Distribution of Giant Viruses and Mobile Genetic Elements in Selected Habitats

Dominik Lücking[1], Alejandro Manzano-Marín[1], Anouk Willemsen[1]

[1] University of Vienna, Centre for Microbiology and Environmental Systems Science, Division of Microbial Ecology, 1030 Vienna, Austria

CeMESS | Centre for Microbiology and Environmental Systems Science

ABSTRACT

Giant viruses (GVs) play crucial ecological roles in oceans, freshwater and soil, yet their temporal and spatial distributions, as well as their interactions within complex systems, remain largely unexplored. In this study, we investigate the use of mobile genetic elements (MGEs) identified on GV genomes to trace potential interaction partners and unravel the evolutionary history of these viruses. Combined with viral, bacterial, and host diversity, we aim to provide a comprehensive understanding of the ecological functions of GVs in both natural and biotechnological ecosystems. Samples from Danish wastewater treatment plants were sequenced and scrutinized for the presence of potential GVs and their hosts. Within the genomes of the identified novel GVs, we detect Ankyrin repeats (ANK), elements known for helping viral replication and evasion of host defense mechanisms. This suggests a complex interaction network of viruses and hosts. Future work aims to unravel this network, by integrating available time series data, tracing ANK repeats and other mobile genetic elements known to interfere with host-virus interactions.

Identification of 78 *Nucleocytoviricota* across 5 orders

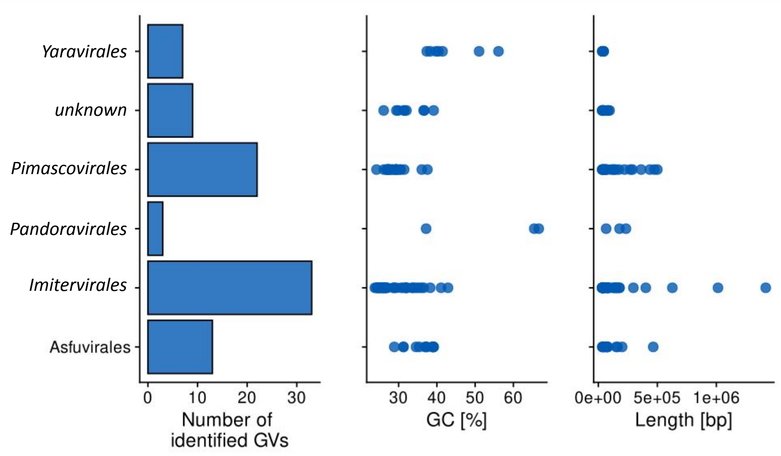

Figure 1: Overview of 78 identified *Nucleocytoviricota*. Displayed are the number of individual viruses, their GC-content (in percentage) and their length (in base pairs). Viruses, for which a taxonomic classification was unsuccessful were labelled as *unknown.*

A

Data

23 metagenomes from wastewater treatment plants in **Denmark** (long and short reads)

Potential hosts of *Nucleocytoviricota* are present in nearly all samples

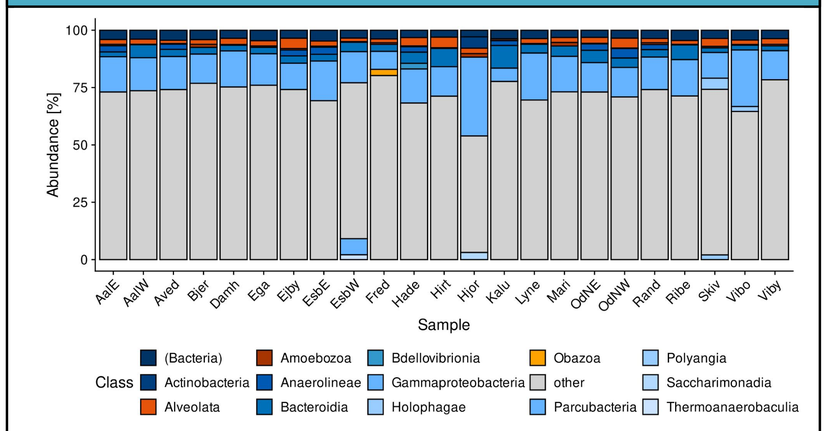

Figure 2: Community composition of the 23 sampling stations. Displayed are the percentage of reads in each metagenome assigned to taxonomic classes. Only classes with >0.5 % abundance are shown. **Potential hosts of *Nucleocytoviricota* are highlighted in shades of orange**, classes with abundances between 0.5 % and 2 % are grouped to *other*.

B

Are the identified giant viruses novel?

Can we identify virus-host interactions and their evolutionary history?

CONCLUSION AND OUTLOOK

- Identification of **78 giant viruses** across 5 orders
- Detection of potential giant virus hosts
- **Ankyrin-repeats** are detected and known for continuously evolving within giant viruses, interfering with antiviral host responses

Outlook:

- Leverage time series data in order to establish virus-host interactions
- Connect the traces of HGT to virus-host pairs in order to entangle their complex ecological and evolutionary interaction network

Phylogenetic analysis suggests we identified previously uncharacterized novel *Nucleocytoviricota*

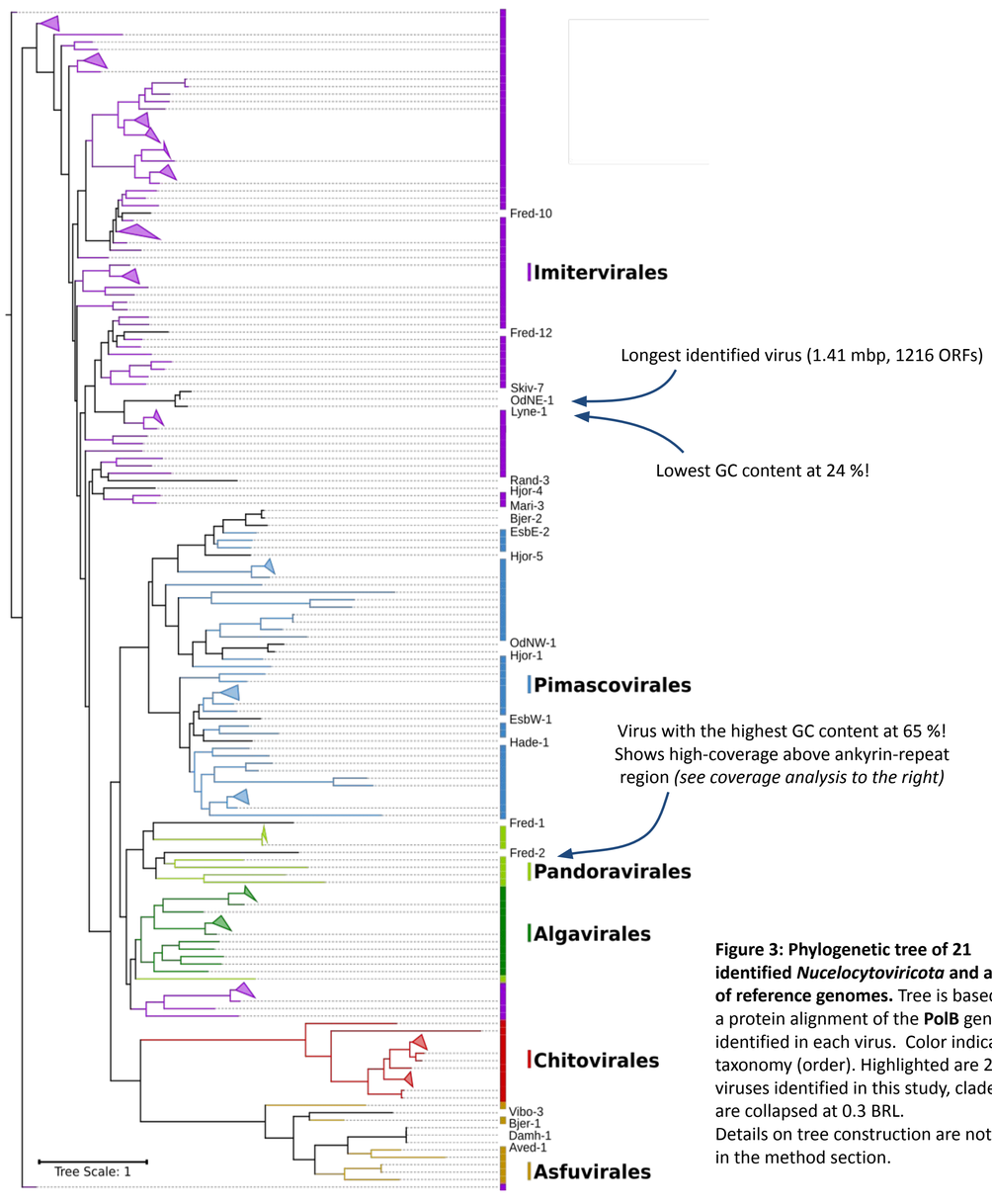

Figure 3: Phylogenetic tree of 21 identified *Nucelocytoviricota* and a set of reference genomes. Tree is based on a protein alignment of the **PolB** gene identified in each virus. Color indicates taxonomy (order). Highlighted are 21 viruses identified in this study, clades are collapsed at 0.3 BRL. Details on tree construction are noted in the method section.

C

Are there traces of horizontal gene transfer?

Can shared genomic elements help us entangle the evolutionary history?

Ankyrin-repeats show increased coverage and are inserted in crucial genomic regions

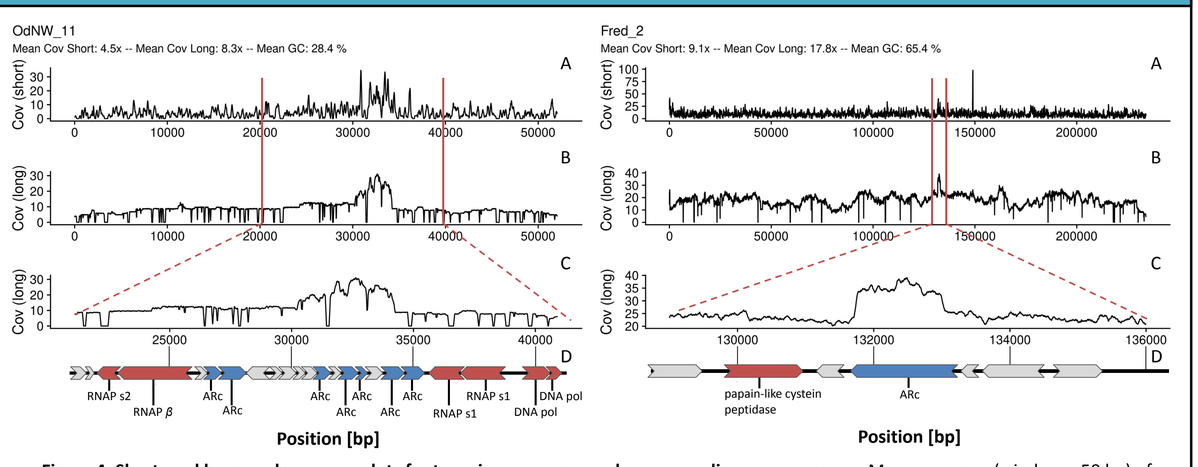

Figure 4: Short- and long read coverage plots for two virus genomes and corresponding genome maps. Mean coverage (window = 50 bp) of short (A: Illumina 2x150bp) and long (B: Oxford Nanopore 1D) reads mapped against virus genomes and a close-up of the high-coverage regions (C). A corresponding genome map of said region is indicated below (D). *RNAP - DNA-dependent RNA-polymerase, ARc - Ankyrin repeat-containing CDS, DNA pol - DNA-polymerase.* Hypothetical genes are shown in grey.

D

Methods

A **Identification and analysis of potential *Nucleocytoviricota*.** Assembled contigs were screened with *ViralRecall, kraken, ncldv_markersearch,* and *diamond blastp.* rRNA genes were predicted using *rnammer-1.2.* Promising contigs were manually searched against *nr* using *blastn* and *blastp* and a final score was given to each contig (https://github.com/dluecking/wwtp_denmark) . This yielded 78 potential nucleocytoviricota which were subsequently analysed. GC content was calculated using *seqkit*.

B **Assessing the community composition of each sampling site.** For each sampling site, Illumina reads were taxonomically classified using *phyloflash*.

C **Phylogenetic tree construction.** First, polB proteins were identified in each contig using *ncldv_markersearch.* This yielded strong signals in 21 out of 78 virus contigs, which were subsequently aligned using *muscle -align* with default parameters. The alignment was cleaned using *trimal* with parameter setting *'automated1'* and finally, a maximum-likelyhood (ML) tree was calculated using *iqtree*,with default bootstrap values (1000) and the LG+F+R7 model. Clades are collapsed by 0.3 BRL.

D **Coverage maps.** Mapping was done for all contigs (viral and non-viral) of one sample in parallel in order to avoid cross-contamination. For Illumina short reads, this was done using *bowtie2* with the following parameters: --very-sensitive-local --no-discordant -I 150 -X 700. Oxford nanopore reads were mapped using *minimap2* with the *-ax map-ont* flag. Coverage per contig was calculated with *samtools depth* and visualized using R. **Genome maps.** Potential viral contigs were annotated using *prokka* with the following parameters: *--evalue '1e-5' --coverage '60'* while using a manually curated set of giant virus proteins in order to improve annotation.

Funding

This project has received funding from the European Union (ERC, CHIMERA, 101039843). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

European Research Council
Established by the European Commission

Acknowledgements

Thanks to AW and AMM for great discussion and support, ED for his map-making skills. Special thanks to CS[1] *et al* for the original sampling and sequencing.

[1]Singleton, Caitlin M., et al. "Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing." Nature communications 12.1 (2021): 2009.