Exploring the diversity and evolution of giant viruses in deep sea sediments using genome-resolved metagenomics

Disa Bäckström

Abstract

Viruses are the most abundant biological entities on this planet, which is impressive considering that they are completely dependent on their hosts for reproduction. Recently the idea of what viruses are has changed dramatically, with the discovery of giant viruses that belong to the Nucleocytoplasmic Large DNA Viruses (NCLDV), such as *Mimiviridae, Marseilleviridae*, and the proposed families Pandoraviruses, and Pithoviruses. Not only are some of these viruses as large as bacteria in size, their genomes also exceed the size of some prokaryotic genomes. The evolutionary path to viral giganticism is not yet fully understood, and several opposing theories have been proposed. The more examples of giant viruses we have to study, the clearer the picture becomes. The rate of discovery, however, is limited by the low capacity of culturing. In an effort to contribute through culture-independent methods, I used genome-resolved metagenomics to retrieve genomes of 23 new members of the NCLDV from deep sea sediment samples that were taken near Loki's Castle hydrothermal vent field. This method has previously been used to study uncultured Bacteria and Archaea, but few successful cases of metagenomic binning of NCLDV have been documented. New methods for refinement and quality control of the binned genomes were developed, combining reads profiling with differential coverage binning, and composition-based cleaning of potentially contaminating sequences. The binned genomes represent several novel clades of NCLDV, the most noteworthy ones distantly related to Pithoviruses and *Marseilleviridae*, and greatly expand their overall diversity. Phylogenetic analysis of their genome content supports the independent evolution of viral giganticism from smaller viruses. Continued use of metagenomics to explore the presence of NCLDV in environmental samples will lead to new insights into their diversity, evolution, and biology.

Dust if you must, but there's not much time,
With rivers to swim, and mountains to climb;
Music to hear, and books to read;
Friends to cherish, and life to lead.

-Rose Milligan 1885

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **Bäckström, D***., Yutin, N*., Jørgensen, S. L., Dharamshi, J., Homa, F., Zaremba-Niedwiedzka, K., Spang, A., Wolf, Y. I., Koonin, E.V., Ettema, T. J.G. (2018). **Novel virus genomes from deep sea sediments expand the ocean megavirome and support independent origins of viral gigantism**; *Submitted.*

*Equal contribution

Reprints were made with permission from the respective publishers.

Other Papers

Yutin, N., **Bäckström, D**., Ettema, T. J. G., Krupovic, M., & Koonin, E. V. (2018). **Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis**. *Virology journal*, *15*(1), 67.

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., **Bäckström, D**., Juzokaite, L., Vancaester, E., Seitz, K. W., Anantharaman, K., Starnawski, P., Kjeldsen, K. U., Stott, M. B., Nunoura, T., Banfield, J. F., Schramm, A., Baker, B. J., Ettema, T. J. G. (2017). **Asgard archaea illuminate the origin of eukaryotic cellular complexity**. *Nature*, *541*(7637), 353.

Contents

Abbreviations

aaRS	aminoacyl tRNA synthethases
bp	base pairs
HGT	horizontal gene transfer
ICTV	International Committee on Taxonomy of Viruses
Kbp	thousand base pairs
LCV	Loki's Castle Virus
Mbp	million base pairs
NCLDV	Nucleocytoplasmic large DNA viruses
ORF	Open reading frame (predicted gene)
ORFan	Predicted gene with no matches in the databases
RDCP	repeat domain containing protein

Preface

I remember the moment in high school when I fell in love with evolution. We were reading an excerpt from *The Blind Watchmaker* by Richard Dawkins, and I was completely taken by a passage about the formation of self-replicating clay crystals and how even inorganic molecules can be subject to selection, and behave in the same way as organic life does. I was dazed by how it all has come together, the vibrant diversity that we see around us.

My next epiphany came in my undergraduate studies, when I was learning about microbial diversity and how most of the diversity of life is not even visible to the naked eye. Every square millimeter of this planet is inhabited by a plethora of organisms, most which are completely uncharacterized. Every single organism has a story to tell about how they came to be, if you look deeply enough. If you zoom into the genetic material, holding the blueprint of each cell, and use sophisticated algorithms, that have been developed by dedicated mathematicians and evolutionary biologists, you can delineate the shared ancestry between everything and piece together how life went from simple bags of self-replicating organic material to this amazing splendor that we find ourselves a part of. That is the mission of researchers who study deep evolution, like we do in the lab where I set out on the academic career.

In this dynamic, ever changing state that we call life I think it is important to accept that nothing lasts forever. People come and go as they find new niches to occupy, lifestyles change with the seasons and the climate, and passions can burn for the longest of times or slowly go out. The forces of life are also not directional. Life operates on a trial and error basis. There is no narrative to life, no intelligent design, destiny or purpose, just matter governed by the principal laws of physics and the process of evolution, and a very long time to both make mistakes and to thrive. It is liberating and terrifying at the same time, right?

With this said I would now like to tell you about what I spent my last three years on.

Introduction

What is a virus?

The living word is divided into three main Domains: Archaea, Bacteria, and Eukaryotes. Viruses are set apart because they cannot multiply on their own (Lwoff 1957). They are tiny, have simple genomes, and they do not encode the components necessary for protein synthesis. The form of the viruses which exists outside of the host cell is called a virion, and it consist the genome of the virus and a protein shell. The virions are not able to do anything until they bump into a cell which they can infect. Then they hijack the host cell's machinery to produce the proteins that make up their particles. Finally, the virions exit the host cell, most often destroying it in the process. Despite this handicap viruses are much more abundant in the environment than cellular life, with one drop of seawater containing millions of virions (Bergh et al. 1989). An illustrating comparison: compared to the estimated number of stars in the universe there are a 10 million-fold more viruses in the oceans (Suttle 2013).

Before the advent of microbiology all disease-causing agents were called virus, from the Greek word meaning '*venom*'. In the end of the 19th-century, it was discovered that some disease-causing agents were so small that they could pass through a filter. Instead of using the term '*ultraviruses*', viruses became the general term for the tiny pathogens, and microbes for the bigger pathogens. As molecular biology progressed it became clear that viruses were different in more ways than their small size. Lwoff (1957) devised four criteria to define viruses: First, they contain only one type of nucleic acids (DNA, or RNA), while cells always contain DNA and different kinds of RNA. Second, when infecting a host cell, it is only the nucleic acids that are important for virus reproduction. Third, viruses are reproduced by transcription and translation of their genetic material inside the host cell, as opposed to binary fission found in cellular reproduction. Fourth, they cannot produce their own energy, in the form of ATP, that cells use for biological reactions.

It is fascinating to think about the origin of viruses, and the huge effect they have on cellular life both on a local and global scale. And it is clear that we have only begun to tap into the wonders and mysteries of viral diversity.

Giant viruses

It took almost a decade from the discovery of the first giant virus, Mimivirus, to get its proper classification. After a pneumonia outbreak at a hospital in Marseille in 1992, medical researchers were trying to identify the cause of the outbreak. Since the bacteria causing pneumonia do not grow well on their own it is common practice to inoculate amoebae (*Acanthamoeba sp*.) with infected environmental samples, and then see what will infect the amoebae. When water from the cooling tower on top of the hospital was tested, a strange coccoid bacterium was observed (Birtles et al. 1997). After many failed isolation attempts, it was put on the shelf and forgotten, until a decade later when electron microscopy revealed that its morphology was similar to the icosahedral virions of nucleocytoplasmic large DNA viruses (NCLDV). The NCLDV are large dsDNA viruses infecting a wide range of eukaryotic hosts, including marine algae (*Phycodnaviridae*), insects (*Ascoviridae*), amphibians (*Iridiviridae*) and mammals (*Poxviridae, Asfaviridae*) (Ivier et al. 2001).

Finally, the viral nature of Mimivirus was recognized, and it baffled the scientific community. The Mimivirus particle size of 400 nm, and its genome size of around 1200 Kbp was unprecedented among viruses (La Scola et al. 2003, Raoult et al. 2004). Following efforts to discover relatives of this strange virus, it became clear that Mimivirus represented a member of a group of viruses that had been completely overlooked. Since then a myriad of novel giant viruses have been discovered, each stranger than the last. They range in virion size from 250 nm, up to 1.5 µm, and genome size of 200 kbp up to over 2 Mbp. Some of their genomes are even larger than some prokaryote genomes (Figure 1). Below is a brief description of the currently characterized giant viruses. There is a lot to be said about these amazing viruses, and all of the intriguing details are not within the scope of this general introduction. For further reading see the reviews by Abergel et al. (2015), Fisher (2016) and Colson et al. (2017).

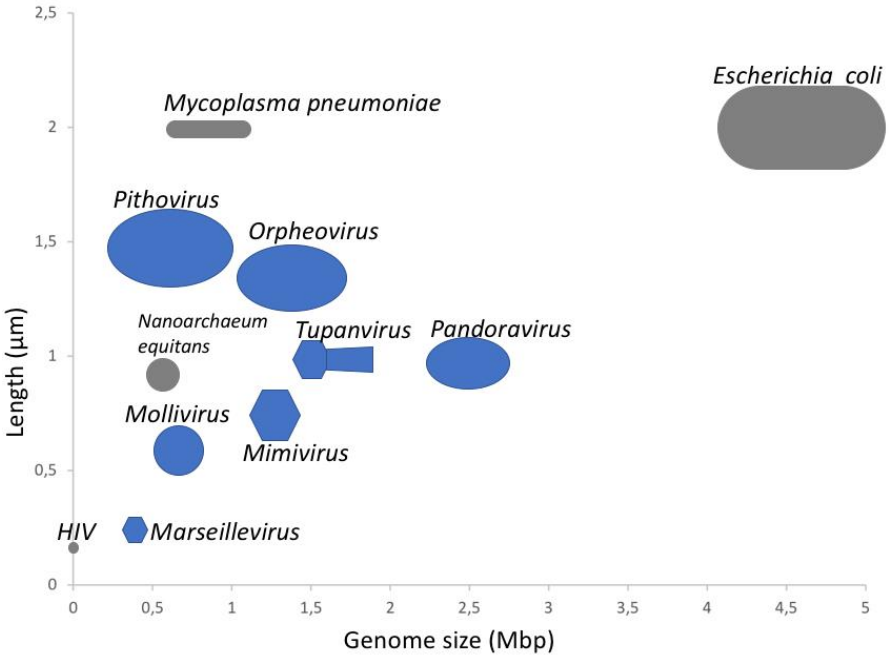

Figure 1. Schematic representation of the sizes and morphology of different giant viruses compared to the virus HIV, the bacteria Escherichia coli and Mycoplasma pneumoniae, and the archaea Nanoarchaeum equitans.

Mimiviridae

Being the first to be discovered, it is not unexpected that Mimiviruses are the most well studied family of giant viruses, comprising over 100 isolated members (Fischer 2016). They have an icosahedral capsid, around 500 nm in diameter, covered in hairy fibrils of 75-140 nm length (La Scola et al. 2003, Yoosuf et al. 2012, Arslan et al. 2011). At one end the icosahedral symmetry is broken by a fibril free structure called the "stargate", where Mimivirus opens to let its contents out once inside the host cytoplasm (Zauberman et al. 2008).

As of 2016 the *Mimiviridae* were divided into three *Acanthamoeba*-infecting lineages, with genomes ranging around 1-1.2 Mbp (Raoult et al. 2004, Yoosuf et al. 2012, Arslan et al. 2011): Lineage A, which includes the original *Acanthamoeba polyphaga mimivirus*, and *Acanthamoeba castelanii mamavirus* (La Scola et al. 2003, Colson et al. 2011a), Lineage B, including *Acanthamoeba polyphaga moumouvirus* (Yoosuf et al. 2012), and Lineage C, including *Megavirus chiliensis* (Arslan et al. 2011). There is also a related branch of smaller viruses that infect marine protists, like *Cafeteria roenbergensis virus* (Colson et al. 2011b), sometimes called "extended Mimiviridae" (Fischer 2016). The *Mimiviridae* family tree has however been

rearranged by the addition of recently discovered lineages, that form new branches in the phylogenies. *Klosneuvirinae* were initially discovered through metagenomic binning (Schulz et al. 2017), and shortly thereafter the Kinetoplastid-infecting *Bodo saltans virus* was found to belong to that lineage (Deeg et al. 2018). The latest addition, Tupanvirus, is a Mimivirus-relative from Brazil with strange morphology: at one side of the icosahedral virion is attached a 550 nm long tail-like structure of unknown function (Abrahao et al. 2018). Clearly there is much left to learn about the *Mimiviridae*.

Marseilleviridae

The first Marseillevirus was discovered from a water cooling tower in Marseille (Boyer et al. 2009). Marseilleviruses are small compared to Mimiviruses, with icosahedral virions of 190-250 nm, covered with short 12 nm fibers or no fibers, and have genomes ranging from 346-386 Kbp (Aherfi et al. 2014a). While they share many features with *Mimiviridae*, including a similar morphology, marker gene phylogenies show that they clearly belong to their own family. Currently five lineages of *Marseilleviridae* have been isolated from all over the world, including France (Boyer et al. 2009), India (Chatterjee and Kondabagil 2017), New Caledonia (Fabre et al. 2017), Tunisia (Aherfi et al. 2014b), and Brazil (Dos Santos et al. 2016, Dornas et al. 2016). This indicates that Marseilleviruses are widespread and diverse.

Pandoraviruses

When researchers were sampling aquatic sediments for new Mimiviruses, they were in for a surprise: a radically different, even larger virus was found infecting the amoebae (Philippe et al. 2013). They were actually first observed in 2008, but were initially regarded as an "extraordinary endocytobiont in Acanthamoeba" with no further classification (Scheid et al. 2008). Ovoid, amphora shaped, the particles of Pandoraviruses are 500 nm wide and 800-1200 nm long, and their genomes are the largest yet observed among viruses. *Pandoravirus salinus* has a 2.5 Mbp long genome with a surprisingly high GC content (61.7%). Among its 2541 predicted genes, only 7% were found to have homologs in the protein databases (Philippe et al. 2013). Presently, six members of the proposed "*Pandoraviridae"* family have been isolated from different corners of the world (Chile, Australia, Germany, France, and New Caledonia), showing a global distribution (Legendre et al. 2017). The genomes of Pandoraviruses are radically different from anything else seen before. 67-73% of their genes lack similarity to anything in the databases. They completely lack the major capsid proteins with double jelly roll fold, that make up the virions of the icosahedral giant viruses. It is currently unknown

what kind of proteins their capsids are made of, or which genes that are involved, and research is ongoing to elucidate the functions of their strange genes (Philippe et al. 2013, Legendre et al. 2017).

Pithoviruses

Pithovirus sibericum was revived from over 30,000 year old Siberian permafrost (Legendre et al. 2014). As with other giant viruses, Pithovirus-like virions had already been observed ten years earlier as a mysterious parasite of amoeba which could not be classified (Michel et al. 2003). It is the virion size record holder at 1.5 µm - similar to the size of *Escherichia coli* cells. Their virions have a similar amphora shape as Pandoraviruses, but with a characteristic striated texture and a honeycomb-like cork in one end. The cork is where the particle opens to fuse their internal membrane with the hosts vacuole membrane and enter the cytoplasm (Legendre et al. 2014). Surprisingly, despite their large virions, Pithovirus genomes are only around 600 Kbp long, and have a low coding density due to 21% the genome consisting of a 150 bp long repeat motif (Legendre et al. 2014).

After the discovery of the "ancient fossil" *P. sibericum*, a contemporary Pithovirus was isolated from sewage water in France: *Pithovirus masiliensis* (Levasseur et al. 2016a). This was followed by isolation of a two-corked relative of Pithovirus named Cedratvirus (Andreani et al. 2016). Despite having slightly smaller genomes (575-589 kb), Cedratvirus genomes have a higher coding density, because they lack the repeats characteristic for Pithoviruses. The repeats seem to have originated after the last common ancestor between Pithoviruses and Cedratviruses, since they are shared between *P. sibericum* and *P. massiliensis*, but not with *Cedratvirus A11* and *Cedratvirus lausanniensis* (Levasseur et al. 2016a, Bertelli et al. 2017). It is a mystery why the repetitive sequences have been conserved. Nonetheless, due to their similarities, Cedratvirus and Pithoviruses have been classified as belonging to the same proposed family "*Pithoviridae"*.

Additional giant viruses

Three additional families of giant viruses are worth brief mentions: Mollivirus, Faustovirus, and Orpheovirus.

Mollivirus sibericum was isolated from the same permafrost sample as *Pithovirus sibericum*, but is totally different. It is shape is spherical, 600 nm in diameter, and it has a genome of 651 Kbp. Mollivirus has a similar replication cycle as Pandoraviruses, with whom it also branches in phylogenetic analyses (Legendre et al. 2015).

Faustovirus was isolated by using by using *Vermamoeba sp.* instead of *Acanthamoeba sp.* It has a 466 Kbp long genome, and 200 nm icosahedral virions (Reteno et al. 2015). The expanding family of Faustoviruses are distantly related to the family *Asfaviridae*, indicating yet another case of a NCLDV lineage independently following the path of viral giganticism (Benamar et al. 2016).

Orpheovirus, found in a French sewer, was also isolated in *Vermamoeba sp* (Andreani et al. 2017). It has a morphology similar to Pandoravirus, with up to 1300 nm long, oval virions, and a genome of 1.5 Mbp. However, its genome content is more similar to Pithoviruses (Andreani et al. 2017), which is interesting since the genome of Orpheovirus is more than twice as large as the Pithoviral genomes.

Classification of giant viruses

The giant viruses are diverse and paraphyletic (Figure 2). The rate of discovery of giant viruses outruns the establishment of an official classification. This far only two families have been approved by the International Committee on Taxonomy of Viruses: *Mimiviridae* and *Marseilleviridae* (ICTV, web catalog 2017:

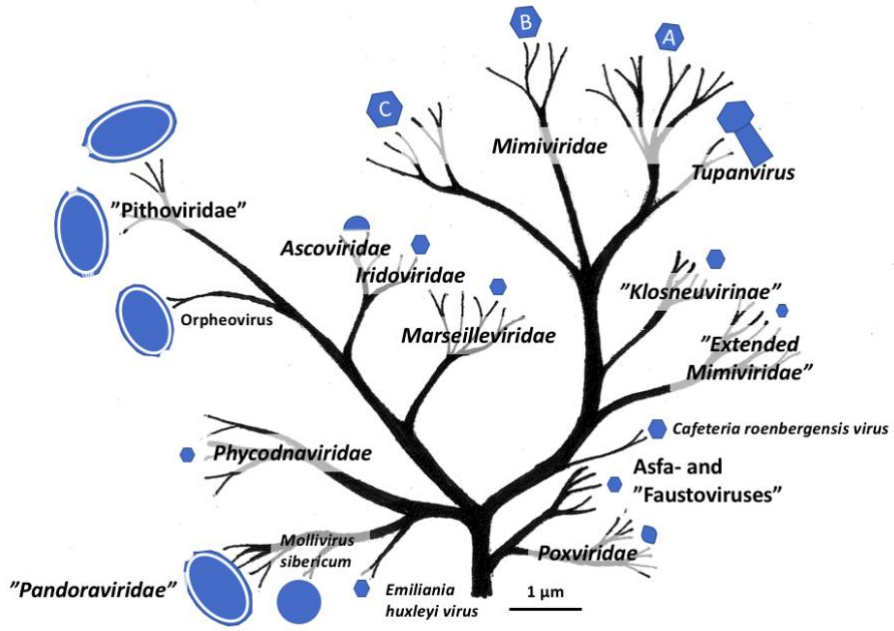

Figure 2. Overview of the NCLDV or "Megavirales" families, including general virion morphologies and average size. The tree is based on the phylogeny in Fisher (2016), with additional lineages added based on phylogenies in Andreani et al. (2017), Schulz et al. (2017), and Abrahao et al. (2018). The placement of the root, branch lengths, and number of branches are all cosmetic.

https://talk.ictvonline.org/taxonomy/).

All of the currently known giant viruses have been shown to share a small set of core genes with the monophyletic group NCLDV, by comparison of conserved marker genes (Yutin et al. 2009, Yutin et al. 2013). Therefore, it has been proposed that the NCLDV order is renamed "Megavirales" (Colson et al. 2013, Colson et al. 2012).

It is clear that we are far from having the whole picture of Megaviral diversity, and the lack of consensus sometimes leads to unnecessary confusion. For example, phylogenetic analyses indicate that Pandoraviruses are related to Phycodnaviruses (Yutin and Koonin 2013), but the researchers that discovered them disagree and claim that they are too different to even be

confidently classified as NCLDV (Legendre et al. 2017). For the intents and purposes of this thesis, to discuss the evolutionary path to viral giganticism, Pandoraviruses are included as a part of NCLDV.

Genome content

Despite the large number of genes with no predicted function, the genome content of giant viruses has been found to reflect their lifestyle. Most giant viruses encode a complete transcription machinery, which they pack inside their virions, and then they replicate and transcribe their DNA directly in the host cytoplasm, followed by the formation of structures called "virus factories" (Claverie et al. 2009). Pandoraviruses and Mollivirus lack several components necessary for DNA replication and many of their genes contain introns, thus they need to invade the host nucleus in the early stages of infection (Legendre et al. 2015, Philippe et al. 2013). Pithoviruses, in contrast to Pandoraviruses, encode a fully functioning transcription machinery and replicate in the host cytoplasm, like *Mimiviridae*. However, functional predictions based on homology do not guarantee that proteins are expressed and contribute to the virus infection cycle, as showed by *Marseilleviridae*. Despite encoding the gene for RNA polymerase, they have been found to not express it, instead employing transcription components from the host nucleus through an unknown process (Fabre et al. 2017).

A central enigma in giant virus research is the source and function of the high number of transcription and translation system components, unprecedented in other viruses. This includes aminoacyl tRNA synthethases (aaRS), tRNAs, and several translation factors (Abrahao et al. 2017). The possible source of these components will be discussed in the next section.

An example of the complexity of giant virus genomes, compared to "regular" viruses, is that they have an extensive mobilome. A mobilome is the collection of mobile genetic elements that is associated with cellular genomes, including transposons, plasmids, and viruses. It is intriguing that the giant viruses have a mobilome of their own. *Acanthamoeba sp.* can be co-infected by smaller viruses, called virophages, that hijack the Mimivirus viral factories and thus inhibit production of the giant virus particles in favor of reproducing the virophage (La Scola et al. 2008, Claverie and Abergel 2009). It has been proposed that Mimiviruses have a defense system against virophages, because of the presence of genomic features similar to the bacterial and archaeal CRISPR-Cas-protection systems against viruses (Levasseur et al. 2016b). In addition, integrated provirophages and mobile plasmid elements, called transpovirons, have been found in some Mimiviral genomes (Desnues et al. 2012). See Diesend et al. (2018) for a review on virophages.

The evolutionary path to viral giganticism

Their unique gene repertoire and cell-like morphology has sparked theories that the "Megavirales" originate from a fourth domain of life, and ignited an intense debate over their classification and evolutionary origins (Boyer et al. 2010a, Colson et al. 2011a & 2012, Moreira & Lopez-Garcia 2009, Philippe et al. 2013). Additionally, it has even been suggested that giant viruses played an important role in eukaryogenesis (Forterre & Gaïa 2016), but this theory is controversial and lacking support (Moreira and Lopez-Garcia 2015). Mainly two scenarios have been debated: One, that giganticism evolved independently in several lineages of smaller viruses among the NCLDV (Yutin et al. 2014, Williams et al. 2011); or two, that they originated from a more complex ancestor and have gone through genome reduction (Claverie and Abergel 2013, Boyer et al. 2010a). With each new genome added to the "Megavirales", more accurate phylogenomic analyses could be conducted, and most of the arguments supporting the "fourth domain of life" or "complex ancestor" scenario have been losing ground. However, the evolutionary path leading to giant genomes and virions is far from resolved. There are still questions left to be answered such as: Why and how have these viruses grown such large genomes and virions? And what is the source and function of their genomic complexity?

Translation system components

Let us first discuss the origin of the translation system components (aaRSs, tRNAs, and translation factors). Are they a remnant of a past as a more complex cellular organism, or were they acquired as an effect of or adaptation to the giant virus lifestyle? The widespread presence of aaRS and tRNA genes in all *Mimiviridae* indicates that the last common ancestor of that family had a more complete set of translation system components, that were gradually lost due to adaptations to different host. This observation could be used to support the "complex ancestor" hypothesis (Raoult et al. 2004, Abrahao et al. 2017). However, careful phylogenomic analysis have given more credibility to the independent acquisition hypothesis (Yutin et al. 2014). To illustrate, the metagenomically binned genomes from Klosenuviruses encode a strikingly large number of aaRSs, tRNAs, and translation factors, but phylogenies of aaRSs in *Klosneuvirinae* and other "Megavirales" show that these genes were not ancestral and monophyletic, but instead originate from various eukaryotes (Schulz et al. 2017). Interestingly, the isolated Klosneuviral representative, *Bodo saltans virus*, has a reduced set of translation system components, and several of the aaRS genes show signs of pseudogenization, which might be an adaptation to the Kinetoplastid host environment (Deeg et al. 2018). It is not possible to prove if an ancestral set of translation system components did exist

for the last common ancestor of the gigantic viruses, and has been gradually lost or replaced by host-derived genes. Thus, the most plausible explanation for the Megaviral translation system components is independent acquisition by horizontal gene transfer from their respective eukaryotic hosts.

"Genomic accordions"

A reason for giant viruses to grow in size, both morphologically and genomically, might have to do with the viral host. Most giant viruses infect amoebae. Amoebae use phagocytosis to feed on Bacteria, and in order to be taken up by that system, the viruses need to be at least 500 nm in diameter (Korn and Weisman 1967). In the case of *Marseilleviridae*, since they are too small to be ingested by phagocytosis, they have been observed to gain entry into the amoeba by aggregating into larger particles, or by interacting with receptors on the host cell surface to induce pinocythosis (Boyer et al. 2009).

Similarly, the genome size might also be an adaptation to the host environment. Amoeba resistant bacteria (ARBs) have been found to have larger genomes than expected for intracellular parasites (Moliner et al. 2010). Similar to giant viruses, they both have an expanded set of repeat domain containing proteins (RDCP), and it is been shown that the number of RDCP genes correlates positively with genome size in "Megavirales" (Shukla et al. 2018). In Mimiviruses the RDCPs are mostly present in the terminal region of the genome (for example observed in Bodo saltans virus; Deeg et al. 2018). Mimivirus has been shown to undergo genome reduction in these regions when grown with Acanthamoeba polyphaga in bacteria free medium, leading to a "naked" virus with no fibrils covering the virion (Boyer et al. 2011). This indicates that expansion of the genome in the form of genes encoding RDCPs plays an important role in the infection cycle of giant viruses, perhaps in competing with the ARBs and/or interacting with the amoebal host's proteins.

The core genome of conserved genes, shared with the last common ancestor of all NCLDV, is only a small fraction of the Megaviral genomes. This indicates that most genes were acquired through HGT. It has been shown that while *Marseilleviridae* and *Poxviridae* have a large proportion of genes derived from their eukaryotic host, in *Mimiviridae* the proportion of genes with bacterial origin is much larger (Boyer et al. 2009, Filée et al. 2008, Filée and Chandler 2010). These observations, and the environment dependent reduction in genome size observed in the experiments by Boyer et al. (2011), led to a proposal that Megaviral genomes behave like accordions (Filée 2013 and 2015). In the accordion-like model of evolution, the common ancestor of all NCLDV was simple, and had a relatively small number of genes. Then, as an adaptation to different host environments, the genomes have gone through several steps of expansion and reduction, leading to mosaic genomes of various sizes, with a patchy distribution of shared genes between lineages.

The mysterious ORFans

Where do all of the unidentified genes come from in giant viruses? A large proportion of the genes of giant viruses, called ORFans, lack homology with anything in the databases and a big part of the ORFans are shared only between members of the same family (Boyer et al. 2010b). Their origin cannot be fully explained by horizontal gene transfer unless we assume that they originate from yet undiscovered, or extinct, organisms. A possible explanation for the ORFans is origin through gene duplication and rapid divergence. This scenario, however, would only be plausible if the genomes of giant viruses have high mutation rates. Comparative studies between the "fossil" *P. sibericum* and the "modern" *P. massiliensis* estimated a mutation rate of $3*10^{-6}$ (Levasseur et al. 2016a), which is faster than some DNA viruses, but slow compared to RNA viruses (Sanjuan et al.2010). However, it is not possible to reliably estimate mutation rates by comparing only two genomes, especially since it is not possible to determine when *P. massiliensis* diverged from *P. sibericum*. Duchene and Holmes (2018) redid the analysis of Levasseur et al. (2016a), taking different relationships between the two Pithoviruses into account. With their more refined mutation rate estimates they made the conclusion that the two viruses could have diverged between 222,000 and $2.35*10^{8}$ years ago, so the mutation rate could either be relatively fast or relatively slow, depending on the timing of their divergence. This result is inconclusive, and it is clear that it makes no sense to draw evolutionary relevant conclusions from mutation rate estimates of giant viruses at this point, so rapid divergence cannot explain the plethora of ORFans.

Legendre (et al. 2017) found that open reading frames (ORFs, another name for predicted genes) were over-predicted in Pandoraviruses because of the high GC content of their genomes. This tricks the gene prediction software to find ORFs in non-coding regions, due to the lack of the AT-rich stop codons. By comparative genomics and proteomics, a stringent annotation protocol was developed, which led to a decrease in the number of predicted genes (40% decrease in *P. salinus*), and discovery of a large number of non-coding RNA transcripts (Legendre et al. 2017). Despite a reduction in predicted ORFs, more than 67% of the Pandoravirus genes are still not similar to anything currently documented in the protein databases. This led Legendre et al. 2017 to suggest a *de novo* gene creation mechanism for the ORFans in Pandoraviruses. *De novo* gene creation is the process when a non-coding part of the DNA becomes transcribed and over time evolves into a functioning protein. As unlikely as it may sound, several studies have confirmed that it happens (see review by Schmitz and Bornberg-Bauer 2017). *De novo* gene creation explains the high number of strain-specific genes that Pandoraviruses have in specific "unstable" regions of their genomes, as shown by extensive comparative genomics and proteomics experiments, however their functions remain a mystery to date (Legendre et al. 2017).

In conclusion, comparative genomics of giant viruses does not paint a straightforward picture about the evolution of NCLDV genome content. The distribution of genes that are interesting for phylogenomic analysis is patchy, and phylogenetic analyses indicate complex histories of lineage specific gains and losses. Differences in virion morphology and genome size between different families are striking. Taken together, the most plausible scenario is that the giganticism and genomic complexity emerged independently in smaller and simpler viruses, through different mechanisms, in order to trick amoebae and other protists into eating them. The path to viral giganticism is far from understood, but with the discovery and characterization of new members of the "Megavirales" the picture becomes clearer.

Virus discovery: culturing and metagenomics

Giant viruses have a global distribution. New members are typically discovered by inoculating cultures of *Acanthamoeba sp.* with environmental samples and looking for signs of virus infection (Pagnier et al. 2013). The method has been refined and automatized to allow for high-throughput isolation of giant viruses, using flow cytometry (Khalil et al. 2016). To this date, over 150 members of giant viruses have been isolated from various environments, including water towers, soil, sewage, rivers, fountains, seawater, and marine sediments (see review by Halary et al. 2016). However, acanthamoebae are not the host of all giant viruses, since some have been found to infect *Vermamoeba vermiformis* (Orpheovirus, and Faustovirus; Andreani et al. 2017, Reteno et al. 2015), and marine protists (*Cafeteria roenbergensis virus*, and *Bodo saltans virus*; Colson et al. 2011b, Deeg et al. 2018). It is likely that many giant viruses remain undiscovered because their host is not known, so the true diversity of giant viruses is not possible to assess through culturing alone.

We have known how to decode nucleotide sequences since the end of the 1970s. During the beginning of this century sequencing techniques have been greatly improved and the costs have dropped significantly (see Goodwin et al. 2016 for a review of modern sequencing techniques). These so-called next-generation sequencing techniques have allowed researchers to study microorganisms and viruses in a new culture-independent way. Sequencing of the 16S or 18S ribosomal RNA gene from the environment have revealed that the majority of microbial diversity has yet to be characterized, and that more than half of the major bacterial and archaeal lineages are represented by only environmental sequences (Hug et al. 2016).

Large scale metagenomic screening of viral diversity indicates that a big part of the Earth's virome remains unexplored (Paez-Espino 2016), and recent advances in sequencing and metagenomics have made it possible to discover new giant viruses through culture independent methods (Schulz et al. 2017, Hingamp et al. 2013). A panel of virology experts recently made a consensus statement that viruses known from only metagenomic sequences should be allowed to be included in the official ICTV taxonomy (Simmonds et al. 2017).

Giant viruses tend to be overlooked in viral metagenomic studies since samples are typically filtered according to the preconception of common virion sizes (Halary et al. 2016). This means that it is necessary to screen full metagenomes for giant virus discovery, and it is necessary to devise methods for separating the giant virus sequences from the non-viral sequences. To aid in giant virus discovery, the program MG-Digger was developed to screen metagenomes for sequences belonging to giant viruses (Verneau et al. 2016), and it has been used to detect for example Orpheovirus in a mine drainage metagenome (Andreani et al. 2018). However, that method will only detect viral sequences that contain previously identified marker genes, and since the

giant virus genomes are too large to be present in metagenomes as one contiguous sequence it will only detect small parts of the genomic information that is present in the samples. A method to retrieve more complete genomes from metagenomes is called 'genome-resolved metagenomics', and it has become a common method to obtain genomes from uncharacterized bacteria and archaea (Albertsen et. al. 2013, Brown et al. 2015, Spang et al. 2017). Up to this date, only one study of Megaviral genomes retrieved by genome resolved metagenomics has been published, and it was from a simple wastewater sludge metagenome (Schulz et al. 2017). The goal of my project was to reconstruct Megaviral genomes from more complex deep-sea sediment metagenomes.

Genome resolved metagenomics

To find out which sequences of a metagenome that belong to a microorganism a procedure called "binning" is used. Imagine the genome of an individual microbe as a jigsaw puzzle. When doing sequencing, the genome it is broken into smaller parts. In a metagenome, the pieces of hundreds of puzzles are mixed. Some of the puzzles are present in many copies, while some are rare. Some of the puzzles are almost identical to others and differ only in a few small parts. Assembly programs are used to piece this information together into a picture, as complete as possible. Only the easy parts of the puzzles can be pieced together. The programs will not know what to do with the pieces that are identical to pieces in other puzzles, or even to parts within the same puzzle. There is also the problem of some pieces disappearing along the way. When there is not enough information the assembly programs will not be able to continue. It is extremely rare to assemble a whole genome at once. What you end up with is a lot of sequence fragments, contigs, and you want to know which individual genome they belong to. This is how binning works. Figure 3 shows a simplified workflow of the procedure from environmental sample to metagenomically binned genome.

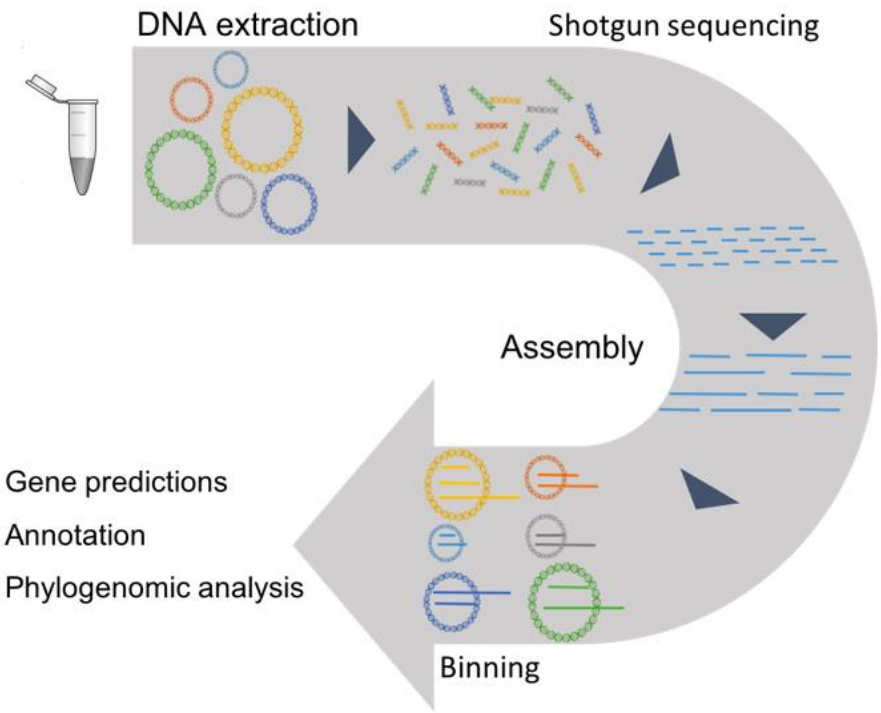

Figure 3. The genome resolved metagenomics workflow: Total DNA is extracted from an environmental sample. The individual microbial genomes, shown as circular DNA molecules, are fragmented during the shotgun sequencing step. The decoded sequence reads, shown as short blue lines, are assembled by assembly programs into contigs (longer blue lines). Binning is the procedure to determine which contigs that belong to which genome.

To separate contigs into binned genomes we make use of algorithms that sort the sequences according to their properties, such as coverage and composition. For example, the program ESOM clusters sequences according to their composition (tetranucleotide frequencies), and visualizes this in the form of a map where "islands" can be manually isolated. The islands represent sequences that have similar composition and are assumed to belong to the same or closely related genomes (Dick et al. 2009). A problem with this method is that some sequences in the same genome can have a different composition then the rest, and then they will end up in another island on the map, and be missed. Vice versa, sequences that happen to have a similar composition, but belong to another species, may be including by mistake.

It is more reliable to bin sequences by combining information about both composition and coverage. Coverage is the number of unique sequence reads that overlap in the assembly of a given sequence. The average coverage of a contig (number of reads * read length / contig size) corresponds to the

abundance of that sequence in the sample, which corresponds to the abundance of the microorganism that the sequence belongs to. If you have several samples from the same environment the microbial composition will differ from sample to sample. If microorganisms of the same strain are presented in several samples, the sequences belonging to their genomes should have similar coverage within the same sample. Information about the co-variation of coverage across samples can thus be used for binning. This is called differential coverage binning, and is used for example by the program CONCOCT (Alneberg et al. 2014).

Since binning is an automatic process, it is common that sequences that do not belong in the genomes of interest get sorted into the bins because they have similar composition or coverage. It is therefore necessary to manually check the binned sequences to see if it makes sense, and remove any contaminating sequences. It is also common that genomes of closely related microbes get sorted into the same bin. This can be detected by looking at the number of conserved single-copy marker genes in the bins. With some luck, the closely related genomes differ slightly in coverage or composition and they can be separated during the more sensitive bin refinement step.

A program for doing bin refinement is Mmgenome, which uses information about coverage, composition (in the form of GC content), presence of conserved marker genes, and reads linkage between contigs (Karst et al. 2016). All this information is visualized on graphs, so that outliers can be removed by selecting sequence clusters. The refinement step is manual, and at this step binning becomes almost artisanal work. You could spend months on refining only one genome, but this is not effective, so it is necessary to define some kind of quality assessment criteria to decide when the bins are clean enough. For example, the bin refinement can be considered done when removing more outliers leads to a decrease in the completeness of the genome, judged by the number of single-copy marker genes. The process of bin refinement and quality assessment is discussed extensively in the supplementary methods of Paper I.

Binning of Megaviral genomes has not been done as extensively as binning of bacterial and archaeal genomes, and it poses some challenges. First of all, the set of conserved single-copy marker genes shared by all giant viruses is small. This gene set is used to judge the completeness, redundancy, and contamination of a binned genome, and when you have only a few conserved marker genes this measure has low resolution. Nine mostly conserved genes have been identified for the NCLDV, but they are not present in all families (Yutin et al. 2010). For future large scale NCLDV binning projects it would be useful to identify a bigger conserved marker gene set for each NCLDV family exclusively. It would make bin refinement and quality assessment easier, even though unexpected exceptions will always exist.

Another factor that can cause problems is the presence of repetitive sequences in the genomes, which is common in for example Pithoviruses

(Legendre et al. 2014). Assembly programs do not handle repetitive sequences well, leading to fragmented assemblies. In my binning endeavors I encountered repetitive sequences in many of the genomes, especially in those related Pithoviruses (see additional file 6 of the manuscript). I managed to improve the assemblies by using the reads profiling software CLARK (Ounit et al. 2015), to extract reads from several metagenomes that matched the binned genomes and co-assemble them. This removed the noise of irrelevant reads in the metagenomes, while at the same time adding information from several samples. Although it was not within the scope of this thesis, it would have been interesting to further explore the repetitive sequences in the binned genomes, to find out how conserved they are between different members. The presence of lineage-specific repeat motifs could perhaps be used for identification, and as another factor in bin refinement.

Paper summary

We were interested in the presence of novel members of "Megavirales" in deep sea sediments from Loki's Castle, a hydrothermal vent area located between Norway and Greenland. Samples from the same site have previously been found to harbor a rich diversity of uncharacterized Archaea (Jørgensen et al. 2013, Spang et al. 2015, Zaremba-Niedzwiedzka et al. 2017). Screening with the NCLDV marker gene for DNAP revealed a huge diversity of novel NCLDV lineages in the Loki's Castle metagenomes.

We retrieved 23 novel NCLDV genomes of high quality. Most of the novel Loki's Castle Virus (LCV) genomes belong to novel lineages only distantly related to Pithoviruses and Orpheovirus, greatly expanding the diversity of this group. Four of the binned genomes are distantly related to *Marseilleviridae*, but all of them have larger genomes. Additionally, two genomes were related to *Klosneauvirinae* and one distantly related to *Iridoviridae*. All of this shows that uncharacterized giant viruses reside in deep-sea sediments.

I did the metagenomic binning and assessment of NCLDV diversity in Loki's Castle and other metagenomes. Since only one case of metagenomic binning of giant viruses has been published this far (Schulz et al. 2017), no standard or recommended methods exist, and I had to develop novel procedures for identifying contaminating sequences, determining the quality of the genome bins, and improving the assemblies. I combined different binning methods with reads profiling, and cleaning based on composition, coverage, and marker gene content, followed by quality assessment by preliminary sequence annotation and phylogenetic analysis of conserved marker genes.

My collaborators Natalya Yutin and Eugene Koonin are experienced in evolutionary analysis of giant viruses, and they did the final annotations, phylogenomic analysis of marker genes and translation system components, and protein sequence clustering. Phylogenetic analysis of the translation system components found in the LCVs showed a polyphyletic origin of these genes, with many examples of horizontal gene transfer from bacteria or members of the *Mimiviridae*.

We have expanded the diversity of "Megavirales", but even though metagenomics revealed the existence of novel viruses and that sequence analysis hints at their genomic potential, it is unclear how complete the genomes are. We also know nothing about their morphology or their lifestyle.

The picture will be incomplete until these viruses are studied in the lab. A remaining mystery concerns who their hosts are, since very few eukaryotic 18S rRNA were found in the Loki's Castle metagenomes, and none of them were from amoeba. It is not possible to determine if the LCVs are hosted by some undetected eukaryote in the sediments, or if they originate from the water column and remain in our samples as a "fossil record". Either scenario could be true for different LCVs, since they represent diverse lineages. Further studies of the deep-sea sediment virosphere are called for. There is no doubt that it will bring many surprises.

Svensk sammanfattning

Vad tänker du på när du hör ordet ”virus”? Antagligen något som är väldigt smått - inte mycket mer än lite genetiskt material omgivet av ett proteinskal. Minimala parasiter som är helt beroende av maskineriet i cellerna som de infekterar. Denna definition av virus var länge standard, men under de senaste 15 åren har bilden ändrats drastiskt. Nu vet vi att det finns jättevirus som är större än vissa bakterier, och förvånande komplexa.

Det första jätteviruset upptäcktes 2003, av Bernard La Scola vid Marseilles universitet. Kollegan Richard Birtles i England hade tio år tidigare identifierat en underlig bakterie, som han isolerat tillsammans med amöbor från ett vattentorn på ett sjukhus efter ett utbrott av lunginflammation. Hur mycket han än försökte så lyckades han inte få bakterien att växa på egen hand, så han skickade ett prov till La Scola, som observerade att det inte alls var en bakterie, utan ett virus som var större än något som setts tidigare. Inte bara partikelstorleken var ovanlig, de hade också väldigt många gener. Viruset döptes till Mimivirus, och snart började fler och fler sorter av jättevirus att upptäckas, när forskarna visste vad de skulle leta efter. Det var som om forskarna hade öppnat Pandoras ask, och allt större och allt konstigare virus började välla fram, som till exempel Pandoravirus, som har rekord i flest gener, och *Pithovirus sibericum*, det största viruset vi känner till, som återupplivades från 30 000 år gammal Sibirisk permafrost.

Hur blev de så stora och komplexa, och var fick de alla sina gener ifrån? Det är fortfarande inte helt klargjort, men ju fler jättevirus som vi känner till och kan jämföra med varandra, desto klarare blir bilden.

Det klassiska sättet att studera mikroorganismer är genom att odla dem på labb, men det är bara cirka 1% av jordens mikroorganismer som går studera på det sättet, eftersom vi inte har listat ut vad resten behöver för förhållanden för att trivas. På senare tid har metoderna för att extrahera och avkoda all DNA direkt från en specifik miljö, så kallad metagenomik, gjort det möjligt att studera hela den mikrobiologiska mångfalden. Min forskning har gått ut på att använda metagenomik för att hitta nya jättevirus i en miljö där de inte har studerats förut, nämligen djuphavssediment från en plats som ligger 3000 meter under havsytan mellan Norge och Grönland.

Jag fiskade fram DNA-sekvenser tillhörande jättevirus genom att söka i metagenomen efter en av deras essentiella gener. Därefter gjorde jag ”binning”, som innebär att gruppera sekvenserna utefter deras egenskaper för att lista ut vilka sekvenser som tillhör individuella genom. Även om många

nya bakterier och arkéer har blivit studerade med hjälp av metagenomisk binning, så fanns det inte väletablerade metoder för jättevirus. Jag blev tvungen att kombinera flera olika metoder för att få bäst resultat, och utveckla nya kriterier för kvalitetskontroll och bortrensning av sekvenser som hamnat fel. I slutändan tog jag fram sekvenser från 23 nya virus, som är avlägset släkt med flera olika jättevirus. Fylogenetisk analys visar att många av dem tillhör helt nya familjer, och deras gen-innehåll stöder teorin om att jättevirus har utvecklats flera gånger självständigt från olika familjer av mindre virus.

Det är högst oklart om dessa virus är aktiva djupt nere i sedimenten, eftersom vi inte kunde hitta några sekvenser från amöbor eller andra eukaryoter som jättevirus vanligtvis infekterar. Det är möjligt att de finns bevarade där som ”fossil”. Förhoppningsvis tar någon sig an utmaningen att återuppliva dem, liksom Pithovirus, så att vi kan lära oss mer om deras biologi.

Acknowledgments

A huge thanks to **Thijs**, for welcoming me into Ettemalab and giving me this academic opportunity. I have learned a lot, both about metagenomics and myself! Thank you **Anja,** co-supervisor one, for noticing those metagenomic sequences that became the start of my project. Thank you **Kasia**, co-supervisor two, for all of the good talks, both science and life-related.

A big thanks to my collaborators **Eugene Koonin** and **Natalya Yutin** for your enthusiastic and thorough work! I'm really glad that I was able to meet **Eugene** at the Ringberg Symposium on Giant Virus Biology, and see that you are not only a brilliant scientist but also a friendly and nice person. Thank you **Matthias Fischer** for organizing the symposium. It gave me a push of inspiration and motivation that was well needed.

My colleagues have all been important, but I need to thank a few of you a bit extra. **Jennah**, thank you for so generously sharing your metagenomes with me. **Joran**, thank you for guiding me through the jungle of binning and bin cleaning. **Felix**, thank you for all of the bioinformatics support and guided programming sessions. **Jennah**, **Felix**, and **Tom**, it's been great to share an office with you! Thank you **Prune**, **Daniel**, and **Arvid** for all of the pre-work or lunch-break ice skating sessions! To the rest of you, it's been a pleasure to have more or less crazy conversations with you all over lunches, fika, and the occasional beer: **Anna-Maria**, **Claudia**, **Courtney**, **Dani**, **Eva**, **Felix**, **Henning**, **Jennah**, **Jonathan**, **Joran**, **Jun-Hoe**, **Laura**, **Maria**, **Martha**, **Max**, **Tom**, **Will**, the rest of **Molevo**, and past colleagues that have moved on to other places.

Last but not least, to my friends, family, and loved ones. I cannot list all of you, but know that if you read this then you are important. I will however mention three people who had to patiently listen to my never-ending rants about my PhD work the most. My sister **Hanna**, I'm so glad that we live so close to each other that we can have dinner whenever. My dear flat mate **Matilda**, it is so calming to have you around, regardless if it's only for a few words over coffee in the morning, or for long mind-clearing talks. **Johan**, my favorite American, thank you for your never ending support. Words are not enough to describe how grateful I am to have you in my life. Finally, thanks a lot for proof reading, **Johan** and **Hanna**!

References

Abergel, C., Legendre, M., & Claverie, J. M. (2015). The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollvirus. *FEMS microbology reviews*, *39*(6), 779-796.

Abrahão, J. S., Araujo, R., Colson, P., & La Scola, B. (2017). The analysis of translation-related gene set boosts debates around origin and evolution of mimiviruses. *PLoSgenetics*, *13*(2), e1006532.

Abrahão, J., Silva, L., Silva, L. S., Khalil, J. Y. B., Rodrigues, R., Arantes, T., ... & Ribeiro, B. (2018). Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nature comunications*, *9*(1), 749.

Aherfi, S., La Scola, B., Pagnier, I., Raoult, D., & Colson, P. (2014a). The expanding family Marseilleviridae. *Virology*, *467*, 27–37.

Aherfi, S., Boughalmi, M., Pagnier, I., Fournous, G., La Scola, B., Raoult, D., & Colson, P. (2014b). Complete genome sequence of Tunisvirus, a new member of the proposed family Marseilleviridae. Archives of virology, 159(9), 2349-2358.

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology*, *31*(6), 533.

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., ... & Quince, C. (2014).Binning metagenomic contigs by coverage and composition. *Nature methods*, *11*(11), 1144-1146.

Andreani, J., Aherfi, S., Bou Khalil, J. Y., Di Pinto, F., Bitam, I., Raoult, D., ... & La Scola, B. (2016). Cedratvirus, a double-cork structured giant virus, is a distant relative of pithoviruses. *Viruses*, *8*(11), 300.

Andreani, J., Khalil, B., Yaacoub, J., Baptiste, E., Hasni, I., Michelle, C., ... & La Scola, B. (2017). Orpheovirus IHUMI-LCC2: A new virus among the giant viruses. *Frontiers in microbiology*, *8*, 2643.

Andreani, J., Verneau, J., Raoult, D., Levasseur, A., & La Scola, B. (2018). Deciphering viral presences: two novel partial giant viruses detected in marine metagenome and in a mine drainage metagenome. *Virology journal*, *15*(1), 66.

Arslan, D., Legendre, M., Seltzer, V., Abergel, C., & Claverie, J. M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proceedings of the National Academy of Sciences*, *108*(42), 17486-17491.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pyshkin, A. V. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, *19*(5), 455-477.

Benamar, S., Reteno, D. G., Bandaly, V., Labas, N., Raoult, D., & La Scola, B. (2016). Faustoviruses: comparative genomics of new Megavirales family members. *Frontiers in microbiology*, *7*, 3.

Bergh, Ø., BØrsheim, K. Y., Bratbak, G., & Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature*, *340*(6233), 467.

Bertelli, C., Mueller, L., Thomas, V., Pillonel, T., Jacquier, N., & Greub, G. (2017). Cedratvirus lausannensis–digging into Pithoviridae diversity. *Environmental microbiology*, *19*(10), 4022-4034.

Birtles, R. J., Rowbotham, T. J., Storey, C., Marrie, T. J., & Raoult, D. (1997). Chlamydia-like obligate parasite of free-living amoebae. *The Lancet*, *349*(9056), 925-926.

Boyer, M., Yutin, N., Pagnier, I., Barrassi, L., Fournous, G., Espinosa, L., ... & Suzan-Monti, M. (2009). Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proceedings of the National Academy of Sciences*, *106*(51), 21848-21853.

Boyer, M., Madoui, M. A., Gimenez, G., La Scola, B., & Raoult, D. (2010a). Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4th domain of life including giant viruses. *PLoS One*, *5*(12), e15530.

Boyer, M., Gimenez, G., Suzan-Monti, M., & Raoult, D. (2010b). Classification and determination of possible origins of ORFans through analysis of nucleocytoplasmic large DNA viruses. *Intervirology*, *53*(5), 310-320.

Boyer, M., Azza, S., Barrassi, L., Klose, T., Campocasso, A., Pagnier, I., ... & Desnues, C. (2011). Mimivirus shows dramatic genome reduction after intraamoebal culture. *Proceedings of the National Academy of Sciences*, *108*(25), 10296-10301.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., ... & Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, *523*(7559), 208.

Chatterjee, A., & Kondabagil, K. (2017). Complete genome sequence of Kurlavirus, a novel member of the family Marseilleviridae isolated in Mumbai, India. *Archives of virology*, *162*(10), 3243-3245.

Claverie, J. M., & Abergel, C. (2013). Open questions about giant viruses. In *Advances in virus research* (Vol. 85, pp. 25-56). Academic Press.

Claverie, J. M., & Abergel, C. (2009). Mimivirus and its virophage. *Annual review of genetics*, *43*, 49-66.

Claverie, J. M., Abergel, C., & Ogata, H. (2009). Mimivirus. In *Lesser Known Large DSDNA Viruses* (pp. 89- 121). Springer, Berlin, Heidelberg.

Colson, P., Yutin, N., Shabalina, S. A., Robert, C., Fournous, G., La Scola, B., … & Koonin, E. V. (2011a). Viruses with more than 1,000 genes: Mamavirus, a new Acanthamoeba polyphaga mimivirus strain, and reannotation of Mimivirus genes. *Genome Biology and Evolution*, *3*(0), 737–42.

Colson, P., Gimenez, G., Boyer, M., Fournous, G., & Raoult, D. (2011b). The giant Cafeteria roenbergensis virus that infects a widespread marine phagocytic protist is a new member of the fourth domain of Life. *PLoS One*, *6*(4), e18935.

Colson, P., De Lamballerie, X., Fournous, G., & Raoult, D. (2012). Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. *Intervirology*, *55*(5), 321-332.

Colson, P., De Lamballerie, X., Yutin, N., Asgari, S., Bigot, Y., Bideshi, D. K., ... & La Scola, B. (2013). "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Archives of virology*, *158*(12), 2517-2521.

Colson, P., La Scola, B., Levasseur, A., Caetano-Anollés, G., & Raoult, D. (2017). Mimivirus: leading the way in the discovery of giant viruses of amoebae. *Nature Reviews Microbiology*, *15*(4), 243.

Deeg, C. M., Chow, C. E. T., & Suttle, C. A. (2018). The kinetoplastid-infecting Bodo saltans virus (BsV), a window into the most abundant giant viruses in the sea. *eLife*, *7*, e33014.

Desnues, C., La Scola, B., Yutin, N., Fournous, G., Robert, C., Azza, S., ... & Raoult, D. (2012). Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proceedings of the National Academy of Sciences*, 201208835.

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., & Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome biology*, *10*(8), R85.

Diesend, J., Kruse, J., Hagedorn, M., & Hammann, C. (2018). Amoebae, giant viruses, and virophages make up a complex, multilayered threesome. *Frontiers in cellular and infection microbiology*, *7*, 527.

Dornas, F. P., Assis, F. L., Aherfi, S., Arantes, T., Abrahão, J. S., Colson, P., & La Scola, B. (2016). A Brazilian Marseillevirus is the founding member of a lineage in family Marseilleviridae. *Viruses*, *8*(3), 76.

Dos Santos, R. N., Campos, F. S., De Albuquerque, N. R. M., Finoketti, F., Côrrea, R. A., Cano-Ortiz, L., ... & Franco, A. C. (2016). A new marseillevirus isolated iSouthern Brazil from Limnoperna fortunei. *Scientific reports*, *6*, 35237.

Duchêne, S., & Holmes, E. C. (2018). Estimating evolutionary rates in giant viruses using ancient genomes. *Virus evolution, 4(1)*, vey006.

Fabre, E., Jeudy, S., Santini, S., Legendre, M., Trauchessec, M., Couté, Y., ... & Abergel, C. (2017). Noumeavirus replication relies on a transient remote control of the host nucleus. *Nature Communications*, *8*, 15087.

Filée, J. (2013). Route of NCLDV evolution: the genomic accordion. *Current opinion in virology*, *3*(5), 595-599.

Filée, J. (2015). Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution. *Frontiers in microbiology*, *6*, 593.

Filée, J., Pouget, N., & Chandler, M. (2008). Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evolutionary Biology*, *8*(1), 320.

Filée, J., & Chandler, M. (2010). Gene exchange and the origin of giant viruses. *Intervirology*, *53*(5), 354-361.

Fischer, M. G. (2016). Giant viruses come of age. *Current Opinion in Microbiology*, *31*, 50–57.

Forterre, P., & Gaïa, M. (2016). Giant viruses and the origin of modern eukaryotes. *Current Opinion in Microbiology*, *31*, 44–49.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333.

Halary, S., Temmam, S., Raoult, D., & Desnues, C. (2016). Viral metagenomics: are we missing the giants? *Current Opinion in Microbiology*, *31*, 34–43.

Hingamp, P., Grimsley, N., Acinas, S. G., Clerissi, C., Subirana, L., Poulain, J., ... & Faust, K. (2013). Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *The ISME journal*, *7*(9), 1678.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C.

J., ... & Suzuki, Y. (2016). A new view of the tree of life. *Nature microbiology*, *1*, 16048.

Iyer, L. M., Aravind, L., & Koonin, E. V. (2001). Common origin of four diverse families of large eukaryotic DNA viruses. *Journal of virology*, *75*(23), 11720-11734.

Jørgensen, S. L., Thorseth, I. H., Pedersen, R. B., Baumberger, T., & Schleper, C. (2013). Quantitative and phylogenetic study of the Deep Sea Archaeal Group in sediments of the Arctic mid-ocean spreading ridge. *Frontiers in microbiology*, *4*, 299.

Karst, S. M., Kirkegaard, R. H., & Albertsen, M. (2016). mmgenome: a toolbox for reproducible genome extraction from metagenomes. *bioRxiv*, 059121.

Khalil, J. Y., Robert, S., Reteno, D. G., Andreani, J., Raoult, D., & La Scola, B. (2016). High-throughput isolation of giant viruses in liquid medium using automated flow cytometry and fluorescence staining. *Frontiers in microbiology*, *7*, 26.

Korn, E. D., & Weisman, R. A. (1967). Phagocytosis of Latex Beads by Acan thamoeba: II. Electron Microscopic Study of the Initial Events. *The Journal of cell biology*, *34*(1), 219-227.

La Scola, B., Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., ... & Raoult, D. (2003). A giant virus in amoebae. *Science*, *299*(5615), 2033-2033.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., … & Raoult, D. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature*, *455*(7209), 100–4.

Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, a., … & Claverie, J.-M. (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proceedings of the National Academy of Sciences*, *111*(11), 4274–4279.

Legendre, M., Lartigue, A., Bertaux, L., Jeudy, S., Bartoli, J., Lescot, M., … Claverie, J.-M. (2015). In-depth study of Mollivirus sibericum , a new 30,000-y-old giant virus infecting Acanthamoeba. *Proceedings of the National Academy of Sciences*, *112*(38), 201510795. http://doi.org/10.1073/pnas.1510795112

Legendre, M., Fabre, E., Poirot, O., Jeudy, S., Lartigue, A., Alempic, J. M., ... & Coute, Y. (2017). Diversity and evolution of the emerging Pandoraviridae family. *bioRxiv*, 230904.

Levasseur, A., Andreani, J., Delerce, J., Bou Khalil, J., Robert, C., La Scola, B., & Raoult, D. (2016a). Comparison of a modern and fossil Pithovirus reveals its genetic conservation and evolution. *Genome biology and evolution*, *8*(8), 2333-2339.

Levasseur, A., Bekliz, M., Chabrière, E., Pontarotti, P., La Scola, B., & Raoult, D. (2016b). MIMIVIRE is a defence system in mimivirus that confers resistance to virophage. *Nature*, *531*(7593), 249.

Michel, R., Schmid, E. N., Hoffmann, R., & Müller, K. D. (2003). Endoparasite KC5/2 encloses large areas of sol-like cytoplasm within Acanthamoebae. Normal behavior or aberration?. *Parasitology research*, *91*(4), 265-266.

Moliner, C., Fournier, P. E., & Raoult, D. (2010). Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. *FEMS microbiology reviews*, *34*(3), 281-294.

Moreira, D., & López-García, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nature Reviews. Microbiology*, *7*(4), 306–11.

Moreira, D., & López-García, P. (2015). Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes?. *Phil. Trans. R. Soc. B*, 370(1678), 20140327.

Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics*, *16*(1), 236.

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., ... & Kyrpides, N. C. (2016). Uncovering Earth's virome. *Nature*, *536*(7617), 425.

Pagnier, I., Reteno, D. G. I., Saadi, H., Boughalmi, M., Gaia, M., Slimani, M., … La Scola, B. (2013). A decade of improvements in mimiviridae and marseilleviridae isolation from amoeba. *Intervirology*, *56*(6), 354–363.

Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., … & Abergel, C. (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*, *341*(6143), 281–6.

Reteno, D. G., Benamar, S., Khalil, J. B., Andreani, J., Armstrong, N., Klose, T., … & La Scola, B. (2015). Faustovirus, an asfarvirus-related new lineage of giant viruses infecting amoebae. *Journal of Virology*, *89*(13), 6585–94.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., … & Claverie, J.-M. (2004). The 1.2-megabase genome sequence of Mimivirus. *Science*, *306*(5700), 1344–50.

Sanjuan, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral mutation rates. *Journal of virology*, *84*(19), 9733-9748.

Scheid, P., Zöller, L., Pressmar, S., Richard, G., & Michel, R. (2008). An extraordinary endocytobiont in Acanthamoeba sp. isolated from a patient with keratitis. *Parasitology research*, *102*(5), 945-950.

Schmitz, J. F., & Bornberg-Bauer, E. (2017). Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. *F1000Research*, *6*.

Schulz, F., Yutin, N., Ivanova, N. N., Ortega, D. R., Lee, T. K., Vierheilig, J., ... & Kyrpides, N. C. (2017). Giant viruses with an expanded complement of translation system components. *Science*, *356*(6333), 82-85

Shukla, A., Chatterjee, A., & Kondabagil, K. (2018). The number of genes encoding repeat domain-containing proteins positively correlates with genome size in amoebal giant viruses. *Virus evolution*, *4*(1), vex039.

Simmonds, P., Adams, M. J., Benkő, M., Breitbart, M., Brister, J. R., Carstens, E. B., ... & Hull, R. (2017). Consensus statement: virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, *15*(3), 161.

Suttle, C. A. (2013). Viruses: unlocking the greatest biodiversity on Earth. *Genome*, *56*(10), 542-544.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., ... & Ettema, T. J. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, *521*(7551), 173-179.

Spang, A, Caceres, EF, & Ettema, T. J. 2017. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science*. *357*(6351).

Verneau, J., Levasseur, A., Raoult, D., La Scola, B., & Colson, P. (2016). MG-Digger: an automated pipeline to search for giant virus-related sequences in metagenomes. *Frontiers in microbiology*, *7*, 428.

Williams, T. A., Embley, T. M., & Heinz, E. (2011). Informational Gene Phylogenies Do Not Support a Fourth Domain of Life for Nucleocytoplasmic Large DNA Viruses. *PLoS ONE*, *6*(6), e21080.

Yoosuf, N., Yutin, N., Colson, P., Shabalina, S. a., Pagnier, I., Robert, C., … & Koonin, E. V. (2012). Related Giant Viruses in Distant Locations and Different Habitats: Acanthamoeba polyphaga moumouvirus Represents a Third Lineage of the Mimiviridae That Is Close to the Megavirus Lineage. *Genome Biology and Evolution*, *4*(12), 1324–1330.

Yutin, N., & Koonin, E. V. (2013). Pandoraviruses are highly derived phycodna-viruses. Biology direct, 8(1), 25.

Yutin, N., Wolf, Y. I., Raoult, D., & Koonin, E. V. (2009). Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virology journal*, *6*(1), 1.

Yutin, N., Colson, P., Raoult, D., & Koonin, E. V. (2013). Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. *Virology journal*, *10*(1), 106.

Yutin, N., Wolf, Y. I., & Koonin, E. V. (2014). Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology*, *466*, 38–52.

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., ... & Ettema, T. J. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, *541*(7637), 353.

Zauberman, N., Mutsafi, Y., Halevy, D. B., Shimoni, E., Klein, E., Xiao, C., ... & Minsky, A. (2008). Distinct DNA exit and packaging portals in the virus Acanthamoeba polyphaga mimivirus. *PLoS biology*, *6*(5), e114.

Paper I

Additional materials

Supplementary materials are available following this link until November 28th 2018:

https://drive.google.com/open?id=1UCIoLmJJjHx7wXSk4b8ensFsLAUMenIR