INSTITUT PASTEUR

Research in Microbiology 159 (2008) 325–331

www.elsevier.com/locate/resmic

Convergent mechanisms of genome evolution of large and giant DNA viruses

Jonathan Filée [a,*], Michael Chandler [b]

[a] *Laboratoire Evolution, Génomes et Spéciation, CNRS UPR 9034, Avenue de la Terrasse, 91198 Gif sur Yvette Cedex, France*
[b] *Laboratoire de Microbiologie et Génétique Moléculaire, CNRS UMR 5100,118 Route de Narbonne, 31062 Toulouse Cedex 04, France*

Received 3 March 2008; accepted 17 April 2008
Available online 10 May 2008

Abstract

We have taken advantage of the availability of the genome sequences of a collection of large and giant viruses infecting bacteria (T4 family) and eukaryotes (NCLDV group) to assess some of the evolutionary forces which might have shaped their genomes. Despite having apparently different ancestors, these two groups of viruses are affected by convergent evolutionary forces. Both types of virus probably originated from a simple and ancient viral ancestor with a small subset of 30–35 genes encoding replication and structural proteins. The genome size and diversity of the descendants most likely grew progressively by (i) lineage-specific gene duplications, (ii) lateral gene transfers of cellular genes and (iii) accretion of diverse families of mobile genetic elements. These results argue against the hypothesis that giant viruses derive from a regressive cell.

Keywords: Large and giant DNA viruses; Genome evolution; Lateral gene transfer; NCLDV; T4

1. Introduction

The concept of «giant viruses» emerged at the end of the 1990s with the discovery of large algal DNA viruses with genomes greater than 300 kb in length [25] and determination of several genome sequences. Several other giant viruses were subsequently discovered (Table 1), reaching a peak of complexity with the discovery of the 1.2 Mb of the Mimivirus genome [19]. By their genome size and complexity, large and giant DNA viruses challenge our current vision of the viral world and the traditional boundaries between viruses and cellular organisms. As these viruses carry many genes closely related to cellular genes, one major question is: to what extent are lateral gene transfers between viruses and cells responsible for this genomic complexity? More generally, what are the evolutionary forces that shape the genome of the large and giant viruses?

2. What are the larger DNA viruses?

Using the Genome Browser at NCBI (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html) we identified the "top 10" viruses according to genome size (Table 1). Most of the larger Eukaryotic DNA viruses belong to the diverse NCLDV group (for nucleocytoplasmic large DNA virus) whose members infect a wide range of eukaryotic hosts including algae (Phycodnaviruses), protists (Mimivirus) and Metazoa (poxviruses, African swine fever vVirus, iridoviruses). NCLDVs are characterised by a large heterogeneity in genome size (between 100 kb and 1.2 Mb) and are thought to be monophyletic based on a small set of 30 common homologous (core) genes [13]. All of the larger bacteriophages are myoviruses characterised by a long and a contractile tail and, furthermore, most belong to the monophyletic group of the T4 phages, a subgroup of the myoviruses. T4 bacteriophages infect a wide range of bacterial hosts from Proteobacteria to Cyanobacteria and their monophyly is based on a core set of 30–40 shared genes [4,26]. Among the small number of known archaeal viruses, the larger examples are halophilic

* Corresponding author. Tel.: +33 5 69 82 37 34.
E-mail address: jonathan.filee@legs.cnrs-gif.fr (J. Filée).

0923-2508/$ - see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.resmic.2008.04.012

Table 1
Genome size of the "top 10" DNA viruses infecting eukaryotes and bacteria

	Host name	Genome size (kb)	Family
Virus name			
Mimivirus	*Acanthamoeba castelanii*	1181	NCLDV
EHV86	*Emilinia huxleyi*	407	NCLDV
Chlorella Viruses	*Chlorella* sp.	288–369	NCLDV
Canary PoxVirus	Birds	360	NCLDV
ESV	*Ectocarpus siliculosus*	336	NCLDV
Shrimp White spot syndrome virus	Crustacea	305	Nimavirus
Fowlpox Virus	*Gallus gallus*	288	NCLDV
Pongine Herpes Virus 4	*Pongo pygmaeus*	241	HerpesVirus
MSEPV	*Melanoplus sanguinepes*	236	NCLDV
Human Herpes Virus 5	*Homo sapiens*	236	HerpesVirus
Phage name			
G	*Bacillus*	498	Myovirus
PhiKZ	*Pseudomonas* sp.	280	Myovirus
P-SSM2	*Prochlorococcus* sp.	252	Myovirus (T4-like)
KVP40	*Vibrio*	245	Myovirus (T4-like)
Aeh1	*Aeromonas*	233	Myovirus (T4-like)
PhiEL	*Pseudomonas*	211	Myovirus
S-PM2	*Synechococcus* sp.	196	Myovirus (T4-like)
c-st	*Clostridium*	186	Myovirus
RB43	*Escherichia coli*	181	Myovirus (T4-like)
P-SSM4	*Prochlorococcus* sp.	178	Myovirus (T4-like)

viruses with genomes of 75–80 kb [24]. As in the case of the bacteriophages these are all also myoviruses.

It is striking that among the 1964 genome sequences currently available in GenBank at http://www.ncbi.nlm.nih.gov/Genbank/, larger DNA viruses are restricted to only two groups: the myovirus (as T4-like phages) and the NCLDVs (as the Mimiviruses, the Phycodnaviruses, and the Poxviruses). Although we cannot exclude that, with our only partial knowledge of the diversity of the viral world, other groups of large and giant viruses have not yet been identified, it seems improbable that there will be numerous members from other families which have systematically escaped detection, whereas the T4 and the NCLDV families have not. Interestingly, PCR- and metagenomic-based studies of natural environments show that NCLDV and T4-like phages are predominant and highly diverse forms of viruses in microbial communities (see for example: [7,11,22]). These observations also suggest that larger DNA viruses are limited to a small subset of viral families (T4, NCLDV…) that have had great evolutionary success, colonizing a large array of biotopes and hosts.

In this paper we defend the notion that the evolutionary success of T4 and NCLDVs is intimately linked to their capacity to sequester a large diversity of genes via gene duplication and lateral exchange with their hosts. In the following sections, we briefly describe the properties of the genomes of the NCLDVs and T4 and present an ensemble of observations supporting a convergent mechanism of genome evolution for these two groups of viruses. We finally describe an evolutionary scenario in which we propose that present T4 and NCLDV genomes have emerged from two ancient and simple ancestral viruses that have progressively grown in genome size by accumulation of duplicated and laterally acquired genes including diverse mobile genetic elements (MGEs).

3. Genome characteristics of NCLDV and T4

The architecture of T4 and NCLDV genomes displays some interesting convergence in terms of gene content. They encode a set of "core" genes shared by almost all family members. Depending on the similarity threshold used in the analysis, the core is composed of 30–35 genes in T4 [5] and in NCLDVs [13]. The majority encode enzymes involved in DNA metabolism and replication, or viral structural proteins. Both T4 and NCLDVs encode a B family DNA polymerase, a type II or, rarely, type Ia DNA topoisomerase, a viral type DNA primase, a sliding clamp/PCNA, a ThyA- or ThyX-type thymidilate synthase and a type I ribonucleotide reductase. NCLDV core genes also include a DNA ligase, a dUTPase and a thymidine kinase (which are only patchily distributed in T4 – notably absent in all T4-like cyanophages), whereas T4 core genes also include a single-strand DNA binding protein (ssb/gp32) absent in the NCLDV core genes. Both T4 and NCLDVs therefore encode a nearly complete DNA replication apparatus in addition to key enzymes involved in the final steps of DNA metabolism. Phylogenetic analysis of the replication genes showed little evidence of lateral gene transfers between the cells and the viruses [4]. In the phylogenetic trees, the viral genes are generally clustered together more often at the base of the tree and far from the related cellular sequences [4]. These trees suggested that the replication genes were present in the ancestors of the T4 and the NCLDVs and that they have evolved independently, rarely affected by lateral gene transfers. The situation for genes encoding proteins involved in DNA metabolism is different: numerous lateral gene transfer events followed by homologous or non-homologous replacement were observed. This is mainly the case for the NCLDVs where at least 12 potential gene transfer events from the host and possible replacements of the original copies were detected. In T4 only two