Available online at www.sciencedirect.com
SCIENCE DIRECT
Virus Research 117 (2006) 156–184
Virus Research
www.elsevier.com/locate/virusres

Evolutionary genomics of nucleo-cytoplasmic large DNA viruses

Lakshminarayan M. Iyer, S. Balaji, Eugene V. Koonin, L. Aravind *

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Available online 21 February 2006

Abstract

A previous comparative-genomic study of large nuclear and cytoplasmic DNA viruses (NCLDVs) of eukaryotes revealed the monophyletic origin of four viral families: poxviruses, asfarviruses, iridoviruses, and phycodnaviruses [Iyer, L.M., Aravind, L., Koonin, E.V., 2001. Common origin of four diverse families of large eukaryotic DNA viruses. J. Virol. 75 (23), 11720–11734]. Here we update this analysis by including the recently sequenced giant genome of the mimiviruses and several additional genomes of iridoviruses, phycodnaviruses, and poxviruses. The parsimonious reconstruction of the gene complement of the ancestral NCLDV shows that it was a complex virus with at least 41 genes that encoded the replication machinery, up to four RNA polymerase subunits, at least three transcription factors, capping and polyadenylation enzymes, the DNA packaging apparatus, and structural components of an icosahedral capsid and the viral membrane. The phylogeny of the NCLDVs is reconstructed by cladistic analysis of the viral gene complements, and it is shown that the two principal lineages of NCLDVs are comprised of poxviruses grouped with asfarviruses and iridoviruses grouped with phycodnaviruses-mimiviruses. The phycodna-mimivirus grouping was strongly supported by several derived shared characters, which seemed to rule out the previously suggested basal position of the mimivirus [Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., Claverie, J.M. 2004. The 1.2-megabase genome sequence of Mimivirus. Science 306 (5700), 1344–1350]. These results indicate that the divergence of the major NCLDV families occurred at an early stage of evolution, prior to the divergence of the major eukaryotic lineages. It is shown that subsequent evolution of the NCLDV genomes involved lineage-specific expansion of paralogous gene families and acquisition of numerous genes via horizontal gene transfer from the eukaryotic hosts, other viruses, and bacteria (primarily, endosymbionts and parasites). Amongst the expansions, there are multiple families of predicted virus-specific signaling and regulatory domains. Most NCLDVs have also acquired large arrays of genes related to ubiquitin signaling, and the animal viruses in particular have independently evolved several defenses against apoptosis and immune response, including growth factors and potential inhibitors of cytokine signaling. The mimivirus displays an enormous array of genes of bacterial provenance, including a representative of a new class of predicted papain-like peptidases. It is further demonstrated that a significant number of genes found in NCLDVs also have homologs in bacteriophages, although a vertical relationship between the NCLDVs and a particular bacteriophage group could not be established. On the basis of these observations, two alternative scenarios for the origin of the NCLDVs and other groups of large DNA viruses of eukaryotes are considered. One of these scenarios posits an early assembly of an already large DNA virus precursor from which various large DNA viruses diverged through an ongoing process of displacement of the original genes by xenologous or non-orthologous genes from various sources. The second scenario posits convergent emergence, on multiple occasions, of large DNA viruses from small plasmid-like precursors through independent accretion of similar sets of genes due to strong selective pressures imposed by their life cycles and hosts.
Published by Elsevier B.V.

Keywords: DNA viruses; Evolution; Poxvirus; Capping enzyme; Triphosphatase; Iridovirus; Coccolithovirus; Phycodnavirus; Primase; Origin of viruses; DNA replication; Origin of DNA replication; Papain-like peptidase; Sugar metabolism; Ubiqutin signaling; Growth factors; Cytokine signaling; Apoptosis

1. Introduction

The origin(s) of viruses had been a topic of intense speculation and debate ever since their discovery (Gibbs et al., 1995; Koonin, 1992). With the first biochemical studies on viruses, it became clear that only two common features were shared by all viruses: (1) their obligate intracellular parasitism; and (2) their virion architecture comprised of a genomic nucleic acid, typically of a single type (either RNA or DNA), packaged into a protein capsid, which in some cases is further associated with outer or inner lipid membranes (Gibbs et al., 1995). Beyond these general features, viruses show tremendous diversity in every respect, including genome size and organization, capsid architecture, mechanisms of propagation, and interactions with host cells. Viruses infect organisms from all three superkingdoms of life (bacteria, archaea, and eukaryotes) and replicate in all known

* Corresponding author. Tel.: +1 301 594 2445; fax: +1 301 480 9241.
E-mail address: aravind@ncbi.nlm.nih.gov (L. Aravind).

0168-1702/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.virusres.2006.01.009

cell types (Wagner and Hewlett, 2003). The extreme diversity of viruses suggests that they must have had multiple evolutionary origins, and the common features observed in all viruses reflect convergences emerging from adaptations to intracellular parasitism. The times and the modes of origins of the various types of viruses and their relationships to cellular genomes remain major issues of debate among evolutionary biologists. Broadly, the early theories of viral origins could be placed in two categories. The first of these sought to place the viruses in the earliest phases of life's evolution and associated them with the primitive precursors of cellular systems (Alstein, 1992; Gibbs et al., 1995). The second group of theories saw viruses as secondary derivatives of cellular systems that underwent drastic degeneration as a consequence of extreme parasitism, or "break away" elements from cellular genomes that survived as minimal parasitic replicons (Gibbs et al., 1995). The two groups of theories are not mutually exclusive: conceivably, some classes of viruses could be primordial whereas others could be later derivatives of "break away" elements from cellular systems. The advent of the first complete genome sequences of viruses did not resolve these debates entirely, but threw considerable light on the actual diversity in the coding capacity of various viruses, the affinities between different viral groups and homologies between viral genes and those of cellular organisms.

The first decade of viral comparative genomics revealed several major assemblages of viruses that were unified on the basis of the evolutionarily conserved proteins of their replication apparatus. Firstly, it became clear that the retroviruses, together with their various relatives such as the hepadnaviruses, plant badnaviruses, and tungroviruses, and the diverse retroposons shared a common ancestor, which encoded a reverse transcriptase (RT) as their principal replication polymerase (Xiong and Eickbush, 1990). The RNA-dependent RNA polymerases (RDRP) of diverse positive strand RNA viruses and several double-stranded(ds) RNA viruses were likewise unified, indicating a common origin for this entire assembly of viruses (Kamer and Argos, 1984; Koonin et al., 1989). At a deeper level, the RTs and RDRPs have been shown to descend from an ancestral replicase that utilized an RNA template (Delarue et al., 1990; Kamer and Argos, 1984; Poch et al., 1989; Xiong and Eickbush, 1990), suggesting that at least these two major classes of viruses might have ultimately descended from an ancient replicon with an RNA genome. This unification also suggested that the diversification of these viruses might be linked to one of the fundamental evolutionary transitions from RNA genomes to the DNA genomes (Forterre, 2002; Leipe et al., 1999; Wintersberger and Wintersberger, 1987).

Similarly, certain assemblages sharing common replication systems also became apparent amongst the DNA viruses. In particular, many small DNA viruses and related plasmids and transposons were unified on the basis of a shared rolling circle replication endonuclease (RCRE), which initiates the eponymous form of replication of these elements (Ilyina and Koonin, 1992; Iyer et al., 2005; Kapitonov and Jurka, 2001). However, the relationships among large dsDNA viruses that have complex genomes with dozens or even hundreds of genes remained far more difficult to elucidate. Amongst the bacteriophages, several major monophyletic groups, such as the lambdoid phages, were identified (Hendrix, 2003). Among the animal large dsDNA viruses, the families *Herpesviridae*, *Baculoviridae*, and *Poxviridae* are obviously monophyletic. The common ancestors of each of these families have been partially reconstructed and, in each case, inferred to have had over 50 genes (Davison et al., 2005; Hughes and Friedman, 2005; Lauzon et al., 2005; McLysaght et al., 2003). Thus, the common ancestral forms of these viral families seem to have already attained considerable complexity—the salient features of replication, gene expression and virion architecture apparently emerged early in their evolution and were retained over vast evolutionary time spans. In contrast, higher-order relationships between various groups of large eukaryotic DNA viruses, if any, remained uncertain. In our previous work, we addressed this issue through comprehensive comparative analysis of the protein sequences encoded by large eukaryotic DNA viruses, followed by cladistic analysis using a character matrix based on the conserved features of these proteins (Iyer et al., 2001). This analysis produced evidence of common ancestry of several families of large eukaryotic DNA viruses, including the animal poxviruses, iridoviruses, and asfarviruses (with a single representative, the African Swine Fever Virus, ASFV), and the phycodnaviruses, which infect phylogenetically diverse algae.

We named this major, monophyletic assemblage of large eukaryotic DNA viruses the Nucleo-Cytoplasmic Large DNA Virus (NCLDV) clade as they either replicate exclusively in the cytoplasm of the host cell or start their life cycle in the host nucleus but complete it in the cytoplasm. Typically, the NCLDVs do not exhibit much dependence on the host replication or transcription systems for completing their replication because, even in viruses like *Paramecium bursaria* Chlorella virus (PBCV), which initiate replication in the nucleus, disruption of a functional host nucleus by irradiation does not abrogate replication (Van Etten et al., 1986). This relative independence of the NCLDVs from the host cells is consistent with the fact that all these viruses encode several conserved proteins performing most key life-cycle processes, such as DNA polymerases, helicases, and DNA clamps for DNA replication, Holliday junction resolvases and topoisomerases for genome manipulation, transcription factors involved in transcription initiation and elongation, ATPase pumps for DNA packaging, and chaperones involved in the capsid assembly (Iyer et al., 2001). In the original analysis, this conserved core was found to include 9 proteins shared by all families of NCLDVs and 22 additional proteins shared by at least three of the four families (Iyer et al., 2001). This suggested that all extant NCLDV families have descended from a common ancestor that already had a fairly complex gene repertoire and was capable of completing its replication cycle in relative autonomy from the cell.

Subsequent to the original description of the NCLDV group, several major developments have occurred, the chief among them being sequencing of the 1.2-megabase genome of the gigantic *Acanthamoeba polyphaga* Mimivirus (Raoult et al., 2004). Analysis of the mimivirus genome showed that it was a new branch of the NCLDV group. In addition, this largest known viral genome contains numerous multi-gene families as