whitelegge JP, Raoult D, Rossmann MG: Structural studies of the Sputnik virophage. J Virol 2010;84:894–897.
69 Kurland CG, Collins LJ, Penny D: Genomics and the irreducible nature of eukaryote cells. Science. 2006;312:1011–1014.
70 Poole AM, Penny D: Evaluating hypotheses for the origin of eukaryotes. Bioessays 2007;29:74–84.
71 Fuerst JA: Intracellular compartmentation in planctomycetes. Annu Rev Microbiol 2005;59:299–328.
72 Novoa RR, Calderita G, Arranz R, Fontana J, Granzow H, Risco C: Virus factories: associations of cell organelles for viral replication and morphogenesis. Biol Cell 2005;97:147–172.
73 Miller S, Krijnse-Locker J: Modification of intracellular membrane structures for virus replication. Nat Rev Microbiol 2008;6:363–374.
74 Simon LD: Infection of *Escherichia coli* by T2 and T4 bacteriophages as seen in the electron microscope: T4 head morphogenesis. Proc Natl Acad Sci USA 1972;69:907.
75 Forterre P, Prangishvili D: The origin of viruses. Res Microbiol 2009;160:466–472.
76 Forterre P: Definition of life : the virus viewpoint. Orig Life Evol Biosph 2010;40:151–160.
77 Bandea C: The origin and evolution of viruses as molecular organisms. Nature Precedings 2009. http://hdl.handle.net/10101/npre.2009.3886.1 (accessed March 26, 2010).
78 Filée J, Bapteste E, Susko E, Krisch HM: A selective barrier to horizontal gene transfer in the T4-type bacteriophages that has preserved a core genome with the viral replication and structural genes. Mol Biol Evol 2006;23:1688–1698.
79 Forterre P: Evolution, viral; in Schaechter M (ed): Encyclopedia of Microbiology, ed 3. Oxford, Elsevier, 2009, pp 370–389.
80 Krupovic M, Ravantti JJ, Bamford DH: Geminiviruses: a tale of a plasmid becoming a virus. BMC Evol Biol 2009;9:112.
81 Soler N, Marguet E, Desnoues N, Keller J, Van Tilbeurgh HN, Sezonov G, Forterre P: Two novel families of plasmids from hyperthermophilic archaea encoding new families of replication proteins. Nucleic Acid Res 2010, in press.
82 Bandea C: A unifying scenario on the origin and evolution of cellular and viral domains. Nature Precedings 2009. http://hdl.handle.net/10101/npre.2009.3888.1 (accessed March 26, 2010).