Review

Giant Viruses – Big Surprises

Nadav Brandes1 and Michal Linial2,*

1 The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel; nadav.brandes@mail.huji.ac.il
2 Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Israel; michall@cc.huji.ac.il
* Correspondence: michall@cc.huji.ac.il; Tel.: +972-02-6585425

Received: date; Accepted: date; Published: date

Abstract: Viruses are the most prevalent infectious agents, populating almost every ecosystem on earth. Most viruses carry only a handful of genes supporting their replication and the production of capsids. It came as a great surprise in 2003 when the first giant virus was discovered and found to have a >1Mbp genome encoding almost a thousand proteins. Following this first discovery, dozens of giant virus strains across several viral families have been reported. Here, we provide an updated quantitative and qualitative view on giant viruses and elaborate on their shared and variable features. We review the complexity of giant virus proteomes, which include functions traditionally associated only with cellular organisms. These unprecedented functions include components of the translation machinery, DNA maintenance, and metabolic enzymes. We discuss the possible underlying evolutionary processes and mechanisms that might have shaped the diversity of giant viruses and their genomes, highlighting their remarkable capacity to hijack genes and genomic sequences from their hosts and environments. This leads us to examine prominent theories regarding the origin of giant viruses. Finally, we present the emerging ecological view of giant viruses, found across widespread habitats and ecological systems, with respect to the environment and human health.

Keywords: Amebae viruses; Viral evolution; Protein domains; Mimivirus; dsDNA viruses; Translation machinery; Pandoravirus; NCLDV.

1. Giant viruses and the viral world

Viruses are cell infecting agents present in almost every ecosystem. Questions regarding viral origin and early evolution alongside all living organisms (Bacteria, Archaea and Eukarya) are still widely open, and relevant theories remain speculative [1-4]. As viruses are exceptionally diverse and undergo rapid changes, it is impossible to construct an ancestral lineage tree for the viral world [5-9]. Instead, virus families are categorized according to the nature of their genetic material, mode of replication, pathogenicity, and structural properties [10].

At present, the viral world is represented by over 8,000 reference genomes [11]. The International Committee on Taxonomy of Viruses (ICTV) provides a universal virus taxonomical classification proposal that covers ~150 families and ~650 genera, with many viruses yet unclassified [12]. This collection provides a comprehensive, compact set of virus representatives.

Inspection of viral genomes reveals that most known viruses have genomes encoding only a few proteins. Actually, 69% of all known viruses have less than 10 proteins encoded in their genomes (Figure 1). It is a common assumption that viruses demonstrate near-optimal genome packing and information compression, presumably in order to maximize their replication rate, number of progenies, and other parameters that increase infectivity [13,14]. However, a debate is still ongoing over the generality of these phenomena [15], and there is a non-negligible percentage of larger viruses (Figure 1). On the far end of the distribution, there are the viruses with hundreds of genes, most of
them belong to giant viruses. Specifically, while only 0.3% of currently known viral proteomes have 500 or more proteins, they encode as much as 7.5% of the total number of viral proteins (Figure 1B).

Figure 1. Number of proteins encoded by viruses. (A) The number of encoded proteins (y-axis) in all 7,959 viral representatives, ranked in descending order. (B) Partitioning of the 7,959 viral proteomes by the number of encoded proteins. The 0.3% viral proteomes with the highest number of proteins (over 500) encode 7.5% of the total number of viral proteins.

2. The discovery of giant viruses

The first giant virus, Acanthamoeba polyphaga mimivirus (APMV), was discovered in 2003 [16]. Its size was unprecedented, being on the scale of small bacteria or archaea cells [17]. Unlike any previously identified virus, APMV could be seen with a light microscope [18,19]. Initially it was mistaken for a bacterium, and recognized as a virus only ten years after its isolation [20]. Up to this day, most of its proteins remain uncharacterized [21,22]. Notably, even more than a decade after the discovery of APMV, the identification of giant viruses is sometimes still involved with confusion, as illustrated in the discovery of the Pandoravirus inopinatum [23] that was initially described as an endoparasitic organism, and Pithovirus sibericum [24] that was also misinterpreted as an archaeal endocytobiont (see discussion in [20,25]).

In the following years from the initial discovery of APMV (2003), many additional giant viral species have been identified and their genomes fully sequenced. Most giant viral genomes have been obtained from large-scale metagenomic sequencing projects covering aquatic ecosystems (e.g., oceans, pools, lakes and cooling wastewater units) [26,27]; others sequenced from samples extracted from underexplored geographical and ecological niches (e.g., the Amazon River, deep seas and forest soils) [28-31]. Despite the accumulation of many more giant virus representatives, the fraction of uncharacterized proteins remains exceptionally high [32]. Many of these uncharacterized proteins were also considered ORFans (i.e., no significant match to any other sequence). However, with proteomes of closely related species, the fraction of ORFans obviously drops. For example, 93% of the Pandoravirus salinus proteins, the first representative of this family [33] were reported as ORFans. However, with the complete proteomes of 5 additional Pandoravirus species (inopinatum,
macleodensis, neocaledonia, dulcis, and quercus) the number of ORFans dropped to 29% (i.e., with a substantial similarity to at least another Pandoravirus protein sequence). Still, the vast majority of Pandoravirus proteins remain uncharacterized.

At present, there are over a hundred giant virus isolates which reveal fascinating and unexpected characteristics. These extreme instances on the viral landscape challenge the current theories on genome size and compactness in viruses, and provide a new perspective on the very concept of a virus and viral origin [4].

3. Definition of giant viruses

Attempts to distinguish giant viruses from other large viruses remain somewhat fuzzy [34,35]. Any definition for giant viruses would necessarily involve some arbitrary threshold, as virus size, whether physical, genomic or proteomic, is clearly a continuum (Figure 2). Giant viruses were initially defined by their physical size as allowing visibility by a light microscope [32]. In this report, we prefer a proteomic definition, even if somewhat arbitrary. We consider giant viruses as Eukaryote-infecting viruses with at least 500 protein-coding genes (Figure 2). Of the 7,959 curated viral genomes (extracted from NCBI Taxonomy complete genomes), 24 represented genomes meet the threshold, among them 5 bacteria-infecting will not be further discussed. The 19 eukaryote-infecting viruses are the genuine giant viruses (Table 1).

Recall that reported proteome sizes are primarily based on automatic bioinformatics tools, which may differ from the experimental expression measurements (e.g., Mimivirus (APMV) [36]). Moreover, physical dimension is not in perfect correlation with the number of proteins or genome size. For example, Pithovirus sibericum, which was recovered from a 30,000-year-old permafrost sample [24], is one of the largest viruses by its physical dimensions (1.5 µm in length and 0.5 µm in diameter). However, it is excluded from this report, as its genome encodes only 467 proteins.

Figure 2. Distribution of viral proteome and genome sizes, colored by host taxonomy. There are 24 represented genomes that meet the threshold of ≥500 proteins among them 5 bacteria-infecting and 19 eukaryote-infecting viruses (dashed red line).
4. Classification of giant viruses and the question of origin

All giant viruses belong to the superfamily of nucleocytoplasmic large DNA viruses (NCLDV), which was substantially expanded following the discoveries of giant viruses [37,38]. The NCLDV superfamily had traditionally been comprised of the following families: Phycodnaviridae, Iridoviridae, Poxviridae, Asfarviridae, and Ascoviridae [39,40], for which a common ancestor had been proposed [41,42]. Following the inclusion of additional giant virus taxonomy groups (Mimiviridae, Pandoravirus and Marseillevirus) into the NCLDV superfamily, there remained only a handful of genes shared by the entire superfamily. Additional disparities in virion shapes and replication modes among NCLDV has led to the conclusion that the superfamily is not necessarily a taxonomic group, and that NCLDV families are more likely to have evolved separately [43-45].

Table 1. Giant viruses

<table>
<thead>
<tr>
<th>Genomea</th>
<th>Accession</th>
<th>Genome length (kb)</th>
<th># of proteins</th>
<th>Hostb</th>
<th>Yearc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mi-Acanthamoeba polyphaga mimivirus</td>
<td>NC_014649</td>
<td>1181.5</td>
<td>979</td>
<td>Pz, Ver</td>
<td>2010</td>
</tr>
<tr>
<td>Mi-Acanthamoeba polyphaga moumouvirus</td>
<td>NC_020104</td>
<td>1021.3</td>
<td>894</td>
<td>Pz, Ver</td>
<td>2013</td>
</tr>
<tr>
<td>Ph-Acanthocystis turfacea chlorella virus 1</td>
<td>NC_008724</td>
<td>288.0</td>
<td>860</td>
<td>Algae</td>
<td>2006</td>
</tr>
<tr>
<td>Mi-Cafeteria roenbergensis virus BV-PW1</td>
<td>NC_014637</td>
<td>617.5</td>
<td>544</td>
<td>Pz</td>
<td>2010</td>
</tr>
<tr>
<td>Pi-Cedravirus A11</td>
<td>NC_032108</td>
<td>589.1</td>
<td>574</td>
<td>Pz</td>
<td>2016</td>
</tr>
<tr>
<td>Ph-Chrysochromulina ericina virus</td>
<td>NC_028094</td>
<td>473.6</td>
<td>512</td>
<td>Algae</td>
<td>2015</td>
</tr>
<tr>
<td>Mi-Megavirus chilensis</td>
<td>NC_016072</td>
<td>1259.2</td>
<td>1120</td>
<td>Pz, Ver</td>
<td>2011</td>
</tr>
<tr>
<td>UC-Mollivirus sibericum</td>
<td>NC_027867</td>
<td>651.5</td>
<td>523</td>
<td>Pz</td>
<td>2015</td>
</tr>
<tr>
<td>Ph-Orpheovirus IHUMI-LCC2</td>
<td>NC_036594</td>
<td>1473.6</td>
<td>1199</td>
<td>Algae</td>
<td>2017</td>
</tr>
<tr>
<td>Pa-Pandoravirus dulcis</td>
<td>NC_021858</td>
<td>1908.5</td>
<td>1070</td>
<td>Pz</td>
<td>2013</td>
</tr>
<tr>
<td>Pa-Pandoravirus inopinatum</td>
<td>NC_026440</td>
<td>2243.1</td>
<td>1839</td>
<td>Pz</td>
<td>2015</td>
</tr>
<tr>
<td>Pa-Pandoravirus macleodensis</td>
<td>NC_037665</td>
<td>1838.3</td>
<td>926</td>
<td>Pz</td>
<td>2018</td>
</tr>
<tr>
<td>Pa-Pandoravirus neocaledonia</td>
<td>NC_037666</td>
<td>2003.2</td>
<td>1081</td>
<td>Pz</td>
<td>2018</td>
</tr>
<tr>
<td>Pa-Pandoravirus quercus</td>
<td>NC_037667</td>
<td>2077.3</td>
<td>1185</td>
<td>Pz</td>
<td>2018</td>
</tr>
<tr>
<td>Pa-Pandoravirus salinus</td>
<td>NC_022098</td>
<td>2473.9</td>
<td>1430</td>
<td>Pz</td>
<td>2013</td>
</tr>
<tr>
<td>Ph-Paramecium bursaria Chlorella virus 1</td>
<td>NC_000852</td>
<td>330.6</td>
<td>802</td>
<td>Algae</td>
<td>1995</td>
</tr>
<tr>
<td>Ph-Paramecium bursaria Chlorella virus AR158</td>
<td>NC_009899</td>
<td>344.7</td>
<td>814</td>
<td>Algae</td>
<td>2007</td>
</tr>
<tr>
<td>Ph-Paramecium bursaria Chlorella virus FR483</td>
<td>NC_008603</td>
<td>321.2</td>
<td>849</td>
<td>Algae</td>
<td>2006</td>
</tr>
<tr>
<td>Ph-Paramecium bursaria Chlorella virus NY2A</td>
<td>NC_009898</td>
<td>368.7</td>
<td>886</td>
<td>Algae</td>
<td>2007</td>
</tr>
</tbody>
</table>

a Families: Mi, Mimiviridae; Ph, Phycodnaviridae; Pi, Pithoviridae; Pa, Pandoraviridae; UC, uncharacterized. b Pz, protozoa; Ver, vertebrates. c Year of genome submission to NCBI.

Two models have been proposed for the evolvement of giant viruses. According to the reductive model, an ancestral cellular genome has reduced in size, leading to dependence of the resulted genome on host cells. The presence of genes carrying cellular functions in almost any giant virus (e.g. translation components) [46] is consistent with this model. An alternative and more accepted theory argues for an expansion model. According to this model, current giant viruses have originated from
smaller ancestral viruses carrying only a few dozens of genes, and through gene duplications and horizontal gene transfer (HGT), have rapidly expanded and diversified [44,47-49]. This model agrees with metagenomic studies and the wave of giant virus discoveries in recent years (e.g. [31]).

To account for the limited number of homologous genes among giant viruses, different HGT mechanisms have been proposed. The amebae host in particular is often described as a melting pot for DNA exchange [50] that leads to chimeric genomes.

Additional important players are virophages, small double-stranded DNA viruses that hitchhike the replication system of giant viruses following coinfection of the host, and are considered parasites of the coinfected giant viruses [51]. A rich network of mobile genetic elements contributes to the host-virus coevolution [52]. Virophages and other mobile elements could facilitate HGT process, including interviral gene transfer, thereby have the potential of shaping the genomes of giant viruses and impact their diversity [20,53,54]. Additional agents that play a role in the rapid dynamics of giant viral genomes are a specific class of canonical transposable elements, which normally act in cellular organisms. The discovery of transpovirons with sequences that are reminiscent to a CRISPR-Cas system propose their contribution to host antivirus [55].

The majority of genes in giant viruses and specifically Mimiviridae have originated from the cells they parasitize mostly amoebal and bacteria. Based on Phylogenetic trees, it is likely that extensive HGT events led to their chimeric genomes. It was also suggested that the spectrum of hosts may be larger than anticipated [56]. Therefore, comparative genomics over giant viruses which infect the same host is unlikely to unambiguously resolve questions of gene origin, namely, whether shared genes have originated from a common viral ancestor or the host. Thus, the degree of similarity among giant viruses infecting different hosts is of a special interest. For example, the phyletic relationship between Mimiviridae (which infect Acanthamoeba) and Phycodnaviridae (infecting algae) was investigated, and it was found that the algae-infecting Chrysochromulina ericina virus (CeV, Table 1) showed moderate resemblance to the algae-infecting Mimivirus [57]. As a result, it was suggested to reclassify CeV as a new clade of Mimiviridae rather than Phycodnaviridae. However, a later discovery of another algae-infecting Phycodnaviridae virus (Heterosigma akashiwo virus, HaV53) has provided a coherent phyletic relationship among Phycodnaviridae, thereby questioning this reclassification [48].

In summary, the taxonomy of giant viruses, as all viruses, is still very unstable, and rapidly updated with new discoveries [30]. The origin and ancestrally of giant viruses have remained controversial with questions of origin are also unresolved [35]. Many newly discovered giant viruses are not compatible with the notion of a single common ancestor, and some giant viruses remain completely undetermined [4].

5. Common features

Despite the ongoing debate on their origin, giant viruses still share some important features. All giant viruses belong to the double-stranded DNA (dsDNA) group, as all NCLDV families. The total genome size of all the giant viruses listed in Table 1 is at least 288 Kbp (Figure 2). These giant viruses are classified into several families: Mimiviridae, Pithoviridae, Pandoraviridae, Phycodnaviridae and the Mollivirus genus [20,24,58].

All amoeba-infecting giant viruses rely on the non-specific phagocytosis by the amoeba host [56]. Interestingly, a necessary condition for phagocytosis is a minimal particle size (~0.6 µm, [59]). As amebae (and related protozoa) are naturally fed on bacteria, it is likely that this minimal size for inducing phagocytosis has become an evolutionary driving force for giant viruses. This fact, together with the largely uncharacterized genomic content of giant viruses, may suggest that much of the content in the genomes of giant viruses serves only for volume filling to increase their physical size.

Giant viruses share not only the cell entry process. When they exist the host cells during lysis, as many as 1,000 virions are released from each lysed host via membrane fusion and active exocytosis [60], which are relatively rare exit mechanisms in viruses.

Other than these genome and cell-biology similarities, other features of giant viruses are mostly family-specific. For example, virion shapes and symmetries, nuclear involvement, duration for the
infection cycle and the steps in virion assembly, substantially vary among viruses from different families [20,61,62].

6. Proteome complexity and functional diversity

The majority of the giant virus proteomes remain with no known function (Figure 3). Actually, the fraction of uncharacterized proteins reaches 65%-85% of all reported proteins in giant viral proteomes, many of them are ORFans. The most striking finding regarding the proteomes of giant viruses is the presence of protein functions that are among the trademarks of cellular organisms, and are never detected among other viruses. To exemplify the complexity of proteome functions in giant viruses, we examine the proteome of the Cafeteria roenbergensis virus (CroV), which infects the marine microzooplankton community in the Gulf of Mexico.

CroV was sequenced in 2010 as the first representative of an algae-infecting virus in the Mimiviridae family. Unexpectedly, despite its affiliation with a recognized viral family, the majority of its proteins show no significant similarity to any other known protein sequence. Of the remaining proteins that show significant BLAST hits to other proteins from all domains of life, 45% are eukaryotic sequences, 22% are from bacteria, and the rest are mostly from other viruses, including other Mimivirus strains. A similar partition of protein origin applies across other members of the Mimiviridae family.

The CroV proteome includes a rich set of genes involved in protein translation [63]. These genes include multiple translation factors, a dozen of ribosomal proteins, tRNA synthetases, and 22 sequences encoding 5 different tRNAs [63]. As the lack of translation potential is considered a hallmark of the virosphere, the presence of translational machinery components raised a debate on the very definition of viruses [64,65]. Similar findings were extended to other giant virus strains of the Tupanvirus genus in the same Mimiviridae family, which were recently isolated in Brazil [66]. The two viruses have 20 ORFs related to tRNA aminoacylation (aaRS), ~70 tRNA sequences decoding the majority of the codons, 8 translation initiation factors, and elongation and release factors. The theory that translation optimization is an evolutionary driving force in viruses [67] may in part explain the curious presence of translation machinery in giant viruses.

In addition to translation, numerous CroV proteins are associated with the transcription machinery. Specifically, The CroV proteome contains several subunits of the DNA-dependent RNA polymerase II, initiation, elongation, and termination factors, the mRNA capping enzyme, and a poly(A) polymerase. Presumably, the virus can activate its own transcription in the viral factory foci in the cytoplasm of its host cell [43].

Another unexpected function detected in CroV is the DNA repair system, specifically of UV radiation damage and base-excision repair. Other DNA-maintenance functions found in CroV include helicase and topoisomerases (type I and II), suggesting a regulation on DNA replication, recombination and chromatin remodeling.

Other rich set of functions related to protein maintenance include chaperons [69] and the ubiquitin-proteasome system [70]. Interestingly, some of these genes seem to be acquired from bacteria (e.g. a homolog of the E. coli heat-shock chaperon). Another rich collection of sugar-, lipid- and amino acid-related metabolic enzymes were also found [17,71] which occupy 13% of the CroV proteome (Figure 3).

It appears that the CroV proteome covers most functions traditionally attributed to cellular organisms, including: protein translation, RNA maturation, DNA maintenance, proteostasis and metabolism. Although CroV exemplifies many widespread functions in giant viruses, each strain has its own unique functional composition. For example, the most abundant group of giant viruses in ocean metagenomes, the Bodo saltans virus (BsV), was recently identified and classified into the same microzooplankton-infecting Mimiviridae family [72]. Unlike the other family members, BsV does not have an elaborate translation apparatus or tRNA genes, but it carries proteins active in cell membrane trafficking and phagocytosis, yet more unprecedented functions discovered in viruses.
Figure 3. Protein function categories in 6 giant virus representatives from three families: Mimiviridae (Mi), Pandoviridae (Pa) and Phycodnaviridae (Ph). In all the proteomes, the majority of proteins are uncharacterized. Short repeated domains are abundant in the proteomes of amebae-infecting giant viruses [71].

7. The emerging ecological view

Viruses are the most abundant entities in nature. In marine and fresh water habitats, there are millions of viruses in each milliliter of water [73]. However, the collection of virus isolates is often sporadic, especially for those without clinical or agricultural relevance. The accelerated pace in the discovery of giant viruses reflects the increasing number of sequencing projects of exotic environmental, including metagenomic projects [31,74].

Giant viruses have been isolated from various environmental niches and distant geographic places, revealing their global distribution and diversity. Current evidence suggests that the representation of giant viruses is underexplored, especially in soil ecosystems [30] and unique ecological niches [75,76]. In fact, ~60% of the giant viral genomes were completed after 2013 (Table 1). Many more virus–host systems that were reported for the last 5 years, are still await isolation and characterization [77].

The hosts of contemporary isolates include mainly protozoa, specifically amoeba (Table 1). However, the prevalence of amoeba as hosts may in part be attributed to sampling bias, specifically to the widespread use of amoebal coculture methods for testing ecological environments [27] (Table 1).

Despite their prevalence, the impact of giant viruses on human health deserve further investigation [21]. A cross-talks between giant viruses and activation of the innate immune cell system in human was reported [78]. Many viruses, including giant viruses were sequenced as part of the large-scale gut microbiome sequencing projects [79] but their composition and dynamic are yet to be determined [80]. Reports on the presence of sequences of giant viruses in the blood, the presence of antibodies against viral proteins, and their association with a broad collection of diseases (e.g. rheumatoid arthritis, unexplained pneumonia, lymphoma) are accumulating [81].

The presence of giant viruses in almost any environment, including extreme niches and manmade sites (e.g., sewage and wastewater plants) suggests that the ecological role of these fascinating viruses and their impact on human health are yet to be determined.

Author Contributions: The authors contributed to writing and manuscript design.

Funding: This research received no external funding.

Acknowledgments: The authors wish to thank the support to M.L. from Yad Hanadiv. We are grateful to Elixir infrastructure tools and resources that were used throughout.
Conflicts of Interest: The authors declare no conflict of interest.

References


