Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by

M. Sc. Josef Vaas

born in: Würzburg, Germany

Oral examination: 26. April 2024

Discovery of Asfarviruses in Next-Generation Sequencing Data

Referees:

Prof. Dr. Dr. h.c. Ralf Bartenschlager

Prof. Dr. Benedikt Brors

Acknowledgements

First, I would like to thank Prof. Dr. Dr. h.c. Ralf Bartenschlager for giving me the opportunity to write this thesis, for his advice and for reviewing it. Second, I am very grateful to Prof. Dr. Benedikt Brors for his thoughtful comments and evaluation of this work. Third, many thanks to Dr. Tim Waterboer for his valuable contribution as a member of my examination committee. Fourth, I would like to express my gratitude to Prof. Dr. Chris Lauber for his thoughtful supervision and always providing advice whenever I needed it, for correcting this thesis and for the willingness to be a part of my examination committee.

I am grateful to Dr. Stefan Seitz for his mentorship, ton of input, his trust, and his support on my journey to become an independent scientist. A special thank you to Prof. Dr. Alexandros Stamatakis for his valuable comments and discussions throughout this thesis.

Finally, I would like to express my appreciation to all members of F170 (DKFZ, Heidelberg), Molecular Virology (University Hospital Heidelberg) and Computational Virology (TWINCORE, Hannover) for sharing their expertise, providing help, and being great company.

Finally, an endless thank you to my family and friends. Without their support, none of this would have been possible.

Contents

Abstract

The *African swine fever virus* (ASFV) causes a highly lethal, haemorrhagic fever in both wild and domestic pigs and is currently panzootic with outbreaks in Africa, Asia and Europe. It poses a tremendous threat to livestock, as infection of a single individual leads to slaughter of the entire pig herd. ASFV is the only official member of the virus family *Asfarviridae* within the order *Asfuvirales*, and the closest relative to date is a mollusc virus, the *Abalone asfa-like virus* (AbALV), which also causes a high mortality disease. The aim of this work was to identify unknown asfarviruses from unprocessed sequencing data, which will help to elucidate the origin of ASFV and assess the risk of the emergence of other severe ASFV-like viruses. Using a highly streamlined profile Hidden Markov Model (HMM)-based approach, 320,000 unprocessed DNA and RNA sequencing datasets from the Sequence Read Archive (SRA) were screened for novel members of the *Asfuvirales*, and those datasets found positive were *de-novo* assembled and searched for viral genomes and hallmark genes. From a sequencing experiment of a marine shell-less mollusc, a circular asfarvirus genome named after the host *Elysia marginata asfarvirus* (EMAV) could have been assembled and comparative genomics was conducted to analyse the protein repertoire, orthologous genes and taxonomy with related viruses. Novel asfarvirus markers were identified in six other molluscan hosts, including a close relative of AbALV from a different abalone species, in two transcriptomic datasets of vertebrates, horse and cow, and in thirteen DNA and RNA datasets of protists. The phylogenetic tree reconstruction revealed 22 new virus sequences in the *Asfuvirales* clade and out of these, 14 sequences clustered in the *Asfarviridae* family. The novel asfarvirus sequences originated from hosts such as protists, molluscs and mammals. The subtree

of the *Asfarviridae* is divided into three clades, showcased by ASFV, AbALV and EMAV, respectively. The closely related protist viruses and the intermixed hosts within that clade could indicate a high likelihood of an ASFV-like virus that has jumped from a protist to an intermediate (mollusc) or mammalian host. Finally, a second, non-canonical major capsid protein (MCP) was detected in five asfarviruses and protein sequence and structure analyses suggested an independent, monophyletic MCP lineage gained in an ancient asfarvirus. My results shed new light on the evolution, possible origins and host reservoirs of ASFV and in the future, it will be crucial to elucidate the pathogenic and zoonotic potential of EMAV and the other novel ASFV-like virus candidates.

Zusammenfassung

Das *Afrikanisches Schweinepest-Virus* (aus dem Englischen *African swine fever virus* - ASFV) verursacht ein tödliches hämorrhagisches Fieber bei Wild- und Hausschweinen und ist derzeit panzootisch in Afrika, Asien und Europa. Da die Infektion eines einzelnen Tieres zur Tötung eines gesamten Schweinebestandes führen kann, stellt es eine enorme Gefahr für die Tierhaltungsbetriebe dar. ASFV ist das einzige offizielle Mitglied der Virusfamilie *Asfarviridae*, innerhalb der Ordnung *Asfuvirales*, und das bisher engste verwandte Virus von ASFV ist ein Molluskenvirus, das *Abalone asfa-like virus* (AbALV), welches ebenfalls eine Krankheit mit einer hohen Sterblichkeit verursacht. Das Ziel dieser Arbeit bestand darin, aus unbearbeiteten Sequenzierdaten unbekannte Asfarviren zu identifizieren, was dazu beitragen kann, den Ursprung von ASFV zu bestimmen und das Risiko des Auftretens anderer gefährlicher ASFV-ähnlicher Viren zu ermitteln. Mithilfe eines hoch optimierten, auf dem *Hidden Markov Model* (HMM) basierenden Hochdurchsatzverfahren wurden 320.000 unbearbeitete DNA- und RNA-Sequenzierdatensätze aus dem *Sequence Read Archive* (SRA) auf neue, unbekannte Mitglieder der *Asfuvirales* durchsucht. Die Treffer wurden *de-novo* assembliert und nach viralen Genomen beziehungsweise Schlüsselgenen gescannt. In einem Sequenzierexperiment eines marinen schalenlosen Weichtieres (*Mollusca*) konnte ein zirkuläres Asfarvirus-Genom detektiert werden, das nach dem Wirt *Elysia marginata asfarvirus* (EMAV) benannt wurde. Es wurde vergleichende Genomik (aus dem Englischen *Comparative Genomics*) durchgeführt, um das Proteinrepertoire, orthologe Gene und Verwandtschaftsbeziehungen zu anderen Viren zu analysieren. Neue Asfarvirus-Schlüsselgene wurden in sechs weiteren Mollusken identifiziert, darunter ein nahes Verwandtes von AbALV aus einer anderen Abalone-Art, in zwei

Transkriptom-Datensätzen von Wirbeltieren, Pferd und Rind, sowie in dreizehn DNA- und RNA-Datensätzen von Protisten. Die Analyse der phylogenetischen Beziehungen in einem Stammbaum ergab 22 neue Virussequenzen in der Klade der *Asfuvirales*, von denen 14 Sequenzen zur Familie der *Asfarviridae* gehörten. Die neuen Asfarvirus-Sequenzen stammten von Wirten wie Protisten, Mollusken und Säugetieren. Der *Asfarviridae*-Teil des Stammbaums ist in drei Kladen unterteilt, die durch ASFV, AbALV beziehungsweise EMAV repräsentiert werden. Die eng verwandten Protistenviren innerhalb dieser Klade und die verschiedenen Wirte könnten auf ein ursprüngliches ASFV-ähnliches Virus hindeuten, das von einem Protisten auf einen Zwischenwirt (Mollusk) oder ein Säugetier übergesprungen ist. In fünf Asfarviren wurde außerdem ein zweites, andersartiges *Major Capsid Protein* (MCP) nachgewiesen, und Analysen der Proteinsequenz und -struktur deuteten auf eine unabhängige, monophyletische Abstammung hin, die aus einem sehr frühen Asfarvirus hervorgegangen sein könnte. Meine Ergebnisse verbessern das Verständnis der Evolution, des möglichen Ursprungs und des Wirtsreservoires von ASFV und in Zukunft wird es von entscheidender Bedeutung sein, das pathogene und zoonotische Risiko von EMAV und den anderen neuen ASFV-ähnlichen Viruskandidaten zu erforschen.

List of figures

List of tables

Abbreviations

aa	amino acids
AAI	average amino acid identity
AbALV	*Abalone asfa-like virus*
ANI	average nucleotide identity
ASFV	*African swine fever virus*
BLAST	Basic Local Alignment Search Tool
DDVD	Data-Driven Virus Discovery
DJR	double jelly roll domain
ElyMa	*Elysia marginata*
EMAV	*Elysia marginata asfarvirus*
FV	*Faustovirus*
HMM	Hidden Markov models
HPC	high-performance computing cluster
ICTV	International Committee on Taxonomy of Viruses
ITRs	inverted terminal repeats
KV	*Kaumoebavirus*
MAGs	metagenome assembled genomes
MCP	major capsid protein
MIUViG	Minimum Information about an Uncultivated Virus Genome
MSAs	multiple sequence alignments

NCBI	National Center for Biotechnology Information
ORF	open reading frame
PSI-BLAST	Position-Specific Iterative BLAST
PSSM	position-specific scoring matrix
PV	*Pacmanvirus*
RdRp	RNA-dependent RNA polymerase
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SRA	Sequence Read Archive
VACV	*Vaccinia virus*
WGS	Whole genome sequencing

1 Introduction

1.1 The *African swine fever virus* - ASFV

The *African swine fever virus* (ASFV) is the best known virus of the *Asfarviridae* family and causes a severe haemorrhagic fever in wild and domestic pigs with a mortality rate of almost 100 % (Li *et al.*, 2022). The virus is responsible for a panzootic with massive outbreaks in Africa, Asia, and Europe as it rapidly spreads in wild boar populations (Ståhl *et al.*, 2023). The transmission occurs through direct contact of susceptible individuals, contaminated pork products, or via arthropod vectors, such as *Ornithodoros* ticks (Gaudreault *et al.*, 2020). Clinical symptoms include high fever, lethargy, diarrhea, respiratory distress and severe pulmonary oedema, and in cases of high pathogenicity the skin exhibits the characteristic haemorrhages (Li *et al.*, 2022). According to the first report in 1921, the geographical origin of ASFV was in Kenya, where it infected warthogs, bushpigs as well as domesticated pigs after they were brought to Africa in the early 20^{th} century (Eustace Montgomery, 1921; Oura *et al.*, 1998). Domesticated pigs are descended from Eurasian and North African wild boars, and ASFV has been shown to have a severe impact on the Eurasian pig lineage (Gifford-Gonzalez and Hanotte, 2011; Michaud *et al.*, 2013). Interestingly, warthogs appear to have developed a tolerance to ASFV, as they do not show clinical symptoms but are able to shed the virus, potentially leading to infection of susceptible animals (Oura *et al.*, 1998; Li *et al.*, 2022). The origin of ASFV is believed to be an *Ornithodoros* tick virus, which then asymptomatically infected wild pigs such as warthogs or bushpigs (Michaud *et al.*, 2013). However,

there is no clear evidence to support this theory. To determine how old ASFV is, a Bayesian approach was performed and the date of the most recent common ASFV ancestor was estimated to be 1700 AD (Michaud *et al.*, 2013). Considering the ancient age of large DNA viruses (Guglielmini *et al.*, 2019), ASFV might be a very young virus. In the same sense, ASFV is rapidly evolving because it has a very high rate of substitution or evolutionary rate compared to other large and small DNA viruses (Michaud *et al.*, 2013).

To date, ASFV spread globally, is panzootic and poses a tremendous threat to livestock, as infection of a single individual leads to slaughter of an entire pig herd (Sauter-Louis *et al.*, 2021; World Organisation for Animal Health WOAH, 2020). For example, in China the largest producer of pork, in 2019 the pig herds were depleted by half and the economical losses were estimated to be more than 140 billion US dollars (World Organisation for Animal Health WOAH, 2020). Although ASFV is a well-known and extensively studied virus, there is currently no treatment or vaccine available and the only preventive measures are quarantine and strict stall keeping (for example: Karger *et al.*, 2019).

1.2 The virus family *Asfarviridae*

Following the discovery of ASFV in the early 20$^{\text{th}}$ century, it was not until 1998 that the International Committee on Taxonomy of Viruses (ICTV) officially proposed the family *Asfarviridae*, with ASFV as the sole member (Eustace Montgomery, 1921; Pringle, 1998). However, as *Asfar* is an acronym for *A*frican *s*wine *f*ever *a*nd *r*elated

virus, it must have been assumed that related viruses exist. Indeed, big progress has been made, novel members have been discovered and according to the ICTV report of 2019, there are four official members of the order *Asfuvirales*: ASFV and three distantly related and recently discovered protist viruses, which are floating genera, not assigned to the family *Asfarviridae* namely *Faustovirus* (FV), *Pacmanvirus* (PV) and *Kaumoebavirus* (KV) (Reteno *et al.*, 2015; Andreani *et al.*, 2017; Bajrai *et al.*, 2016). The term *Asfu* comes from *Asfarviridae* and Faustovirus which was the first protist asfuvirus that has been discovered in 2015 (Reteno *et al.*, 2015). The protist asfuviruses were isolated through co-cultivation with their amoebic hosts. FV and KV with *Vermamoeba vermiformis* and PV with *Acanthamoeba castellanii*. In 2020, ASFV's closest relative to date, the *Abalone asfa-like virus* (AbALV) was discovered and may be the fifth official member of the *Asfuvirales* and the first with a molluscan host, called abalone (Matsuyama *et al.*, 2020). Similar to ASFV, AbALV causes a severe disease that leads to mass mortality of its marine mollusc hosts. The genome sizes of asfuviruses range from 170 kb (ASFV genotypes: 170 - 194 kb) to more than 460 kb (FV) and AbALV has a genome size of approximately 280 kb (Alonso *et al.*, 2018; Reteno *et al.*, 2015; Matsuyama *et al.*, 2023). For more facts and figures, see Table 1 on page 35. In recent years, there have been remarkable discoveries of novel metagenome assembled genomes (MAGs) of asfuviruses (Schulz *et al.*, 2020b; Moniruzzaman *et al.*, 2020). However, the hosts of these MAGs largely remain elusive, as samples with various environmental origins like fresh water served as sequencing material.

Taken together, the few known asfuviruses have a wide variety of eukaryotic uni- and multicellular hosts, in ubiquitous environments, but ASFV is still unmated in

its lineage. Because distant relatives of ASFV have only recently been discovered, primarily facilitated by large-scale metagenomic studies, I hypothesize that additional asfarviruses exist in a possibly wide range of hosts, including hidden host reservoirs of ASFV.

1.3 Towards the origin of ASFV and large DNA viruses

From an evolutionary point of view, ASFV and the other asfuviruses are of special interest. They belong to an old phylum of double-stranded DNA (dsDNA) viruses with giant genomes, the largest known to date, called *Nucleocytoviricota* or formerly nucleo-cytoplasmic large DNA viruses – NCLDV (Iyer *et al.*, 2001; Iyer *et al.*, 2006). The evolution of *Nucleocytoviricota* dates back to a last common ancestor and the emergence of eukaryotic large dsDNA viruses even predates the origin of modern eukaryotes (Guglielmini *et al.*, 2019). A common feature is their enormous genome size that can be as large as 2,500 kb and code for up to several hundred proteins (Boratto *et al.*, 2020; Philippe *et al.*, 2013). The prefix *Nucleocyto* derives from the fact that these viruses replicate in either the host cell's nucleus or in the cytoplasm, forming viral factories (Schulz *et al.*, 2022).

All *Nucleocytoviricota* share a major capsid protein (MCP) with its main structural element the double jelly roll domain (DJR) what is a unification feature of the whole *Bamfordvirae* kingdom within the *Varidnaviria* realm (Krupovic and Bamford, 2008; Xian and Xiao, 2020; Koonin *et al.*, 2020). A major component of the virions are the MCPs and many viruses with DJR MCP form icosahedral capsids, such as the ASFV (Krupovic *et al.*, 2022; Koonin *et al.*, 2020; Wang *et al.*, 2019; Carrascosa

et al., 1984). However, it is known that certain *Nucleocytoviricota* have multiple MCP copies and some even have major virion proteins completely independent of the DJR MCP like the pandoraviruses (Mönttinen *et al.*, 2021; Krupovic *et al.*, 2020). Identifying the MCP of a new virus is essential for characterising its virion structure and understanding its evolutionary history.

Besides that, the *Nucleocytoviricota* form a very diverse phylogenetic group, potentially divided into over 30 virus families, with a host spectrum that includes protists, invertebrates, vertebrates, mammals and humans (Aylward *et al.*, 2021; Sun *et al.*, 2020; Matsuyama *et al.*, 2023). Many of these viruses cause serious maladies in their hosts, for example, members of the *Poxviridae* cause disease in a variety of mammals including the severe smallpox in humans, which was globally eradicated in the 1980s after a successful vaccination campaign (Brennan *et al.*, 2022). In addition, the ASFV is a deadly scourge for pigs worldwide and in the context of future pandemics, knowing its evolutionary history and potential related viruses is crucial for minimizing spill-over risk and early prevention.

It is hypothesised that special mobile genetic elements, known as polintons, which are special transposons that integrate into the host genome and also encode for virus capsid proteins form the evolutionary basis of all *Nucleocytoviricota* (Koonin *et al.*, 2015a; Koonin *et al.*, 2015b). That they encode for major and minor capsid proteins suggest that they can form functional virions making them virus-like transposable elements, blurring the boundary between transposons and viruses (Yutin *et al.*, 2015; Koonin and Krupovic, 2017; Yutin *et al.*, 2013). Recently, Starrett *et al.*, 2021 suggested the classification system for animal-tropic viruses that derive from polinton-

like elements and named them *Adintovirus* after the similarity to adenoviruses and the ability to express an integrase. Their genome sizes range from 14 to 40 kb, and due to their ability to integrate into the host genomes, endogenous viral sequences are thought to play a crucial role in the evolution of unicellular and multicellular eukaryotic genomes (Bellas *et al.*, 2023; Moniruzzaman and Aylward, 2023). These viruses appear to be associated with animals from different groups, but they are also involved in the evolution of giant viruses.

Evolutionary trajectories and taxonomy require genetic markers and unlike prokaryotes, whose taxonomy is often based on the similarity of 16S ribosomal RNA (rRNA), viruses lack a common genetic marker (Weisburg *et al.*, 1991). However, in *Nucleocytoviricota* it has been shown that there are various genes that are well conserved and common to most viral members, including the MCP, DNA and RNA polymerases, the helicase SF2, the packaging ATPase A32 or the late transcription factor VLTF3 (for example: Yutin *et al.*, 2009). Subsequently, these core genes or hallmark genes can be utilized for the discovery of novel asfarviruses, their taxonomic classification and the reconstruction of reliable phylogenies.

The revolution of genome sequencing technologies leads to massively and constantly increasing amounts of unprocessed sequencing data publicly available in the Sequence Read Archive (SRA) that is hosted by the National Center for Biotechnology Information (NCBI) (Leinonen *et al.*, 2011). For each dataset, metadata such as the organism, tissue or location of the sequenced sample is stored and can be analysed and exploited. In many of the millions of sequencing datasets genetic by-catch from viruses that infected the sequenced sample have been captured. Therefore, these ge-

nomic data bare an unprecedented value for the discovery of new viruses and formed the basis of a new field of computational virology research called Data-Driven Virus Discovery (DDVD) (Roux *et al.*, 2019; Lauber and Seitz, 2022; Gregory *et al.*, 2019; Edgar *et al.*, 2022). DDVD is uncoupled from the collection and processing of biological samples and utilizes highly distributed computing, such as high-performance or cloud computing, to screen large amounts of raw sequencing data of any origin for the presence of viral by-catch. Moreover, the easy-to-use bioinformatics toolkit for identifying, annotating and analysing viral sequence data continues to grow, making DDVD more sophisticated and applicable to many more scientists (Lu and Peng, 2021). The DDVD approach already resulted in the discovery of numerous novel and often highly divergent viruses (for example: Lauber *et al.*, 2017; Tisza *et al.*, 2019; Lauber *et al.*, 2019; Moniruzzaman *et al.*, 2020; Schulz *et al.*, 2020b; Lauber *et al.*, 2021). Here, the screening method was adjusted to screen for asfarviruses in datasets from different animal groups.

In summary, asfarviruses have an unclear, wide-ranging history and evolution, and known virus hallmark genes were used for powerful, large-scale computational screens of the massive sequencing data repository SRA to detect novel members of the asfarviruses.

1.4 Aim of this work

The objective of this study was to analyse unprocessed sequencing data with a deep mining, high-throughput virus discovery approach which was already successfully tested on RNA viruses (Lauber *et al.*, 2017; Lauber *et al.*, 2019; Lauber *et al.*, 2021), in order to identify previously unknown asfuviruses, which includes the highly infectious, lethal and standalone pig virus ASFV. Sequencing data were systematically screened and the majority of available genomic and transcriptomic sequencing runs for vertebrates, molluscs, mites and ticks, and protists were searched using viral hallmark genes as search queries. Subsequently, a refined and effective large virus assembly method was crucial and involved downloading raw sequencing data, trimming low quality bases, mapping against a host genome, and assembly and super-assembly with multiple assembly tools. This assembly workflow was benchmarked by screening for unknown relatives of the pandemic Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) from the order *Nidovirales*, which contains viruses with the largest known RNA virus genomes (Lauber *et al.*, 2021). In contrast to previous environmental data screens, newly discovered viruses are also characterised with the metadata stored in the SRA, such as host, tissue, disease, breed/strain, sex, date and origin of sample collection, and if the virus sequence was found in multiple datasets. The overall goal is to discover novel ASFV-like viruses in potential new hosts and to perform comparative genomics with reference viruses, to reconstruct phylogenetic relationships. This will help to better understand these viruses, evaluate their pandemic risk, and ultimately trace the origin of ASFV.

2 Material and methods

This chapter lists all the software tools, packages, databases and computer systems used to ensure the reproducibility of the results of this work.

2.1 Bioinformatics tools, software and high-performance computing system

All the following tools and tasks were either run on a local LINUX device with an Ubuntu 18.04.3 operation system or if more computing resources were required, on the high-performance computing cluster (HPC) of the DKFZ in Heidelberg, Germany. According to the developers' webpage, "[...] the cluster runs the IBM Spectrum LSF workload management software and is based on CentOS. As of November 2020, it consists of more than 150 servers and provides access to more than 4500 CPU cores [...]." (ODCF DKFZ, Heidelberg, 2020). This HPC was used to scale up the processes that were initially implemented and tested on a local computer system.

Bioinformatics workflows were scripted in Python, data processing and plotting was done in R, and many small tasks were solved with bash scripts.

2.2 Virus discovery workflow

This chapter explains how novel asfarviruses were screened and discovered.

2.2.1 Search for novel asfarviruses in unprocessed sequencing data

The basic principles of this virus discovery method have already been established and successfully tested for other virus groups by Prof. Dr. Chris Lauber and Dr. Stefan Seitz. There is only a brief description at this point and for more details please refer to Lauber *et al.*, 2017; Lauber *et al.*, 2019 or Lauber *et al.*, 2021. The screening workflow is called *Virushunter* and its basic principle is visualised in Figure 1.

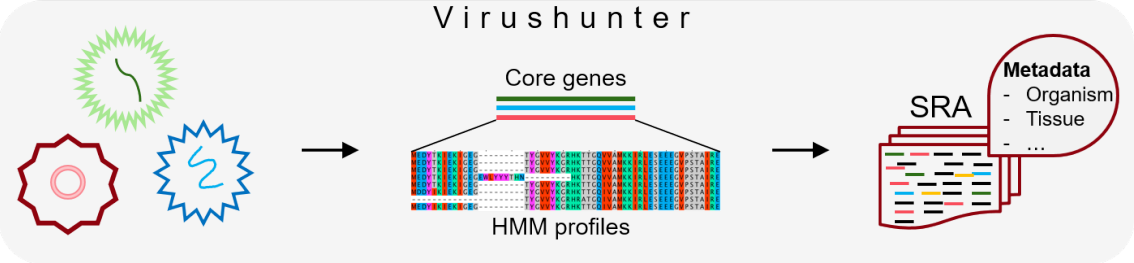

Figure 1: The screening process for novel asfarviruses in raw sequencing data. The virus icons on the left indicate a virus group of interest with shared hallmark genes. These genes can then be used to build profile HMMs, which are then utilized to screen the raw sequencing data repository the Sequence Read Archive (SRA), which also contains metadata.

The screening process is based on the homology of profile Hidden Markov models (HMM) with raw sequencing data from many different eukaryotes. Profile HMMs are based on multiple sequence alignments (MSAs) and incorporate a position-specific scoring system that ranks the probability of observing each of the 20 amino acids at that position and adds gap penalties. These features can be used to sensitively

search sequence databases for remote homology (Eddy, 1998; Steinegger *et al.*, 2019). However, this text does not focus on the actual screening process, but on the selection of the best protein queries, which datasets were screened and the analysis of the results.

Ideal queries for the screens are viral hallmark or marker genes, which are genes shared by all members of a virus group. For RNA viruses, screening is often based on the RNA-dependent RNA polymerase (RdRp) as this is an ubiquitous gene for many RNA viruses. For large DNA viruses it is not trivial to select reasonable query proteins due to the high diversity of these viruses. To evaluate the initial queries and profile HMMs, a test screen was performed using a toy dataset of 461 SRA projects. A part of the dataset contained large DNA virus sequences and the other did not. Ten hallmark genes of *Nucleocytoviricota* (Iyer *et al.*, 2001; Yutin *et al.*, 2009; Koonin and Yutin, 2018) were tested as queries including NCLDV major capsid protein (MCP), A2 late transcription factor (A2Ltf), DEAD/SNF2-like helicase (SF2), DNA-directed RNA polymerase small and large subunit, DNA polymerase family B C terminus and N terminus, D5 DNA primase/helicase, small subunit ribonuclease reductase and A32 packaging ATPase. The screen was highly specific for most of this markers, but the highest sensitivity was achieved with the MCP. Therefore, the reference asfarvirus MCPs were used to build accurate profile HMMs, which were then used to screen the raw sequencing data.

For screening, sequencing data from relevant animal groups were selected, further reading in the results section 3.2.1 on page 24. The NCBI's SRA can help to locate all sequencing projects related to a particular animal group using the NCBI Taxonomy

Browser (Schoch *et al.*, 2020), and the metadata contains important information for characterising potential virus hits. This includes information such as the sample material, the tissue, where the sample was collected, or whether the host organism was in a noticeable state of health.
After the screen, the list of asfarvirus positive sequencing projects was quality controlled and only hits with an E-value below 1×10^{-5} and percent identity below 90 % against the closest RefSeq virus were used for subsequent *de-novo* assembly.

Originally, the *Virushunter* screen was followed by a seed-driven assembly approach called *Virusgatherer*. This means that only those reads that matched the initial search profiles were considered for iterative assembly. This approach is highly effective for assembling novel RNA virus genomes with a monopartite genome structure (refer to Lauber *et al.*, 2017 and Lauber *et al.*, 2019). However, the larger genomes of the asfarviruses made seed-driven assembly inapplicable and a *non-targeted* assembly approach was established as part of this work.

2.2.2 Non-targeted *de-novo* assembly

The non-targeted *de-novo* virus assembly approach (Figure 2) includes downloading the whole sequencing dataset, trimming sequencing adapters and bases of low quality using Trimmomatic v. 0.39 (Bolger *et al.*, 2014). The quality-controlled reads were then aligned to the host's genome, if it was available in NCBI Taxonomy (Schoch *et al.*, 2020), using Bowtie 2 v. 2.3.4.1 and SAMtools v. 1.7 (Langmead and Salzberg, 2012; Danecek *et al.*, 2021). All non-host reads were retained for *de-novo* assembly. The

assembling tools MEGAHIT (v. 1.2.9; Li *et al.*, 2015), SPAdes (v. 3.11.1; Bankevich *et al.*, 2012) and CAP3 (version data: 12/21/07; Huang, 1999) were used and finally a super-assembly step combined the intermediate contigs of each assembling tool. All FASTA sequences were handled and edited with SeqKit (Shen *et al.*, 2016).

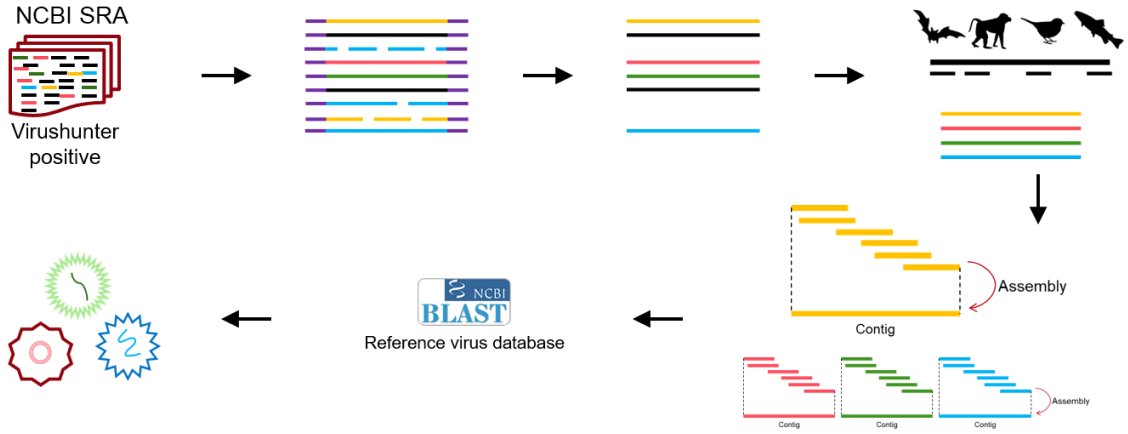

Figure 2: Schematic overview of the non-targeted *de-novo* virus assembly. *Virushunter* positive raw sequencing datasets were downloaded, quality controlled, trimmed, mapped to a host genome and then assembled by multiple assembling tools. Virus sequences were then detected using BLAST or HMMsearch.

This non-targeted *de-novo* assembly workflow was initially evaluated with the assembly of novel members of the *Nidovirales* order, to solve any potential issues. Nidoviruses are the RNA viruses with the largest known genomes, making them a well-suited group for testing. Whether the concept of non-targeted assembly has any advantages over the original seed-driven assembly approach will be the subject of section 3.1 on page 21. The *Virushunter* screening process was performed as described above, but the viral RdRp was used as the search query.

The semi-automated non-targeted *de-novo* assembly workflow was written in the Python programming language and initially it was deployed on a local computing system with restricted computational resources. To enable parallelisation and

greater automation, the process was implemented on the DKFZ HPC cluster. After *Virushunter* screening, a list of SRA identifiers is the starting point for downstream analysis (Figure 3, top right). The NCBI E-utilities toolbar was used to fetch the metadata on each SRA dataset. Based on this, the host genome was downloaded for read mapping, as described above. A wrapper script was used to pass all the information for the individual assembly jobs and run them in parallel (Figure 3, center). A second script was then used to assess the progress of the assembly and, if errors occurred, the jobs were re-run from the point of failure (Figure 3, right). The last step was to remove redundant data in order to free up disk space.

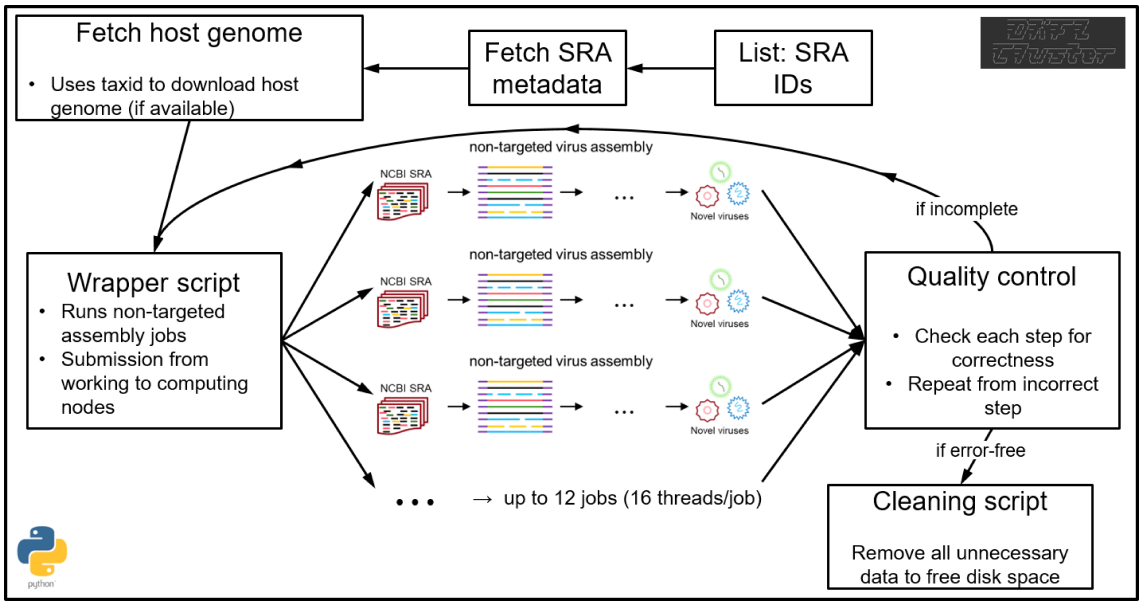

Figure 3: Once the basic virus assembly workflow was tested and implemented, there were several steps of improvement and automation. Scripts were added to retrieve metadata and download the host genome sequence. A wrapper script has been implemented to enable multiple assembly jobs to run simultaneously on the computing cluster. Finally, scripts were programmed to check for potential errors and clean up unnecessary data.

2.2.3 Downstream analysis of assembled contigs

After screening thousands of sequencing experiments for asfarviruses and performing non-targeted *de-novo* assembly the resulting contigs were analysed. The assembled sequences were categorised as viral depending on matches found in the non-redundant nucleotide (nr/nt) or non-redundant protein (nr) sequence databases of the NCBI using Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990), either with nucleotide or protein sequences after open reading frame (ORF) prediction (minimum length 60 aa; ORFfinder, v. 0.4.3, https://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/ORFfinder/). Moreover, *ncldv_markersearch* was utilized to screen for the presence of asfarvirus marker genes, as published by Moniruzzaman *et al.*, 2020 and the following markers were included: RNR - ribonuclease reductase, RNAPS + RNAPL - DNA-directed RNA polymerase small + large subunit, D5 - DNA primase/helicase, A32 - packaging ATPase, VLTF3 - Poxvirus late transcription factor VLTF3, MCP - NCLDV major capsid protein, PolB - DNA polymerase family B, mRNAc - mRNA guanylyltransferase, SF2 - DEAD/SNF2-like helicase. BLAST searches against the protein RefSeq database were used to verify potential viral hits and approval was only granted if the closest relative was an asfarvirus or related sequence. For visual genome comparison and to assess contigs' overlaps Easyfig 2.2.2 was used (Sullivan *et al.*, 2011). Finally, Bowtie2 was used to map reads to the virus contigs and the read coverage was plotted using R and ggplot2 (Langmead and Salzberg, 2012; Hadley, 2016). The assembly quality was assessed by visual inspection with the Integrative Genomics Viewer - IGV (Robinson *et al.*, 2011).

In order to obtain larger fragments of the viral genomes, following the detection of asfarvirus markers, the contigs were re-inspected for other virus genes. Only the most promising datasets with the largest asfarvirus contigs were selected. The re-inspection was conducted with tBLASTx (Camacho *et al.*, 2009) using the asfuvirus genomes as query (including the novel asfarvirus genome of *Elysia marginata asfarvirus* (EMAV), see section 3.2.3 on page 30) and run it against the entirety of the assembled contigs as subject. The E-value cut-off was 1×10^{-4}. The resulting viral sequences were verified with BLAST and the non-redundant protein database of the NCBI. To determine the alignment coverage, the contigs were mapped against the closest virus genome using tBLASTx.

2.2.4 Matrix-based screening of protein database using PSI-BLAST

After screening the unprocessed molluscan sequencing data for asfarviruses, sequences related to adintoviruses were found in large quantities. This led to the decision to search for footprints of these viruses in the NCBI non-redundant protein database using Position-Specific Iterative BLAST (PSI-BLAST). PSI-BLAST is based on a position-specific scoring matrix (PSSM) and can be used to detect distant homologous proteins by updating the matrix with new hits from the previous iteration (Altschul *et al.*, 1997). The principle is illustrated in Figure 4.

For the PSI-BLAST search the following marker genes of *Branchiostoma lancelet adintovirus* were used as queries: adenain (maturation protease), protein-primed polymerase B, retrovirus-like integrase, hexon, penton, ftsk (DNA packaging) and

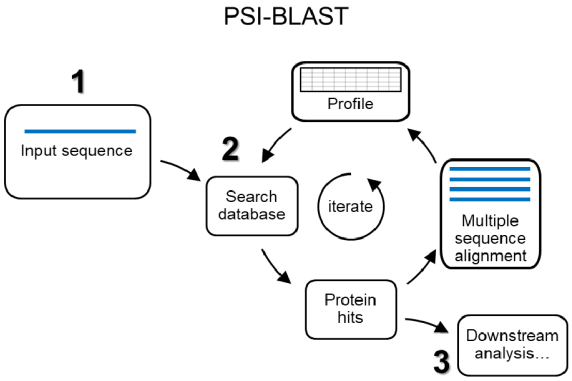

Figure 4: After inputting a query protein sequence (1) a protein database is searched (2) and the resulting hits are either extracted (3) or used to build a position-specific scoring matrix (PSSM) for re-screening. In every subsequent iteration the new hits are added to the previous matrix and the following search will yield a higher diversity of proteins. Therefore, PSI-BLAST can detect distantly related proteins.

gasdermin. If there were no significant differences between the results of the current and previous iterations, the searches were stopped. Taxonomy assignment of the obtained sequences has been performed by MMseq2 (Steinegger and Söding, 2017) and for visualisation Pavian has been used (Breitwieser and Salzberg, 2020).

The adintovirus sequences were found in database entries of various organisms and based on the sizes of the nucleotide sequences, it is likely that the adintovirus genes are on host genome contigs. Given the expression of an integrase, this can be expected. First, only the contigs of the adenain query are plotted (Section 3.3, Figure 22, part B and Figure 23) but then the underlying nucleotide sequences of all the PSI-BLAST queries were considered for subsequent experiments (refsec:Polinton-like viruses widely distributed in various animal groups, Figure 24). Only nucleotide contigs with a minimum of four different adintovirus markers were selected and the number of integration sites of adintovirus gene clusters was identified. For that purpose, the nucleotide sequences behind the protein hits were extracted using the

NCBI E-utilities toolbar. This approach failed for some of the proteins, hence the slightly different number of hits. Nucleotide BLAST searches were then performed to determine the exact location of the adintovirus genes on the host contig. If two adintovirus markers were less than 30 kb apart, it was assumed that they originated from the same integration cluster. Otherwise, they were considered to be from different integration sites at different positions in the genome.

2.3 Methods of comparative genomics

Comparative genomics studies the commonalities and differences between different organisms based on their genome sequences. Here, the genomes of known and novel asfuviruses will be investigated.

2.3.1 Characterization of virus genomes

Reference virus genome sequences were obtained from NCBI virus database and the ORFs were detected using the NCBI ORFfinder with a minimum length of 60 aa (v. 0.4.3, https://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/ORFfinder/). GC-content was calculated using SeqKit (Shen *et al.*, 2016). To detect orthologous genes Proteinortho was utilized (Lechner *et al.*, 2011) and tRNAs were searched using ARAGORN with default parameters (Laslett and Canback, 2004). To calculate the average amino acid identity (AAI) the tool CompareM was utilized and the E-value, percent sequence identity and percent alignment length thresholds were set to 0.001,

25 % and 60 %, respectively (https://github.com/dparks1134/CompareM). For visual genome comparison, synteny plots were plotted with the R package gggenomes (Hackl *et al.*, 2023). The functional annotation of the putative proteins was done with EggNOG, which comes with an annotated orthology resource of thousands of viruses (Cantalapiedra *et al.*, 2021). The proteins are then categorized into 25 clusters of orthologous genes (COG) groups within four protein function supergroups (Tatusov *et al.*, 1997). The icons of the animal host habitus included in many figures were obtained from PhyloPic, which is a database for silhouette images of organisms (https://www.phylopic.org/).

2.3.2 Phylogenetic tree reconstructions

First, the novel asfarvirus sequences that were discovered as part of this work, were combined with a broad selection of *Nucleocytoviricota* reference viruses for phylogenetic analysis. A Bayesian tree was constructed using a combination of six different marker genes, namely MCP - NCLDV major capsid protein, RNAPS - DNA-directed RNA polymerase small subunit, PolB - DNA polymerase family B, VLTF3 - Poxvirus Late Transcription Factor VLTF3, D5 - DNA primase/helicase and A32 - Packaging ATPase. Reference NCBI RefSeq virus genomes were obtained in 03/2023 and the best matching marker gene against the respective ncldv_markersearch profile (Moniruzzaman *et al.*, 2020) was selected and further utilized. Protein MSAs were aligned using MAFFT v7.310 (Katoh *et al.*, 2002) and trimAl 1.2rev59 (Capella-Gutiérrez *et al.*, 2009). The Bayesian phylogenetic tree reconstruction was built with BEAST v2.7.1 (Bouckaert *et al.*, 2019) using OBAMA Bayesian Amino Acid Model

Averaging, optimized relaxed clock and besides that default settings. Visualization of the tree was performed with ITOL (Letunic and Bork, 2021).
The MCP-based tree was built from MCPs from asfuvirus MAGs discovered from environmental data (Moniruzzaman *et al.*, 2020; Schulz *et al.*, 2020b), and all the novel sequences discovered as part of this work. As there were multiple MCP proteins detected on five asfarviruses genomes, a third tree was constructed from primary/group I and secondary/group II MCPs, as described above. The further analyses performed with these MCPs are described in the next section.

2.3.3 Analysis of asfarvirus major capsid proteins (MCPs)

For all the MCP sequences that were found to be expressed on the asfuvirus genomes protein structure models were predicted. The ColabFold (v1.5.2) was used, which is a method that combines AlphaFold2 with MMseqs2 and makes it available on *Google Colaboratory* (Mirdita *et al.*, 2022). For verification also other structural prediction tools like SWISS-MODEL and Phyre2 were used, respectively (Waterhouse *et al.*, 2018; Kelley *et al.*, 2015). For visualization of the protein structure models UCSF ChimeraX was used and the coloring is based on the secondary structure properties (Meng *et al.*, 2023). Subsequently, to uncover their structural similarities, protein structure models were aligned with DALI in the "All against all structure comparison" mode (Holm *et al.*, 2023). After that, ITOL was used to plot the structural similarity tree (Letunic and Bork, 2021).

3 Results and discussion

The aim of this work was to detect novel asfarviruses in publicly available and unprocessed next-generation sequencing data, in order to expand the current understanding of asfarvirus diversity and their potential pandemic risk. Prior to this, the assembly workflow was established and tested on a screen for unknown relatives of SARS-CoV-2, which belong to the diverse order of *Nidovirales*. In the search for asfarviruses, there were false-positive projects containing adintovirus marker genes, and the large number of hits prompted me to investigate them more closely. Interestingly, adintoviruses share an evolutionary history with *Nucleocytoviricota* (Koonin *et al.*, 2015a; Koonin *et al.*, 2015b; Starrett *et al.*, 2021).

3.1 Proof-of-concept - Non-targeted *de-novo* assembly of nidoviruses

To benchmark the non-targeted *de-novo* assembly workflow and to test how the results compare to the seed-driven assembly approach, a proof-of-concept study was conducted using a screen on novel members of the *Nidovirales*. Within the nidoviruses there are serious human pathogens with a high potential of spill over from animal to humans like SARS-CoV-1 or SARS-CoV-2. Based on current knowledge, nidoviruses have the largest genomes in the RNA virus world, but they are still very small compared to the genomes of large DNA viruses. However, the assembly of new nidoviruses provided an ideal opportunity to prove the concept of the newly

established non-targeted *de-novo* assembly approach. Only the comparison between non-targeted and seed-driven assembly will be discussed here for any details on the individual findings and their relevance refer to Lauber *et al.*, 2021. Just in brief, likewise studies that utilize highly-customized genome assembly, can contribute to future pandemics preparedness, as members of the *Nidovirales* have a notable risk of crossing species barriers and emerging as severe human pathogens.

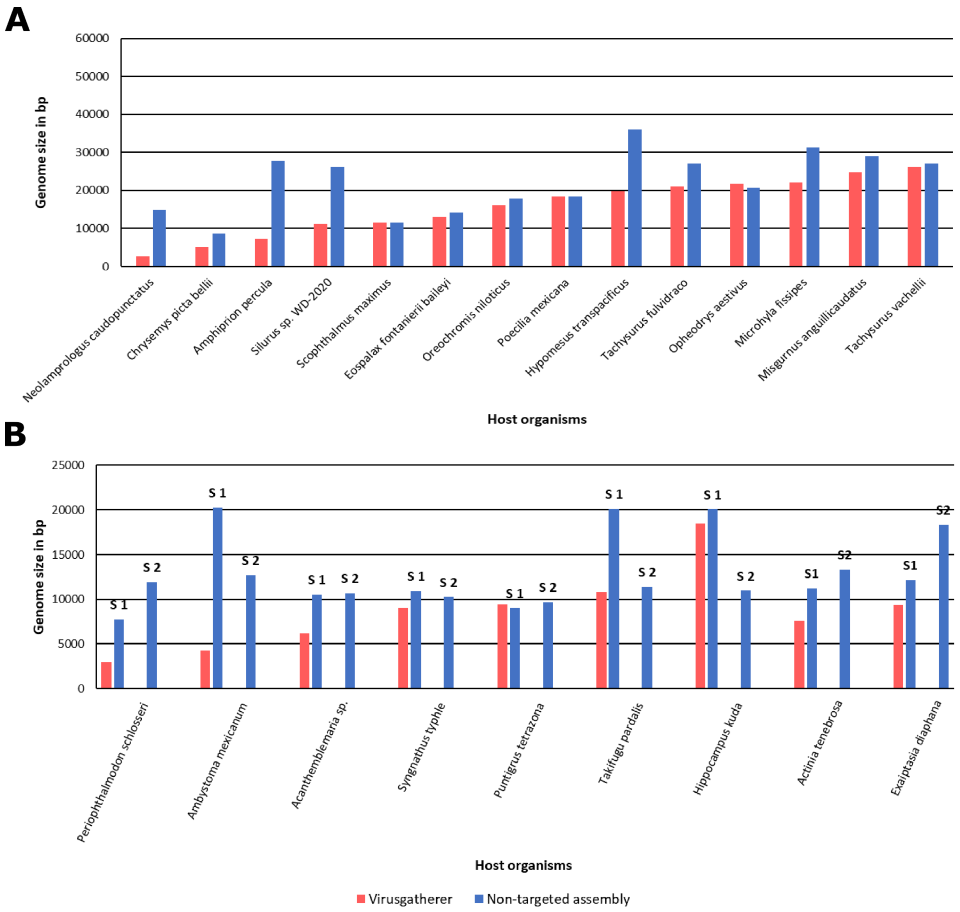

Figure 5: The discovery of new members of the *Nidovirales* was used to compare non-targeted and seed-driven assembly approaches. The genome sizes after the original seed-driven approach are shown in red, and after the non-targeted assembly in blue. Part **A** and **B** show the newly discovered monopartite and bipartite nidovirus genomes, respectively. The use of a non-targeted assembly approach in most cases resulted in longer virus contigs as the blue bars outnumber the red bars, especially for segmented, bipartite genomes, providing a significant advantage over the seed-driven approach.

The genome lengths of the newly discovered monopartite viruses were longer when using the non-targeted assembly approach compared to the original seed-driven assembly approach (Figure 5, part A). A similar result was obtained for the newly discovered bipartite nidovirus genomes (Figure 5, part B). In this case, the entire second segment of the virus genome was missing. In more detail, the RdRp was used as the seed for the assembly, hence the genome segment that lacks the RdRp was missing for all viruses and can only be uncovered by non-targeted assembly.

In the context of large DNA virus genomes a comparable observation was made during the assembly: the longer the genome, the greater the likelihood that it will be fragmented, for example due to areas of low sequencing coverage. Therefore, the non-targeted approach is the preferred method for large and segmented RNA viruses, as well as for large DNA viruses, specifically the asfarviruses. However, when screening thousands of sequencing experiments, it is not possible to perform non-targeted assembly on all of them. In conclusion, this results show that a combination of both methods may be the best choice for assembling large or segmented RNA or DNA viruses.

3.2 Discovery of novel asfarviruses

To learn more about the origin, whether there are any other relatives of ASFV or a hidden host reservoir a large-scale screen for unknown asfarviruses in the SRA, the NCBI's repository of unprocessed sequencing data, was performed.

3.2.1 Screen for novel asfarviruses

In the systematic and large-scale screen, a total of approximately 320,000 unprocessed DNA and RNA sequencing experiments were searched for unknown asfarviruses. The DNA sequencing datasets were analyzed for the detection of new ful-length virus genomes or genome fragments, while in the RNA datasets novel asfarvirus marker genes were searched. To characterise potential novel viral hits, this approach involves the projects' metadata, such as the organism, tissue or location of the sequenced sample.

SRA datasets of the following host groups were analyzed (Figure 6). The first group of organisms screened, were close and distant relatives of the swine including RNA data of all vertebrates and excluding the most redundant model and laboratory organisms such as human, rhesus monkey, Norway rat, mouse and zebrafish. There were also screens in DNA sequencing experiments of the superorder *Laurasiatheria*, which are placental mammals including ungulates, carnivores and whales. As ASFV is not only transmitted by direct contact between individuals but also by tick bites (for example: Chastagner *et al.*, 2020), the next group was *Acari*, consisting of mites and ticks. Since AbALV is a mollusc virus, DNA and RNA sequencing data from

molluscs and gastropods were screened, too. Finally, as there are also asfarviruses with protist hosts, the DNA and RNA sequencing data from protists which are a large collection of mainly unicellular eukaryotic organisms were searched. For this protist screen, all eukaryotes except *Metazoa*, *Streptophyta*, *Fungi* and over-represented *Plasmodium falciparum* and *Plasmodium vivax* were included.

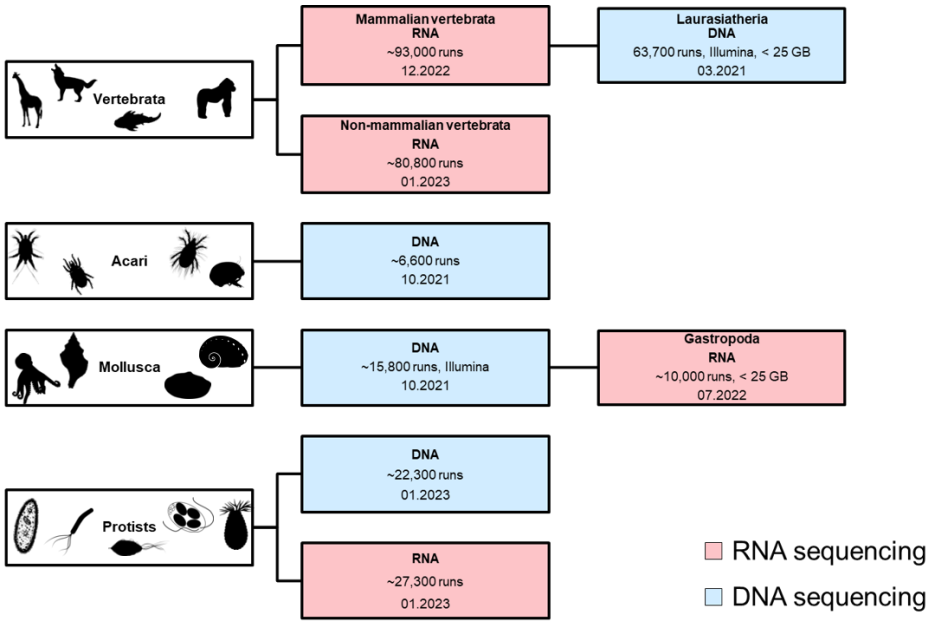

Figure 6: Approximately 320,000 unprocessed sequencing datasets were screened to find novel asfarvirus sequences. The rows indicate the animal groups of the screened data. Red boxes show RNA and blue boxes show DNA sequencing experiments. The boxes contain the date of screening, amount of datasets and information on how the projects were prefiltered.

3.2.2 Novel asfarvirus sequences in molluscs, vertebrates and protists

After screening hundreds of thousands of SRA experiments from a large variety of organisms, I found novel asfarvirus markers in 22 sequencing runs from molluscs, vertebrates and protist (Figure 7). The molluscs mainly inhabit marine but also freshwater environments as in the case of *Galba truncatula* and *Bithynia siamensis*

goniomphalos. In *Verconia norba* and *Elysia marginata* (ElyMa), a complete set of ten markers was found, strongly suggesting novel asfarviruses in these shell-less molluscs. The *ElyMa* dataset is from whole genome sequencing (WGS) and a full viral genome was recovered after *de-novo* assembly, for more details see section 3.2.3 on page 30. In a RNA seqeuencing dataset of *Haliotis diversicolor*, a species of abalone, a novel and almost complete set of markers was found that lacked the D5 DNA primase/helicase. The novel markers of the *Haliotis diversicolor* asfarvirus shared a sequencing similarity of 60 % with the markers from AbALV. Interestingly, *Haliotis diversicolor* was tested as not susceptible to the reference AbALV and no genome replication occurred in experimental infection (Matsuyama *et al.*, 2021). However, these results suggest a close relative of AbALV as a new pathogen of the *Haliotis diversicolor* species. The WGS sequencing dataset of the pacific oyster (*Crassostrea gigas*) contained two novel asfar markers but it was not possible to assemble any meaningful contigs or a draft virus genome. In other molluscan projects, only few markers have been found.

In thirteen protist datasets, asfarvirus marker genes were found and from many of the DNA sequencing datasets a nearly complete set of markers was recovered. Among the protist hits, there were many members of the *Rhizaria* for example *Vulvinaria inflata*, *Uvigerina striata*, *Cassidulina limbata*, *Bolivinella seminuda* and *Bolivina* species. In sequencing data from *Sargassum hemiphyllum*, a brown alga and in *Delisea pulchra* a red alga, asfarvirus markers were detected. *Loxodes* is a ciliate, that was found asfarvirus positive with a full set of markers. The exact species of other protists were not annotated.

The asfarvirus screen in vertebrate sequencing projects detected the horse (*Equus caballus*) and the cow (*Bos taurus*) with eight and one markers, respectively. The

sample material of the horse (NCBI BioSample: SAMN22133205) was bone marrow mononuclear cells that were cultured in autologous normal equine synovial fluid and the sample of the cow (NCBI BioSample: SAMN04240345) was female bovine blastocysts from in-vitro serum of Holstein cow. It is uncertain whether these asfarvirus sequences originate from ASFV-like viruses that infect these mammalian hosts. In this case, further investigation is necessary before drawing any conclusions.

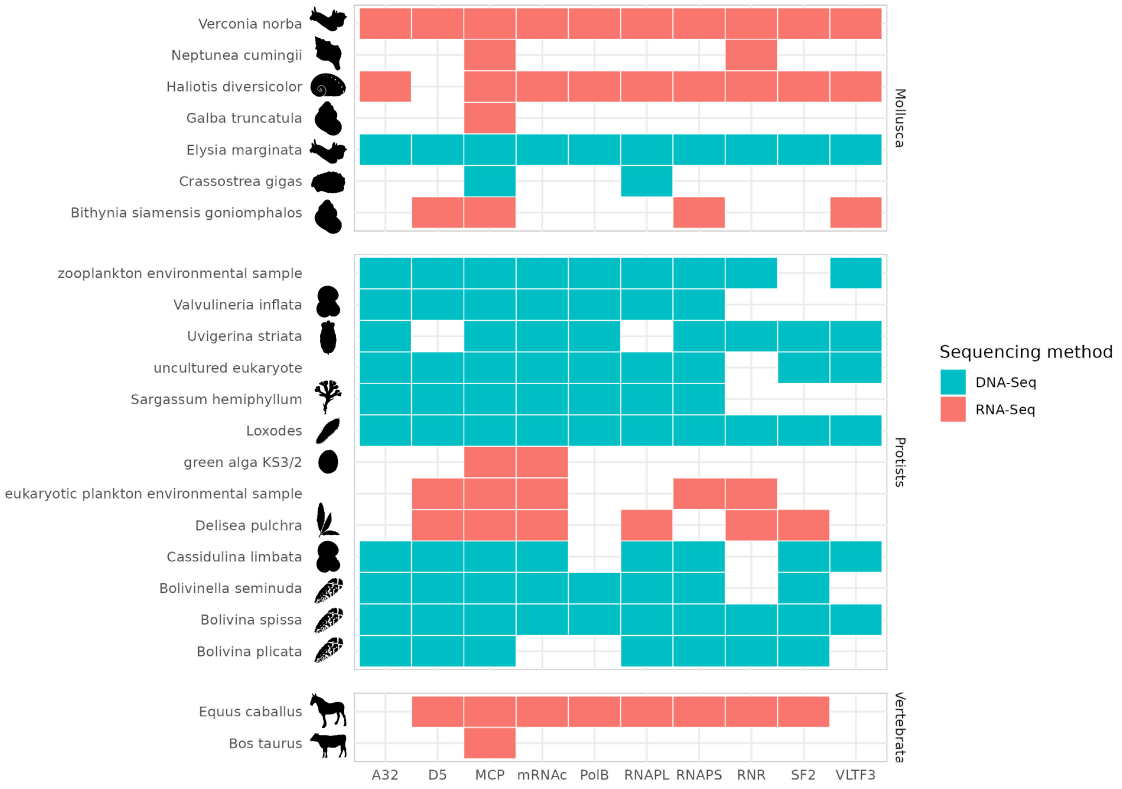

Figure 7: Asfarvirus marker genes found in sequencing data of various mollusc, vertebrate and protist species. The color fill represents the sequencing method. The icons resemble the habitus of the host species. Projects with an almost complete set of markers are *Verconia norba, Haliotis diversicolor, Elysia marginata*, the *Rhizaria* species and the vertebrate project of *Equus caballus*.

Except for the protist datasets, which mostly originate from metagenomic datasets, many asfarvirus markers were found in RNA sequencing data (Figure 7). Perhaps there are more viral transcripts present in the sample than viral genomic DNA, or it

could be that the sequencing coverage is lower in WGS compared to RNA-Seq.
Some of the novel virus sequences (Figure 7) are non-complete proteins and shorter than reference proteins. Only the MCPs are discussed here, as they were used as phylogenetic markers. Especially in the case of *Neptunea cumingii*, *Crassostrea gigas*, *Galba truncatula* and *Bithynia siamensis goniomphalos* the MCP lengths are only around 105, 120, 170 and 180 amino acids (aa), respectively. For *Verconia norba*, *Haliotis diversicolor* and *ElyMa* likely full-length MCPs were assembled. Among the assemblies from the protist datasets, the *Cassidulina limbata*, *Sargassum hemiphyllum* and *Loxodes* MCPs are below 200 aa and the others might be complete proteins, too. The asfarvirus MCPs from vertebrates show protein lengths of 110 aa (*Equus caballus*), whereas the *Bos taurus* sequences was likely full-length with 490 aa.
MSAs with reference proteins from members of the *Nucleocytoviricota* were then used to determine whether the novel MCPs were authentic, reliable and if they could be used for phylogeny. Regardless of their lengths, all novel sequences aligned neatly to the reference MCPs, confirming their authenticity.

Then, the sizes of the nucleotide contigs of the asfarvirus markers were analyzed (Figure 8, part A). In most cases, the viral marker contigs are smaller than 10,000 bp and only the protist projects zooplankton sample and *Cassidulina limbata*, yielded contigs larger than 10 kb. Overall, the markers' contigs have a low read coverage, mostly below 30. This often correlates with small contig sizes as low coverage leads to poor assembly outcomes, while high coverage leads to better results. Again, one exception is the zooplankton sample that has a decent coverage for most of the contigs what can be correlated to the increased contig sizes. A few contigs stand out, for example one from a project of protist *green alga KS32* with a value of 130

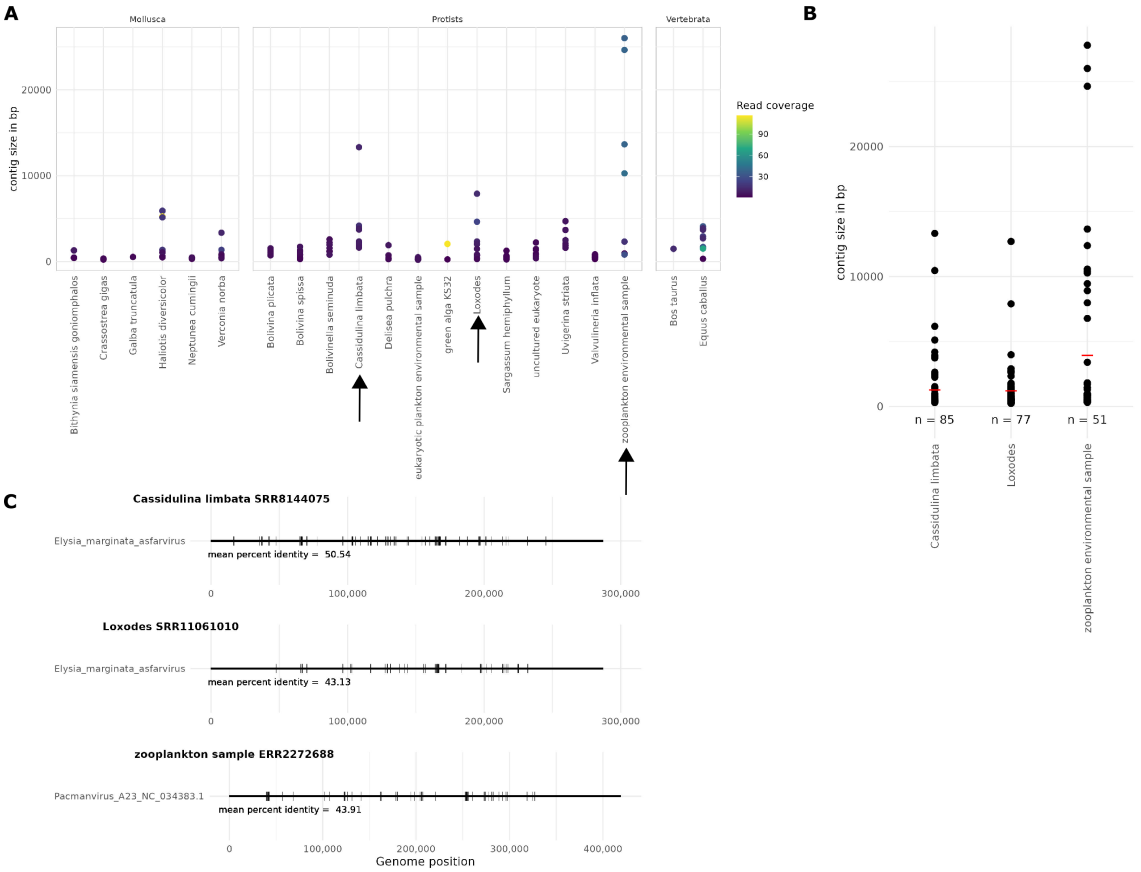

Figure 8: **A:** The sizes of the marker contigs found in various molluscan, protist and vertebrate species. The results from *ElyMa* are not included and will be discussed later. The y-axis depicts the length of the nucleotide contigs in bp and the longest contigs were recovered from the protist project of a zooplankton sample. The read coverage is indicated by the color of the data points, from purple to yellow. The arrows mark the datasets with the largest asfarvirus contigs, which were further screened for other virus genes.
B: The number of viral contigs and their sizes were increased after searching for all virus genes, not just markers. The red lines indicate the average contig sizes. **C:** Alignment of the virus genes with the closest related virus genome. It can be seen that the virus contigs match all across the reference genomes, indicating the presence of full-length virus genomes in the raw data.

and one contig from vertebrate *Equus caballus* with a value of 68. To prove that the marker hits are valid the assembled contigs from projects with the largest marker contigs were re-screened for other viral proteins. A tBLASTx search of the asfuvirus genomes against the entirety of contigs revealed many more viral sequences and the contigs' sizes could be increased (Figure 8, part B). Around 50 to 85 viral contigs

were recovered, but many of which were only a few kilobases in size. All viral contigs were aligned to the genome of the closest relative (EMAV and PV) and the contigs match all across the reference genomes, which provides evidence for the presence of genetic material from probable full-size virus genomes even if no meaningful draft genomes or fragments could be assembled (Figure 8, part C).
In summary, these results strongly suggest that there are various asfarviruses in molluscs, protists and perhaps even vertebrates that have not yet been reported. Further investigation of these novel sequences and periodic screening can help fill the gaps in the understanding of asfarviruses.

3.2.3 Assembly of the *Elysia marginata asfarvirus* - EMAV

The *ElyMa* hit is from a BioProject with six different sequencing runs (BioProject: PRJDB3267; SRA: DRR238951 - DRR238956) and originated from one biological specimen of the whole body of *Elysia ornata*, which was collected in Japan (BioSample SAMD00025083). For any reason, the registered species of the GenBank accession (BMAT00000000.1) changed from *Elysia ornata* to *Elysia marginata*, which is the reason why the virus was named *Elysia marginata asfarvirus*. Non-targeted *de-novo* and super-assembly of all of the six unprocessed sequencing datasets were performed and two large viral contigs with a length of 270,017 bp and 82,540 bp were discovered. The 83 kb contig was spanning the ends of the 270 kb contig with a large overlap (Figure 9) and thereafter manual assembly yielded one single draft genome with a length of 286,835 bp. This assembly was a challenging task, as the terminal parts of the genome contained a plethora of redundant and repeating sequences and genes

(Figure 9 and Figure 10, bottom). The genome's ends contain mating read pairs that align across the terminal parts, which suggests that the genome might be circular and not just with inverted terminal repeats (ITRs) or covalently closed hairpin termini as has been described for other members of the *Asfuvirales* and *Nucleocytoviricota* (Geballa-Koukoulas *et al.*, 2020; Geballa-Koukoulas *et al.*, 2021).

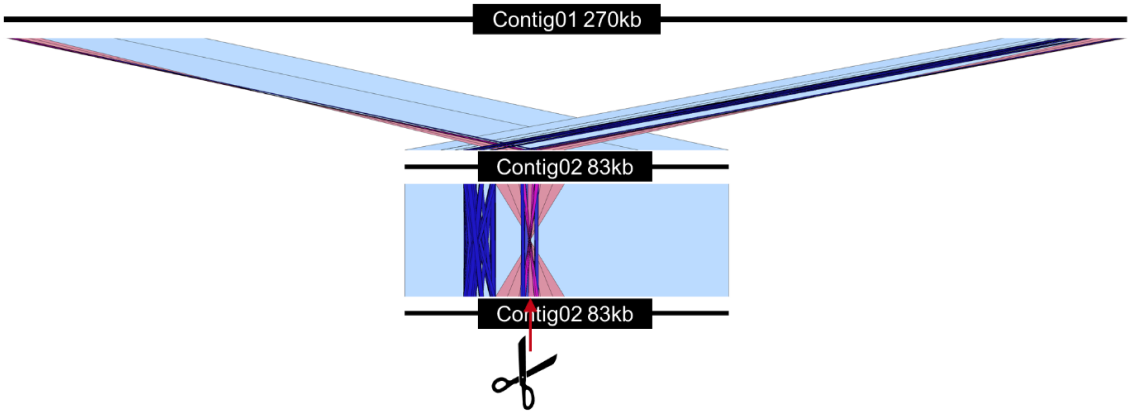

Figure 9: Alignment of the contigs of the novel EMAV genome to visualize the overlapping ends. This provides evidence for the circularity of the genome. The top line shows the large 270 kb contig and the two bottom lines the self-aligning 83 kb contig. The scissors indicate the position between the ITRs and where the genome has been cut for linear display. The blue lines show linear homology and the red lines show inverted homology. The darker the color, the lower the similarity between the two contigs.

When performing BLAST searches (Altschul *et al.*, 1990) with proteins encoded on the EMAV genome, some were already in the GenBank database, although labelled as proteins of host *ElyMa*. This might be the reason why recently several genes of the EMAV have already been notified as novel virus genes (Kao *et al.*, 2023; Hannat *et al.*, 2023; Zhao *et al.*, 2023b; Matsuyama *et al.*, 2023). However, the correct annotation, the complete viral genome and its characterization were still missing. A GenBank entry of the virus genome was requested and can be accessed under the accession number BK063403.

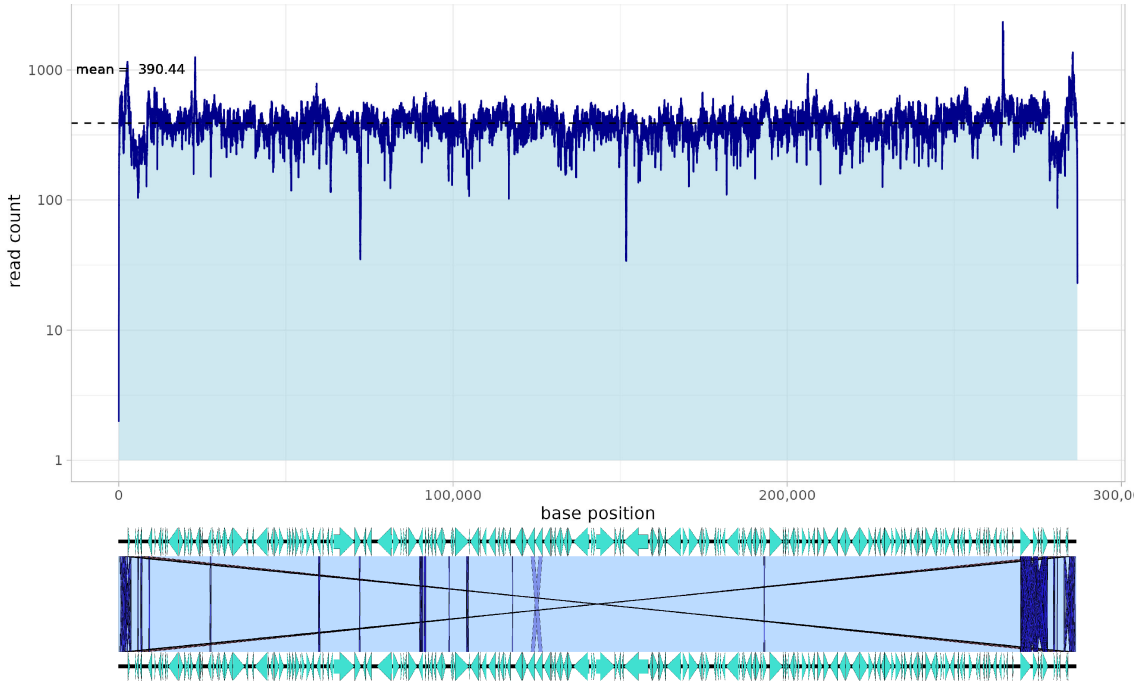

Figure 10: Read coverage of the full-length EMAV genome summed up for six sequencing runs of the sequencing project of *ElyMa* (PRJDB3267). The y-axis shows the read count and is logarithmic. The read coverage is fairly high but declines at the terminal parts of the genome due to terminal redundancies, which can also be seen in the self-alignment at the bottom. The coverage is congruent with the genome maps. In the genome maps, the green arrows show ORFs and the darker the linking line the lower the similarity.

To assess the reliability of the assembled genome sequence the read coverage was examined for all of the six sequencing runs (DRR238951 - DRR238956, Figure 10). The average read coverage of the whole genome in each of the runs ranged from 21 to 212 and the combined read coverage depth average was 390. This high abundance of viral genetic material in the biological sample of *ElyMa* suggests an active replication of the virus. The overall read coverage is very high, but declines towards the ends of the genome due to terminal redundancies and repeats. Moreover, the EMAV genome also shows numerous self-complementary regions that is visible in the self-alignment (Figure 10, bottom). The other asfuviruses also show a similar pattern of self-complementarity, especially in the case of ASFV (data not shown). EMAV

shows ITRs with a length of around 8.5 kb, which are tandem repeat regions with complementary orientation. For better visualisation, the genome was linearized and cut between the ITRs (Figure 9, Figure 10, bottom). Other asfarviruses also contain ITR regions but they are significantly shorter, for example ASFV, AbALV, KV or FV with a size of around 2.1, 0.78, 1.1 and 0.69 kb, respectively (Rodríguez *et al.*, 2015; Geballa-Koukoulas *et al.*, 2021; Geballa-Koukoulas *et al.*, 2020). Similarly, also poxviruses have ITR regions with sizes in between 1 and 17 kb. Interestingly, many ITRs of poxvirus might be involved in host immune evasion (Brennan *et al.*, 2022). It remains to be verified if the ITRs of EMAV have a similar function.

3.2.4 Quality and completeness of the EMAV genome - MIUViG

To assess the quality and completeness of the EMAV genome, I considered the criteria of the "Minimum Information about an Uncultivated Virus Genome (MIUViG)". These MIUViG criteria have been established by a variety of high-ranking members of the scientific community (Roux *et al.*, 2019). First, the genome sequence is a single contig without gaps or ambiguities (Figure 10). Second, the completeness and circularity of the genome was demonstrated by the overlapping ends of the contigs (Figure 9). Adding to that, the genome completeness can be assessed by genome length comparison with relatives with a size deviation threshold of a maximum of 10 %. EMAV and AbALV both have a putative molluscan host, which supposedly makes them close relatives and their genome sizes differ by only 6 kb or 2 % (Table 1). Third, the presence of a complete set of marker genes also supports the assumption that the genome is complete (Figure 7 and Figure 14). Finally, I

performed a comprehensive manual review to annotate putative gene functions and transcriptional units, which will be the topic of the following sections. In conclusion, I propose the EMAV genome to be complete and finished according to the MIUViG criteria.

3.2.5 Comparative genomics of asfuviruses

Comparative genomics is a biological research discipline that compares DNA or RNA sequences of different species to analyse the genes encoded, gene order, gene function, and phylogenetic relationships. Identifying similarities and differences between them improves understanding of their evolutionary relationships. Here, the novel genome of the EMAV will be compared to the other *Asfuvirales*.

3.2.5.1 Genomes facts and figures

In the ICTV report of 2019 there are four members of the order *Asfuvirales*, namely ASFV, *Fausto-* (FV), *Pacman-* (PV) and *Kaumoebavirus* (KV). In 2020, the *Abalone asfa-like virus* (AbALV) was discovered and might be the fifth *Asfuvirales* member and the first asfarvirus with a molluscan host, however the official proof is missing (Matsuyama *et al.*, 2020). Here, I propose the genome sequence of a second asfarvirus found in a shell-less mollusc called *ElyMa* hence named *Elysia marginata asfarvirus* (EMAV). The features of the known and novel asfuviruses will now be discussed.

Table 1: Basic characteristics of EMAV and *Asfuvirales* genomes. Depicted are the genome length, GC content, number of ORFs, ORF densities, orthologous genes, number of MCPs, tRNAs and the natural hosts. The order is according to their genome sizes.

Virus	Genome length in kb	GC content in %	Number of ORFs	ORFs per kb	Ortholog. genes	MCPs	Genome organisation	tRNAs	Host
ASFV	186	38.4	164	0.88	64	1	linear	0	mammalia
AbALV	281	31.3	345	1.23	57	2	linear	0	mollusca
EMAV	**287**	**55.7**	**190**	**0.66**	**69**	**2**	**circular**	**0**	**mollusca**
KV	351	43.7	429	1.22	46	1	linear	0	protist
PV	419	33.6	444	1.06	85	1	linear	1	protist
FV	466	39.8	506	1.09	88	1	linear	0	protist

The characteristics of the current asfuviruses including AbALV and EMAV are shown in Table 1. ASFV has the smallest genome size depending on the genotype ranging in between 170 kb and 194 kb (Alonso *et al.*, 2018), while amoeba asfuviruses have the largest, with FV reaching almost 470 kb (Reteno *et al.*, 2015). The GC content of the six virus genomes range in between 31 % for AbALV and 56 % for EMAV. For example, in the group of around 2,000 *Nucleocytoviricota* MAGs discovered in the study of Schulz *et al.*, 2020b, the GC content is very variable and ranges from 17 % of a cafeteriavirus to 67 % of *Squirrelpox virus*. For PV one isoleucine tRNA was predicted (Andreani *et al.*, 2017) and the other viruses do not appear to encode for tRNAs. There are above 20 genotypes of ASFV, the virus variant Liv 33 was chosen as reference (Zhao *et al.*, 2023a; Chastagner *et al.*, 2020).

The numbers of ORFs and the genome sizes were correlated and there are two groups with similar ORF densities (Figure 11). One group consists of ASFV and EMAV (Figure 11, lower part) with a moderate ORF per kilobase value of 0.9 and 0.7, respectively. AbALV and the amoeba viruses form another group with ratios ranging from 1.1 to 1.2. Interestingly, the AbALV and EMAV differ considerably with respect to the the number of ORFs, despite their similar genome length. Especially, the terminal parts of AbALV are packed with ORFs much tighter than in the case of EMAV (Figure 11 and synteny plot in Figure 14). Further investigation is required to test, if for example the genes of EMAV undergo complex splicing, as for example genes of FV (Louazani *et al.*, 2018).

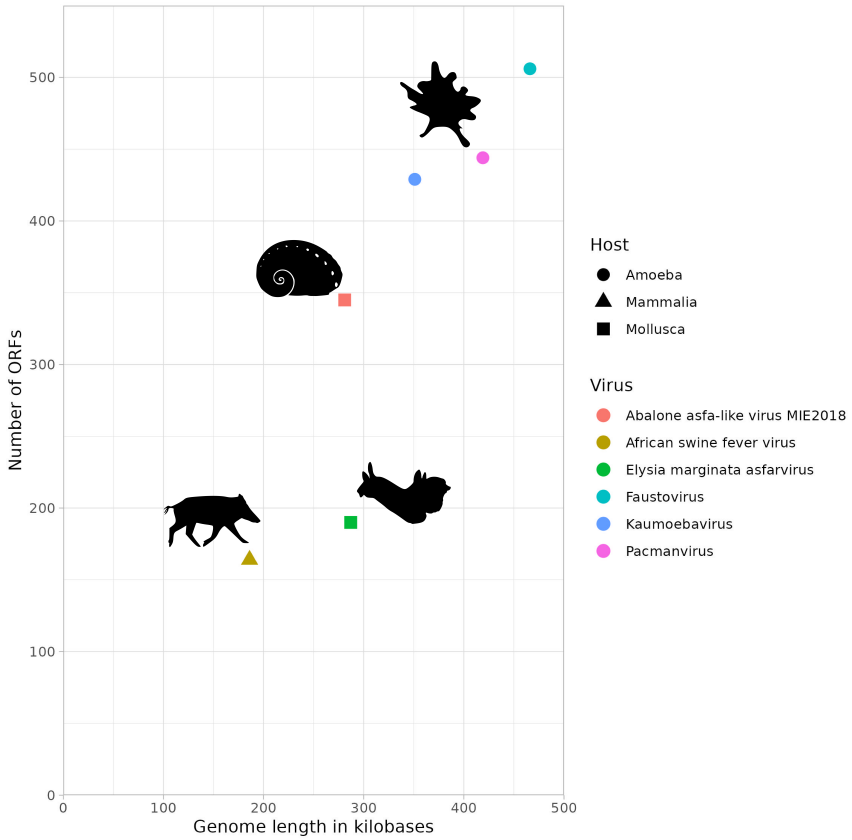

Figure 11: Comparison of the ORF densities of the asfuviruses show that ASFV and EMAV form one group with similar ORF per kilobase of genome ratios. AbALV and the protist viruses form another group. The putative natural hosts are indicated with a symbol.

3.2.5.2 Quantifying similarity using average amino acid identity (AAI)

One method to assess the similarity of viruses and subsequently taxonomic relations is the whole-genome average nucleotide identity (ANI), but due to the low inter-species nucleotide similarity of asfuviruses, the average amino acid identity (AAI) was utilized. The AAI is the mean value of pairwise protein similarities and based on

that the operational taxonomic units at the virus species level were defined. ASFV and EMAV had the highest AAI of 39 % so slightly more than AbALV and EMAV with an AAI of 38 % (Figure 12). When evaluating that EMAV appeared to be closer related to ASFV than AbALV, it should be considered that there was a standard deviation of around ± 4.5. Overall, based on this AAI analysis EMAV, ASFV and AbALV formed one cluster and the amoeba viruses formed another. ASFV and KV had the lowest AAI of only 31 % and for some of the viruses the AAI value was based on the comparison of only very few genes due to high viral sequence divergence (Figure 12, numbers in tiles).

Taken together, this analysis shows that supposedly closely related asfuviruses are still highly divergent and share only a few proteins with the relatives known to date. Either the evolutionary steps by which these viruses evolved were very large, they occurred very long ago or the linking viruses have simply not yet been discovered.

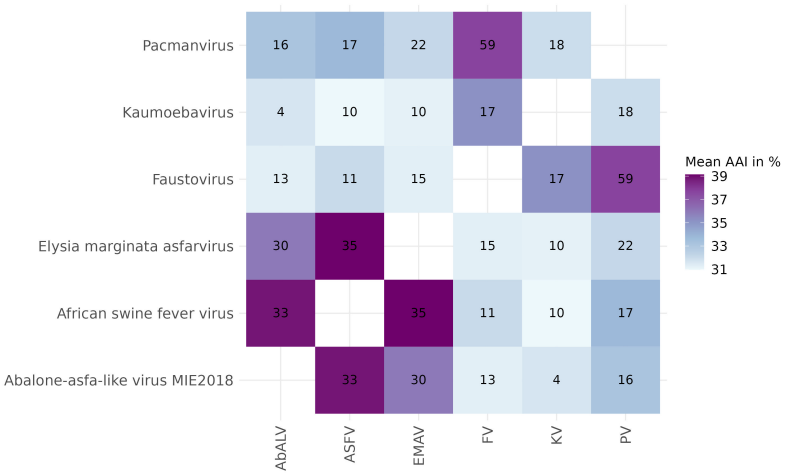

Figure 12: The mean pairwise AAI for the asfuviruses range from 31 % to 39 % and EMAV is closest related with ASFV and AbALV. The darker the fill color the higher the pairwise AAI. The numbers in the tiles show the amount of proteins that were compared within 25 % sequence identity and 60 % alignment length thresholds.

3.2.5.3 Detection of orthologous/shared genes

For further characterization, the proportion of orthologous genes was quantified including EMAV and all *Asfuvirales* members. Orthologous genes, also known as shared genes, are similar genes that are common to multiple viruses. Following that, the ratio of orthologous genes to the total number of genes was determined (Figure 13).

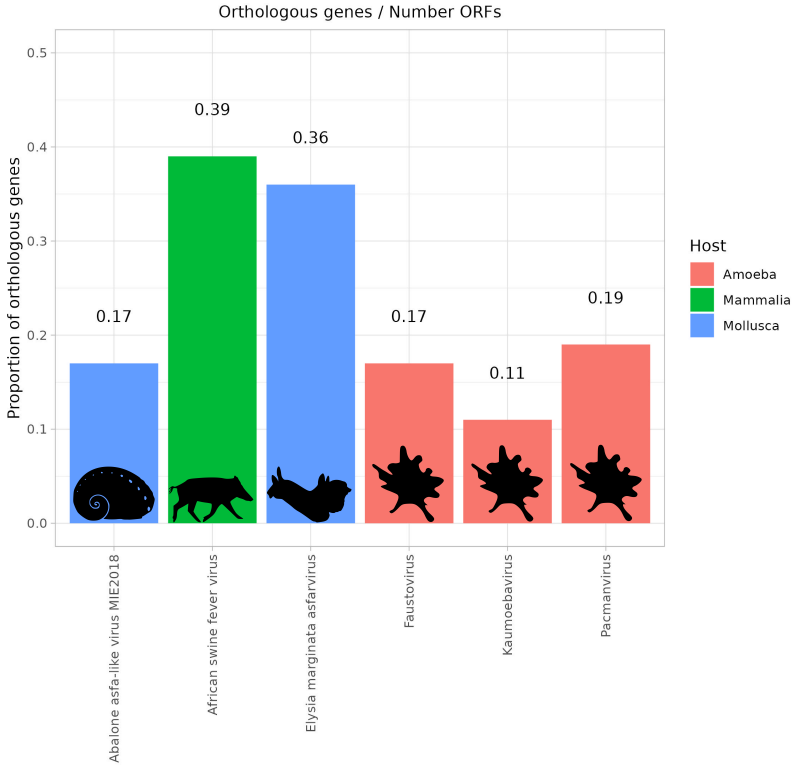

Figure 13: Orthologous genes are genes that are shared with related reference viruses. ASFV shares almost 40 % and KV only 10 % of their total genes. The colors of the bars and the icons indicate the putative host groups of the viruses.

EMAV and ASFV have a similar ratio of orthologous genes of around 0.4 and AbALV and the protist viruses also have similar values, ranging from 0.1 to 0.2 (Figure 13). A previous analysis (Figure 11) showed that AbALV has a larger number of ORFs

than EMAV, despite their similar sized genomes. Here, it was demonstrated that many of the genes of AbALV appear to be unique to the virus, while EMAV has twice the proportion of shared genes compared to AbALV (Figure 13). Table 1 provides an overview of the number of orthologous genes per virus.
Taken together, most of the genes seem to be non-orthologous and unique to the individual viruses, which was observed for other *Nucleocytoviricota*, as well (Mönttinen *et al.*, 2021). Some of the unique genes may be host specific or originate from non-coding proteins. Alternatively, many genes may have diverged over very long evolutionary timescales, making sequence similarity no longer detectable.

3.2.5.4 Synteny shared between asfuviruses

Next, the gene content and synteny of the three complete asfarvirus genomes were analyzed using a tBLASTx-based graphical comparison (Figure 14). Comparing the genome structure of a novel virus to that of its closest relatives is significant for its classification. The labelled ORFs are viral hallmark genes that are characteristic proteins, shared by most members of a virus group and can therefore be used to evaluate the completeness of a virus and to build phylogenetic reconstructions. EMAV encodes for a complete set of ten markers. The colors of the ORFs indicate the 73 inter-species groups of orthologous proteins, which are proteins of similar type. When comparing the three virus genomes in terms of ORF order, orientation, and predicted orthologous groups, many of the features are similar (Figure 14).

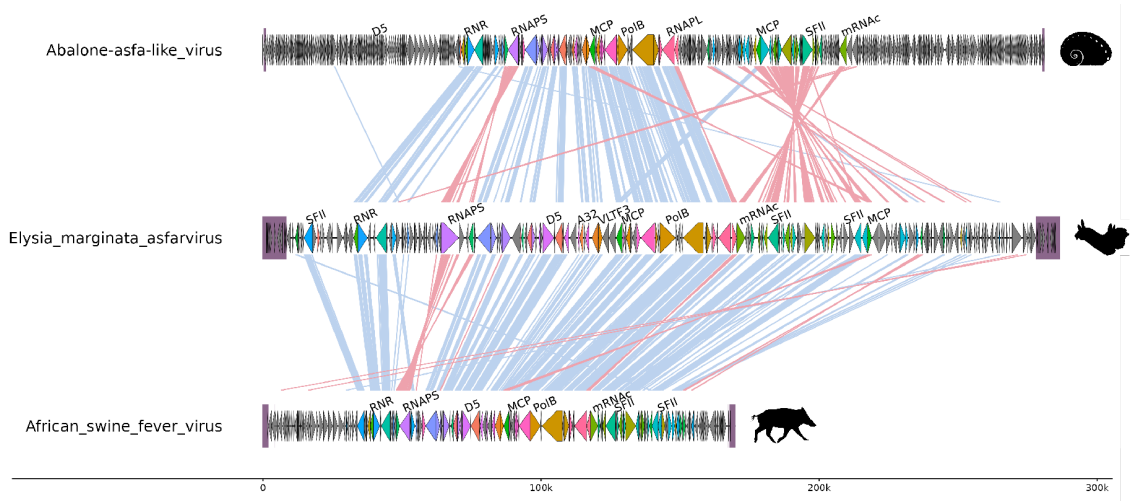

Figure 14: Synteny of the AbALV, EMAV and ASFV based on amino acid similarity. The lines connect the homologous regions and it can be seen that in particular the central parts of the genomes, are well conserved. A blue line shows a linear homology and lines in red indicate inversions. The asfarviral hallmark genes are labelled. The colors of the ORFs show the 73 orthologous groups, which are inter-species proteins of similar type. ITRs are labelled with purple bars at the outer parts of the genomes. All of the viruses encode for a complete or almost complete set of marker genes. The marker genes are: D5 - DNA primase/helicase, RNR - Ribonuclease Reductase, RNAPS + RNAPL - DNA-directed RNA polymerase small + large subunit, SF2 - DEAD/SNF2-like helicase, A32 - Packaging ATPase, MCP - NCLDV major capsid protein, VLTF3 - Poxvirus Late Transcription Factor VLTF3, PolB - DNA polymerase family B, mRNAc - mRNA guanylyltransferase.

In more detail, one conserved genome cassette is in the center of the genomes and contains markers like A32, VLTF3, one MCP version, PolB and mRNAc (Figure 14). A similar order can be observed in the genomes of the amoeba viruses PV, FV and KV, as well (data not shown). The RNAPS region of EMAV is inverted compared to ASFV and AbALV (Figure 14, purple ORFs connected by red lines at approximately base position 75 kb). On top of that, when comparing AbALV with EMAV a 25 kb fraction that is located at the 3'-end is inverted and covers the markers mRNAc, a second MCP and one SF2. This inversion seems to be unique for AbALV. Taken together, the virus genomes have a homologous genome structure in the center, whereas the terminal parts differ in terms of ORF density and length of terminal

repeats (Figure 14, terminal purple bars). When comparing the links between the viruses, especially the outer regions of EMAV show homology, whereas AbALV and ASFV have large terminal parts without. This leads to the assumption that the ends of the genomes are particularly prone to modification.
Moreover, towards the ends of the genomes there are many paralogous gene repeats especially in the case of AbALV and ASFV but also EMAV. Many of these repeating genes are akyrin repeats and appear to be a common feature of asfarviruses and *Nucleocytoviricota* and might have evolved from duplication and tandem repeat gain, but their functions remain unknown (Erdozain *et al.*, 2023; Matsuyama *et al.*, 2023). For ASFV, it has been reported that the terminal regions contain genes of the multigene families, which are highly variable between virus variants and are vigorously exchanged by homologous recombination (Michaud *et al.*, 2013). In general, nucleocytoviruses share a highly versatile genome architecture known as genome plasticity, which can occur through duplication, deletion or lateral gene transfer (Mönttinen *et al.*, 2021), and this can be confirmed from these results.

Finally, it is worth noting that there are two gene variants of the MCP encoded by EMAV and AbALV. In contrast, ASFV has only one MCP (Figure 14, two connecting lines of EMAV MCPs merge into the single MCP of ASFV at approximate base position 100 kb). Also, PV, FV and KV only have one MCP. These EMAV and AbALV MCP variants will be analysed in more detail in section 3.2.7 on page 50.

3.2.5.5 Functional annotation and bipartite network analysis

Next, a sequence-based, functional annotation of the ORFs found on the *Asfuvirales* genomes was performed and all viruses commonly encode proteins involved in transcription, RNA processing, replication, post-translational modifications, chromatin structure and dynamics, transport and metabolism of nucleotides, coenzymes and amino acids (Figure 15). These proteins appear to be ubiquitous for the entire asfuviruses. On the other hand, some genes are not present in EMAV, AbALV and ASFV and can only be found in the amoeba viruses (FV, KV, PV), including genes for products involved in translation ribosomal structure and biogenesis (Figure 15, row 2). Proteins in the categories of defense mechanisms and cell wall/envelope biogenesis were found exclusively in EMAV and protist viruses. EMAV has the highest proportion of annotated genes with around 23 % and the AbALV the lowest with only eight percent. Seventeen percent of the ASFV genes can be assigned a putative function and the amoeba viruses have only ten to thirteen percent. However, the vast majority of encoded proteins cannot be assigned a putative function (Figure 15, row 1). Some proteins without assigned function represent very short ORFs (sORFs), which are located particularly at the terminal regions of the genomes (see Figure 14). It remains to be determined, whether these short ORFs code for proteins of functional relevance or are non-coding (Finkel *et al.*, 2018). The large proportion of genes with unknown function, highlights the need for a better characterization of their genetic repertoire, perhaps involving additional *Asfuvirales* members discovered in the future, to improve the understanding of their highly complex genomes.

Figure 15: Sequence-based, functional annotation of asfuvirus ORFs. The top line shows the number of proteins that could not be assigned any putative function. The largest number of assigned proteins are from the category replication, recombination and repair. It appears that certain categories of proteins are present in all viruses, while others are found only in individual viruses.

Following this, a bipartite network analysis was carried out involving EMAV and the five members of the *Asfuvirales*. The orthologous genes, which are genes shared by at least two viruses, are shown together with details of the genes' functions (Figure 16). Not necessarily, all the proteins with assigned function are orthologous genes and therefore the numbers may differ from numbers in Figure 15.

In total, there are 113 orthologous genes and twenty-four genes are shared by all of the viruses (Figure 16, large cluster in the center). Many of these proteins are involved in DNA and RNA processing, transcription and transport processes. ASFV, AbALV and EMAV exclusively share eleven proteins (Figure 16, cluster bottom center) and only three of these genes have been functionally assigned, one to posttranslational modification, protein turnover and chaperones, the second to carbohydrate transport and metabolism and the third to replication, recombination and repair.

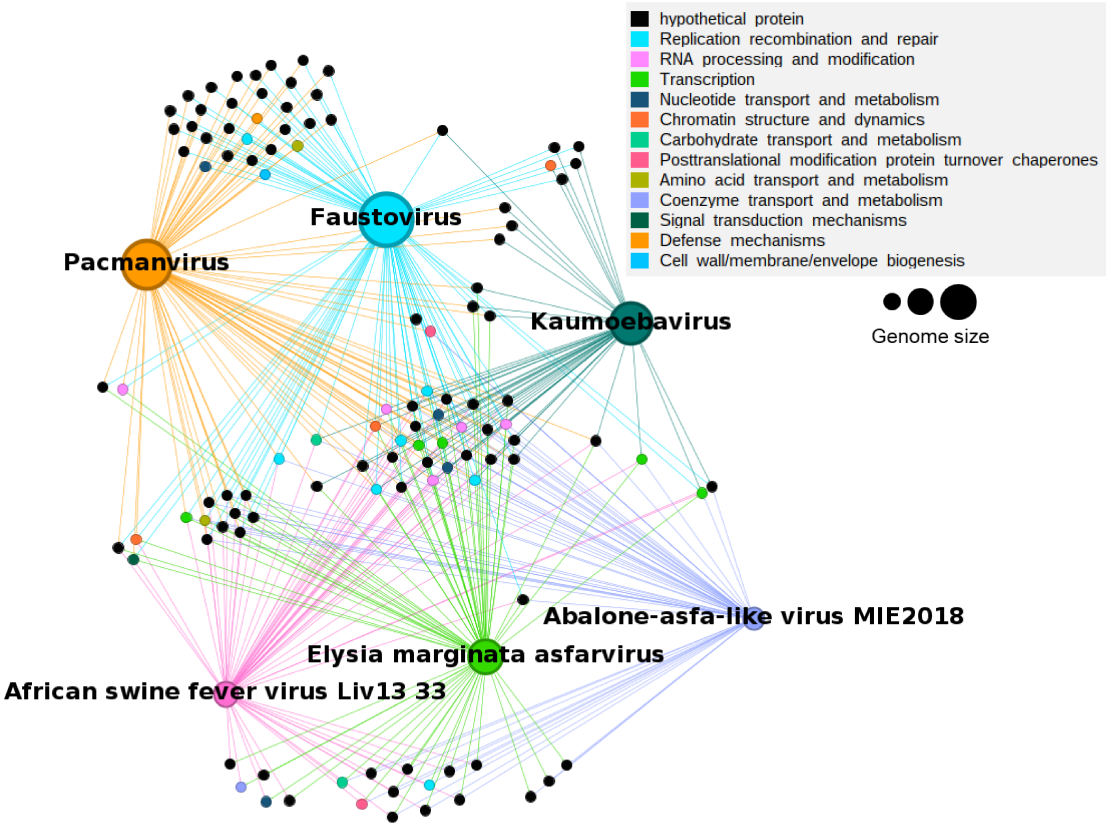

Figure 16: Bipartite network analysis of EMAV and the five asfuviruses. The sizes of the large virus dots indicate the genome sizes. The small dots represents the orthologous genes and their colors indicate their putative function. Small black dots are proteins of unknown function. 113 genes are shared by at least two viruses and the entirety of the viruses share 24 genes, which are the proteins that cluster in the center. The three asfarviruses, ASFV, AbALV and EMAV share 11 genes. PV and FV have 30 genes in common and KV shares only three genes with PV and FV. The taxonomic units can be estimated from the number of shared genes.

Moreover, ASFV, AbALV and EMAV do not share any proteins that influence cellular housekeeping tasks such as transcription or RNA processing and modification but do share such proteins with the other viruses at a higher taxonomic rank like the order. FV and PV share 30 genes whereas KV shares only three genes with the other amoeba viruses. In the context of virology, an understanding of which genes have which functions and how they are shared within a group of viruses is very insightful. Furthermore, by simply comparing how many genes are shared with related viruses,

this network provides a good understanding of the taxonomic relationships between viruses. It can be assumed that ASFV, EMAV and AbALV belong to one taxonomic unit and the amoeba viruses form another with PV + FV mated to KV. Nevertheless, more details about the accurate phylogenetic relations will be the topic of the next chapter.

3.2.6 Phylogenetic analysis of novel asfuvirus sequences

Defining the operational taxonomic units at the virus species level was one aim of the AAI analysis (section 3.2.5.2 on page 37) and the bipartite network analysis (section 3.2.5.5 on page 43). To overcome any remaining uncertainties about the taxonomic relations between EMAV, ASFV and AbALV, a phylogenetic analysis that included *Nucleocytoviricota* references was carried out based on six marker genes, namely MCP, RNAPS, PolB, VLTF3, D5 and A32 (Figure 17).
Now, EMAV appears to be in one clade with AbALV, the other molluscan virus, and both form the sister group to the ASFV (Figure 17, bottom). According to these results, the three viruses ASFV, AbALV and EMAV belong to a *Asfarviridae*-like family. The PV and FV form a distinct group closely related to the *Asfarviridae*, while the KV is separated from both. Altogether, they form the monophyletic clade of the *Asfuvirales*. As the molluscan asfarvirus AbALV and EMAV form a well supported monophyletic lineage, the conclusion is that EMAV is closest related to AbALV.

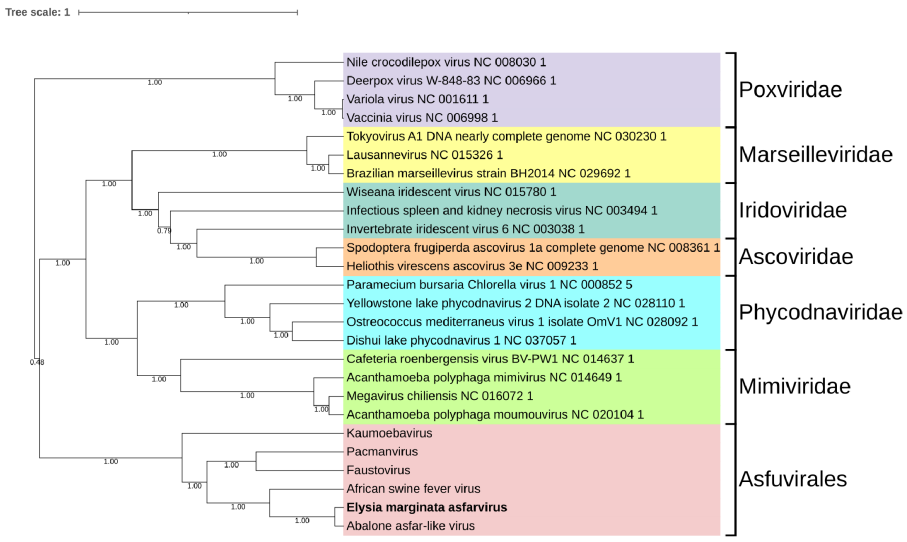

Figure 17: Phylogenetic tree of EMAV and *Nucleocytoviricota* references based on the six marker genes MCP, RNAPS, PolB, VLTF3, D5 and A32. EMAV is the closest relative of AbALV and both together form a sister group to ASFV. Including PV, FV and KV they form the order of *Asfuvirales*. The other *Nucleocytoviricota* references are located in a subtree within the sistering clade. The *Poxviridae* form the outgroup.

Next, an MCP-based phylogenetic analysis (Figure 18) was conducted to resolve the evolutionary relationships of the new virus sequences discovered in protists, molluscs, vertebrates and other *Nucleocytoviricota* references and MAGs (O'Leary *et al.*, 2016; Schulz *et al.*, 2020b; Moniruzzaman *et al.*, 2020). The analysis consists of 79 MCPs, 22 of which are novel sequences discovered as part of this study (Figure 18, red dots). The *Asfuvirales* subtree (Figure 18, bottom) is composed of three major clades prototyped by KV, FV+PV and ASFV, AbALV, EMAV, respectively. This MCP phylogeny confirms that FV and PV are closely related, nested in one clade, while KV sits in a sister clade. This relation has already been seen in many of the analyses stated above. The large majority of viruses in these amoeba virus clades have protist hosts or are isolated from MAGs. It is assumed, that a subset of the viruses from MAGs also have protist hosts whose identities remain unknown. This clade includes the yellow-labelled MCP from a *Bos taurus* project and within the

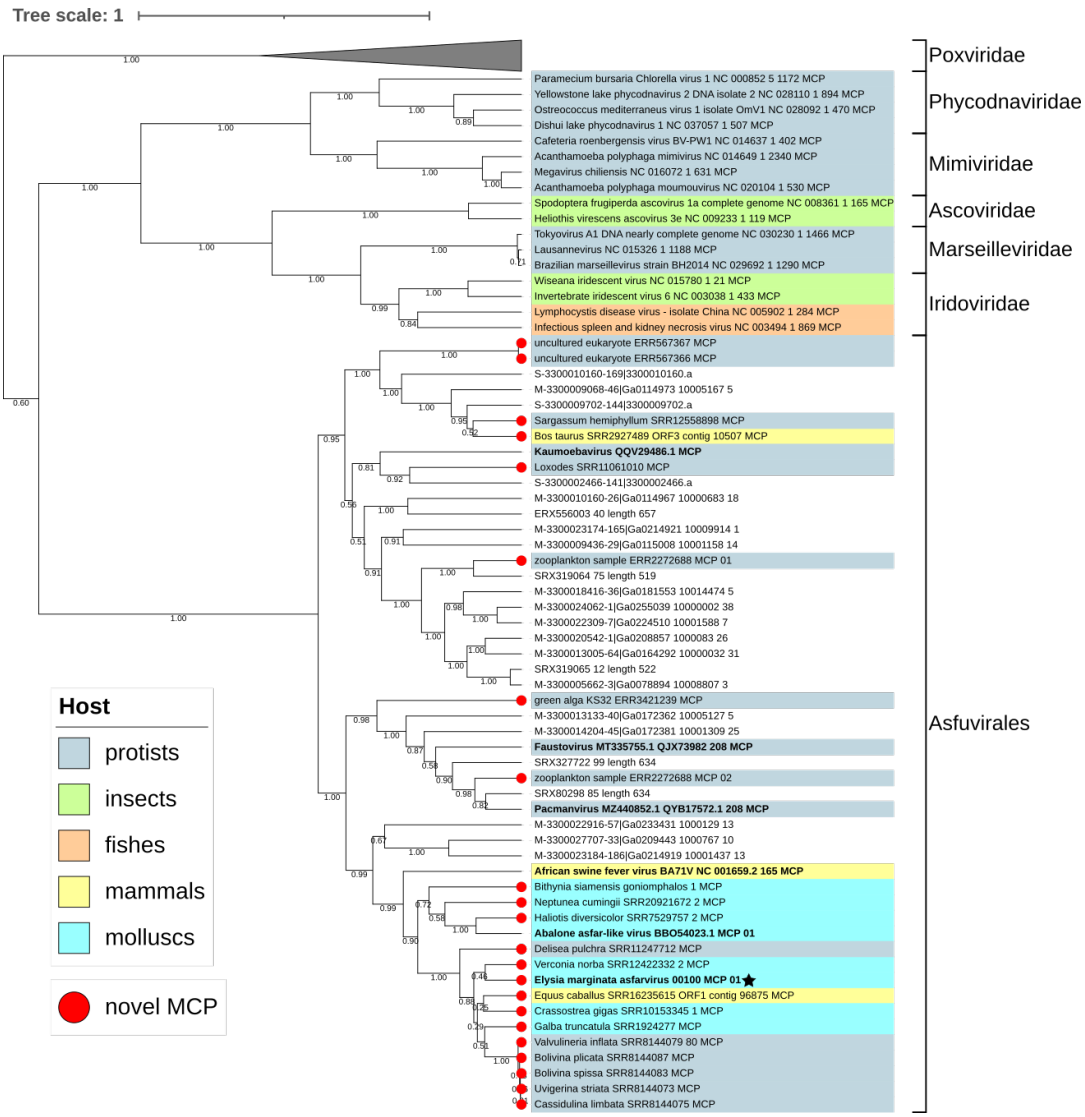

Figure 18: Bayesian tree of 79 MCP sequences from *Nucleocytoviricota* and asfarviruses. The top part shows the *Poxviridae* outgroup and the *Nucleocytoviricota* references, and the bottom part represents the clade of the *Asfuvirales*. Novel asfuvirus MCPs are labelled with a red dot. The boldly labeled MCPs are the references and the colors indicate the putative hosts. The uncolored sequences come from MAGs, such as the MCPs from Moniruzzaman *et al.*, 2020 (named after the SRA experiment in which the sequences were detected), and from Schulz *et al.*, 2020b. The MCPs from mollusc sequencing data are colored in blue and cluster together within the clade of the *Asfarviridae* and ASFV (bottom). The MCPs highlighted in yellow are the sequences found in vertebrates. The MCP from horse groups with the molluscan sequences and the MCP from the cow locates within the clade of the viruses from water metagenomes and algae. Moreover, at the very bottom and within the *Asfarviridae*-like sequences, there are novel MCPs detected in protists like *Rhizaria*, which are tangled up with mollusc asfarviruses.

molluscan sequences, there is the MCP from the horse, *Equus caballus*, project. However, if these MCPs originate from *bona-fide* viruses, that infect these hosts remains elusive and requires further investigation. Within the putative *Asfarviridae* clade (Figure 18, bottom) are ASFV, AbALV, EMAV and many novel sequences from sequencing data of mollusc (blue) and protist (grey) hosts. All the novel sequences from mollusc hosts are specific to the *Asfarviridae* and they cluster in two major clades, one with AbALV and the other with EMAV, suggesting more AbALV-like and EMAV-like viruses. Moreover, that the MCPs from molluscan viruses and ASFV share one branch implies a common ancestry of these viruses.

According to the overrepresentation of *Asfuvirales* members identified in protists, the existence of a large and hidden viral diversity in these rather primitive eukaryotes is suggested. The viral phylogeny, in which the KV and FV+PV clades form basal (for example outgroup) lineages to the ASFV clade, further suggests that this might apply to ASFV-like viruses as well and that an ancestral ASFV-like virus may have jumped from a protist into a vertebrate or an intermediate host. Adding to that, even though ticks (*Acari*) were also proposed to serve as hosts of ancestral ASFV-like viruses (Michaud *et al.*, 2013), my screen revealed no evidence for that. However, the fact that still no close relatives of ASFV were found indicates that further screening efforts aiming at the discovery of close relatives of ASFV and involving a wide variety of host groups are required to validate the hypothesis of an asfarvirus origin and reservoir in protists.

3.2.7 Secondary MCP acquisition of several *Asfuvirales* members

During the analysis AbALV, EMAV and the viruses from the *Loxodes*, *Haliotis diversicolor* and *Verconia norba* samples appeared to encode a second, divergent MCP-like protein, contrary to ASFV, which only has the canonical MCP gene. Generally, many members of the *Nucleocytoviricota* phylum encode multiple MCPs, and some members of the *Mimiviridae* or *Phycodnaviridae* even up to ten (Koonin and Yutin, 2019; Schulz *et al.*, 2020b). Recently, Hannat *et al.*, 2023 also reported that AbALV encodes two ASFV p72 capsid homologs, which can be confirm and even extended to other asfuviruses. The MCP variants are referred to as group I being the canonical MCP, closest related to the ASFV p72 capsid, and group II, which is the secondary MCP.

All the detected group I and group II MCPs were combined with MCP references for a joined phylogenetic analysis (Figure 19). It revealed that the five group II MCPs form a well-supported monophyletic sister lineage to the *Asfuvirales* group I MCPs and occupy an intermediate phylogenetic position between group I MCPs of *Asfuvirales* and structural proteins of *Poxviridae* members. A second group I MCP was identified in the *Cassidulina limbata* sample, indicating that genetic material from two different viruses may be present in that sample. Several of the group I and group II MCPs were retrieved from RNA-Seq datasets (Figure 7), suggesting that they are expressed. In conclusion, these phylogenetic relationships can be reconciled with a single gain of the group II MCP via gene duplication or from an external source in an ancient *Asfuvirales* ancestor that existed before the split of the lineages leading to KV, FV, PV and ASFV, followed by diversification of the group II MCP genes. The group II MCP was subsequently lost in an ancestor of ASFV, possibly due to adaptation to the

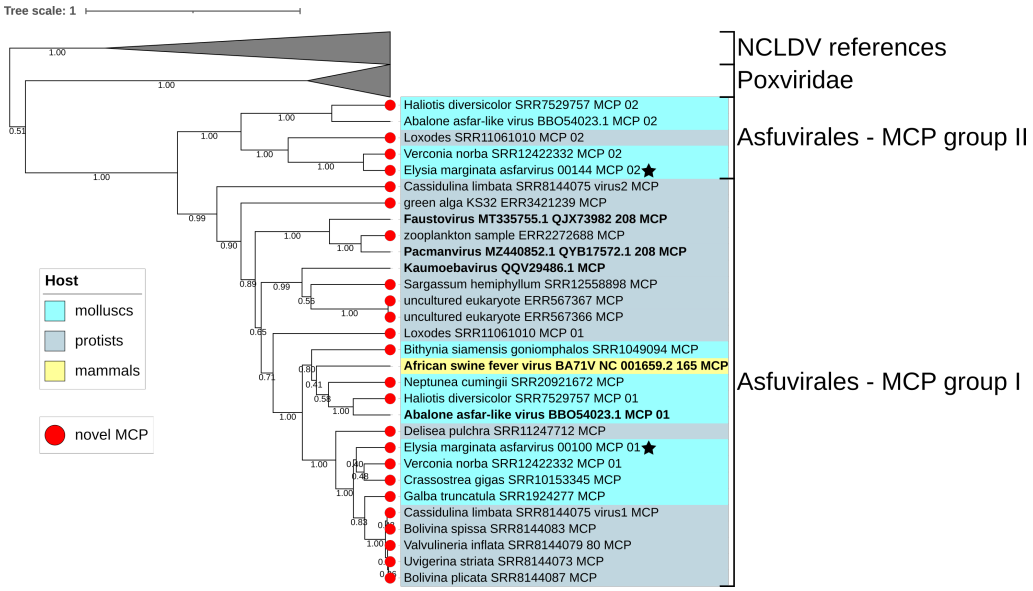

Figure 19: Bayesian phylogenetic tree reconstruction of all the MCPs found in known and novel *Asfuvirales*. The colors show the hosts of the viruses. The red dots indicate novel MCPs. The large bottom clade includes all the group I/canonical MCPs and above the group II/secondary MCPs are located. The close phylogenetic proximity of the group II MCPs indicates a common evolutionary history within the *Asfarviridae*. *Poxviridae* and *Nucleocytoviricota* references form the outgroups.

mammalian host. An alternative, less parsimonious scenario would involve multiple independent group II MCP acquisition events followed by convergent evolution in the different group II MCP-containing viruses. A wider sampling of complete genome sequences of group II-encoding viruses, enabling analyses of gene order and possible similarities of genomic integration sites of group II MCP genes, is required to fully resolve the evolutionary trajectories of the group I and group II MCP genes during *Asfuvirales* evolution.

For further characterization of the novel protein variants, the protein structures were predicted and revealed that all MCPs share a similar double jelly roll domain (DJR) consisting of a double beta-barrel core (Figure 20, dashed boxes). Nevertheless, the peripheral parts of the models differ and some have large beta sheet structures at the

upper domain and putative membrane interacting alpha helices at the lower domain as reported for the ASFV MCP (Andrés *et al.*, 2020). The structural protein D13 of *Vaccinia virus* (VACV) served as an outgroup because it shows structural homology to the major capsid proteins of *Nucleocytoviricota* (Hyun *et al.*, 2022; Figure 20, bottom right). As was already seen in the phylogeny, the group II MCP structures (Figure 20, second column) are more similar to each others than compared to the group I MCPs, especially when considering the peripheral parts. Moreover, that the ASFV MCP is the canonical protein is also supported by the similarity of ASFV MCP with the other group I proteins.

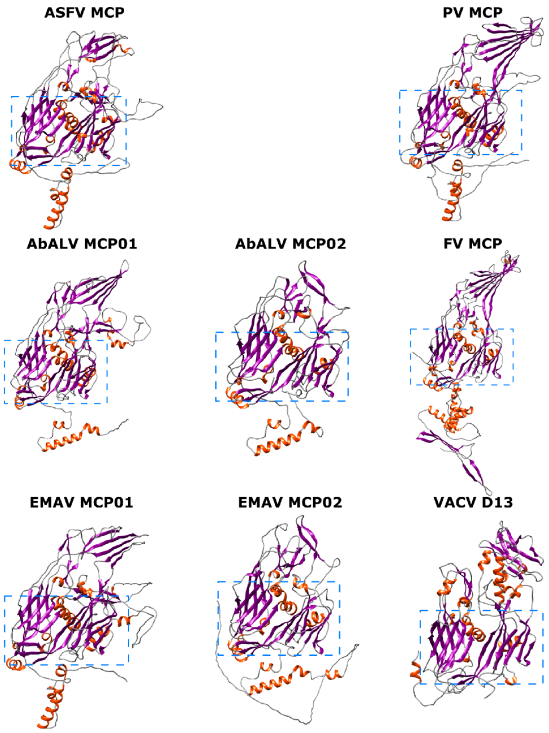

Figure 20: Predicted structural models of the asfuvirus MCPs including the EMAV MCPs and the VACV D13 structural protein. The first column shows the canonical MCPs of the asfarviruses and in the second are the group II MCPs of AbALV and EMAV. The last column contains the MCPs of the other references, PV, FV and VACV, respectively. The dashed boxes show the part of the models that represent the DJR, which is an important feature of this type of capsid proteins. However, in the upper and lower domains there are differences in the arrangement of beta folds and alpha helices.

A structure-based phylogenetic analysis that was built on protein structure models and not on amino acid sequences validated that the group I MCPs form a sister clade to the group II MCPs of AbALV and EMAV (Figure 21). This is in line with previous results (Figure 19 and Figure 20) and undermines the hypothesis that the group II MCPs have a common ancestry, derived from a single gain event. Overall, it is assumed that structural features are better preserved than amino acid sequences. The KV MCP locates outside of the *Asfuvirales* clade and also in previous analysis KV positioned as a stand out virus. Nonetheless, the quality of the KV MCP model was poor and may also cause this odd positioning.

The conserved structures of group I/primary and group II/secondary MCPs imply that these proteins have similar functions. FV, PV and KV form icosahedral virus particles with a diameter of around 250 nm very similar to that of the well-studied ASFV, in which the capsids are built from MCP p72 and minor capsid protein (mCP) p49 (Reteno *et al.*, 2015; Andreani *et al.*, 2017; Bajrai *et al.*, 2016; García-Escudero *et al.*, 1998; Epifano *et al.*, 2006; Wang *et al.*, 2019). According to that and as EMAV and AbALV also express a mCP (Hannat *et al.*, 2023), a similar virus particle structure can be assumed.

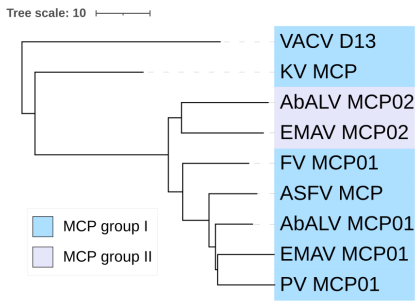

Figure 21: Structure-based phylogenetic analysis of the predicted structural models of the group I and group II asfuvirus MCPs. The group II MCPs form a sister group to the group I MCPs, which is in line with protein sequence-based phylogeny (Figure 19) and indicates a common origin. The structural protein D13 of VACV served as the outgroup.

In conclusion, the multiplication of MCP proteins was found in various known and novel asfarviruses and the similarity of the protein sequences and structures point towards a common origin within the *Asfuvirales*. Whether the group II MCPs are expressed, whether they form functional proteins and, if so, how they impact the virus particle structure remains elusive and requires further investigation.

3.3 Adintoviruses widely distributed in various animal groups

Adintoviruses are derived from polintons, which may form the evolutionary basis of all *Nucleocytoviricota* (Koonin *et al.*, 2015a; Koonin *et al.*, 2015b; Starrett *et al.*, 2021). Especially, when performing the asfarvirus screen in sequencing data of *Mollusca* and *Acari* there were many adintovirus hits. From around 100 *Virushunter* positive and *de-novo* assembled sequencing experiments, 55 contained non-integrated adintovirus contigs, which sizes were analysed (Figure 22, part A). In general, adintovirus genome sizes range in between 14 and 40 kb (Bellas *et al.*, 2023) and for some of the detected adintovirus contigs, detailed analysis revealed complete virus genomes with ITRs and a total length of around 20 kb.

As there were numerous adintovirus hits in the asfarvirus screen, subsequent tests were performed to assess whether the hits are unspecific or come from sequence similarity of *Nucleocytoviricota* and adintoviruses. HMMsearch analyses revealed that there is sequence homology between *Nucleocytoviricota* marker HMMs and adintovirus sequences (adintovirus genomes obtained from NCBI virus and Welch *et al.*, 2019). The asfarvirus markers MCP, A32, D5, PolB, RNR and VLTF3 showed homology to adintovirus sequences and asfarvirus markers RNAPS, RNAPL and mRNAc did not. It remains to be investigated whether the markers, particularly the MCPs, are orthologs or if the overlap is a result of random sequence similarity. Nevertheless, this homology might be the reason for the adintovirus hits in the asfarvirus screen.

The high number of adintoviruses in mollusc sequencing data prompted a screening of the NCBI non-redundant protein database utilizing PSI-BLAST, which is a method for detecting distantly related proteins. The query protein in a PSI-BLAST pilot screen was adenain from *Branchiostoma lancelet adintovirus*, which is a hallmark gene of the adintoviruses, and after the fourth PSI-BLAST iteration almost 1,475 closely and distantly related proteins were obtained. Next, the taxonomy of the adintovirus-related proteins was assigned and the Sankey diagram showed that the largest fraction originate from *Eukaryota* (Figure 23). Among the eukaryotes, the *Arthropoda* are abundant with insects and many spiders (*Aranea*), especially from the genus *Trichonephila*. Besides that, there are many mollusc hits, which is expected after finding many adintovirus sequences in the molluscan raw sequencing data.

In summary, many of these organisms may be potential new hosts and some are already known to be associated with adintoviruses, for example, nematodes, members of the *Rhabditidae*, proving the reliability of this method (Starrett *et al.*, 2021; Welch *et al.*, 2019). Interestingly, a significant number of hits account for viruses and as can be anticipated, there are adintoviruses, along with numerous closely related viruses from the *Adenoviridae* family. There are even further distant viruses from the *Asfuvirales*, and taken together with the HMMsearch results (see above), these findings may explain why adintoviruses appear in the asfarvirus screen.

In conclusion, further investigation is required, to determine the number of new virus sequences, if they come from *bona fide*, from endogenized or proviruses and if the integrated sequences are still infectious or not. However, these results highlight that PSI-BLAST can be utilized for the detection of distantly related novel virus sequences in protein databases and to determine virus host association after virus-host gene transfer.

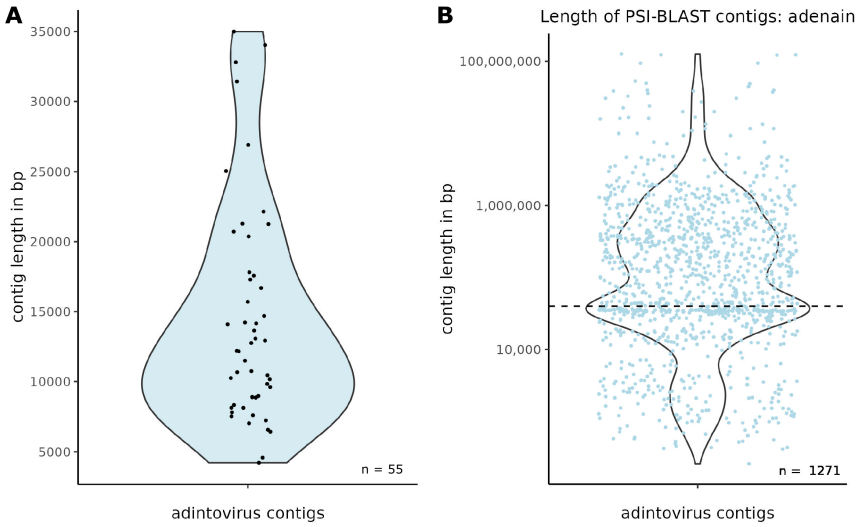

Figure 22: **A:** Contig sizes of 55 adintovirus sequences detected after screening unprocessed DNA sequencing data of molluscs. The largest proportion of contigs were around 10 kb and the largest being around 35 kb. **B**: The sizes of the 1,271 adintovirus nucleotide contigs detected with *Branchiostoma lancelet adintovirus* adenain PSI-BLAST in the NCBI non-redundant protein database after the fourth iteration. The horizontal dashed line indicates the 40 kb adintovirus genome size cut-off, where many contigs accumulate. Above that line it is likely that the adintovirus genes are integrated into a host genome, either endogenized or as provirus.

Now, for all PSI-BLAST protein hits the underlying nucleotide sequences were fetched and the sizes of these contigs have a large variation as they range from hundreds to billions of base pairs (Figure 22, part B). From 1,475 PSI-BLAST hits only 1,271 nucleotide sequences could be fetched. These sequences may have varying origins, as contigs of less than approximately 40 kb may represent *bona fide* viruses while contigs above this size likely originate from adintoviruses that have been integrated into the host genome as endogenized or provirus (Figure 22, part B, dashed line indicates size cut-off). Interestingly, the largest proportion of virus contigs accumulate below this size cut-off and may represent complete adintovirus genomes. One example of a large 1.6 Mbp adintovirus contig (from Figure 22) was selected and will be analyzed in more detail (Figure 24, part B).

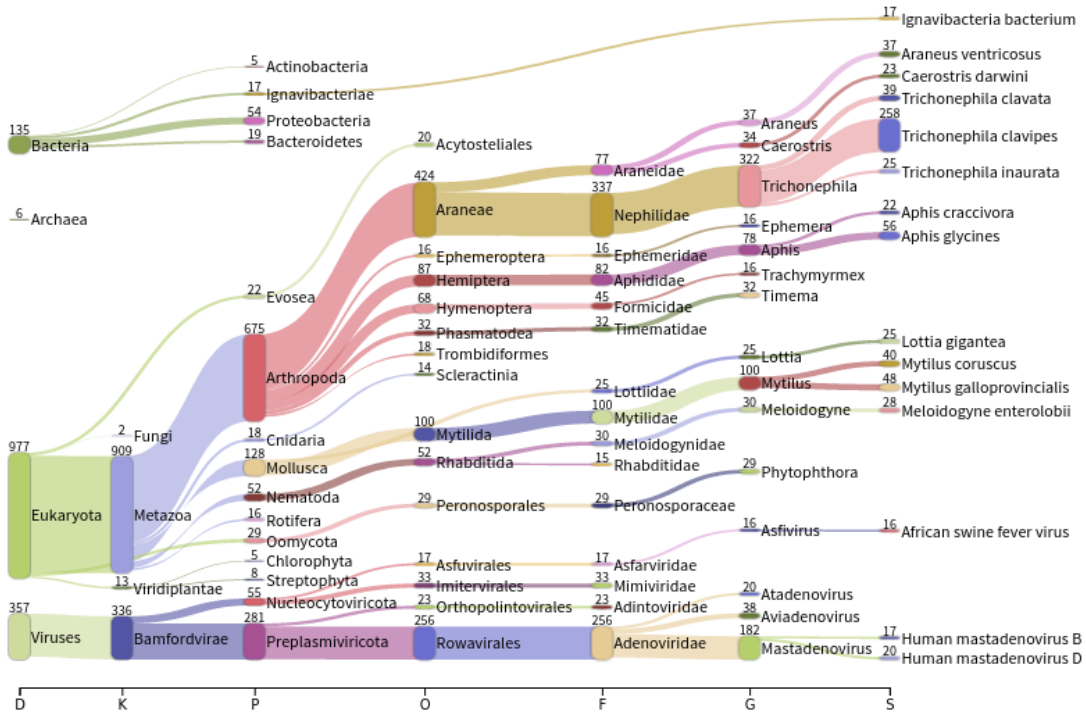

Figure 23: Taxonomy of 1,475 contigs found after *Branchiostoma lancelet adintovirus* adenain PSI-BLAST. The x-axis shows the taxonomic levels: superkingdom - D, kingdom - K, phylum - P, order - O, family - F, genus - G and species - S. Within the eukaryotes there are many hits from *Arthropoda*, especially spiders (*Aranea*), *Mollusca* and *Nematoda*. Some of the organisms are already known adintovirus hosts and many might be novel. As expected, also many adinto- and adenovirus hits were detected. That there are also asfarvirus hits might explain the adintoviruses appearing in the asfarvirus screen.

Because, an authentic virus genome encodes more than one adintovirus marker, PSI-BLAST searches were carried out also with other markers: polymerase B, retrovirus-like integrase, hexon, penton, ftsk and gasdermin. After PSI-BLAST and fetching the nucleotide sequences, 824 contigs that have at least four out of seven adintovirus markers were obtained.

Further investigation revealed that there are genome contigs with multiple locations of adintovirus genes. From 824 genome contigs, 578 were found with only one single cluster of adintovirus markers and 246 genome contigs showed multiple locations of adintovirus clusters (Figure 24). The minimum distance between two clusters

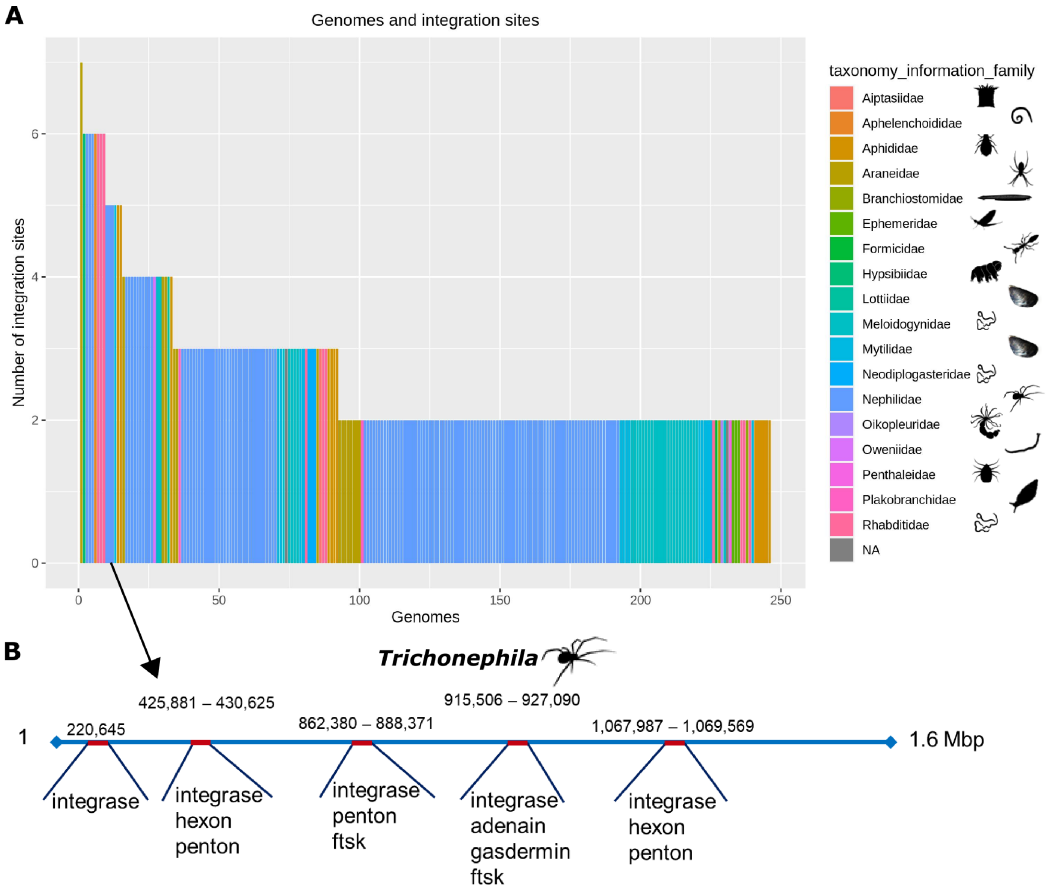

Figure 24: **A**: The numbers of integration sites of adintovirus genes are shown for 246 genome contigs with two or more sites. The colors represent the host family and it can be seen that members of the *Nephilidae* represent the largest group of adintovirus hosts with any number of integration sites. One member of the *Araneidae* had up to seven integration sites. **B**: One example of a 1.6 Mbp genome contig from *Trichonephila clavipes*, which includes five adintovirus gene clusters. Information on the markers and base position of the integration sites is also provided. The sites differ considering the marker composition and the sizes.

was 30 kb. With seven different integration sites on one genome contig, a member of the *Araneidae* family showed the highest number of integration sites. Moreover, many overrepresented *Nephilidae* members (Figure 23) are also appearing again and range from two to six adintovirus integration sites (Figure 24, part A, blue). One example of a *Trichonephila clavipes* genome contig with five adintovirus integration sites was selected and it can be seen that the sites are not identical in terms of

marker composition and adintovirus gene cluster sizes (Figure 24, part B). According to Starrett *et al.*, 2021, adintovirus integrases are believed to be highly site-specific. However, the numerous integration sites with varying virus gene patterns and sizes imply a more random mechanism of integration. Additionally, the varying number of integration sites within the same species (Figure 24, part A), suggests that other factors must be involved, as well. Finally, the time after integration and subsequent convergent evolution might also play an important role.

The results of this section lead to the conclusion that PSI-BLAST can be used as an effective tool for detecting distantly related viruses, as demonstrated here for adintoviruses. Not only, did it identify previously known adintovirus hosts, but it also revealed novel adintovirus genomes and various species that were not yet associated with adintoviruses. Finally, this results might also hint how asfarviruses are tangled up with other members from the *Nucleocytoviricota* or even the *Bamfordvirae* as in the case of the adintoviruses. This chapter raises a number of questions that require further investigation and should be followed up next.

4 Conclusions and outlook

In the previous section 3 many of my results have already been discussed. Here, after a brief overview of my main findings, I will try to draw a broader picture of my results, place them in the context of current scientific knowledge, and give some concluding remarks. This will help to answer remaining research questions and to design future projects on data-driven virus discovery of large DNA viruses.

The aim of this study was to identify previously unknown asfuviruses from unprocessed sequencing data, which will help to improve our understanding of this group of viruses and by elucidating the origin of ASFV, assess the risk of the emergence of other severe ASFV-like viruses. That ASFV has relatives in restricted virus hosts such as molluscs and protists has only recently been discovered. Now, my results suggest that there are several other asfarviruses with molluscan hosts and one complete asfarvirus genome was discovered in the shell-less marine mollusc *Elysia marginata*. Furthermore, a newly discovered virus closely related to AbALV was found to infect another abalone species, *Haliotis diversicolor*, and the viruses' marker genes share approximately 60 % sequence similarity. Moreover, I detected asfuviral hallmark genes in sequencing data of other mammals, namely horse and cow, but the authenticity of an origin of the sequences from *Asfuvirales*-like viruses infecting these mammalian hosts, remains to be verified. With the discovery of markers of thirteen novel asfuviruses, protists were the largest host group. A second MCP-like protein was found in the genome of EMAV, AbALV and among the novel asfarvirus markers, and protein sequence and structure analysis revealed that they form a sister lineage to the canonical MCPs of the *Asfuvirales*, possibly derived from an ancient

gain event in a common ancestor. However, many of my findings highlight that there is still a lack of knowledge regarding this group of viruses and a greater diversity of asfuviruses is very likely. In the future, it will be crucial to elucidate the pathogenic and zoonotic potential of the novel asfuvirus candidates and to assess the risk of virus spill-over.

4.1 Difficulties in assembling asfarviruses

During my search for asfarviruses in unprocessed sequencing data, I noticed that in molluscan and mammalian sequencing projects, asfuvirus sequences are more prevalent in transcriptomics datasets than in WGS datasets, suggesting that viral transcripts may be more abundant than viral genomic data (Figure 7). This does not include the protist projects, as many of these are derived from metagenomic datasets, which are likely to be carried out at higher sequencing depths. Considering the principle of viral infection in a host cell and the dogma of translation, the number of viral transcripts is significantly higher than the number of viral genomes, resulting in improved detectability when sequenced (Li *et al.*, 2022). However, in many cases, the asfarvirus markers are close to the limit of sequence detection resulting in low read coverage, quality and length of the virus contigs (Figure 8). As learned from EMAV, successful sequencing and assembly of a full virus genome requires a high number of genome copies, and capturing repetitive, redundant, and tandem regions is particularly challenging. Despite six WGS sequencing runs on the single *ElyMa* sample and a very high sequencing coverage of up to 200, the automatic computational assembly and recovery of the complete viral genome failed. In the

end, the genome was only completed by manual assembly of the two remaining virus contigs. Very similar results were obtained in benchmarking experiments performed by Schulz *et al.*, 2020a using giant viruses spiked into genomic data, and only high viral genome copy numbers, yielding a read coverage of 165, resulted in sufficient assembly quality and apparently long reads have an advantage over short. Also, difficulties in sequencing of the ASFV are known and especially that read coverage declines towards the terminal parts of the genome where most of the redundant regions are (for example: Forth *et al.*, 2019). In summary, genome sequencing and assembly of full-length and complete asfuviruses is challenging even when sequencing coverage is moderately high. This should be taken into account in the design of future studies, as it has a significant impact on the success of data-driven large DNA virus discovery.

4.2 High complexity of large DNA viruses

The discovery of giant viruses in 2003 and subsequent years challenged previous perceptions, as viruses were found to be similar in size and genome length to the smallest and simplest cellular organisms (Raoult *et al.*, 2004; Philippe *et al.*, 2013). Another milestone was the discovery of virophages, which are viruses that cannot infect a host cell independently but rely on a giant virus co-infecting the protist and exploiting the giant virus virion factories (La Scola *et al.*, 2008). The *Nucleocytoviricota* exhibit vast diversity, with prototype viruses defying many common classification features. An example for this is the *Pandoravirus*, which lost the characteristic DJR MCP and gained a major virion protein that evolved independently

from a bacterial enzyme, likely acquired through horizontal gene transfer (HGT) (Krupovic *et al.*, 2020). Over the last decade, significant progress has been made in characterising giant viruses and *Nucleocytoviricota*, resulting in an increased understanding of the true virus complexity and this process still continues (for example: Yutin and Koonin, 2012; Brandes and Linial, 2019; Gaïa and Forterre, 2023). Also in this study, it was found that the knowledge of asfarviruses and *Nucleocytoviricota* is still limited, with results often difficult to classify and draw conclusions from due to the lack of context. For instance, when comparing closely related asfarviruses like AbALV, EMAV and ASFV, their genomes appear to differ significantly in genome architecture and structure and also protein sequence similarity (Figure 14 and Figure 12). Moreover, a significant proportion of virus proteins encoded by asfarvirus genomes have unknown functions and are dissimilar to those found in large-scale protein databases (Figure 15). Finally, this suggests the importance to conduct further research to accurately compare and interpret results and today's findings will provide the context that is required for future interpretations and conclusions.

4.3 Concluding remarks on *Elysia marginata asfarvirus*

One key finding of this work is the discovery of the full-length asfarvirus genome EMAV, from the marine shell-less mollusc, *ElyMa*, that is the second molluscan asfarvirus after AbALV. Based on my findings, the *Asfarviridae* subtree is divided into three clades, showcased by ASFV, AbALV and EMAV, respectively (Figure 18). This provides evidence for a completely new virus lineage within the *Asfarviridae*

that is prototyped by EMAV. Moreover, the EMAV clade has a sister clade composed of sequences from mammalian, molluscan and protist asfuviruses, which proposes viruses of various hosts within the same branch. This intersection of host groups within one virus clade could imply a high likelihood of viruses jumping hosts, for example from a protist to a distant host such as a pig. Here, it is important to note that even *Elysia* and *Haliotis*, the hosts of EMAV and AbALV respectively, are not very closely related. Both belong to the *Gastropoda* class, but do not share a more basal taxonomic relationship. The hypothesis of viruses jumping hosts will be further discussed in the following section 4.4. Furthermore, it would be very informative to use a Bayesian approach similar to Michaud *et al.*, 2013 to estimate the age of the most recent viral ancestors of EMAV and AbALV, and to assess which virus evolved first. Finally, it remains unclear whether EMAV can cause a high mortality disease in *ElyMa* sea slugs similar to AbALV in abalones or ASFV in pigs (Matsuyama *et al.*, 2020; Li *et al.*, 2022). My *in-silico* analysis of EMAV led to important conclusions, but to uncover the epidemiological, pathogenic and zoonotic potential of EMAV, it is necessary to conduct wet-lab molecular virology experiments such as infection assays, cell culture, biochemical assays, and imaging.

4.4 Unicellular hosts as potential source of ASFV

Protists appeared to be overrepresented as hosts given the novel and known asfuviruses and I suggest that unicellular eukaryotes still hide a large diversity of *Asfuvirales*. In this sense, and based on the intertwined phylogenetic position of the protist asfuviruses in the *Asfarviridae* clade (Figure 18), one hypothesis is that an ASFV-like

virus could have spilled from a protist to an intermediate for example molluscan or mammalian host. Additionally, ASFV has a very high evolutionary rate of around 1×10^{-5} substitutions per site per year and this is a rate similar to that of some RNA viruses, which are by nature more prone to modification than DNA viruses (Michaud *et al.*, 2013). For example, the evolutionary rate of vertebrate herpesviruses is with 1×10^{-9} substitutions per site per year several orders of magnitude lower than that of ASFV (Michaud *et al.*, 2013). This could imply the ability of rapid interspecies host adaptations of an ancient ASFV-like virus. Furthermore, the assumption that ASFV shared a long-term coevolution with its host is contradicted by the fact that the most recent common ancestor is only 300 years old, which rather suggests an exceptional spill-over event that occurred around that time (Michaud *et al.*, 2013). Another possibility is that ticks served as natural reservoir for ASFV-like viruses, as the virus can replicate and be transmitted during blood feeding from ticks to swine (for example: Forth *et al.*, 2020). However, the asfarvirus screen carried out as part of this work found no evidence of ASFV-like viruses in ticks, but highly supports the conception that asfarviruses originate from or have a protist host reservoir. Lastly and considering both theories, ASFV-like viruses may have originated from an interaction between both, ticks and protists, which subsequently paved the ground for an ancient spill-over to *Mammalia*. To uncover missing evidence for either hypotheses, it will be crucial to conduct further screening of sequencing data from a variety of hosts but especially protists, molluscs and ticks.

4.5 MCP diversity of *Nucleocytoviricota*

Due to the discovery of a second MCP variant in five asfarviruses, including the full viral genomes EMAV and AbALV, it was of particular relevance to analyse the number of MCPs in other members of the *Nucleocytoviricota*. It is already known that *Nucleocytoviricota* can express multiple marker genes on one virus genome and that gene duplication may even be a major driver of virus genome gigantism (Koonin and Yutin, 2019; Machado *et al.*, 2023). Nevertheless, I performed preliminary HMM-based MCP searches in *Nucleocytoviricota* reference viruses and many genomes from *Mimiviridae* and *Phycodnaviridae* code multiple MCPs and it can be concluded that the variants originate from at least three to four distinct groups that are consistently present in multiple representatives. Furthermore, in the reference virus genomes that I examined, there were candidates encoding up to eight different MCPs and Schulz *et al.*, 2020b reported that some *Nucleocytoviricota* even have as many as 20 variants. At first glance, the preliminary phylogenetic tree of all identified *Nucleocytoviricota* MCPs revealed that the evolutionary trajectories of many MCPs can be traced, allowing for conclusions for example about duplication events. Examining the protein sequences and structures of the different *Nucleocytoviricota* MCPs can provide valuable insights and should be a focus of future research.

4.6 Outlook

In summary, this text outlines how data-driven virology can help to discover new viruses, including full-length virus genomes by deep-mining unprocessed sequencing data repositories or protein databases. Given the ongoing growth of sequencing data at explosive pace, it is crucial to prioritize regular screening efforts to provide necessary additions and context for asfuvirus classification. Future discoveries may provide evidence for either hypothesis on the origin of ASFV and the likelihood of the emergence of similar viruses. Moreover, the screening method that was streamlined for the search for novel members of the *Asfuvirales* can now be adapted to screen for other large DNA viruses in the future. For example, *Herpesviridae* are a vast group of viruses, which can have a significant impact on the health of both animals and humans as infection is often severe, persistent and lifelong (Knipe and Howley, 2013). Surveying wildlife data for unknown herpesviruses and investigating their spread or if they jump hosts could be highly rewarding, as there are also zoonotic infections in humans (Tischer and Osterrieder, 2010; Hu *et al.*, 2022). Subsequently, screening data of patients who suffer from symptoms of unknown origin could also be a promising approach to reveal virus-disease associations that were previously unknown. However, there are many other interesting groups of large DNA viruses that lack fundamental understanding because of missing diversity and context, making them worth to further investigate.

Bibliography

Alonso, C., M. Borca, L. Dixon, Y. Revilla, F. Rodriguez, and J. M. Escribano. (2018). ICTV virus taxonomy profile: Asfarviridae. *Journal of General Virology*. 99 613–614.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. (1990). Basic local alignment search tool. *Journal of Molecular Biology*. 215 403–410.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*. 25 3389–3402.

Andreani, J., J. Y. B. Khalil, M. Sevvana, S. Benamar, F. Di Pinto, I. Bitam, P. Colson, T. Klose, M. G. Rossmann, D. Raoult, and B. La Scola. (2017). Pacmanvirus, a New Giant Icosahedral Virus at the Crossroads between Asfarviridae and Faustoviruses. *Journal of Virology*. 91 1–11.

Andrés, G., D. Charro, T. Matamoros, R. S. Dillard, and N. G. Abrescia. (2020). The cryo-EM structure of African swine fever virus unravels a unique architecture comprising two icosahedral protein capsids and two lipoprotein membranes. *Journal of Biological Chemistry*. 295 1–12.

Aylward, F. O., M. Moniruzzaman, A. D. Ha, and E. V. Koonin. (2021). A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLOS Biology*. 19 1–18.

Bajrai, L. H., S. Benamar, E. I. Azhar, C. Robert, A. Levasseur, D. Raoult, and B. La Scola. (2016). Kaumoebavirus, a new virus that clusters with Faustoviruses and Asfarviridae. *Viruses*. 8 1–10.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*. 19 455–477.

Bellas, C., T. Hackl, M. S. Plakolb, A. Koslová, M. G. Fischer, and R. Sommaruga. (2023). Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses. *Proceedings of the National Academy of Sciences of the United States of America*. 120 1–10.

Bolger, A. M., M. Lohse, and B. Usadel. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 30 2114–2120.

Boratto, P. V. M., G. P. Oliveira, T. B. Machado, A. C. S. P. Andrade, J.-P. Baudoin, T. Klose, F. Schulz, S. Azza, P. Decloquement, E. Chabrière, P. Colson, A. Levasseur, B. La Scola, and J. S. Abrahão. (2020). Yaravirus: A novel 80-nm virus infecting *Acanthamoeba castellanii*. *Proceedings of the National Academy of Sciences*. 1–8.

Bouckaert, R., T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, M. Matschiner, F. K. Mendes, N. F. Müller, H. A. Ogilvie, L. Du Plessis, A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C. H. Wu, D. Xie, C. Zhang, T. Stadler, and A. J. Drummond. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*. 15 1–28.

Brandes, N. and M. Linial. (2019) Giant viruses-big surprises. *Viruses*. 11 1–12.

Breitwieser, F. P. and S. L. Salzberg. (2020). Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics*. 36 1303–1304.

Brennan, G., A. M. M. Stoian, H. Yu, M. J. Rahman, S. Banerjee, J. N. Stroup, C. Park, L. Tazi, and S. Rothenburg. (2022). Molecular Mechanisms of Poxvirus Evolution. *mBio*. 14 1–13.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*. 10 1–9.

Cantalapiedra, C. P., A. Hernandez-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*. 38 5825–5829.

Capella-Gutiérrez, S., J. M. Silla-Martínez, and T. Gabaldón. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 25 1972–1973.

Carrascosa, J. L., J. M. Carazo, A. L. Carrascosa, N. García, A. Santisteban, and E. Viñuela. (1984). General morphology and capsid fine structure of African swine fever virus particles. *Virology*. 132 160–172.

Chastagner, A., R. P. de Oliveira, E. Hutet, M. L. Dimna, F. Paboeuf, P. Lucas, Y. Blanchard, L. Dixon, L. Vial, and M.-F. L. Potier. (2020). Coding-Complete Genome Sequence of an African Swine Fever Virus Strain Liv13/33 Isolate from Experimental Transmission between Pigs and Ornithodoros moubata Ticks. *Microbiology Resource Announcements*. 9 1–3.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*. 10 1–4.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*. 14 755–763.

Edgar, R. C., J. Taylor, V. Lin, T. Altman, P. Barbera, D. Meleshko, D. Lohr, G. Novakovsky, B. Buchfink, B. Al-Shayeb, J. F. Banfield, M. de la Peña, A. Korobeynikov, R. Chikhi, and A. Babaian. (2022). Petabase-scale sequence alignment catalyses viral discovery. *Nature*. 1–6.

Epifano, C., J. Krijnse-Locker, M. L. Salas, J. Salas, and J. M. Rodríguez. (2006). Generation of Filamentous Instead of Icosahedral Particles by Repression of African Swine Fever Virus Structural Protein pB438L. *Journal of Virology*. 80 11456–11466.

Erdozain, S., E. Barrionuevo, L. Ripoll, P. Mier, and M. A. Andrade-Navarro. (2023). Protein repeats evolve and emerge in giant viruses. *Journal of Structural Biology*. 215 1–11.

Eustace Montgomery, R. (1921). On A Form of Swine Fever Occurring in British East Africa (Kenya Colony). *Journal of Comparative Pathology and Therapeutics*. 34 159–191.

Finkel, Y., N. Stern-Ginossar, and M. Schwartz. (2018). Viral Short ORFs and Their Possible Functions. *Proteomics*. 18 1–8.

Forth, J. H., L. F. Forth, J. King, O. Groza, A. Hübner, A. S. Olesen, D. Höper, L. K. Dixon, C. L. Netherton, T. B. Rasmussen, S. Blome, A. Pohlmann, and M. Beer. (2019). A Deep-Sequencing Workflow for the Fast and Efficient Generation of High-Quality African Swine Fever Virus Whole-Genome Sequences. *Viruses*. 11 1–18.

Forth, J. H., L. F. Forth, S. Lycett, L. Bell-Sakyi, G. M. Keil, S. Blome, S. Calvignac-Spencer, A. Wissgott, J. Krause, D. Höper, H. Kampen, and M. Beer. (2020). Identification of African swine fever virus-like elements in the soft tick genome provides insights into the virus' evolution. *BMC Biology*. 18 1–18.

Gaïa, M. and P. Forterre. (2023). From Mimivirus to Mirusvirus: The Quest for Hidden Giants. *Viruses*. 15 1–11.

García-Escudero, R., G. Andrés, F. Almazán, and E. Viñuela. (1998). Inducible gene expression from African swine fever virus recombinants: analysis of the major capsid protein p72. *Journal of virology*. 72 3185–3195.

Gaudreault, N. N., D. W. Madden, W. C. Wilson, J. D. Trujillo, and J. A. Richt. (2020). African Swine Fever Virus: An Emerging DNA Arbovirus. *Frontiers in Veterinary Science*. 7 1–17.

Geballa-Koukoulas, K., J. Andreani, B. La Scola, and G. Blanc. (2021). The Kaumoebavirus LCC10 Genome Reveals a Unique Gene Strand Bias among "Extended Asfarviridae". *Viruses*. 13 1–13.

Geballa-Koukoulas, K., H. Boudjemaa, J. Andreani, B. L. Scola, and G. Blanc. (2020). Comparative Genomics Unveils Regionalized Evolution of the Faustovirus Genomes. *Viruses*. 12 1–16.

Gifford-Gonzalez, D. and O. Hanotte. (2011). Domesticating Animals in Africa: Implications of Genetic and Archaeological Findings. *Journal of World Prehistory*. 24 1–23.

Gregory, A. C., A. A. Zayed, N. Conceição-Neto, B. Temperton, B. Bolduc, A. Alberti, M. Ardyna, K. Arkhipova, M. Carmichael, C. Cruaud, C. Dimier, G. Domínguez-Huerta, J. Ferland, S. Kandels, Y. Liu, C. Marec, S. Pesant, M. Picheral, S. Pisarev, J. Poulain, J. É. Tremblay, D. Vik, S. G. Acinas, M. Babin, P. Bork, E. Boss,

C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, A. I. Culley, B. E. Dutilh, and S. Roux. (2019). Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*. 177 1109–1123.

Guglielmini, J., A. C. Woo, M. Krupovic, P. Forterre, and M. Gaia. (2019). Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proceedings of the National Academy of Sciences*. 116 19585–19592.

Hackl, T., M. J. Ankenbrand, and B. van Adrichem. (2023). gggenomes: A Grammar of Graphics for Comparative Genomics. *R package version 0.9.12.9000*.

Hadley, W. (2016) ggplot2: Elegant Graphics for Data Analysis. *R package*.

Hannat, S., B. La Scola, J. Andreani, and S. Aherfi. (2023). Asfarviruses and Closely Related Giant Viruses. *Viruses*. 15 1015.

Holm, L., A. Laiho, P. Törönen, and M. Salgado. (2023). DALI shines a light on remote homologs: One hundred discoveries. *Protein Science*. 32 1–18.

Hu, G., H. Du, Y. Liu, G. Wu, and J. Han. (2022). Herpes B virus: History, zoonotic potential, and public health implications. *Biosafety and Health*. 4 213–219.

Huang, X. (1999). CAP3: A DNA Sequence Assembly Program. *Genome Research*. 9 868–877.

Hyun, J., H. Matsunami, T. G. Kim, and M. Wolf. (2022). Assembly mechanism of the pleomorphic immature poxvirus scaffold. *Nature Communications*. 13 1–10.

Iyer, L. M., L. Aravind, and E. V. Koonin. (2001). Common Origin of Four Diverse Families of Large Eukaryotic DNA Viruses. *Journal of Virology*. 75 11720–11734.

Iyer, L. M., S. Balaji, E. V. Koonin, and L. Aravind. (2006). Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Research*. 156–184.

Kao, S., C.-F. Kao, W. Chang, and C. Ku. (2023). Widespread Distribution and Evolution of Poxviral Entry-Fusion Complex Proteins in Giant Viruses. *Microbiology Spectrum*. 11 1–158.

Karger, A., D. Pérez-Núñez, J. Urquiza, P. Hinojar, C. Alonso, F. B. Freitas, Y. Revilla, M. F. Le Potier, and M. Montoya. (2019). An Update on African Swine Fever Virology. *Viruses*. 11 1–14.

Katoh, K., K. Misawa, K. I. Kuma, and T. Miyata. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*. 30 3059–3066.

Kelley, L. A., S. Mezulis, C. M. Yates, M. N. Wass, and M. J. Sternberg. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*. 10 845–858.

Knipe, D. M. and P. Howley. (2013). *Fields virology: Sixth edition*: Wolters Kluwer Health Adis (ESP).

Koonin, E. V., V. V. Dolja, M. Krupovic, A. Varsani, Y. I. Wolf, N. Yutin, F. M. Zerbini, and J. H. Kuhn. (2020). Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews*. 84 1–33.

Koonin, E. V., V. V. Dolja, and M. Krupovic. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology*. 479-480 2–25.

Koonin, E. V. and M. Krupovic. (2017). Polintons, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. *Current opinion in virology*. 25 7–15.

Koonin, E. V., M. Krupovic, and N. Yutin. (2015). Evolution of double-stranded DNA viruses of eukaryotes: From bacteriophages to transposons to giant viruses. *Annals of the New York Academy of Sciences*. 1341 10–24.

Koonin, E. V. and N. Yutin. (2018). Multiple evolutionary origins of giant viruses. *F1000Research. NLM (Medline)*. 7 1–12.

Koonin, E. V. and N. Yutin. (2019) Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. *Advances in Virus Research*. 103 167–202.

Krupovic, M. and D. H. Bamford. (2008). Virus evolution: how far does the double β-barrel viral lineage extend? *Nature Reviews Microbiology*. 6 941–948.

Krupovic, M., K. S. Makarova, and E. V. Koonin. (2022). Cellular homologs of the double jelly-roll major capsid proteins clarify the origins of an ancient virus kingdom. *Proceedings of the National Academy of Sciences of the United States of America*. 119 1–10.

Krupovic, M., N. Yutin, and E. Koonin. (2020). Evolution of a major virion protein of the giant pandoraviruses from an inactivated bacterial glycoside hydrolase. *Virus Evolution*. 6 1–8.

La Scola, B., C. Desnues, I. Pagnier, C. Robert, L. Barrassi, G. Fournous, M. Merchat, M. Suzan-Monti, P. Forterre, E. Koonin, and D. Raoult. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature*. 455 100–104.

Langmead, B. and S. L. Salzberg. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 9 357–359.

Laslett, D. and B. Canback. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*. 32 11–16.

Lauber, C., M. Seifert, R. Bartenschlager, and S. Seitz. (2019). Discovery of highly divergent lineages of plant-associated astro-like viruses sheds light on the emergence of potyviruses. *Virus Research*. 260 38–48.

Lauber, C. and S. Seitz. (2022). Opportunities and Challenges of Data-Driven Virus Discovery. *Biomolecules*. 12 1–10.

Lauber, C., S. Seitz, S. Mattei, A. Suh, J. Beck, J. Herstein, J. Börold, W. Salzburger, L. Kaderali, J. A. Briggs, and R. Bartenschlager. (2017). Deciphering the Origin and Evolution of Hepatitis B Viruses by Means of a Family of Non-enveloped Fish Viruses. *Cell Host and Microbe*. 22 387–399.

Lauber, C., J. Vaas, F. Klingler, P. Mutz, A. E. Gorbalenya, R. Bartenschlager, and S. Seitz. (2021). Deep mining of the Sequence Read Archive reveals bipartite coronavirus genomes and inter-family Spike glycoprotein recombination. *bioRxiv*. 1–37.

Lechner, M., S. Findeiß, L. Steiner, M. Marz, P. F. Stadler, and S. J. Prohaska. (2011). Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics*. 12 1–9.

Leinonen, R., H. Sugawara, and M. Shumway. (2011). The Sequence Read Archive. *Nucleic Acids Research*. 39 D19–D21.

Letunic, I. and P. Bork. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*. 49 W293–W296.

Li, D., C. M. Liu, R. Luo, K. Sadakane, and T. W. Lam. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 31 1674–1676.

Li, Z., W. Chen, Z. Qiu, Y. Li, J. Fan, K. Wu, X. Li, M. Zhao, H. Ding, S. Fan, and J. Chen. (2022). African Swine Fever Virus: A Review. *Life*. 12 1–43.

Louazani, A. C., E. Baptiste, A. Levasseur, P. Colson, and B. La Scola. (2018). Faustovirus E12 Transcriptome Analysis Reveals Complex Splicing in Capsid Gene. *Frontiers in microbiology*. 9 1–10.

Lu, C. and Y. Peng. (2021). Computational Viromics: Applications of the Computational Biology in Viromics Studies. *Virologica Sinica*. 36 1256–1260.

Machado, T. B., A. C. R. Picorelli, B. L. de Azevedo, I. L. M. de Aquino, V. F. Queiroz, R. A. L. Rodrigues, J. João Pessoa Araújo, L. S. Ullmann, T. M. dos Santos, R. E. Marques, S. L. Guimarães, A. C. S. P. Andrade, J. S. Gularte, M. Demoliner, M. Filippi, V. M. A. G. Pereira, F. R. Spilki, M. Krupovic, F. O. Aylward, L.-E. Del-Bem, and J. S. Abrahão. (2023). Gene duplication as a major force driving the genome expansion in some giant viruses. *Journal of Virology*. 97 1–16.

Matsuyama, T., I. Kiryu, M. Inada, T. Takano, Y. Matsuura, and T. Kamaishi. (2021). Susceptibility of four abalone species, haliotis gigantea, haliotis discus discus, haliotis discus hannai and haliotis diversicolor, to abalone asfa-like virus. *Viruses*. 13 1–16.

Matsuyama, T., I. Kiryu, T. Mekata, T. Takano, K. Umeda, and Y. Matsuura. (2023). Pathogenicity, genomic analysis and structure of abalone asfa-like virus: evidence for classification in the family Asfarviridae. *The Journal of general virology*. 104 1–10.

Matsuyama, T., T. Takano, I. Nishiki, A. Fujiwara, I. Kiryu, M. Inada, T. Sakai, S. Terashima, Y. Matsuura, K. Isowa, and C. Nakayasu. (2020). A novel Asfarvirus-like virus identified as a potential cause of mass mortality of abalone. *Scientific Reports*. 10 1–12.

Meng, E. C., T. D. Goddard, E. F. Pettersen, G. S. Couch, Z. J. Pearson, J. H. Morris, and T. E. Ferrin. (2023). UCSF ChimeraX: Tools for Structure Building and Analysis. *Protein Science*. 1–13.

Michaud, V., T. Randriamparany, and E. Albina. (2013). Comprehensive Phylogenetic Reconstructions of African Swine Fever Virus: Proposal for a New Classification and Molecular Dating of the Virus. *PLoS ONE*. 8 1–14.

Mirdita, M., K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*. 19 679–682.

Moniruzzaman, M. and F. O. Aylward. (2023). Endogenous DNA viruses take center stage in eukaryotic genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*. 120 1–8.

Moniruzzaman, M., C. A. Martinez-Gutierrez, A. R. Weinheimer, and F. O. Aylward. (2020). Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nature Communications*. 11 1–11.

Mönttinen, H. A., C. Bicep, T. A. Williams, and R. P. Hirt. (2021). The genomes of nucleocytoplasmic large DNA viruses: viral evolution writ large. *Microbial Genomics*. 7 1–17.

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R.

Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*. 44 D733–D745.

ODCF DKFZ, Heidelberg. (2020) Compute Cluster and Developer Services.

Oura, C. A., P. P. Powell, E. Anderson, and R. M. Parkhouse. (1998). The pathogenesis of African swine fever in the resistant bushpig. *Journal of General Virology*. 79 1439–1443.

Philippe, N., M. Legendre, G. Doutre, Y. Couté, O. Poirot, M. Lescot, D. Arslan, V. Seltzer, L. Bertaux, C. Bruley, J. Garin, J. M. Claverie, and C. Abergel. (2013). Pandoraviruses: Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*. 341 281–286.

Pringle, C. R. (1998). Virology Division News Virus Taxonomy-San Diego 1998. *VDN Virology Division News Arch Virol*. 143 1–12.

Raoult, D., S. Audic, C. Robert, C. Abergel, P. Renesto, H. Ogata, B. La Scola, M. Suzan, and J. M. Claverie. (2004). The 1.2-megabase genome sequence of Mimivirus. *Science*. 306 1344–1350.

Reteno, D. G., S. Benamar, J. B. Khalil, J. Andreani, N. Armstrong, T. Klose, M. Rossmann, P. Colson, D. Raoult, and B. L. Scola. (2015). Faustovirus, an Asfarvirus-Related New Lineage of Giant Viruses Infecting Amoebae. *Journal of Virology*. 89 6585–6594.

Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. (2011). Integrative genomics viewer. *Nature Biotechnology*. 29 24–26.

Rodríguez, J. M., L. T. Moreno, A. Alejo, A. Lacasta, F. Rodríguez, and M. L. Salas. (2015). Genome Sequence of African Swine Fever Virus BA71, the Virulent Parental Strain of the Nonpathogenic and Tissue-Culture Adapted BA71V. *PLOS ONE*. 10 1–22.

Roux, S., E. M. Adriaenssens, B. E. Dutilh, E. V. Koonin, A. M. Kropinski, M. Krupovic, J. H. Kuhn, R. Lavigne, J. R. Brister, A. Varsani, C. Amid, R. K. Aziz, S. R. Bordenstein, P. Bork, M. Breitbart, G. R. Cochrane, R. A. Daly, C. Desnues, M. B. Duhaime, J. B. Emerson, F. Enault, J. A. Fuhrman, P. Hingamp, P. Hugenholtz, B. L. Hurwitz, N. N. Ivanova, J. M. Labonté, K. B. Lee, R. R. Malmstrom, M. Martinez-Garcia, I. K. Mizrachi, H. Ogata, D. Páez-Espino, M. A. Petit, C. Putonti, T. Rattei, A. Reyes, F. Rodriguez-Valera, K. Rosario, L. Schriml, F. Schulz, G. F. Steward, M. B. Sullivan, S. Sunagawa, C. A. Suttle, B. Temperton, S. G. Tringe, R. V. Thurber, N. S. Webster, K. L. Whiteson, S. W. Wilhelm, K. E. Wommack, T. Woyke, K. C. Wrighton, P. Yilmaz, T. Yoshida, M. J. Young, N. Yutin, L. Z. Allen, N. C. Kyrpides, and E. A. Eloe-Fadrosh. (2019). Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature biotechnology*. 37 29–37.

Sauter-Louis, C., F. J. Conraths, C. Probst, U. Blohm, K. Schulz, J. Sehl, M. Fischer, J. H. Forth, L. Zani, K. Depner, T. C. Mettenleiter, M. Beer, and S. Blome. (2021). African Swine Fever in Wild Boar in Europe—A Review. *Viruses*. 13 1–30.

Schoch, C. L., S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. McVeigh, K. O'Neill, B. Robbertse, S. Sharma, V. Soussov, J. P. Sullivan, L. Sun, S. Turner, and I. Karsch-Mizrachi. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database: The Journal of Biological Databases and Curation*. 2020 1–21.

Schulz, F., C. Abergel, and T. Woyke. (2022). Giant virus biology and diversity in the era of genome-resolved metagenomics. *Nature Reviews Microbiology*. 1–16.

Schulz, F., J. Andreani, R. Francis, H. Boudjemaa, J. Y. Bou Khalil, J. Lee, B. La Scola, and T. Woyke. (2020). Advantages and Limits of Metagenomic Assembly and Binning of a Giant Virus. *mSystems*. 5 1–10.

Schulz, F., S. Roux, D. Paez-Espino, S. Jungbluth, D. A. Walsh, V. J. Denef, K. D. McMahon, K. T. Konstantinidis, E. A. Eloe-Fadrosh, N. C. Kyrpides, and T. Woyke. (2020). Giant virus diversity and host interactions through global metagenomics. *Nature*. 578 432–436.

Shen, W., S. Le, Y. Li, and F. Hu. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*. 11 1–11.

Ståhl, K., A. Boklund, T. Podgórski, T. Vergne, J. C. Abrahantes, A. Papanikolaou, G. Zancanaro, and L. Mur. (2023). Epidemiological analysis of African swine fever in the European Union during 2022. *EFSA Journal*. 21 1–50.

Starrett, G. J., M. J. Tisza, N. L. Welch, A. K. Belford, A. Peretti, D. V. Pastrana, and C. B. Buck. (2021). Adintoviruses: a proposed animal-tropic family of midsize eukaryotic linear dsDNA (MELD) viruses. *Virus Evolution*. 7 1–10.

Steinegger, M., M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, and J. Söding. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*. 20 1–15.

Steinegger, M. and J. Söding. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*. 35 1026–1028.

Sullivan, M. J., N. K. Petty, and S. A. Beatson. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics*. 27 1009–1010.

Sun, T.-W., C.-L. Yang, T.-T. Kao, T.-H. Wang, M.-W. Lai, and C. Ku. (2020). Host Range and Coding Potential of Eukaryotic Giant Viruses. *Viruses*. 12 1–20.

Tatusov, R. L., E. V. Koonin, and D. J. Lipman. (1997). A genomic perspective on protein families. *Science*. 278 631–637.

Tischer, B. K. and N. Osterrieder. (2010). Herpesviruses - a zoonotic threat? *Veterinary microbiology*. 140 1–8.

Tisza, M. J., D. V. Pastrana, N. L. Welch, B. Stewart, A. Peretti, G. J. Starrett, Y.-y. S. Pang, S. R. Krishnamurthy, P. A. Pesavento, D. H. McDermott, P. M. Murphy, J. L. Whited, B. Miller, J. M. Brenchley, S. P. Rosshart, B. Rehermann, J. Doorbar, B. A. Ta'ala, O. Pletnikova, J. Troncoso, B. Bolduc, S. M. Resnick, M. B. Sullivan, A. Varsani, A. M. Segall, and C. B. Buck. (2019). Discovery of several thousand highly diverse circular DNA viruses. *bioRxiv*. 9 1–26.

Wang, N., D. Zhao, J. Wang, Y. Zhang, M. Wang, Y. Gao, F. Li, J. Wang, Z. Bu, Z. Rao, and X. Wang. (2019). Architecture of African swine fever virus and implications for viral assembly. *Science*. 366 640–644.

Waterhouse, A., M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. De Beer, C. Rempfer, L. Bordoli, R. Lepore, and T. Schwede. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*. 46 W296–W303.

Weisburg, W. G., S. M. Barns, D. A. Pelletier, and D. J. Lane. (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology*. 173 697–703.

Welch, N. L., M. J. Tisza, A. Belford, D. V. Pastrana, Y.-Y. S. Pang, J. T. Schiller, P. An, P. G. Cantalupo, J. M. Pipas, S. Koda, K. Subramaniam, T. B. Waltzek, C. Bian, Q. Shi, Z. Ruan, T. Fei, F. Ng, G. J. Starrett, and C. B. Buck. (2019).

Identification of "Missing Link" Families of Small DNA Tumor Viruses. *bioRxiv*. 1–19.

World Organisation for Animal Health WOAH. (2020). Panorama 2020-1 : African swine fever: responding to the global threat. *Bulletin de l'OIE*. 2020 1–91.

Xian, Y. and C. Xiao. (2020). Current capsid assembly models of icosahedral nucleocytoviricota viruses. *Advances in virus research*. 108 1–36.

Yutin, N. and E. V. Koonin. (2012). Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. *Virology Journal*. 9 1–18.

Yutin, N., D. Raoult, and E. V. Koonin. (2013). Virophages, polintons, and transpovirons: A complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virology Journal*. 10 1–15.

Yutin, N., S. Shevchenko, V. Kapitonov, M. Krupovic, and E. V. Koonin. (2015). A novel group of diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biology*. 13 1–14.

Yutin, N., Y. I. Wolf, D. Raoult, and E. V. Koonin. (2009). Eukaryotic large nucleocytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virology Journal*. 6 1–13.

Zhao, D., E. Sun, L. Huang, L. Ding, Y. Zhu, J. Zhang, D. Shen, X. Zhang, Z. Zhang, T. Ren, W. Wang, F. Li, X. He, and Z. Bu. (2023). Highly lethal genotype I and II recombinant African swine fever viruses detected in pigs. *Nature Communications*. 14 1–10.

Zhao, H., H. Hikida, Y. Okazaki, and H. Ogata. (2023). A 1.5 Mb continuous endogenous viral element region in the arbuscular mycorrhizal fungus Rhizophagus irregularis. *bioRxiv*. 1–9.