Intervirology

Intervirology 2010;53:284–292
DOI: 10.1159/000312913

Published online: June 15, 2010

Origin and Evolution of Eukaryotic Large Nucleo-Cytoplasmic DNA Viruses

Eugene V. Koonin Natalya Yutin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Md., USA

Key Words
Bacteriophage · Eukaryogenesis · Nucleo-cytoplasmic large DNA viruses, evolution · Phylogenetic analysis

Abstract
Background/Aims: The nucleo-cytoplasmic large DNA viruses (NCLDV) constitute an apparently monophyletic group that consists of 6 families of viruses infecting a broad variety of eukaryotes. A comprehensive genome comparison and maximum-likelihood reconstruction of NCLDV evolution reveal a set of approximately 50 conserved genes that can be tentatively mapped to the genome of the common ancestor of this class of eukaryotic viruses. We address the origins and evolution of NCLDV. ***Results:*** Phylogenetic analysis indicates that some of the major clades of NCLDV infect diverse animals and protists, suggestive of early radiation of the NCLDV, possibly concomitant with eukaryogenesis. The core NCLDV genes seem to have originated from different sources including homologous genes of bacteriophages, bacteria and eukaryotes. These observations are compatible with a scenario of the origin of the NCLDV at an early stage of the evolution of eukaryotes through extensive mixing of genes from widely different genomes. ***Conclusions:*** The common ancestor of the NCLDV probably evolved from a bacteriophage as a result of recruitment of numerous eukaryotic and some bacterial genes, and concomitant loss of the majority of phage genes except for a small core of genes coding for proteins essential for virus genome replication and virion formation.

Introduction

Viruses are ubiquitous parasites of all cellular life forms. As a group, they are united by their intracellular reproduction and reliance on the host cell translation system, but not necessarily by common origin [1]. Indeed, not a single gene is represented in the genomes of all known viruses, although a small group of 'viral hallmark genes' encoding some of the key proteins involved in genome replication and virion structure formation are shared by extremely diverse subsets of viruses [2, 3]. Thus, viruses as a class of biological agents are not monophyletic, at least not within the traditional concept of monophyly. Nevertheless, several large groups of viruses infecting diverse hosts do appear to share common ancestry in the strict sense – that is, to have evolved from a single ancestral virus – which is indicated by the conservation of sets of genes encoding proteins responsible for many functions essential for virus reproduction.

One of the most expansive apparently monophyletic viral divisions currently includes six families of eukaryotic viruses with large DNA genomes that are collectively denoted nucleo-cytoplasmic large DNA viruses (NCLDV; table 1) [4, 5]. The best known of these viral families, Poxviridae, is a large assemblage of animal viruses that includes a major human pathogen, the smallpox virus, important animal pathogens, such as rabbit myxoma virus, as well as vaccinia virus, one of the best characterized models of molecular biology [6–8]. Another family of the NCLDV that recently became the focus

KARGER

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

© 2010 S. Karger AG, Basel

Accessible online at:
www.karger.com/int

Eugene V. Koonin
National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894 (USA)
E-Mail koonin@ncbi.nlm.nih.gov

Table 1. The six families of NCLDV

Virus family	Hosts	Genome size range, kb	Replication site
Phycodnaviridae	Green algae; algal symbionts of paramecia and hydras	150–400	Nucleus and cytoplasm
Poxviridae	Animals: insects, reptiles, birds, mammals	130–380	Cytoplasm
Asfarviridae	Mammals	170	Cytoplasm
Asco- and Iridoviridae	Invertebrates and non-mammalian vertebrates	100–220	
Ascoviridae	Insects, mainly noctuids	150–190	Nucleus and cytoplasm
Iridoviridae	Insects, cold-blooded vertebrates	100–220	Nucleus and cytoplasm
Mimiviridae	Acanthamoeba; algae (probably); corals (possibly)	1,180	Cytoplasm
Marseillevirus	Acanthamoeba; algae (probably)	370	Nucleus and cytoplasm (?)

of much attention and fascination is the Mimiviridae, which so far includes two closely related giant viruses isolated from Acanthamoeba – Mimivirus and Mamavirus. With their genomes being slightly larger than 1 megabase, these viruses are undisputed genome size record holders in the virosphere, exceed numerous parasitic bacteria, and approach the genome size of the simplest free-living prokaryotes [9–13].

The NCLDV infect animals and diverse unicellular eukaryotes, and either replicate exclusively in the cytoplasm of the host cells, or possess both cytoplasmic and nuclear stages in their life cycle (table 1). The NCLDV typically do not strongly depend on the host replication or transcription systems for completing their replication [6, 14]. In line with this relative independence of virus reproduction from the host cell functions (apart from translation, of course), the NCLDV encode several conserved proteins that mediate most of the processes essential for viral reproduction. These key proteins include DNA polymerases, helicases and primases responsible for DNA replication, Holliday junction resolvases and topoisomerases involved in genome DNA processing and maturation, transcription factors that function in transcription initiation and elongation, ATPase pumps mediating DNA packaging, chaperones involved in capsid assembly, and capsid proteins themselves [4, 5, 15]. Although several viral hallmark genes [3] are shared by NCLDV and other large DNA viruses, such as herpesviruses and baculoviruses, the conservation of the entire set of core genes clearly demarcates the NCLDV as a distinct class of viruses [5].

Recently, a novel giant virus, denoted Marseillevirus, has been isolated from Acanthamoeba. Genome analysis of Marseillevirus indicated that it represents a putative new family of NCLDV that appears to be distantly related to iridoviruses and ascoviruses [16]. In addition, comparative genomic analysis revealed probable gene exchange between Marseillevirus and Mimiviruses, an observation that suggests a role of amoeba as a 'melting pot' of giant virus evolution.

We performed a new comparison of the updated collection of NCLDV genomes and constructed clusters of orthologous NCLDV genes (NCVOGs), 177 of which were represented in two or more viral families [15]. The NCVOGs were employed for phylogenetic analysis and for reconstruction of the ancestral viral gene set. Here we review the results of these analyses in the context of the origin and evolution of the NCLDV, and attempt to decipher the origins of the NCLDV genes that are mapped to the last common ancestral virus.

Cross-Mapping of the Phylogenetic Trees of NCLDV and Eukaryotes

As in the major divisions of cellular life forms [17, 18], very few genes are represented in all sequenced NCLDV genomes. The original comparative genomic analysis revealed 9 universal NCLDV genes [4], and the latest update that took into account the newly discovered viral families showed that 5 genes remained common to all known NCLDV [15] (table 2). In order to derive a maximally robust phylogeny of the NCLDV, we analyzed phylogenetic trees of both the universal and the nearly universal genes. These trees had somewhat conflicting topologies; however, given the results of previous studies that pointed to an origin of the NCLDV from a single ancestral virus ([4, 5] and see below), we assumed that the discrepancies between the tree topologies of the highly conserved NCLDV genes were caused by phylogenetic analysis artifacts rather

Table 2. The reconstructed core gene set of the common ancestor of the NCLDV

NCVOG name [15]	Vaccinia virus gene number[a]	Functional category	Number of viral families/ genomes	Number of genomes present in a cluster						NCVOG annotation	Probable origin
				Pox-viridae	Asfar-viridae	Phycodna-viridae	Mimi-viridae	Irido- and Asco-viridae	Mar-seille-virus		
NCVOG0076	A18R	DNA replication, recombination and repair	6/38	20/20	1/1	9/10	2/2	5/11	1/1	DNA or RNA helicases of superfamily II (COG1061)	Bacterial
NCVOG0023	D5R	DNA replication, recombination and repair	6/45	20/20	1/1	10/10	2/2	11/11	1/1	D5-like helicase-primase	Bacteriophage/ plasmid
NCVOG0038	E9L	DNA replication, recombination and repair	6/45	20/20	1/1	10/10	2/2	11/11	1/1	DNA polymerase elongation subunit family B	Eukaryotic/ archaeal
NCVOG0037		DNA replication, recombination and repair	6/15	1/20	1/1	8/10	2/2	2/11	1/1	DNA topoisomerase II	Eukaryotic
NCVOG0276	F4L	Nucleotide metabolism	6/29	13/20	1/1	10/10	2/2	2/11	1/1	Ribonucleotide reductase small subunit	Eukaryotic
NCVOG1353	I4L	Nucleotide metabolism	6/24	3/20	1/1	10/10	2/2	7/11	1/1	Ribonucleoside diphosphate reductase, large subunit	Eukaryotic?
NCVOG0052	E10R	Virion structure and morphogenesis	6/44	20/20	1/1	9/10	2/2	11/11	1/1	Disulfide (thiol) oxidoreductase; Erv1/Alr family (pfam04777)	Eukaryotic (no homologs in prokaryotes)
NCVOG0236	D9R, D10R	Transcription and RNA processing	6/29	20/20	1/1	1/10	2/2	4/11	1/1	Nudix hydrolase (D10 ortholog)	Eukaryotic (?)
NCVOG0262	A2L	Transcription and RNA processing	6/45	20/20	1/1	10/10	2/2	11/11	1/1	Poxvirus late transcription factor VLTF3 like	Uncertain (no obvious homologs outside NCLDV)
NCVOG1164	A1L	Transcription and RNA processing	6/44	20/20	1/1	10/10	2/2	10/11	1/1	A1L transcription factor/late transcription factor VLTF-2	Uncertain (no obvious homologs outside NCLDV)
NCVOG0271	A24R	Transcription and RNA processing	6/36	20/20	1/1	1/10	2/2	11/11	1/1	DNA-directed RNA polymerase subunit beta	Eukaryotic
NCVOG0274	J6R	Transcription and RNA processing	6/36	20/20	1/1	1/10	2/2	11/11	1/1	DNA-directed RNA polymerase subunit alpha	Eukaryotic
NCVOG0272	E4L	Transcription and RNA processing	6/39	18/20	1/1	8/10	2/2	9/11	1/1	Transcription factor S-II (TFIIS)-domain-containing protein	Eukaryotic
NCVOG1117	D1R	Transcription and RNA processing	6/33	20/20	1/1	8/10	2/2	1/11	1/1	mRNA capping enzyme large subunit	Eukaryotic
NCVOG1361		Uncharacterized	6/11	2/20	1/1	1/10	2/2	4/11	1/1	T5orf172 domain (pfam10544)	Bacteriophage [39]
NCVOG0022	D13L	Virion structure and morphogenesis	6/45	20/20	1/1	10/10	2/2	11/11	1/1	NCLDV major capsid protein (pfam03340 for Poxviridae; pfam04451 for others)	Bacteriophage? No homologs in cellular life forms
NCVOG0249	A32L	Virion structure and morphogenesis	6/45	20/20	1/1	10/10	2/2	11/11	1/1	A32-like packaging ATPase	Bacteriophage [40]
NCVOG0278	A22R	DNA replication, recombination and repair	5/36	20/20	0/1	9/10	2/2	4/11	1/1	RuvC, Holliday junction resolvases; cl00243. Extended Pox_A22, Poxvirus A22 family (pfam04848)	Bacterial/bacteriophage, possibly, through the mitochondrion [41]
NCVOG1060	G5R	DNA replication, recombination and repair	5/35	20/20	0/1	1/10	2/2	11/11	1/1	FLAP-like endonuclease XPG (cd00128)	Eukaryotic
NCVOG0319	J2R	Nucleotide metabolism	5/20	15/20	1/1	1/10	2/2	0/11	1/1	Thymidine kinase	Eukaryotic
NCVOG0330	VACWR012, VACWR207	Signal transduction regulation	5/26	16/20	0/1	4/10	2/2	3/11	1/1	RING-finger-containing E3 ubiquitin ligase (COG5432: RAD18)	Eukaryotic (no prokaryotic homologs)
NCVOG0261	A7L	Transcription and RNA processing	5/35	20/20	1/1	0/10	2/2	11/11	1/1	Poxvirus early transcription factor (VETF), large subunit (pfam04441)	Uncertain, no obvious homologs outside NCLDV
NCVOG0273		Transcription and RNA processing	5/15	0/20	1/1	1/10	2/2	10/11	1/1	Divergent DNA-directed RNA polymerase subunit 5	Eukaryotic/archaeal
NCVOG0034[b]	A50R	DNA replication, recombination and repair	4/19	11/20	1/1	6/10	0/2	0/11	1/1	ATP-dependent DNA ligase (pfam01068, PRK01109)	Apparently, polyphyletic: different origins (eukaryotic or bacteria/bacteriophage) in different groups of the NCLDV [29]

NCVOG0004		DNA replication, recombination and repair	4/6	2/20	1/1	0/10	2/2	0/11	1/1	AP (apurinic) endonuclease family 2 - bacterial	Bacterial
NCVOG1192		DNA replication, recombination and repair	4/13	1/20	1/1	9/10	2/2	0/11	0/1	YqaJ viral recombinase family (pfam09588)	Bacteriophage
NCVOG1068	F2L	Nucleotide metabolism	4/30	17/20	1/1	8/10	0/2	4/11	0/1	dUTPase (cl00493)	Eukaryotic
NCVOG0320	A48R	Nucleotide metabolism	4/21	4/20	1/1	5/10	0/2	11/11	0/1	pfam02223: thymidylate kinase	Eukaryotic
NCVOG0040	H1L	Other metabolic functions	4/30	19/20	0/1	6/10	2/2	3/11	0/1	cd00127, DSPc, dual specificity phosphatases; Ser/Thr and Tyr protein phosphatases	Eukaryotic
NCVOG1127		Transcription and RNA processing	4/11	0/20	1/1	7/10	2/2	0/11	1/1	Transcription initiation factor IIB	Archaeal/eukaryotic
NCVOG0010		Uncharacterized	4/11	2/20	0/1	1/10	2/2	6/11	0/1	pfam02498: Bro-N; BRO family, N-terminal domain. This family includes the N-terminus of baculovirus BRO and ALI motif proteins	Bacteriophage [39]
NCVOG0211	F9L, L1R	Virion structure and morphogenesis	4/34	20/20	1/1	0/10	2/2	11/11	0/1	Myristylated IMV envelope protein (pfam02442: lipid membrane protein of large eukaryotic DNA viruses)	Uncertain, no detectable homologs outside NCLDV
NCVOG0035		DNA replication, recombination and repair	3/7	3/20	0/1	0/10	2/2	2/11	0/1	NAD+-dependent DNA ligase (smart00532)	Bacteriophage/bacterial [29]
NCVOG0024		DNA replication, recombination and repair	3/4	0/20	1/1	0/10	2/2	0/11	1/1	Superfamily II helicase related to herpesvirus replicative helicase (origin-binding protein UL9), pfam03121	Bacteriophage
NCVOG0036	H6R	DNA replication, recombination and repair	3/23	20/20	0/1	1/10	2/2	0/11	0/1	DNA topoisomerase I	Bacterial
NCVOG0267	I8R	DNA replication, recombination and repair	3/23	20/20	1/1	0/10	2/2	0/11	0/1	RNA-helicase DExH-NPH-II	Eukaryotic
NCVOG0009		Host-virus interactions	3/4	2/20	1/1	0/10	0/2	1/11	0/1	pfam00653: BIR (baculovirus inhibitor of apoptosis protein repeat) domain	Eukaryotic
NCVOG0012	A33R, A44R, A40R	Host-virus interactions	3/20	18/20	1/1	1/10	0/2	0/11	0/1	C-type lectin: smart00034, cd03594,cd03593, pfam00059, cd00037, pfam05966	Eukaryotic
NCVOG1360	VACWR011, VACWR208	Miscellaneous	3/18	15/20	0/1	0/10	2/2	1/11	0/1	KilA domain (pfam04383); always present at protein N-termini except for Mimiviruses. In some proteins, followed by a RING-finger domain	Bacteriophage [39]
NCVOG1115	D4R	Other metabolic functions	3/23	20/20	0/1	0/10	2/2	0/11	1/1	Uracil-DNA glycosylase	Bacterial
NCVOG0246		Other metabolic functions	3/4	0/20	1/1	1/10	2/2	0/11	0/1	Ulp1-like	Eukaryotic
NCVOG1088		Transcription and RNA processing	3/13	0/20	1/1	0/10	0/2	11/11	1/1	RNA ligase	Bacterial/bacteriophage
NCVOG1424		Uncharacterized	3/6	3/20	0/1	0/10	2/2	1/11	0/1	Uncharacterized domain; found downstream of KilA, BRO, and MSV199 domains. Also found in some baculoviruses	Uncertain; not detected outside NCLDV and baculoviruses
NCVOG1122	G9R, J5L, A16L	Virion structure and morphogenesis	3/31	20/20	0/1	0/10	2/2	9/11	0/1	Myristylated protein; pfam03003, DUF230	Uncertain (no obvious homologs outside NCLDV)
NCVOG0256	H3L	Other metabolic functions	2/22	20/20	0/1	0/10	2/2	0/11	0/1	IMV envelope protein, glycosyltransferase	Uncertain (bacterial and/or eukaryotic); possibly, different origins in poxviruses and Mimiviruses
NCVOG0329		Other metabolic functions	2/3	0/20	1/1	0/10	2/2	0/11	0/1	UBCc, ubiquitin-conjugating enzyme E2 (cd00195)	Eukaryotic
NCVOG0059		Other metabolic functions	2/3	0/20	1/1	0/10	2/2	0/11	0/1	FtsJ-like methyltransferase family proteins (pfam01728)	Eukaryotic

[a] The systematic gene names are for the Copenhagen strain of vaccinia virus, except for the names starting with VACWR which are from the WR strain because the respective genes are missing/disrupted in the Copenhagen strain. [b] The ATP-dependent DNA ligase gene is included for the illustration/discussion purposes although it is unlikely to have been present in ancestral NCLDV.

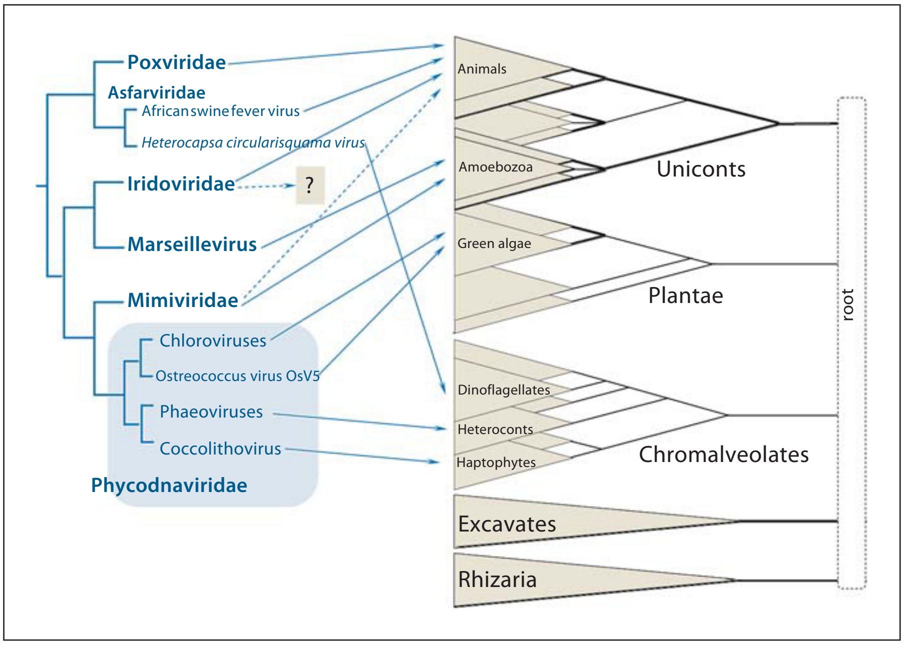

Fig. 1. Cross-mapping of the phylogenetic trees of the NCLDV and eukaryotes, the chordopoxvirus and chlorovirus branches are collapsed. The eukaryotic tree is shown as a multifurcation of 5 supergroups. Lines connect viruses with their host organisms. Solid lines show established virus-host relationships; broken lines show putative relationships inferred from metagenomic data. The short broken line from Iridoviridae is to indicate that, according to metagenomic results, members of this family probably infect marine unicellular eukaryotes but the exact unicellular hosts are not known. Adapted from [15, 20].

than genuinely different evolutionary trajectories of these genes (although some exceptions are possible [15]). Thus, to produce a 'species tree' of the NCLDV, we derived a consensus of the trees for individual conserved genes (fig. 1). The best supported consensus tree topology reveals three major divisions of the NCLDV: (1) the recently discovered Marseillevirus clustered with iridoviruses and ascoviruses, with the latter confidently placed inside the family Iridoviridae; (2) Mimiviruses clustered with phycodnaviruses, and (3) poxviruses clustered with asfarviruses.

When an NCLDV tree was constructed using a completely different approach that was based on the comparison of the patterns of representation (phyletic patterns) of viruses in NCVOGs [19], the resulting tree topologies were generally compatible with the topology of the sequence-based consensus tree, indicating that evolution of the gene repertoire of the NCLDV largely mirrored the evolution of the conserved core genes [15, 16]. There was one notable exception to this congruence, namely, clustering of Marseillevirus with the Mimivirus that suggests extensive gene exchange between these viruses that reproduce in the same amoebal host [16].

Although viruses of unicellular eukaryotes are still poorly characterized, the hosts of known NCLDV span much of the phylogenetic diversity of eukaryotes. The best current representation of eukaryotic phylogeny appears to be a multifurcation of five (or possibly four) major supergroups (fig. 1). The earliest events of eukaryotic radiation and, accordingly, the root of the tree remain murky [20–22]. Cross-mapping of the NCLDV and eukaryotic trees reveals a complex network structure where members of the same NCLDV branch often infect organisms that belong to different eukaryotic supergroups (fig. 1). For instance, the phycodna-Mimivirus clade of NCLDV spans three eukaryotic supergroups, and the pox-asfarvirus clade spans at least two supergroups (fig. 1). Beyond doubt, this assessment of the diversity of the NCLDV host range only scratches the surface, as indicated by the discovery of an extensive unexplored diversity of homologs of the key genes of the NCLDV in marine metagenomic sequences [23–25]. The discovery of numerous environmental sequences that are homologous to genes of Mimiviruses, iridoviruses, phycodnaviruses and asfarviruses suggests that not only the large divisions but even individual families of the NCLDV (with the possible exception of Poxviridae) infect highly diverse hosts including both animals and unicellular organisms from different supergroups [25, 26].

The complex network connecting the different lineages of the NCLDV with the host lineages (fig. 1) clearly indicates that, on a large scale, viruses of this class did not coevolve with their hosts. Two possible scenarios of NCLDV evolution could account for the observed virus-host mapping:

(1) origin of the common ancestor of the NCLDV in a distinct eukaryotic lineage, for instance, Amoebozoa

which are known to host diverse and complex NCLDV, and subsequently, diverged after acquisition by other hosts via horizontal virus transfer;

(2) the main lineages of the NCLDV radiated from the ancestral lineage prior to the divergence of the eukaryotic supergroups, and this primordial virus diversity was subsequently 'sampled' by evolving hosts.

Of course, mixed evolutionary scenarios are also readily imaginable. Given that horizontal virus transfer between taxonomically distant hosts remains a speculative possibility and the indications of the early origin of the NCLDV (see below), which was probably concomitant with eukaryogenesis, the ancient divergence scenario appears most plausible.

Reconstruction of Evolution of the NCLDV Gene Repertoire

The species tree derived as a consensus phylogeny of the conserved NCLDV genes (fig. 1) was employed as the scaffold to reconstruct the core gene repertoires of ancestral viruses as well as gene loss and gain events during the evolution of the NCLDV. Original reconstructions of the NCLDV gene repertoire evolution were performed using a simple maximum parsimony approach. Recently, we used a more sophisticated maximum-likelihood methodology developed by Csuros and Miklos [27] to map 47 genes to the common ancestor of the NCLDV (table 1) and reconstruct progressively growing gene repertoires for other ancestral viruses (fig. 2). These are highly conservative reconstructions because no approach will assign to ancestral forms genes that survived in only one of the progeny lineages let alone those that were lost in all extant lineages. Nevertheless, the reconstructed gene repertoire seems to cover most, if not all, of the core functions characteristic of this class of viruses. This indicates that the common viral ancestor of all known NCLDV already possessed the relative autonomy from the host cell that is the distinguishing feature of this class of viruses. Such functions include the basal machineries for replication, transcription and transcript processing (such as the capping and decapping enzymes), enzymes required for DNA precursor synthesis (thymidine kinase and thymidylate kinase), the two major virion proteins, the central enzymes of virion morphogenesis (protease and disulfide oxidoreductase), and even some proteins implicated in virus-cell interaction such as a RING-finger ubiquitin ligase subunit (table 2).

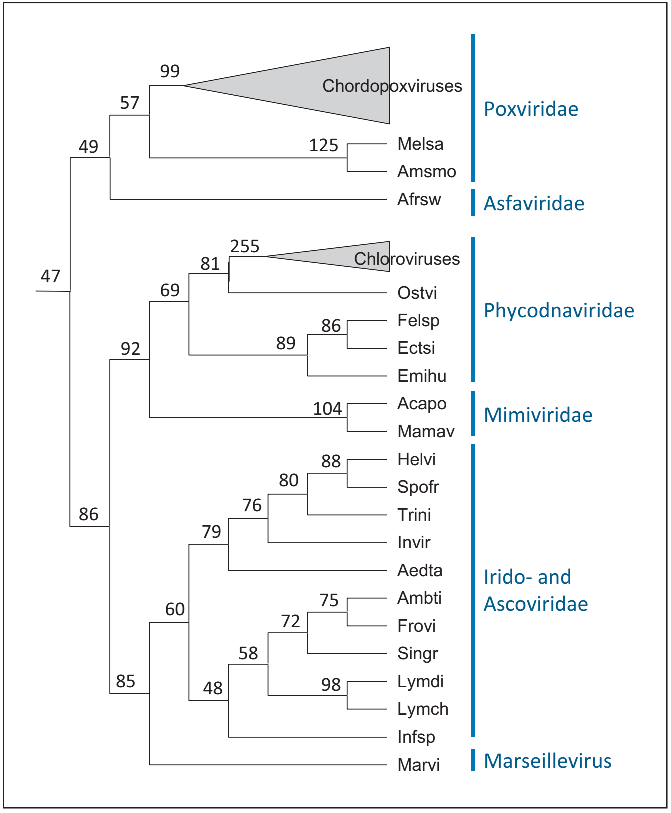

Fig. 2. Reconstruction of the evolution of the NCLDV gene repertoire. Numbers at internal nodes indicate the number of genes assigned to the given ancestral form with high confidence [15]. Amsmo = Amsacta moorei entomopoxvirus; Melsa = melanoplus sanguinipes entomopoxvirus; Helvi = heliothis virescens ascovirus 3e; Trini = trichoplusia ni ascovirus 2c; Spofr = spodoptera frugiperda ascovirus 1a; Afrsw = African swine fever virus; Aedta = aedes taeniorhynchus iridescent virus (Invertebrate iridescent virus 3); Invir = invertebrate iridescent virus 6; Lymdi = lymphocystis disease virus 1; Lymch = lymphocystis disease virus isolate China; Infsp = infectious spleen and kidney necrosis virus; Singr = Singapore grouper iridovirus; Frovi = frog virus 3; Ambti = ambystoma tigrinum virus; Acapo = acanthamoeba polyphaga mimivirus; Mamav = Mamavirus; Emihu = emiliania huxleyi virus 86; Felsp = Feldmannia species virus; Ectsi = ectocarpus siliculosus virus 1; Ostvi = ostreococcus virus OsV5; Marvi = Marseillevirus.

Some of the core functions are prone to non-orthologous gene displacement [28] among the NCLDV, sometimes showing complex patterns of evolution. A case in point is the DNA ligase that is an essential activity for DNA replication. The reconstruction of the ancestral NCLDV gene repertoire tentatively defines the ATP-dependent ligase as an ancestral gene; however, Mimiviruses, entomopoxviruses and some iridoviruses lack the

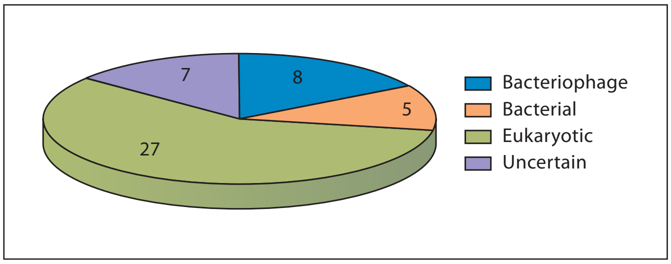

Fig. 3. Breakdown of the 47 genes mapped to the ancestral NCLDV genome by likely origin. The data are from table 2.

ATP-dependent ligase and instead encode an NAD-dependent ligase that is characteristic of bacteria and also found in some bacteriophages. In addition, a considerable number of NCLDV from different families, including some poxviruses and the majority of iridoviruses, encode no DNA ligase at all. Phylogenetic analysis of ATP-dependent and NAD-dependent ligases yielded unexpected results: the NAD-dependent ligases of the NCLDV, although quantitatively less prevalent than the ATP-dependent ligases, turned out to be monophyletic, whereas the ATP-dependent ligases showed diverse phylogenetic affinities, with monophyly confidently rejected. The most likely interpretation of these findings seems to be that the ancestral NCLDV encoded an NAD-dependent ligase, probably of bacteriophage origin, but this ancestral gene was repeatedly and independently lost and replaced with the gene for an ATP-dependent ligase in several viral lineages (table 2) [29]. This case study reveals the remarkable complexity of the NCLDV evolution that is augmented by the possibility of complementation of some of the viral functions by cellular analogs, as recently demonstrated experimentally for the poxvirus DNA ligase [30], and is only partially captured by reconstructions based on patterns of gene presence-absence.

Given the inherent conservative character of the reconstruction and the complications caused by non-orthologous gene displacement, the actual genome size and complexity of the ancestral NCLDV is a wide-open question. The 47 genes mapped to the ancestral genome in the present reconstruction comprise only the core of most highly conserved, essential viral genes involved in key functions. The ancestral NCLDVs undoubtedly reproduced in unicellular eukaryotes, and this type of hosts support the propagation of extant giant viruses, such as the Mimiviruses [13, 31], that actively absorb genes from the eukaryotic hosts as well as bacterial endosymbionts [32, 33]. Thus, it cannot be ruled out that the common ancestor of all extant NCLDV was a highly complex, possibly even a giant virus [16].

Origins of Ancestral NCLDV Genes

All the complications notwithstanding, the reconstruction of the gene composition of the common ancestor of the NCLDV is a relatively straightforward task. In contrast, the origin of this ancestral virus remains enigmatic. We examined homologs and phyletic patterns of the inferred set of ancestral genes of the NCLDV in an attempt to decipher their likely origins (table 2). Definitive inference of gene origins requires a comprehensive phylogenetic analysis that is beyond the scope of the present review. However, in many cases, even examination of the taxonomic composition of the most similar homologs of a gene allows one to determine its most likely origin, especially when all or nearly all homologs belong to the same taxon [see, for instance, 34, 35]. We therefore compared representative sequences of the 47 putative ancestral proteins from all NCLDV families to the non-redundant protein sequence database at the NCBI [36] using the BLASTP program, with multiple PSI-BLAST iterations where required [37, 38], and manually examined the results using the Taxonomy Report feature of the NCBI BLAST server, in an attempt to infer the likely origin of each gene. For most of the putative ancestral genes, the taxonomic distribution of the highly conserved homologs turned out to be obviously skewed, allowing confident inference of the most likely origin that, in several cases, was also supported by previous detailed analyses (table 2).

The majority of the ancestral genes of the NCLDV showed a clear eukaryotic affinity but a substantial minority appeared to be of bacteriophage origin and a few genes of bacterial origin (table 2 and fig. 3). The genes of apparent bacteriophage origin encode some of the key proteins involved in viral replication, such as the DNA primase-helicase, NAD-dependent ligase and Holliday junction resolvase, and DNA packaging in the capsid, namely the packaging ATPase. The major capsid protein itself is most likely of the same origin (table 2). All these genes, with the possible exception of the DNA ligase, are viral hallmark genes that are shared by diverse viruses [3]. Genes of inferred eukaryotic origin encode proteins involved in functions that are related to the cytoplasmic site of the NCLDV replication, such as the RNA polymerase subunits, and the specifics of eukaryotic molecular biology, such as the capping and decapping enzymes or ubiquitin ligase (table 2).

These observations are most compatible with a scenario for the origin of the NCLDV under which the ancestral virus of this class evolved from a bacteriophage by replacement of many (probably most) of the phage genes, primarily by genes acquired from the eukaryotic hosts. Only a small core of phage genes encoding virus-specific functions for which no functional analog exists in cellular life forms survived in the NCLDV genomes. It is notable that even the principal enzyme of DNA replication, the DNA polymerase, was apparently replaced by the eukaryotic counterpart. Nevertheless, this scenario is compatible with the principle of the continuity of evolution in the virus world, *Omnis virus e virus* [3].

Concluding Remarks

The recent expansion of virology into the study of viruses infecting unicellular eukaryotes resulted in the unexpected discovery of giant viruses that belong to three families: Mimiviridae, Phycodnaviridae, and the putative novel family represented by Marseillevirus. Phylogenetic analysis of the expanded class of NCLDV and cross-mapping of the phylogenetic trees of NCLDV and their eukaryotic hosts suggest an early origin and primary radiation of the NCLDV, possibly concomitant with eukaryogenesis. Phylogenomic reconstruction maps approximately 50 genes to the last common ancestor of the extant NCLDV. However, this is a conservative reconstruction. A distinct possibility is that the ancestral virus of this class was indistinguishable from its modern members in terms of genetic complexity. The core NCLDV genes seem to have originated from different sources, with the majority affined with eukaryotic homologs but a substantial minority derived from bacteriophage genes. These observations are compatible with the principle of the evolutionary continuity of the viral world, whereby the common ancestor of the NCLDV evolved from a bacteriophage as a result of recruitment of numerous eukaryotic and some bacterial genes, and concomitant loss of the majority of the ancestral phage genes. Only a small core of genes coding for proteins that are essential for virus genome replication and virion formation and which have no functional analogs in cellular life forms survived in the NCLDV. Subsequent evolution of the NCLDV included lineage-specific recruitment of numerous additional genes from both the eukaryotic hosts and bacteria. Gene duplication was also prominent, especially in giant viruses, as well as loss of ancestral genes, especially in animal viruses with smaller genomes, resulting in the extensive genomic diversity observed among the extant NCLDV.

Acknowledgment

The authors' research is supported by the DHHS intramural program (NIH, National Library of Medicine).

References

1 Raoult D, Forterre P: Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol 2008;6:315–319.
2 Forterre P: The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 2006;117:5–16.
3 Koonin EV, Senkevich TG, Dolja VV: The ancient virus world and evolution of cells. Biol Direct 2006;1:29.
4 Iyer LM, Aravind L, Koonin EV: Common origin of four diverse families of large eukaryotic DNA viruses. J Virol 2001;75: 11720–11734.
5 Iyer LM, Balaji S, Koonin EV, Aravind L: Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res 2006;117: 156–184.
6 Moss B: Poxviridae: the viruses and their replication; in Fields BN, Knipe DM, Howley PM, Griffin DE (eds): Fields Virology. Philadelphia, Lippincott, Williams & Wilkins, 2001, vol 2, pp 2849–2884.
7 Lefkowitz EJ, Wang C, Upton C: Poxviruses: past, present and future. Virus Res 2006;117: 105–118.
8 Hughes AL, Irausquin S, Friedman R: The evolutionary biology of poxviruses. Infect Genet Evol 2010;10:50–59.
9 La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Merchat M, Suzan-Monti M, Forterre P, Koonin E, Raoult D: The virophage as a unique parasite of the giant Mimivirus. Nature 2008;455:100–104.
10 Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM: The 1.2-megabase genome sequence of Mimivirus. Science 2004;306: 1344–1350.
11 Claverie JM, Abergel C: Mimivirus and its virophage. Annu Rev Genet 2009;43:49–66.
12 Claverie JM, Abergel C, Ogata H: Mimivirus. Curr Top Microbiol Immunol 2009;328:89–121.
13 Suzan-Monti M, La Scola B, Raoult D: Genomic and evolutionary aspects of Mimivirus. Virus Res 2006;117:145–155.
14 Van Etten JL: Unusual life style of giant chlorella viruses. Annu Rev Genet 2003;37:153–195.
15 Yutin N, Wolf YI, Raoult D, Koonin EV: Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 2010;6:223.

16 Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa M, Robert C, Azza A, Sun S, Rossmann MG, Suzan-Monti M, La Scola B, Koonin EV, Raoult D: Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimaeric microorganisms. Proc Natl Acad Sci USA 2009; 106:21848–21853.

17 Koonin EV: Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 2003;1:127–136.

18 Charlebois RL, Doolittle WF: Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res 2004;14:2469–2477.

19 Wolf YI, Rogozin IB, Grishin NV, Koonin EV: Genome trees and the tree of life. Trends Genet 2002;18:472–479.

20 Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW: The tree of eukaryotes. Trends Ecol Evol 2005;20:670–676.

21 Keeling PJ: Genomics: deep questions in the tree of life. Science 2007;317:1875–1876.

22 Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AG, Roger AJ: Phylogenomic analyses support the monophyly of excavata and resolve relationships among eukaryotic 'Supergroups'. Proc Natl Acad Sci USA 2009; 106:3859–3864.

23 Monier A, Claverie JM, Ogata H: Taxonomic distribution of large DNA viruses in the sea. Genome Biol 2008;9:R106.

24 Monier A, Larsen JB, Sandaa RA, Bratbak G, Claverie JM, Ogata H: Marine Mimivirus relatives are probably large algal viruses. Virol J 2008;5:12.

25 Kristensen DM, Mushegian AR, Dolja VV, Koonin EV: New dimensions of the virus world discovered through metagenomics. Trends Microbiol 2010;18:11–19.

26 Ogata H, Toyoda K, Tomaru Y, Nakayama N, Shirai Y, Claverie JM, Nagasaki K: Remarkable sequence similarity between the dinoflagellate-infecting marine girus and the terrestrial pathogen African swine fever virus. Virol J 2009;6:178.

27 Csuros M, Miklos I: Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. Mol Biol Evol 2009;26:2087–2095.

28 Koonin EV, Mushegian AR, Bork P: Non-orthologous gene displacement. Trends Genet 1996;12:334–336.

29 Yutin N, Koonin EV: Evolution of DNA ligases of nucleo-cytoplasmic large DNA viruses of eukaryotes: a case of hidden complexity. Biol Direct 2009;4:51.

30 Paran N, De Silva FS, Senkevich TG, Moss B: Recruitment of cellular DNA ligase I to cytoplasmic vaccinia virus factories complements viral DNA replication in the absence of the viral ligase. Cell Host Microbe 2009;6: 563–569.

31 Claverie JM, Grzela R, Lartigue A, Bernadac A, Nitsche S, Vacelet J, Ogata H, Abergel C: Mimivirus and Mimiviridae: giant viruses with an increasing number of potential hosts, including corals and sponges. J Invertebr Pathol 2009;101:172–180.

32 Filee J, Pouget N, Chandler M: Phylogenetic evidence for extensive lateral acquisition of cellular genes by nucleocytoplasmic large DNA viruses. BMC Evol Biol 2008;8:320.

33 Filee J, Siguier P, Chandler M: I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends Genet 2007; 23:10–15.

34 Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin EV: Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. Nucleic Acids Res 2005;33:4626–4638.

35 Podell S, Gaasterland T: DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biol 2007;8:R16.

36 Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, John Wilbur W, Yaschenko E, Ye J: Database resources of the national center for biotechnology information. Nucleic Acids Res 2010; 38(Database issue):D5–D16.

37 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.

38 Altschul SF, Koonin EV: PSI-BLAST: a tool for making discoveries in sequence databases. Trends Biochem Sci 1998;23:444–447.

39 Iyer LM, Koonin EV, Aravind L: Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal apses transcription factors. Genome Biol 2002;3:RESEARCH0012.

40 Iyer LM, Makarova KS, Koonin EV, Aravind L: Comparative genomics of the FTSK-HERA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. Nucleic Acids Res 2004;32:5260–5279.

41 Garcia AD, Aravind L, Koonin EV, Moss B: Bacterial-type DNA Holliday junction resolvases in eukaryotic viruses. Proc Natl Acad Sci USA 2000;97:8926–8931.