

**Virom-Charakterisierung anhand unprozessierter lokaler und öffentlich
verfügbarer Sequenzierungsdaten**

**Virome characterization from unprocessed local and publicly available
sequencing data**

Masterarbeit
zur Erlangung des Hochschulgrades
Master of Science (M.Sc.)

im Studiengang Biomedizinische Datenwissenschaft
der Medizinischen Hochschule Hannover

vorgelegt von
Sergej Ruff
Geboren am 23.12.1997
In Kustanai, Kasachstan

Hannover, Oktober 2024

Die vorliegende Masterarbeit wurde im Zeitraum vom 08.04.2024 bis 08.10.2024 im Twincore, Zentrum für Experimentelle und Klinische Infektionsforschung GmbH angefertigt.

Erstgutachter/in: **Jun. Prof. Dr. Chris Lauber**

Leiter der Forschungsgruppe Computergestützte Virologie
TWINCORE – Zentrum für Experimentelle und
Klinische Infektionsforschung

Zweitgutachter/in: **Prof. Dr. rer. nat. Helena U. Zacharias**

Leiterin Bereich Klinische Datenwissenschaften
Peter L. Reichertz Institut für Medizinische Informatik
der TU Braunschweig und der Medizinischen
Hochschule Hannover

Abgabedatum:
07.10.2024

Zusammenfassung

Viren sind allgegenwärtig. Virales Genmaterial findet sich in verschiedenen Formen auf dem Land, im Wasser, in Lebewesen und deren Mikrobiota. Heute wissen wir, dass Viren nicht nur Krankheitserreger sind, sondern auch positive Effekte auf ihren Wirt oder die Umgebung haben können. Sie tragen zur Evolution des Wirts bei oder verbleiben als Kommensalen ohne wesentlichen Einfluss auf die Gesundheit des Wirts. Die Virosphäre ist jedoch bis heute nicht vollständig erforscht, und Schätzungen zufolge liegt die Anzahl unentdeckter Viren im sechsstelligen Bereich. Traditionelle Methoden zur Identifizierung neuer Viren beruhen auf der Probenentnahme, Verarbeitung und Sequenzierung viraler Genome im Labor. Neue datenbasierte Ansätze ergänzen dies, indem sie Sequenzdaten aus öffentlichen Repositoryn nach viralen Sequenzen durchsuchen. Im Rahmen dieser Masterarbeit habe ich zwei Datensätze erhalten: Sequenzierdaten aus der Leber von Patienten nach einer Lebertransplantation sowie Sequenzierdaten aus Säugetieren, die vom Sequence Read Archive Repository heruntergeladen wurden. Eine zweistufige, datengestützte Pipeline zur Identifizierung neuer Viren wurde auf beide Datensätze angewendet. Im ersten Schritt, genannt VirusHunter, werden neue und bekannte Viren in unprozessierten Sequenzierdaten mithilfe eines Profil-Hidden-Markov-Modells identifiziert. Im zweiten Schritt, VirusGatherer, erfolgt die Assemblierung der potenziell viralen Sequenzen. Mit dem Ziel zu untersuchen, ob virale Sequenzen in den Lebersequenzierdaten vorhanden sind und ob diese möglicherweise mit Leberabstössungen assoziiert sind, wurde die Pipeline angewendet. Dabei konnten nur in einem kleinen Anteil der Sequenzierdaten virale Sequenzen identifiziert werden, die jedoch sehr kurz waren. Dies deutet darauf hin, dass entweder keine Viren in den Lebersequenzierdaten vorhanden waren oder dass virale Sequenzen größtenteils durch die Vorverarbeitungsschritte im Labor entfernt wurden. In den Säugetierdaten wurden potenzielle virale Sequenzen von RNA-Viren bereits identifiziert, jedoch noch nicht assembliert. Durch die Anwendung der Pipeline konnten hier über 6.000 virale Sequenzen assembliert werden. Diese

Zusammenfassung

Sequenzen wiesen Ähnlichkeiten mit Viren aus 65 RNA-Virusfamilien auf, die größtenteils charakteristische Viren umfassen, die bei Nutztieren vorkommen. Der Großteil der Sequenzen war im Vergleich zu den viralen Referenzsequenzen neuartig. Gleichzeitig waren jedoch die meisten Sequenzen kurz, sodass nur 444 sowohl neuartig im Vergleich zu Referenz als auch länger als 1,000 Nukleotide waren. Eine funktionelle Annotation der längsten Sequenzen hat gezeigt, dass die Pipeline in der Lage ist, echte virale Sequenzen zu identifizieren und zu assemblieren. Zusätzlich zur Analyse der Daten wurde im Rahmen dieser Masterarbeit ein R-Paket namens Virusparies entwickelt. Dieses Paket ermöglicht die Durchführung von deskriptiver Statistik, Visualisierungen und weiterer Prozessierung des Outputs von VirusHunter und VirusGatherer.

Abstract

Viruses are omnipresent. Viral genetic material can be found in various forms on land, in water, in living organisms, and in their microbiota. Today, we know that viruses are not just pathogens; they can also have positive effects on their hosts or the environment. They contribute to the evolution of their hosts or remain as commensals with little significant influence on the health of the host. However, the virosphere is still not fully explored, and estimates suggest that the number of undiscovered viruses is in the six-figure range.

Traditional methods for identifying new viruses rely on sample collection, processing, and sequencing of viral genomes in the lab. New data-driven approaches complement these by searching for viral sequences in publicly available sequence data repositories.

In this thesis, I received two datasets: sequencing data from the livers of patients after liver transplantation and sequencing data from mammals, downloaded from the Sequence Read Archive repository. A two-stage, data-driven pipeline for identifying new viruses was applied to both datasets.

In the first step, called VirusHunter, new and known viruses are identified in unprocessed sequencing data using a profile Hidden Markov Model. In the second step, VirusGatherer assembles the potentially viral sequences. The pipeline was applied to investigate whether viral sequences were present in the liver sequencing data and if they might be associated with liver rejections.

Only a small proportion of viral sequences were identified in the sequencing data, and they were very short. This suggests that either no viruses were present in the liver sequencing data, or that viral sequences were largely removed during pre-processing steps in the lab.

In the mammalian data, potential viral sequences from RNA viruses had already been identified but had not yet been assembled. By applying the pipeline, over 6,000 viral sequences were assembled. These sequences showed similarities to viruses from 65 RNA virus families, most of which include viruses commonly found in livestock. The majority of the sequences were novel

Abstract

compared to the viral reference sequences. However, most of the sequences were short, resulting in only 444 sequences that were both novel and longer than 1,000 nucleotides. A functional annotation of the longest sequences showed that the pipeline is capable of identifying and assembling authentic viral sequences.

In addition to analyzing the data, an R package called Virusparies was developed as part of this thesis. This package provides functions for descriptive statistics, visualizations, and further processing of the output from VirusHunter and VirusGatherer.

Table of Contents

1	Introduction	1
1.1	The need for virus discovery	1
1.1.1	The increasing frequency of pandemics and their viral origins	2
1.1.2	The virome: viruses within the body.....	3
1.1.3	Tracing the evolutionary history and surrogates viruses.....	7
1.2	History of viral discovery.....	9
1.3	VirusHunter and VirusGatherer	12
1.3.1	Virus ‘hunting’ in raw (unprocessed) sequencing data	14
1.3.2	Viral taxonomy	16
1.3.3	‘Gathering’ viral reads for assembly.....	20
1.3.4	Comparison to other approaches.....	21
1.4	Aim of thesis	22
2	Material and Methods	24
2.1	Code Availability Statement.....	24
2.2	Data Availability Statement.....	24
2.3	Software used.....	24
2.4	Sequencing data.....	25
2.4.1	Taubert liver transplant data	25
2.4.2	Mammalian data	25
2.5	VirusHunter.....	27
2.5.1	SRA download	27
2.5.2	Preprocessing of FASTQ files.....	27
2.5.3	Sensitive homology-based detection of viral sequence reads in unprocessed data	27
2.5.4	Filtering against contaminants and viral reference sequences	28
2.5.5	Output of VirusHunter	29
2.6	VirusGatherer	30
2.6.1	Assembly of Contigs	30
2.6.2	Compare against viral reference database	31
2.6.3	Output of VirusGatherer	31
2.7	Virusparies.....	32
2.8	Prediction and visualization of predicted functional domains	32
3	Results	34
3.1	Characterization of the virome after liver transplantation	34

Table of Contents

3.1.1 The majority of hits identified in the liver sequencing data are not significant.....	34
3.1.2 Only a small fraction of the initial dataset detected possible viral contigs .	36
3.1.3 All contigs identified in the liver sequencing data are short	37
3.2 Characterization and assembly of RNA viruses from mammalian samples	40
3.2.1 Over 6,000 contigs found in mammalian sequencing data	40
3.2.2 VirusGatherer assembled contigs from over 300 SRA experiments, identifying 69 RNA virus families	42
3.2.3 Most of assembled contigs originate from novel viruses	44
3.2.4 Four-fifths of all assembled contigs measure less than 1,000 nucleotides	47
3.2.5 Mammalian hosts are mainly farm animals and Old World monkeys	52
3.2.6 Unclassified Viral Families and the Specificity of Hidden Markov Model Profiles	53
4 Discussion.....	57
4.1 Discussion of the results obtained from Taubert sequencing data.....	58
4.2 Discussion of the results obtained from screening and assembling mammalian viral sequences.....	60
4.3 Virusparies.....	64
4.4 Outlook	66
5 List of references	68
6 Acknowledgements.....	81
7 Appendix.....	83

List of Abbreviations

Abbreviation	Meaning
AIDS	Acquired immunodeficiency syndrome
ARPANET	Advanced Research Projects Agency Network
AWS	Amazon Web Services
BLAST	Basic Local Alignment Search Tool
chr	Character
DAAs	Direct-acting antiviral (regimens)
DDVD	Data-driven virus discovery
EnvGP	Envelope glycoproteins
GBV-B	George Baker Virus B
GP	Glycoprotein
HCC	Hepatocellular carcinoma
HCV	Hepatitis C virus
HDAg	Hepatitis delta antigen
HERVs	Human endogenous retroviruses
HPV	Human papillomavirus
HSPs	High-scoring segment pairs
ICTV	International Committee on Taxonomy of Viruses
ITS	Internal transcribed spacer
IQR	Interquartile range
MCP	Major capsid protein
MERS	Middle east respiratory syndrome
MHH	Hannover Medical School
MSA	Multiple sequence alignments
NCBI	National Center for Biotechnology Information
NiRAN	(Nidovirus) RNA-dependent RNA-polymerase-associated Nucleotidyltransferase
N-protein	Nucleocapsid
NIH	National Institute of Health
NSP	Non-structural proteins
ORF	Open reading frame
P/V-PP	Phosphoprotein
PAM	Pulmonary alveolar macrophages
PCR	Polymerase chain reaction
pHMM	Profile Hidden Markov Model
PRRSV	Porcine reproductive and respiratory syndrome virus

List of Abbreviations

PRRSV-1	Porcine reproductive and respiratory syndrome virus 1
PRRSV-2	Porcine reproductive and respiratory syndrome virus 2
Q1/3	Quartile 1/3
RdRp	RNA-dependent RNA polymerase
Rep	Replication initiator protein
SRA	Sequence Read Archive
STI	Sexually transmitted infection
TCP/IP	Transmission Control Protocol/Internet Protocol
TMV	Tobacco mosaic virus

List of Figures

Figure 1: Growth of the Sequence Read Archive database.....	11
Figure 2: Schematic workflow of VirusHunterGatherer pipeline.....	13
Figure 3: Mammalian sequencing data - Distribution of viral groups detected across query sequences for all profiles	26
Figure 4: Taubert sequencing data - Distribution of viral reference E-values for each viral family detected by VirusHunter.....	35
Figure 5: Taubert sequencing data – VirusHunter results.....	37
Figure 6: Taubert sequencing data – VirusGatherer results.	38
Figure 7: Mammalian sequencing data - Distribution of viral reference E-values for each viral family (VirusGatherer).....	41
Figure 8: Mammalian sequencing data - Distribution of the number of SRA experiments detecting viral families.	43
Figure 9: Mammalian sequencing data - Protein sequence identities to the nearest known reference virus for each viral family.	45
Figure 10: Mammalian sequencing data – Distribution of the length of the contigs in each viral family.	48
Figure 11: Mammalian sequencing data – Functional annotation of the top ten largest contigs.....	51
Figure 12: Distribution of contigs with assigned viral reference families.....	54
Figure 13: Screening for Flaviviridae RNA-dependent RNA polymerase (RdRp) - Distribution of the number of SRA experiments detecting viral families.....	56

List of Tables

Table 1: Taubert sequencing data	25
Table 2: Columns of the VirusHunter hittable	29
Table 3: Columns of the VirusGatherer hittable	31
Table 4: Taubert sequencing data - Closest viral family and subject found among large DNA, small DNA, and RNA viruses.....	39
Table 5: Eleven non-RNA viruses identified in mammalian sequencing data during RNA virus screening.	44
Table 6: Top ten viral families with the highest number of contigs aligning to them. .	47
Table 7: Mammalian sequencing data - Closest viral family and subject found.....	52
Table 8: VirusHunter and VirusGatherer dependencies	83
Table 9: Profiles used for RNA virus screening in VirusHunter.....	86
Table 10: Profiles used for large DNA virus screening in VirusHunter.....	87
Table 11: Profiles used for small DNA virus screening in VirusHunter	89
Table 12: Taubert VirusGatherer – Summary statistics of the distribution of sequence identity percentages to the closest viral reference sequences.....	89
Table 13: Taubert VirusGatherer – Summary statistics of the distribution of contig lengths.....	90
Table 14: Mammalian sequencing data – Summary Statistics of the distribution of viral reference E-values for each viral family (VirusGatherer).....	90
Table 15: Mammalian sequencing data - Summary Statistics of the distribution of protein sequence identities to the nearest known reference virus for each viral family.	92
Table 16: Mammalian sequencing data – Number and percentage of contigs above and below 90% sequence identity across all phyla.....	94
Table 17: Mammalian sequencing data - Summary Statistics of the distribution of contig lengths for each viral family.	94
Table 18: Mammalian sequencing data – Number and percentage of contigs with sequence identity below 90% and length above 1,000 nucleotides, grouped by phyla and family.....	96
Table 19: Mammalian sequencing data – Top ten longest contigs: host, SRA experiment, length, closest viral match with phylum and sequence identity	98
Table 20: Mammalian sequencing data – Number and percentage (relative to total contig count) of contigs found in each host.....	98

1 Introduction

1.1 The need for virus discovery

Viruses are omnipresent. In aquatic environments, a single millilitre can contain up to hundreds of millions of viral particles¹; while in soil, a gram can hold as many as a billion². However, one need not dig into the earth or dive into the ocean to encounter viruses, as the human body itself is a viral ecosystem, housing over 10^{13} viral particles³. With an estimated 10^{31} viral particles globally, viruses outnumber all other life forms combined – including animals, plants, fungi, bacteria, and archaea^{4,5}.

Viruses also exhibit remarkable structural diversity, displaying a range of genomic configurations that set them apart from other life forms. While most organisms utilize double-stranded DNA to store genetic information, viruses can possess either DNA or RNA, existing in forms that are single-stranded or double-stranded, linear or circular⁶. Their genomes may consist of a single continuous piece of nucleic acid or be segmented, and they can be classified as positive-sense (read in the 5' to 3' direction) or negative-sense (read from 3' to 5')⁶.

While viruses are abundant and exhibit notable genomic diversity, our historical understanding of them has often been overly simplistic, framing them primarily as pathogenic agents without considering any potential beneficial roles they might play in their hosts or ecosystems. Contemporary research has begun to reshape this perspective. In addition to their ability to cause disease and pandemics, viruses may also contribute positively to the environment. For example, viruses lyse ~40% of marine prokaryotes daily, stimulating the plankton growth⁷. In humans, viruses have also played a role in shaping our evolution. Around 8% of the human genome consists of human endogenous retroviruses (HERVs), which resulted from the integration of these viruses into our genome. Although these viral sequences are no longer infectious, some still produce transcripts and have been linked to various cancers, including ovarian cancer⁸, prostate cancer⁹, breast cancer, lymphoma¹⁰, and melanoma¹¹. On the other hand, HERVs also play beneficial roles, such as contributing to human placental morphogenesis¹².

Despite these insights, the virosphere remains largely unexplored, as do its complex interactions with hosts and ecosystems. It is estimated that 1.67 million unknown viral species exist, with up to 827,000 potentially capable of infecting humans^{13,14}.

Virus discovery focuses on identifying these new and unknown viruses. The thesis explores the history of viral discovery and introduces a modern, data-driven approach that leverages publicly available, unprocessed data. Specifically, it presents VirusHunter, a tool for data-driven virus discovery (DDVD), alongside VirusGatherer, which assembles identified viral contigs. The competition landscape to VirusHunter and VirusGatherer is also be discussed, concluding the Introduction with the goals of the master thesis.

But first, I begin by addressing a key question: '*Why do we need virus discovery?*'

1.1.1 The increasing frequency of pandemics and their viral origins

Between 1918 and 1919, an avian influenza virus that had previously infected only waterfowl began to spread to farm poultry and pigs. Eventually, it breached the species barrier, posing a global threat to humans. Known as the 'Spanish' influenza H1N1, this pandemic infected one-third of the global population with more than 50 million deaths¹⁵. Historically, major pandemics like the 'Spanish Flu' were rare, occurring perhaps once or twice per century. However, in recent decades, the frequency of pandemics has significantly increased due to factors such as globalization, increased air travel, population growth, and human encroachment into wildlife habitats¹⁶.

The first two decades of the 21st century alone have witnessed six major pandemics. The first was the SARS pandemic from 2002 to 2004, caused by the SARS-CoV virus from the *Coronaviridae*¹⁷ family. This was followed by the H1N1 'swine flu' pandemic in 2009. Between 2012 and 2016, several viral outbreaks emerged, including Middle East Respiratory Syndrome (MERS), the Ebola virus epidemic, and the Zika virus outbreak. Most recently, the global COVID-19 pandemic, caused by SARS-CoV-2, emerged in 2019¹⁶.

Notably, all six major pandemics of this century have been caused by viral infections, with the majority originating from animals and transmitted primarily through airborne means. Viruses that emerge through zoonosis pose a significant threat because the new host species lacks existing immunity^{18,19}. As the host's immune system has not previously encountered these viruses, this lack of immunity, combined with the airborne transmission, increasing population densities, and intensified human-animal interactions, facilitates rapid viral spread across populations.

Effective prevention of global pandemics relies on the prompt identification of the causative virus. Early identification is critical for the development of diagnostic tools, vaccines, and treatments, as these depends on knowing the infectious agent and its pathogenic mechanisms. Monitoring existing viruses and discovering new ones has long been feasible through methods like metagenomic analyses of wastewater and other environments^{20–24}. However, viral discovery is notably more challenging than the identification of bacteria or fungi. Bacteria can often be identified using cultures and 16S rRNA sequencing^{25–28}, and fungi through internal transcribed spacer (ITS) region sequencing^{29,30}. In contrast, viruses lack standardized culture methods and universal sequencing markers, complicating the discovery process³¹. Instead, thorough sampling of the viral community, genome sequencing, and comparison with existing reference genomes are necessary. This is complicated by the absence of yet uncharacterized viruses in reference databases and the difficulty of detecting sequences with low homology to known viruses³². Yet, despite these challenges, rapid monitoring of viral activity is crucial, as climate change and global interconnectedness elevate our contact with both humans and animals, heightening the risk of new outbreaks and potential pandemics^{33–35}.

1.1.2 The virome: viruses within the body

Besides the identification of viruses during possible pandemic and disease scenarios, viral discovery could also enhance our understanding of the virome—the collective population of eukaryotic and prokaryotic viruses inhabiting the host body. Traditionally, viruses were viewed solely as disease-causing agents with no benefits to their hosts or the environment. However, this perception has radically shifted.

Today, we recognize that each human harbors a virome consisting of over 10^{13} viral particles³, with each body site creating its own niche, and each niche harboring viruses that may have no pathogenic effects or even offer beneficial effects to the host³⁶.

For example, the digestive system houses the most abundant population of viruses, most of which target the bacteria in our gut³⁶. These viruses, known as bacteriophages, inject their genome into bacterial cells, after which they can take one of two pathways. In the lytic cycle, they exploit the host's replication machinery to produce more viral particles, which are released by lysing/destroying the host cell. Alternatively, in the lysogenic cycle, the bacteriophage genome integrates into the host's genome and remains dormant, switching to the lytic cycle when stress signals indicate a threat to the virus's genomic stability³⁶. This integration may benefit the host by inducing mutations that help bacterial adaptation³⁷, or by providing immunity against further viral infections through mechanisms like CRISPR-Cas systems³⁸. On the other hand, direct lysis of bacteria controls the composition and functionality of the microbiome by removing non-resistant strains, while bacteriophage composition adapts to bacterial resistances, driving antagonistic co-evolution^{39–41}. Direct lysis has also been shown to contribute to human innate immunity, as some T4 phages can associate with the mucosal surface of the gastrointestinal tract, preventing bacterial infection of epithelial cells by eliminating invading bacteria⁴². This means the virome can play a direct role in establishing an innate immune barrier.

Some eukaryotic viruses can integrate their genomes into host cells, including human cells. About 8% of the human genome consists of human endogenous retroviruses (HERVs), which are the result of viral integrations that have occurred throughout human evolution. While these HERVs are no longer infectious, they can still produce transcripts linked to several cancers, such as ovarian cancer⁸, prostate cancer⁹, breast cancer, lymphoma¹⁰, and melanoma¹¹. Conversely, some HERVs have beneficial roles, such as contributing to human placental development¹².

Some viruses can remain latent but may reactivate under certain conditions. Human herpesvirus 6, which affects nearly 90% of the population, can cause a range of symptoms when active, including reduced production of blood cells in the

bone marrow (myelosuppression), inflammation of the lung (pneumonitis) and brain tissue (encephalitis), and skin rashes. Additionally, herpesvirus 6, along with adenoviruses, can play a role in graft-versus-host disease, where donor immune cells attack the recipient's cells after transplantation^{43–45}. Conversely, while viruses from the families *Redondoviridae*⁴⁶ and *Anelloviridae*⁴⁷ may become more abundant in immunosuppressed individuals, no health issues are currently known to be associated with them³⁶. This illustrates the dual nature of viruses, where some can cause significant health problems while others may coexist peacefully within the host.

Various host-related factors can further influence the virome's composition and abundance over time. Under normal circumstances, the virome develops shortly after birth^{48,49}, remains stable through adulthood, and decreases in old age⁵⁰, but smoking, for instance, has been shown to alter the composition of phages in the lung⁵¹. Dietary habits also influence the virome, with many gut viruses originating from the food we consume, including plant-based viruses from the family *Virgaviridae*⁵². One study examined how different diets affect the virome and found that individuals with similar diets had more closely related viromes than those with differing diets⁵³. A shift in diet throughout life can also influence the gastrointestinal virome. Research that fed mice a high-fat diet observed a decrease in viruses from the families *Myoviridae*, *Siphoviridae*, and *Podoviridae*, alongside an increase in *Microviridae*. As previously mentioned, phages can be either lytic or lysogenic, with most phages in the virome typically in the lysogenic state. However, this project also detected a shift toward more lytic viral communities, suggesting that phages can influence bacterial composition in response to dietary changes⁵⁴.

Recent work in both twins and non-twins emphasize the greater influence of shared environmental factors compared to genetic factors in shaping the virome^{55–57}. Although genetics can affect the virome, especially in cases of immunodeficiencies⁵⁸, environmental factors like geography and social interactions play a crucial role. As an illustration, findings indicate that individuals from the same household typically have more similar oral viromes compared to those from different households, underscoring the impact of shared living environments on viral exposure⁵⁹. The exchange of viruses is not confined to humans; it crosses species barriers. Pets of individuals with COVID-19 have tested positive for the SARS-CoV-

2 N gene and, in some cases, developed antibodies, indicating that humans can transmit the virus to their pets⁶⁰. Furthermore, humans themselves are not the natural reservoirs for SARS-CoV-2. Instead, the viruses responsible for major pandemics over the last 20 years—SARS-CoV-1, SARS-CoV-2, and MERS—are zoonotic diseases, originating in animals and transmitted to humans⁶¹. This exchange of SARS-CoV-2 from animals to humans and from humans to animals illustrates the intricate web of viral transmission that connects different species.

Geography impacts the virome, as evidenced by the lower virome diversity observed in individuals from the highly urbanized city of Hong Kong compared to those in the less urban, partially rural Yunnan province in China. Notably, residents of Hong Kong exhibited an increased presence of phages targeting *Lactobacillus* and *Lactococcus* bacteria⁶². Similarly, research indicated that Australian children in the Northern Territory with diarrhea displayed a higher prevalence of viruses from the *Picornaviridae* and *Adenoviridae* families compared to their counterparts in Melbourne⁶³. The environment not only shapes our virome but is also in return influenced by viruses. For instance, viruses lyse 40% of marine prokaryotes each day, stimulating plankton growth⁷. Furthermore, they regulate the global carbon cycle by breaking down biomass, which contributes to new dissolved carbon sources in soil and water⁶⁴.

In conclusion, viruses establish highly diverse and abundant populations within their hosts, interacting in various ways: they can cause disease, benefit the host, or have no known pathogenic effects. These interactions occur either directly or by altering the composition and functionality of the host's microorganisms. Conversely, the host can influence its virome through changes in habits or environment. The environment affects the host, microorganisms, and virome, and is itself influenced by the virome. This creates a complex network of interactions. Many viruses remain unknown, leaving the complex mechanisms of this network largely unexplored. Virus discovery through metagenomic analysis of the virome and environmental samples can help identify these unknown viruses, leading to a better understanding of the virome and its interactions with other organisms and the environment.

1.1.3 Tracing the evolutionary history and surrogates viruses

Virus discovery plays a crucial role in identifying surrogate viruses. The disease mechanisms and host responses of certain viruses cannot be directly studied in their natural hosts. This limitation complicates the development of effective treatments and vaccines. In such cases, studying closely related viruses can provide a valuable alternative for research.

A pertinent example is the hepatitis C virus (HCV), a positive-sense single-stranded RNA virus belonging to the genus *Hepacivirus* within the family *Flaviviridae*, part of the phylum *Kitrinoviricota*⁶⁵. HCV infections can lead to the accumulation of scar tissue in the liver due to chronic inflammation (liver fibrosis), followed by impaired liver function (decompressed cirrhosis), and potentially culminating in liver cancer (hepatocellular carcinoma, or HCC)⁶⁶. According to the Global Hepatitis Report 2024, ~50 million people are living with HCV infection, with nearly one million new cases reported in 2022 alone⁶⁶. HCV is globally prevalent, with injection drug users particularly at risk, a situation exacerbated by the opioid epidemic in the United States⁶⁷.

Current treatment for chronic HCV infection in both children and adult populations primarily target viral replication and polyprotein processing through direct-acting antiviral (DAAs) regimens, such as sofosbuvir/declatasvir, sofosbuvir/velpatasvir, or glecaprevir/pibrentasvir⁶⁶. Despite the availability of these treatments, access remains restricted in under-resourced areas, among uninsured individuals, and particularly within high-risk populations, such as injection drug users, who face elevated risks of reinfection^{68,69}. This stark reality underscores the pressing needs for a long-term solution, namely, the development of an effective vaccine against HCV.

The clinical development of vaccines often relies on animal models to study immune responses and disease progression⁷⁰. Historically, chimpanzees have served as the primary animal models for HCV, because they are the only other species, aside humans that can naturally harbor the virus. The first HCV clone was developed in chimpanzees^{71,72}, yielding significant insights into the role of CD8+ and CD4+ T cells during HCV infection^{73,74}. Despite these advances, the use of chimpanzees in biomedical research is now severely restricted following decision by the National Center for Research to halt funding for chimpanzee breeding⁷⁵ and

their designation as endangered species by the U.S Fish and Wildlife Service in 2015^{76,77}. Consequently, the National Institute of Health (NIH) has ceased supporting such studies⁷⁸.

In response, alternative research strategies have emerged to sustain HCV studies. One approach involves exploring and developing new model organisms, particularly in smaller animal models, although this has proven challenging due to the restricted host range of HCV⁷⁹. Another strategy relies on the discovery of new viruses that are closely related to HCV as potential replacement for both HCV and chimpanzees. To date, over 250 viruses from the family *Flaviviridae* have been identified, with a diverse array of hosts, including primates and other mammals such as equines, canines, and bovines⁷⁹. The discovery of the first *Hepacivirus* infecting non-mammalian hosts, such as catsharks, in 2016 significantly broadened our understanding of *Flaviviridae* diversity. Previously, *Hepaciviruses* were known to infect only mammals, and prior to 2011, exclusively humans and primates^{80,81}. Among primates, the George-Baker virus B (GBV-B) has emerged as a valuable research model system. GBV-B, also a positive-sense single stranded RNA virus from the family *Flaviviridae* and genus *Hepacivirus*, was first identified in tamarins that developed hepatitis following exposure to serum from a patient named George Baker^{79,82}. Tamarins offers advantages in laboratory settings due to their manageable size, and their immune response post-infection shows similarities to humans. A virus more closely related to HCV than GBV-B is the *equine hepacivirus*, which has also been identified in canines (*canine hepacivirus*). Additional non-primate hepaciviruses have been also identified in bats, cattle, and rodents; however, none have been as effective as GBV-B to substitute for HCV⁷⁹. Thus, GBV-B remains the current surrogate for HCV albeit with notable limitations. For instance, chimeric viruses combining HCV and GBV-B sequences exhibit reduced efficacy⁷⁹. Moreover, tamarin breeding is costly, and while chronic infection is a hallmark of HCV in humans, tamarins rarely develop chronic infection after GBV-B infection, posing challenges for studying long-term disease dynamics. Similarly to chimpanzees, research using tamarins also faces ethical concerns about animal welfare, the necessitate for specialized facilities as well as expert veterinary care, and high costs⁸³. These limitations highlight the need of optimizing current methods and identifying additional surrogate viruses for HCV.

The role of viral discovery is of note in this regard. Prior to 2011, hepaciviruses were believed to infect only humans and chimpanzees. However, the discovery of new mammalian and non-mammalian hepaciviruses has revolutionized our understanding of the family *Flaviviridae* and enabled ongoing HCV research despite the restrictions of chimpanzee models. Viral discovery enables the identification of both novel and known viruses, enhances our understanding of their evolutionary relationship, and allows for the use of closely related viruses as effective surrogates for other family members.

1.2 History of viral discovery

The discovery of viruses can be tracked back to the pioneering work of Russian biologist Dmitri Iosifovich Ivanovsky in the late 19th century^{84,85}. While studying at the University of Saint Petersburg, Ivanovsky investigated a disease affecting tobacco crops in Ukraine and Bessarabia, which he termed ‘pock disease’. He was sent to Crimea to study tobacco plants afflicted with brown spots, consistent with the disease he had described. The plants also displayed dark-green and yellow areas on their leaves, symptoms now recognized as tobacco mosaic disease⁸⁶.

This disease had first been described by Adolf Mayer, who noted that it could infect nearby tobacco plants but lost its infectious quality if bacteria were filtered from the plant sap⁸⁷. To test this, Ivanovsky employed a porcelain Chamberland filter, designed to retain bacteria and other microorganisms, while allowing smaller particles to pass through. He then inoculated healthy plants with the filtered homogenate from affected plants. The inoculated plants soon exhibited the same distinct mosaic pattern on their leaves. Ivanovsky published the discovery of the tobacco mosaic virus (TMV) in 1892, revealing the existence of infectious agents smaller than bacteria⁸⁶. In 1939, Martinus Willem Beijerinck independently identified the tobacco mosaic virus and introduced the concept of a virus as infectious particles that could only replicate inside living cells^{88,89}. Seven years after Ivanovsky and the discovery of the first (plant) virus, the presence of viruses in animals was confirmed⁹⁰. Notably, up to this point, only the presence of small infectious particles could be proven, but their identity, structure, and sequence remained unknown.

It was not until 1935 that Wendell Meredith Stanley managed to crystallize the infectious particles of TMV⁹¹, and another six years later, John Desmond Bernal and Isidor Fankuchen used X-ray diffraction on these crystallized particles, leading to the first description of the viral structure's size and shape in 1941⁹². Meaning it took 49 years between Ivanovsky's first viral discovery and the first description of the viral structure of the same virus, and it would take until 1977 for the invention of Sanger sequencing, which allowed for the identification of the viral genome⁹³. Today, viruses can be identified by both their structure and nucleic acid sequences, with new technologies such as polymerase chain reaction (PCR)⁹⁴ and second- and third-generation sequencers⁹⁵ enabling faster virus discovery on a larger scale. High-throughput sequencers led to metagenomics, enabling the sequencing and analysis of entire microbial populations, including complete viral communities from specific environments. In 2002, Breitbart et al. conducted the first study of environmental viral communities, identifying a large and diverse collection of phage sequences in seawater⁹⁶. Another significant development took place on January 1, 1983, when the Advanced Research Projects Agency Network (ARPANET) adopted the Transmission Control Protocol/Internet Protocol (TCP/IP), enabling communication between computers on different networks and paving the way for the modern Internet⁹⁷. There is no longer a need to distribute viral genome data on physical media, which previously limited the amount of information that could be shared or accessed. Instead, public repositories like the Sequence Read Archive (SRA) have been established, enabling researchers to share unprocessed genome data openly⁹⁸. New sequencing technologies have significantly increased the volume of sequencing data generated simultaneously. As of September 2024, approximately 91.91 petabases have been uploaded to SRA, with 53 petabases being open access⁹⁹ (Figure 1).

Previous studies that mined publicly available sequencing data repositories have demonstrated that viral genes and genome sequences are often detected as by-products when sequencing a host, even when the research was not intended to study viruses^{100–103}. Lauber et al. highlighted that raw sequencing data from public repositories, like those in the SRA, offer new opportunities for virus discovery using computational approaches¹⁰⁴. Traditionally, identifying viruses—both known and novel—depended on collecting biological samples and performing laboratory-based

processing and analysis, which constrained research to the data labs could physically handle^{102–105}. In contrast, data-driven methods tap into these vast archives of unprocessed sequencing data, leveraging powerful, parallelized computing to uncover viral sequences retrospectively. This allows researchers to examine far more data than was possible with conventional lab-based techniques. Lauber et al. also noted that viral sequences can be present in samples collected for purposes other than viral identification. Conventional viral discovery often focuses on predefined pathogens, specific hosts, or geographically restricted areas, which can result in unnoticed viral presence. By analyzing large volumes of existing data without focusing on predefined factors, data-driven approaches can capture viral diversity that conventional methods may miss. Besides sequencing data, detailed metadata about the study, host, and host tissue is usually also provided, enabling assignment of host information to the identified virus. Free access to public sequencing data also reduces costs compared to traditional laboratory-based viral discovery, as it eliminates laboratory expenses and requires only personnel and computing resources. For these reasons, data-driven virus discovery (DDVD) offers a novel approach to exploring the natural diversity of viruses, complementing traditional discovery methods that rely on wet lab experiments¹⁰⁴.

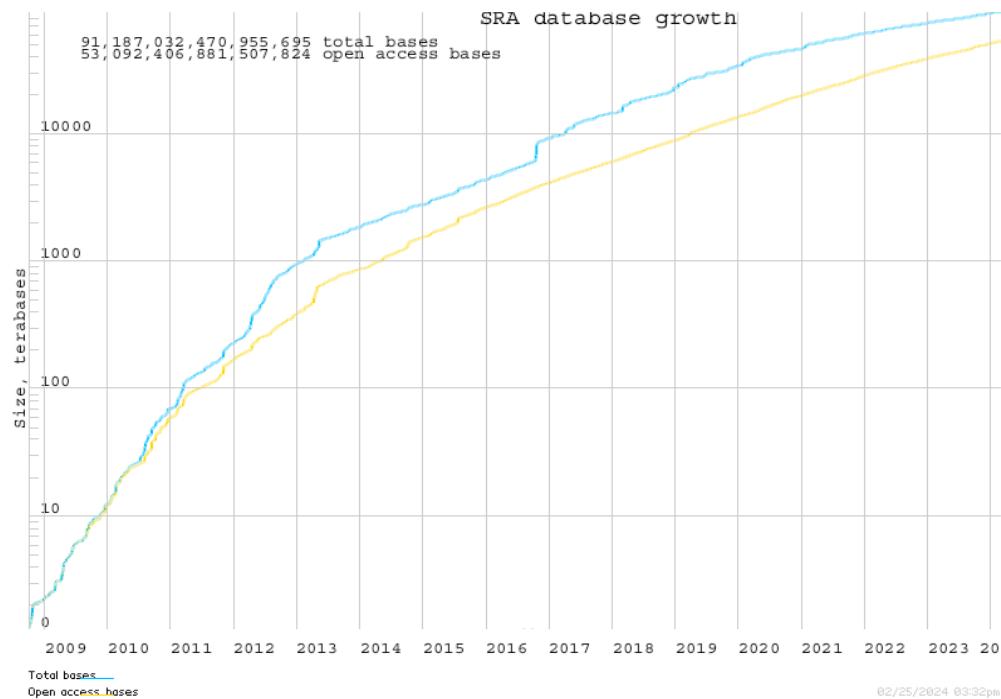
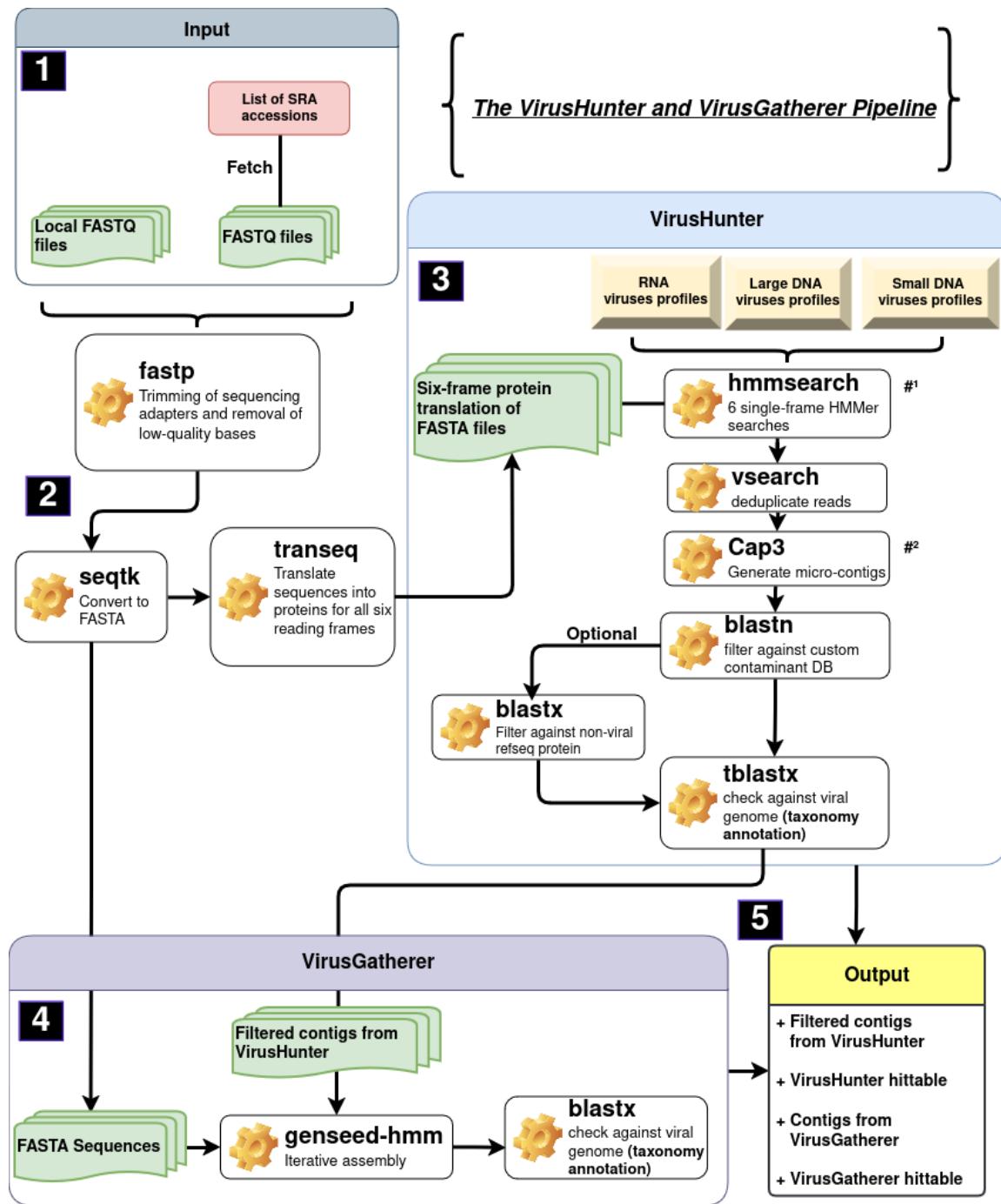


Figure 1: Growth of the Sequence Read Archive database from June 5, 2007, to February 25, 2024.⁹⁹

1.3 VirusHunter and VirusGatherer

Various computational solutions have emerged to facilitate virus discovery^{102,109–115}. For instance, Zayed et al. used RNA-dependent RNA polymerase (RdRp) sequences from 28 terabases of RNA data to uncover 5,500 novel RNA virus contigs by aligning them with RdRp motifs, demonstrating the potential of finding novel virus sequences in large-scale datasets¹¹⁶. Similarly, the Serratus platform - the first cloud-based system for viral discovery - analyzed 5.7 million samples, totaling 10.2 petabases, identifying 132,000 new viruses using conserved regions of the RdRp catalytic core from publicly available SRA data¹⁰². Additionally, another software solution (DAMIAN) has attempted to integrate viral discovery with cohort analysis, focusing on identifying which novel viruses could be linked to specific diseases in patients¹¹⁵. These examples underscore the significant role computational tools play in expanding our understanding of viral diversity.

The VirusHunter and VirusGatherer pipeline is another advancement in data-driven virus discovery (DDVD) (Figure 2), designed for the efficient detection and assembly of viral sequences from large-scale datasets¹⁰⁴. VirusHunter identifies viral sequences in raw (either local or SRA-downloaded unprocessed) sequencing data by matching them against protein profiles using a profile Hidden Markov Model^{117,118} (pHMM) indicative of a specific virus group. VirusGatherer then assembles overlapping reads identified by VirusHunter into longer viral contigs or complete genomes. The following subchapter details this dual approach (VirusHunter and VirusGatherer pipelines; hereafter sometimes referred to as ‘VirusHunterGatherer’).



#1 > Only HMM-positive reads are used in the subsequent pipeline steps following the HMM profile search.
#2 > Micro-contigs generated from Cap3 assembly are used in the filtering step.

Figure 2: Schematic workflow of VirusHunterGatherer pipeline. This pipeline facilitates viral discovery by processing either local sequencing files (FASTQ) or downloading sequencing data from public datasets, such as those from the NCBI SRA database⁹⁸, using SRA accession numbers (1). Initially, FASTQ files are preprocessed to remove sequencing artifacts and then converted into FASTA format (2). VirusHunter translates nucleotide sequences into protein sequences across all six reading frames. The translated sequences are then screened using profile Hidden Markov Models^{117,118} (pHMMs; profiles of either DNA or RNA viruses), in order to detect viral reads. After deduplicating the detected viral reads, the remaining sequences are assembled into short viral segments called 'micro-contigs'. Custom contaminants and optional non-viral reference sequences are filtered out, and taxonomic annotation is applied by pairwise aligning the viral contigs against a reference viral database (3). The contigs generated by VirusHunter, along with the quality-controlled initial input data are then passed to VirusGatherer, which performs iterative assembly to construct longer viral contigs or complete genomes (4). These assembled contigs undergo a final taxonomic annotation to identify the viral taxa present. The final outputs include contigs and hittables, which are documenting the best viral sequences (5).

1.3.1 Virus ‘hunting’ in raw (unprocessed) sequencing data

The Basic Local Alignment Search Tool (BLAST) has become a standard for identifying regions of similarity between query sequences and databases using pairwise alignment^{116,117}. While widely used for certain viral discovery applications, such as computational pipelines like VirusSeeker¹¹¹, BLAST lacks the sensitivity to detect distant evolutionary relationships^{32,121,122}, complicating the viral discovery that exhibit low sequence identity to reference genomes. To address this, alternative methods have been developed to detect evolutionary relationships among sequences with low identity^{117,118,123–125}. Instead of comparing a single sequence to another, these methods involve comparing the input sequence against a collection of related query sequences or a representation of that collection with the same characteristics.

Profile Hidden Markov Models (pHMM), as implemented in VirusHunter, exemplify this approach. Here, multiple sequence alignments (MSA) of protein domains, widely conserved within a virus group of interest, are constructed, representing the collection of sequences. However, rather than using the collection of sequences directly, the MSA is converted into pHMM, which use a position-specific scoring system to generate a probabilistic representation of the sequences¹²¹.

Probabilistic representation, in this context, involves capturing the probability of a specific amino acid appearing at each position across a set of aligned sequences used in the MSA, as well as the probability of transitioning to a particular amino acid at the next position. As a result, the models can be used to compare against new sequences, determining how likely a given amino acid from the input sequence is to appear at each position in the viral protein domain¹²¹. It is this ability of pHMMs to assess the probability of a query sequence aligning to a viral protein family, rather than individual sequences, that increases their sensitivity in comparison to BLAST¹²². That is why in a previous study, VirusHunterGatherer identified homologs with as low as 35% sequence identity to viral reference genomes^{100,126–128}.

Both locally generated sequencing data (in FASTQ format) and raw sequencing data downloaded from the National Center for Biotechnology Information (NCBI) SRA database can serve as input for VirusHunter (Figure 2, part 1). After quality control, including adapter trimming and removal of low-quality bases, the nucleotide sequences are translated into protein sequences (in FASTA format) across all six reading frames (Figure 2, part 2). Alignment is conducted at the proteome level because homologous viral proteins often diverge beyond recognition at the nucleotide level^{129,130}. This is particularly true for RNA viruses, where high mutation rates cause significant sequence divergence, even among copies of the same virus within a single host¹³¹. Viral proteins, on the other hand, are the functional building blocks of biology and tend to preserve their structure and function over long evolutionary periods¹³². This conservation allows for the comparison of viral protein sequences, even among distantly related viruses^{133–136}. The translated protein sequences are aligned against profiles of RNA, small DNA, or large DNA viruses (Figure 2, part 3). Ideally, these sequences align only, when they share the same protein family as the profiles used. Duplicate sequences resulting from wet-lab processes, such as duplicates from PCR amplification¹³⁷, are removed, and the remaining reads are assembled into ‘micro-contigs’.

The ‘micro-contigs’ are run through two BLAST filtering steps, one of which is optional; followed by a final BLAST step to filter out remaining non-viral contigs; and add taxonomic annotations to each identified viral contig. The initial filtering step removes sequences that match predefined custom contaminants, such as proteins from the hosts that might falsely align with short viral contigs due to high sequence similarity¹³⁸. The optional filter includes databases of host sequences, enabling the removal of host-specific sequences. Finally, the remaining contigs are aligned against a database of known viral reference sequences to eliminate any residual non-viral sequences. The output of VirusHunter includes the filtered contigs and an exported table file called ‘hittable,’ which contains the top hits from the homology search (pHMM positive reads) along with the final taxonomic annotation based on the alignment against the viral database (Figure 2, part 5).

1.3.2 Viral taxonomy

The International Committee on Taxonomy of Viruses (ICTV) serves as the global authority for the classification of viruses, organizing them into taxonomic categories based on shared characteristics¹³⁹. Its primary goal is to establish a universally accepted system for viral taxonomy, promoting consistency and clarity in virus identification. As of June 19, 2024, the ICTV recognizes six viral realms, including *Riboviria*^{141,137}, which encompasses RNA viruses (MSL #39; release v3, June 19, 2024)¹⁴². At the time of this writing, the classification system includes 10 kingdoms, 18 phyla, 2 subphyla, 41 classes, 81 orders, 11 suborders, 314 families, 200 subfamilies, 3,522 genera, 84 subgenera, and 14,690 species¹⁴².

The input data is aligned with profiles of small DNA, large DNA, or RNA viruses. RNA virus profiles are defined by the six phyla within the *Orthornavirae* kingdom, under the *Riboviria* realm^{141,137}. These include *Lenarviricota*, *Pisuviricota*, *Kitrinoviricota*, *Duplornaviricota*, *Negarnaviricota*, as first described by Wolf et al.¹⁴³, and the recently added phylum *Ambiviricota*¹⁴⁴. As of June 19, 2024, the ICTV recognizes nine families, four classes, and five orders within the *Lenarviricota* phylum, comprising 749 genera and 2,806 species. Viruses in this phylum are positive-strand RNA viruses. The family *Leviviridae*¹⁴⁵ is unique as the only known group of positive-strand RNA viruses that infect bacteria. Other families in this phylum, such as *Narnaviridae* and *Mitoviridae*, infect eukaryotes¹⁴⁰.

The *Pisuviricota* phylum is composed of 38 viral families, organized into 8 suborders and 7 orders, spanning 3 classes. It encompasses a total of 1,044 species, grouped into 82 subgenera and 197 genera¹⁴². *Picornavirales* is the largest order within this phylum, which is why *Pisuviricota* is also referred to as the 'picornavirus supergroup'¹⁴³. This order comprises positive-sense RNA viruses, with *Picornaviridae*¹⁴⁶ and *Secoviridae*¹⁴⁷ as some of the earliest identified families. *Picornaviridae* are known to infect a diverse array of hosts, including mammals, birds, reptiles, amphibians, and bony fishes. *Secoviridae* target dicotyledonous plants. Other families can also infect invertebrates such as insects¹⁴⁸. *Nidovirales* is the second-largest order and includes 14 families, such as *Coronaviridae*¹⁷, *Tobaniviridae*¹⁴⁹, and *Roniviridae*¹⁵⁰. These viruses are notable for having some of the largest known RNA genomes¹⁵¹. Members of the *Coronaviridae* family are notable here, as they have gained increased attention in the 21st century due to their

role in several major pandemics: the emergence of SARS-CoV in 2002, the MERS-CoV outbreak in 2012, and the SARS-CoV-2 pandemic in 2019¹⁶. Not all viruses within the *Pisuviricota* phylum have a single-stranded positive genome. For instance, viruses from the order *Durnavirales* possess double-stranded RNA genomes and infect a range of eukaryotic hosts, including fungi, plants, and both vertebrates and invertebrates¹⁴⁰.

Another phylum containing positive-strand RNA viruses is *Kitrinoviricota*, which, unlike *Lenarviricota*, includes only viruses that infect eukaryotes. Across 91 genera, 21 families, 6 orders, and 4 classes, there are 746 species that belong to *Kitrinoviricota*¹⁴². Members of the *Flaviviridae* family include known human pathogens, such as the hepatitis C virus (genus: *Hepacivirus*), which causes chronic liver disease and cirrhosis. Other members, such as those in the genus *Orthoflavivirus*, are transmitted to humans through the bites of infected ticks and mosquitoes and include pathogens like the *Orthoflavivirus flavi* (yellow fever virus), *Orthoflavivirus dengue* (dengue virus), *Orthoflavivirus zikaense* (zika virus), *Orthoflavivirus japonicum* (japanese encephalitis virus), and *Orthoflavivirus nilense* (west nile virus)¹⁵². Besides flavivirus-like viruses, the *Kitrinoviricota* phylum also includes the class *Alsuviricetes*, formerly known as the 'alphavirus supergroup'¹⁴³. While most viruses in this class infect plants, exceptions include *Paslahepevirus balayani* from the *Hepeviridae* family, known as hepatitis E virus¹⁵³, and *Rubivirus rubella* from the *Matonaviridae* family, known as the rubella virus¹⁵⁴.

Members of *Duplornaviricota* are double-stranded RNA viruses characterized by a conserved capsid protein structure. Their capsid is organized into an unusual T=1 lattice (also known as Pseudo-T=2 lattice), composed of 60 homo- or heterodimers of the capsid protein subunits¹⁵⁵. To date, 321 viral species have been identified across 39 genera, 22 families, 3 orders, and 3 classes¹⁴². Among those 22 families is *Cystoviridae*, which contains the only double stranded RNA viruses that can infect prokaryotes¹⁵⁶. Plant-, fungi-, and invertebrate-infecting viruses are dominant in this phylum, with one exception: the *Reoviridae* family, which, for example, contains the human pathogen rotavirus A¹⁵⁷.

Negarnaviricota comprises negative-sense, single-stranded RNA viruses that infect all hosts except prokaryotes. *Negarnaviricota* comprises 1,473 negative-strand RNA viruses across 264 genera, 37 families, and 6 classes¹⁴². Notable

examples include *Rabies lyssavirus* (rabies virus), *Zaire ebolavirus* (ebola virus), and the influenza viruses, which belong to the *Orthomyxoviridae* family¹⁵⁸.

Ambiviricota is the most recently discovered RNA virus phylum, currently comprising only 20 species distributed across 4 families: *Dumbiviridae*, *Quambiviridae*, *Trimbiviridae*, and *Unambiviridae*¹⁴². These viruses have circular RNA genomes with at least two non-overlapping open reading frames in an ambisense orientation and are known to infect fungi¹⁴⁴.

The differentiation between small DNA viruses and large DNA viruses is not an official classification but rather a system conceived by the Computational Virology Research Group to categorize DNA viruses with genomes smaller than 10,000 kb as small DNA viruses and those larger than 10,000 kb as large DNA viruses. Beyond this, most small DNA viruses are single-stranded, whereas large DNA viruses are predominantly double-stranded.

Large DNA viruses include six phyla: *Peploviricota*, *Uroviricota*, *Taleaviricota*, *Nucleocytoviricota*, *Preplasmaviricota*, and *Dividoviricota*. *Peploviricota* and *Uroviricota* belong to the kingdom *Heunggongvirae*, which is part of the realm *Duplodnaviria*¹⁴². Both phyla share only one class. *Peploviricota* contains just 133 species across 23 genera and 3 families, whereas *Uroviricota* encompasses 4,840 viral species distributed across 1,497 genera and 74 families¹⁴². An example for the former include herpesviruses such as Epstein-Barr virus, Kaposi's sarcoma virus, and herpes simplex virus type 1 and 2, all of which can infect humans^{159–161}. An example of the latter is members of the class *Caudoviricetes*, part of *Uroviricota*, which are the most abundant phages in the human virome³⁶. *Taleaviricota* currently comprises 32 archaeal viral species, organized into 14 genera and 5 families¹⁴².

Both the *African swine fever virus*, which spread out of Africa in 2007 and is currently spreading through the European pig population¹⁶², and poxviruses (*Poxviridae*), including *Orthopoxvirus variola* (smallpox), belong to the class *Pokkesviricetes*, part of the phylum *Nucleocytoviricota*¹⁶³. So far, 132 species, organized into 58 genera and 14 families, belong to the phylum *Nucleocytoviricota*. *Preplasmaviricota*, which are also part of the realm *Viridnaviria*, include 146 species in 28 genera, and 16 families¹⁴². Most known among them are the *Adenoviridae*, with hosts ranging from mammals and birds to reptiles, amphibians, and fish, depending on the genus¹⁶⁴. The last large DNA virus phylum, *Dividoviricota*,

consists of the order *Halopanivirales*, which infect thermophilic bacteria and archaea, and contains only 9 species across 3 genera¹⁴².

Cressdnaviricota is the first phylum of small DNA viruses, comprising 1,490 viral species, 266 genera, and 23 families¹⁴². This phylum is characterized by eukaryotic viruses with circular single-stranded DNA genomes that encode the replication initiator protein (Rep), enabling genome replication through a rolling-circle mechanism¹⁶⁵. The families *Geminiviridae*¹⁶⁶ and *Nanoviridae*¹⁶⁷ include plant-infecting viruses, while *Bacilladnaviridae* infect diatoms¹⁶⁸, and *Circoviridae* are found in mammals, birds, and fish¹⁶⁹. *Hoffneiviricota* is a phylum comprising a single class and order. It includes 60 prokaryotic viral species, 32 genera, and 3 families: *Plectoviridae*¹⁷⁰, *Paulinoviridae*¹⁷¹, and *Inoviridae*¹⁷². Similarly, *Phixviricota* also feature just one class and order. This phylum encompasses 22 species and 7 genera, with *Microviridae* as its sole family¹⁷³. The penultimate small DNA viruses phylum, *Cossaviricota*, includes 440 species and 90 genera across 4 families (*Bidnaviridae*¹⁷⁴, *Polyomaviridae*¹⁷⁵, *Papillomaviridae*¹⁷⁶, and *Parvoviridae*¹⁷⁷), while the last phylum, *Saleviricota*, comprises 16 species and 3 genera within a single family (*Pleolipoviridae*¹⁷⁸). All viruses in the *Pleolipoviridae* family target *Halobacteria*, which are halophilic archaea¹⁷⁸. Human papillomavirus (HPV), a prevalent sexually transmitted infection (STI), can lead to warts and, in some cases, penile, vaginal, anal, or cervical cancer¹⁷⁹. Other members of the *Papillomaviridae* family have been documented in mammals, birds, and fish¹⁷⁶. Similarly, *Polyomaviridae* and *Parvovirinae* have been observed in mammals, birds, and fish^{175,177}. However, *Densovirina*, a subfamily within *Parvoviridae*, target invertebrates such as insects and crustaceans¹⁷⁷.

VirusHunter does not exclusively utilize the defined phyla from small DNA, large DNA, and RNA viruses. Profiles were also constructed for families such as *Anelloviridae*⁴⁷ (small DNA viruses), *Yaraviridae*¹⁸⁰ (large DNA viruses), and *Birnaviridae*¹⁸¹ and *Permutotetraviridae*¹⁸² (RNA viruses). None of these families are yet assigned to a specific phylum. In addition to the known phyla for small DNA, large DNA, and RNA viruses, there are other phyla do not fit into these categories. Currently, only the phylum *Artverviricota* falls into this category. *Artverviricota* also belongs to the realm *Riboviria* like the six phyla defined as RNA viruses. However, *Artverviricota* is distinct in that it falls under the kingdom *Pararnaviriae*, which

includes viruses with RNA genomes that utilize a reverse transcriptase¹⁴². Examples include the family *Hepadnaviridae*¹⁸³ and the genus *Lentivirus*, some members of which are associated with acquired immunodeficiency syndrome (AIDS)¹⁸⁴.

1.3.3 ‘Gathering’ viral reads for assembly

Viral discovery extends beyond VirusHunter. Although viral ‘micro-contigs’ are identified at this stage, the contigs may not represent a complete protein-coding sequence, let alone a coding-complete viral genome, which is required to establish a new viral taxon¹⁸⁵. In response to this requirement, the viral ‘micro-contigs’ identified in VirusHunter serve as seeds for a progressive assembly in the VirusGatherer step of the pipeline, aimed at generating long viral contigs, or coding-complete viral genomes. Internally, VirusGatherer implements the seed-based assembly tool GenSeed-HMM¹⁸⁶ (Figure 2, part 4). A key advantage of GenSeed-HMM is its ability to accept input as nucleotide sequences, protein sequences, or pHMMs. GenSeed-HMM begins by conducting similarity searches with different tools depending on the input data: BLASTN for nucleotides, TBLASTN for proteins, and hmmsearch for pHMMs. These searches retrieve sequences where the seeds aligns with the initial input data used at the start of the pipeline (Figure 2, part 2). The sequences are then assembled into longer contigs using third-party assemblers like CAP3¹⁸⁷, Newbler¹⁸⁸, Velvet¹⁸⁹, SOAPdenovo¹⁹⁰, or ABySS¹⁹¹. Further contig assembly continues iteratively, with each round using contig ends from the previous iteration for homology searches against the initial seed sequences. Overlapping regions from the sequences identified in the homology search are merged with the contig ends. The resulting contig ends then serve as starting points for the next round of assembly. This allows the contig to grow in length as long as new sequences can be found in the homology search step. If no new sequences are found, GenSeed-HMM performs three additional extensions, trimming 25% of the contig end in each iteration before proceeding with the extension. If no new sequences are found for further extension in these three additional iterations, or if the contig length or number of iterations reaches a user-defined maximum, the assembly process stops. The resulting contigs are then aligned against a viral database, similar to the final VirusHunter step, to remove any remaining non-viral

contigs and perform taxonomic annotation. The final output is the above-mentioned VirusHunter output, the viral contigs generated from the assembly in VirusGatherer, and a VirusGatherer ‘hittable’ with the best results for the taxonomic annotation and assembly (Figure 2, part 5).

1.3.4 Comparison to other approaches

The RNA-dependent RNA polymerase (RdRp) is a conserved protein found in all RNA viruses, linking them evolutionarily¹⁹². Because of its presence in all RNA viruses, computational tools such as Serratus search for RdRp in sequences to distinguish RNA viruses from non-RNA virus sequences^{102,110,112,114}. This enables the identification of both known and novel RNA viruses and helps monitor and anticipate for potential health crises arising from RNA viruses, at the cost of not being able to detect DNA viruses. In contrast, the VirusHunter and VirusGatherer pipeline is designed to detect both RNA and DNA viruses by aligning query sequences with dedicated profiles for each virus type. Serratus, with its ability to process 5.7 million samples, appears to handle more samples than previous works utilizing VirusHunterGatherer. However, it relies on Amazon Web Services (AWS), incurring a cost of 2,350 USD per petabase and requiring AWS infrastructure¹⁰². In contrast, VirusHuntergatherer can be installed and run on non-commercial high-performance computing infrastructure, avoiding the need for AWS and associated costs. Notably, Serratus exhibits decreased sensitivity when the sequence identity between RdRp motive and query sequences falls below 60 %, whereas earlier studies using VirusHunterGatherer successfully identified divergent viruses with as little as 35% protein sequence identity^{100,126–128}. VirusHunter can be time-consuming due to its reliance on multiple alignments against both viral protein/nucleotide sequence databases and profile models (Figure 2, part 3). VirusGatherer, in contrast, reduces the assembly time compared to conventional *de novo* assemblers by focusing on assembling only viral contigs based on seed sequences, rather than attempting to assembly every query sequence. On top of that, VirusGatherer’s use of Genseed-HMM¹⁸⁶, unlike other seed-based assemblers, may enable the assembly of longer contigs beyond the conserved domain used in homology searches (for example RdRp), while also generating fewer chimeric sequences¹⁹³. Generating clear and accessible reports should also be a high priority to make virus discovery data both comprehensible and useful for informed decision-making and

further application. While DAMIAN¹¹⁵ provides tabular reports designed for easy interpretation by diagnosticians, and Serratus¹⁰² offers graphical reports via a web interface for users without programming expertise, VirusHunter and VirusGatherer's hittables may require additional downstream analysis, such as data visualization and summary statistics, which can be challenging for people lacking skills in programming and statistics.

1.4 Aim of thesis

The aim of this thesis is to identify known and novel viruses in two datasets: a human clinical sequencing dataset and a public mammalian dataset. For this, I applied the VirusHunterGatherer pipeline, which conducts a sensitive homology search to detect viral sequences from raw sequencing data and assembles longer viral contigs.

A dataset was provided by the Taubert working group from the Department of Gastroenterology, Hepatology, Infectious Diseases, and Endocrinology at Hannover Medical School (MHH). This dataset comprises sequencing data from liver biopsies of liver transplant patients, where the samples indicate either no rejection or rejection mediated by T-cell or antibody responses. The complete VirusHunter and VirusGatherer pipeline was applied to this dataset, covering all three viral groups—small DNA, large DNA, and RNA viruses—to detect viral presence and assemble viral contigs from the liver sequencing data. The primary focus of this analysis is to identify viruses present in the patient samples and to assess whether these identified viruses are associated with transplant rejection and, if so, whether they can be linked to one of the two types of rejection: T-cell-mediated or antibody-mediated responses.

The Computational Virology Research Group also supplied an existing VirusHunter hittable, generated from a previous search for RNA viruses in public sequencing data from mammalian samples. In this thesis, a selected subset of virus-positive SRA entries from that hittable is utilized as the second dataset, and longer viral contigs are assembled with VirusGatherer, which has not been done before. Both the provided hittable and the analyzed subset represent a substantial collection

of public unprocessed sequencing data. The assembly of longer viral contigs allows for the potential discovery of both known and novel viruses. To ensure that the identified contigs are indeed viral in nature, functional annotation of the viral proteins is performed for the ten longest viral contigs identified.

Finally, one of the outcomes of this master's thesis is the development of the R¹⁹⁴ package Virusparies¹⁹⁵. Virusparies provides functions to subset and process hittables, calculate summary statistics, and create plots and graphical tables for VirusHunter and VirusGatherer data.

2 Material and Methods

2.1 Code Availability Statement

All code, programs, and scripts developed for this study are available through the 'Computational Virology Research Group,' led by Jun. Prof. Dr. Chris Lauber. The VirusHunterGatherer tool, used in this study, can be accessed on GitHub at <https://github.com/lauberlab/VirusHunterGatherer>. An R¹⁹⁴ package, named 'Virusparies'¹⁹⁵, developed during this thesis for processing and visualizing VirusHunterGatherer output and generating summary statistics, is also available at <https://github.com/SergejRuff/Virusparies>.

2.2 Data Availability Statement

All data generated in this study, including the VirusHunter and VirusGatherer output tables, are publicly available in a dedicated GitHub repository dedicated to this master's thesis. This repository contains all datasets used in the analysis and the results presented in this work. The repository can be accessed at <https://github.com/SergejRuff/MasterThesis>.

2.3 Software used

Viral discovery from raw and unprocessed sequencing data was first performed via VirusHunter, which conducts a homology search against profile Hidden Markov Models¹¹⁸ (pHMMs) of proteins specific to a virus group to identify potential viral sequences. Following this, VirusGatherer then assembles the identified viral sequences (micro-contigs) into complete viral genomes or larger yet incomplete viral contigs. The 'Aeternitas' computing cluster at Twincore was utilized to concurrently execute VirusHunterGatherer across multiple runs. VirusHunterGatherer is implemented in Perl¹⁹⁶ but was executed using Snakemake, a Python-based workflow management system¹⁹⁷. Each pipeline component was represented by a rule managed by Snakemake. A configuration file in YAML format specified paths for filter databases and input data on the 'Aeternitas' server. Both local FASTQ files and a list of SRA accessions were used as input, with the

configuration file allowing optional filtering of host sequences. All dependencies required for running VirusHunterGatherer are listed (Table 8 in Appendix).

2.4 Sequencing data

Two datasets were analyzed for viral discovery in large DNA, small DNA, and RNA viruses, with each dataset sourced from different research groups. All samples in these two sets were derived from distinct viral host organisms.

2.4.1 Taubert liver transplant data

Patient liver sequencing data from the Taubert group (Department of Gastroenterology, Hepatology, Infectious Diseases, and Endocrinology at MHH) was provided locally, encompassing 11 folders with a total of 323 samples distributed across 557 FASTQ files (Table 1).

Table 1: Taubert sequencing data was provided locally in 11 folders, consisting of 323 samples distributed unevenly across these folders. For some samples, Read 1 and Read 2 were stored in separate FASTQ files, resulting in a total of 557 FASTQ files. Analysis was completed for small DNA, large DNA, and RNA viruses for all datasets.

Taubert Data				
Folder	Number of Samples	Completed		
		Small DNA viruses	Large DNA viruses	RNA viruses
15-0001	36	✓	✓	✓
16-0149	25	✓	✓	✓
16-0271	22	✓	✓	✓
17-0238	32	✓	✓	✓
18-0199	48	✓	✓	✓
18-0219	35	✓	✓	✓
18-0220	27	✓	✓	✓
19-0130	27	✓	✓	✓
20-0055	45	✓	✓	✓
20-0056	4	✓	✓	✓
20-0057	22	✓	✓	✓

2.4.2 Mammalian data

The Computational Virology Research Group provided an existing VirusHunter

hittable with data analyzed from January to March 2023. No viral contig assembly via VirusGatherer was performed on this dataset. It consisted of 34,337 unique SRA accessions, containing viral reads detected in mammalian hosts. Unlike the Taubert dataset, this data was downloaded from NCBI. The dataset was filtered for significant viral matches, requiring an E-value below $1e^{-5}$ and viral sequence identity less than 90%, ensuring novel virus discovery. A minimum of four reads was required for contig assembly during the VirusGatherer stage. After filtering, 1,666 unique SRA accessions with 29 profiles remained (Figure 3).

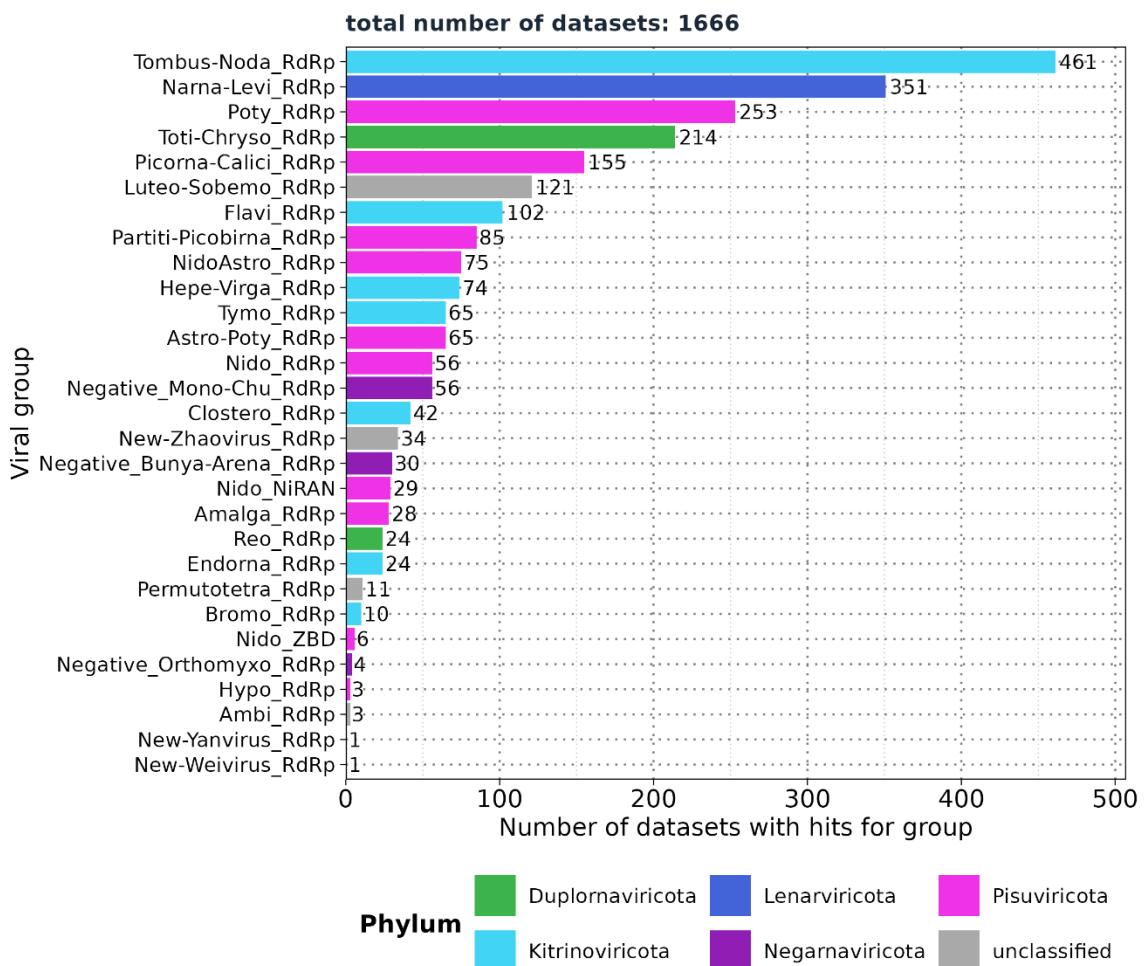


Figure 3: Mammalian sequencing data - Distribution of viral groups detected across query sequences for all profiles identified after filtering the existing hittable. 1,666 unique SRA accessions with 29 profiles were found after applying filters for E-value < $1e^{-5}$, number of reads > 4, and viral sequence identity against the viral BLAST reference < 90%.

Eight profiles of interest were selected for further analysis: Flavi_RdRp (found in 102

out of 1,666 SRA accessions), NidoAstro_RdRp (75), Hepe-Virga_RdRp (74), Nido_RdRp (56), Negative_Mono-Chu_RdRp (56), Negative_Bunya-Arena_RdRp (30), Nido_NiRAN (29), and Negative_Orthomyxo_RdRp (4). These profiles were reanalyzed using VirusHunter, focusing on RNA virus detection, and the resulting micro-contigs were subsequently assembled into larger contigs using VirusGatherer for the first time in this thesis.

2.5 VirusHunter

VirusHunter was employed to identify viral sequences in raw sequencing datasets using a sensitive homology search with pHMMs. This section outlines the steps involved in VirusHunter implementation and use.

2.5.1 SRA download

VirusHunterGatherer supports both locally available FASTQ files and a list of SRA accessions as input files. If only locally available FASTQ files is provided, the FASTQ files are used as directly. Otherwise, SRA accession numbers are processed using the SRA Toolkit (with the prefetch and fastq-dump tools) to download sequencing data from the National Center for Biotechnology Information (NCBI) SRA database⁹⁸. Both types of input were used in this thesis: Taubert's data was available locally, while mammalian data was downloaded from NCBI SRA database.

2.5.2 Preprocessing of FASTQ files

Trimming of sequencing adapters and removal of low-quality bases were performed using fastp¹⁹⁸. Adapter sequences were auto detected and removed, and reads shorter than 15 bases or containing more than 5 'N' bases were discarded. Paired-end reads were combined into a single FASTQ file. The processed files were converted to FASTA format using seqtk¹⁹⁹. The sequences were then translated into protein sequences across six reading frames using transeq²⁰⁰.

2.5.3 Sensitive homology-based detection of viral sequence reads in unprocessed data

Detection of viral reads in unprocessed sequencing data was accomplished using the hmmsearch tool¹¹⁷, which allows for the identification of viral sequences

by querying a set of pHMMs. The protein sequences translated from each of the six reading frames were applied in six separate HMMer searches, each against query pHMMs specific to protein domains associated with RNA, small DNA, or large DNA viruses. Profiles for RNA viruses predominantly featured RNA-dependent RNA polymerase (RdRp) domains¹⁹². However, for hepatitis delta virus and nidoviruses, the profile included hepatitis delta antigen (HDAg)^{201,202} domains for hepatitis delta virus and RNA-dependent RNA-polymerase-associated Nucleotidyltransferase (NiRAN)²⁰³ domain for nidoviruses. For large DNA viruses, profiles mainly comprised major capsid protein (MCP) domains^{204,205}, whereas small DNA viruses queried mostly against the replication initiator protein (Rep) domains²⁰⁶. A complete list of all profiles is provided (Table 9, 10, and 11 in Appendix). The size of the target database for E-value calculations was set to six times the number of reads, accounting for each reading frame. An E-value of 10 was applied to report hits in the sequence search. The significance of each search result was then assessed using two criteria: the number of hits and the minimum E-value observed. If the total number of hits met or exceeded a threshold of two, or if the globally best E-value was below 0.01, the hits were considered significant. Score, E-value, and query profile for each significant hit were then exported. FASTA files with non-significant hits were deleted and not used further.

2.5.4 Filtering against contaminants and viral reference sequences

Further filtering was performed only on data with significant hits. First, vsearch²⁰⁷ deduplicated hits by removing sequences that had identical length and nucleotide content. Deduplicated reads were then assembled to micro contigs via CAP3 assembler¹⁸⁷. Only those sequences that had a minimum overlap length of 20 nucleotides and a maximum overhang percentage length of 75 % were considered. The overhang percentage is calculated by dividing the total length of all overhangs by the length of the overlap, then multiplying by 100. As part of the master's thesis, modifications were made to the VirusHunter pipeline to make the filter against host sequences optional. Consequently, only the filters against custom contaminants and the viral database were applied. Filtering against custom contaminants was performed using BLASTN. The contaminant database included ubiquitin sequences, and any matches with an E-value above $1e^{-4}$ were excluded.

Afterwards, the remaining contigs were filtered against a viral database via TBLASTX¹²⁰, and only matches with an E-Value lower than one were retained.

2.5.5 Output of VirusHunter

VirusHunter exported the sequences of the filtered contigs in a compressed FASTA file. Furthermore, results for hits found at each step of the pipeline were documented in separate hittables files for each sequencing experiment. These individual results were then compiled into a final VirusHunter hittable. A VirusHunter hittable contained 11 to 15 characteristic columns, depending on whether the input sequencing data was local or downloaded from the NCBI SRA database (Table 2).

Table 2: Columns of the VirusHunter hitable , including data types (chr' for character, 'num' for numeric, and 'date' for date values in ISO 8601 format) and three representative example values for each column. The 'run_id' column (yellow) is specific to locally processed input data. In contrast, columns related to SRA identifiers and host information are exclusive to SRA data downloaded from NCBI (violet). For each viral read detection, the best HMM profile match, its E-value, and the number of hits are recorded in separate columns. After comparing contigs against viral reference database, the best-matching viral subject and taxonomy, the best E-value, and the percentage of sequence identity with the viral reference are documented in their respective columns. Lastly, the date of analysis is also included.

Column	Type	Examples
run_id	<chr>	21320_13_S5_L002_R1_001, 22642_12_S4_R1_001, 10014-14_R1
SRA_run	<chr>	DRR248913, SRR12507950, SRR14180567
SRA_sample	<chr>	DRS256618, SRS7251525, SRS8660194
SRA_study	<chr>	DRP009016, SRP107302, SRP313939
host_taxon	<chr>	Macaca fascicularis, Tursiops truncatus, Ovis aries
host_taxid	<num>	9541, 9739, 9940
num_hits	<num>	65, 1, 1
best_E	<num>	6e-05, 2.3, 0.00018
best_query	<chr>	Flavi_RdRp, Negative_Bunya-Arena_RdRp, Flavi_RdRp
ViralRefSeq_E	<num>	1.3e-37, 0.38, 3.8e-21
ViralRefSeq_ident	<num>	80.5, 73.3, 81.6
ViralRefSeq_aLen/sLen	<chr>	231 / 234, 45 / 125, 147 / 150
ViralRefSeq_contigs	<num>	3, 1, 1
ViralRefSeq_subject	<chr>	NC_031950.1 Guereza hepacivirus, complete sequence, KY288905.1 Zika virus strain M genome, NC_077023.1 Pestivirus sp. isolate ovine/It/338710-3/2017 polyprotein (QKUO2 complete cds
ViralRefSeq_taxonomy	<chr>	taxid:1354498 Hepacivirus colobi Hepacivirus Flaviviridae Amarillovira taxid:64320 Orthoflavivirus zikae Orthoflavivirus Flaviviridae Amarillovira taxid:31657 Pestivirus Flaviviridae Amarillovira Flasuviricetes Kitrinoviricota Ortho
date_analyzed	<date>	2024-06-24, 2024-06-24, 2024-06-24

For Data downloaded from NCBI, the hittable included columns with identifiers for each SRA run, sample, and study identifier. Additionally, the hittable featured columns with details about the host taxon and the host taxon identifier. For

each SRA run, the best match against the viral reference database was documented, providing details such as the viral reference taxonomy name, the percentage of sequence identity with the viral reference, and the E-value of the best match. The E-value and the name of the best HMM profile match were also documented in separate columns, along the number of viral contigs identified for each SRA run.

For locally processed data, the hittable omitted the columns related to SRA identifiers and host information, replacing them with a single column for the Identifiers of the local FASTQ files.

2.6 VirusGatherer

The identified viral contigs found in the VirusHunter stage serve as seeds for the progressive assembly of full-length viral genome sequences. This section details how VirusGatherer was utilized in this work.

2.6.1 Assembly of Contigs

For VirusGatherer, fastp¹⁹⁸ was used to trim adapters and remove low-quality bases from each unprocessed FASTQ file, which was then converted to FASTA format. Unlike the VirusHunter stage, no translation into protein for each reading frame was performed. The viral contigs were assembled into larger contigs via CAP3¹⁸⁷, considering only sequences with an overhang percentage of 99% or less, and a minimum overlap length of 20 bases for assembly. These were then utilized as seeds for a targeted assembly of viral contigs found in the FASTQ files, with the help of a modified version of Genseed-HMM¹⁸⁶. The progressive assembly was performed using CAP3, with a maximum of 100 iterations. Contig ends with a length of 30 were used for further assembly, and the maximum contig length was set to 50.000. For CAP3, the minimum percentage identity for an overlap was set to 85%, with the minimum overlap length required being 20 nucleotides. Additionally, the maximum allowable overhang was set to 99%. For the initial similarity search, BLASTN¹²⁰ with an E-value cutoff of 1e⁻³, a percentage identity of 85%, and a word size of 7 was applied. Lastly, the modified version of Genseed-HMM performed deduplication in each iteration to decrease CAP3 assembly time.

2.6.2 Compare against viral reference database

Assembled contigs were compared against NCBI's database of reference viral proteins using BLASTX¹²⁰. For the BLASTX queries, the maximum number of reported high-scoring segment pairs (HSPs)²⁰⁸ – which are local alignments with the highest scores and no gaps – was set to one. Additionally, the maximum number of target sequences returned per query was also set to one.

2.6.3 Output of VirusGatherer

VirusGatherer returned the sequences of the assembled viral contigs/ whole genomes. Furthermore, hittables similar to the VirusHunter hittables were generated (Table 3). Virusgatherer hittables also contained columns related to SRA identifiers and host information, or the identifiers for the locally available FASTQ files, depending on the input used. Unique columns in VirusGatherer hittables included viral contig name and its length, as well as details of the best viral match from the viral reference database comparison.

Table 3: Columns of the VirusGatherer hittable , including data types (‘chr’ for character, ‘num’ for numeric, and ‘date’ for date values in ISO 8601 format) and three representative example values for each column. The ‘run_id’ column (yellow) is specific to locally processed input data. In contrast, columns related to SRA identifiers and host information are exclusive to SRA data downloaded from NCBI (violet). The contig identifier and its length after assembly are reported. After comparing contigs against viral reference database, the best-matching viral subject and taxonomy, the best E-value, and the percentage of sequence identity with the viral reference are documented in their respective columns. Lastly, the date of analysis is also included.

Gatherer Hittable Structure		
Column	Type	Examples
run_id	<chr>	21320_13_S5_L002_R1_001, 10014-14_R1, 10014-14_R1
SRA_run	<chr>	ERR10568066, ERR10569187, SRR13364366
SRA_sample	<chr>	ERS14300147, ERS14300285, SRS7974787
SRA_study	<chr>	ERP143068, ERP143068, SRP300585
host_taxon	<chr>	Sus scrofa, Sus scrofa, Manis javanica
host_taxid	<num>	9823, 9823, 9974
contig_id	<chr>	ERR10568066_cap3_Contig-1, ERR10569187_cap3_Contig-2, SRR13364366_cap3_Conti
contig_len	<num>	1903, 987, 486
ViralRefSeq_E	<num>	0.000681, 1.29e-45, 4.16e-15
ViralRefSeq_ident	<num>	29.762, 35.685, 35.398
ViralRefSeq_aLen	<num>	84, 241, 113
ViralRefSeq_subject	<chr>	acc:YP_004821526 MHC class I protein [Yokapox virus], acc:YP_010056903 RNA-binding
ViralRefSeq_taxonomy	<chr>	acc:YP_073558 RNA-dependent DNA polymerase [lymphocystis disease virus-China]
ViralRefSeq_taxonomy	<chr>	taxid:1076255 Centapoxvirus Chordopoxvirinae Poxviridae Chitovirales Pokkesviricetes
ViralRefSeq_taxonomy	<chr>	taxid:2027899 Myranavirus phabba Myrnavirus Ceeclamvirinae Caudoviricetes Uroviric
ViralRefSeq_taxonomy	<chr>	taxid:256729 Lymphocystis disease virus
ViralRefSeq_taxonomy	<chr>	2 Lymphocystivirus Alphairidovirinae Iridoviridae Pimascovirales Megaviricetes Nucleo
date_analyzed	<date>	2024-06-26, 2024-06-26, 2024-06-30

2.7 Virusparies

One of the outcomes of this master's thesis was the development of the R¹⁹⁴ package Virusparies¹⁹⁵. Virusparies provides functions to subset and process hittables, calculate summary statistics, and create plots and graphical tables for VirusHunter and VirusGatherer hittables. Both import of hittables into R and export of results in a user-specified file format was handled by Virusparies. Virus family names were extracted from the 'ViralRefSeq_taxonomy' column for each observation of the hittables via the VhgPreprocessTaxa function. Where no family name is present, but it is possible to infer the phylum from other information in the 'ViralRefSeq_taxonomy' column, 'unclassified' followed by the phylum name was used. If inferring the phylum was not possible, only 'unclassified' was assigned to the observation. The processed taxonomy data was subsequently used to group data for plots and summary statistics calculations. Boxplots were generated to visualize the distribution of E-values ('ViralRefSeq_E'), identity percentages ('ViralRefSeq_ident'), and contig lengths ('contig_len') for each group. The sum of hits for each virus group and the distribution of viral groups detected across query sequences were plotted in bar charts. Crucially, each dataset was filtered to include only observations with an E-value of 1e⁻⁵ or lower in the 'ViralRefSeq_E' column before plotting. When E-values were plotted, the negative logarithm of the threshold served as a cutoff line instead, and no filtering was applied. In some cases where E-values were visualized, E-values of exactly zero resulted in infinite values when transforming to their negative logarithm. To address this, all E-values of zero were replaced with the smallest E-value greater than zero. If the smallest E-value was above the cutoff (1e⁻⁵), zeros were replaced with the cutoff multiplied by ten raised to the power of negative ten. The mode, median, mean, standard deviation, and first (Q1) and third (Q3) quartiles were calculated for viral E-values, identity percentages, and contig lengths, and the results were summarized in tables.

2.8 Prediction and visualization of predicted functional domains

For the viral screening of mammalian sequencing data, the three longest contigs from each RNA virus phylum were chosen to examine their predicted functional domains. The analysis utilized contigs assembled by VirusGatherer. Protein domains for each contig were predicted with InterProScan²⁰⁹ by comparing the

assembled sequences against the InterPro database²¹⁰, using an E-value threshold of $< 1e^{-5}$ for filtering. To ensure the most relevant matches, the protein match with the lowest E-value was selected, as multiple matches from different databases often corresponded to the same protein. Contigs with an open reading frame (ORF) comprising less than half of the sequence length were excluded from further analysis, leaving only 10 contigs. The length of each protein domain was multiplied by 3 to reflect the nucleic acid sequence length since VirusGatherer returns nucleic acid output. The selected viral genomes were visualized using the gggenomes R package (version 1.0.1)²¹¹.

3 Results

This study examined two datasets, comprising liver sequencing data from patients post-liver transplantation (Taubert sequencing data) and mammalian sequencing data obtained from the Sequence Read Archive. I employed VirusGatherer to investigate whether viral sequences were present in the liver sequencing data and if they might be associated with liver rejections. The viral hits from the mammalian sequencing data were assembled for the first time using VirusGatherer, and the top ten longest contigs were then subjected to functional annotation to confirm their viral characteristics.

3.1 Characterization of the virome after liver transplantation

In this study, I employed the VirusHunter and VirusGatherer pipeline to characterize the virome of liver sequencing data. This data, provided by the Taubert working group, included 323 patient samples across 557 FASTQ files. VirusHunter performed a sensitive sequence homology search within this data using profile Hidden Markov Models based on conserved protein sequences typical of for RNA, large DNA, and small DNA viruses (Table 9, 10, and 11 in Appendix).

The sequences (micro-contig) identified during the screening acted as seeds in the VirusGatherer step, enabling the progressive assembly of longer contigs up to complete viral genomes. Both pipeline steps included a pairwise comparison of contigs against a viral database and produced tabular output capturing the best results from screening, assembly, and comparison. I incorporated this output into Virusparies¹⁹⁵, an R¹⁹⁴ package developed for this master thesis, to visualize the results and generate summary statistics.

3.1.1 The majority of hits identified in the liver sequencing data are not significant

Using the described approach, I obtained five FASTQ files corresponding to four samples that matched at least once against a small DNA virus profile in the VirusHunter step, representing only 0.89% of the total input files. For large DNA virus profiles, I retrieved 165 files (29.62%), corresponding to 142 samples, and for RNA viruses, there were 17 files (3.05%), totaling 13 samples. The output from VirusHunter and VirusGatherer was deemed significant only when the comparison

Results

against the viral reference database yielded an E-value smaller than $1e^{-5}$. In the run against small DNA virus profiles, the micro-contigs aligned with three viral families in the reference database; however, none of these alignments had an E-value below $1e^{-5}$ and, therefore, could not be classified as significant (Figure 4, part A).

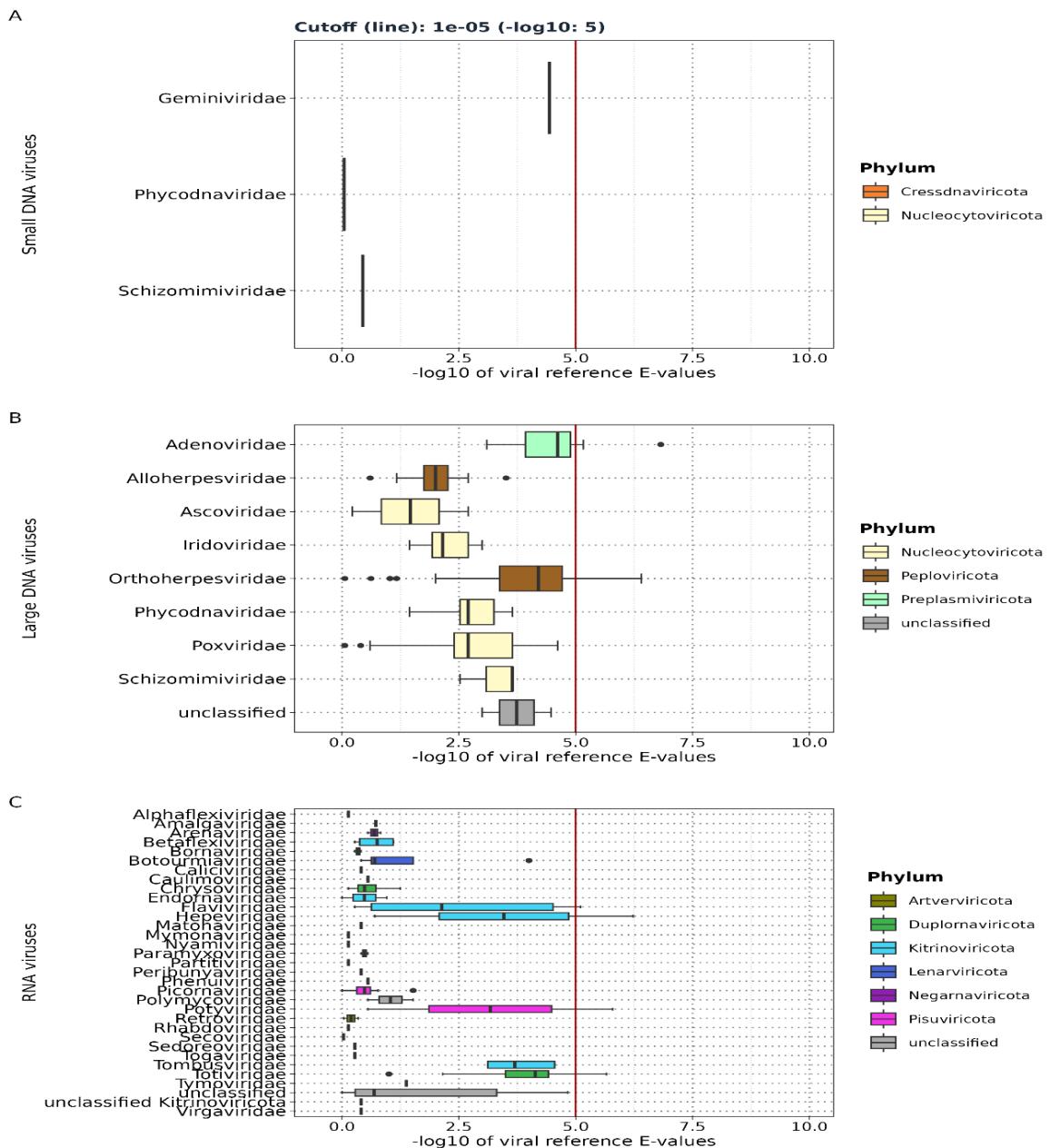


Figure 4: Taubert sequencing data - Distribution of viral reference E-values for each viral family detected by VirusHunter runs against profiles for small DNA viruses (A), large DNA viruses (B), and RNA viruses (C). E-values from contig alignment against viral reference database are transformed to their negative logarithm base 10. The vertical red line in the plot marks the $1e^{-5}$ cutoff: values right of the line ($E\text{-value} < 1e^{-5}$) are significant, while those on the left ($E\text{-value} > 1e^{-5}$) are not.

VirusHunter identified eight large DNA virus families, along with one unclassified group for which no taxonomic information was available (Figure 4, part B). Among them, only the alignments against two families returned E-value below $1e^{-5}$: *Adenoviridae*¹⁶⁴ from the phylum *Preplasmaviricota* and *Orthoherpesviridae*¹⁵⁹ from the phylum *Peploviricota*. For RNA viruses, 31 families matched, along with one unclassified family from the phylum *Kitrinoviricota* and one unclassified group without phylum information (Figure 4, part C). Statistically significant alignments were observed against the families *Flaviviridae*¹⁵² and *Hepeviridae*¹⁵³ from the phylum *Kitrinoviricota*, as well as *Potyviridae*²¹² from the phylum *Duplornaviricota*, and *Totiviridae*²¹³ from *Pisuviricota*.

3.1.2 Only a small fraction of the initial dataset detected possible viral contigs

Next, I focused on the number of files (and samples) that yielded significant matches ($E\text{-value} < 1e^{-5}$) during the alignment against the viral database, as documented in the VirusHunter hittable. As noted earlier, screening for small DNA viruses returned no significant hit. The search for small DNA viruses in the Taubert sequencing data yielded results from 3 out of 165 files (1.81%) matching *Adenoviridae* and 8 files (4.85%) matching *Orthoherpesviridae* (Figure 5, part A). The number of samples matched the number of files, with only 11 remaining files having at least one alignment deemed statistically significant. This represented 1.97% of the original 557 files and amounts to just 3.4% of the 323 samples. Only four files contained micro-contigs identified as potential RNA viruses, with the number of files corresponding to the number of samples. Each of these files/samples aligned with a distinct RNA virus family (Figure 5, part F). This accounted for 0.71% of the original files and 1.23% of the samples with potential RNA virus sequences. The assembly of micro-contigs in VirusHunter depends on adequate read coverage. These micro-contigs may subsequently serve as seeds for assembling larger contigs. In the patient liver sequencing data, I observed relatively few reads aligning to specific viral groups. For large DNA viruses, 18 reads aligned with *Orthoherpesviridae*, while 5 reads aligned with *Adenoviridae* (Figure 5, part B). RNA viruses exhibited a higher number of reads, with *Totiviridae* and *Potyviridae* exceeding the read counts for *Orthoherpesviridae* and *Adenoviridae* (Figure 5, part E). Among RNA viruses, *Kitrinoviricota* showed the fewest reads. All viral micro-contigs in the large DNA virus

screening showed an exact match to known viruses (Figure 5, part C), while *Totiviridae* was the only family with a protein sequence identity below 90% (Figure 5, part E).

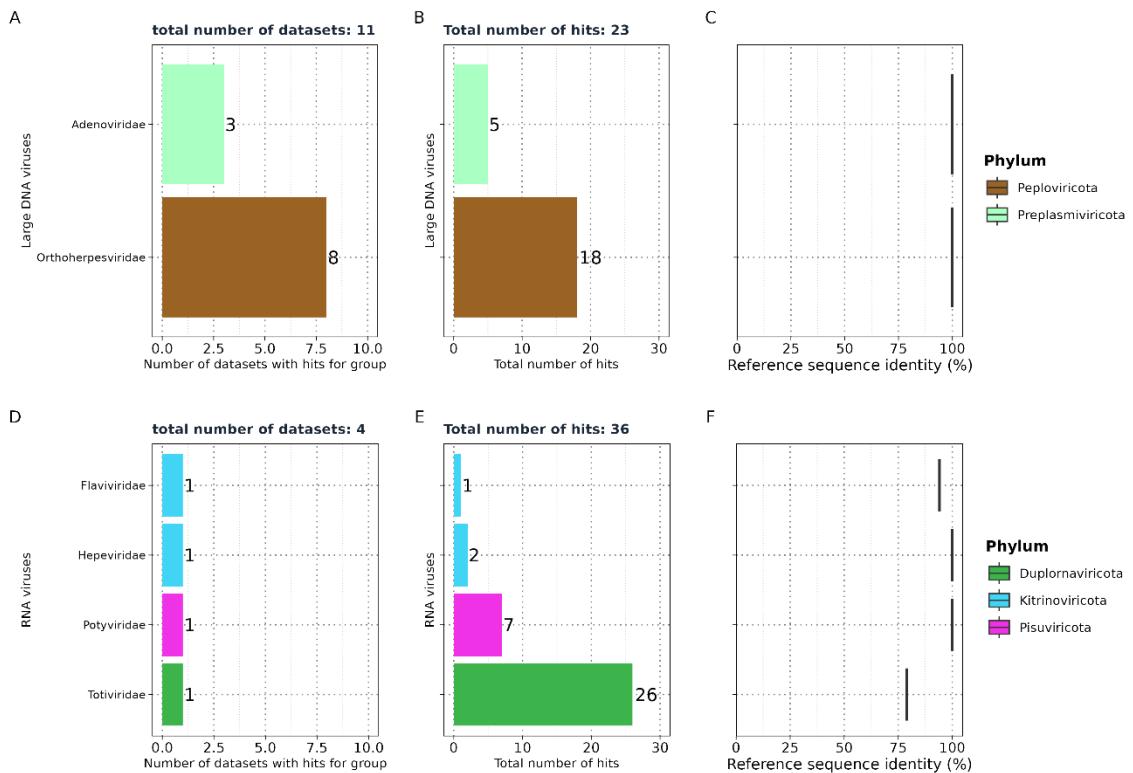


Figure 5: Taubert sequencing data – VirusHunter results. A, D: Distribution of the number of files detecting viral families across large DNA, and RNA viruses. B, E: Number of reads per virus family (labelled left) in VirusHunter. C, F: Protein sequence identities to the nearest known reference virus for each viral family (labelled left).

3.1.3 All contigs identified in the liver sequencing data are short

VirusGatherer assembled longer viral contigs by leveraging viral sequences identified in VirusHunter as seeds. It aligned these sequences in an internal pair aligner against a viral database, selecting only contigs with an E-value smaller than $1e^{-3}$ for assembly. This relaxed E-value criterion, compared to the significance cutoff of $1e^{-5}$, enabled the inclusion of viral hits that were not deemed significant in VirusHunter. Consequently, the assembly of small DNA viruses produced a contig aligning with the family *Hepadnaviridae*¹⁸³ from the phylum *Artvervirocota* (Figure 6,

Results

part A), exhibiting a protein sequence identity of 95% (Figure 6, part B) and a contig length of 70 (Figure 6, part C), despite no statistically significant alignments against a viral reference sequence in the VirusHunter step. *Flaviviridae*, *Potyviridae*, and *Adenoviridae*, detected by VirusHunter (Figure 5), were not found by VirusGatherer (Figure 6). Additionally, while both *Totiviridae*, *Hepeviridae*, and *Orthoherpesviridae* were identified as significant matches in VirusHunter and also appear in Gatherer, contigs also aligned with the families *Poxviridae*¹⁶³ and *Kolmioviridae*²¹⁴ (Figure 6, part D,G).

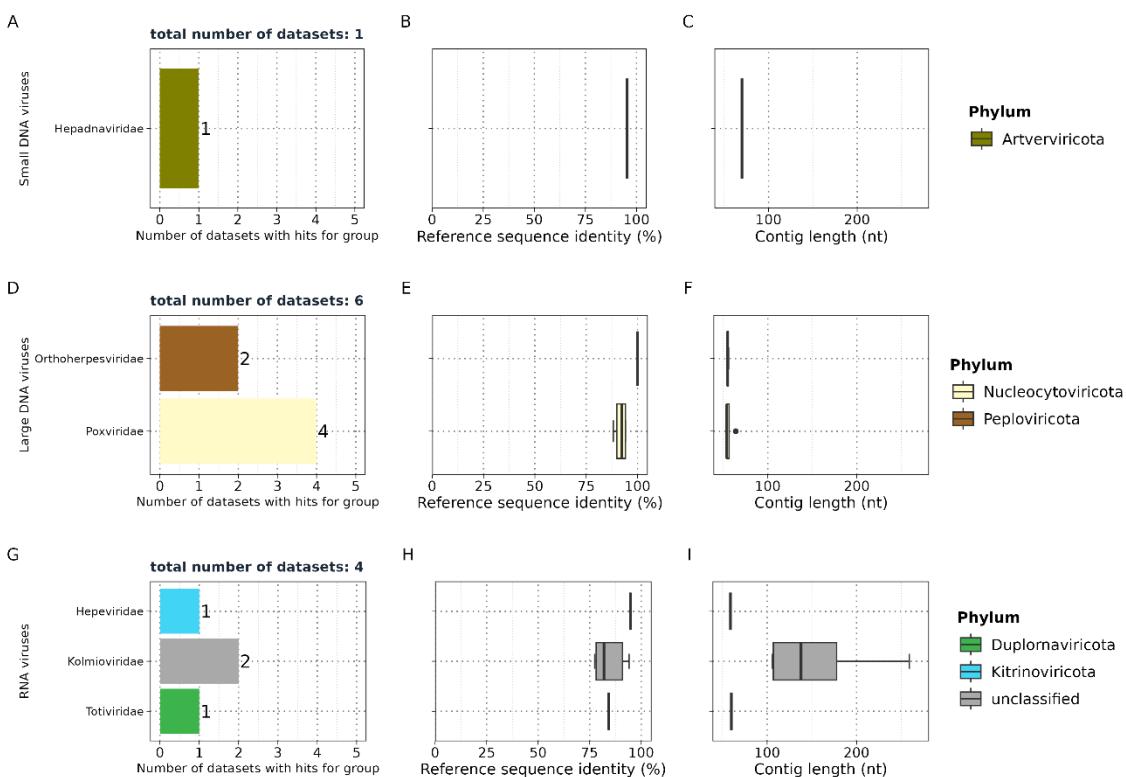


Figure 6: Taubert sequencing data – VirusGatherer results. A, D, G: Distribution of the number of files detecting viral families across large DNA, small DNA, and RNA viruses. B, E, H: Protein sequence identities to the nearest known reference virus for each viral family (labelled left). C, F, I: Length of the contigs in each viral family (labelled left).

The alignment of 133 micro-contigs to *Poxviridae* surpassed that of *Orthoherpesviridae* (76 micro-contigs without filtering), but none of these alignments

were considered significant, as they did not meet the cutoff criteria. The *Kolmioviridae* family, which includes hepatitis D viruses, was not detected in VirusHunter but appeared in VirusGatherer. The contigs aligning with *Kolmioviridae* in VirusGatherer did not have significant hits in VirusHunter, with a median of 0.39 (IQR: 0.28 to 0.55, min/max = 0.03/1). Across six datasets, six contigs were identified as potentially representing large DNA viruses, while four files revealed ten contigs associated with RNA viruses. Out of the 17 viral contigs, seven (41.17%) displayed less than 90% protein sequence identity to the closest known virus (Figure 6, part B, E, H; table 12 in Appendix). All contigs were shorter than 300 nucleotides, with the largest measuring 259 nucleotides and aligning with the *Kolmioviridae* family, while the shortest contig was 59 nucleotides, associated with *Orthoherpesviridae* (Figure 6, part C, F, I; table 13 in appendix).

Although VirusHunterGatherer categorized the contigs into various viral families using human liver sequencing data as input, the closest related viruses appeared to be animal-infecting viruses (Table 4).

Table 4: Taubert sequencing data - Closest viral family and subject found among large DNA, small DNA, and RNA viruses , with the number of contigs ('Count' column) aligned to each viral subject.

Taubert - Found subjects		
Viral reference taxonomy	Viral reference subject	Count
Large DNA Viruses		
Orthoherpesviridae	Human betaherpesvirus 5	1
Orthoherpesviridae	Panine betaherpesvirus 2	1
Poxviridae	Cotia virus SPAn232	1
Poxviridae	Eptesipox virus	2
Poxviridae	Skunkpox virus	1
Small DNA Viruses		
Hepadnaviridae	Hepatitis B virus	1
RNA Viruses		
Hepeviridae	Paslahepevirus balayani	1
Kolmioviridae	Hepatitis delta virus	3
Kolmioviridae	Snake deltavirus F18-5	3
Kolmioviridae	hepatitis D virus 1	2
Totiviridae	Saccharomyces cerevisiae virus L-BC (La)	1

Among the potential RNA viruses, nine contigs aligned with hepatitis D or hepatitis E viruses (*Paslahepevirus balayani*). The contigs related to hepatitis E showed high protein sequence identity, with a maximum of 94.74%. In contrast, five out of eight contigs aligning with hepatitis D displayed less than 90% sequence identity to the reference sequences. For contigs matching *Orthoherpesviridae*, the closest related viruses were betaherpesvirus 5 (human) and betaherpesvirus 2 (primates), both exhibiting 100% sequence identity. However, both contigs were among the shortest, measuring only 54 and 56 nucleotides, respectively.

3.2 Characterization and assembly of RNA viruses from mammalian samples

A pre-existing VirusHunter hittable with 1,666 experiments, containing hits for RNA viruses from mammalian samples, served as the second data source in this thesis (Material and Methods 2.4.2; Figure 3). Since no assembly has been performed previously, I selected a subset of data matching eight profile Hidden Markov Models. This subset included 19,113 micro-contigs, which VirusGatherer used as seeds for assembly.

3.2.1 Over 6,000 contigs found in mammalian sequencing data

I discovered 8,232 contigs aligning to 128 known families and 6 unclassified viruses (Figure 7; table 14 in Appendix). Of these, 6,209 contigs had alignments against viral reference sequences with an E-value smaller than $1e^{-5}$. *Pisuviricota* exhibited the highest number of significant alignment between contigs and viral reference sequences, with 2,212 out of 2,553 contigs (86.64%) having an E-value below $1e^{-5}$. *Lenarviricota* followed; with 686 out of 773 contig alignments deemed significant (88.75%).

Results

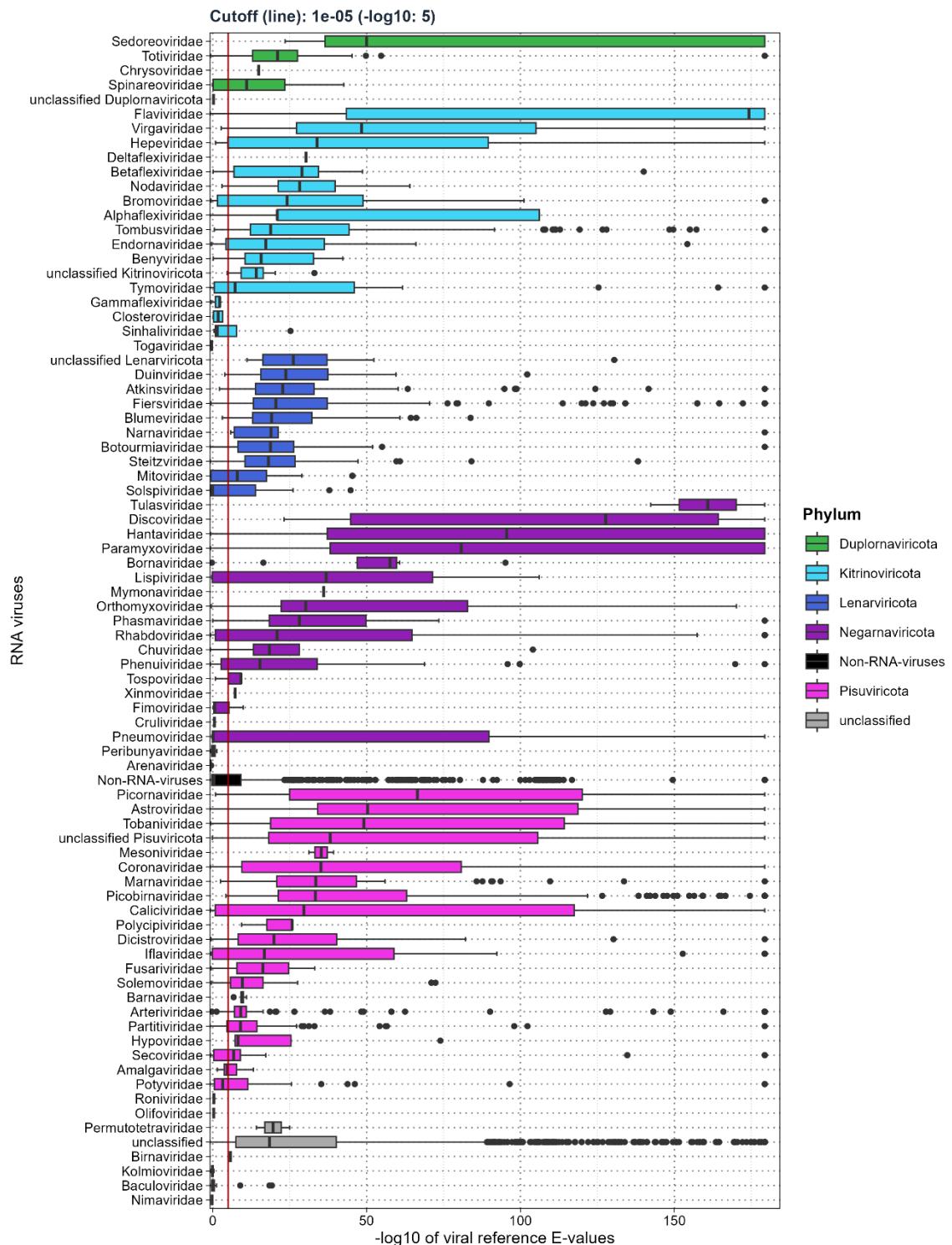


Figure 7: Mammalian sequencing data - Distribution of viral reference E-values for each viral family (VirusGatherer). E-values from contig alignment against viral reference database are transformed to their negative logarithm base 10. The vertical red line in the plot marks the $1e-5$ cutoff: values right of the line ($E\text{-value} < 1e-5$) are significant, while those on the left ($E\text{-value} > 1e-5$) are not.

Next were *Kitrinoviricota* (453 out of 552; 86.78%), *Negarnaviricota* (209 out of 281; 74.37%), and *Duplornaviricota* (115 out of 124; 92.74%). Of the remaining viral contigs, including unclassified contigs, 2,534 out of 3,979 (63.68%) had E-values below $1e^{-5}$ in the alignment against a viral reference database.

3.2.2 VirusGatherer assembled contigs from over 300 SRA experiments, identifying 69 RNA virus families

The assembly generated 6,209 viral contigs with an E-value below $1e^{-5}$ from 332 experiments (Figure 8). Within the phylum *Pisuviricota*, I identified 21 viral families. Dominating this group was *Arteriviridae*²¹⁵, which comprised 1,082 of the 2,212 viral contigs (48.92%) found across 36 of the 332 datasets. *Picobirnaviridae*²¹⁶ followed with 312 contigs (14.10%), and *Astroviridae*²¹⁷ contributed 234 contigs (10.58%). Collectively, these three families accounted for 73.6% of all *Pisuviricota* contigs. Excluding unclassified entries, *Arteriviridae* and *Picobirnaviridae* had the highest counts among all phyla, with *Fiersviridae*²¹⁸ adding 284 viral contigs (41.40%) from a total of 686 *Lenarviricota* contigs. *Steitzviridae*²¹⁹ appeared in 16 samples, contributing 114 contigs (16.61%), while *Blumeviridae*²²⁰ was present in 12 samples with 70 contigs (10.20%). The remaining *Lenarviricota* families each accounted for less than 10%.

In the *Negarnaviricota* phylum, a similar trend emerged, with *Paramyxoviridae*²²¹ leading at 66 contigs (31.57%), followed by *Rhabdoviridae*²²² with 41 contigs (19.62%) and *Phenuiviridae*²²³ with 34 contigs (16.27%). Other families in this phylum contributed less than 10% each. *Totiviridae* was the predominant family within the *Duplornaviricota*, representing 79 of the 115 contigs (70%) found across 22 samples. *Flaviviridae* was the most frequently detected family in *Kitrinoviricota*, appearing in more samples (94) than all other families combined. Despite this prevalence, it accounted for only 38.41% of the total viral contigs (174 out of 453), slightly surpassing *Tombusviridae*²²⁴, which had 136 contigs (30.02%); the remaining families contributed less than 10% each.

Results

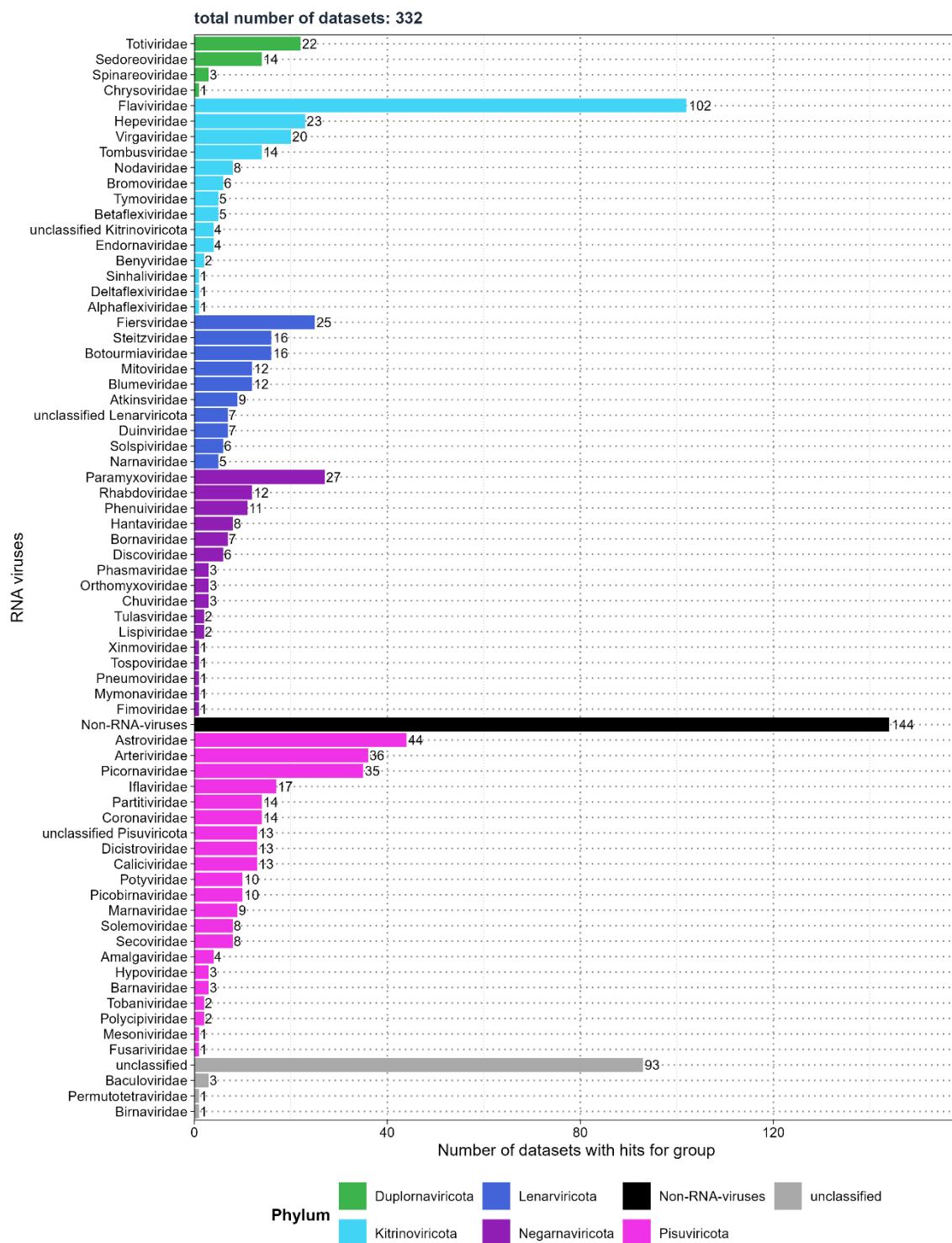


Figure 8: Mammalian sequencing data - Distribution of the number of SRA experiments detecting viral families.

Additionally, 144 samples included 11 families of potential non-RNA viruses (Table 5). *Retroviridae*¹⁸⁴, belonging to the phylum *Artvervirocota* within the kingdom *Pararnavirae*, was detected despite the primary focus of my screening and

assembly on RNA virus families from the kingdom *Orthornavirae*. The remaining families comprised DNA viruses within the RNA virus assembly. The comparison of E-values between non-RNA viruses and RNA viruses revealed differences in their distributions. Non-RNA viruses exhibited lower $-\log_{10}$ median E-values of 0.43 (IQR: -0.3 to 9.13). The median E-value from VirusGatherer prior to transformation was 0.37 (IQR: $7.41e^{-10}$ to 1.99). RNA viruses displayed higher $-\log_{10}$ median E-values of 14.87 (IQR: 6.55 to 38.01), with the original median E-value being $1.34e^{-15}$ (IQR: $9.74e^{-39}$ to $2.85e^{-7}$).

Table 5: Eleven non-RNA viruses identified in mammalian sequencing data during RNA virus screening.

Virus Families	
Family 1-6	Family 7-11
Alloherpesviridae	Poxviridae
Baculoviridae	Retroviridae
Iridoviridae	Schizomimiviridae
Mimiviridae	Steigviridae
Orthoherpesviridae	unclassified Uroviricota
Phycodnaviridae	—

3.2.3 Most of assembled contigs originate from novel viruses

A substantial 80.59% of viral contigs (5,004 out of 6,209) exhibited less than 90% protein sequence identity with their closest known virus, suggesting they come from new viral sequences (Figure 9; table 15 in Appendix). Starting with the highest percentage, the phylum *Lenaviricota* had 97.81% classified as novel, followed closely by *Kitrinoviricota* at 75.72%. *Duplornaviricota* and *Negarnaviricota* ranked third and fourth, with 73.04% and 70.33%, respectively. Although *Pisuviricota*

Results

contained the largest total number of contigs among RNA phyla, only 57.05% were classified as novel. Lastly, 98% of all unclassified contigs displayed protein sequence identities below 90%, and non-RNA viruses had no contigs with identities of 90% or higher (Table 16 in Appendix).

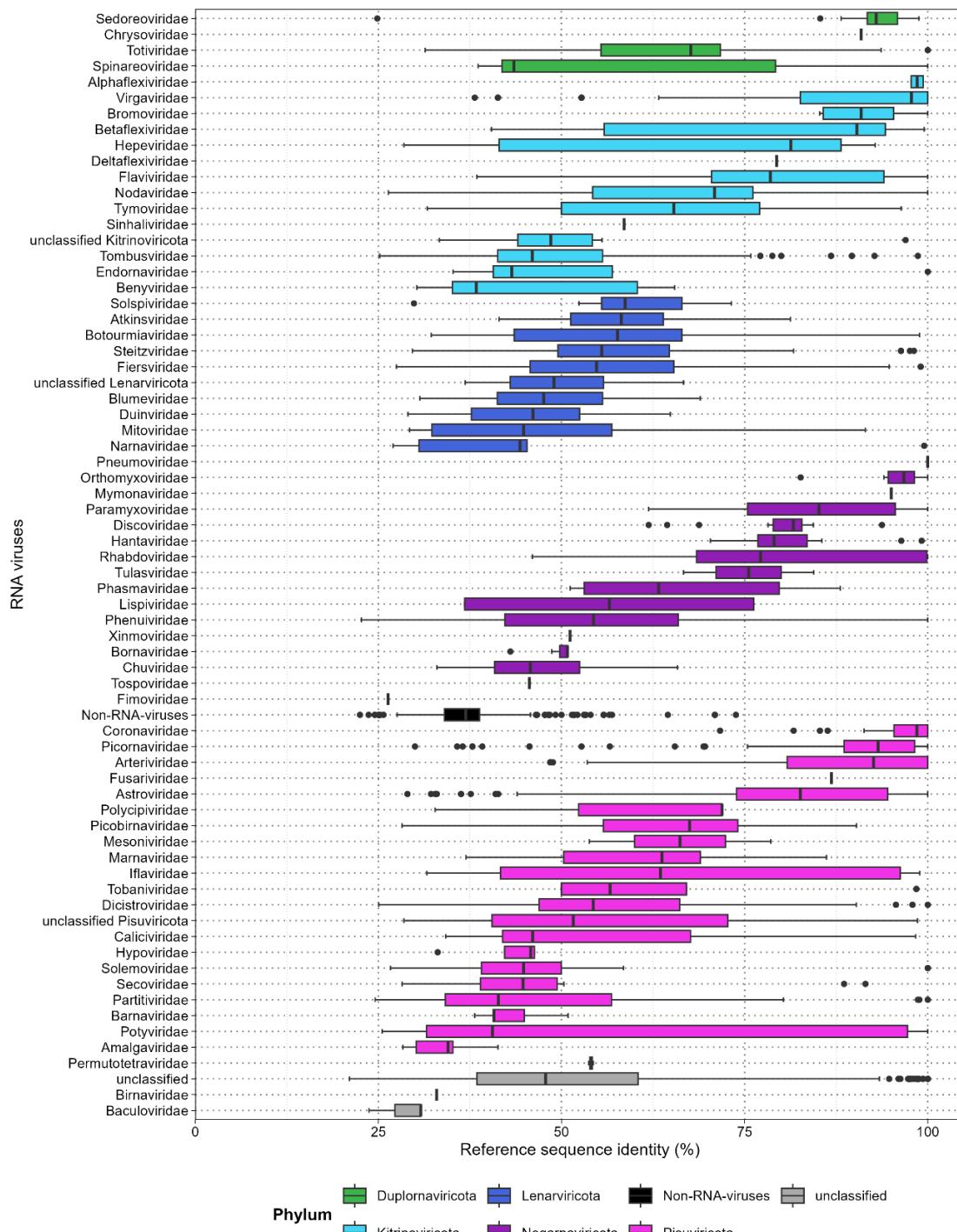


Figure 9: Mammalian sequencing data - Protein sequence identities to the nearest known reference virus for each viral family.

Among the 10 families with the highest number of contigs, five—*Picobirnaviridae*, *Fiersviridae*, *Tombusviridae*, *Steitzviridae*, and *Partitiviridae*²²⁵—demonstrated that over 90% of their viral contigs were classified as novel sequences (Table 6). *Steitzviridae* and *Fiersviridae* belong to the phylum *Lenarviricota*, while *Picobirnaviridae* and *Partitiviridae* are part of *Pisuviricota*, and *Tombusviridae* falls under *Kitrinoviricota*. Despite *Arteriviridae* having the highest total number of contigs, only 36.14% were novel sequences, representing the second-lowest percentage, just above *Picornaviridae*. Together, these 10 families accounted for 42.25% of all viral contigs identified. Additionally, when including *Astroviridae* from *Pisuviricota* and *Flaviviridae* from *Kitrinoviricota*, three phyla emerged as dominant: *Lenarviricota*, *Kitrinoviricota*, and the most represented, *Pisuviricota*.

I further explored the studies from which the novel sequences originated. Among the top three viral families, 391 novel contigs aligned with *Arteriviridae*, with a substantial portion (347) originating from study ERP130049²²⁶. This study investigated infected cell lines derived from the kidneys of *Chlorocebus sabaeus*, where porcine reproductive and respiratory syndrome virus 1 (PRRSV) was introduced in vitro to evaluate alternative cell lines as replacements for primary pulmonary alveolar macrophages (PAM) in PRRSV isolation and growth²²⁷.

Of the 310 novel contigs aligning with the family *Picobirnaviridae*, 290 were derived from the same study (SRP352646²²⁸), which identified picobirnaviruses in saliva samples from backyard swine in a South African province²²⁹. The novel contigs aligning with *Fiersviridae* primarily originate from two studies. In this case, SRP352646 provided 122 contigs that aligned with *Leviviridae*¹⁴⁵ species and single-stranded RNA phages. Another study (SRP273560²³⁰) examined extracellular vesicles from cheese-making by-products²³¹. Thus, while the host information in the SRA indicates cattle (*Bos taurus*), the actual analysis was performed on dairy products.

Overall, the top ten viral families included 1,676 novel contigs, which accounted for 26.99% of the 6,209 contigs identified. The five most prevalent hosts were primarily domesticated pigs and cattle, comprising 963 contigs (57.46% of the 1,676 contigs), followed by the Horseshoe bat and the Middle East blind mole-rat.

Table 6: Top ten viral families with the highest number of contigs aligning to them. Contigs with sequence identities below 90%, and those equal to or above, are given, along with the median, 25th, and 75th percentiles (quantile 1 and 2) of reference sequence identity of sequence identities for each family.

Viral taxonomy	Identity < 90%	Identity ≥ 90%	Total	Median	Q1	Q3
Arteriviridae	391	691	1082	92.59	80.81	100.00
Picobirnaviridae	310	2	312	67.50	55.70	74.07
Fiersviridae	277	7	284	54.78	45.70	65.35
Astroviridae	135	99	234	82.59	73.88	94.52
Flaviviridae	116	58	174	78.52	70.49	94.01
Tombusviridae	134	2	136	46.00	41.30	55.60
Steitzviridae	109	5	114	55.49	49.52	64.72
Picornaviridae	37	69	106	93.23	88.58	98.21
unclassified Pisuviricota	85	10	95	51.61	40.52	72.70
Partitiviridae	82	4	86	41.38	34.15	56.82

3.2.4 Four-fifths of all assembled contigs measure less than 1,000 nucleotides

The median contig length varied significantly, ranging from 104 nucleotides (IQR: 97 to 114) in the family *Arteriviridae* to a single contig measuring 10,373 nucleotides that aligned with *Pneumoviridae*²³² from the phylum *Negarnaviricota* (Table 17 in Appendix). Notably, *Arteriviridae* had both the lowest median length and the shortest individual contig at 78 nucleotides, alongside the second-longest contig at 18,735 nucleotides (Figure 10). The longest contig overall measured 26,579 nucleotides and aligned with the family *Coronaviridae*¹⁷. Across all viral phyla, only 19.72% (1,225) of the contigs exceeded 1,000 nucleotides in length. *Pisuviricota* accounted for the largest portion, contributing 258 contigs (21.06%). Among these, 202 contigs exhibited sequence identities below 90% and lengths greater than 1,000 nucleotides. Over half of these contigs were associated with the families *Picobirnaviridae* (46), *Arteriviridae* (36), and *Astroviridae* (31) (Table 18 in Appendix).

Results

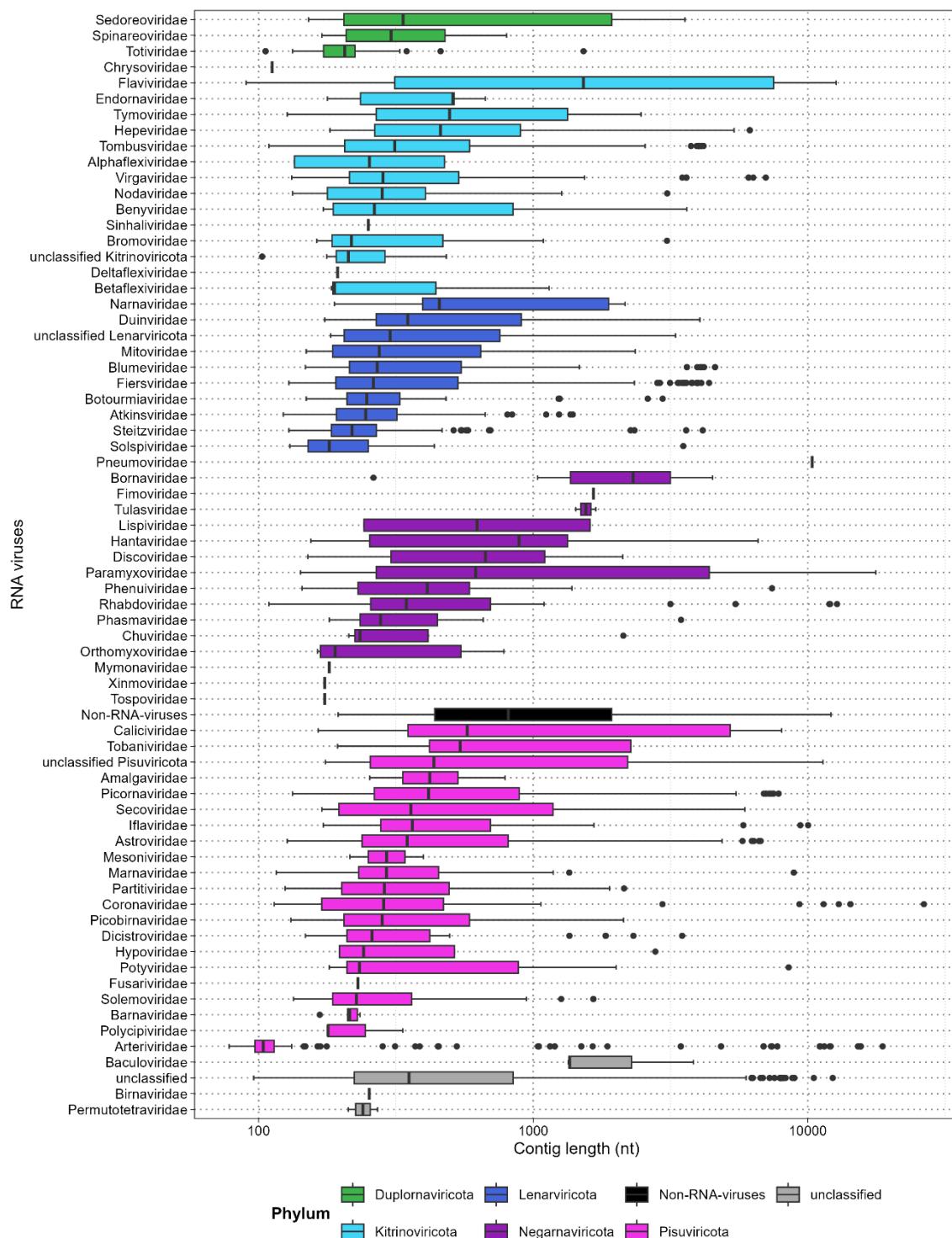


Figure 10: Mammalian sequencing data – Distribution of the length of the contigs in each viral family.

Following *Pisuviricota*, *Kitrinoviricota* had the second-highest number of contigs longer than 1,000 nucleotides, totalling 148 (12.08%), with 103 classified as novel. The majority of longer contigs with low sequence identity in *Kitrinoviricota*

were from the families *Flaviviridae* (58), *Tombusviridae* (28), and *Hepeviridae* (5). Ninety-seven contigs contributed to *Lenarvirocota* (7.92%), with 95 being novel. These primarily aligned with the families *Fiersviridae* (41), *Blumeviridae* (14), and *Duinviridae*²²⁰ (11). The top ten viral families with the most contigs were predominantly from these three phyla, which also dominated the count of contigs longer than 1,000 nucleotides. *Negarnaviricota* accounted for 54 contigs (4.41%), of which 42 were novel, while *Duplornaviricota* had only 10 contigs (0.82%), with two identified as novel. The *Duplornaviricota* contigs aligned with the families *Sedoreoviridae*¹⁵⁷ and *Totiviridae*, while the dominant families in *Negarnaviricota* were *Paramyxoviridae* (12), *Bornaviridae*²³³ (6), and *Hantaviridae*^{234,235} (5).

In total, 444 contigs (36.24%) from RNA virus phyla were both novel and exceeded 1,000 nucleotides in length. These contigs had a median length of 2,619 nucleotides (IQR = 1,587 to 4,844), with lengths ranging from 1,001 to 26,479 nucleotides.

To ensure that the identified contigs represent genuine viral sequences, a functional annotation of the viral proteins for the ten longest contigs originating from different RNA virus phyla was performed. The lengths of the top ten longest contigs varied from 3,569 to 26,479 nucleotides (Figure 11; table 19 in Appendix). Remarkably, all contigs aligned to different virus families, and with the exception of one, each exhibited a protein sequence identity below 90%. What is more, the functional annotation consistently identified the RNA-dependent RNA polymerase (RdRp), which served as the profile for screening, while most contigs also found characteristic protein domains specific to their respective virus families.

As an illustration, The longest contig corresponds to a novel alphacoronavirus identified in a sample from Harrison's large-eared giant mastiff bat (*Otomops harrisoni*). This sample was obtained during a metatranscriptomic analysis of oral and rectal swabs collected from bats inhabiting caves in Kenya²³⁶. The second largest contig, measuring 18,735 nucleotides, was one of the few identified novel sequences of porcine reproductive and respiratory syndrome virus 2 found in wild boars (*Sus scrofa*)^{237,238}, rather than in green monkeys (*Chlorocebus sabaeus*). Typical proteins of *Coronaviridae* and *Arteriviridae* were identified in both contigs, aligning with the correct open reading frames (ORF) and arranged in the expected order, with only previously known proteins found²³⁹. The longest contig

identified ORF1a along with its known non-structural proteins (NSP), while ORF1b included the RdRp, non-structural protein 14 (NSP14), and helicase; the subsequent open reading frames (ORFs) also contained the structural proteins. Similarly, the contig aligning to the family *Arteriviridae* identified ORF1a and ORF1b with their associated proteins, as well as envelope glycoproteins 2-5 arranged (EnvGP2a, 3 and 5) in the correct order²³⁹.

Lastly, the third largest contig was classified as Mount Mabu Lophuromys virus 1 from the *Paramyxoviridae* family. When compared to the known genome structure of this virus, the characteristic ORFs were identified, containing the nucleocapsid (N-protein), phosphoprotein (P/V-PP), glycoprotein (GP), and the final ORF with the RdRp²⁴⁰. A closer examination of the sequence revealed a region in the ORF containing the phosphoprotein that resembled the RNA editing site (TTAAAAAAAGGCA), previously observed in Mount Mabu Lophuromys virus 1. This sequence is characteristic of the conserved motif (YTAAAARRGGCA) found in all members of the genera *Henipavirus*, *Morbillivirus*, and *Jeilongviruses*²⁴⁰.

Results



Figure 11: Mammalian sequencing data – Functional annotation of the top ten largest contigs. This gene map was created via the gggenomes R package²¹¹.

3.2.5 Mammalian hosts are mainly farm animals and Old World monkeys

I was particularly interested in the most prevalent viral subjects assigned to each contig, so I extracted the top ten (Table 7). Porcine reproductive and respiratory syndrome virus (PRRSV) and PRRSV-2 were identified 987 times within the family *Arteriviridae*, representing approximately 91.22% of all *Arteriviridae*-aligned contigs, with 298 classified as novel. Notably, these two viruses and *Otarine picobirnavirus* were the only ones where the total number of contigs differed from those with a protein sequence identity below 90%. Among the hosts, African green monkeys (*Chlorocebus sabaeus*), often used in biomedical research²⁴¹, appeared in 981 contigs, of which 293 might be novel. Wild boars (*Sus scrofa*) seemed to host 4 contigs (three of which were novel), while two novel contigs were identified in Bushy-tailed woodrats (*Neotoma cinerea*). Another rodent arterivirus was found in Old World monkeys of the genus *Chlorocebus* (46, *Chlorocebus sabaeus*) and Eastern gray squirrels (2, *Sciurus carolinensis*). Closer inspection revealed that all instances of PRRSV-1 and PRRSV-2 found in green monkeys originated from the same study (ERP130049²²⁶), which involved infecting a cell line derived from the fetal kidney tissue of African green monkeys in vitro with PRRSV-1²²⁷. Meaning that these infections did not occur naturally.

Table 7: Mammalian sequencing data - Closest viral family and subject found , with the number of contigs ('Count' column) aligned to each viral subject. The number of contigs with sequence identity below 90% to the closest viral reference is also provided.

Top 10 Viral Subjects			number of contigs	number of contigs below 90% identity
Viral taxonomy	Viral subjects			
Arteriviridae	Porcine reproductive and respiratory syndrome virus	740	171	
Arteriviridae	Porcine reproductive and respiratory syndrome virus 2	247	127	
Picobirnaviridae	<i>Otarine picobirnavirus</i>	90	88	
Picobirnaviridae	Picobirnavirus Equ3	68	68	
Astroviridae	Mamastrovirus 18	57	57	
Flaviviridae	Hepacivirus glareoli	50	50	
Arteriviridae	Rodent arterivirus	48	48	
Picobirnaviridae	Picobirnavirus dog/KNA/2015	46	46	
Picobirnaviridae	Picobirnavirus green monkey/KNA/2015	40	40	
Picobirnaviridae	Chicken picobirnavirus	35	35	

Of the top 10 viral subjects, half were found in the family *Picobirnaviridae*, comprising 279 viral contigs, with 277 having a protein sequence identity below 90%. *Otarine picobirnavirus*, typically associated with sea lions, was identified 90

times. In this study, the primary host appears to be domesticated pigs (*Sus scrofa domesticus*, 85 contigs), a finding also noted by Chauhan et al.²⁴². Additionally, 4 contigs were found in donkeys (*Equus asinus*) and 1 in cattle (*Bos taurus*), suggesting that all instances of this aquatic mammal virus occurred in farm animals. Chicken picobirnavirus was detected in domesticated pigs (32), donkeys (2), and cattle (1). The dog variant was observed in pigs (44) and donkeys (2), while the variant associated with green monkeys seems to have been found in 38 pigs and 2 donkeys. Lastly, the variant primarily associated with horses (*Equus caballus*) was detected in pigs (60), rats (*Rattus*, 2), and cattle (3). Remarkably, it was also found in two Rhesus macaques (*Macaca mulatta*) and one Sunda pangolin (*Manis javanica*).

Mamastrovirus 18, typically associated with bats²⁴³, was identified in three bat species (*Miniopterus africanus*, *Rhinolophus*, and *Myotis tricolor*). Novel contigs aligning to a hepacivirus were found in Middle East blind mole-rats (46, *Nannospalax ehrenbergi*), Eastern deer mice (2, *Peromyscus maniculatus*), and one each in muskrats (*Ondatra zibethicus*) and Desmarest's spiny pocket mice (*Heteromys desmarestianus*).

Analyzing the complete set of 6,209 contigs and their distribution among the host taxa in which they were identified revealed that the majority originated from farm animals (Table 20 in Appendix). Notably, 38.67% of all contigs were found in cattle, while domesticated pigs accounted for 17.28%. Additionally, *Chlorocebus sabaeus*, primarily represented by samples from a single study, contributed 17.06% of the contigs. The remaining contigs were predominantly associated with bats and rodents.

3.2.6 Unclassified Viral Families and the Specificity of Hidden Markov Model Profiles

I generated all figures and tables presented (except Figure 11) and performed preprocessing and calculations of summary statistics in the results section using Virusparies, the R package I created for this master's thesis. Contigs were grouped into different viral families on the y-axis of each figure by extracting the virus family from the *ViralRefSeq_taxonomy* column in both VirusHunter and

Results

VirusGatherer hittables. This extraction was handled internally by Virusparies functions.

Values in this column consist of a taxid and taxonomic information about the detected virus sequence, derived from the alignment of query sequences against the viral database, and are separated by the "|" character (e.g., *taxid:1354498|Hepacivirus* *colobi|Hepacivirus|Flaviviridae|Amarillovirales|Flasuviricetes|Kitrinoviricota|Orthornavirae|Riboviria*).

Occasionally, the VirusHunter and VirusGatherer pipelines fail to extract information about the viral family, resulting in contigs being grouped into the 'unclassified' category or 'unclassified' followed by the phylum name if phylum information is still present. I observed that approximately one-third of all contigs provided no information about family or phylum, while 3.35% provided information about the phylum but not the virus family (Figure 12). Only 2,894 out of the 6,209 viral contigs provided information about the family.

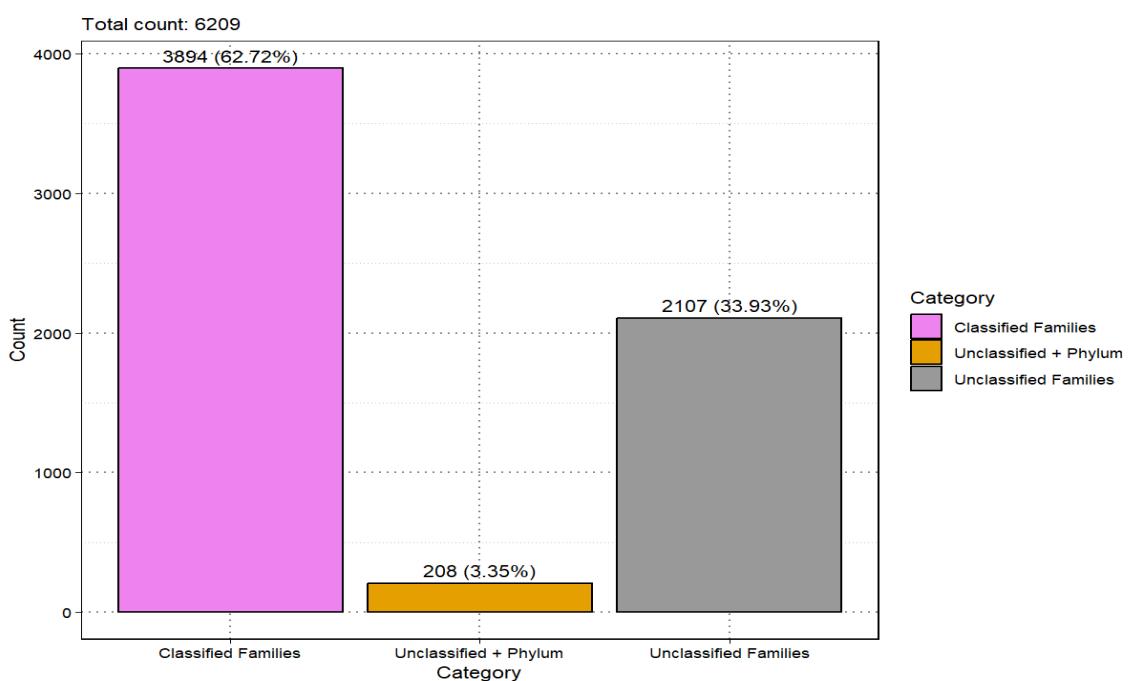


Figure 12: Distribution of contigs with assigned viral reference families (Classified Families) compared to contigs where only phylum information (Unclassified + Phylum) or no reference information (Unclassified Families) is available.

VirusHunter was employed to screen for viral sequences within unprocessed data by utilizing profiles of conserved proteins characteristic of specific virus families. The identified contigs were subsequently assembled using VirusGatherer. A total of eight profiles were applied (see Materials and Methods), and the results were consolidated into a single output file.

When running VirusHunter independently, it became apparent that the profiles used were not exclusively specific to their corresponding virus families. For example, when VirusHunter was run with SRA experiments aligned solely against the RNA-dependent RNA polymerase (RdRp) of the *Flaviviridae* family (Flavi_RdRp), 14 other families were identified within my VirusHunter hittable (Figure 13). Notably, some of these identified profiles—such as NidoAstro_RdRp, Nido_NiRAN, and Negative_Bunya-Arena_RdRp—correspond to the eight profiles utilized in this study. This suggests that the run for Flavi_RdRp also contributed to the assembly of contigs belonging to families associated with the other profiles

Results

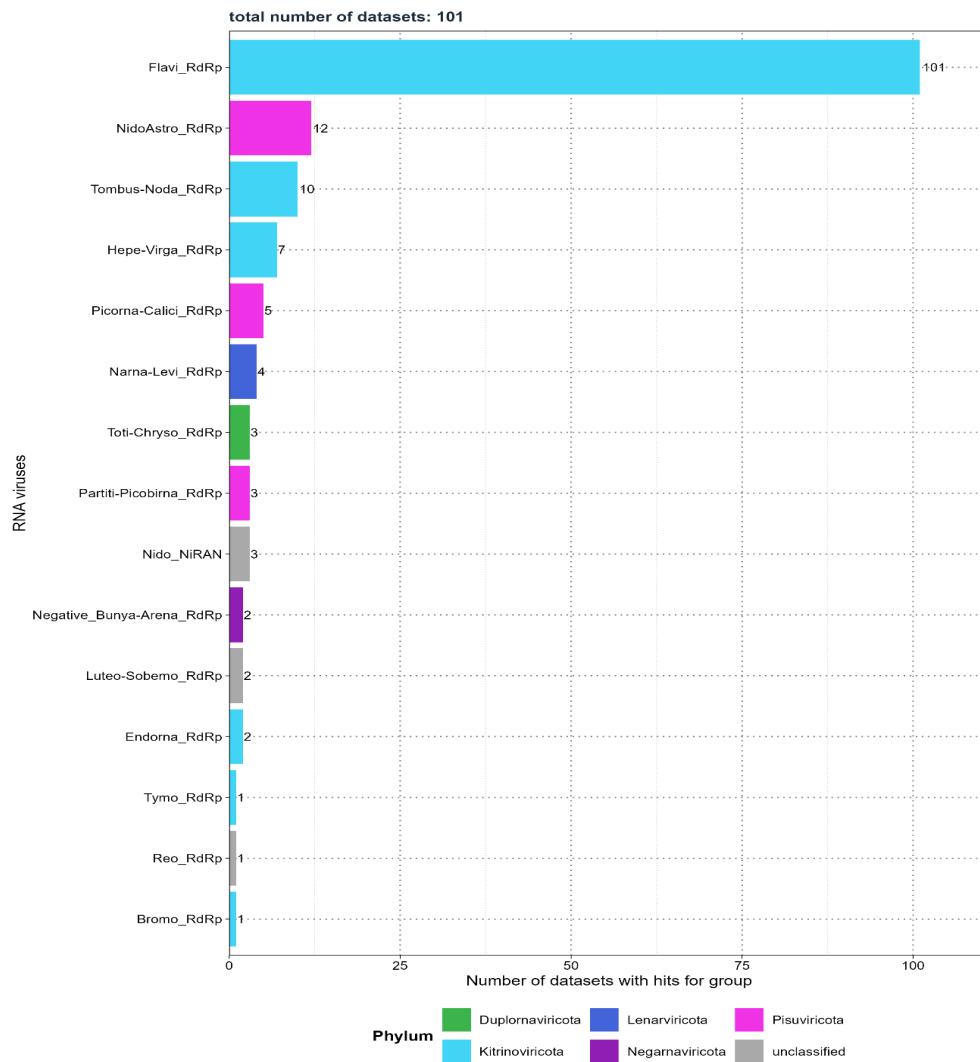


Figure 13: Screening for Flaviviridae RNA-dependent RNA polymerase (RdRp) - Distribution of the number of SRA experiments detecting viral families.

4 Discussion

Currently, an estimated 1.67 million viral species are still unknown^{13,14}. Half of these species may be capable of infecting humans, with some potentially contributing to future outbreaks as the frequency of pandemics rises¹⁶. Other viruses are part of the virome. Meaning that living organisms, including humans harbor viruses in their cells or within the microorganisms that comprise their microbiota. The virome comprises viruses that can cause disease, particularly in immunosuppressed individuals, while others coexist as commensal viruses in our body without known pathogenicity^{36,43,44}. A few even integrate into the genome, contributing to the evolution of their hosts¹². Factors such as the host's lifestyle choices and environmental conditions influence the composition of the virome^{51–54}. Conversely, the virome affects the host, their microbiota, and the surrounding environment^{7,64}, resulting in a complex network of interactions.

Virus discovery aims to identify these novel and previously unknown viruses. Allowing us to improve our understanding of the evolutionary history of viruses, their disease-causing mechanisms, and the intricate relationships they have with their host and environment. This new knowledge could have practical applications, such as helping to find surrogate viruses when the original virus may not be suitable for biomedical research. Historically, virus discovery relied on sample collection and wet lab analysis, with progress tied to technological advancements. X-ray diffraction enabled structural analysis⁹², polymerase chain reaction (PCR) and Sanger sequencing allowed to study the viral genomes^{93,94}, and next-generation sequencing made large-scale metagenomics analysis possible, enabling the identification of viral sequences directly from environmental samples⁹⁶.

Internet access, along with the ability to upload large files to public repositories, has allowed researchers to share vast amounts of sequencing data. One such repository is the Sequencing Read Archive (SRA), which hosts 53 petabases of open-access sequencing data^{98,99}. It has been shown that sequencing data deposited in the SRA often come from samples that may have been infected by one or multiple viruses at the time of collection. This suggests that the SRA, with its large amount of unprocessed sequencing data, could contain unnoticed viral sequences – some from known viruses and others potentially novel – that were undetected

during initial sampling. New Data-Driven Virus Discovery (DDVD) approaches leverage the SRA to screen for viral sequences in deposited sequencing data, enabling the monitoring of viruses from existing samples. DDVD is primarily computational; performing homology searches for virus-positive samples, aligning them to viral reference genomes, and assembling longer viral contigs. These approaches offer advantages such as lower costs compared to wet lab methods and the ability to use metadata associated with the sequencing data to gain insights about the host¹⁰⁴. While not a replacement for wet lab experiments, DDVD tools serve as complementary approaches.

The aim of this thesis is to characterize novel and known viruses in sequencing data from liver biopsies of liver transplant patients provided by the Taubert working group, and to assess whether the identified viruses are associated with liver rejection. It further involves the characterization, assembly, and functional annotation of mammalian RNA viruses from existing screenings.

To archive this, I applied the VirusHunterGatherer pipeline developed by the Computational Virology Research Group at Twincore, a computational workflow designed for data-driven virus discovery. VirusHunter performs a sensitive homology search for conserved viral protein families to identify virus-positive samples from unprocessed sequencing data, either locally available or fetched from a public repository. Then, VirusGatherer assembles longer viral contigs and classifies them. The pipeline is employed on the Taubert dataset across all three viral groups (Small DNA, Large DNA, and RNA viruses), while only screening for RNA viruses in the dataset with the mammalian samples. The following subsections discuss the results.

4.1 Discussion of the results obtained from Taubert sequencing data

The results from the screening of the Taubert sequencing data, as detailed in section 3.1.1, indicate that only a small fraction of the initial dataset was capable of detecting viral sequences. Among the 557 FASTQ files (representing 323 samples) provided by the Taubert working group, only 0.89% (5 files) had micro-contigs associated with small DNA viruses, while 3.05% (17 files) were linked to RNA viruses. The largest proportion was found for large DNA viruses, at 29.62% (165

files). However, upon excluding all non-significant hits (E -value $> 1e^{-5}$), no files remained for small DNA viruses. For the screenings against RNA viruses and large DNA viruses, the number of files detecting potential viral micro-contigs was low (maximum of 11), as was the number of viral families identified (maximum of four, Figure 5, part A, D). Observed read coverage was low, with a maximum of 26 reads (Figure 5, part B, E). The low number of files detecting reads and potential viral contigs may suggest two possibilities: 1) the samples could genuinely contain few viral sequences, or 2) viral sequences may have been present in the liver samples but went undetected, possibly because the sequencing approach was not optimal for detecting viruses. There may have been no specific enrichment for viral sequences, or the viral sequences could have been lost during pre-processing steps before sequencing^{106–108,244,245}. Both scenarios are plausible, as the samples were not originally collected for viral discovery. At this stage, no definitive conclusions can be made regarding which of the two possibilities is correct. Further investigation would be needed to determine whether the low detection rate is due to a genuine lack of viral sequences in the samples or the result of decisions made during sample pre-processing.

VirusGatherer identified 17 virus-positive contigs with sequence similarity to six viral families. However, alignments against the families *Hepadnaviridae*¹⁸³ and *Poxviridae*¹⁶³ were considered not significant in VirusHunter, while *Flaviviridae*¹⁵², *Potyviridae*²¹², and *Adenoviridae*¹⁶⁴, detected by VirusHunter, were not found by VirusGatherer (Figure 6). This may be due to the low number of contigs aligning to these families in the VirusHunter step. The assembly of longer contigs typically requires sufficient coverage, and if coverage is too low, successful assembly may not occur. Among the remaining contigs, some aligned with betaherpesviruses and hepatitis viruses, with the majority having animal-infecting viruses as their closest match (Table 4). Importantly, the maximum length of all contigs was under 300 nucleotides (Figure 6, part C, F, I). Assigning taxonomy based on homology searches may prove challenging when the sequences used for alignment are short and potentially diverge from the reference sequence, as they contain limited taxonomic information²⁴⁶.

Since no host subtraction was performed (See 2 Material and Methods) and the host genome is typically much more abundant than viral sequences, there is the

possibility that short sequences from the host may align with the viral reference sequences. Previous work has shown the challenges associated with distinguishing viral sequences from host sequences, especially those with E-values near the borderline of significance¹¹¹. In my analysis, it is noteworthy that most of the contigs exhibited E-values close to the 1e⁻⁵ cutoff.

That is further compounded by the fact that some human sequences are of viral origin. Such host sequences could potentially act as contaminant^{247–249}.

Furthermore, out of the 6 potential viral families identified in VirusGatherer, 3 (*Hepadnaviridae*, *Kolmioviridae*²¹⁴, and *Poxviridae*) were deemed non-significant or not detected in VirusHunter, appearing only in VirusGatherer due to its use of a more lenient E-value threshold during assembly. Given the higher significance cutoff applied in my analysis, it is likely that the VirusGatherer results may not be truly valid. Functional annotation of protein domains can offer support in assessing a sequence's viral origin by identifying known viral protein domains. However, due to their short contig length, the contigs presented here are unlikely to contain meaningful protein information. Meaning it is difficult to ascertain whether these contigs are genuinely viral in nature.

Overall, the small proportion of virus-positive contigs might indicate that samples were not infected by viruses, as a genuine viral infection would likely produce a higher abundance of viral sequences, which VirusHunterGatherer would have been more likely to detect. However, this remains speculative, and definitive conclusions cannot be made without more detailed information on the pre-processing of the samples, which was not provided. Furthermore, the few contigs that were found are too short to make precise statements about whether they are truly viral in nature or whether they are associated with potential liver rejection. The results were presented to the Taubert working group, allowing them to make informed decisions based on the available data.

4.2 Discussion of the results obtained from screening and assembling mammalian viral sequences

This study utilized VirusGatherer to assemble 6,209 contigs from 332 SRA experiments derived from mammalian samples. The analysis revealed alignments

with 65 RNA virus families and 11 non-RNA virus families (Figure 8, table 5), predominantly represented by the three phyla: *Lenarviricota*, *Kitrinoviricota*, and *Pisuviricota*. These phyla accounted for a large portion of the longer contigs ($\geq 1,000$ nucleotides, Figure 10). The top ten virus families, based on the total number of contig alignments, were also primarily represented by these three phyla (table 6).

The observed prevalence of these phyla may be partially influenced by the host composition, given that over half of the samples originated from farm animals. The viruses detected appear to align with those typically infecting livestock²⁵⁰. For example, the family *Picobirnaviridae* was primarily identified in a study analyzing domestic pigs in South Africa²⁴². Similarly, families such as *Astroviridae*, *Picornaviridae*, and *Tombusviridae* are recognized as pathogens in farm animals, with most of their hosts in this project being either pigs or cattle (Table 20 in Appendix). It is also worth noting that half of the ten hosts contributing to the longest contigs are either sheep, pigs, or cattle (Table 19 in Appendix).

Approximately 80% of the assembled contigs appear to be novel, as they exhibit less than 90% protein sequence identity to the closest viral reference (Figure 9). However, it should be acknowledged that nearly 80% of these contigs are relatively short, with lengths under 1,000 nucleotides. Only 444 contigs seem to meet the criteria for both novelty and length, of which approximately 45% are associated with *Pisuviricota*, 23% with *Kitrinoviricota*, and 21% with *Lenarviricota*. As a result, these three phyla are the most dominant among the novel, longer contigs.

The functional annotation of the longest contigs indicates that VirusHunter and Virusgatherer can successfully identify and assemble viral sequences (Figure 11). Viral proteins characteristic of specific virus families were accurately located within the expected open reading frames (ORF) and arranged in the correct order across various families. As expected, each contig contained at minimum the domain for the RNA-dependent RNA polymerase (RdRp), used as profiles to screen for specific RNA virus families. Beyond that, the third longest contig aligning to mabu lophuromys virus 1 (*Paramyxoviridae*²²¹) is noteworthy because it not only contains the expected protein domains but also features the characteristic RNA editing site in the phosphoprotein-containing open reading frame, supporting the notion that these contigs are of genuine viral origin²⁴⁰.

The same applies to the longest contig, which aligns to an alphacoronavirus from a Harrison's large-eared giant mastiff bat sample. For instance, characteristic proteins associated with the *Coronaviridae*¹⁷ family are identified, including ORF1a, along with its non-structural proteins, ORF1b, RdRp, and spike proteins²³⁹. Furthermore, all but one of the top ten longest contigs exhibit protein sequence identities below 90%. This suggests that some of the identified contigs, such as the one aligning with the alphacoronavirus, could represent not only novel contigs but also genuine novel viral genomes. Further investigation is needed to determine whether the contigs are coding complete, along with a phylogenetic analysis. None of this was done here due to time constraints.

Of particular relevance is the second longest contig, which not only contains the expected viral proteins but is also one of the few contigs aligning to the porcine reproductive and respiratory syndrome virus (PRRSV) found in its natural host. PRRSV is known for infecting wild hogs and domesticated pigs, causing reproductive disorders in sows and respiratory signs in piglets and fattener pigs²⁵¹. This virus is significant due to the substantial economic losses it inflicts on the pig industry, with the latest outbreak occurring in 2019 in Russia²⁵². However, in this project, the majority of PRRSV cases were identified in *in vitro* infected cells from Old World monkeys²²⁷. Meaning, these infections did not occur naturally. While the SRA-provided host taxon information indicates that the samples came from *Chlorocebus sabaeus*, it is worth noting that they were derived solely from cells, and no monkeys were reported as infected at the time of the experiment. This applies also to a portion of contigs associated with picobirnaviruses, as one study analyzed dairy products but provided *Bos taurus* as the host²³¹.

In both cases, the limitations of utilizing SRA host information become apparent: taxonomic annotation is constrained by the options available to the submitters of each sequencing dataset. If submitters cannot indicate whether samples originate from a host, are derived from a host, or are byproducts of the host (such as milk products from cattle), or if they incorrectly provide sample information despite the available options, then the host information lacks granularity. This situation requires users of VirusHunterGatherer to trace the results back to the original work to determine whether the sample truly originates from the host. As these contigs are typically of little interest for virus discovery, curating the identified

contigs would be necessary unless the scientific question specifically requires contigs found in cultivated cells. In this project, removing the aforementioned *Picobirnaviridae*, PRRSV, and potentially irrelevant contigs could reduce the total number of contigs by at least one-sixth.

Contigs of particular interest align with the families *Flaviviridae*, *Tombusviridae*²²⁴, *Coronaviridae*, *Picobirnaviridae*, *Astroviridae*²¹⁷, and *Arteriviridae*. As mentioned, *Coronaviridae* and PRRSV (from *Arteriviridae*) contributed the longest contigs (Figure 11), which contain the majority of characteristic protein domains and exhibit low protein sequence identity to the closest viral references. *Flaviviridae* was represented in the most SRA experiments (Figure 8) and included among the highest number of contigs, many of which appear to be both long (>1,000 nucleotides) and novel sequences. Both *Astroviridae* and *Tombusviridae* contain known viruses found in farm animals and also featured among the longest contigs with protein sequence identities below 90%. Interestingly, contigs aligning to *Picobirnaviridae* were associated with domesticated pigs and contained viruses related to livestock; however, a significant proportion of these viruses have sea lions as their natural hosts, with currently no known explanation for how a virus from an aquatic mammal could spread to livestock²⁴².

This study acknowledges the limitations and potential inaccuracies that may arise with VirusHunterGatherer. First, VirusHunter can only detect viral sequences that yield a hit during screening. For viruses with continuous genomes, this may not pose a significant issue. However, segmented viruses require individual profiles for each segment. In this study, I referenced at least one analysis suggesting a potential prevalence of influenza A virus in domesticated pigs. Without segment-specific profiles for Influenza A, it is possible that this study missed significant portions of segmented viruses, such as influenza A, if infections were indeed present. Second, VirusHunterGatherer failed to retrieve taxonomy information in at least one-third of the cases (Figure 12), resulting in the absence of data regarding the closest virus in those instances. Third, this study observed cross-reaction of non-RNA viruses (table 5), which were detected with lower E-values during the screening and assembly of RNA viruses. Lastly, profiles are not exclusive to a single viral family. Screening against one profile can yield contigs aligning to multiple families (Figure 13). Consequently, runs for specific profiles may inadvertently contribute to the assembly

of contigs associated with other families, ultimately increasing the total number of assembled contigs.

In summary, of the 6,000 assembled contigs, the predominant virus families and phyla align with viruses that infect farm animals, reflecting the large number of livestock hosts. Some families are prevalent because most contigs align with them, they return the longest sequences, contain the most novel sequences, or a combination of these factors. At least one-sixth of the contigs originate from studies that examined viruses in cultivated cells or host by-products rather than their natural hosts. This analysis reveals certain limitations of VirusHunterGatherer screening but also demonstrates its ability to identify and assemble viral contigs with characteristic viral protein domains.

4.3 Virusparies

As mentioned in subsection 1.3.4, tools like Serratus¹⁰² and DAMIAN¹¹⁵ offer the advantage of providing clear, well-organized reports in both tabular and graphical formats. This is crucial, as the vast amount of data produced when screening for viruses in public datasets can be overwhelming. For instance, the VirusGatherer hittable for mammalian samples alone contained 8,232 rows and 13 columns before filtering, making it impractical to interpret directly. VirusHunterGatherer, until now, lacked a dedicated tool for visualizing or summarizing results, requiring extensive downstream analysis. This process demanded proficiency in both programming and statistics, as well as a deep understanding of the structure and purpose of each column in the hittables to generate meaningful reports.

I present the R¹⁹⁴ package Virusparies¹⁹⁵, designed to enable users to visualize, process, and summarize the hittable outputs from VirusHunterGatherer. Virusparies includes plot functions that generate boxplots to show the distribution of significant and non-significant hits based on E-value cutoffs (Figure 4 and 7), as well as the distribution of contig lengths (Figure 10) and the distribution of results based on reference sequence identity to the viral reference (Figure 9). It can also generate bar plots displaying the number of files or samples that aligned to specific taxa

(Figure 8) and the number of reads found (Figure 5, part B and E), enabling the user to specify the taxa rank of interest (e.g., "family").

Virusparies offers functions to generate summary statistics (Table 14, 15 and 17 in Appendix) and extract specific information, such as the closest viral subject (Table 4) or the number of contigs below or above specific criteria (e.g., the number of contigs below 90% sequence identity and above, Fig 6). Hittables can also be further subset based on criteria like specific viral taxa, E-value, contig length, or sequence identity. For each summary statistic or processed hittable, users have the option to generate a graphical table suitable for inclusion in reports or published work.

Each function is highly modular, allowing users to customize the output of Virusparies functions to their preferences. Additionally, all plot functions and functions generating graphical tables return a ggplot or gt object, respectively, enabling users to further modify plots and graphical tables if Virusparies does not meet their needs.

Virusparies was created as part of this project to provide an additional tool for VirusHunterGatherer, enabling users to generate quick visualizations and summary statistics that support informed decision-making. The decision was made to create Virusparies as an R package to make the tools accessible to a larger audience, particularly since some wet-lab scientists have a rudimentary understanding of the R language through courses offered by their institutes, university programs, or self-study. To enhance accessibility for this group, each function was designed as a one-liner, in contrast to the layered syntax used in ggplot2²⁵³ and gt²⁵⁴ packages, which can alienate inexperienced R programmers due to their complexity.

Even experienced R users benefit from Virusparies, as they can save time by utilizing the functions instead of writing their own code from scratch. They can also build on the existing code by customizing the R package to their preferences or adding layers to ggplot2 or gt objects.

Reproducibility is considered a fundamental requirement in both traditional scientific fields and computational research. However, many scientific studies present challenges to reproducibility, making it difficult or even impossible to replicate results^{255,256}. In bioinformatics, methodological reproducibility often falters

because code is frequently not shared publicly alongside the study. Additionally, technical hurdles, such as incompatibilities between different versions of software programs or libraries, can further hinder reproducibility. Virusparies enhances code reusability and shareability among different users and institutions, allowing them to easily cite the package in their work²⁵⁷. It improves methodological reproducibility by packaging code with clear documentation, including usage instructions, examples, and datasets, ensuring that results can be replicated by others.

4.4 Outlook

The VirusHunter and VirusGatherer pipelines successfully identified and assembled over 6,000 contigs. This analysis indicates that at least some of these contigs are novel and contain viral proteins characteristic of known viruses, suggesting that they likely represent genuine viral entities.

The next steps should focus on curating these contigs and eliminating any unwanted entries, particularly those that do not originate from the host, such as cells derived from hosts or by-products of the host. This process will require tracing the samples back to the original work to determine the virus's source, the data processing methods used, and any relevant information about the virus or host.

Following the curation of these contigs, the next phase will involve phylogenetic analysis and comprehensive taxonomic classification. Currently, only basic taxonomic annotations have been conducted, relying on protein sequence identity comparisons with the nearest viruses in the reference database. A more thorough taxonomic classification will require a detailed examination of the viral sequences within the taxonomic system. This process typically requires reconstructing a phylogenetic tree to accurately classify the viral sequences within the broader context of viral taxonomy²⁵⁸. A taxonomic classification can also assess whether the viral sequences are genuinely novel, rather than merely novel in comparison to the viral reference.

Future work should also investigate whether novel viruses possess new viral proteins or only known ones, as well as determine the proportion of novel sequences that are genuine viral sequences versus those representing contaminants. Especially regarding the contigs aligning with picobirnaviruses from sea lions,

despite their identification in domesticated pigs, it is important to determine if these sequences genuinely originate from viruses infecting sea lions. If they do, further investigation is needed to understand how transmission from sea lions to pigs could occur, given their geographic differences.

Ultimately, the assembled contigs could potentially serve as a valuable resource for the Experimental Virology Research Group's current and future projects. For instance, Li Chuin Chong's research on predicting the spillover risk of RNA viruses could benefit from these newly assembled contigs, either as a new dataset or by augmenting existing data for her own project.

5 List of references

1. Bergh O, Børshem KY, Bratbak G, Heldal M. High abundance of viruses found in aquatic environments. *Nature*. 1989;340(6233):467-468. doi:10.1038/340467a0
2. Swanson M m., Fraser G, Daniell T j., Torrance L, Gregory P j., Taliansky M. Viruses in soils: morphological diversity and abundance in the rhizosphere. *Ann Appl Biol*. 2009;155(1):51-60. doi:10.1111/j.1744-7348.2009.00319.x
3. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biol*. 2016;14(8):e1002533. doi:10.1371/journal.pbio.1002533
4. Mushegian AR. Are There 1031 Virus Particles on Earth, or More, or Fewer? *J Bacteriol*. 2020;202(9):e00052-20. doi:10.1128/JB.00052-20
5. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage | PNAS. Accessed October 3, 2024. <https://www.pnas.org/doi/full/10.1073/pnas.96.5.2192>
6. Flint J, Racaniello VR, Rall GF, Hatzioannou T, Skalka AM. *Principles of Virology, Volume 1: Molecular Biology*. John Wiley & Sons; 2020.
7. O'Malley MA. The ecological virus. *Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci*. 2016;59:71-79. doi:10.1016/j.shpsc.2016.02.012
8. Wang-Johanning F, Liu J, Rycaj K, et al. Expression of multiple human endogenous retrovirus surface envelope proteins in ovarian cancer. *Int J Cancer*. 2007;120(1):81-90. doi:10.1002/ijc.22256
9. Goering W, Ribarska T, Schulz WA. Selective changes of retroelement expression in human prostate cancer. *Carcinogenesis*. 2011;32(10):1484-1492. doi:10.1093/carcin/bgr181
10. Contreras-Galindo R, Kaplan MH, Leissner P, et al. Human Endogenous Retrovirus K (HML-2) Elements in the Plasma of People with Lymphoma and Breast Cancer. *J Virol*. 2008;82(19):9329-9336. doi:10.1128/JVI.00646-08
11. Serafino A, Balestrieri E, Pierimarchi P, et al. The activation of human endogenous retrovirus K (HERV-K) is implicated in melanoma cell malignant transformation. *Exp Cell Res*. 2009;315(5):849-862. doi:10.1016/j.yexcr.2008.12.023
12. Mi S, Lee X, Li X, et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*. 2000;403(6771):785-789. doi:10.1038/35001608
13. Anthony SJ, Epstein JH, Murray KA, et al. A strategy to estimate unknown viral diversity in mammals. *mBio*. 2013;4(5):e00598-00513. doi:10.1128/mBio.00598-13
14. Carroll D, Daszak P, Wolfe ND, et al. The Global Virome Project. *Science*. 2018;359(6378):872-874. doi:10.1126/science.aap7463
15. Taubenberger JK, Kash JC, Morens DM. The 1918 influenza pandemic: 100 years of questions answered and unanswered. *Sci Transl Med*. 2019;11(502):eaau5485. doi:10.1126/scitranslmed.aau5485
16. Viral Pandemics in the Past Two Decades: An Overview - PubMed. Accessed October 3, 2024. <https://pubmed.ncbi.nlm.nih.gov/34660399/>
17. Woo PCY, de Groot RJ, Haagmans B, et al. ICTV Virus Taxonomy Profile: Coronaviridae 2023. *J Gen Virol*. 2023;104(4):001843. doi:10.1099/jgv.0.001843
18. Mandl JN, Ahmed R, Barreiro LB, et al. Reservoir Host Immune Responses to Emerging Zoonotic Viruses. *Cell*. 2015;160(1):20-35. doi:10.1016/j.cell.2014.12.003
19. Khalil AM, Martinez-Sobrido L, Mostafa A. Zoonosis and zoonanthroponosis of emerging respiratory viruses. *Front Cell Infect Microbiol*. 2024;13. doi:10.3389/fcimb.2023.1232772
20. Cantalupo PG, Calgua B, Zhao G, et al. Raw sewage harbors diverse viral

- populations. *mBio*. 2011;2(5):e00180-11. doi:10.1128/mBio.00180-11
21. Rosario K, Duffy S, Breitbart M. Diverse circovirus-like genome architectures revealed by environmental metagenomics. *J Gen Virol*. 2009;90(Pt 10):2418-2424. doi:10.1099/vir.0.012955-0
22. Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M. Metagenomic analysis of viruses in reclaimed water. *Environ Microbiol*. 2009;11(11):2806-2820. doi:10.1111/j.1462-2920.2009.01964.x
23. Muscatt G, Cook R, Millard A, Bending GD, Jameson E. Viral metagenomics reveals diverse virus-host interactions throughout the soil depth profile. *mBio*. 14(6):e02246-23. doi:10.1128/mbio.02246-23
24. Sunagawa S, Acinas SG, Bork P, et al. Tara Oceans: towards global ocean ecosystems biology. *Nat Rev Microbiol*. 2020;18(8):428-445. doi:10.1038/s41579-020-0364-5
25. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A*. 1985;82(20):6955-6959.
26. Clarridge JE. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clin Microbiol Rev*. 2004;17(4):840-862. doi:10.1128/CMR.17.4.840-862.2004
27. Petti CA, Polage CR, Schreckenberger P. The role of 16S rRNA gene sequencing in identification of microorganisms misidentified by conventional methods. *J Clin Microbiol*. 2005;43(12):6123-6125. doi:10.1128/JCM.43.12.6123-6125.2005
28. Abayasekara LM, Perera J, Chandrasekharan V, et al. Detection of bacterial pathogens from clinical specimens using conventional microbial culture and 16S metagenomics: a comparative study. *BMC Infect Dis*. 2017;17(1):631. doi:10.1186/s12879-017-2727-8
29. Schoch CL, Seifert KA, Huhndorf S, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci*. 2012;109(16):6241-6246. doi:10.1073/pnas.1117018109
30. Irinyi L, Serena C, Garcia-Hermoso D, et al. International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database—the quality controlled standard tool for routine identification of human and animal pathogenic fungi. *Med Mycol*. 2015;53(4):313-337. doi:10.1093/mmy/myv008
31. Fitzpatrick AH, Rupnik A, O'Shea H, Crispie F, Keaveney S, Cotter P. High Throughput Sequencing for the Detection and Characterization of RNA Viruses. *Front Microbiol*. 2021;12. doi:10.3389/fmicb.2021.621719
32. Brenner SE, Chothia C, Hubbard TJP. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A*. 1998;95(11):6073-6078.
33. He Y, Liu WJ, Jia N, Richardson S, Huang C. Viral respiratory infections in a rapidly changing climate: the need to prepare for the next pandemic. *eBioMedicine*. 2023;93. doi:10.1016/j.ebiom.2023.104593
34. Carlson CJ, Albery GF, Merow C, et al. Climate change increases cross-species viral transmission risk. *Nature*. 2022;607(7919):555-562. doi:10.1038/s41586-022-04788-w
35. Gupta S, Rouse BT, Sarangi PP. Did Climate Change Influence the Emergence, Transmission, and Expression of the COVID-19 Pandemic? *Front Med*. 2021;8:769208. doi:10.3389/fmed.2021.769208
36. Liang G, Bushman FD. The human virome: assembly, composition and host interactions. *Nat Rev Microbiol*. 2021;19(8):514-527. doi:10.1038/s41579-021-00536-5
37. Harrison E, Brockhurst MA. Ecological and Evolutionary Benefits of Temperate Phage: What Does or Doesn't Kill You Makes You Stronger. *BioEssays News Rev Mol Cell Dev Biol*. 2017;39(12). doi:10.1002/bies.201700112
38. Bellas CM, Anesio AM, Barker G. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. *Front Microbiol*.

- 2015;6:656. doi:10.3389/fmicb.2015.00656
39. De Sordi L, Lourenço M, Debarbieux L. "I will survive": A tale of bacteriophage-bacteria coevolution in the gut. *Gut Microbes*. 2019;10(1):92-99. doi:10.1080/19490976.2018.1474322
40. Scanlan PD, Hall AR, Blackshields G, et al. Coevolution with bacteriophages drives genome-wide host evolution and constrains the acquisition of abiotic-beneficial mutations. *Mol Biol Evol*. 2015;32(6):1425-1435. doi:10.1093/molbev/msv032
41. Pal C, Maciá MD, Oliver A, Schachar I, Buckling A. Coevolution with viruses drives the evolution of bacterial mutation rates. *Nature*. 2007;450(7172):1079-1081. doi:10.1038/nature06350
42. Barr JJ, Auro R, Sam-Soon N, et al. Subdiffusive motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters. *Proc Natl Acad Sci*. 2015;112(44):13675-13680. doi:10.1073/pnas.1508355112
43. Zerr DM, Boeckh M, Delaney C, et al. HHV-6 reactivation and associated sequelae after hematopoietic cell transplantation. *Biol Blood Marrow Transplant J Am Soc Blood Marrow Transplant*. 2012;18(11):1700-1708. doi:10.1016/j.bbmt.2012.05.012
44. Pichereau C, Desseaux K, Janin A, et al. The complex relationship between human herpesvirus 6 and acute graft-versus-host disease. *Biol Blood Marrow Transplant J Am Soc Blood Marrow Transplant*. 2012;18(1):141-144. doi:10.1016/j.bbmt.2011.07.018
45. Feghoul L, Chevret S, Cuinet A, et al. Adenovirus infection and disease in paediatric haematopoietic stem cell transplant patients: clues for antiviral pre-emptive treatment. *Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis*. 2015;21(7):701-709. doi:10.1016/j.cmi.2015.03.011
46. Abbas A, Taylor LJ, Collman RG, Bushman FD, ICTV Report Consortium. ICTV Virus Taxonomy Profile: Redondoviridae. *J Gen Virol*. 2021;102(1):001526. doi:10.1099/jgv.0.001526
47. Varsani A, Opriessnig T, Celer V, et al. Taxonomic update for mammalian anelloviruses (family Anelloviridae). *Arch Virol*. 2021;166(10):2943-2953. doi:10.1007/s00705-021-05192-x
48. Lim ES, Rodriguez C, Holtz LR. Amniotic fluid from healthy term pregnancies does not harbor a detectable microbial community. *Microbiome*. 2018;6(1):87. doi:10.1186/s40168-018-0475-7
49. Breitbart M, Haynes M, Kelley S, et al. Viral diversity and dynamics in an infant gut. *Res Microbiol*. 2008;159(5):367-373. doi:10.1016/j.resmic.2008.04.006
50. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe*. 2020;28(5):724-740.e8. doi:10.1016/j.chom.2020.08.003
51. Gregory AC, Sullivan MB, Segal LN, Keller BC. Smoking is associated with quantifiable differences in the human lung DNA virome and metabolome. *Respir Res*. 2018;19(1):174. doi:10.1186/s12931-018-0878-9
52. Garmaeva S, Gulyaeva A, Sinha T, et al. Stability of the human gut virome and effect of gluten-free diet. *Cell Rep*. 2021;35(7):109132. doi:10.1016/j.celrep.2021.109132
53. Minot S, Sinha R, Chen J, et al. The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Res*. 2011;21(10):1616-1625. doi:10.1101/gr.122705.111
54. Schulfer A, Santiago-Rodriguez TM, Ly M, et al. Fecal Viral Community Responses to High-Fat Diet in Mice. *mSphere*. 2020;5(1):e00833-19. doi:10.1128/mSphere.00833-19
55. Moreno-Gallego JL, Chou SP, Di Rienzi SC, et al. Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins. *Cell Host Microbe*. 2019;25(2):261-272.e5. doi:10.1016/j.chom.2019.01.019
56. Rothschild D, Weissbrod O, Barkan E, et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature*. 2018;555(7695):210-215. doi:10.1038/nature25973
57. Reyes A, Blanton LV, Cao S, et al. Gut DNA viromes of Malawian twins discordant

- for severe acute malnutrition. *Proc Natl Acad Sci U S A.* 2015;112(38):11941-11946. doi:10.1073/pnas.1514285112
58. Orth G. Genetics of epidermodysplasia verruciformis: Insights into host defense against papillomaviruses. *Semin Immunol.* 2006;18(6):362-374. doi:10.1016/j.smim.2006.07.008
59. Robles-Sikisaka R, Ly M, Boehm T, Naidu M, Salzman J, Pride DT. Association between living environment and human oral viral ecology. *ISME J.* 2013;7(9):1710-1724. doi:10.1038/ismej.2013.63
60. Kuroda Y, Watanabe K, Yamamoto T, et al. Pet Animals Were Infected with SARS-CoV-2 from Their Owners Who Developed COVID-19: Case Series Study. *Viruses.* 2023;15(10):2028. doi:10.3390/v15102028
61. Van Brussel K, Holmes EC. Zoonotic disease and virome diversity in bats. *Curr Opin Virol.* 2022;52:192-202. doi:10.1016/j.coviro.2021.12.008
62. Zuo T, Sun Y, Wan Y, et al. Human-Gut-DNA Virome Variations across Geography, Ethnicity, and Urbanization. *Cell Host Microbe.* 2020;28(5):741-751.e4. doi:10.1016/j.chom.2020.08.005
63. Holtz LR, Cao S, Zhao G, et al. Geographic variation in the eukaryotic virome of human diarrhea. *Virology.* 2014;468-470:556-564. doi:10.1016/j.virol.2014.09.012
64. Frontiers | The “Regulator” Function of Viruses on Ecosystem Carbon Cycling in the Anthropocene. Accessed October 3, 2024. <https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.858615/full>
65. Simmonds P, Becher P, Bukh J, et al. ICTV Virus Taxonomy Profile: Flaviviridae. *J Gen Virol.* 2017;98(1):2-3. doi:10.1099/jgv.0.000672
66. Global hepatitis report 2024: action for access in low- and middle-income countries. Accessed October 3, 2024. <https://www.who.int/publications/i/item/9789240091672>
67. Zibbell JE, Iqbal K, Patel RC, et al. Increases in hepatitis C virus infection related to injection drug use among persons aged ≤30 years - Kentucky, Tennessee, Virginia, and West Virginia, 2006-2012. *MMWR Morb Mortal Wkly Rep.* 2015;64(17):453-458.
68. Mazhnaya A, Meteliuk A, Barnard T, Zelenev A, Filippovich S, Altice FL. Implementing and scaling up HCV treatment services for people who inject drugs and other high risk groups in Ukraine: An evaluation of programmatic and treatment outcomes. *Int J Drug Policy.* 2017;47:187-195. doi:10.1016/j.drugpo.2017.07.023
69. Dev A, Sundararajan V, Sievert W. Ethnic and cultural determinants influence risk assessment for hepatitis C acquisition. *J Gastroenterol Hepatol.* 2004;19(7):792-798. doi:10.1111/j.1440-1746.2004.03381.x
70. Singh K, Mehta S. The clinical development process for a novel preventive vaccine: An overview. *J Postgrad Med.* 2016;62(1):4-11. doi:10.4103/0022-3859.173187
71. Yanagi M, Purcell RH, Emerson SU, Bukh J. Transcripts from a single full-length cDNA clone of hepatitis C virus are infectious when directly transfected into the liver of a chimpanzee. *Proc Natl Acad Sci U S A.* 1997;94(16):8738-8743. doi:10.1073/pnas.94.16.8738
72. Kolykhalov AA, Agapov EV, Blight KJ, Mihalik K, Feinstone SM, Rice CM. Transmission of hepatitis C by intrahepatic inoculation with transcribed RNA. *Science.* 1997;277(5325):570-574. doi:10.1126/science.277.5325.570
73. Shoukry NH, Sidney J, Sette A, Walker CM. Conserved hierarchy of helper T cell responses in a chimpanzee during primary and secondary hepatitis C virus infections. *J Immunol Baltim Md 1950.* 2004;172(1):483-492. doi:10.4049/jimmunol.172.1.483
74. Shoukry NH, Grakoui A, Houghton M, et al. Memory CD8+ T cells are required for protection from persistent hepatitis C virus infection. *J Exp Med.* 2003;197(12):1645-1655. doi:10.1084/jem.20030239
75. Knight A. The beginning of the end for chimpanzee experiments? *Philos Ethics Humanit Med.* 2008;3(1):16. doi:10.1186/1747-5341-3-16
76. U.S. Fish and Wildlife Service Finalizes Rule Listing All Chimpanzees as Endangered Under the Endangered Species Act | U.S. Fish & Wildlife Service. February

- 19, 2022. Accessed October 3, 2024. <https://www.fws.gov/press-release/2015-06/us-fish-and-wildlife-service-finalizes-rule-listing-all-chimpanzees>
77. Endangered and Threatened Wildlife and Plants; Listing All Chimpanzees as Endangered Species. Federal Register. June 16, 2015. Accessed October 3, 2024. <https://www.federalregister.gov/documents/2015/06/16/2015-14232/endangered-and-threatened-wildlife-and-plants-listing-all-chimpanzees-as-endangered-species>
78. NIH Will No Longer Support Biomedical Research on Chimpanzees. National Institutes of Health (NIH). November 18, 2015. Accessed October 3, 2024. <https://www.nih.gov/about-nih/who-we-are/nih-director/statements/nih-will-no-longer-support-biomedical-research-chimpanzees>
79. Berggren KA, Suzuki S, Ploss A. Animal Models Used in Hepatitis C Virus Research. *Int J Mol Sci.* 2020;21(11):3869. doi:10.3390/ijms21113869
80. Shi M, Lin XD, Vasilakis N, et al. Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses. *J Virol.* 2016;90(2):659-669. doi:10.1128/JVI.02036-15
81. Hartlage AS, Cullen JM, Kapoor A. The Strange, Expanding World of Animal Hepacivirus. *Annu Rev Virol.* 2016;3(1):53-75. doi:10.1146/annurev-virology-100114-055104
82. Simons JN, Pilot-Matias TJ, Leary TP, et al. Identification of two flavivirus-like genomes in the GB hepatitis agent. *Proc Natl Acad Sci U S A.* 1995;92(8):3401-3405. doi:10.1073/pnas.92.8.3401
83. Herron ICT, Laws TR, Nelson M. Marmosets as models of infectious diseases. *Front Cell Infect Microbiol.* 2024;14. doi:10.3389/fcimb.2024.1340017
84. Lvov DK, Alkhovsky SV, Zhirnov OP. 130th anniversary of virology. *Probl Virol.* 2022;67(5):357-384. doi:10.36233/0507-4088-140
85. Artenstein AW. The discovery of viruses: advancing science and medicine by challenging dogma. *Int J Infect Dis.* 2012;16(7):e470-e473. doi:10.1016/j.ijid.2012.03.005
86. Ivanowski D. Concerning the mosaic disease of the tobacco plant. Published online 1942.
87. Mayer A. Concerning the mosaic disease of Tobacco. Published online 1942.
88. Bos L. Beijerinck's work on tobacco mosaic virus: historical context and legacy. *Philos Trans R Soc Lond B Biol Sci.* 1999;354(1383):675-685.
89. Beijerinck MW. Concerning a contagium vivum fluidum as a cause of the spot-disease of Tobacco leaves. Published online 1942.
90. Loeffler F, Frosch P. Summarischer Bericht über die Ergebnisse der Untersuchungen der Commission zur Erforschung der Maul-und Klauenseuche. *Dtsch Med Wochenschr.* 1897;23:617-617.
91. Stanley WM, Loring HS. The isolation of crystalline tobacco mosaic virus protein from diseased tomato plants. *Science.* 1936;83(2143):85-85.
92. Bernal JD, Fankuchen I. X-ray and crystallographic studies of plant virus preparations: I. Introduction and preparation of specimens II. Modes of aggregation of the virus particles. *J Gen Physiol.* 1941;25(1):111.
93. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.* 1977;74(12):5463-5467. doi:10.1073/pnas.74.12.5463
94. Saiki RK, Scharf S, Faloona F, et al. Enzymatic Amplification of β-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia. *Science.* 1985;230(4732):1350-1354. doi:10.1126/science.2999980
95. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. *Hum Immunol.* 2021;82(11):801-811. doi:10.1016/j.humimm.2021.02.012
96. Breitbart M, Salamon P, Andresen B, et al. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A.* 2002;99(22):14250-14255. doi:10.1073/pnas.202488399
97. Leiner BM, Cerf VG, Clark DD, et al. A brief history of the internet. *SIGCOMM Comput Commun Rev.* 2009;39(5):22-31. doi:10.1145/1629607.1629613

98. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Res.* 2011;39(suppl_1):D19-D21. doi:10.1093/nar/gkq1019
99. Getting Started. Accessed October 3, 2024.
<https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/>
100. Chong LC, Lauber C. Viroid-like RNA-dependent RNA polymerase-encoding ambiviruses are abundant in complex fungi. *Front Microbiol.* 2023;14. doi:10.3389/fmicb.2023.1144003
101. Lauber C, Vaas J, Klingler F, et al. Deep mining of the Sequence Read Archive reveals bipartite coronavirus genomes and inter-family Spike glycoprotein recombination. Published online October 20, 2021:2021.10.20.465146. doi:10.1101/2021.10.20.465146
102. Edgar RC, Taylor B, Lin V, et al. Petabase-scale sequence alignment catalyses viral discovery. *Nature.* 2022;602(7895):142-147. doi:10.1038/s41586-021-04332-2
103. Lauber C, Seitz S, Mattei S, et al. Deciphering the Origin and Evolution of Hepatitis B Viruses by Means of a Family of Non-enveloped Fish Viruses. *Cell Host Microbe.* 2017;22(3):387-399.e6. doi:10.1016/j.chom.2017.07.019
104. Lauber C, Seitz S. Opportunities and Challenges of Data-Driven Virus Discovery. *Biomolecules.* 2022;12(8):1073. doi:10.3390/biom12081073
105. Wertheim JO, Hostager R, Ryu D, et al. Discovery of Novel Herpes Simplexviruses in Wild Gorillas, Bonobos, and Chimpanzees Supports Zoonotic Origin of HSV-2. *Mol Biol Evol.* 2021;38(7):2818-2830. doi:10.1093/molbev/msab072
106. Shi M, Lin XD, Tian JH, et al. Redefining the invertebrate RNA virosphere. *Nature.* 2016;540(7634):539-543. doi:10.1038/nature20167
107. Shi M, Lin XD, Chen X, et al. The evolutionary history of vertebrate RNA viruses. *Nature.* 2018;556(7700):197-202. doi:10.1038/s41586-018-0012-7
108. Nga PT, Parquet M del C, Lauber C, et al. Discovery of the First Insect Nidovirus, a Missing Evolutionary Link in the Emergence of the Largest RNA Virus Genomes. *PLOS Pathog.* 2011;7(9):e1002215. doi:10.1371/journal.ppat.1002215
109. Fu P, Wu Y, Zhang Z, Qiu Y, Wang Y, Peng Y. VIGA: a one-stop tool for eukaryotic virus identification and genome assembly from next-generation-sequencing data. *Brief Bioinform.* 2023;25(1):bbad444. doi:10.1093/bib/bbad444
110. Chen G, Tang X, Shi M, Sun Y. VirBot: an RNA viral contig detector for metagenomic data. *Bioinforma Oxf Engl.* 2023;39(3):btad093. doi:10.1093/bioinformatics/btad093
111. Zhao G, Wu G, Lim ES, et al. VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology.* 2017;503:21-30. doi:10.1016/j.virol.2017.01.005
112. Cheng DQ, Kolundžija S, Lauro FM. Global phylogenetic analysis of the RNA-dependent RNA polymerase with OrViT (OrthornaVirae Tree). *Front Virol.* 2022;2. doi:10.3389/fviro.2022.981177
113. Pérot P, Bigot T, Temmam S, Regnault B, Eloit M. Microseek: A Protein-Based Metagenomic Pipeline for Virus Diagnostic and Discovery. *Viruses.* 2022;14(9):1990. doi:10.3390/v14091990
114. RdRp-based sensitive taxonomic classification of RNA viruses for metagenomic data | *Briefings in Bioinformatics* | Oxford Academic. Accessed October 3, 2024.
<https://academic.oup.com/bib/article/23/2/bbac011/6523411?login=false>
115. Alawi M, Burkhardt L, Indenbirken D, et al. DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples. *Sci Rep.* 2019;9(1):16841. doi:10.1038/s41598-019-52881-4
116. Zayed AA, Wainaina JM, Dominguez-Huerta G, et al. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science.* 2022;376(6589):156-162. doi:10.1126/science.abm5847
117. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol.* 2011;7(10):e1002195. doi:10.1371/journal.pcbi.1002195
118. Eddy SR. Profile hidden Markov models. *Bioinforma Oxf Engl.* 1998;14(9):755-

763. doi:10.1093/bioinformatics/14.9.755
119. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2
120. C C, G C, V A, et al. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10. doi:10.1186/1471-2105-10-421
121. Oliveira LS, Gruber A. Rational design of profile hidden Markov models for viral classification and discovery. *Bioinforma Internet.* Published online 2021.
122. Park J, Karplus K, Barrett C, et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods1. *J Mol Biol.* 1998;284(4):1201-1210. doi:10.1006/jmbi.1998.2221
123. Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences increase the detection of homology between sequences. *J Mol Biol.* 1997;273(1):349-354. doi:10.1006/jmbi.1997.1288
124. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389-3402.
125. John B, Sali A. Detection of homologous proteins by an intermediate sequence search. *Protein Sci.* 2004;13(1):54-62. doi:10.1110/ps.03335004
126. Lauber C, Zhang X, Vaas J, et al. Deep mining of the Sequence Read Archive reveals major genetic innovations in coronaviruses and other nidoviruses of aquatic vertebrates. *PLOS Pathog.* 2024;20(4):e1012163. doi:10.1371/journal.ppat.1012163
127. Wyler E, Lauber C, Manukyan A, et al. Comprehensive profiling of wastewater viromes by genomic sequencing. Published online December 19, 2022:2022.12.16.520800. doi:10.1101/2022.12.16.520800
128. Aghebatrafat AA, Lauber C, Merkel K, et al. Evolutionary Insight into the Association between New Jersey Polyomavirus and Humans. *Viruses.* 2023;15(11):2248. doi:10.3390/v15112248
129. Balaji S, Srinivasan N. Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: Inferences on protein evolution. *J Biosci.* 2007;32(1):83-96. doi:10.1007/s12038-007-0008-1
130. Krupovic M, Bamford DH. Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. *Curr Opin Virol.* 2011;1(2):118-124. doi:10.1016/j.coviro.2011.06.001
131. Domingo E, García-Crespo C, Lobo-Vega R, Perales C. Mutation Rates, Mutation Frequencies, and Proofreading-Repair Activities in RNA Virus Genetics. *Viruses.* 2021;13(9):1882. doi:10.3390/v13091882
132. Abroi A, Gough J. Are viruses a source of new protein folds for organisms?— virosphere structure space and evolution. *Bioessays.* 2011;33(8):626-635.
133. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol.* 2001;307(4):1113-1143. doi:10.1006/jmbi.2001.4513
134. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods.* 2019;16(7):603-606. doi:10.1038/s41592-019-0437-4
135. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 1986;5(4):823-826. doi:10.1002/j.1460-2075.1986.tb04288.x
136. Balaji S, Srinivasan N. Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. *Protein Eng.* 2001;14(4):219-226. doi:10.1093/protein/14.4.219
137. Rochette NC, Rivera-Colón AG, Walsh J, Sanger TJ, Campbell-Staton SC, Catchen JM. On the causes, consequences, and avoidance of PCR duplicates: Towards a theory of library complexity. *Mol Ecol Resour.* 2023;23(6):1299-1318. doi:10.1111/1755-0998.13800

138. Rappoport N, Linial M. Viral Proteins Acquired from a Host Converge to Simplified Domain Architectures. *PLOS Comput Biol*. 2012;8(2):e1002364. doi:10.1371/journal.pcbi.1002364
139. Gorbalenya AE, Krupovic M, Mushegian A, et al. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat Microbiol*. 2020;5(5):668-674. doi:10.1038/s41564-020-0709-x
140. Koonin EV, Kuhn JH. Code assigned: 2019.006 G. Published online 2019.
141. Kuhn JH, Abe J, Adkins S, et al. Annual (2023) taxonomic update of RNA-directed RNA polymerase-encoding negative-sense RNA viruses (realm Riboviria: kingdom Orthornavirae: phylum Negarnaviricota). *J Gen Virol*. 2023;104(8):001864. doi:10.1099/jgv.0.001864
142. Current ICTV Taxonomy Release | ICTV. Accessed October 4, 2024. <https://ictv.global/taxonomy>
143. Wolf YI, Kazlauskas D, Iranzo J, et al. Origins and Evolution of the Global RNA Virome. *mBio*. 2018;9(6):e02329-18. doi:10.1128/mBio.02329-18
144. Kuhn JH, Botella L, de la Peña M, et al. Ambiviricota, a novel ribovirian phylum for viruses with viroid-like properties. *J Virol*. 2024;98(7):e0083124. doi:10.1128/jvi.00831-24
145. Olschoorn RC, Van Duin J. Leviviridae—Positive Sense RNA Viruses—Positive Sense RNA Viruses (2011)—International Committee on Taxonomy of Viruses (ICTV). *Int Comm Taxon Viruses ICTV*. Published online 2017.
146. Zell R, Delwart E, Gorbalenya AE, et al. ICTV Virus Taxonomy Profile: Picornaviridae. *J Gen Virol*. 2017;98(10):2421-2422. doi:10.1099/jgv.0.000911
147. Fuchs M, Hily JM, Petrik K, et al. ICTV Virus Taxonomy Profile: Secoviridae 2022. *J Gen Virol*. 2022;103(12):001807. doi:10.1099/jgv.0.001807
148. Valles SM, Chen Y, Firth AE, et al. ICTV Virus Taxonomy Profile: Iflavirusidae. *J Gen Virol*. 2017;98(4):527-528. doi:10.1099/jgv.0.000757
149. Ziebuhr J, Baric RS, Baker S, et al. Reorganization of the family Coronaviridae into two families, Coronaviridae (including the current subfamily Coronavirinae and the new subfamily Letovirinae) and the new family Tobaniviridae (accommodating the current subfamily Torovirinae and three other subfamilies), revision of the genus rank structure and introduction of a new subgenus rank. *Propos 2017013 0808 2018 Int Comm Taxon Viruses*. Published online 2017.
150. Walker PJ, Cowley JA, Dong X, et al. ICTV Virus Taxonomy Profile: Roniviridae. *J Gen Virol*. 2021;102(1):001514. doi:10.1099/jgv.0.001514
151. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ. *Nidovirales*: Evolving the largest RNA virus genome. *Virus Res*. 2006;117(1):17-37. doi:10.1016/j.virusres.2006.01.017
152. Simmonds P, Becher P, Bukh J, et al. ICTV Virus Taxonomy Profile: Flaviviridae. *J Gen Virol*. 2017;98(1):2-3. doi:10.1099/jgv.0.000672
153. Purdy MA, Drexler JF, Meng XJ, et al. ICTV Virus Taxonomy Profile: Hepeviridae 2022. *J Gen Virol*. 2022;103(9):001778. doi:10.1099/jgv.0.001778
154. ICTV Virus Taxonomy Profile: Matonaviridae 2022 | Microbiology Society. Accessed October 4, 2024. <https://www.microbiologyresearch.org/content/journal/jgv/10.1099/jgv.0.001817>
155. Mata CP, Luque D, Gómez-Blanco J, et al. Acquisition of functions on the outer capsid surface during evolution of double-stranded RNA fungal viruses. *PLoS Pathog*. 2017;13(12):e1006755.
156. Poranen MM, Mäntynen S, ICTV Report Consortium. ICTV Virus Taxonomy Profile: Cystoviridae. *J Gen Virol*. 2017;98(10):2423-2424. doi:10.1099/jgv.0.000928
157. Matthijnssens J, Attoui H, Bányai K, et al. ICTV Virus Taxonomy Profile: Sedoreoviridae 2022. *J Gen Virol*. 2022;103(10):001782. doi:10.1099/jgv.0.001782
158. Diot C, Cosentino G, Rameix-Welti MA. Ribonucleoprotein transport in Negative Strand RNA viruses. *Biol Cell*. 2023;115(1):2200059. doi:10.1111/boc.202200059
159. Gatherer D, Depledge DP, Hartley CA, et al. ICTV Virus Taxonomy Profile: Herpesviridae 2021. *J Gen Virol*. 2021;102(10):001673. doi:10.1099/jgv.0.001673

160. Rymo L, Forsblom S. Cleavage of Epstein-Barr virus DNA by restriction endonucleases EcoRI, HindIII and BamI. *Nucleic Acids Res.* 1978;5(4):1387-1402. doi:10.1093/nar/5.4.1387
161. Russo JJ, Bohenzky RA, Chien MC, et al. Nucleotide sequence of the Kaposi sarcoma-associated herpesvirus (HHV8). *Proc Natl Acad Sci U S A.* 1996;93(25):14862-14867. doi:10.1073/pnas.93.25.14862
162. Njau EP, Domelevo Entfellner JB, Machuka EM, et al. The first genotype II African swine fever virus isolated in Africa provides insight into the current Eurasian pandemic. *Sci Rep.* 2021;11(1):13081. doi:10.1038/s41598-021-92593-2
163. McInnes CJ, Damon IK, Smith GL, et al. ICTV Virus Taxonomy Profile: Poxviridae 2023. *J Gen Virol.* 2023;104(5):001849. doi:10.1099/jgv.0.001849
164. Benkő M, Aoki K, Arnberg N, et al. ICTV Virus Taxonomy Profile: Adenoviridae 2022. *J Gen Virol.* 2022;103(3):001721. doi:10.1099/jgv.0.001721
165. Krupovic M, Varsani A, Kazlauskas D, et al. Cressdnnaviricota: a Virus Phylum Unifying Seven Families of Rep-Encoding Viruses with Single-Stranded, Circular DNA Genomes. *J Virol.* 2020;94(12):e00582-20. doi:10.1128/JVI.00582-20
166. Fiallo-Olivé E, Lett JM, Martin DP, et al. ICTV Virus Taxonomy Profile: Geminiviridae 2021. *J Gen Virol.* 2021;102(12):001696. doi:10.1099/jgv.0.001696
167. Thomas JE, Gronenborn B, Harding RM, et al. ICTV Virus Taxonomy Profile: Nanoviridae. *J Gen Virol.* 2021;102(3):001544. doi:10.1099/jgv.0.001544
168. Varsani A, Krupovic M, Varsani A. Code assigned: 2022.002 F.
169. Breitbart M, Delwart E, Rosario K, Segalés J, Varsani A, ICTV Report Consortium. ICTV Virus Taxonomy Profile: Circoviridae. *J Gen Virol.* 2017;98(8):1997-1998. doi:10.1099/jgv.0.000871
170. Knezevic P, Adriaenssens EM, ICTV Report Consortium. ICTV Virus Taxonomy Profile: Plectroviridae. *J Gen Virol.* 2021;102(5):001597. doi:10.1099/jgv.0.001597
171. Krupovic M, Turner D, Morozova V, et al. Bacterial Viruses Subcommittee and Archaeal Viruses Subcommittee of the ICTV: update of taxonomy changes in 2021. *Arch Virol.* 2021;166(11):3239-3244. doi:10.1007/s00705-021-05205-9
172. Knezevic P, Adriaenssens EM, ICTV Report Consortium. ICTV Virus Taxonomy Profile: Inoviridae. *J Gen Virol.* 2021;102(7):001614. doi:10.1099/jgv.0.001614
173. Doore SM, Fane BA. The microviridae: Diversity, assembly, and experimental evolution. *Virology.* 2016;491:45-55. doi:10.1016/j.virol.2016.01.020
174. Krupovic M, Koonin EV. Evolution of eukaryotic single-stranded DNA viruses of the Bidnaviridae family from genes of four other groups of widely different viruses. *Sci Rep.* 2014;4(1):5347. doi:10.1038/srep05347
175. Moens U, Calvignac-Spencer S, Lauber C, et al. ICTV Virus Taxonomy Profile: Polyomaviridae. *J Gen Virol.* 2017;98(6):1159-1160. doi:10.1099/jgv.0.000839
176. Van Doorslaer K, Chen Z, Bernard HU, et al. ICTV Virus Taxonomy Profile: Papillomaviridae. *J Gen Virol.* 2018;99(8):989-990. doi:10.1099/jgv.0.001105
177. Cotmore SF, Agbandje-McKenna M, Canuti M, et al. ICTV Virus Taxonomy Profile: Parvoviridae. *J Gen Virol.* 2019;100(3):367-368. doi:10.1099/jgv.0.001212
178. Liu Y, Dyall-Smith M, Oksanen HM. ICTV Virus Taxonomy Profile: Pleolipoviridae 2022. *J Gen Virol.* 2022;103(11):001793. doi:10.1099/jgv.0.001793
179. Morshed K, Polz-Gruszka D, Szymański M, Polz-Dacewicz M. Human Papillomavirus (HPV) – Structure, epidemiology and pathogenesis. *Otolaryngol Pol.* 2014;68(5):213-219. doi:10.1016/j.otpol.2014.06.001
180. Pv de MB, Gp O, Js A. "Yaraviridae": a proposed new family of viruses infecting Acanthamoeba castellanii. *Arch Virol.* 2022;167(2). doi:10.1007/s00705-021-05326-1
181. Delmas B, Attoui H, Ghosh S, et al. ICTV virus taxonomy profile: Birnaviridae. *J Gen Virol.* 2019;100(1):5-6. doi:10.1099/jgv.0.001185
182. Adams MJ, Carstens EB. Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2012). *Arch Virol.* 2012;157:1411-1422.
183. Magnus L, Mason WS, Taylor J, et al. ICTV Virus Taxonomy Profile: Hepadnaviridae. *J Gen Virol.* 2020;101(6):571-572. doi:10.1099/jgv.0.001415

184. Coffin J, Blomberg J, Fan H, et al. ICTV Virus Taxonomy Profile: Retroviridae 2021. *J Gen Virol.* 2021;102(12):001712. doi:10.1099/jgv.0.001712
185. Adriaenssens EM, Roux S, Brister JR, et al. Guidelines for public database submission of uncultivated virus genome sequences for taxonomic classification. *Nat Biotechnol.* 2023;41(7):898-902. doi:10.1038/s41587-023-01844-2
186. Alves JMP, de Oliveira AL, Sandberg TOM, et al. GenSeed-HMM: A Tool for Progressive Assembly Using Profile HMMs as Seeds and its Application in Alpavirinae Viral Discovery from Metagenomic Data. *Front Microbiol.* 2016;7. doi:10.3389/fmicb.2016.00269
187. Huang X, Madan A. CAP3: A DNA Sequence Assembly Program. *Genome Res.* 1999;9(9):868-877. doi:10.1101/gr.9.9.868
188. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 2005;437(7057):376-380. doi:10.1038/nature03959
189. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008;18(5):821-829. doi:10.1101/gr.074492.107
190. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler | GigaScience | Oxford Academic. Accessed October 5, 2024. <https://academic.oup.com/gigascience/article/1/1/2047-217X-1-18/2656146>
191. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABYSS: A parallel assembler for short read sequence data. *Genome Res.* 2009;19(6):1117-1123. doi:10.1101/gr.089532.108
192. Venkataraman S, Prasad BVLS, Selvarajan R. RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution. *Viruses.* 2018;10(2):76. doi:10.3390/v10020076
193. J G, Jf Q, Y S, et al. Review, Evaluation, and Directions for Gene-Targeted Assembly for Ecological Analyses of Metagenomes. *Front Genet.* 2019;10. doi:10.3389/fgene.2019.00957
194. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2021.
195. Ruff S. Virusparies: Data Visualisations for VirusHunterGatherer hittables output. Published online 2024. <https://github.com/SergejRuff/Virusparies>
196. Christiansen T, Wall L, Orwant J. *Programming Perl: Unmatched Power for Text Processing and Scripting.* O'Reilly Media, Inc.; 2012.
197. Mölder F, Jablonski KP, Letcher B, et al. Sustainable data analysis with Snakemake. Published online January 18, 2021. doi:10.12688/f1000research.29032.1
198. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinforma Oxf Engl.* 2018;34(17):i884-i890. doi:10.1093/bioinformatics/bty560
199. Li H. lh3/seqtk. Published online October 3, 2024. Accessed October 4, 2024. <https://github.com/lh3/seqtk>
200. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* 2000;16(6):276-277.
201. Poisson F, Roingeard P, Baillou A, et al. Characterization of RNA-binding domains of hepatitis delta antigen. *J Gen Virol.* 1993;74(11):2473-2478.
202. Chang MF, Chen CH, Lin SL, Chen CJ, Chang SC. Functional domains of delta antigens and viral RNA required for RNA packaging of hepatitis delta virus. *J Virol.* 1995;69(4):2508-2514.
203. Posthuma CC, Te Velthuis AJ, Snijder EJ. Nidovirus RNA polymerases: complex enzymes handling exceptional RNA genomes. *Virus Res.* 2017;234:58-73.
204. Webby R, Kalmakoff J. Sequence comparison of the major capsid protein gene from 18 diverse iridoviruses. *Arch Virol.* 1998;143:1949-1966.
205. Chee M, Rudolph SA, Plachter B, Barrell B, Jahn G. Identification of the major capsid protein gene of human cytomegalovirus. *J Virol.* 1989;63(3):1345-1353.
206. Castellano MM, Sanz-Burgos AP, Gutiérrez C. Initiation of DNA replication in a eukaryotic rolling-circle replicon: identification of multiple DNA-protein complexes at the geminivirus origin. *J Mol Biol.* 1999;290(3):639-652.

207. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*. 2016;4:e2584. doi:10.7717/peerj.2584
208. Zhang H. Alignment of BLAST high-scoring segment pairs based on the longest increasing subsequence algorithm. *Bioinformatics*. 2003;19(11):1391-1396. doi:10.1093/bioinformatics/btg168
209. Jones P, Binns D, Chang HY, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30(9):1236-1240. doi:10.1093/bioinformatics/btu031
210. Paysan-Lafosse T, Blum M, Chuguransky S, et al. InterPro in 2022. *Nucleic Acids Res.* 2023;51(D1):D418-D427. doi:10.1093/nar/gkac993
211. Thomas Hackl and Markus J. Ankenbrand and Bart van Adrichem. gggenomes: A Grammar of Graphics for Comparative Genomics. Published online 2024.
212. ICTV Virus Taxonomy Profile: Potyviridae 2022 | Microbiology Society. Accessed October 4, 2024.
<https://www.microbiologyresearch.org/content/journal/jgv/10.1099/jgv.0.001738>
213. Hillman BI, Cohen AB. Totiviruses (Totiviridae). In: *Encyclopedia of Virology: Volume 1-5, Fourth Edition*. Elsevier; 2020:648-657.
214. Kuhn JH, Babaian A, Bergner LM, et al. ICTV Virus Taxonomy Profile: Kolmioviridae 2024. *J Gen Virol*. 2024;105(2):001963. doi:10.1099/jgv.0.001963
215. Brinton MA, Gulyaeva AA, Balasuriya UBR, et al. ICTV Virus Taxonomy Profile: Arteriviridae 2021. *J Gen Virol*. 2021;102(8):001632. doi:10.1099/jgv.0.001632
216. Delmas B, Attoui H, Ghosh S, et al. ICTV virus taxonomy profile: Picobirnaviridae. *J Gen Virol*. 2019;100(2):133-134. doi:10.1099/jgv.0.001186
217. Gough RE, McNulty MS. Astroviridae. *Poult Dis*. Published online 2008:392-397. doi:10.1016/B978-0-7020-2862-5.50038-6
218. Adler BA, Chamakura K, Carion H, et al. Multicopy suppressor screens reveal convergent evolution of single-gene lysis proteins. *Nat Chem Biol*. 2023;19(6):759-766. doi:10.1038/s41589-023-01269-7
219. Callanan J, Stockdale SR, Adriaenssens EM, et al. Rename one class (Leviviricetes-formerly Allassoviricetes), rename one order (Norzivirales-formerly Levivirales), create one new order (Timlovirales), and expand the class to a total of six families, 420 genera and 883 species. *RNA Bacteriophages Divers Abundance Appl*. Published online 2021:11.
220. Callanan J, Stockdale SR, Adriaenssens EM, et al. Leviviricetes: expanding and restructuring the taxonomy of bacteria-infecting single-stranded RNA viruses. *Microb Genomics*. 2021;7(11):000686. doi:10.1099/mgen.0.000686
221. Scheid A. Chapter 16 Paramyxoviridae. In: Nermut MV, Steven AC, eds. *Perspectives in Medical Virology*. Vol 3. Animal Virus Structure. Elsevier; 1987:233-252. doi:10.1016/S0168-7069(08)70098-3
222. ICTV Virus Taxonomy Profile: Rhabdoviridae | Microbiology Society. Accessed October 4, 2024.
<https://www.microbiologyresearch.org/content/journal/jgv/10.1099/jgv.0.001020>
223. Sasaya T, Palacios G, Briese T, et al. ICTV Virus Taxonomy Profile: Phenuiviridae 2023. *J Gen Virol*. 2023;104(9):001893. doi:10.1099/jgv.0.001893
224. Sit TL, Lommel SA. Tombusviridae. In: eLS. John Wiley & Sons, Ltd; 2015:1-9. doi:10.1002/9780470015902.a0000756.pub3
225. ICTV Virus Taxonomy Profile: Partitiviridae | Microbiology Society. Accessed October 4, 2024.
<https://www.microbiologyresearch.org/content/journal/jgv/10.1099/jgv.0.000985>
226. NextSeq 500 sequencing; RNASeq of MARC-145 or MA-104 cells infected w... - SRA - NCBI. Accessed October 4, 2024.
[https://www.ncbi.nlm.nih.gov/sra/ERX5747122\[accn\]](https://www.ncbi.nlm.nih.gov/sra/ERX5747122[accn])
227. Cook GM, Brown K, Shang P, et al. Ribosome profiling of porcine reproductive and respiratory syndrome virus reveals novel features of viral gene expression. Rodnina MV, Davenport MP, eds. *eLife*. 2022;11:e75668. doi:10.7554/eLife.75668
228. Sus scrofa domesticus saliva - SRA - NCBI. Accessed October 4, 2024.

- [https://www.ncbi.nlm.nih.gov/sra/SRX13519353\[accn\]](https://www.ncbi.nlm.nih.gov/sra/SRX13519353[accn])
229. Chauhan RP, San JE, Gordon ML. Metagenomic Analysis of RNA Fraction Reveals the Diversity of Swine Oral Virome on South African Backyard Swine Farms in the uMgungundlovu District of KwaZulu-Natal Province. *Pathogens*. 2022;11(8):927. doi:10.3390/pathogens11080927
230. small RNA sequencing of Bos taurus - SRA - NCBI. Accessed October 4, 2024. [https://www.ncbi.nlm.nih.gov/sra/SRX8824842\[accn\]](https://www.ncbi.nlm.nih.gov/sra/SRX8824842[accn])
231. Sukreet S, Braga CP, An TT, et al. Isolation of extracellular vesicles from byproducts of cheesemaking by tangential flow filtration yields heterogeneous fractions of nanoparticles. *J Dairy Sci*. 2021;104(9):9478-9493.
232. ICTV Virus Taxonomy Profile: Pneumoviridae | Microbiology Society. Accessed October 4, 2024. <https://www.microbiologyresearch.org/content/journal/jgv/10.1099/jgv.0.000959>
233. ICTV Virus Taxonomy Profile: Bornaviridae | Microbiology Society. Accessed October 4, 2024. <https://www.microbiologyresearch.org/content/journal/jgv/10.1099/jgv.0.001613>
234. Laenen L, Vergote V, Calisher CH, et al. Hantaviridae: Current classification and future perspectives. *Viruses*. 2019;11(9):788.
235. Kuhn JH, Schmaljohn CS. A brief history of Bunyaviral family Hantaviridae. *Diseases*. 2023;11(1):38.
236. SUOHC1004 - SRA - NCBI. Accessed October 4, 2024. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR14579882>
237. RNA-Seq of sus scrofa: white blood cells - SRA - NCBI. Accessed October 4, 2024. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18700679>
238. Wu Q, Han Y, Wu X, et al. Integrated time-series transcriptomic and metabolomic analyses reveal different inflammatory and adaptive immune responses contributing to host resistance to PRRSV. *Front Immunol*. 2022;13. doi:10.3389/fimmu.2022.960709
239. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ. Nidovirales: Evolving the largest RNA virus genome. *Virus Res*. 2006;117(1):17-37. doi:10.1016/j.virusres.2006.01.017
240. Vanmechelen B, Bletsa M, Laenen L, et al. Discovery and genome characterization of three new Jeilongviruses, a lineage of paramyxoviruses characterized by their unique membrane proteins. *BMC Genomics*. 2018;19(1):617. doi:10.1186/s12864-018-4995-0
241. Warren WC, Jasinska AJ, García-Pérez R, et al. The genome of the vervet (*Chlorocebus aethiops sabaeus*). *Genome Res*. 2015;25(12):1921-1933. doi:10.1101/gr.192922.115
242. Chauhan RP, San JE, Gordon ML. Metagenomic Analysis of RNA Fraction Reveals the Diversity of Swine Oral Virome on South African Backyard Swine Farms in the uMgungundlovu District of KwaZulu-Natal Province. *Pathogens*. 2022;11(8):927. doi:10.3390/pathogens11080927
243. He B, Li Z, Yang F, et al. Virome Profiling of Bats from Myanmar by Metagenomic Analysis of Tissue Samples Reveals More Novel Mammalian Viruses. *PLoS ONE*. 2013;8(4):e61950. doi:10.1371/journal.pone.0061950
244. Wertheim JO, Hostager R, Ryu D, et al. Discovery of Novel Herpes Simplexviruses in Wild Gorillas, Bonobos, and Chimpanzees Supports Zoonotic Origin of HSV-2. *Mol Biol Evol*. 2021;38(7):2818-2830. doi:10.1093/molbev/msab072
245. Käfer S, Paraskevopoulou S, Zirkel F, et al. Re-assessing the diversity of negative strand RNA viruses in insects. *PLoS Pathog*. 2019;15(12):e1008224. doi:10.1371/journal.ppat.1008224
246. Sczyrba A, Hofmann P, Belmann P, et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat Methods*. 2017;14(11):1063-1071. doi:10.1038/nmeth.4458
247. Selitsky SR, Marron D, Hollern D, et al. Virus expression detection reveals RNA-sequencing contamination in TCGA. *BMC Genomics*. 2020;21(1):79. doi:10.1186/s12864-

020-6483-6

248. Host Subtraction, Filtering and Assembly Validations for Novel Viral Discovery Using Next Generation Sequencing Data - PMC. Accessed October 5, 2024. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4476701/>
249. Kojima S, Kamada AJ, Parrish NF. Virus-derived variation in diverse human genomes. *PLOS Genet.* 2021;17(4):e1009324. doi:10.1371/journal.pgen.1009324
250. Kwok KTT, Nieuwenhuijse DF, Phan MVT, Koopmans MPG. Virus Metagenomics in Farm Animals: A Systematic Review. *Viruses.* 2020;12(1):107. doi:10.3390/v12010107
251. Butler JE, Lager KM, Golde W, et al. Porcine reproductive and respiratory syndrome (PRRS): an immune dysregulatory pandemic. *Immunol Res.* 2014;59(1):81-108. doi:10.1007/s12026-014-8549-5
252. Raev S, Yuzhakov A, Bulgakov A, et al. An Outbreak of a Respiratory Disorder at a Russian Swine Farm Associated with the Co-Circulation of PRRSV1 and PRRSV2. *Viruses.* 2020;12(10):1169. doi:10.3390/v12101169
253. Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Published online 2016.
254. Richard Iannone and Joe Cheng and Barret Schloerke and Ellis Hughes and Alexandra Lauer and JooYoung Seo and Ken Brevoort and Olivier Roy. gt: Easily Create Presentation-Ready Display Tables. Published online 2024. <https://gt.rstudio.com/> / <https://github.com/rstudio/gt>
255. Ivie P, Thain D. Reproducibility in Scientific Computing. *ACM Comput Surv.* 2018;51(3):63:1-63:36. doi:10.1145/3186266
256. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature.* 2016;533(7604):452-454. doi:10.1038/533452a
257. Marwick B, Boettiger C, Mullen L. Packaging Data Analytical Work Reproducibly Using R (and Friends). *Am Stat.* 2018;72(1):80-88. doi:10.1080/00031305.2017.1375986
258. Simmonds P, Adriaenssens EM, Zerbini FM, et al. Four principles to establish a universal virus taxonomy. *PLOS Biol.* 2023;21(2):e3001922. doi:10.1371/journal.pbio.3001922

6 Acknowledgements

I would like to take the time to thank some people without whom none of this would have been possible.

First, I am deeply grateful to Prof. Dr. Chris Lauber. Thank you for letting me do my master's thesis with you. Thank you for the help and tips you gave me when I had issues. You are an incredible supervisor. Furthermore, I appreciate all the discussions we had during our weekly Thursday meetings (and outside of them as well). I learned a lot about virus discovery, and specifically about data-driven virus discovery. It was an honor to have the opportunity to write my thesis in this field, and an even greater honor to contribute to it through Virusparies and the two works on which I have the privilege of being a co-author. Thank you for the experience. I hope to have the chance to work with you and your team again.

To Prof. Dr. Helena U. Zacharias, my thanks for taking on the role of second supervisor on short notice. I would also like to thank both of you, Chris and Helena, for being excellent lecturers during the master's program. Thank you, Chris, for providing me with introductory knowledge in computer science, and thank you, Helena, for offering an amazing introductory course on data science and data analysis.

I hope to stay in contact with both of you in the future as I start my PhD at TiHo with Prof. Dr. Klaus Jung. I would also like to express my deepest appreciation to him. He understood that a significant portion of my attention needed to be devoted to my master's thesis rather than my role as a student assistant offering both support and understanding. Here's to a great future boss and a great humane being.

I also did not forget the great contributions of Li Chuin Chong, who guided me throughout this journey. She showed me how to run all the software used in this thesis, provided help and tips when needed, and proofread the thesis offering valuable suggestions. Thank you for everything.

A big THANKS goes out to Prof. Dr. Richard Taubert and his working group for providing the liver sequencing data used in this thesis, as well as to the entire scientific community that uploads their sequencing data on SRA, a small part of which I utilized here.

Acknowledgements

Another person who read my text and provided support is my best friend of 12 years, Vua Lam. Thank you for that, and thank you for being one of the best people I have had the pleasure to know. Here's to a great friendship - всегда вместе, друг.

Last but not least, I would like to thank the rest of the computational virology team, Isra Nurhassen and Nicolai Böker, as well as the entire Twincore community for the great working atmosphere, the kind people, and the interesting presentations held every Monday and Wednesday.

As a whole, THANK YOU ALL!

7 Appendix

Table 8: VirusHunter and VirusGatherer dependencies

Name	Version	Build
_libgcc_mutex	0.1	main
_openmp_mutex	5.1	1_gnu
bcrypt	3.2.0	py36h7b6447c_0
blas	1.0	mkl
blast	2.12.0	pl5262h3289130_0
bowtie2	2.2.5	py36h6bb024c_5
brotli	0.7.0	py36h27cf23_1003
bzip2	1.0.8	h7b6447c_0
c-ares	1.19.0	h5eee18b_0
ca-certificates	2024.3.11	h06a4308_0
cap3	10.2011	h779adbc_3
certifi	2021.5.30	py36h06a4308_0
cffi	1.14.6	py36h400218f_0
charset-normalizer	2.0.4	pyhd3eb1b0_0
cryptography	35.0.0	py36hd23ed53_0
curl	7.88.1	h5eee18b_0
docutils	0.17.1	py36h06a4308_1
dropbox	5.2.1	py36_0
emboss	6.6.0	he06d7ca_1
entrez-direct	16.2	he881be0_1
expat	2.4.9	h6a678d5_0
fastp	0.23.2	hb7a2d85_2
filechunkio	1.6	py36_0
fontconfig	2.14.1	h52c9d5c_1
freetype	2.12.1	h4a9f257_0
ftputil	3.2	py36_0
giflib	5.1.4	h14c3975_1
hmmer	3.3.2	h87f3376_2
icu	58.2	he6710b0_3
idna	3.3	pyhd3eb1b0_0
intel-openmp	2022.1.0	h9e868ea_3769
isa-l	2.30.0	h7f8727e_0
jpeg	9e	h5eee18b_1
krb5	1.19.4	h568e23c_0
ld_impl_linux-64	2.38	h1181459_1
libcurl	7.88.1	h91b91d3_0
libdeflate	1.17	h5eee18b_0
libedit	3.1.20221030	h5eee18b_0
libev	4.33	h7f8727e_1
libffi	3.3	he6710b0_2

Appendix

libgcc-ng	11.2.0	h1234567_1
libgd	2.2.5	h8e06009_4
libgomp	11.2.0	h1234567_1
libiconv	1.16	h7f8727e_2
libidn2	2.3.4	h5eee18b_0
libnghttp2	1.46.0	hce63b2e_0
libpng	1.6.39	h5eee18b_0
libsodium	1.0.18	h7b6447c_0
libssh2	1.10.0	h8f2d780_0
libstdcxx-ng	11.2.0	h1234567_1
libtiff	4.1.0	hecacb30_2
libunistring	0.9.10	h27cf23_0
libuuid	1.41.5	h5eee18b_0
libwebp	1.0.1	h8e7db2f_0
libxml2	2.9.14	h74e7548_0
lz4-c	1.9.4	h6a678d5_0
mkl	2020.2	256
mkl-service	2.3.0	py36he8ac12f_0
mkl_fft	1.3.0	py36h54f3939_0
mkl_random	1.1.1	py36h0573a6f_0
ncbi-ncbi-ncbi-sdk	2.11.2	hff44eed_0
ncurses	6.4	h6a678d5_0
numpy	1.19.2	py36h54aff64_0
numpy-base	1.19.2	py36hfa32c7d_0
openssl	1.1.1w	h7f8727e_0
pandas	1.1.5	py36ha9443f7_0
paramiko	2.8.1	pyhd3eb1b0_0
pcre	8.45	h295c915_0
perl	5.26.2	h14c3975_0
perl-app-cpanminus	1.7044	pl526_1
perl-archive-tar	2.32	pl526_0
perl-carp	1.38	pl526_3
perl-common-sense	3.74	pl526_2
perl-compress-raw-bzip2	2.087	pl526he1b5a44_0
perl-compress-raw-zlib	2.087	pl526hc9558a2_0
perl-constant	1.33	pl526_1
perl-exporter	5.72	pl526_1
perl-extutils-makemaker	7.36	pl526_1
perl-file-path	2.16	pl526_0
perl-file-temp	0.2304	pl526_2
perl-io-compress	2.087	pl526he1b5a44_0
perl-io-zlib	1.10	pl526_2
perl-json	4.02	pl526_0
perl-json-xs	2.34	pl526_1
perl-list-moreutils	0.15	pl526_1

Appendix

perl-parent	0.236	pl526_1
perl-pathools	3.75	pl526h14c3975_1
perl-scalar-list-utils	1.52	pl526h516909a_0
perl-uri	1.71	pl526_3
perl-xml-libxml	2.0132	pl526h7ec2d77_1
perl-xml-namespacesupport	1.12	pl526_0
perl-xml-sax	1.02	pl526_0
perl-xml-sax-base	1.09	pl526_0
pip	21.2.2	py36h06a4308_0
psutil	5.8.0	py36h27cf23_1
pycparser	2.21	pyhd3eb1b0_0
pynacl	1.4.0	py36h7b6447c_1
pyopenssl	22.0.0	pyhd3eb1b0_0
pysftp	0.2.9	py36_0
pysocks	1.7.1	py36h06a4308_0
python	3.6.13	h12debd9_1
python-dateutil	2.8.2	pyhd3eb1b0_0
pytz	2021.3	pyhd3eb1b0_0
pyyaml	5.4.1	py36h27cf23_1
readline	8.2	h5eee18b_0
requests	2.27.1	pyhd3eb1b0_0
seqtk	1.3	h5bf99c6_3
setuptools	58.0.4	py36h06a4308_0
six	1.16.0	pyhd3eb1b0_1
snakemake	3.13.3	py36_0
spades	3.13.1	0
sqlite	3.41.2	h5eee18b_0
sra-tools	2.10.0	pl526he1b5a44_0
tk	8.6.12	h1ccaba5_0
urllib3	1.26.8	pyhd3eb1b0_0
vsearch	2.18.0	h95f258a_0
wget	1.21.3	h0b77cf5_0
wheel	0.37.1	pyhd3eb1b0_0
wrapt	1.12.1	py36h7b6447c_1
xz	5.4.2	h5eee18b_0
yaml	0.2.5	h7b6447c_0
zlib	1.2.13	h5eee18b_0
zstd	1.5.5	hc292b87_0

Appendix

Table 9: Profiles used for RNA virus screening in VirusHunter

Index	RNA virus profile
1	Amalga_RdRp
2	Astro-Poty_RdRp
3	Birnaviridae_RdRp
4	Bromo_RdRp
5	Clostero_RdRp
6	Endorna_RdRp
7	Hepe-Virga_RdRp
8	Hypo_RdRp
9	Luteo-Sobemo_RdRp
10	Narna-Levi_RdRp
11	Negative_Bunya-Arena_RdRp
12	Negative_Mono-Chu_RdRp
13	Negative_Orthomyxo_RdRp
14	New-Qinvirus_RdRp
15	New-Weivirus_RdRp
16	New-Yanvirus_RdRp
17	New-Zhaovirus_RdRp
18	Partiti-Picobirna_RdRp
19	Permutotetra_RdRp
20	Picornavirus-Calici_RdRp
21	Poty_RdRp
22	Reo_RdRp
23	Tombus-Noda_RdRp
24	Toti-Chryso_RdRp
25	Tymo_RdRp
26	Flavi_RdRp
27	NidoAstro_RdRp
28	Nido_NiRAN
29	Delta_HDAg

Table 10: Profiles used for large DNA virus screening in VirusHunter

Index	Large DNA virus profile
1	Adeno_hexon
2	Adeno_penton
3	Adeno_pIIIa
4	Adeno_pTP
5	Herpes_POLcore
6	Herpes_TER1
7	Herpes_UL19Cterm
8	Herpes_UL19Nterm
9	Herpes_UL29
10	Herpes_UL5
11	Herpes_UL6
12	NCLDV_packATPase
13	NCLDV_A2Ltf
14	NCLDV_ssuRiboReduc
15	Ascoviruses_mcp
16	Asfarviruses_mcp
17	Aureococcusviruses_mcp
18	Cafeteriaviruses_mcp
19	Chloroviruses_mcp
20	Chrysomulinaviruses_mcp
21	Coccolithoviruses_mcp
22	Faunusviruses_mcp
23	Faustoviruses_mcp
24	Iridoviruses_mcp
25	Klosneuviruses_mcp
26	Marseilleviruses_mcp
27	MeML13_mcp
28	MeML17_mcp
29	MeML23_mcp
30	MeML24_mcp
31	MeML25_mcp
32	MeML26_mcp
33	MeML28_mcp
34	MeML29_mcp
35	MeML30_mcp
36	MeML31_mcp
37	MeML32_mcp
38	MeML33_mcp
39	MeML34_mcp
40	MeML35_mcp
41	MeML36_mcp

Index	Large DNA virus profile
42	MeML38_mcp
43	MeML39_mcp
44	MeML3_mcp
45	MeML40_mcp
46	MeML41_mcp
47	MeML42_mcp
48	MeML43_mcp
49	MeML44_mcp
50	MeML45_mcp
51	MeML46_mcp
52	MeML47_mcp
53	MeML48_mcp
54	MeML49_mcp
55	MeML4_mcp
56	MeML50_mcp
57	MeML51_mcp
58	MeML52_mcp
59	MeML54_mcp
60	MeML55_mcp
61	MeML56_mcp
62	MeML57_mcp
63	MeML58_mcp
64	MeML59_mcp
65	MeML5_mcp
66	MeML60_mcp
67	MeML61_mcp
68	MeML62_mcp
69	MeML63_mcp
70	MeML64_mcp
71	MeML65_mcp
72	MeML66_mcp
73	MeML67_mcp
74	MeML68_mcp
75	Megamimiviruses_mcp
76	Namaoviruses_mcp
77	OLPG_mcp
78	Phaeoviruses_mcp
79	Pithocedratviruses_mcp
80	Poxviruses_mcp
81	Prasinoviruses_mcp
82	Prymnesioviruses_mcp
83	Raphidoviruses_mcp
84	Solumviruses_mcp
85	Sylvanviruses_mcp

Index Large DNA virus profile	
86	Tetraselmisviruses_mcp
87	YLMG_mcp
88	YLPG_mcp

Table 11: Profiles used for small DNA virus screening in VirusHunter

Index Small DNA virus Profile	
1	Anello_ORF1core
2	Circo_Rep
3	Gemini_Rep
4	Hepadna-Nackedna_TP
5	Nano_MRep
6	Nano_NS1
7	Papilloma_E1
8	Parvo_NS1
9	Polyoma_LTag
10	Genomo_Rep
11	Redondo_Rep
12	Smaco_Rep

Table 12: Taubert VirusGatherer – Summary statistics of the distribution of sequence identity percentages to the closest viral reference sequences.

Viral reference taxonomy	Median	Q1	Q3	Mean	SD	Min	Max
RNA Viruses							
Hepeviridae	94.7	94.7	94.7	94.7	NA	94.7	94.7
Totiviridae	84.2	84.2	84.2	84.2	NA	84.2	84.2
Kolmioviridae	81.9	78.1	90.9	84.2	6.9	77.4	94.1
Large DNA Viruses							
Orthoherpesviridae	100.0	100.0	100.0	100.0	0.0	100.0	100.0
Poxviridae	92.3	89.9	94.1	91.7	2.9	88.2	94.1
Small DNA Viruses							
Hepadnaviridae	95.5	95.5	95.5	95.5	NA	95.5	95.5

Appendix

Table 13: Taubert VirusGatherer – Summary statistics of the distribution of contig lengths.

Viral reference taxonomy	Median	Q1	Q3	Mean	SD	Min	Max
RNA Viruses							
Kolmioviridae	138	107	178	154	56	106	259
Totiviridae	60	60	60	60	NA	60	60
Hepeviridae	59	59	59	59	NA	59	59
Large DNA Viruses							
Orthoherpesviridae	55	54	56	55	1	54	56
Poxviridae	54	54	56	56	5	54	64
Small DNA Viruses							
Hepadnaviridae	70	70	70	70	NA	70	70

Table 14: Mammalian sequencing data – Summary Statistics of the distribution of viral reference E-values for each viral family (VirusGatherer). E-values from contig alignment against viral reference database are transformed to their negative logarithm base 10.

Viral reference taxonomy	Median	Q1	Q3	Mean	SD	Min	Max
Flaviviridae	174.25	43.44	179.39	119.75	67.51	-0.85	179.39
Tulasviridae	160.86	151.59	170.12	160.86	26.20	142.33	179.39
Discoviridae	127.66	44.77	164.29	108.73	62.46	23.15	179.39
Hantaviridae	95.52	37.24	179.39	101.27	74.15	-0.71	179.39
Paramyxoviridae	80.75	38.15	179.39	95.09	62.11	-0.76	179.39
Picornaviridae	66.52	24.96	120.06	78.67	64.13	0.92	179.39
Bornaviridae	57.55	46.94	59.69	50.45	29.42	-0.28	95.12
Astroviridae	50.35	34.05	118.71	76.34	60.46	-1.00	179.39
Sedoreoviridae	49.97	36.50	179.39	88.04	63.11	23.52	179.39
Tobaniviridae	49.13	18.78	114.26	68.91	73.60	-0.68	179.39
Virgaviridae	48.40	27.23	105.02	75.62	62.32	2.70	179.39
unclassified Pisuviricota	38.22	18.17	105.64	65.78	64.50	-0.18	179.39
Lispiviridae	36.84	-0.18	71.45	40.73	47.61	-0.34	106.06
Mymonaviridae	36.06	36.06	36.06	36.06	0.00	36.06	36.06
Mesoniviridae	35.23	33.22	37.23	35.23	5.68	31.21	39.24
Coronaviridae	35.17	9.47	80.77	57.10	61.22	-0.87	179.39
Hepeviridae	33.90	4.94	89.55	56.23	64.49	0.92	179.39
Marnaviridae	33.48	20.82	46.69	40.66	33.92	2.52	179.39

Appendix

Viral reference taxonomy	Median	Q1	Q3	Mean	SD	Min	Max
Picobirnaviridae	33.31	21.27	62.99	54.55	50.35	4.26	179.39
Deltaflexiviridae	30.32	30.32	30.32	30.32	NA	30.32	30.32
Orthomyxoviridae	30.20	22.26	82.78	55.28	53.31	-0.46	170.16
Caliciviridae	29.63	0.78	117.45	62.31	72.18	-0.77	179.39
Betaflexiviridae	29.00	6.90	34.35	32.80	43.68	0.11	140.06
Nodaviridae	28.26	21.22	39.74	30.50	18.98	3.00	64.09
Phasmaviridae	28.18	18.35	49.75	45.51	55.24	0.00	179.39
unclassified Lenarvircota	26.17	16.30	37.10	33.38	30.68	11.15	130.47
Polycipiviridae	25.81	17.58	25.81	20.32	9.50	9.35	25.81
Bromoviridae	24.20	1.49	48.86	43.13	60.24	-0.63	179.39
Duinviridae	23.74	15.64	37.40	28.25	19.46	3.88	102.19
Atkinsviridae	22.75	13.88	32.92	33.20	35.22	2.15	179.39
Totiviridae	21.06	12.92	27.56	22.93	20.57	-0.70	179.39
Alphaflexiviridae	20.95	20.95	106.15	50.65	51.45	-0.94	106.15
Rhabdoviridae	20.91	0.89	64.71	45.59	59.26	-0.76	179.39
Fiersviridae	20.47	13.22	37.21	33.42	37.08	-0.53	179.39
Dicistroviridae	19.88	8.27	40.28	29.23	32.29	-0.60	179.39
Permutotetraviridae	19.59	16.90	22.28	19.59	7.61	14.21	24.97
Blumeviridae	19.09	12.98	32.23	24.95	18.76	3.11	83.79
Narnaviridae	18.94	6.95	21.24	46.48	74.62	5.88	179.39
Tombusviridae	18.80	12.22	44.27	37.15	41.49	0.49	179.39
Botourmiaviridae	18.76	8.13	26.34	22.01	25.53	-0.92	179.39
unclassified	18.47	7.54	40.15	34.86	44.35	-0.95	179.39
Chuviridae	18.45	13.24	28.11	32.61	41.27	-0.80	104.03
Steitzviridae	18.09	10.47	26.75	22.26	20.08	-0.85	138.21
Endornaviridae	17.22	4.27	36.27	28.85	42.11	-0.43	154.22
Iflaviridae	16.78	-0.15	58.91	39.38	53.97	-0.64	179.39
Fusariviridae	16.26	7.85	24.68	16.26	23.80	-0.57	33.10
Benyviridae	15.70	10.48	32.77	20.18	16.51	0.19	42.28
Phenuiviridae	15.34	2.70	33.95	30.74	44.79	-0.79	179.39
Chrysoviridae	14.94	14.94	14.94	14.94	0.00	14.94	14.94
unclassified Kitrinoviricota	14.10	9.24	16.41	14.26	8.31	4.70	33.03
Spinareoviridae	10.99	0.12	23.42	14.55	16.09	-0.15	42.56
Solemoviridae	9.62	5.74	16.36	14.17	16.94	-0.48	72.39
Barnaviridae	9.37	9.24	9.96	9.31	1.40	6.77	10.97
Tospoviridae	9.20	5.04	9.20	6.43	4.80	0.89	9.20
Arteriviridae	9.08	7.08	10.92	14.05	28.43	-0.26	179.39
Partitiviridae	9.03	4.59	14.32	13.72	21.45	-0.93	179.39
Hypoviridae	8.27	7.29	25.42	24.45	33.03	7.29	73.97
Mitoviridae	7.96	-0.53	17.43	10.74	11.25	-0.53	45.41
Xinmoviridae	7.31	7.31	7.31	7.31	NA	7.31	7.31
Tymoviridae	7.31	0.50	45.99	37.12	58.61	-0.94	179.39
Secoviridae	6.79	0.31	9.00	27.77	57.09	-0.70	179.39
Birnaviridae	5.73	5.73	5.73	5.73	NA	5.73	5.73

Appendix

Viral reference taxonomy	Median	Q1	Q3	Mean	SD	Min	Max
Amalgaviridae	4.73	3.76	7.75	6.00	3.35	1.46	13.22
Potyviridae	3.28	0.51	11.41	18.96	42.09	-0.93	179.39
Gammaflexiviridae	2.30	0.90	2.30	1.37	1.61	-0.49	2.30
Closteroviridae	1.74	0.25	3.20	1.70	1.73	0.12	3.20
Sinhaliviridae	1.44	0.83	7.76	7.14	12.10	0.43	25.26
Fimoviridae	0.62	0.38	5.27	3.56	5.51	0.14	9.92
Cruliviridae	0.51	0.51	0.51	0.51	NA	0.51	0.51
Non-RNA-viruses	0.43	-0.30	9.13	12.15	27.16	-1.00	179.39
Roniviridae	0.33	0.33	0.33	0.33	NA	0.33	0.33
Olfoviridae	0.28	0.28	0.28	0.28	NA	0.28	0.28
unclassified Duplornaviricota	0.24	0.24	0.24	0.24	NA	0.24	0.24
Pneumoviridae	0.17	-0.05	89.78	59.76	103.60	-0.28	179.39
Peribunyaviridae	0.12	-0.35	0.73	0.21	1.09	-0.83	1.35
Solspiviridae	-0.04	-0.63	13.87	7.04	11.11	-0.90	44.79
Kolmioviridae	-0.11	-0.35	0.13	-0.11	0.68	-0.59	0.37
Baculoviridae	-0.11	-0.34	0.49	1.63	5.28	-0.97	19.21
Nimaviridae	-0.23	-0.23	-0.23	-0.23	NA	-0.23	-0.23
Togaviridae	-0.36	-0.36	-0.36	-0.36	NA	-0.36	-0.36
Arenaviridae	-0.80	-0.82	-0.73	-0.75	0.16	-0.87	-0.52

Table 15: Mammalian sequencing data - Summary Statistics of the distribution of protein sequence identities to the nearest known reference virus for each viral family.

Viral reference taxonomy	Median	Q1	Q3	Mean	SD	Min	Max
Pneumoviridae	100.0	100.0	100.0	100.0	NA	100.0	100.0
Alphaflexiviridae	98.5	97.7	99.4	98.5	0.9	97.7	99.4
Coronaviridae	98.5	95.4	100.0	96.3	5.7	71.6	100.0
Virgaviridae	97.8	82.6	100.0	87.0	19.5	38.2	100.0
Orthomyxoviridae	96.7	94.6	98.2	95.6	5.3	82.6	100.0
Mymonaviridae	95.0	95.0	95.0	95.0	0.0	95.0	95.0
Picornaviridae	93.2	88.6	98.2	87.9	16.6	30.0	100.0
Sedoreoviridae	93.0	91.7	95.9	91.0	13.6	24.9	98.8
Arteriviridae	92.6	80.8	100.0	90.0	10.9	48.5	100.0
Bromoviridae	90.9	85.7	95.3	91.1	5.2	85.2	100.0
Chrysoviridae	90.9	90.9	90.9	90.9	0.0	90.9	90.9
Betaflexiviridae	90.3	55.8	94.2	75.8	24.3	40.4	99.5
Fusariviridae	86.8	86.8	86.8	86.8	NA	86.8	86.8
Paramyxoviridae	85.2	75.4	95.6	84.7	12.1	61.9	100.0
Astroviridae	82.6	73.9	94.5	81.2	16.5	29.0	100.0
Discoviridae	81.7	78.9	82.8	79.2	8.2	61.9	93.8
Hepeviridae	81.3	41.5	88.2	67.6	23.9	28.5	92.8

Appendix

Viral reference taxonomy	Median	Q1	Q3	Mean	SD	Min	Max
Deltaflexiviridae	79.4	79.4	79.4	79.4	NA	79.4	79.4
Hantaviridae	79.0	76.8	83.5	81.5	8.3	70.3	99.2
Flaviviridae	78.5	70.5	94.0	78.6	16.4	38.4	100.0
Rhabdoviridae	77.2	68.4	99.9	81.1	17.4	46.0	100.0
Tulasviridae	75.6	71.1	80.0	75.6	12.6	66.7	84.4
Polycipiviridae	71.9	52.3	71.9	58.9	22.6	32.7	71.9
Nodaviridae	70.9	54.3	76.1	67.4	23.3	26.4	100.0
Totiviridae	67.6	55.4	71.7	65.2	12.3	31.4	100.0
Picobirnaviridae	67.5	55.7	74.1	64.5	13.1	28.2	90.3
Mesoniviridae	66.2	60.0	72.4	66.2	17.5	53.8	78.6
Tymoviridae	65.3	50.0	77.0	63.7	20.0	31.7	96.4
Marnaviridae	63.7	50.3	68.9	60.6	12.3	37.0	86.2
Iflaviridae	63.5	41.7	96.2	66.9	24.3	31.6	98.9
Phasmaviridae	63.3	53.1	79.7	66.9	15.4	51.2	88.0
Solspiviridae	58.7	55.5	66.4	59.4	10.1	29.9	73.2
Sinhaliviridae	58.5	58.5	58.5	58.5	NA	58.5	58.5
Atkinsviridae	58.2	51.2	63.9	57.3	8.8	41.5	81.2
Botourmiaviridae	57.6	43.5	66.4	56.2	16.6	32.2	98.9
Tobaniviridae	56.6	50.0	67.1	64.6	21.1	50.0	98.4
Lispiviridae	56.5	36.8	76.2	56.5	22.8	36.8	76.2
Steitzviridae	55.5	49.5	64.7	57.5	13.8	29.6	98.1
Fiersviridae	54.8	45.7	65.3	56.1	14.5	27.5	99.0
Phenuiviridae	54.4	42.3	65.9	55.3	19.9	22.7	100.0
Dicistroviridae	54.4	47.0	66.1	57.9	18.4	25.1	100.0
Permutotetraviridae	54.0	53.9	54.2	54.0	0.5	53.7	54.3
unclassified Pisuviricota	51.6	40.5	72.7	58.8	21.1	28.5	98.6
Xinmoviridae	51.2	51.2	51.2	51.2	NA	51.2	51.2
Bornaviridae	50.8	49.8	50.8	49.4	2.9	43.0	50.8
unclassified Lenarviricota	49.0	43.0	55.8	49.6	9.2	36.9	66.7
unclassified Kitrinoviricota	48.5	44.0	54.2	53.0	19.1	33.3	97.0
unclassified	47.8	38.5	60.4	50.4	15.6	21.0	100.0
Blumeviridae	47.6	41.2	55.6	48.4	10.3	30.7	68.9
Duinviridae	46.1	37.7	52.5	44.6	9.4	29.0	64.9
Caliciviridae	46.1	41.9	67.6	55.9	20.8	34.2	98.3
Tombusviridae	46.0	41.3	55.6	49.4	13.7	25.2	98.6
Hypoviridae	45.8	42.2	46.3	42.7	6.4	33.1	46.3
Chuviridae	45.7	40.9	52.5	47.6	13.7	33.0	65.8
Tospoviridae	45.6	45.6	45.6	45.6	0.0	45.6	45.6
Mitoviridae	44.8	32.3	56.9	46.6	14.3	29.3	91.5
Solemoviridae	44.8	39.1	50.0	47.9	15.5	26.7	100.0
Secoviridae	44.8	38.9	49.4	47.7	19.5	28.2	91.5
Narnaviridae	44.3	30.5	45.3	49.3	29.2	27.0	99.5
Spinareoviridae	43.5	41.9	79.2	59.7	28.0	38.6	100.0
Endornaviridae	43.2	40.7	56.9	51.7	20.2	35.2	100.0

Appendix

Viral reference taxonomy	Median	Q1	Q3	Mean	SD	Min	Max
Partitiviridae	41.4	34.1	56.8	46.7	17.7	24.6	100.0
Barnaviridae	40.8	40.7	44.9	42.9	4.7	38.2	50.9
Potyviridae	40.6	31.6	97.2	60.0	33.1	25.5	100.0
Benyviridae	38.4	35.1	60.3	45.9	15.9	30.2	65.5
Non-RNA-viruses	36.9	34.0	38.8	37.5	6.9	22.5	73.8
Amalgaviridae	34.5	30.2	35.2	33.8	4.8	28.3	41.3
Birnaviridae	33.0	33.0	33.0	33.0	NA	33.0	33.0
Baculoviridae	30.8	27.3	30.8	28.4	4.1	23.7	30.8
Fimoviridae	26.3	26.3	26.3	26.3	NA	26.3	26.3

Table 16: Mammalian sequencing data – Number and percentage of contigs above and below 90% sequence identity across all phyla.

Phylum	Number of contigs	Contigs with sequence identity < 90%	Contigs with sequence identity >= 90%	Percentage < 90%	Percentage >= 90%
Artvervirocota	11	11	0	100.00	0.00
Duplornavirocota	115	84	31	73.04	26.96
Kitrinovirocota	453	343	110	75.72	24.28
Lenarvirocota	686	671	15	97.81	2.19
Negarnavirocota	209	147	62	70.33	29.67
Nucleocytovirocota	313	313	0	100.00	0.00
Peplovirocota	5	5	0	100.00	0.00
Pisuvirocota	2212	1262	950	57.05	42.95
Urovirocota	92	92	0	100.00	0.00
unclassified	2113	2076	37	98.25	1.75

Table 17: Mammalian sequencing data - Summary Statistics of the distribution of contig lengths for each viral family.

Viral reference taxonomy	Median	Q1	Q3	Mean	SD	Min	Max
Pneumoviridae	10373	10373	10373	10373	NA	10373	10373
Bornaviridae	2311	1418	3160	2319	1435	262	4504
Fimoviridae	1659	1659	1659	1659	NA	1659	1659
Tulasviridae	1560	1495	1626	1560	185	1430	1691
Flaviviridae	1524	313	7511	3724	4343	90	12688
Baculoviridae	1357	1350	2597	2179	1436	1342	3837
Lispiviridae	926	242	1611	926	790	242	1611
Hantaviridae	889	254	1334	1722	2275	155	6588
Non-RNA-viruses	812	438	1930	1585	1866	195	12139

Appendix

Viral reference taxonomy	Median	Q1	Q3	Mean	SD	Min	Max
Discoviridae	670	304	1128	781	570	151	2118
Paramyxoviridae	617	268	5039	3729	5943	142	17699
Caliciviridae	575	350	5275	2645	2964	165	8028
Tobaniviridae	542	455	2267	1102	976	194	2267
Endornaviridae	510	235	515	405	196	178	670
Tymoviridae	496	268	1335	892	833	127	2468
Hepeviridae	460	266	898	1185	1693	182	6148
Narnaviridae	455	396	1881	1017	928	189	2163
unclassified Pisuviricota	435	255	2208	1875	2848	175	11367
Amalgaviridae	421	336	535	461	193	254	790
Picornaviridae	416	264	890	1368	2243	133	7831
Phenuiviridae	412	230	586	862	1690	144	7409
Secoviridae	370	196	1328	1067	1612	170	5903
Iflaviridae	363	279	699	1168	2377	172	10027
unclassified	353	223	845	872	1264	96	12329
Duinviridae	349	268	907	915	1157	174	4043
Astroviridae	348	238	811	904	1435	127	6735
Rhabdoviridae	345	256	699	1632	3277	109	12782
Sedoreoviridae	336	204	1933	1013	1165	152	3569
Tombusviridae	313	206	586	743	996	109	4172
Mesoniviridae	306	261	352	306	129	215	398
Alphaflexiviridae	305	135	475	305	196	135	475
Spinareoviridae	305	212	488	385	243	170	801
unclassified Lenarvircota	302	205	768	712	914	183	3297
Marnaviridae	292	232	452	561	1234	116	8894
Partitiviridae	287	201	494	498	505	125	2142
Coronaviridae	286	170	472	2452	5506	114	26479
Virgaviridae	284	214	542	1220	1973	132	7031
Nodaviridae	282	178	405	538	819	133	3077
Picobirnaviridae	282	205	586	503	465	131	2134
Phasmaviridae	278	235	482	764	1197	181	3455
Mitoviridae	275	186	660	643	731	149	2354
Blumeviridae	270	214	547	895	1352	148	4595
Benyviridae	264	187	844	1019	1486	172	3630
Fiersviridae	262	191	532	716	1070	129	4373
Dicistroviridae	259	210	420	505	676	148	3487
Birnaviridae	253	253	253	253	NA	253	253
Sinhaliviridae	251	251	251	251	NA	251	251
Botourmiaviridae	248	210	327	455	588	149	2962
Hypoviridae	246	197	918	869	1279	197	2786
Atkinsviridae	246	192	319	352	306	123	1394
Permutotetrviridae	242	227	256	242	42	212	271
Chuviridae	234	224	712	702	951	213	2129
Potyviridae	233	210	883	1044	2013	181	8519

Appendix

Viral reference taxonomy	Median	Q1	Q3	Mean	SD	Min	Max
Fusariviridae	230	230	230	230	NA	230	230
Solemoviridae	227	186	361	393	401	134	1656
Steitzviridae	219	184	269	454	768	129	4144
Bromoviridae	218	185	469	655	953	163	3071
Barnaviridae	214	211	229	212	24	167	234
unclassified Kitrinoviricota	212	192	303	264	141	103	483
Totiviridae	206	172	224	222	159	106	1525
Deltaflexiviridae	194	194	194	194	NA	194	194
Orthomyxoviridae	190	168	545	352	251	164	782
Betaflexiviridae	189	186	468	403	363	184	1142
Solspiviridae	182	152	251	387	786	130	3521
Mymonaviridae	181	181	181	181	0	181	181
Polycipiviridae	179	179	257	231	90	179	335
Tospoviridae	174	174	174	174	0	174	174
Xinmoviridae	174	174	174	174	NA	174	174
Chrysoviridae	112	112	112	112	0	112	112
Arteriviridae	104	97	114	371	1744	78	18735

Table 18: Mammalian sequencing data – Number and percentage of contigs with sequence identity below 90% and length above 1,000 nucleotides, grouped by phyla and family

Phylum	Viral reference family	Contigs length >= 1000 nt & sequence identity < 90%	% of length >= 1000 nt & sequence identity < 90%
Artvervirocota	Retroviridae	7	100.00
Duplornaviricota	Sedoreoviridae	1	50.00
Duplornaviricota	Totiviridae	1	50.00
Kitrinoviricota	Flaviviridae	58	56.31
Kitrinoviricota	Tombusviridae	28	27.18
Kitrinoviricota	Hepeviridae	5	4.85
Kitrinoviricota	Virgaviridae	5	4.85
Kitrinoviricota	Tymoviridae	3	2.91
Kitrinoviricota	Nodaviridae	2	1.94
Kitrinoviricota	Benyviridae	1	0.97
Kitrinoviricota	Betaflexiviridae	1	0.97
Lenarviricota	Fiersviridae	41	43.16
Lenarviricota	Blumeviridae	14	14.74
Lenarviricota	Duinviridae	11	11.58
Lenarviricota	Steitzviridae	8	8.42
Lenarviricota	Mitoviridae	7	7.37
Lenarviricota	Botourmiaviridae	6	6.32
Lenarviricota	Atkinsviridae	4	4.21

Appendix

Phylum	Viral reference family	Contigs length >= 1000 nt & sequence identity < 90%	% of length >= 1000 nt & sequence identity < 90%
Lenarvirocota	unclassified Lenarvirocota	2	2.11
Lenarvirocota	Narnaviridae	1	1.05
Lenarvirocota	Solspiviridae	1	1.05
Negarnaviricota	Paramyxoviridae	12	28.57
Negarnaviricota	Bornaviridae	6	14.29
Negarnaviricota	Hantaviridae	5	11.90
Negarnaviricota	Discoviridae	4	9.52
Negarnaviricota	Phenuiviridae	4	9.52
Negarnaviricota	Rhabdoviridae	4	9.52
Negarnaviricota	Lispiviridae	2	4.76
Negarnaviricota	Tulasviridae	2	4.76
Negarnaviricota	Chuviridae	1	2.38
Negarnaviricota	Fimoviridae	1	2.38
Negarnaviricota	Phasmaviridae	1	2.38
Nucleocytoviricota	Iridoviridae	73	74.49
Nucleocytoviricota	Mimiviridae	16	16.33
Nucleocytoviricota	Phycodnaviridae	7	7.14
Nucleocytoviricota	Poxviridae	2	2.04
Peploviricota	Alloherpesviridae	1	100.00
Pisuviricota	Picobirnaviridae	46	22.77
Pisuviricota	Arteriviridae	36	17.82
Pisuviricota	Astroviridae	31	15.35
Pisuviricota	unclassified Pisuviricota	29	14.36
Pisuviricota	Caliciviridae	12	5.94
Pisuviricota	Partitiviridae	11	5.45
Pisuviricota	Picornaviridae	11	5.45
Pisuviricota	Dicistroviridae	5	2.48
Pisuviricota	Marnaviridae	4	1.98
Pisuviricota	Iflaviridae	3	1.49
Pisuviricota	Potyviridae	3	1.49
Pisuviricota	Secoviridae	3	1.49
Pisuviricota	Solemoviridae	3	1.49
Pisuviricota	Tobaniviridae	3	1.49
Pisuviricota	Coronaviridae	1	0.50
Pisuviricota	Hypoviridae	1	0.50
Uroviricota	unclassified Uroviricota	78	98.73
Uroviricota	Steigviridae	1	1.27
unclassified	unclassified	467	99.36
unclassified	Baculoviridae	3	0.64

Appendix

Table 19: Mammalian sequencing data – Top ten longest contigs: host, SRA experiment, length, closest viral match with phylum and sequence identity

SRA experiment	Phylum	Viral reference taxonomy	viral subject	Host	contig length	Sequence Identity [%]
SRR14579882	Pisuviricota	Coronaviridae	Alphacoronavirus sp.	Otomops harrisoni	26479	71.648
SRR18700679	Pisuviricota	Arteriviridae	Porcine reproductive and respiratory syndrome virus 2	Sus scrofa	18735	85.817
SRR12430148	Negarnaviricota	Paramyxoviridae	Mount Mabu Lophuromys virus 1	Heteromys desmarestianus	17699	61.900
SRR19431078	Negarnaviricota	Rhabdoviridae	Obodhang virus	Bos taurus	12782	48.205
SRR15647801	Kitrinoviricota	Flaviviridae unclassified Pisuviricota	Border disease virus	Ovis aries	12688	85.186
SRR17345014	Pisuviricota		Posavirus 2	Sus scrofa domesticus	11367	34.233
SRR16151768	Negarnaviricota	Pneumoviridae	human respiratory syncytial virus	Macaca mulatta	10373	100.000
SRR18779479	Kitrinoviricota	Virgaviridae	Nephila clavipes virus 4	Ovis aries	7031	41.320
SRR17345012	Kitrinoviricota	Hepeviridae	Paslahepevirus balayani	Sus scrofa domesticus	6148	87.775
SRR13364364	Duplornaviricota	Sedoreoviridae	Stretch Lagoon orbivirus	Rhinolophus	3569	24.855

Table 20: Mammalian sequencing data – Number and percentage (relative to total contig count) of contigs found in each host

Host	Number of contigs	Percentage
Bos taurus	2401	38.67
Sus scrofa domesticus	1073	17.28
Chlorocebus sabaeus	1059	17.06
Sus scrofa	317	5.11
Otomops harrisoni	186	3.00
Rhinolophus	126	2.03
Manis javanica	111	1.79
Rattus norvegicus	70	1.13
Myotis riparius	69	1.11
Nannospalax ehrenbergi	67	1.08
Macaca mulatta	58	0.93
Equus asinus	50	0.81
Neotoma cinerea	44	0.71
Capra hircus	42	0.68
Rattus	38	0.61
Rhizomys pruinosus	38	0.61
Sciurus carolinensis	35	0.56
Blarina brevicauda	32	0.52
Ovis aries	32	0.52
Miniopterus africanus	31	0.50
Capreolus capreolus	29	0.47
Chlorocebus aethiops	27	0.43
Macaca fascicularis	25	0.40
Monodelphis domestica	23	0.37
Rhinolophus ferrumequinum	23	0.37

Appendix

Host	Number of contigs	Percentage
<i>Peromyscus maniculatus</i>	22	0.35
<i>Chiroptera</i>	21	0.34
<i>Hipposideros caffer</i>	19	0.31
<i>Rhinolophus landeri</i>	13	0.21
<i>Miniopterus natalensis</i>	11	0.18
<i>Notomys alexis</i>	11	0.18
<i>Heteromys desmarestianus</i>	10	0.16
<i>Rhinolophus affinis</i>	9	0.14
<i>Bos indicus</i>	8	0.13
<i>Canis lupus familiaris</i>	8	0.13
<i>Mesocricetus auratus</i>	8	0.13
<i>Ondatra zibethicus</i>	8	0.13
<i>Tamias striatus</i>	8	0.13
<i>Myotis tricolor</i>	6	0.10
<i>Bubalus bubalis</i>	5	0.08
<i>Chinchilla lanigera</i>	5	0.08
<i>Marmota monax</i>	5	0.08
<i>Chaetodipus intermedius</i>	4	0.06
<i>Myodes glareolus</i>	4	0.06
<i>Callithrix jacchus</i>	3	0.05
<i>Cavia porcellus</i>	3	0.05
<i>Tursiops truncatus</i>	3	0.05
<i>Equus caballus</i>	2	0.03
<i>Mustela putorius furo</i>	2	0.03
<i>Oryctolagus cuniculus</i>	2	0.03
<i>Antechinus flavipes</i>	1	0.02
<i>Cervus nippon</i>	1	0.02
<i>Papio anubis</i>	1	0.02

**Eidesstattliche Erklärung gemäß § 8 Abs. 7 der Prüfungsordnung für den
Masterstudiengang Biomedizinische Datenwissenschaft**

Hiermit erkläre ich, dass ich die vorliegende Masterarbeit (Virome characterization from unprocessed local and publicly available sequencing data) selbstständig und ohne die Hilfe Dritter verfasst und ausschließlich mit den angegebenen Hilfeleistungen und Hilfsmitteln angefertigt habe. Direkt oder indirekt übernommene Inhalte oder Daten aus anderen Quellen sind als solche kenntlich gemacht.

Die Richtlinien der „Grundsätze der Medizinischen Hochschule Hannover zur Sicherung guter wissenschaftlicher Praxis und Verfahrensregeln für den Umgang mit wissenschaftlichem Fehlverhalten“ sowie die „Richtlinie der Medizinischen Hochschule Hannover (MHH) zum Umgang mit textgenerierender künstlicher Intelligenz (KI) bei der Erstellung wissenschaftlicher Dokumente“ habe ich eingehalten.

Die Arbeit wurde bisher nicht für eine Prüfung oder für einen ähnlichen Zweck zur Beurteilung eingereicht oder veröffentlicht.

Ich erkläre mich damit einverstanden, dass die Arbeit mit Hilfe eines Plagiatserkennungsdienstes auf enthaltene Plagiate überprüft werden kann und auch die Primärdaten stichprobenartig kontrolliert werden können.

Ich versichere, dass ich die vorstehenden Angaben nach bestem Wissen vollständig und der Wahrheit entsprechend gemacht habe.

Hannover, 7.10.2024

KuFF

Ort, Datum

Unterschrift

Erklärung zur Nutzung von KI bei der Erstellung der Masterarbeit

Wurden spezielle KI-basierte oder KI-unterstützte Konzepte erstellt, wissenschaftliche Methodiken oder Arbeitsweisen für das Studiendesign verwendet, sind diese im Methoden-Abschnitt beschrieben.

Wurden textgenerierende KI bei der Erstellung des Manuskriptes bzw. des Diskussionsteils eingesetzt, werden im Folgenden die Gründe für die Verwendung, der Umfang und die betroffenen Textstellen genannt.

Nein. Es wurde keine KI bei der Erstellung der Masterarbeit genutzt.

Ich bestätige, diese generierten Texte geprüft und überarbeitet zu haben und verantworte deren Inhalte.

Hannover, 7.10.2024

Muff

Ort, Datum

Unterschrift