PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2009) **364**, 2263–2274
doi:10.1098/rstb.2009.0036

Review

The not so universal tree of life *or* the place of viruses in the living world

Harald Brüssow*

Chemin de la Chaumény 13, CH-1814, La Tour de Peilz, Switzerland

Darwin provided a great unifying theory for biology; its visual expression is the universal tree of life. The tree concept is challenged by the occurrence of horizontal gene transfer and—as summarized in this review—by the omission of viruses. Microbial ecologists have demonstrated that viruses are the most numerous biological entities on earth, outnumbering cells by a factor of 10. Viral genomics have revealed an unexpected size and distinctness of the viral DNA sequence space. Comparative genomics has shown elements of vertical evolution in some groups of viruses. Furthermore, structural biology has demonstrated links between viruses infecting the three domains of life pointing to a very ancient origin of viruses. However, presently viruses do not find a place on the universal tree of life, which is thus only a tree of cellular life. In view of the polythetic nature of current life definitions, viruses cannot be dismissed as non-living material. On earth we have therefore at least two large DNA sequence spaces, one represented by capsid-encoding viruses and another by ribosome-encoding cells. Despite their probable distinct evolutionary origin, both spheres were and are connected by intensive two-way gene transfers.

Keywords: universal tree; viruses; phages

1. *E PLURIBUS UNUM*: ABOUT UNITY IN BIOLOGY

When reading the books of Aristotle on animals, one gets the impression that the roots of biological research go back to classical Greece. His books offer careful descriptions, sharp reasoning, the beginning of experimentation and entertaining errors. Despite this hopeful start, it took nearly 2000 years before biology became an established branch of science. Today we continue Linné's eighteenth-century efforts to put a systematic grid on the diversity of life, which has been for biologists a source of aesthetic fascination and scientific despair. The closer biologists looked into the biosphere, the more species were discovered. To date, zoologists have described more than a million animal species and botanists a quarter of a million flowering plant species (May 1988; Brusca & Brusca 2003), it is now the turn of microbiologists to add perhaps millions of species to the long list of life. How can biologists cope conceptually and technically with this enormous species number? A deep sigh of relief came for biologists already in 1859 with the publication of Charles Darwin's book 'On the Origin of Species'. Suddenly, biologists had a unifying theory for their branch of science. One could even argue that the holy grail of a great unifying theory was achieved by Darwin and Wallace at a time when Maxwell was unifying physics, the older sister of biology, at the level of the electromagnetic field theory. In contrast to Maxwell's book, Darwin's grand design was expressed in simple sentences comprehensible to a lay audience. Understandably, his theory became a scientific and philosophical revolution.

Interestingly, the only non-verbal information in Darwin's book was a single figure with a hypothetical tree of descent for organisms. In the late nineteenth century, Ernst Haeckel quickly popularized the tree by putting names on the branches. Darwin's tree has been grown into a veritable forest. From the very concept Darwin's tree was daring: not only extant, but also all extinct organisms found, in principle, a place on this tree. It was from the very beginning designed as a universal tree.

This intuitive insight has been vindicated by the 150 years of biological research performed since Darwin first put forward his theory. Organisms so small that they were mere specks in the microscope of the nineteenth century biologists have now found a place on this tree of life. Darwin lacked any knowledge about the physical basis of heredity. Nevertheless, when the genetics, molecular biology and genomics revolutions successively rolled over biology, his theory stood the test of time and could incorporate the new discoveries (Padian 2008). Understandably, the tree became quite complex and modern depictions resemble now a rather impenetrable thicket (Doolittle 1999). The complexity of modern tree displays does not only reflect increasing knowledge and mere numbers of species. When analysing prokaryotic genomes, microbiologists actually questioned the linear descent concept of vertical evolution, a basic tenet of Darwin's concept (Doolittle & Bapteste 2007). Pictorially, many branches of the tree became interconnected. The transition of the tree of

* haraldbruessow@yahoo.com

One contribution of 11 to a Theme Issue 'The network of life: genome beginnings and evolution'.

life into a network of life is a major subject of this special issue. The problems—practical and theoretical—are serious and concern the relative contribution of vertical versus horizontal evolution in the tree of life.

Biologists are now seeking a new paradigm describing the genetic relationships that link the world of the living. This task becomes even more difficult due to the current inflow of sequence information from metagenomic projects sifting through the world's oceans, digging into the ocean's floors, or looking into the gut of animals. We have learned more about biological diversity then we could have imagined just a few years ago. At the moment we are still in a database explosion phase. When the dust has settled, the number of described bacterial species may have increased to millions. However, one important thing has not changed, the basic belief that all organisms, big or small, will most likely find their place on a universal network of life. In current thinking, all forms of life are derived from LUCA, the 'last universal common ancestor' of life (Woese 1998). The unifying concept of biology is thus surprisingly vigorous. Who would have believed that in Linné's time when individual species were still considered the result of unique acts of divine creation? The political motto 'E pluribus unum' is thus also a valid description for contemporary biology. Diversity is not an evolutionary accident, but the organizing principle in biology, without which evolution would not occur. Diversity and unity are thus just two sides of the same biological coin.

2. VIRUSES: LIFE BEYOND LUCA?

For theoretical biologists, there remains a nagging question about the unity and diversity of life. Is every organism on Earth just a variation on a single theme which started with the haphazard choice of LUCA? Theoretical biologists anticipate that LUCA was only one of several designs for early life. Do we have information about other forms of life not derived from LUCA? To answer this question, do we need to recover organisms that do not reside on our planet? One might expect answers from these extraterrestrial life forms—if they exist. Are they built on molecular principles that follow the same traditions as those seen on Earth? This would then indicate a remarkably parallel evolution (life as a cosmic necessity) or a planetary seeding event in the inner solar system (panspermia). The question whether life is a cosmic orphan has occupied human imagination and provides a strong incentive to finance costly space explorations.

Some biologists argue that we do not need exobiology and space missions to analyse life forms not derived from LUCA. A few scientists asked whether viruses are alien life forms, which challenges the concept of the universal tree of life (Forterre 2006). Biologists thinking about the origin of viruses have indeed re-defined LU*C*A as the last *c*ellular, and not *c*ommon, ancestor of life stressing the independence of viruses. The fact that viruses have only recently been studied to address the LUCA question has historical reasons. Viruses cause diseases when they infect humans, animals or plants. From the very beginning of microbiology, there was thus a strong incentive to study viruses for their medical or economic importance. Therefore, viral evolution was studied so we could learn how viruses escape from immune surveillance, or switched to new hosts. Research on bacterial viruses (phages) lacked this incentive and took another path. Following the advice from Max Delbrück, phage biologists studied biological phenomena with one simple experimental system (Brock 1990). This reductionist principle—introduced by a physicist that had turned into a biologist—led to the foundation of molecular biology. Basic mechanisms underlying the functioning of phage T4 and lambda in its host *Escherichia coli* became the early focus of phage research and not the diversity and evolution of phages.

3. EARLY IDEAS ABOUT PHAGE EVOLUTION

Escherichia coli phages were the first biological organisms for which genome sequences were determined. The comparison of the genomes from different coliphages, initially studied by heteroduplex mapping in the electron microscope, led to the influential modular hypothesis of phage evolution (Susskind & Botstein 1978). In this model, the genome of phages is perceived as a random assortment of gene clusters or modules where each module fulfils a given function. Phage lambda, for example, is composed of about 10 such modules directing DNA packaging, head morphogenesis, or lysogeny establishment, to quote some of them (Casjens *et al.* 1992). Intriguingly, the gene map order of the morphogenesis modules from lambda strictly parallels the topological order of structural elements in the phage particle (head, neck, tail, tail fibre genes) (Casjens & Hendrix 1974) as if genes, proteins and morphological structure still needed an ordered physical alignment to lead to functional phage particles. One might wonder whether this ordered gene assortment reflects a since-forgotten way of translating genetic information into a protein structure.

The function of many phage modules can be fulfilled by gene sets that lack even low sequence identity—there might be 10 alleles per module in the case of lambdoid coliphages. Lambda phages represent combinations of the different modules. In this way, an amazing diversity of lambda phages can be assorted by a permutation principle. Horizontal evolution by lateral exchange of gene modules was thus described in phages long before it became a hot subject in bacterial evolution models. There was some discussion whether phage modules are the actual units of genetic exchange. Some groups reported on DNA repeats between the modules (e.g. promoters), which could mediate the modular exchanges by homologous recombination (Clark *et al.* 2001). Other groups suggested that recombination occurs everywhere between phage genomes and the impression of modular exchanges is only the result of selection when all non-functional recombination events, not respecting the borders of functional modules, are discarded for lack of competitiveness (Juhala *et al.* 2000).

According to the modular hypothesis, phages do not show vertical evolution in the Darwinian sense and it makes no sense to draw an evolutionary tree for phages. Despite these early exciting results challenging common concepts of evolution, phage biologists

hesitated to claim a different evolutionary origin for viral genomes. The reasons were historical and philosophical. Some researchers perceived viruses more as a complex crystal made up of biomolecules than as living material. Furthermore, phages were seen as genetic material that was derived from the bacterial genome and had only achieved a certain degree of genetic independence while being totally dependent from the cellular translation machinery. Hypotheses came mainly in two forms. One 'top–down' hypothesis of regressive evolution commonly referred to as the 'reduction hypothesis' perceived phages somewhat as an extremely reduced bacterial genome—gene loss accompanied the transformation into a cellular parasite (Forterre 2009). While the 'bottom–up' hypothesis also called the 'escape hypothesis' proposes phages as the aggregation of genes which had escaped cellular control (Claverie 2006). These genes achieved autonomy from the bacterial chromosome with respect to their replication, somewhat like bacterial plasmids. At some point this selfish DNA acquired a protective shell for the transport of its genetic material from one bacterial cell to another to facilitate the crossing of a hostile chemical environment. Neither hypothesis has many adherents today, but the basic philosophical question remained unanswered: are viruses alive?

4. WHAT IS LIFE?

Viruses are by common definition neither organisms nor alive. Therefore, it was traditionally believed that viruses cannot tell us much about the origin and the evolution of life. However, this belief may be an easy solution for a dilemma which is not just about semantics, but about a valid definition of life. Biologists have a tacit understanding of life and their definition is in many ways philosophically not very sound. A frog or a sunflower is living material while a rock or clay is not. However, this emotional understanding of the obvious does not replace a clear definition. What about frogs that are completely frozen during hibernation? Are these metabolically frozen frogs alive? What about plant seeds with little or no metabolic activity? Does it make sense to distinguish alive and dead by allowing an intermediate category of biological material that has the potential to return to life? When distinguishing storage forms for potential life from actual living forms, we must be aware that we create a rear door for viruses to enter the living world. When speaking about viruses, one tacitly imagines the virion, the inert transport vehicle of the viral genome as the virus. Biologists have recently argued that the living part of the virus is not the virion, but the viral factory within the infected cell with all its associated biosynthetic activity. The virion would then correspond to the plant seed that becomes metabolically active when falling on fertile ground (in the case of a virus, a susceptible cell). If one objects that viral metabolism depends on the synthetic machinery of the host cell, one must take care not to lose obligate intracellular bacteria from the living category when designing a too narrow definition of life.

It is interesting that biologists have not yet elaborated a broadly accepted answer to the question 'What is life?'. It was actually a physicist–philosopher (Erwin Schrödinger) who provided an influential approach to this issue when writing in the 1940s a book under this title. Biologists in contrast offered context-dependent life definitions. This has an old tradition in philosophy. For example, Aristotle has already distinguished in his influential treatise 'De anima' different souls for humans, animals and plants ranging from a conscious to a nutritive soul. What modern scientists need is a definition of life that embraces all forms of living matter to the exclusion of all inanimate material. Biologists had a difficult time with a general definition of life. Life definitions embraced in various combinations entities executing a particular set of chemical and physiological reactions leading to irritability, reproduction, heredity and a metabolism. Tacitly the classical monothetic Aristotelian definition of a class was replaced by the polythetic concept of family resemblance as expressed by the philosopher Wittgenstein, where individual characteristics are neither necessary nor sufficient. Therefore, not all what is commonly perceived as living material fulfils all defining criteria. This philosophical distinction might not satisfy the scientist who is used to thinking in well-defined classes and categories. In addition, virologists have further blurred the definition of the lower end of living material by describing sub-viral entities like viroids and prions. Viroids (Owens 1999) are plant pathogens consisting only of a short, but highly structured RNA molecule that mediates pathology probably via the silencing of plant genes. No proteins are encoded by the viroid RNA and no protein shell protects the viroid RNA during transmission of the pathogen. An even more exotic pathogen is the prion causing spongiform encephalopathies like BSE in cattle and Creutzfeld-Jacob disease in humans. The name prion is a contraction of *pro*tein and *in*fectious. In the currently favoured model, the prion represents a mis-folded cellular protein that transmits its infectivity (i.e. its misfolding) in a strain-specific way without containing any nucleic acid genome. Despite their exotic nature, both sub-viral agents cause common diseases. In analogy to Aristotle's metaphysics (initially meaning his writings following his physics books), virologists have now created a meta-biology with chapters on prions following conventional viruses in virology textbooks (Prusiner 1996). The inclusion of these infectious agents has given biology a fuzzy border. Classical viruses are thus not at the bottom of the ladder of material aspiring for a definition of life.

The life definition was further compounded by the futile search for the '*vis vitalis*', the unique 'living force' that distinguishes living from non-living matter. Religion and mythology have imagined gods breathing life into clay (the parallelism with ideas of the molecular biologist Cairns on the role of clays for the replication of early genetic material is fascinating). In the early twentieth century, the *vis vitalis* was invoked by some German biologists (Driesch) as a reaction against the very materialistic biology in the second half of the nineteenth century (Haeckel) that saw only physico-chemical laws behind the most complex biological phenomena. In an influential textbook from the 1980s entitled 'The Vital Force', Frank

Harold defined the absorption of environmental physical and chemical energy and its transformation into a proton or electric gradient across a biomembrane, which then powers all processes of the living cell, as the very essence of bioenergetics. Is the chemi-osmotic process the vital force, the essence of the living state?

However, biologists do not think that the basic problems of the definition of life will be solved by philosophical reasoning. They believe in the explicative power of a careful interpretation derived from newly acquired data. Over the last decade, new data have accumulated that allows biologists to address the question of the evolutionary origin of viruses and thus perhaps the origin of life on earth in scientific terms. Significant progress came from genomics, structural biology and the discovery of unusual viruses.

5. GENOMICS REVEALS ELEMENTS OF VERTICAL EVOLUTION IN PHAGES

The heydays of classical phage research were already over when systematic and large-scale sequencing efforts started in the 1990s. Initially, genomics researchers considered phage genomes as too small and too dull to be informative (Brüssow & Hendrix 2002). Instead of beginning with phage sequencing, the major sequencing centres started directly with bacterial genomes. The first systematic phage sequencing was performed by applied microbiologists (Brüssow & Desiere 2001), who were investigating the problem of phage infection: Phage interfered with industrial food fermentation, particularly in the dairy industry (Brüssow 2001), which uses Gram-positive bacteria as starter cultures. Despite the large evolutionary gap between the host bacteria, dairy phages closely resembled phage lambda infecting a Gram-negative host. A further surprise was that systematic sequencing efforts with dairy phages revealed substantial elements of vertical evolution in phages (Desiere *et al*. 1999). The alignments of the phage morphogenesis modules revealed gradients of relatedness that ranged from close to lesser degrees of DNA sequence identity. Phage genomes were identified that lacked DNA sequence, but still showed appreciable protein sequence identity. Further relatedness was revealed through conserved gene order in the absence of any sequence similarity (Lucchini *et al*. 1999). Distinct lineages of lambdoid phages were differentiated in Gram-positive bacteria (Sfi21- and Sfi11-like phages) that resembled those defined in Gram-negative bacteria (HK97- and lambda-like phages) (Brüssow & Desiere 2001). With respect to the morphogenesis module, the streptococcal phage Sfi21 resembles coliphage HK97. The structural gene order from streptococcal phage Sfi11 resembled that of coliphage lambda. Notably, the similarities in the structural genes were greater between phages Sfi21 and HK97 (or Sfi11 and lambda) than those between HK97 and lambda, which infect the same host species *E. coli*. Clearly, phages did not co-evolve with their bacterial hosts, but followed their own evolutionary trajectories. For the major head protein, weak protein sequence identity was detected not only between lambda-like phages infecting Gram-negative and Gram-positive bacteria, but even with phages infecting Archaea (Pfister *et al*. 1998). It was deduced that the structural module of lambda-like phages had evolved before prokaryotic cells had split into Bacteria and Archaea, suggesting substantial antiquity for these phage genes.

6. THE PROBLEM OF A PHAGE TREE

Encouraged by these observations, microbiologists tried to develop phylogenetic trees for phages based on the alignment of individual phage proteins. Success was limited because of the lack of phage proteins that were conserved across different classes of phages. Even within more defined classes of phages like the lambda supergroup, few proteins were sufficiently conserved to be useful for tree-building. Trees have been constructed using the phage terminase and the phage portal protein, notably two proteins that are involved in the packaging of DNA into the phage capsid. However, not much evolutionary insight was gained from these analyses. More promising was tree-building using the major head gene sequences. With increasing distance from the reference T4 *E. coli* phage, Pseudo, Schizo and Exo T-even phages were classified (Tétart *et al*. 2001) that defined a large Myoviridae phage family infecting host bacteria as distinct as Proteobacteria and Cyanobacteria. Large-scale ecological surveys across the world's oceans yielded marine phage isolates that defined six further groups of T-even phages on this tree. The marine T4-type phages were so abundant in the oceans and so diverse with respect to their sequence that the scientists spoke about a ubiquitous component of dark matter in the biosphere (Filée *et al*. 2005). However, despite the wide reach of this tree, it still described the ramifications in a single, albeit widespread phage group. Unsatisfied with these tree-building exercises, genomics-oriented microbiologists tried tree-building in the absence of signature sequences so that phages belonging to different morphological groups could be compared. To achieve that goal, these scientists used the whole set of protein sequences encoded by a phage to calculate genetic relationships between different phage groups. This 'Phage Proteome Tree' allowed a grouping of phages that was largely compatible with taxonomical classification (Rohwer & Edwards 2002). Very distinct phages differing in genome type and morphology were connected in this tree, but no links to cellular organisms were described. However, the association of phages that use chemically different forms of nucleic acids as components of their genome is problematic since genome researchers and phage taxonomists argued strongly for a polyphyletic origin of phages (Fauquet 1999; Koonin *et al*. 2006). Newer reticulate classification schemes for phages based on gene content analysis clearly distinguished phages with dsDNA, ssDNA, dsRNA and ssRNA genomes (Lima-Mendez *et al*. 2008). In the DNA categories, at least five distinct evolutionary clusters of phages were detected and the idea of a single phage tree was rejected.

7. LARGE VIRUSES: LINKS TO THE UNIVERSAL TREE?

Small viruses are under pressure to fulfil all essential viral functions with a minimal genome size. Economical use of the genetic material led to the use of different reading frames on the same strand and even the use of the opposite DNA strand for coding proteins. It is likely that under such selection pressures, information about the origin of the viral genes has been erased. Virologists have therefore turned to the genomes of large viruses in order to determine the origin of viral genes. The best investigated of the large phages is T4 phage from *E. coli*. T4 encodes 300 proteins in its 170 kb genome. Despite decades of research, a function could be defined for only about half of the T4 genes. A mere 60 T4 genes are essential in laboratory growth (Miller *et al.* 2003). The sequencing of further T4-like coliphages has defined additional subgroups within the T-even phage group (Tétart *et al.* 2001). Comparative T-even phage genomics has revealed a core of conserved genes represented by contiguous blocks of structural genes and DNA replication genes, separated by hyperplastic regions containing mostly novel genes of unknown function and origin (Filée *et al.* 2006; Comeau *et al.* 2007). The unassigned T4 genes are clustered in a few designated genome regions, they are encoded on one strand and they represent particularly small open reading frames (ORFs). The direction of transcription is the same as that of the flanking conserved genes. Most are found among early and middle genes, and many of them encode lethal functions for *E. coli* (Comeau *et al.* 2007; Zuber *et al.* 2007). Comparative genomics revealed that the gene pool of T4 coliphages is clearly distinct from that of its host *E. coli* (Zuber *et al.* 2007). Few T4 proteins, those mainly involved in DNA metabolisms, are homologous with *E. coli* proteins. A phylogenetic tree analysis suggested that these T4 enzymes branched off before the split between Eukarya and Bacteria (Miller *et al.* 2003), suggesting a very ancient origin of T4 phages.

Virologists subsequently examined even larger virus genomes to identify potential links with the universal tree. For example, mimivirus infects amoeba and has an exceptionally large 1.2 Mb genome for a virus (Raoult *et al.* 2004). This genome size is substantially larger than that of many intracellular bacteria and blurs further the distinction between the viral and the cellular world. In contrast to other viruses, mimivirus encodes a few proteins involved in protein synthesis that belong to a set of universally conserved genes in cellular life. When these proteins were concatenated for a tree analysis, mimivirus became a fourth branch on the universal tree of life located next to the Eukarya split. Analyses by other groups strongly disagreed with this view (Moreira & Brochier-Armanet 2008), suggesting that mimivirus acquired many of these genes by horizontal gene transfer from its amoebal hosts. Indeed, the evolutionary relationships of mimivirus are complex. It shows genes with bacterial affinity which were attributed to the feeding mode of the host amoeba eating bacteria. Tree analyses with other proteins placed mimivirus next to algal viruses of the Phycodna group and related viruses have been found with high frequency and diversity in the oceans (Monier *et al.* 2008). There are also indications that mimivirus can cause pneumonia in humans still further extending the host range of this versatile and promiscuous giant virus.

8. VIRAL HALLMARK GENES AND THE ORIGIN OF DNA

Genomics researchers have identified a number of viral hallmark genes that are shared by many diverse groups of viruses, but never by all of them (Koonin *et al.* 2006). These genes have distant homologues in cellular organisms. Forterre speculated that these viral-specific proteins were either directly invented in an ancient viral world or were recruited from ancient cellular lineages that existed parallel to LUCA, but which are now extinct (Forterre 2002; Forterre & Gribaldo 2007). Examples of viral hallmark genes include the jelly-roll capsid protein (see below) and a few DNA replication (helicase, primase) and packaging genes (terminase). From the comparison with cellular hallmark genes, Forterre derived two different spheres of genomes. On one side you find ribosome-encoding organisms, constituting the 'universal' tree of cellular life, and on the other side are capsid-encoding viruses, constituting the virosphere (Raoult & Forterre 2008). Cellular organisms lack capsid genes, while viruses lack the protein translation apparatus. In some recent speculations, the cellular world traces back into the RNA world with LUCA, which had already invented translation and RNA genomes, but not DNA genomes and cell membranes as yet. There are also speculations that viruses have introduced DNA into LUCA. Forterre proposed a hypothesis about three RNA cells which led to the current three ribosomal lineages. Each of these three RNA cells acquired a distinct DNA virus. According to that hypothesis, viruses contributed not only the modern dsDNA genome to the RNA cells (Forterre 2006), but also introduced the enzymes for DNA replication into the three primordial cells (Forterre 1999). The idea is not as farfetched as it sounds at first and other researchers came up with similar suggestions (Villareal & De Filippis 2000). Viruses show a great flexibility with the chemistry of their genomes. Uridine-containing deoxynucleotides have been located in phages (Takahashi & Marmur 1963), chemically modified bases are plentiful, single and double-stranded DNA and RNA are used as well as viruses that switch regularly in their life cycle between RNA and DNA genomes. The modification of contemporary viral genomes is interpreted as an answer to the pressure of cellular nucleases that would otherwise digest unprotected genomes. At the very beginning of biological evolution, this pressure might have been the incentive for viruses to play with the chemistry of nucleic acids. Viral dsDNA had the added advantage of a greater chemical stability than RNA, allowing the faithful replication of larger genomes and DNA might then have been co-opted by the cell of the RNA world. It is an interesting observation that DNA replication enzymes such as DNA polymerase, primase and helicase are not orthologous in Bacteria and Archaea despite the very similar

DNA replication mechanisms used in both organisms. A common ancestral state is therefore unlikely. Forterre argues that they were probably derived from two different DNA viruses. In fact, there is good evidence that mitochondria use DNA polymerases derived from T3/T7-like prophages that had invaded the *Rickettsia* ancestors of the mitochondria. The argument was extended to chloroplast DNA polymerase and nuclear-encoded DNA polymerase in eukaryotic cells (Filée & Forterre 2005). Instead of phages deriving genes from the cellular gene pool, genomics has provided evidence for a gene flow from phages into the cellular domain (McGeoch & Bell 2005).

Large viruses elaborate complex virus-induced structures in the infected host cell, including an intracellular viral factory which somewhat resembles a cell nucleus (Suzan-Monti *et al.* 2007). This has led to the speculation that the eukaryotic nucleus might be a viral invention (Claverie 2006).

9. SATELLITE VIRUSES

In mimivirus-infected cells, researchers found virus production sites (viral factories) which release mature progeny viral particles into the cell. Recently, a much smaller virus was seen associated with the mimivirus intracellular viral factory. It was called Sputnik and had an 18 kb genome (La Scola *et al.* 2008). Most of its genes had no database matches while three were related to mimivirus proteins. Sputnik cannot grow on uninfected amoeba. However, when it grows in mimivirus-infected amoeba, it substantially decreases the yield of mimivirus and the lysis of the amoeba. Notably, this is the first virus which grows on another virus. Paradoxically, the fact that it can get 'sick' makes mimivirus more 'alive' (Pearson 2008) and pushes this virus even more across the borderline traditionally separating viruses from organisms.

Satellite viruses are known from animals and plants (Talianski & Palukaitis 1999). Their small genomes do not encode for a capsid protein and the satellite virus relies on a helper virus for encapsidation. Satellite viruses like the adeno-associated virus (Berns 1996) divert resources from the helper virus, but Sputnik makes mimivirus literally 'ill' leading to the production of morphologically aberrant mimivirus. Virologists have also shown that defective interfering particles from negative strand RNA animal viruses reduce the infectivity of the helper virus (Roux 1999). However, in that case a truncated genome of the helper virus lacking the viral RNA polymerase gene competed very efficiently with the replication of the full-sized genome. All these discoveries demonstrate that we have to account for virus–virus competition when describing the virus–cell interaction.

10. STRUCTURAL BIOLOGY ALLOWS A DEEP LOOK INTO THE PAST

The jelly-roll capsid protein is an interesting viral hallmark gene. The main structural feature of this protein is a pair of eight-stranded viral β-barrels or jelly rolls. The two jelly rolls have the same topology, but no apparent sequence conservation. If they are the result of gene duplication, the event must be very ancient. This protein associates in trimers and forms a basic structural element composing the viral capsid. When analysing this protein in the *E. coli* phage PRD1, structural biologists observed striking similarities with the hexon structure from animal adenovirus (Benson *et al.* 1999). The similarity between phage and adenovirus goes even farther. Both capsids show the same lattice type. Pentameric proteins occupy the vertices of their capsids, to which fibre proteins are attached. Both viral genomes are linear dsDNA with inverted terminal repeats. Both viruses have terminal proteins, covalently linked to the 5′-end of the DNA, which are used as primers for DNA replication. The researchers judged that these observations cannot be explained by convergence and they argued for a very ancient link between viruses infecting two distinct domains of life. They predicted that jelly roll proteins will be found in viruses infecting other forms of cellular life. Soon it was shown that a Phycodnavirus that infects *Chlorella* (a unicellular photosynthetic alga) had a capsid protein similar to that of phage PRD1 (Nandhagopal *et al.* 2002). Similarities went beyond this shared protein fold. The algal virus, when attached to its hosts, digests the cell wall around the attachment point, injects its DNA and leaves its empty capsid on the cell surface. This infection mechanism is quite unusual for a virus infecting eukaryotes, but typical for viruses infecting prokaryotes. This initial prediction has been confirmed as the same capsid fold has been described in an archaeal virus that infects *Sulfolobus*, which lives in an extreme environment characterized by low pH and high temperatures (Khayat *et al.* 2005). The three-dimensional structure of these two proteins could be overlaid with only minimal deviation in space (<2 Å). These observations are now a strong argument for a common origin of the capsid in viruses infecting all three domains of cellular life. One is tempted to postulate that the ancestor virus, which first displayed this capsid fold, 'lived' before the three domains of life had separated. Like the shared gene map argument, structural biology provides new tools in the exploration of distant evolutionary relationships, which are so ancient that all sequence similarity has been erased.

Dennis Bamford (2003) formulated a hypothesis that the capsid structure and the genome packaging machinery are the 'self' (or in the terminology of Aristotle, the 'soul') of the virus, which was faithfully inherited from the viral ancestor. Other traits of the virus such as recognition and multiplication in a given host necessitate adaptation of the virus to the cellular host. Proteins involved in that task tend to be acquired horizontally, typically from the host or from other viruses exploiting the same host. These genes belong to the 'non-self' part of the viral genomes which do not show ancient relationships. This hypothesis fits nicely with the genome categorization of viral genes and ideas from Forterre on viral evolution.

Structural biologists have also identified a second lineage of viral capsid proteins that is likewise distributed across all three domains of life. A closely related fold is found in the head proteins from *E. coli* phages

HK97 and T4 (Fokine *et al.* 2005), representing two different taxonomical groups of phages, namely Siphoviridae and Myoviridae. A related protein fold was identified in a virus-like particle from the Archaeon *Pyrococcus furiosus* and in animal herpesvirus (Akita *et al.* 2007). The wide phyletic distribution of a second viral protein fold suggests that the ancestor of the three domains (Commonote) might already have been infected by at least two ur-viruses defining two distinct lineages.

11. FROM ORPHANS ...

The morphological structure of bacterial viruses is quite peculiar. The prototype is represented by a capsid containing the viral genome, a tail for injecting the genome into the bacterial cell and a base plate with tail fibres for the identification of the appropriate target cell (Miller *et al.* 2003). This structure combines the advantage of a genuine gene container, with a mechanical conduit for guiding the DNA into the target cell, linked to a sophisticated sensor, which differentiates target from non-target cells. The tailed virus model is so efficient that nearly every physical viral particle is also an infectious virus. In morphologically less well-defined animal viruses, frequently only one out of 1000 physical particles is actually an infectious virus. It is not obvious from what cellular material these sophisticated base plate structures could have been derived (Kanamaru *et al.* 2002). Some phages show a distinctly different morphology including tail-less capsids, membrane-enveloped phage particles and filamentous phages. However, the selective advantage of the tailed phage construction model is overwhelming as more than 96 per cent of all described bacterial viruses are tailed phages (Caudovirales) (Ackermann 1996).

Viruses from the other two domains of life are morphologically less uniform. Caudovirales are also found in the Euryarchaeota subgroup of Archaea (Pfister *et al.* 1998). Some Euryarchaeota also have a high percentage of genes of probable bacterial origin, leading to the suggestion that Caudovirales entered Euryarchaeota by interdomain genetic exchanges. In the other branch of Archaea, the Crenarchaeota, a bewildering morphological diversity of viruses, was identified (Prangishvili *et al.* 2006). Strange archaeal viruses resembled spindles (Fuselloviridae) or are lemon-shaped. Some develop two-tailed spindles decorated with terminal hooks (Bicaudoviridae), which remarkably finished their morphogenesis outside of the infected cell (Häring *et al.* 2005). Such cell-independent 'growth' of viruses is a novelty in the viral world and might be reminiscent of an earlier greater autonomy of viruses from the cellular hosts. In Crenarchaeota, there are furthermore bottle-shaped (Ampullaviridae) and droplet-shaped (Guttaviridae) viruses, others are linear viruses without (Rudiviridae) and with lipid-containing envelopes (Lipothrixviridae). Some linear viruses end with a claw, which clamps the virus onto the pili of the archaeal host. There are also enveloped spherical viruses (Globuloviridae) and spherical archaea viruses that contain internal membranes like Tectivirus PRD1, infecting *E. coli*. It is not clear how this morphological diversity of viruses could have originated in the cellular world.

12. ... TO ORFANS

Viruses from Crenarchaeota display unique morphologies and their genomes contain mainly unknown genes. For example, Acidianus bottle-shaped virus, an Ampullavirus, showed over its 57 ORFs only three which had significant database hits (Peng *et al.* 2007). Similarities, if they occurred, were with DNA and RNA transaction enzymes (ATPases, packaging enzymes, DNA precursor metabolism, RNA modification enzymes, glycosylases), never with structural proteins.

Another archaeal virus, a Globulovirus, likewise did not have a significant hit in the sequence databases suggesting a substantial genetic isolation of the archaeal virosphere (Häring *et al.* 2004). Also in some bacterial viruses (e.g. *Mycobacteria* phages), the majority of the ORFs are novel (Pedulla *et al.* 2003). This observation contrasts with the sequence analyses from bacterial genomes where only 10 per cent of the ORFs generally lack matches with the database (Edwards & Rohwer 2005). One might argue that much more of the global bacterial metagenome has been sampled, while the viral metagenome is still mostly unexplored. However, substantial metagenome analyses of viral DNA sequences have recently been conducted in various environments. One large study comprised 15 million sequences from nine biomes, terrestrial and aquatic (Dinsdale *et al.* 2008). The most extensively explored system was ocean water collected over a decade and representing the major oceanic regions of the world. In the largest of the ocean studies, more than 91 per cent of the sequences from the viral DNA fraction did not have a significant hit in the sequence databases (Angly *et al.* 2006). From this observation, we have to conclude that the viral DNA sequence sphere is very large. Microbial ecologists working in the oceans provided data that independently support this conclusion (Wommack & Colwell 2000). The first big surprise was the discovery of large numbers of viruses in coastal water. In eutrophic estuarine water, 10^7 viral particles were counted per millilitre of water. This is 10 times the amount of bacteria in this ecosystem. From the data of many ecological surveys, it was calculated that viruses are by far the most abundant 'biological entities' in the world's oceans yielding a global level of more than 10^{30} viruses. Viruses are not only numerous, they are also a major cause of microbial mortality in the sea, rivaled only by grazing from protists (Breitbart *et al.* 2008; Danovaro *et al.* 2008; Jürgens & Massana 2008). Viruses play an important role in the ocean ecosystem by maintaining the genetic diversity of microbes according to the 'killing the winning fraction' concept (Thingstad & Lignell 1997; Wommack & Colwell 2000). In addition, viruses power the microbial loop that maintains nutrients in the microbial world, preventing their flow into the marine food chain. Therefore, viruses play a major role in biogeochemistry. This role is not limited to

the open sea. The deep-sea floor covers approximately 65 per cent of the Earth's surface and the prokaryotic biomass in the top 10 cm of the ocean sediment contains an estimated half of the total microbial carbon on Earth. Recent ecological surveys documented a deep viral impact on this benthic ecosystem (Danovaro *et al.* 2008): approximately 80 per cent of the prokaryotes in the sediment succumb to viral infection.

If one combines the large number of ORFans in viral metagenome analyses and the sheer number of viruses in the biosphere, it is possible that the viral sequence space exceeds that of their prokaryotic hosts in size. These data are simply not compatible with the older concept that the viral genes escaped from cells. While the invention of translation gave LUCA world dominance and forced viruses into an existence of parasites of cellular life, it did not lead to a massive gene loss in the viral sequence space (Brüssow 2007*a*).

13. GENE FLOW BETWEEN VIRUSES AND CELLS

Classically, phages and bacteria were interpreted in the predator–prey framework. Phages could not extinguish the host cells because they relied on the translation and energy production capacities of their bacterial hosts. Should they wipe out their bacterial host, the phage would also go extinct. In contrast, it is not clear why bacteria should not force phages into extinction. The conundrum is normally answered by reference to an arms race between phages and bacteria, which can be inferred from the analysis of the highly variable genes. Genes encoding restriction enzymes (as defence against foreign DNA intrusion) and lipopolysaccharides (as phage receptor) are hotspots of *E. coli* genetic diversity, which is interpreted as response to phage infection pressure. However, this cannot be the whole story. During *in vitro* co-evolution experiments, phages frequently lose the race against bacteria. Are there reasons that bacteria live under natural conditions better with phages than without them? Indeed, in recent years, phage–bacterium interaction has received a much more positive interpretation than previously. Phages were identified as a source of bacterial genetic diversity. When different bacterial strains, belonging, for example, to the same species of lactic acid bacteria, are compared, they typically differ for approximately 10 per cent of their gene content (Berger *et al.* 2007). Prophage DNA frequently account for a third of the differences. If phages import useful genes for bacteria, they could be of selective advantage for the lysogen. This is particularly true considering most bacteria do not undergo sexual reproduction. In principle, bacteria should therefore be clonal organisms displaying less genetic variability than sexual organisms. This is not the case. Bacterial species show even greater genetic variability than eukaryotic species—this comparison is, however, somewhat compromised by the ambiguous bacterial species definition. Anyway, bacteria have discovered powerful methods for introducing genetic variation and phages are probably a form of 'infectious sex' for bacteria. 'Unprotected sex' is also dangerous for bacteria since they would accumulate too much prophage DNA. Indeed, some bacteria harbour many prophages and appear nearly as phages in bacterial disguise, but this seems to be recent events such as in the emerging food pathogen *E. coli* O157 (Canchaya *et al.* 2003). Over evolutionary times bacteria have maintained a relatively small genome size despite the constant bombardment with DNA from temperate phages. There must be forces acting against this accumulation of prophage DNA. Genomics showed a trend for prophage DNA decay (occurrence of non-infectious prophages, prophage remnants and isolated phage genes in bacterial genomes; Canchaya *et al.* 2003). This process can already be read from the *E. coli* O157 genome. Phage biologists proposed a non-discriminating DNA deletion process in bacteria against the accumulation of alien DNA (Lawrence *et al.* 2001). If essential bacterial genes are deleted, the clone is lost from the population; if, however, prophage sequences are deleted, the clone might have an increased fitness with respect to non-lysogenic cells by the very fact that they have less DNA to replicate. However, this advantage of non-lysogenic over lysogenic cells has never been demonstrated experimentally (Denou *et al.* 2008). Classical data with laboratory strains of *E. coli* actually showed that the possession of a prophage was advantageous for the cell (Edlin & Kudrna 1975). Also, the genome analysis of pathogenic bacteria was in that respect quite revealing. A surprisingly high percentage of bacterial virulence factors turned out to be prophage-encoded. Medical microbiologists documented a positive selection for prophage-containing *Streptococcus pyogenes* pathogens in the human population over the last 80 years. What had happened? In high GC content Gram-positive bacteria, the phage-encoded virulence factors tend to be next to the right phage attachment site (Brüssow *et al.* 2004). Phages have most likely acquired bacterial genes by an imprecise prophage excision process from the bacterial chromosome. By serial integration/excision events, a temperate phage with a broad host range can explore a relatively large bacterial sequence space and import valuable genes, which cannot be acquired that easily by other genetic means. Poly-lysogeny (the possession of multiple prophages) opens up the possibility of genetic variation by a permutation principle and puts bacterial pathogens exploiting prophages on an evolutionary fast lane (Canchaya *et al.* 2004).

In double infections, phage–phage DNA recombination events can take place, allowing bacteria access to the large viral DNA sequence space. These viral genetic crosses might also be the possible origin of a particular form of phage DNA, which was called by the Pittsburgh phage group 'morons' (Juhala *et al.* 2000; Hendrix & Casjens 2008). As this 'more of DNA' was located internally in the phage genome, imprecise integration/excision events are an unlikely interpretation for this extra DNA. The 'morons' in lambdoid phages represent separate genetic units, flanked by their own transcriptional signals (Hendrix *et al.* 2000). Comparative genomics thus demonstrated that phage DNA became an important motor for the evolution of bacterial pathogenicity. If phages confer fitness genes to bacterial commensals or free-living

bacteria, there may be a good reason why having a phage could be of selective advantage to many bacteria, not as an individual (which is anyway not well defined in a non-sexual population), but as a population.

Indeed, when the gene content in the virus fraction of environmental samples was analysed by metagenome analysis, a substantial amount of apparent bacterial genes was detected in the virome (Dinsdale *et al*. 2008). The magnitude of microbial metabolic capacities contained in the viral DNA fraction suggested to these researchers that viruses serve as a repository for storing and sharing genes among their microbial hosts. Phages are ideal gene carriers because their capsid shows some tolerance towards extra DNA packaging. In fact, phages must carry a minimal amount of DNA to package their capsids with DNA correctly, and are therefore not penalized to carry extra DNA (morons) in their genome.

The role of prophages for the ecological adaptation of commensal bacteria (e.g. by conferring fitness factors) has not been intensively studied. Some prophages were shown to mediate the transition from planktonic to aggregated growth or biofilms. The analysis of phages infecting cyanobacteria has demonstrated that phages can also carry photosynthesis genes in their genomes. Phages frequently inhibit, in the infection process, the transcription of the host genes. Some photosynthesis proteins are, however, particularly sensitive to light damage. Phages would therefore lose their energy basis if these genes are not transcribed during infection. The easiest way of assuring their expression is to encode them on the phage genome; this has recently been demonstrated experimentally (Lindell *et al*. 2005). So far cyanophages have been mainly studied under lytic infection conditions. It is therefore not clear whether these phages can coexist with their hosts and whether these imported photosynthesis genes could be of selective advantage to the bacterial host. It is interesting to note that photosynthetic capacities are found in bacteria that are phylogenetically unrelated. Horizontal gene transfer must therefore have played a major role in the evolution of bacterial photosynthesis. Photosynthesis gene-carrying phages are ideal gene shuttling devices.

What is the advantage for phages in this scenario? If a bacterial clone has a selective advantage, it will grow out. A prophage contained in this clone will also passively grow out with the replication of the bacterial chromosome. In the context of the selfish DNA concept, this might be enough 'incentive' for phages to cooperate. Part of the gene flow between bacteria and phages might thus be explained by mutual benefit.

The quest for food is one of the driving forces of biological evolution. The involvement of phages in bacterial virulence was recently rationalized within this theoretical framework. The arguments are as follows (Brüssow 2007*b*): in ecological surveys, bacteriophages and protist grazers are major causes of bacterial mortality. Genomics suggests that phages evolved well before eukaryotic protists. Bacteria were thus initially only confronted with phage predators. When eukaryotic protists evolved, bacteria were caught between two types of predator. One successful anti-grazing strategy of bacteria was the elaboration of toxins that would kill the protist grazer. The released cell content would, in addition, feed bystander bacteria. To fight grazing protists, bacteria cooperate with those phage predators that can coexist with them in the form of lysogeny. Lysogeny was perhaps initially a resource management strategy of phages that could not maintain infection chains. Subsequently, lysogeny evolved into a bacterium–prophage coalition attacking protists, which became a food source for them. When protists evolved into multicellular animals, the lysogenic bacteria tracked their evolving food source. This hypothesis could explain why a frequent scheme of bacterial pathogenicity is the survival in phagocytes and why a significant fraction of bacterial pathogens show prophage-encoded virulence genes and why some virulence factors of animal pathogens are still active against unicellular eukaryotes. Game theory would see bacterial pathogenicity as one playing option in the stone-scissor-paper conflict entertained between phages, bacteria and protists.

Similar arguments might be developed for the complex relationship between viruses and animals (Bushman 2002). When describing the large amount of retroviral elements in the human genome, some researchers quibbled that humans derive genetically as well from apes as from viruses. Like bacterial genomes that contain prophages, mammalian genomes contain many provirus sequences from retroviruses. As animal genomes were not selected for maintaining a small size, mammals accumulated a huge amount of provirus remnants. Our own genome is literally littered with their terminal repeats. It is unknown what role this DNA plays and whether it is of selective advantage to us. Like prophages carrying genes of apparent bacterial origin, some retroviruses carry animal genes that can be of dire consequence (oncogenes). However, it has been demonstrated that the development of morphological structures crucial for mammals like the placenta is influenced by provirus-encoded genes (Schulte *et al*. 1996; Mi *et al*. 2000).

A thorough analysis of the origin and the breadth of the viral sequence space could thus not only lead to interesting theoretical insights into the beginning of genome evolution, which might be hard to come by otherwise, but also to insights of medical importance.

The author acknowledges many discussions with Patrick Forterre, David Prangishvili, Roger Hendrix, Henry Krisch and Dennis Bamford during two Les Treilles meetings on the evolution of phages.

REFERENCES

Ackermann, H. W. 1996 Frequency of morphological phage descriptions in 1995. *Arch. Virol.* **141**, 209–218. (doi:10.1007/BF01718394)

Akita, F., Chong, K. T., Tanaka, H. *et al.* 2007 The crystal structure of a virus-like particle from the hyperthermophilic archaeon *Pyrococcus furiosus* provides insights into the evolution of viruses. *J. Mol. Biol.* **368**, 1469–1483. (doi:10.1016/j.jmb.2007.02.075)

Angly, F. E. *et al.* 2006 The marine viromes of four oceanic regions. *PLoS Biol.* **4**, 2121–2131.

Bamford, D. H. 2003 Do viruses form lineages across different domains of life? *Res. Microbiol.* **154**, 231–236. (doi:10.1016/S0923-2508(03)00065-2)

Benson, S. D., Bamford, J. K. H., Bamford, D. H. & Burnett, R. M. 1999 Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. *Cell* **98**, 825–833. (doi:10.1016/S0092-8674(00)81516-0)

Berger, B., Pridmore, R. D., Barretto, C., Delmas-Julien, F., Schreiber, K., Arigoni, F. & Brüssow, H. 2007 Similarity and differences in the *Lactobacillus acidophilus* group identified by polyphasic analysis and comparative genomics. *J. Bacteriol.* **189**, 1311–1321. (doi:10.1128/JB.01393-06)

Berns, K. I. 1996 Parvoviridae: the viruses and their replication. In *Fields virology* (eds B. N. Fields, D. M. Knipe & P. M. Howley), pp. 2173–2197, Philadelphia, PA: Lippincott-Raven Publishers.

Breitbart, M., Middelboe, M. & Rohwer, F. 2008 Marine viruses: community dynamics, diversity and impact on microbial processes. In *Microbial ecology of the oceans* (ed. D. L. Kirchman), pp. 443–480, Hoboken, NJ: Wiley-Blackwell.

Brüssow, H. 2001 Phages of dairy bacteria. *Annu. Rev. Microbiol.* **55**, 283–303. (doi:10.1146/annurev.micro.55.1.283)

Brüssow, H. 2007*a* *The quest for food: a natural history of eating.* New York, NY: Springer Scientific Publisher.

Brüssow, H. 2007*b* Bacteria between protists and phages: from antipredation strategies to the evolution of pathogenicity. *Mol. Microbiol.* **65**, 583–589. (doi:10.1111/j.1365-2958.2007.05826.x)

Brüssow, H. & Desiere, F. 2001 Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. *Mol. Microbiol.* **39**, 213–222. (doi:10.1046/j.1365-2958.2001.02228.x)

Brüssow, H. & Hendrix, R. W. 2002 Phage genomics: small is beautiful. *Cell* **108**, 13–16. (doi:10.1016/S0092-8674(01)00637-7)

Brüssow, H., Canchaya, C. & Hardt, W. D. 2004 Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **68**, 560–602. (doi:10.1128/MMBR.68.3.560-602.2004)

Brock, T. D. 1990 *The emergence of bacterial genetics.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Brusca, R. C. & Brusca, G. J. 2003 *Invertebrates.* Sunderland, MA: Sinauer Associates Publishers.

Bushman, F. 2002 *Lateral DNA transfer.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brüssow, H. 2003 Prophage genomics. *Microbiol. Mol. Biol. Rev.* **67**, 238–276. (doi:10.1128/MMBR.67.2.238-276.2003)

Canchaya, C., Fournous, G. & Brüssow, H. 2004 The impact of prophages on bacterial chromosomes. *Mol. Microbiol.* **53**, 9–18. (doi:10.1111/j.1365-2958.2004.04113.x)

Casjens, S. & Hendrix, R. W. 1974 Comments on the arrangement of the morphogenetic genes of bacteriophage lambda. *J. Mol. Biol.* **90**, 20–23. (doi:10.1016/0022-2836(74)90253-8)

Casjens, S., Hatfull, G. & Hendrix, R. W. 1992 Evolution of dsDNA tailed-bacteriophage genomes. *Sem. Virol.* **3**, 383–397.

Clark, A. J., Inwood, W., Cloutier, T. & Dhillon, T. S. 2001 Nucleotide sequence of coliphage HK620 and the evolution of lambdoid phages. *J. Mol. Biol.* **311**, 657–679. (doi:10.1006/jmbi.2001.4868)

Claverie, M. 2006 Viruses take center stage in cellular evolution. *Genome Biol.* **7**, 110. (doi:10.1186/gb-2006-7-6-110)

Comeau, A. M., Bertrand, C., Letarov, A., Tétart, F. & Krisch, H. M. 2007 Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology* **362**, 384–396. (doi:10.1016/j.virol.2006.12.031)

Danovaro, R., Dell'Anno, A., Corinaldesi, C., Magagnino, M., Noble, R., Tamburini, C. & Weinbauer, M. 2008 Major viral impact on the functioning of benthic deep-sea ecosystems. *Nature* **454**, 1084–1087. (doi:10.1038/nature07268)

Denou, E. *et al.* 2008 The role of prophage for genome diversification within a clonal lineage of *Lactobacillus johnsonii*: characterization of the defective prophage LJ771. *J. Bacteriol.* **190**, 5806–5813. (doi:10.1128/JB.01802-07)

Desiere, F., Lucchini, S. & Brüssow, H. 1999 Comparative sequence analysis of the DNA packaging, head, and tail morphogeneis modules in the temperate *cos*-site *Streptococcus thermophilus* bacteriophage Sfi21. *Virology* **260**, 244–253. (doi:10.1006/viro.1999.9830)

Dinsdale, E. A. *et al.* 2008 Functional metagenomic profiling of nine biomes. *Nature* **452**, 629–632. (doi:10.1038/nature06810)

Doolittle, W. F. 1999 Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129. (doi:10.1126/science.284.5423.2124)

Doolittle, W. F. & Bapteste, E. 2007 Pattern pluralism and the tree of life hypothesis. *Proc. Natl Acad. Sci. USA* **104**, 2043–2049. (doi:10.1073/pnas.0610699104)

Edlin, G. L. & Kudrna, R. 1975 Lambda lysogens of *E. coli* reproduce more rapidly than non-lysogens. *Nature* **255**, 735–737. (doi:10.1038/255735a0)

Edwards, R. & Rohwer, F. 2005 Viral metagenomics. *Nat. Rev. Microbiol.* **3**, 504–510. (doi:10.1038/nrmicro1163)

Fauquet, C. M. 1999 Taxonomy, classification and nomenclature of viruses. In *Encyclopedia of virology* (eds A. Granoff & R. G. Webster), pp. 1730–1756, San Diego, CA: Academic Press.

Filée, J. & Forterre, P. 2005 Viral proteins functioning in organelles: a cryptic origin? *Trends Microbiol.* **13**, 510–513. (doi:10.1016/j.tim.2005.08.012)

Filée, J., Tétart, F., Suttle, C. A. & Krisch, H. M. 2005 Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc. Natl Acad. Sci. USA* **102**, 12 471–12 476. (doi:10.1073/pnas.0503404102)

Filée, J., Bapteste, E., Susko, E. & Krisch, H. M. 2006 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.* **23**, 1688–1696. (doi:10.1093/molbev/msl036)

Fokine, A. *et al.* 2005 Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc. Natl Acad. Sci. USA* **102**, 7163–7168. (doi:10.1073/pnas.0502164102)

Forterre, P. 1999 Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. *Mol. Microbiol.* **33**, 457–465. (doi:10.1046/j.1365-2958.1999.01497.x)

Forterre, P. 2002 The origin of DNA genomes and DNA replication proteins. *Curr. Opin. Microbiol.* **5**, 525–532. (doi:10.1016/S1369-5274(02)00360-0)

Forterre, P. 2006 Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. *Proc. Natl Acad. Sci. USA* **103**, 3669–3674. (doi:10.1073/pnas.0510333103)

Forterre, P. 2009 Origin and evolution of viruses. *Encyclopedia of Microbiology.* Oxford, UK: Elsevier.

Forterre, P. & Gribaldo, S. 2007 The origin of modern terrestrial life. *HFSP J.* **1**, 156–168. (doi:10.2976/1.2759103)

Häring, M., Peng, X., Brügger, K., Rachel, R., Stetter, K. O., Garrett, R. A. & Prangishvili, D. 2004 Morphology and genome organization of virus PSV of the hyperthermophilic archaeal genera *Pyrobaculum* and *Thermoproteus*: a novel virus family, the Globuloviridae. *Virology* **323**, 233–242. (doi:10.1016/j.virol.2004.03.002)

Häring, M., Vestergaard, G., Rachel, R., Chen, L., Garrett, R. A. & Prangishvili, D. 2005 Independent virus development outside a host. *Nature* **436**, 1101–1102. (doi:10.1038/4361101a)

Hendrix, R. W. & Casjens, S. R. 2008 The role of bacteriophages in the generation and spread of bacterial patogens. In *Horizontal gene transfer in the evolution of pathogenesis* (eds M. Hensel & H. Schmidt). New York, NY: Cambridge University Press.

Hendrix, R. W., Lawrence, J. G., Hatfull, G. H. & Casjens, S. 2000 The origins and ongoing evolution of viruses. *Trends Microbiol.* **8**, 504–508. (doi:10.1016/S0966-842X(00)01863-1)

Jürgens, K. & Massana, R. 2008 Protistan grazing on marine bacterioplankton. In *Microbial ecology of the oceans* (ed. D. L. Kirchman), pp. 383–442. Hoboken, NJ: Wiley-Blackwell.

Juhala, R. J., Ford, M. E., Duda, R. L., Youlton, A., Hatfull, G. & Hendrix, R. W. 2000 Genomic sequences of bacteriophage HK97 and HK022: perversive genetic mosaicism in the lambdoid bacteriophages. *J. Mol. Biol.* **299**, 27–51. (doi:10.1006/jmbi.2000.3729)

Kanamaru, S., Leiman, P. G., Kostyuchenko, V. A., Chipman, P. R., Mesyanzhinov, V. V., Arisaka, F. & Rossmann, M. G. 2002 Structure of the cell-puncturing device of bacteriophage T4. *Nature* **415**, 553–557. (doi:10.1038/415553a)

Khayat, R., Tang, L., Larson, E. T., Lawrence, C. M., Young, M. & Johnson, J. E. 2005 Structure of an archaeal virus capsid protein reveals a common ancestry to eukaryotic and bacterial viruses. *Proc. Natl Acad. Sci. USA* **102**, 18 944–18 949. (doi:10.1073/pnas.0506383102)

Koonin, E. V., Senkevich, T. G. & Dolja, V. V. 2006 The ancient Virus World and evolution of cells. *BMC Biol. Direct* **1**, e29.

La Scola, B., Desnues, C., Pagnier, I. *et al.* 2008 The virophage as a unique parasite of the giant mimivirus. *Nature* **455**, 100–104. (doi:10.1038/nature07218)

Lawrence, J. G., Hendrix, R. W. & Casjens, S. 2001 Where are the pseudogenes in bacterial genomes? *Trends Microbiol.* **9**, 535–540. (doi:10.1016/S0966-842X(01)02198-9)

Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. 2008 Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* **25**, 762–777. (doi:10.1093/molbev/msn023)

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. 2005 Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**, 86–89. (doi:10.1038/nature04111)

Lucchini, S., Desiere, F. & Brüssow, H. 1999 Comparative genomics of *Streptococcus thermophilus* phage species supports a modular evolution theory. *J. Virol.* **73**, 8647–8656.

May, R. M. 1988 How many species are there on earth? *Science* **241**, 1441–1449. (doi:10.1126/science.241.4872.1441)

McGeoch, A. & Bell, S. D. 2005 Eukaryotic/archaeal primase and MCM proteins encoded in a bacteriophage genome. *Cell* **120**, 167–168. (doi:10.1016/j.cell.2004.11.031)

Mi, S. *et al.* 2000 Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**, 785–789.

Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T. & Rüger, W. 2003 Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.* **67**, 86–156. (doi:10.1128/MMBR.67.1.86-156.2003)

Monier, A., Claverie, J.-M. & Ogata, H. 2008 Taxonomic distribution of large DNA viruses in the sea. *Genome Biol.* **9**, R106. (doi:10.1186/gb-2008-9-7-r106)

Moreira, D. & Brochier-Armanet, C. 2008 Giant viruses, giant chimeras: the multiple evolutionary histories of mimivirus genes. *BMC Evol. Biol.* **8**, e12.

Nandhagopal, N. *et al.* 2002 The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus. *Proc. Natl Acad. Sci. USA* **99**, 14 758–14 763. (doi:10.1073/pnas.232580699)

Owens, R. A. 1999 Viroids. In *Encyclopedia of virology* (eds A. Granoff & R. G. Webster), pp. 1928–1937, San Diego, CA: Academic Press.

Padian, K. 2008 Darwin's enduring legacy. *Nature* **451**, 632–634. (doi:10.1038/451632a)

Pearson, H. 2008 'Virophage' suggests viruses are alive. *Nature* **454**, 677. (doi:10.1038/454677a)

Pedulla, M. L. *et al.* 2003 Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**, 171–182. (doi:10.1016/S0092-8674(03)00233-2)

Peng, X., Basta, T., Häring, M., Garrett, R. A. & Prangishvili, D. 2007 Genome of the *Acidianus* bottle-shaped virus and insights into the replication and packaging mechanism. *Virology* **364**, 237–243. (doi:10.1016/j.virol.2007.03.005)

Pfister, P., Wasserfallen, A., Stettler, R. & Leisinger, T. 1998 Molecular analysis of *Methanobacterium* phage psiM2. *Mol. Microbiol.* **30**, 233–244. (doi:10.1046/j.1365-2958.1998.01073.x)

Prangishvili, D., Forterre, P. & Garrett, R. A. 2006 Viruses of Archaea: a unifying view. *Nat. Rev. Microbiol.* **4**, 837–848. (doi:10.1038/nrmicro1527)

Prusiner, S. B. 1996 Prions. In *Fields virology* (eds B. N. Fields, D. M. Knipe & P. M. Howley), pp. 2901–2950, Philadelphia, PA: Lippincott- Raven Publishers.

Raoult, D. & Forterre, P. 2008 Redefining viruses: lessons from mimivirus. *Nat. Rev. Microbiol.* **6**, 315–319. (doi:10.1038/nrmicro1858)

Raoult, D. *et al.* 2004 The 1.2-magabase genome sequence of mimivirus. *Science* **306**, 1344–1350. (doi:10.1126/science.1101485)

Rohwer, F. & Edwards, R. 2002 The Phage Proteome Tree: a genome-based taxonomy for phage. *J. Bacteriol.* **184**, 4529–4535. (doi:10.1128/JB.184.16.4529-4535.2002)

Roux, L. 1999 Defective interfering viruses. In *Encyclopedia of virology* (eds A. Granoff & R. G. Webster), pp. 371–375, San Diego, CA: Academic Press.

Schulte, A. M., Lai, S., Kurtz, A., Czubayko, F., Riegel, A. T. & Wellstein, A. 1996 Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus. *Proc. Natl Acad. Sci. USA* **93**, 14 759–14 764. (doi:10.1073/pnas.93.25.14759)

Susskind, M. M. & Botstein, D. 1978 Molecular genetics of bacteriophage P22. *Microbiol. Rev.* **42**, 385–413.

Suzan-Monti, M., La Scola, B., Barrassi, L., Espinosa, L. & Raoult, D. 2007 Ultrastructural characterization of the giant volcano-like virus factory of *Acanthamoeba polyphaga* mimivirus. *PLOS One* **3**, e328.

Takahashi, I. & Marmur, J. 1963 Replacement of thymidylic acid by deoxyuridylic acid in the deoxyribonucleic acid of a transducing phage from *Bacillus subtilis*. *Nature* **197**, 794–795. (doi:10.1038/197794a0)

Talianski, M. E. & Palukaitis, P. F. 1999 Satellite RNAs and satellite viruses. In *Encyclopedia of Virology* (eds A. Granoff & R. G. Webster), pp. 1607–1615. San Diego, CA: Academic Press.

Tétart, F., Desplats, C., Kutateladze, M., Monod, C., Ackermann, H. W. & Krisch, H. M. 2001 Phylogeny of the major head and tail genes of the wide-ranging T4-type bacteriophages. *J. Bacteriol.* **183**, 358–366. (doi:10.1128/JB.183.1.358-366.2001)

Thingstad, T. F. & Lignell, R. 1997 Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat. Microb. Ecol.* **13**, 19–27. (doi:10.3354/ame013019)

Villareal, L. P. & DeFilippis, V. R. 2000 A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. *J. Virol.* **74**, 7079–7084. (doi:10.1128/JVI.74.15.7079-7084.2000)

Woese, C. 1998 The universal ancestor. *Proc. Natl Acad. Sci. USA* **95**, 6854–6859. (doi:10.1073/pnas.95.12.6854)

Wommack, K. E. & Colwell, R. 2000 Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114. (doi:10.1128/MMBR.64.1.69-114.2000)

Zuber, S., Ngom-Bru, C., Barretto, C., Bruttin, A., Brüssow, H. & Denou, E. 2007 Genome analysis of phage JS98 defines a fourth major subgroup of T4-like phages in *Escherichia coli*. *J. Bacteriol.* **189**, 8206–8214.