Virology 513 (2018) 114–128

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/virology

Sequence analysis of malacoherpesvirus proteins: Pan-herpesvirus capsid module and replication enzymes with an ancient connection to "*Megavirales*"

Arcady Mushegian[a,*], Eli Levy Karin[b], Tal Pupko[b]

[a] Division of Molecular and Cellular Biosciences, National Science Foundation, 2415 Eisenhower Avenue, Alexandria, VA 22314, USA
[b] Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel

ARTICLE INFO

Keywords:
Herpesvirales
Virus evolution
Virus phylogeny

ABSTRACT

The order *Herpesvirales* includes animal viruses with large double-strand DNA genomes replicating in the nucleus. The main capsid protein in the best-studied family *Herpesviridae* contains a domain with HK97-like fold related to bacteriophage head proteins, and several virion maturation factors are also homologous between phages and herpesviruses. The origin of herpesvirus DNA replication proteins is less well understood. While analyzing the genomes of herpesviruses in the family *Malacohepresviridae*, we identified nearly 30 families of proteins conserved in other herpesviruses, including several phage-related domains in morphogenetic proteins. Herpesvirus DNA replication factors have complex evolutionary history: some are related to cellular proteins, but others are closer to homologs from large nucleocytoplasmic DNA viruses. Phylogenetic analyses suggest that the core replication machinery of herpesviruses may have been recruited from the same pool as in the case of other large DNA viruses of eukaryotes.

1. Introduction

The order *Herpesvirales* consists of animal viruses with large linear double-strand (ds) DNA genomes of 125–145 kbp, encoding between 70 and 200 proteins. Herpesviruses are characterized by the following phenotypic attributes (Davison et al., 2009; Pellett et al., 2011): the mature particle contains the genome within a T = 16 icosahedron capsid, composed of 162 capsomers, of which one (the portal) has a distinct structure, specialized for mediating virus DNA entry and exit from the capsid; the capsid is surrounded by a proteinaceous tegument and further wrapped in a lipid envelope that contains several virus-encoded transmembrane proteins; the viral genome contains directed or inverted DNA repeats; the virus DNA replication mechanism generates head-to-tail concatemers that are later cleaved and individually packed into pre-made capsids.

These characteristic molecular and morphological features have been used to assign members to the order *Herpesvirales* even in the absence of the genome sequence information (reviewed in Davison, 2010). In the genomic era, however, the state-of-the-art taxonomy of herpesviruses is based on phylogenetic analyses of virus nucleotide and amino acid sequences (Davison et al., 2009; Davison, 2010; Pellett et al., 2011). For certain subsets of herpesviruses, additional phylogenetic signal can be extracted from the information about gene synteny and by estimating the minimal amount of rearrangements required to convert the gene order between two species (Hannenhalli et al., 1995; Bourque and Pevzner, 2002; Larget et al., 2005).

Analysis of these and other molecular features resulted in establishing three families within *Herpesvirales* that are currently recognized by the International Committee for Taxonomy of Viruses (ICTV; Davison et al., 2009; Pellett et al., 2011). The *Herpesviridae* family (GenBank TaxID 10292, with 73 completely sequenced virus genomes as of May 2017) includes viruses that infect amniotes; typically, a pair of virus species from this family shares dozens of protein-coding genes recognized by high amino acid sequence similarity (Davison, 2010). The family *Alloherpesviridae* (TaxID 548682, 8 completely sequenced genomes) includes viruses of fishes and amphibians. These viruses have been reported to share only 13 core orthologous genes with each other, suggesting that this group is more divergent than the family *Herpesviridae* (van Beurden and Engelsma, 2012). The third family, *Malacoherpesviridae* (TaxID 548685), includes two ICTV-approved members, *Ostreid herpesvirus 1* (OsHV-1) and *Haliotid herpesvirus 1* (HaHV-1), which infect bivalve and gastropod molluscs, respectively. In addition to the herpesvirus-like morphology of their virions and capsids, as well as the genomic repeat features, these two viruses were reported to have 39 orthologous genes in common (Savin et al., 2010), some of which are related to enzymes that play essential roles in other herpesviruses, for example, the large subunit of virus terminase ATPase and family B replicative DNA polymerase. However, no gene products with similarity

* Corresponding author.
E-mail address: mushegian2@gmail.com (A. Mushegian).

http://dx.doi.org/10.1016/j.virol.2017.10.009
Received 20 August 2017; Received in revised form 8 October 2017; Accepted 9 October 2017
0042-6822/ Published by Elsevier Inc.

to herpesvirus structural proteins have been reported in *Malacoherpesviridae*. Interestingly, a putative virus related to mollusc herpesviruses has been detected by sequence database searches, among the contigs co-assembled with the portions of the draft genome of a cephalochordate *Branchiostoma floridae* (Savin et al., 2010). Taken together, these genomic data suggest that the order *Herpesvirales* is characterized by higher genomic diversity, broader host range, and, likely, deeper phylogeny than was assumed just a decade ago.

In contrast to our understanding of the phylogenetic relationships within *Herpesivirales*, the evolutionary origin of this order as a whole remains uncertain. The picture is clearer in the case of structural proteins. One domain, "the floor domain", in the large multi-domain capsid proteins of vertebrate herpesviruses, whose structure has been determined by a cryoEM approach, is structurally similar to bacteriophage capsid proteins with the HK97 fold (Baker et al., 2005; Yu et al., 2017). Moreover, mechanistic similarities in the maturation pathways of these diverse virus particles have been noted (Casjens and King, 1975; Booy et al., 1991; Davison, 2002; Mettenleiter et al., 2009; Veesler and Johnson, 2012), and sequence comparisons have established the ancestral relationship of the capsid maturation protease between phages and *Herpesviridae* (Cheng et al., 2004; Liu and Mushegian, 2004). That relationship is further supported by the demonstration that the high-resolution structure of phage protease is folded very similarly to the herpesvirus assemblin proteases, whose structure has been determined earlier (Fokine and Rossmann, 2016). Along with the long-known monophyly of the large (ATPase) subunit of phage and virus terminases (Mitchell et al., 2002), these observations point at the common origin of the entire head/capsid formation module in *Herpesviridae* and in a subset of tailed viruses of bacteria and archaea.

The state of affairs is different for genes involved in virus DNA replication and expression. For several established families of eukaryotic dsDNA viruses, which, unlike herpesviruses, replicate partly or fully in the cytoplasm (Nucleocytoplasmic Large DNA Viruses, or NCLDV), considerable evidence of the monophyletic origin has been derived from comparative genome analyses (Iyer et al., 2006; Yutin et al., 2009; Koonin and Yutin, 2010). The assembly of large DNA viruses consisting of NCLDV and several related lineages has been proposed as a candidate order "*Megavirales*" (Colson et al., 2013), not yet accepted by ICTV. Large dsDNA viruses replicating in the nucleus, such as herpesviruses and baculoviruses, have several replication enzymes homologous to those of NCLDV (see below), but the evolutionary scenario linking them together has not been established.

In this work, we analyzed the gene repertoire of malacoherpesviruses, using predicted open reading frames (ORFs) of HaHV-1 (synonym *Abalone herpesvirus* Victoria/AUS/2009) as the starting point. Our study characterized the enlarged repertoire of conserved malacoherpesvirus genes, and for many of them we identified similar sequences in other viruses or in the genomes of cellular organisms. Analysis of sequence similarities and phylogenetic inference on those gene families expand the common gene core of *Herpesvirales,* support the hypothesis of the common ancestry of their morphogenetic module, and suggest that the herpesvirus replication module may have been recruited from the same ancient gene pool as the replication genes of other large DNA viruses.

2. Methods

Sequence database searches were mostly performed in July-December of 2016, except for the results presented in Fig. 1 and the accompanying text, for which the searches were repeated in June-September of 2017. Protein queries representing every predicted ORF in HaHV-1 (GenBank Taxonomy ID 860344) were used to query the nucleotide sequence databases (NT and dbEST) and the non-redundant protein sequence database (NR) at NCBI. Searches were done with the BLAST family of programs (Altschul et al., 1997) with the "Composition-based statistics" option. The HHPred server (Söding et al., 2005; Alva et al., 2016) was used to compare profile-Hidden Markov Models (HMM) of virus proteins to profile-HMMs built from the entries in the NCBI CDD database (Marchler-Bauer et al., 2017), with multiple sequence alignment generation method set to PSI-BLAST, re-aligning the results with MAC set to 0.3, and the remaining options set to default values. Multiple sequence alignments were computed using the MUSCLE program (Edgar, 2004), joining closely related (family level and below) sequences, and combining the alignments obtained at that first step using the -profile option of MUSCLE. When distantly related proteins were to be aligned with structural constraints, the PROMALS3D server was used (Pei and Grishin, 2014). Spatial structures were visualized with the PyMOL legacy build 0_99rc6 (https://sourceforge.net/projects/pymol/files/Legacy/).

Phylogenetic inference was done using a local installation of the PhyML program as well as the PhyML 3.0 web server (Guindon et al., 2010) with the LG substitution model, most other parameters estimated from the data, and bootstrap replicates performed to assess the support of the internal partitions in the tree. The iTOL v.3 server (Letunic and Bork, 2016) was employed for tree examination and visualization.

Statistical tests for relaxed selective constraints in the lineages leading to *B.floridae* and *Capitella teleta* were conducted using RELAX (Wertheim et al., 2015). For each set of orthologous proteins, corresponding nucleotide sequences were extracted from GenBank, codons were aligned using the PRANK program (Löytynoja, 2014), and for each multiple sequence alignment, two p-values were computed (one for a relaxation signal in the lineage leading to *B. floridae* and one for a relaxation signal in the lineage leading to *C. teleta*). These p-values were computed using a chi-square distribution in a likelihood ratio test, as proposed in RELAX. In order to account for multiple hypotheses testing and control the false discovery rate, p-values were adjusted using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), as implemented in the R function p.adjust.

3. Results and discussion

3.1. Nearly half of malacoherpesvirus gene products are conserved

We compared 118 putative proteins encoded by the genome of HaHV-1 to the protein and nucleotide sequence databases, as well as the databases of known conserved protein domains (the CDD database at NCBI). The identities of the best statistically supported matches in viruses and cellular organisms were recorded; some of the matches had low statistical significance, but could be validated by observing the conservation of characteristic sequence motifs. The combined results of these analyses are presented in Table 1.

Nearly 60 proteins in HaHV-1 have homologs in the better-studied malacoherpesvirus, OsHV-1 (and often also in the unclassified, but highly similar, malacoherpesvirus infecting scallops, *Chlamys acute necrobiotic virus;* data not shown). The BLASTP pairwise sequence identity within malacoherpesvirus sequences was generally between 20% and 40%, suggesting considerable divergence of these viruses. Some amount of synteny between the two virus genomes was observed, usually in blocks of 2–4 genes.

As reported before (Savin et al., 2010), substantial portions of the malacoherpesvirus genomes also show homology to a part of the draft genome assembly of the cephalochordate *B. floridae.* Most of the herpesvirus-like matches in the *B. floridae* genome are mapped to a single contig; some of these genes have been predicted by the original genome annotation, while others have not been reported by Savin et al. (2010), but could be identified by matching malacoherpesvirus protein sequences to the translations of the putative intergenic regions. We also detected a collection of homologous genes in a draft genome assembly of another marine invertebrate, annelid *C. teleta*. In this case too, at least some of the genes appear to be located on the same contig (Table 1).

Taken together, these matches provide evidence for evolutionary