Review TRENDS in Microbiology Vol.12 No.10 October 2004

The evolution of large DNA viruses: combining genomic information of viruses and their hosts

Laura A. Shackelton and Edward C. Holmes

Department of Zoology, Tinbergen Building, University of Oxford, South Parks Road, Oxford, UK OX1 3PS

Research describing the patterns and processes of evolution in large DNA viruses has been hampered by low levels of sequence similarity between virus families and the absence of a viral fossil record. By analysing patterns of genome organization, the similarities among host and viral genes, and virus and host phylogenies it is now possible to show that DNA viruses have undergone intergenomic lateral gene transfer and intragenomic gene duplication during their evolution. Many viral proteins are also homologous to cellular proteins, suggesting extensive host gene capture. This new, combined approach to studying viral evolution enhances our ability to infer the origins of virus and host genes as well as the complex evolutionary history of large DNA viruses.

With the increasing availability of genome sequence data and improvements in methods of phylogenetic analysis, the evolutionary relationships within many families of DNA viruses have been established [1–3]. Yet, although we have a good understanding of their relatively recent evolution, accurately reconstructing the phylogenetic history of DNA viruses over greater time periods, or their large-scale patterns of genome evolution, has proven to be a far greater challenge. In particular, the evolutionary relationships among host and viral genes are difficult to determine, although central to understanding the emergence of living systems.

Although a definitive picture of viral origins is likely to remain elusive for some time (Box 1), the answers to many other questions surrounding DNA virus evolution are now within our grasp. Conventional sequence-based phylogenetic analysis is often difficult for distantly related taxa, such as families of DNA viruses, because so many nucleotide and amino acid changes have accumulated that practically all phylogenetic resolution is lost. Although large DNA viruses are less prone to error than RNA viruses (which use RNA polymerases without known repair activity), the estimated substitution rate for DNA viruses like herpes simplex virus type 1 is $\sim 3.5 \times 10^{-8}$ substitutions per site per year [4]. This is an order of magnitude greater than the evolutionary rate of mammalian nuclear genes [5], so that methods that do not rely on

Corresponding author: Edward C. Holmes (Edward.Holmes@zoo.ox.ac.uk).
Available online 26 August 2004

Box 1. The origin of DNA viruses

Numerous theories exist to explain the origins of DNA viruses. Some propose that the ancestors of modern viruses appeared before cells [45,47]. For example, it has been suggested that 'micelles', which had the tendency to trap and protect nucleic acids, formed from prebiotic material [47]. As the number of trapped sequences increased and ribozyme activity evolved, the micelles became protoviruses, developed proteins and frequently fused, sharing information. Eventually, some evolved into protocells. It has also been proposed that viruses were the first to use DNA to protect their genetic material from degradation by the RNA-encoded enzymes of RNA cells [48]. Another theory is that viruses arose through the reductive evolution of pathogenic microorganisms or primitive cells living before the last universal common ancestor (LUCA) [49]. The cells would have reduced the size of their genome substantially, keeping only those genes needed for persistence as parasites, a process in part recapitulated by endosymbiotic bacteria [50].

An alternative theory for the origin of DNA viruses is that they are pieces of cellular DNA that 'escaped' from cellular control. This DNA, which probably encoded replication machinery, then gained regulatory and packaging genes, enabling movement between cells [31]. A cellular protein might also have evolved to form an icosahedral shell. The capsid genes would then be replicated by the cellular host and accumulate additional beneficial genes [51]. In turn, this genomic flexibility of viruses and their ability to transfer genes between cells would have enabled early cells to 'share successful evolutionary experiments' [51].

The absence of a viral fossil record makes it difficult to test these theories, so inferences are often made from indirect sources. Some clues come from the similarities between viruses that infect organisms from different domains of life. This diversity of hosts could mean that viruses diverged before the LUCA, then evolving in parallel with cells [34,45]. The similarities in capsid structure, replication mechanism and/or genome organization between some eukaryotic DNA viruses and phages, and some eukaryotic DNA viruses and archaeal DNA viruses, has also been taken as evidence of their antiquity [8,52].

More definitive information regarding the later evolution of DNA viruses comes from comparing the phylogenies of DNA viruses with those of their host species. A significant match between host and virus phylogenies is indicative of virus–host cospeciation. If cospeciation is observed, then the timescale of viral evolution can often be inferred, as it must have occurred over the same period as that of their hosts. This technique was used to infer approximate divergence times for mammalian herpesviruses over the last ~200 million years [53]. However, although these studies clearly reveal the relative age of some DNA viruses, we do not know whether cospeciation can be extended back to the early history of eukaryotes and prokaryotes, which might be expected if DNA viruses are remnants of the earliest life forms.

 0966-842X/$ - see front matter doi:10.1016/j.tim.2004.08.005

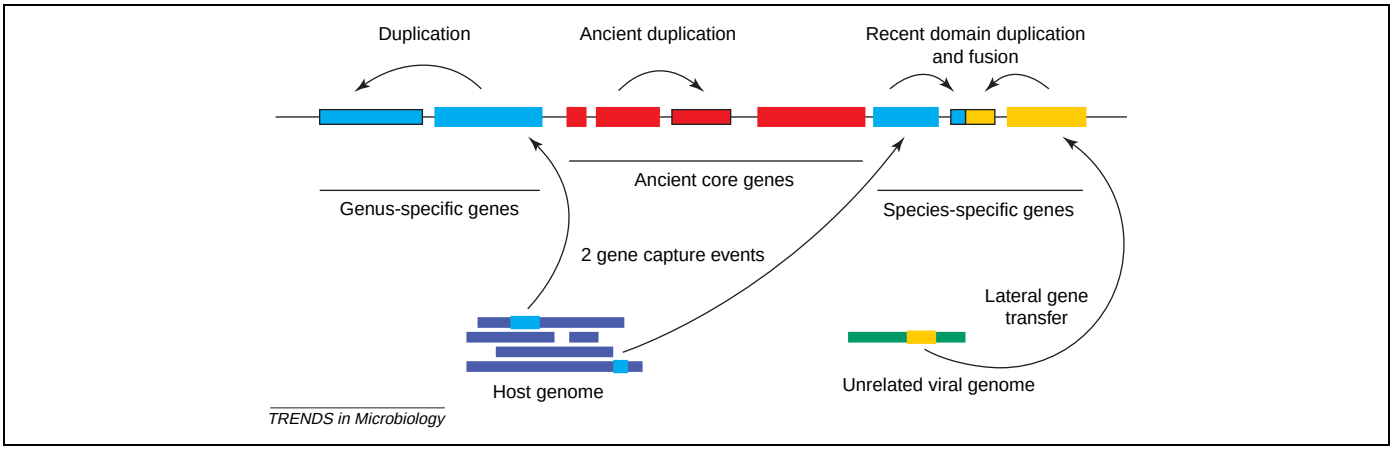


Figure 1. The possible routes through which large eukaryotic DNA viruses have acquired some of their hundreds of genes. Core viral genes that have been inherited vertically from an ancient viral ancestor are shown in red. Genes captured from the genome of a cellular host species are depicted in blue, and genes laterally transferred from other viruses are shown in yellow. Genes that have originated through intragenomic duplication are outlined in black.

sequence similarity are needed to reveal the distant evolutionary relationships among DNA viruses. One possibility is to use protein tertiary structure (where known), because this does not change as rapidly as the underlying amino acid sequence [6]. For example, adenoviruses and bacteriophage PRD1 share a unique major coat protein arrangement not found in other viruses [7]. However, the applicability of these methods is limited in the absence of realistic models of protein structure evolution.

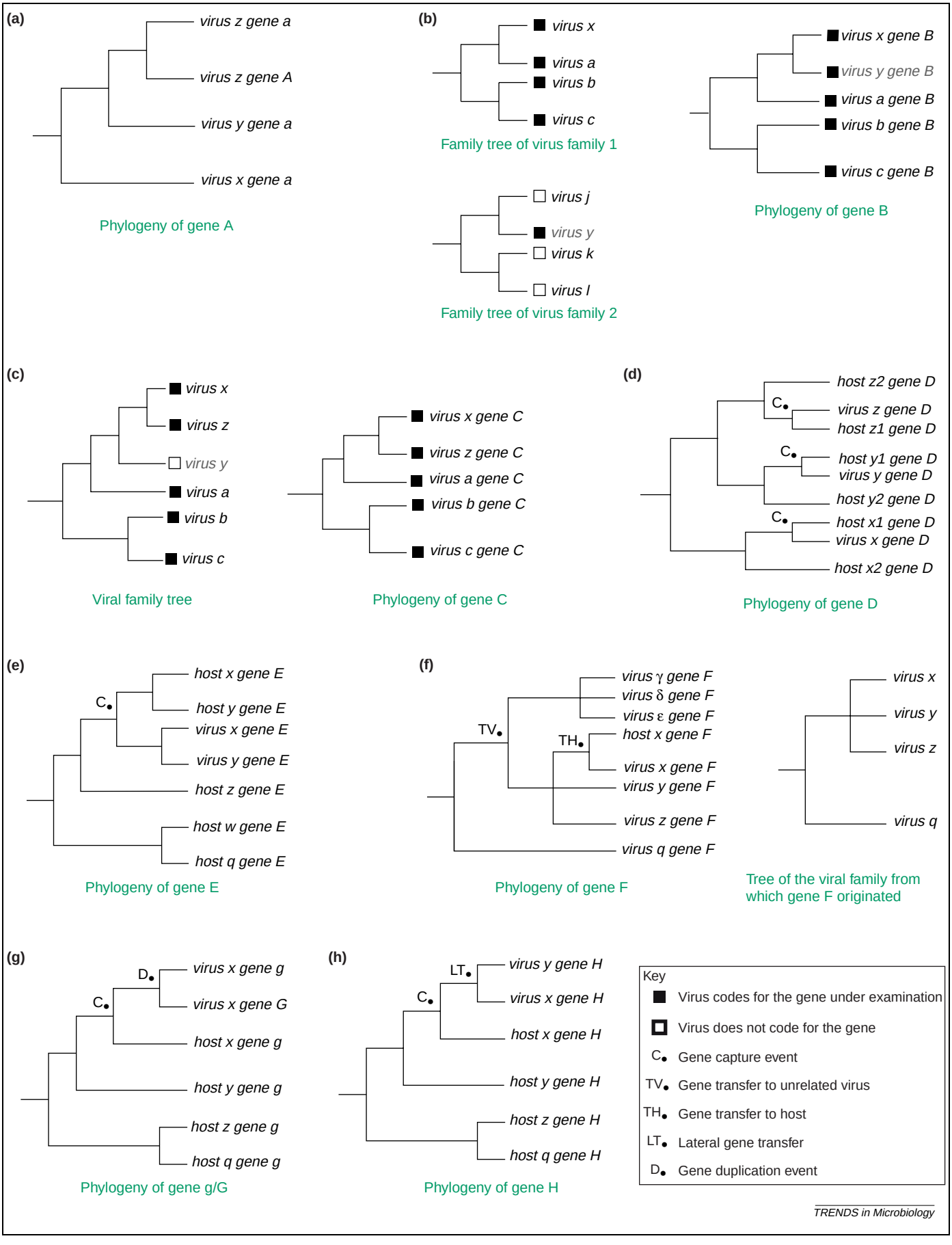
Another limitation of sequence-based phylogenies is that single gene trees often differ from one another because of lateral gene transfer, gene loss and gene duplication. Consequently, an accurate portrayal of viral phylogenetic relationships will only come from studies utilizing whole genomes, including both primary sequence comparisons and detailed analyses of genome organization. Such studies undertaken to date have revealed that phylogenies of baculoviruses and herpesviruses based on gene order, and phylogenies of baculoviruses based on gene content, agree with and add complementary information to those created using protein and whole-genome primary sequence data [3,8]. The goal for the near future is to extend these new approaches to reveal interfamily relationships.

As well as providing an accurate picture of phylogenetic relationships, comparisons among distantly related virus genomes give important insights into the basic mechanisms of genome evolution. Because even the genomes of viruses within the same family can differ in size and content, it is reasonable to assume that large DNA viruses have acquired new genes during their evolution. It is important to note here that we consider viruses with genomes over 100 kb and the intermediate-sized adenoviruses to be 'large' DNA viruses. Small DNA viruses, with genomes under 10 kb, probably have a different evolutionary history, are perhaps subject to additional constraints on genome sizes and therefore are not the subject of this review. Sources for novel genes in large DNA viruses include the duplication of the virus's own genes, lateral gene transfer between viruses, and host gene capture (Figures 1 and 2). Strikingly, the frequency of each of these fundamental evolutionary processes has not yet been determined, although their respective rates can be revealed through genomic comparisons and phylogenetic analyses, and such comparisons are being undertaken for bacteria [9] and eukaryotes [10].

The extent of sequence similarity between the proteins of large DNA viruses and those of cellular organisms is of special interest. Many of the cellular homologues of viral proteins are involved in either host immune function or nucleic acid metabolism and replication. With accurate phylogenetic analyses, we can reconstruct the history of these viral genes and determine their evolutionary relationships with cellular homologues. Botstein first proposed that phage genomes were constructed through recombination of genetic elements from several ancestral viruses, a process dubbed 'modular evolution' [11]. This was later expanded so that the sources of phage genes could include the genomes of other phages and their bacterial hosts [12]. We believe that a similar theory is applicable to all large DNA viruses: that they are collections of modules acquired from other viruses, host cells and intragenomic duplication events.

Gene transfer from cells to viruses

The similarities between many viral and cellular proteins suggest that viruses have 'captured' cellular genes, retaining those that increase their fitness. Phylogenies including both viral and cellular homologues will help to determine the direction of transfer (Figure 2). If, on the trees, the viral genes fall in clusters of cellular genes, they were probably captured by viruses. Conversely, if the cellular proteins fall into viral clusters, the cellular capture of viral genes is more likely. Examining the genomic position of genes and their composition can provide additional evidence for the direction of transfer. If the gene is located in different regions among related viruses or has mixed patterns of presence and absence, the case for exogenous acquisition is strengthened, although gene loss after duplication and lateral transfer must also be considered.


(a)
virus z gene a
virus z gene A
virus y gene a
virus x gene a
Phylogeny of gene A
(b)
virus x
virus a
virus b
virus c
Family tree of virus family 1
virus j
virus y
virus k
virus l
Family tree of virus family 2
virus x gene B
virus y gene B
virus a gene B
virus b gene B
virus c gene B
Phylogeny of gene B
(c)
virus x
virus z
virus y
virus a
virus b
virus c
Viral family tree
virus x gene C
virus z gene C
virus a gene C
virus b gene C
virus c gene C
Phylogeny of gene C
(d)
host z2 gene D
C
virus z gene D
host z1 gene D
C
host y1 gene D
virus y gene D
host y2 gene D
C
host x1 gene D
virus x gene D
host x2 gene D
Phylogeny of gene D
(e)
host x gene E
C
host y gene E
virus x gene E
virus y gene E
host z gene E
host w gene E
host q gene E
Phylogeny of gene E
(f)
virus γ gene F
virus δ gene F
virus ε gene F
TV
TH
host x gene F
virus x gene F
virus y gene F
virus z gene F
virus q gene F
Phylogeny of gene F
virus x
virus y
virus z
virus q
Tree of the viral family from which gene F originated
(g)
D
virus x gene g
C
virus x gene G
host x gene g
host y gene g
host z gene g
host q gene g
Phylogeny of gene g/G
(h)
LT
virus y gene H
C
virus x gene H
host x gene H
host y gene H
host z gene H
host q gene H
Phylogeny of gene H
Key
Virus codes for the gene under examination
Virus does not code for the gene
C Gene capture event
TV Gene transfer to unrelated virus
TH Gene transfer to host
LT Lateral gene transfer
D Gene duplication event
TRENDS in Microbiology

Genes involved in immune function

Large DNA viruses encode numerous proteins that are homologous to cellular proteins involved in immune function. Many of these viral homologues modulate the host's immune system, often by mimicking or interfering with their corresponding cellular proteins. For example, among the human herpesviruses are viral proteins homologous to cytokines and chemokines, receptors of these molecules, apoptosis-related proteins, major histocompatibility complex proteins, complement system proteins, Fc receptors and immunoglobulin superfamily proteins [13]. Several large DNA viruses encode interleukin-10 (IL-10) homologues, often with over 80% identity to cellular IL-10 molecules [14]. Some of these viral IL-10 molecules have been shown to suppress the cell-mediated immune response [15].

Other viral proteins are homologous to cellular apoptosis inhibitors. For example, γ-herpesvirus proteins homologous to cellular bcl2 and FLIP block mitochondrial-induced and death-receptor-induced apoptosis, respectively. Interestingly, whereas cellular bcl2 can be cleaved by caspases and converted to an apoptosis inducer, the viral homologues lack the caspase target region [13]. Other viruses encode proteins with regions homologous to cellular proteins involved in regulating the expression of immunity genes. For example, African swine fever virus (ASFV) encodes a protein with a region homologous to cellular IκB. IκB inhibits cellular NF-κB factors – the transcription factors for several cytokines, chemokines and cytokine receptors [16,17].

Viral proteins might be homologous to their cellular counterparts throughout the protein or only to specific domains. This pattern of conservation can be indicative of the role that viral proteins have in immune modulation. For example, some poxvirus tumour necrosis factor receptors, although homologous to their cellular counterparts in the N-terminal ligand-binding region, differ in their C-terminal region and encode a signal peptide. This results in their secretion, enabling them competitively to bind the immune-signalling ligands in the extracellular space [18].

To test the assumption that viruses acquired these immune modulation genes from cells, phylogenetic analyses have been undertaken [19]. A tree of mammalian IL-10 sequences and homologues from Epstein–Barr virus, equine herpesvirus, baboon herpesvirus 2 and poxvirus Orf suggested that three separate gene acquisition events by viruses occurred during mammalian evolution [20]. Likewise, the baculovirus inhibitor of apoptosis homologues has seemingly originated from multiple cellular capture events [21]. However, a phylogeny including the vaccinia and variola virus homologues to CD47 – an Ig protein involved in T-cell activation [22] – shows the viral proteins falling together outside the rat and human cluster. This suggests a single capture event by a poxviral ancestor before primate and rodent orders diverged, although a larger sample of taxa are required to confirm this [20]. Finally, a recent case of host gene capture was observed in a mammalian 2β-1,6-*N*-acetylglucosaminyltransferase-mucin type protein and its only known viral homologue, the bovine herpesvirus 4 Bo17 protein [23]. A phylogenetic analysis revealed that the gene was captured by the virus after the split between African buffalo and cattle at ~1.5 million years ago. Since that time the viral gene, which is thought to reduce the ability of immune cells to target infected cells, has evolved 20 to 30 times faster than its cellular counterpart [23]. This study also highlighted the potential problems of combining genes that are both divergent and that have very different rates of substitution. Foremost among these is long-branch attraction, when lineages that have experienced elevated rates of substitution cluster together regardless of the true topology, thereby causing phylogenetic error. To avoid this problem, it is advisable to infer trees using a variety of models of evolution and outgroups.

Genes involved in nucleic acid metabolism and replication

Large DNA viruses also encode many nucleic acid metabolism and replication enzymes homologous to cellular proteins. For example, iridoviruses encode homologues of cellular DNA polymerases, ribonucleotide reductase (RR) subunits, DNA-dependent RNA polymerase II subunits, DNA topoisomerase II, thymidylate synthase, helicases and exoribonuclease [24]. Similarly, all human herpesviruses have homologues to cellular DNA polymerase, helicase and primase, uracil-DNA glycosylase and the RR large subunit [25]. Because DNA replication is obviously essential for viruses, it is likely that DNA polymerases were present in their earliest ancestors. This is supported by the observation that phylogenies constructed with viral polymerase proteins generally match current viral taxonomic classifications. However, phylogenies of viral RRs (which increase the availability of deoxyribonucleotides by catalysing the reduction of ribonucleotides) do not agree with accepted viral taxonomic classifications, indicating that they were acquired independently by different virus species from their individual hosts [24].

Strikingly, few cellular homologues for structural virion capsid proteins have been detected. If these genes were indeed captured from cells, viruses are likely to have

Figure 2. Model phylogenetic tree topologies indicating the source of viral genes. **(a)** If the viral gene groups with other genes in the same viral genome, it probably arose through gene duplication. **(b)** If the viral gene clusters with genes from a different viral family, lateral transfer from one virus to another probably occurred. **(c)** If the viral gene clusters with genes from the same family, it is probably ancestral. However, when most species in a family possess a gene and others do not, both intrafamily lateral transfer and ancestral origin followed by gene loss events (as shown here) must be considered. The case for ancestral origin is supported by comparing the viral gene tree to the accepted viral family phylogeny. **(d,e)** If the viral gene groups with cellular genes, host capture is the most probable scenario. As shown in (d), when the gene is in different positions in the viral genomes, and the viral gene tree does not match the viral species phylogeny, the case for multiple capture events is supported. Alternatively, as shown in (e), when the gene is in the same position in the viral genomes, and the viral gene tree matches the viral species phylogeny, the case for capture by an ancestral virus is supported. **(f)** If the viral gene appears at the base of a phylogenetic cluster, it has an ancient origin and was passed onto viral and/or cellular organisms. **(g,h)** Finally, it is possible that a combination of the above scenarios has occurred in the long history of a viral gene, as shown in the following examples. (g) Gene g was captured by virus x from host x and duplicated. The duplicated gene gave rise to gene G. (h) Gene H was captured by virus x from host x and subsequently transferred to virus y.

employed them for such unique functions that they might have evolved too far from their cellular counterparts to detect any sequence similarity [26]. Hence, we might only be able to detect those viral homologues whose evolution has been highly constrained because they perform essential functions, or because their benefit to the virus stems from their similarity to cellular proteins.

Finally, it is important to determine the mechanisms by which viruses capture cellular genes. One possibility is direct recombination, although there are two major objections to this process; first, that many large DNA viruses do not enter the cell nucleus, so that cellular and viral genomes are never in close proximity, and, second, that most viral homologues are intronless [19,27]. Although direct recombination and subsequent deletion of introns could explain the latter, it does not provide a solution to the former objection. Alternatively, intronless cDNA copies of spliced cellular mRNAs might have been inserted into the viral genome [28], although the origin of the necessary reverse transcriptase is problematic. Possibilities include infectious retroviruses also present in the cell (which have been observed in the genomes of herpesviruses [29] and fowlpox virus [30]) or one of the cell's endogenous retroviruses [14].

Gene transfer from viruses to cells?

The phylogenies of DNA polymerases from cellular organisms are often incompatible with traditional species phylogenies, particularly as the cellular polymerases are frequently intermixed with viral ones. Indeed, there have been suggestions that DNA polymerase trees of cellular organisms cannot be connected without including viral polymerase sequences [31]. Phylogenies also support the theory that DNA polymerase genes have been transferred between viruses and cells, although, crucially, these trees are often unrooted, so the direction of the transfers is unknown [32,33]. One example concerns the B-family of DNA polymerases, which includes the eukaryotic polymerases α and δ, as well as the polymerases of large DNA viruses such as the phycodnaviruses (which infect microalgae), herpesviruses, ascoviruses and iridoviruses [33]. Phylogenetic analyses positioned the phycodnavirus polymerases near the base of the eukaryotic polymerase δ cluster [31,32,34]. Also of note is that the B1 polymerase of the archaeon *Halobacterium salinarum* clusters with haloviral DNA polymerases and not with the other archaeal polymerases. This provides convincing evidence that the *H. salinarum* polymerase has a viral origin [33]. Figure 3 summarizes the current evidence regarding the evolutionary relationships among cellular and viral B-family polymerases.

Viruses also encode A-family polymerases. Phylogenies suggest that this family appeared first in bacteria or bacteriophages and was later transferred to other viruses. Endocytosed bacteria which gave rise to mitochondria might also have transferred A-family polymerases to eukaryotic genomes [33]. The displacement of nuclear genes by organelle genes has occurred frequently during the course of evolution [35]. Once the gene was transferred, the mitochondrial polymerase seems to have been replaced by a T3/T7 phage-like polymerase, suggesting phage-to-cell gene transfer [33]. The current mitochondrial replicative helicase DnaB might also have T3/T7 phage origins [34].

Despite some important uncertainties in the phylogenetic analyses, these studies strongly support the idea that cells and viruses have contributed genetically to each other's evolution (Box 2).

Consequences of viral–cellular gene exchange

Host immune protein homologues are often specific to just a few viruses and are found at disparate places in viral phylogenies. Thus, many probably have a role in the adaptation of viruses to their particular hosts [25]. Given the ubiquity and usefulness of immune homologues in modulating host defences, it is curious that RNA viruses have not captured similar proteins. A plausible explanation is the constraint on RNA virus genome size, itself probably a function of the low fidelity of RNA-dependent RNA polymerases and the subsequent error threshold [36].

DNA virus homologues of cellular proteins are involved in intricate cellular modulation strategies. By controlling cellular inflammatory and immune responses, these viruses often prevent both their own elimination and damage to the host. This might even result in a benign latent infection, as in the case of herpesviruses. By contrast, RNA viruses rely on their high mutation rate to escape host immune pressure [37]. However, at present we do not know which factors contribute to the difference in gene acquisition tendencies between large and small DNA viruses. It is possible that capsid size poses a

Box 2. Viruses as the origin of the eukaryotic nucleus?

Could an ancient virus – related to modern poxviruses or African swine fever virus - have given rise to the eukaryotic nucleus through a symbiotic relationship with ancient archaea [54,55]? The basis of this theory is that DNA viruses and eukaryotes share such features as linear chromosomes, mRNA capping and a separation of transcription from translation, and that phylogenetic trees often depict eukaryotes evolving from an archaeal-like ancestor after the divergence of the archaeal and bacterial domains [56]. Consequently, it has been proposed that the eukaryotic cell nucleus was derived from a virus [43,55,57], and that the nuclear envelope is a remnant of the viral envelope [54]. Other genes, such as those necessary for translation, would have been captured by the new nucleus from the cytoplasmic archaeal genome. The later engulfment of a eubacterium, and thus the acquisition of many of its genes [58], would have been possible once the viral membrane fusion proteins had been incorporated into the cellular membrane. Subsequent loss of the remaining cytoplasmic prokaryotic genetic material would have resulted in a eukaryotic cell [55].

Phylogenetic analysis is crucial to testing this theory. To replicate their genome, eukaryotes use a complex of polymerases α and either δ or ε, or both. Whereas viruses such as the herpesviruses and phycodnaviruses use a δ-like polymerase, two orthopoxviruses – vaccinia and variola – employ an α-like polymerase. Phylogenies show that the α-like polymerases of eukaryotes and these orthopoxviruses are related, suggesting to some that the α- and δ-polymerases originated from the viral and archaeal ancestors, respectively [54]. A phylogeny of the guanylyltransferase domains of mRNA-capping enzymes in eukaryotes and some large DNA viruses also indicates that they share a common ancestor [55]. However, such analyses are often inconclusive because of uncertainty over the position of the root and the sample of taxa analysed.

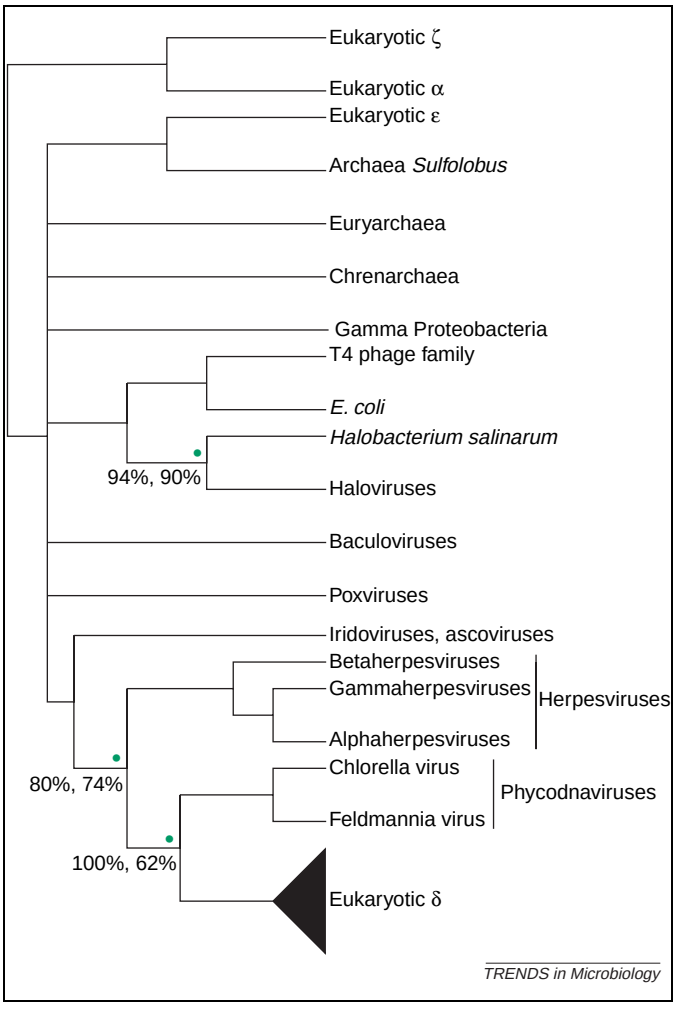


Figure 3. Schematic representation of the phylogenetic relationships of B-family polymerases. This phylogeny is one probable summary of analyses presented by several authors [32–34]. Nodes of special interest for this discussion are marked with dots. Bootstrap values from at least two separate analyses are shown for well-supported nodes.

constraint but this begs the question of how large DNA viruses developed increased capsid capacities. It is also likely that some small single-stranded DNA viruses have a mutation rate approaching that of RNA viruses [38], so that they are subject to an error threshold which might in turn limit genome size. Alternatively, the gene order in small DNA viruses is often temporally or functionally structured, suggesting that this order cannot be disrupted, and the presence of overlapping reading frames might further limit the addition of exogenous genes.

Many of the virally encoded immune homologues have consequences for the host beyond the maintenance of the viral infection. For example, according to the 'molecular mimicry' hypothesis, the host's immune system recognizes the viral homologue as foreign, therein stimulating an immune reaction to the viral protein and, consequently, an autoimmune reaction to the cellular homologue [39,40]. Although there are clear associations between certain infections and specific autoimmune diseases, the definitive mechanism by which the pathogens induce autoimmune reactions has not been established [40,41]. Other viral homologues of cellular proteins manipulate patterns of cellular growth. For example, Epstein–Barr virus encodes a guanosine nucleotide-binding protein-coupled receptor which is similar to cellular CXC chemokine receptors except that it does not need a ligand to be activated. Expression of the viral receptor causes cellular transformation and secretion of vascular growth factor, which leads to angiogenesis [42].

The origins of novel viral genes

As well as vertical transmission and gene capture from host cells, large DNA viruses can also acquire genes through lateral gene transfer between viruses and gene duplication.

Viral genes can be classified as either core genes or genus- or species-specific genes. Core genes are conserved among all family members and were presumably present in that family's last common ancestor. By contrast, genus- and species-specific genes were acquired subsequent to the divergence of the genera or species in question. However, some genes that appear to be genus- or species-specific might be ancestral and then lost in particular taxa. Examining the genomic position of genes might help to resolve this problem. For example, viral homologues of human ORF12 are found in human herpesvirus (HHV)-8, murine herpesvirus (MHV)-68 and herpesvirus saimiri (HVS)-2. However, there are no such homologues in ateline herpesvirus 3 and *Macaca mulatta* rhadinovirus – related viruses with different primate hosts. Because the gene is in the same position in the genomes of HHV-8, MHV-68 and HVS-2, gene loss seems to be the most probable explanation of this pattern [19]. In many large DNA viruses, core genes are located near the centre of the genome whereas genus- and species-specific genes are located near the ends (Figure 1). The former are usually involved in general functions like replication and encapsidation, whereas the latter provide unique host interactions and are most likely to have been acquired more recently [1,2,9].

Gene duplication has presumably taken place among both genus- and species-specific genes and core genes of large DNA viruses, particularly as they often fall into related gene pairs and multigene families [43,44]. Following duplication, gene sequences diverge and can assume different functions. As a case in point, many genes in the E4 region of mastadenoviruses and atadenoviruses might have arisen from multiple duplications of only two genes: 34K and either an F-box or a dUTPase gene [1]. By contrast, the dUTPase gene of herpesviruses underwent intragenic duplication, resulting in the duplication of essential domains [44].

Lateral transfer among viruses – which can result from recombination between viruses co-infecting the same cell – is an additional source of genes. In accordance with the original modular evolution theory, it is possible that novel viruses arise entirely from the combination of genes from established viral lineages [45]. Although lateral transfer between bacteriophages is frequently observed, the extent to which it has occurred in the evolution of eukaryotic DNA viruses is less certain. Detecting lateral transfer between viruses in the same family is again complicated by the possibility of ancestral acquisition followed by gene loss in some members of the family. However, detection of

closely related homologues in genera from two different families (such as those found in certain *Adenoviridae* and *Herpesviridae* genera) supports the action of lateral transfer [1].

Perspectives: the power of comparative genomics

The study of DNA virus evolution is set to enter an exciting new phase. The increasing availability of genome sequence data and new methods for comparative genomics and phylogenetics, as well as an increasing number of viral protein structures, will doubtless provide a more comprehensive picture of viral evolution. In particular, gene order and gene content are now being used to infer wide-ranging phylogenetic trees with the use of neighbour pair and breakpoint distance analysis, and matrices recording gene absence or presence, respectively [3,46]. This is clearly a major area of growth in evolutionary bioinformatics and might provide important new insights into the distant phylogenetic relationship among DNA viruses.

By conducting searches for proteins similar to viral proteins and using significant matches to create protein phylogenies, we can determine the sources of viral genes. This will be particularly important in testing the hypothesis that large DNA viruses are a collection of genes arising from host capture, lateral transfer between viruses and duplication events. It will also enable us to determine the frequency of each of these events, help to explain why DNA viruses differ so substantially in size and clarify the role of viruses in the evolution of eukaryotic proteins. To do this effectively, however, homologue searches and phylogenetic analyses must include other proteins in the viral genome, the proteins of other viruses and proteins of cellular organisms. The position of the viral gene within the resulting tree should reveal its origin (Figure 2).

To accurately infer phylogenetic relationships among genes, and reconstruct these potentially complex evolutionary events, it is essential that we broaden our sequence sampling pool. Although viral phylogenies are informative, only by including both cellular and viral proteins can the direction of transfer be investigated. Many previous analyses have been limited to vertebrates and their viruses. To form a clearer picture of viral evolution, we must therefore search for viral homologues among all available sequences from eubacteria, archaea, eukaryotes and other viruses.

The transfer of genes, both between cells and viruses and among different viruses, along with the duplication and divergence of such genes, appears to have been frequent in large DNA viruses, representing an elaborate form of modular evolution. In these circumstances, the complicated evolutionary history of DNA viruses might be better represented by multiple phylogenies, or even net-like structures, than simple single-gene phylogenies. Only through the use of such 'genomic networks' will we accurately depict the complex history of DNA viruses.

Acknowledgements

We thank the Rhodes Trust and the Wellcome Trust for financial support, and a variety of reviewers for valuable comments.

References

1 Davison, A.J. *et al*. (2003) Genetic content and evolution of adenoviruses. *J. Gen. Virol.* 84, 2895–2908
2 Gubser, c. *et al*. (2004) Poxvirus genomes: a phylogenetic analysis. *J. Gen. Virol.* 85, 105–117
3 Herniou, E.A. *et al*. (2001) Use of whole genome sequence data to infer baculovirus phylogeny. *J. Virol.* 75, 8117–8126
4 Sakaoka, H. *et al*. (1994) Quantitative analysis of genomic polymorphism of herpes simplex virus type 1 strains from six countries: studies of molecular evolution and molecular epidemiology of the virus. *J. Gen. Virol.* 75, 513–527
5 Li, W. (1997) *Molecular Evolution*, Sinauer Associates, Inc.
6 Matthews, B.W. and Rossmann, M.G. (1985) Comparison of protein structures. *Methods Enzymol.* 115, 397–420
7 Bamford, D.H. *et al*. (2002) Evolution of viral structure. *Theor. Popul. Biol.* 61, 461–470
8 Hannenhalli, S. *et al*. (1995) Genome sequence comparison and scenarios for gene rearrangements: a test case. *Genomics* 30, 299–311
9 Koonin, E.V. *et al*. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* 55, 709–742
10 Koonin, E.V. *et al*. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* 5, R7
11 Botstein, D. (1980) A theory of modular evolution for bacteriophages. *Ann. N. Y. Acad. Sci.* 354, 484–490
12 Campbell, A. and Botstein, D. (1983) Evolution of the lambdoid phages. In *Lambda II* (Hendrix, R.W. *et al*., eds), Cold Spring Harbor Laboratory
13 Raftery, M. *et al*. (2000) Herpesvirus homologues of cellular genes. *Virus Genes* 21, 65–75
14 Fleming, S.B. *et al*. (2000) Sequence and functional analysis of a homolog of interleukin-10 encoded by the parapoxvirus orf virus. *Virus Genes* 21, 85–95
15 Spencer, J.V. *et al*. (2002) Potent immunosuppressive activities of cytomegalovirus-encoded interleukin-10. *J. Virol.* 76, 1285–1292
16 Powell, P.P. *et al*. (1996) An IκB homolog encoded by African swine fever virus provides a novel mechanism for downregulation of proinflammatory cytokine responses in host macrophages. *J. Virol.* 70, 8527–8533
17 Revilla, Y. *et al*. (1998) Inhibition of nuclear factor κB activation by a virus-encoded IκB-like protein. *J. Biol. Chem.* 273, 5405–5411
18 Hu, F.Q. *et al*. (1994) Cowpox virus contains two copies of an early gene encoding a soluble secreted form of the type II TNF receptor. *Virology* 204, 343–356
19 Bugert, J.J. and Darai, G. (2000) Poxvirus homologues of cellular genes. *Virus Genes* 21, 111–133
20 Hughes, A.L. (2002) Origin and evolution of viral interleukin-10 and other DNA virus genes with vertebrate homologues. *J. Mol. Evol.* 54, 90–101
21 Hughes, A.L. (2002) Evolution of inhibitors of apoptosis in baculoviruses and their insect hosts. *Infect. Genet. Evol.* 2, 3–10
22 Liu, Y. *et al*. (2002) Signal regulatory protein (SIRPalpha), a cellular ligand for CD47, regulates neutrophil transmigration. *J. Biol. Chem.* 277, 10028–10036
23 Markine-Goriaynoff, N. *et al*. (2003) The core 2β-1,6-*N*-acetylglucosaminyltransferase-mucin encoded by the bovine herpesvirus 4 was acquired from an ancestor of the African buffalo. *J. Virol.* 77, 1784–1792
24 Tidona, C.A. and Darai, G. (2000) Iridovirus homologues of cellular genes - implications for the molecular evolution of large DNA viruses. *Virus Genes* 21, 77–81
25 Holzerlandt, R. *et al*. (2002) Identification of new herpesvirus gene homologs in the human genome. *Genome Res.* 12, 1739–1748
26 McGeoch, D. and Davison, J. (1995) Origins of DNA viruses. In *Molecular Basis of Virus Evolution* (Gibbs, A. *et al*., eds), Cambridge University Press
27 Grabherr, R. *et al*. (1992) The DNA polymerase gene from chlorella viruses PBCV-1 and NY-2A contains an intron with nuclear splicing sequences. *Virology* 188, 721–731
28 Sekiguchi, J. and Shuman, S. (1997) Ligation of RNA-containing duplexes by vaccinia DNA ligase. *Biochemistry* 36, 9073–9079
29 Brunovskis, P. and Kung, H.J. (1995) Retrotransposition and herpesvirus evolution. *Virus Genes* 11, 259–270

30 Singh, P. *et al*. (2003) Reticuloendotheliosis virus sequences within the genomes of field strains of fowlpox virus display variability. *J. Virol.* 77, 5855–5862
31 Villarreal, L.P. (1999) DNA virus contribution to host evolution. In *Origin and Evolution of Viruses* (Domingo, E. *et al*., eds), Academic Press
32 Villarreal, L.P. and DeFillippis, V.R. (2000) A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. *J. Virol.* 74, 7079–7084
33 Filée, J. *et al*. (2002) Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J. Mol. Evol.* 54, 763–773
34 Filée, J. *et al*. (2003) The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. *Res. Microbiol.* 154, 237–243
35 Timmis, J.N. *et al*. (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* 5, 123–135
36 Holmes, E.C. (2003) Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol.* 11, 543–546
37 Chaston, T.B. and Lidbury, B.A. (2001) Genetic 'budget' of viruses and the cost to the infected host: a theory on the relationship between the genetic capacity of viruses, immune evasion, persistence and disease. *Immunol. Cell Biol.* 79, 62–66
38 Truyen, U. *et al*. (1995) Evolution of the feline-subgroup parvoviruses and the control of canine host-range *in vivo*. *J. Virol.* 69, 4702–4710
39 Fujinami, R.S. and Oldstone, M.B. (1985) Amino acid homology between the encephalitogenic site of myelin basic protein and virus: mechanism for autoimmunity. *Science* 230, 1043–1045
40 Benoist, C. and Mathis, D. (2001) Autoimmunity provoked by infection: how good is the case for T cell epitope mimicry? *Nat. Immunol.* 2, 797–801
41 Flodstrom-Tullberg, M. (2003) Viral infections: their elusive role in regulating susceptibility to autoimmune disease. *Microbes Infect.* 5, 911–921
42 Bais, C. *et al*. (1998) G-protein-coupled receptor of Kaposi's sarcoma-associated herpesvirus is a viral oncogene and angiogenesis activator. *Nature* 391, 86–89
43 Blasco, R. (1995) Evolution of poxviruses and African swine fever virus. In *Molecular Basis of Virus Evolution* (Gibbs, A. *et al*., eds), Cambridge University Press
44 McGeoch, D. and Davison, J. (1999) Molecular evolutionary history of the herpesviruses. In *Origin and Evolution of Viruses* (Domingo, E. *et al*., eds), Academic Press
45 Bamford, D.H. (2003) Do viruses form lineages across different domains of life? *Res. Microbiol.* 154, 231–236
46 Hyink, O. *et al*. (2002) Whole genome analysis of the Epiphyas postvittana nucleopolyhedrovirus. *J. Gen. Virol.* 83, 957–971
47 Krisch, H.M. (2003) The view from Les Treilles on the origins, evolution and diversity of viruses. *Res. Microbiol.* 154, 227–229
48 Forterre, P. (2002) The origin of DNA genomes and DNA replication proteins. *Curr. Opin. Microbiol.* 5, 525–532
49 Forterre, P. (2003) The great virus comeback - from an evolutionary perspective. *Res. Microbiol.* 154, 223–225
50 Moran, N.A. (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 108, 583–586
51 Hendrix, R.W. *et al*. (2000) The origins and ongoing evolution of viruses. *Trends Microbiol.* 8, 504–508
52 Peng, X. *et al*. (2001) Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses. *Virology* 291, 226–234
53 McGeoch, D.J. *et al*. (1995) Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. *J. Mol. Biol.* 247, 443–458
54 Takemura, M. (2001) Poxviruses and the origin of the eukaryotic nucleus. *J. Mol. Evol.* 52, 419–425
55 Bell, P.J. (2001) Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus? *J. Mol. Evol.* 53, 251–256
56 Woese, C.R. *et al*. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* 87, 4576–4579
57 Moss, B. (2001) Poxviridae: the viruses and their replication. In *Field's Virology* (Vol. 4) (Knipe, D. and Howley, P., eds), Lippincott Williams and Wilkins
58 Martin, W. and Muller, M. (1998) The hydrogen hypothesis for the first eukaryote. *Nature* 392, 37–41