Virom-Charakterisierung anhand unprozessierter lokaler und öffentlich verfügbarer Sequenzierungsdaten

Virome characterisation 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, <Monat (ausgeschrieben) Jahr (Zahl)>

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:

Table of Contents

List of Abbreviations

Abbreviations	Meaning
pHMM	Profile Hidden Markov Model
SRA	Sequence Read Archive

Table of Figures

Index of Tables

1 Introduction

1.1 The need for virus discovery

Viruses are omnipresent. In water, even a single milliliter can contain up to hundreds of millions of aquatic viruses [Source: High abundance of viruses found in aquatic environments]; and on land, one could dig up soil and find up to a billion viruses in a single gram [Source: Viruses in soils: morphological diversity and abundance in the rhizosphere]. But one would not need to dig up earth or dive deep into the ocean to find viruses, as the human body itself is a viral ecosystem for over 10^{13} viral particles. With an estimated 10^31 viral particles, viruses outnumber every organism in the domains of life – animals, plants, fungi, bacteria, and archaea [Quelle 1: Evolutionary relashionship among diverse bacteriophages and prophages. Quelle 2: Are there 1031 virus particles on earth, or more, or fewer?].

All life forms can be infected by viruses [Buch Virologie], and they all share a common strategy for genome replication and gene expression: double-stranded DNA stores genetic information, which is transcribed into single-stranded RNA and, in some cases, translated into protein. Viruses, on the other hand, use both DNA and RNA in various forms for their genomes, along with all possible intermediates. Viral genomes may be composed of RNA or DNA, existing as single or double-stranded, and can be linear or circular [BUCH Virologie]. Additionally, viral genomes can consist of a single continuous piece of nucleic acid or be divided into multiple segments. They may be positive-sense, meaning the genome is read in the 5' to 3' direction, or negative-sense, where it is read from 3' to 5' [BUCH Virologie].

Our understanding of viruses has evolved over time, with the traditional view seeing them as merely disease-causing and offering no benefits to their hosts or the environment. Today, this view has changed radically. Alongside their ability to cause disease and pandemics, viruses can also benefit the environment, as they, for example, lyse 40% of marine prokaryotes each day, which stimulates the growth of plankton [The ecological virus]. In humans, 8% of our genome consists of human endogenous retroviruses (HERVs), which resulted from the integration of these viruses into our genome. Although no longer infectious, some HERVs still produce transcripts, which have been associated with various cancers, including ovarian cancer [Expression of multiple human endogenous retroviruses surface envelope

proteins in ovarian cancer], prostate cancer [selective changes of retroelements expression in human prostate cancer], breast cancer and lymphoma [Human Endogenous Retrovirus K elements in plasma of people with Lymphoma and brest cancer], and melanoma [The activation of human endogenous retrovirus K is implicated in melanoma of malignant transformation]. On the other hand, HERVs also benefit us by being involved in human placental morphogenesis [Syncytin is captive retroviral envelope protein involved in human placental morphogenesis].

The virosphere remains largely unexplored, as do its complex interactions with hosts and the environment. It was estimated, that 1.67 million unknown viral species exist, and among them, up to 827,000 could potentially infect humans [The global virome project, A strategy to extimate unknown viral diversity in mammals].

Virus discovery focuses on identifying these new and unknown viruses. The Introduction explores the history of viral discovery and introduces a modern, data-driven approach that leverages publicly available, unprocessed data. It presents VirusHunter, a tool for data-driven viral discovery, alongside VirusGatherer, which assembles identified viral contigs. The alternatives to VirusHunter and VirusGatherer will also be discussed, concluding the Introduction with the goals of the master thesis. But first, I will begin the master's thesis by asking the question: 'Why do we need virus discovery?'

1.1.1 Pandemics are increasing and they are caused by viruses

Between 1918 and 1919, an avian influenza virus, previously exclusively infecting waterfowl, spread to farm poultry and pigs. Eventually, it breached the species barrier, posing a global threat to humans. It infected one-third of the global population with more than 50 million deaths. Historically, major pandemics like the 'Spanish Flu' of 1918-1919 were rare, occurring 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 habitats.

The first 20 years of the 21st century have seen six major pandemics alone. The first was the SARS pandemic from 2002 to 2004, caused by the SARS-CoV virus from the *Coronaviridae* family. This was followed by the H1N1 'swine flu' pandemic

in 2009. Between 2012 and 2016, the world faced several viral outbreaks, including the Middle East Respiratory Syndrome (MERS), the Ebola virus pandemic, and the Zika virus pandemic. Most recently, the global COVID-19 pandemic, caused by SARS-CoV-2, emerged in 2019.

Notably, all six major pandemics of this century have been caused by viral infections, with the majority originating from animals and being transmitted through airborne means. Viruses that emerge through zoonosis face no existing immunity in the new host species, as the host's immune system has not previously encountered them. This lack of immunity, combined with airborne transmission, increasing population densities, and intensified human-animal interactions, allows these viruses to spread rapidly across entire populations.

Effective prevention of global pandemics relies on the prompt identification of the infecting virus, as developing diagnostic tools, vaccines, and treatments depends on knowing the virus and its disease mechanisms. Monitoring existing viruses and discovering new ones has long been feasible through methods like metagenomic analyses of wastewater [**Raw sewage harbors diverse viral populations,Diverse circovirus-like genome architecture revealed by environmental metagenomics, Eukaryotic viruses in wastewater samples from US;].** However, viral discovery is considerably more challenging than identifying bacteria or fungi. While bacteria can be identified through cultures and 16S rRNA sequencing, and fungi through internal transcribed spacer (ITS) region sequencing, viruses lack standardized culture methods and universal sequencing markers. 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 viruses. 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 [**Did Climate change influence the emergence, transmission and expression of the covid-19 pandemic**].

1.1.2 The virome: ‘our’ viruses

Besides the identification of viruses during possible pandemic and disease scenarios, viral discovery could 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 particles, 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 host.

For instance, the digestive system contains the most abundant population of viruses, most of which target the bacteria in our gut. 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 stability [**The human virome: assembly, composition and host interactions**]. The integration of the viral genome can benefit the host, provided that lysis does not occur, since the insertion of new sequences and transposition can lead to mutations and subsequent adaptations [**Ecological and evolutionary Benefits of temperate Phage: What does or doesn’t kill you makes you stronger**], or protect the bacteria from further infection by granting immunity through phage-encoded factors such as CRISPR-Cas sequences targeting other phages [**Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions**]. 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 [**“I will survive”: a tale of bacteriophage-bacteria coevolution in the gut; Explaining microbial population genomics through phage predation; Coevolution with viruses drives the evolution of bacterial mutation rates: Coevolution with bacteriophages drives genome-wide host evolution and constrains the**

acquisition of abiotic-beneficial mutations]. 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 bacteria [**Subdiffuse motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters**]. 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 cancer [Expression of multiple human endogenous retroviruses surface envelope proteins in ovarian cancer], prostate cancer [Selective changes of retroelements expression in human prostate cancer], breast cancer and lymphoma [Human Endogenous Retrovirus K elements in plasma of people with lymphoma and breast cancer], and melanoma [The activation of human endogenous retrovirus K is implicated in melanoma of malignant transformation]. Conversely, some HERVs have beneficial roles, such as contributing to human placental development [Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis]. 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 [**Adenovirus infection and disease in paediatric hsc transplant; The complex relashionship between human herpesvirus 6 and acute GvHD; HHV-6 Reactivation and associated sequelae after HCT**]. In contrast, viruses such as *Redondoviridae* and *Anelloviridae* are generally controlled by the host’s immune system and are not commonly associated with disease in healthy or immunocompromised individuals [human virome: assembly ….].

Moreover, various host-related factors can influence the virome's composition and abundance over time. Under normal circumstances, the virome develops shortly after birth [**Viral diversity and dynamics in an infant gut;Amniotic fluid from healthy term pregnancies does not harbor a detectable microbial community**], remains stable through adulthood, and decreases in old age [**The gut virome database reveals age-dependent patterns of virome diversity in human gut**], but smoking, for instance, has been shown to alter the composition of phages in the lung [**The human gut virome: Inter-individual variation and dynamic reponse to diet**]. Dietary habits also influence the virome, with many gut viruses originating from the food we consume, including plant-based viruses from the family *Virgaviridae*. 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 diets. A shift in diet throughout life can also influence the gastrointestinal virome. A study 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, the study also detected a shift toward more lytic viral communities, suggesting that phages can influence bacterial composition in response to dietary changes [**Fecal viral community response to high fat diet in mice**].

Recent studies in both twins and non-twins emphasize the greater influence of shared environmental factors compared to genetic factors in shaping the virome [**Gut DNA viromes of malawia twins discordant for severe acute malnutrition; Enviroment dominates over host genetics in shaping human gut microbiota**;**Virome diversity correlates with insetinal microbiome diversity in adult monozygotic twins**]. Although genetics can affect the virome, especially in cases of immunodeficiencies [**Genetics of epidermodyplasia verruciformis: Insight into host defense against paillomavirus**], environmental factors like geography and social interactions play a crucial role. For example, one study found that the oral virome of individuals from the same household was more similar compared to unrelated individuals, underscoring the impact of shared living environments on viral exposure. 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 pets [**Pet Animals Were infected with SARS-Cov 19 from their Owners**]. However, 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 humans [**Zootonic disease and virome diversity in bats**]. Geography impacts the virome, with individuals in the highly urbanized city of Hong Kong showing significantly lower virome diversity compared to those in the less urban, partially rural Yunnan province in China. Additionally, Hong Kong residents had an increased presence of phages targeting *Lactobacillus* and *Lactococcus* bacteria [**Human-Gut-DNA Virome Variantions across geography,Ethnicity, and Urbanisation**]. Another study found that Australian children in the Northern Territory with diarrhea had a higher prevalence of viruses from the *Picornaviridae* and *Adenoviridae* families compared to children from Melbourne [**Geographic variation in the eukaryotic virome of human diarrhea**]. The environment shapes our virome, yet viruses also play a crucial role in influencing their surroundings. Viruses lyse 40% of marine prokaryotes each day, a process that stimulates the growth of plankton [The Ecological Virus]. Furthermore, viruses regulate the global carbon cycle by breaking down biomass, which contributes to new dissolved carbon sources in soil and water [**The regulator function of viruses on Ecosystem Carbon Cycling in the Anthropocene**].

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 can also aid in identifying surrogate viruses. The disease mechanisms and host responses of some viruses cannot be directly studied in their natural hosts, complicating the development of treatments and vaccines. In such cases, studying closely related viruses can provide a valuable alternative for research. An example of such a virus is the hepatitis C virus (HCV), a positive-sense-single-stranded RNA virus from the genus *Hepacivirus* within the *Flaviviridae* family, part of the phylum *Kitrinoviricota **[ICTV profile Flaviviridae]***. An HCV infection can result in accumulation of scar tissue in the liver due to chronic inflammation (liver fibrosis), followed by impaired liver function (decompressed cirrhosis), and potentially develop into liver cancer (hepatocellular carcinoma, or HCC). According to the global hepatitis report 2024, 50 million people are living with HCV infection, with nearly 1 million new cases reported in 2022 alone [**Golbal hepatitis report 2024**]. HCV is prevalent globally, with drug users particularly at risk. In the United States, the opioid epidemic has further increased HCV infection rates among individuals under 30 who use injection drugs [**Increase in hepatitis C virus infection related to injection drug use**]. Current treatment for chronic hepatitis C infection in both children and adults relies on inhibiting components of the viral replication machinery and polyprotein processing using direct-acting antiviral regimens, such as sofosbuvir/declatasvir,sofosbuvir/velpatasvir, or glecaprevir/pibrentasvir [**Hepatistic report 2024**]. Yet, despite the availability of these treatments, access is still restricted in under-resourced areas, for those without sufficient health coverage, and among injection drug users who face a higher risk of reinfection [**ethnic and cultural determinants influence risk assessment of hepatitis c acquision;Implementanting scaling up HCV treatment services for people who inject drugs and other high risk groups in ukraine**]. Since certain populations are at higher risk of reinfection and may not afford repeated treatments, developing vaccines as a long-term solution is necessary.

Clinical studies for vaccine development typically involve a phase where the vaccine is tested on animal models [**The clinical development process for a novel preventive vaccine: an overview**]. For HCV, chimpanzees have been used as

animal models because they are the only species, besides humans, that naturally hosts the virus. In chimpanzees, the first HCV clone was developed [**Transcripts from single full-length cDNA clone of hepatitis C viruses are infected when directly transfected into the liver of a chimpanzee; Transmission of hepatitis C by intrahepatic inoculation with transcribed RNA**], and significant insights into the role of CD8+ and CD4+ cells during infection were obtained [**Memory Cd8 T cells are required for protection from persistent hepatitiss c virus infection;conserved hierarchy of helper T cell responses in a chimpanzee during primary and secondary hepatitis c infections.**]. Despite these advances, chimpanzee use in biomedical research is now heavily restricted following the national center for research' decision to halt funding for chimpanzee breeding [The **beginning of the end for chimpanzee experiments**] and their 2015 designation as endangered by the U.S fish and wildlife service [US Fish and Wildlife serice **finalizes Rule listing all chimanzees as eendangered**]. Consequently, the national institute of health (NIH) has since ceased support for such studies as well [**NIH will no longer support biomedical research on chimpanzees**].

Since then, alternative approaches have been sought to ensure continued HCV research. One approach involves discovering and developing new model organisms, particularly in small animals, but this has proven challenging due to HCV's limited host range [**Animal Models used in HCV research**]. Another strategy relies on discovering new viruses closely related to HCV as potential replacement for HCV and chimpanzees. So far, over 250 viruses from the Flaviviridae family have been discovered, with hosts being both primates and other mammals such as horses, dogs, and bovines [**Animal Models used in HCV research**]. In 2016, the discovery of the first *Hepacivirus* infecting non-mammals in catsharks significantly broadened our understanding of *Flaviviridae* diversity, as *Hepaciviruses* had previously been known to infect only mammals, and before 2011, exclusively humans and primates [**Divergent Viruses Discovered in Arthropods and Vertebrates revise the evolutionary history of Flaviviridae and related viruses; The Strange expanding world of animal Hepaciviruses**]. As for primates, the George Baker Virus B (GBV-B) has shown to be highly valuable in research. GBV-B, a positive-sense single stranded RNA virus of the family *Flaviviridae* and genus *Hepacivirus*, was first identified in tamarins that developed

hepatitis after exposure to serum from a patient named George Baker [**Animal Models used in HCV research, Identification of two flavivirus-like genomes in the GB hepatitis agent**]. Tamarins are more manageable in laboratory setting due to their smaller size and their immune response after infection shows similarities to humans. A virus more closely related to HCV than GBV-B is the *equine hepacivirus*, which has also been found in canines (*canine hepacivirus*). *Hepaciviruses* have been also identified in other mammals, such as bats, cattle, and rodents. However, these non-primate *hepaciviruses* have not yet proven to be effective substitutes for HCV [**Animal Models used in HCV research]**. For that reason, GBV-B serves as the current surrogate for HCV in research, but it has its limitations. Chimeric vectors combining HCV and HBV-B sequences exhibit reduced efficacy [**Animal Models used in HCV research]**.Moreover, tamarin breeding is costly, and while chronic infection is a common symptom of HCV, tamarins rarely develop it naturally after GBV-B infection, with current lab methods proving inadequate for establishing chronic infection. Similarly to research involving chimpanzees, studies with tamarins raise ethical concerns about animal welfare and necessitate specialized facilities and expert veterinary care, driving up costs [**Marmosets as models of infectious diseases**].This highlight the need for further optimizing current methods, and also identifying additional surrogate viruses for HCV.

The role of viral discovery is of note here. Prior to 2011, *hepaciviruses* were thought to infect only humans and chimpanzees. However, the discovery of new mammalian and non-mammalian *hepaciviruses* has revealed the diverse nature of *Flaviviridae* and enabled continued HCV research despite the restrictions on chimpanzees as models. Meaning, viral discovery enables us to identify both new and known viruses, understand their evolutionary history and relationship, and use closely related viruses as surrogates for other members of their family.

1.2 History of viral discovery

The history of virus discovery begins with the work of Russian biologist Dmitri Iosifovich Ivanovsky in the late 19th century. While studying at the University of Saint Petersburg, Ivanovsky investigated a disease impacting tobacco plant crops in Ukraine and Bessarabia, and named it "pock virus". Later, he was sent to Crimea to investigate tobacco plants affected by brown spots, which were consistent with the disease he had described. The plants also exhibited dark-green and yellow areas

on their leaves, mirroring the symptoms of tobacco mosaic disease. This disease was first described by Adolf Mayer, who noted that it could infect nearby tobacco plants but would lose its infectious quality if bacteria were filtered out from the plant sap. To test this hypothesis, Ivanovsky used a porcelain Chamberland filter, which retained bacteria and other microorganisms but allowed particles smaller than bacteria to pass through. He then inoculated unaffected plants with the filtered homogenate from affected plants. The inoculated plants soon exhibited the same distinct mosaic pattern on their leaves, demonstrating that the disease was caused by an agent smaller than bacteria. The discovery of the tobacco mosaic virus was published by Ivanovsky in 1892, revealing the existence of infectious agents smaller than bacteria. In 1989, Martinus Willem Beijerinck independently identified the tabacco mosaic virus and introduced the concept of a virus as infectious particles that could only replicate inside living cells. Seven years after Ivanovsky and the discovery of the first (plant) virus, the presence of viruses in animals was confirmed. 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 TMV, 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 1941. 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 genome. Today, viruses can be identified by both their structure and nucleic acid sequences, with new technologies such as polymerase chain reaction (PCR) and second- and third-generation sequencers 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 a. conducted the first study of environmental viral communities, identifying a large and diverse collection of phage sequences in seawater. 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 Internet. 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 openly. 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 access **[Siehe Figure** …].

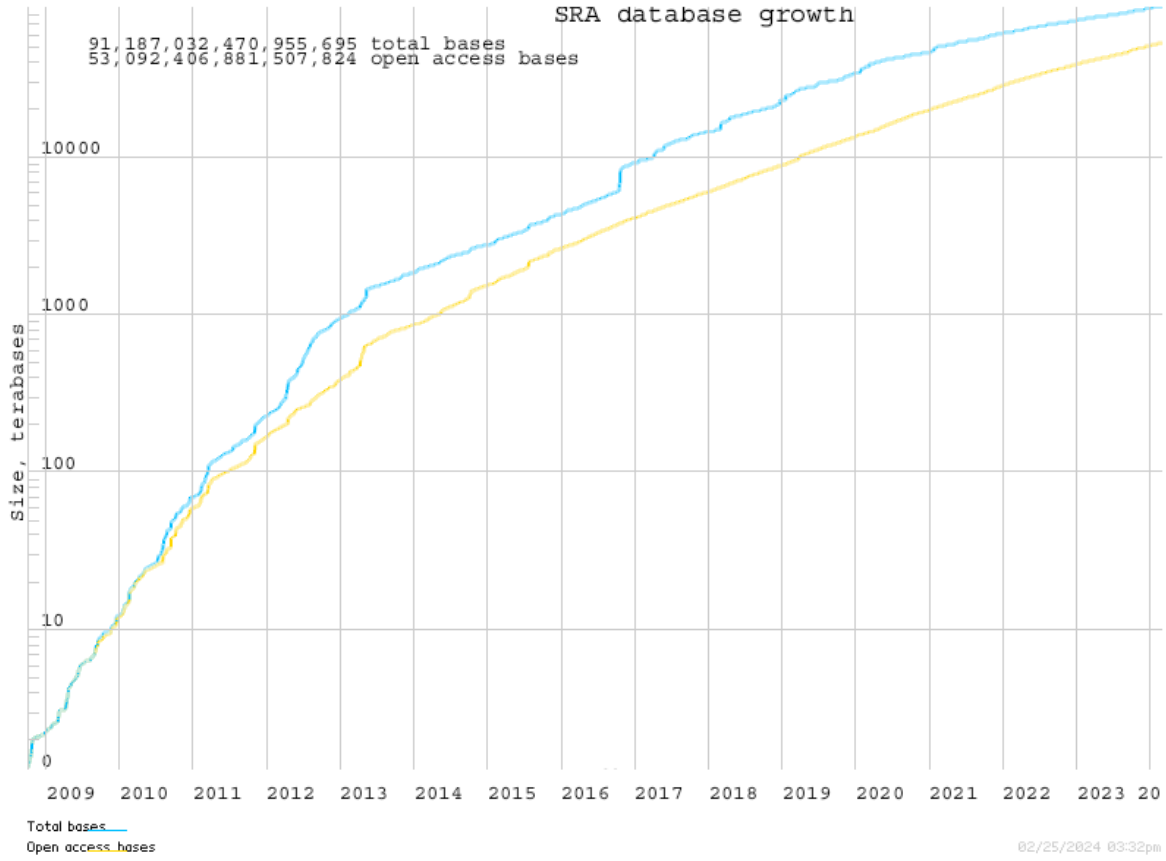

Figure 1: Growth of the Sequence Read Archive database. Since 2007, the total base count has expanded to 91.91 petabases, with 53 petabases being open access.

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 **[Quelle]**. Lauber et al. highlighted that raw sequencing data from public repositories, like those in the SRA, offer new opportunities for virus discovery

using computational approaches. 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. 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, DDVD offers a novel approach to exploring the natural diversity of viruses, complementing traditional discovery methods that rely on wet lab experiments.

1.3 VirusHunter and VirusGatherer

Various computational solutions have emerged for virus discovery [**Quelle**]. Zayed et al. compared RNA-dependent RNA polymerase (RdRp) sequences across 28 terabases of RNA data from the global ocean. By aligning these sequences with the RdRp motif, they uncovered 5,500 new RNA viruses, demonstrating the potential of finding novel viruses in large datasets. Another significant development, Serratus, the first cloud-based platform for viral discovery, analyzed 5.7 million samples, totaling 10.2 petabases. By aligning these samples with three conserved regions of the RdRp catalytic core, Serratus identified 132,000 new viruses from publicly available SRA data. Additionally, another software solution has attempted to

integrate viral discovery with cohort analysis, focusing on identifying which novel viruses could be linked to specific diseases in patients.

The VirusHunter and VirusGatherer pipeline represents another advancement in data-driven virus discovery [See fi**gur**e …]. VirusHunter identifies highly conserved viral sequences, either from local or SRA-downloaded unprocessed sequencing data, by matching them against protein profiles in a profile Hidden Markov Model indicative of a specific virus group. Thus, it is the VirusHunter step that carries out the core step of DDVD. The reads identified by VirusHunter as viral sequences are then used as seed in the next stage, VirusGatherer, where overlapping reads are assembled to reconstruct longer viral contigs or complete viral genomes. The following subchapter details the VirusHunter and Gatherer pipeline (hereafter sometimes referred to as `VirusHunterGatherer`).

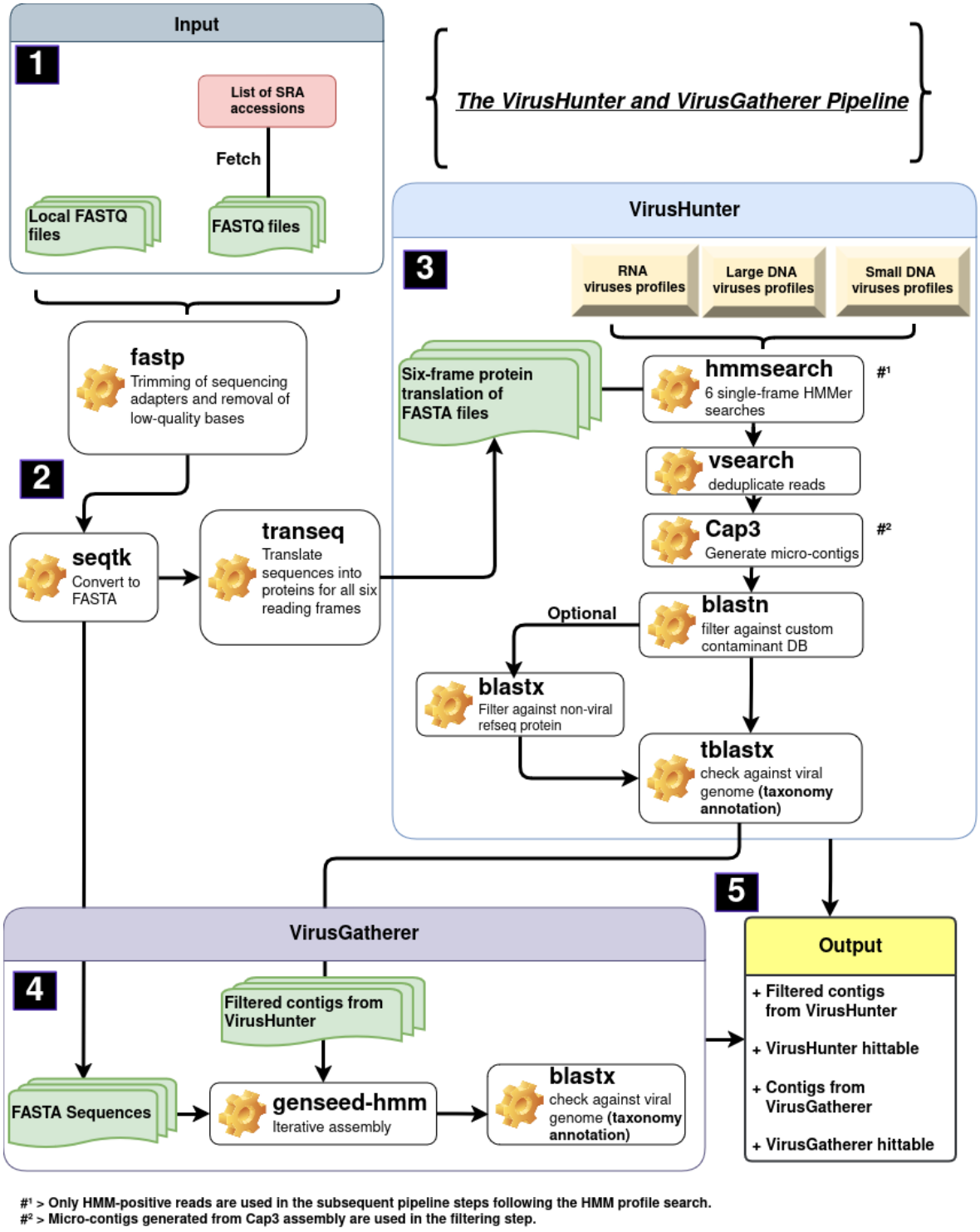

Figure 1: The VirusHunter and VirusGatherer pipeline can process local FASTQ files or download sequencing data from the NCBI SRA database using SRA accessions (1). For both VirusHunter and VirusGatherer, the FASTQ files must be processed and converted into FASTA format (2). For VirusHunter, the nucleotide sequences need to be translated into proteins across all six reading frames. (3) A profile hidden Markov model, with profiles for either DNA or RNA viruses, is employed to detect viral reads within the translated six reading frames. After deduplicating the detected viral reads, they are assembled into micro-contigs. Custom contaminants and, optionally, non-viral reference sequences are filtered out. Finally, taxonomic annotation is conducted by pairwise aligning the remaining viral contigs against a viral database. (4) The contigs from the VirusHunter output, along with the initial input data that has been quality controlled and converted into FASTA format without protein translation, are used for iterative assembly of longer viral contigs. A final taxonomic annotation is then performed on these longer contigs. (5) The outputs from VirusHunter and VirusGatherer include contigs and hittables, which document the best results.

1.3.1 ‘Hunting’ for viruses in unprocessed data

The Basic Local Alignment Search Tool (BLAST) and its variants have become standard tools for finding regions of similarity between a query sequence and a database. They are widely used to identify novel sequences and have been employed in viral discovery [Quelle], including computational pipelines like VirusSeeker [Quelle]. However, BLAST has been shown to lack the sensitivity required to detect distant relationships between sequences [Quelle]. This complicates virus discovery, as distantly related viruses may exhibit low sequence identity with the viral reference genome used for alignment. To address this, alternative methods have been developed to detect evolutionary relationships among sequences with low identity [Quelle zu PSI BLAST, pHMM und ISS]. 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, as implemented in VirusHunter, exemplifiy 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 mentioned above. Rather than using the collection of sequences directly, the MSA is converted into profile hidden markov models, which use a position-specific scoring system to generate a probabilistic representation of the sequences.

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 domain. It is this ability of profile hidden markov models to asses the probability of a query sequence aligning to a viral

protein family, rather than individual sequences, that increases their sensitivity in comparison to BLAST. That is why in a previous study, VirusHunter identified homologs with as low as 35 % identity to viral reference genomes.

Both locally available sequencing data and downloaded sequencing data from the NCBI SRA database can serve as input for VirusHunter. For the latter, a list of SRA accessions that need to be retrieved should be provided **(Figure … 1)**. After trimming potential sequencing adapters and removing low-quality bases, VirusHunter converts the FASTQ files into FASTA, removing the quality indicators in the fourth line and retaining only the sequences and headers. It then translates the nucleotide sequences into proteins for all six reading frames (**Figure 1 2**). Alignment is conducted at the proteome level because homologous viral proteins often diverge beyond recognition at the nucleotide level [Double-stranded DAN virusws: 20 families and only five different architectural principles for virion assembly; Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: inference of protein evolution]. 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 host [Quelle: Mutation rates,mutation frequencies, and Proof-Repair Activities in RNA virus Genomes.]. Viral proteins, on the other hand, are the functional building blocks of biology and tend to preserve their structure and function over long evolutionary periods [Are viruses a source of new protein folds for organisms –Virosphere structure space and evolution]. This conservation allows for the comparison of viral protein sequences, even among distantly related viruses [Use of database of structureal alignments and phylogenetic trees in investigating the relashionshio between sequence and structural variability among homologous proteins, Comparison of solvent-inaccessible cores of homologous proteins; The relation between the divergence of sequence and structure of protein; Evolution of function in protein superfamilies, from structure perspective;protein-level assembly increases proteins equence recovery from metgenomc samples manyfold]. The translated query sequences are aligned against profiles of RNA, small DNA, or large DNA viruses (**fig 1 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 amplification [On the causes, consequences, and avoidance of PCR

duplicates: towards a theory of library complexity], 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 similarity [Viral proteins acquired from host convergence to simplified Domain Architecture]. 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.

1.3.2 Viral taxonomy

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. These include *Lenarviricota*, *Pisuviricota*, *Kitrinoviricota*, *Duplornaviricota*, *Negarnaviricota*, as first described by Wolf et al., and the recently added phylum *Ambiviricota*. As of September 9, 2023, the International Committee on Taxonomy of Viruses (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* 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 eukaryotes.

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 genera. *Picornavirales* is the largest order within this

phylum, which is why *Pisuviricota* is also referred to as the 'picornavirus supergroup' [Quelle: Origins of evolution of the global RNA virome]. This order comprises positive-sense RNA viruses, with *Picornaviridae* and *Secoviridae* 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 insects [Quelle zu iflaviviridae]. *Nidovirales* is the second-largest order and includes 14 families, such as *Coronaviridae*, *Tobaniviridae*, and *Roniviridae*. These viruses are notable for having some of the largest known RNA genomes [Quelle:Nidovirales; evolving the largest RNA virus genome]. 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 2019. 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 invertebrate.

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*. 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). Besides flavivirus-like viruses, the *Kitrinoviricota* phylum also includes the class *Alsuviricetes*, formerly known as the 'alphavirus supergroup.' While most viruses in this class infect plants, exceptions include *Paslahepevirus balayani* from the *Hepeviridae* family, known as hepatitis E virus, and *Rubivirus rubella* from the *Matonaviridae* family, known as the *rubella virus*.

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 subunits. To date, 321 viral species have been identified across 39 genera, 22 families, 3 orders, and 3 classes. Among those 22 families is *Cystoviridae*, which contains the only double stranded RNA viruses that can infect prokaryotes. 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 A.

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 classes. Notable examples include *Rabies lyssavirus* (rabies virus), *Zaire ebolavirus* (Ebola virus), and the influenza viruses, which belong to the *Orthomyxoviridae* family.

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

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*. 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 families. An example for the former include herpesviruses such as Epstein-Barr virus, Kaposi's sarcoma virus, and herpes simplex virus types 1 and 2, all of which can infect humans. An example of the latter is members of the class *Caudoviricetes*, part of *Uroviricota*, which are the most abundant phages in the human virome [Quelle: The humane virome: assembly,composition and host interactions]. *Taleaviricota* currently comprises 32 archaeal viral species, organized into 14 genera and 5 families.

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

Cressdnaviricota is the first phylum of small DNA viruses, comprising 1,490 viral species, 266 genera, and 23 families. This phylum is characterized by eukaryotic viruses with circular single-stranded DNA genomes that encode the Rep protein, enabling genome replication through a rolling-circle mechanism. The *Geminiviridae* and *Nanoviridae* families include plant-infecting viruses, while *Bacilladnaviridae* infect diatoms, and *Circoviridae* are found in mammals, birds, and fish. *Hoffneiviricota* is a phylum comprising a single class and order. It includes 60 prokaryotic viral species, 32 genera, and 3 families: *Plectoviridae*, *Paulinoviridae*, and *Inoviridae*. Similarly, *Phixviricota* also feature just one class and order. This

phylum encompasses 22 species and 7 genera, with *Microviridae* as its sole family. The penultimate small DNA viruses phylum, *Cossaviricota*, includes 440 species and 90 genera across 4 families (*Bidnaviridae*, *Polyomaviridae*, *Papillomaviridae*, and *Parvoviridae*), while the last phylum, *Saleviricota*, comprises 16 species and 3 genera within a single family (*Pleolipoviridae*). All viruses in the *Pleolipoviridae* family target *Halobacteria*, which are halophilic archaea. Human papillomavirus (HPV), a prevalent sexually transmitted infection (STI), can lead to warts and, in some cases, penile, vaginal, anal, or cervical cancer. Other members of the *Papillomaviridae* family have been documented in mammals, birds, and fish. Similarly, *Polyomaviridae* and *Parvovirinae* have been observed in mammals, birds, and fish. However, *Densovirina*, a subfamily within *Parvoviridae*, target invertebrates such as insects and crustaceans.

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* (small DNA viruses), *Yaraviridae* (large DNA viruses), and *Birnaviridae* and *Permutotetraviridae* (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 *Artvericota* falls into this category. *Artvericota* also belongs to the realm *Riboviria*, like the six phyla defined as RNA viruses. However, *Artvericota* is distinct in that it falls under the kingdom *Pararnaviriae*, which includes viruses with RNA genomes that utilize a reverse transcriptase. Examples include the family *Hepadnaviridae* and the genus *Lentivirus*, some members of which are associated with acquired immunodeficiency syndrome (AIDS).

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 taxon [Guidelines for public database submission of uncultivated virus

genome sequences for taxonomic classification]. 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-HMM [GenSeed-HMM: A tool for progressive assembly using profile HMMs …] (**fig 1 4**). A key advantage of GenSeed-HMM it 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**). The sequences are then assembled into longer contigs using third-party assemblers like Cap3, Newbler, Velvet, SOAPdenovo, or ABySS. 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 (**fig 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 evolutionarily. 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. 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 profiled 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 infrastructure. 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 27 % protein sequence identity.

1.4 Aim of thesis

The aim of this thesis is to apply the VirusHunterGatherer pipeline developed by the Computational Virology Research Group at Twincore on two different datasets. 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 first dataset provided by the working group Taubert from the Department of Gastroenterology, Hepatology, Infectious Diseases, and Endocrinology at MHH consists of sequencing data from liver biopsies of liver transplant patients. The complete VirusHunter and VirusGatherer pipeline is employed on this dataset across all three viral groups (Small DNA, Large DNA, and RNA viruses) to identify viruses in the liver sequencing data and assemble the resulting viral contigs. The focus is on examining whether viruses present in the patient data could be linked to transplant rejection.

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, including coding-complete viral genomes, are assembled with VirusGatherer, which has not been done before. The assembly of longer viral contigs allows for the potential discovery of both known and novel viruses, which relies on generating coding-complete viral genomes.

Finally, one of the outcomes of this master's thesis was the development of the R package Virusparies. 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 as part of this work have been provided to the 'Computational Virology Research Group,' led by Jun. Prof. Dr. Chris Lauber. The VirusHunterGatherer tool, which was utilized in this study, is available for download on GitHub at https://github.com/lauberlab/VirusHunterGatherer. Additionally, an R package named 'Virusparies,' which includes functions for processing and visualizing VirusHunterGatherer output and generating summary statistics, was developed as part of this master's thesis. The package is available at https://github.com/SergejRuff/Virusparies.

2.2 Data Availability Statement

All data generated during this study, including the VirusHunter and VirusGatherer output tables, are available in a dedicated GitHub repository created for this master's thesis. This repository contains the specific 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 in raw and unprocessed sequencing data was first performed via VirusHunter, which conducts a homology search with profile Hidden Markov Models (pHMMs) of proteins specific to a virus group to identify potential viral sequences. Following this, VirusGatherer utilized these identified sequences, or micro contigs, from the VirusHunter step as seeds to assemble the complete viral genomes or larger yet incomplete viral contigs. Twincore´s 'Aeternitas' computing cluster was utilized to concurrently execute VirusHunterGatherer across multiple runs. VirusHunterGatherer is implemented in Perl but was executed using Snakemake, a Python-based workflow management system. Meaning, each component of the VirusHunterGatherer pipeline was represented by a rule, managed by Snakemake. A configuration file in .yaml format specified the paths of filter databases and input

data on the 'Aeternitas' server. Both local FASTQ files and a list of SRA accessions, which are downloaded and converted to FASTQ format, are supported as input. Additionally, the configuration file allowed users to enable or disable the filter against host sequences. A table listing all necessary dependencies for runnning VirusHunterGatherer is provided (**Table x**).

2.4 Sequencing data

Two data sets were utilized for viral discovery of large DNA, small DNA, and RNA viruses, with each data set originating from different working groups. The samples in all two sets were derived from distinct viral host organisms.

2.4.1 Taubert liver transplant data

The working group Taubert from the Department of Gastroenterology, Hepatology, Infectious Diseases, and Endocrinology at MHH provided patient liver sequencing data. The data was provided locally in 11 folders, encompassing a total of 323 samples distributed across 557 FASTQ files (**see Table X**).

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

Total Number of Samples: **323** | Total Number of Files: **557**

		Completed		
Folder	Number of Samples	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	✅	✅	✅

Viral discovery was conducted for Small DNA, large DNA, and RNA viruses utilizing VirusHunter. Furthermore, VirusGatherer was employed to assemble viral contigs from the micro-contigs identified during the Hunter step.

2.4.2 Mammalia 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 to this point. The data consisted of SRA data downloaded from NCBI, containing viral reads detected in mammalian hosts, with a total of 34,337 unique SRA accessions. This is in contrast to the Taubert sequencing data, which consisted of locally provided FASTQ data. The hittable was filtered to include only rows where the E-value from the BLAST search against the viral database was below 1e-5, ensuring that only significant results were retained. Additionally, to exclude non-novel viruses, the Hittable was filtered to include only entries with viral sequenced identity below 90% against the viral BLAST database. A minimum of four reads, as identified in comparison with the profiles during VirusHunter, was also required to ensure that some contigs would be assembled during the Virusgatherer step.

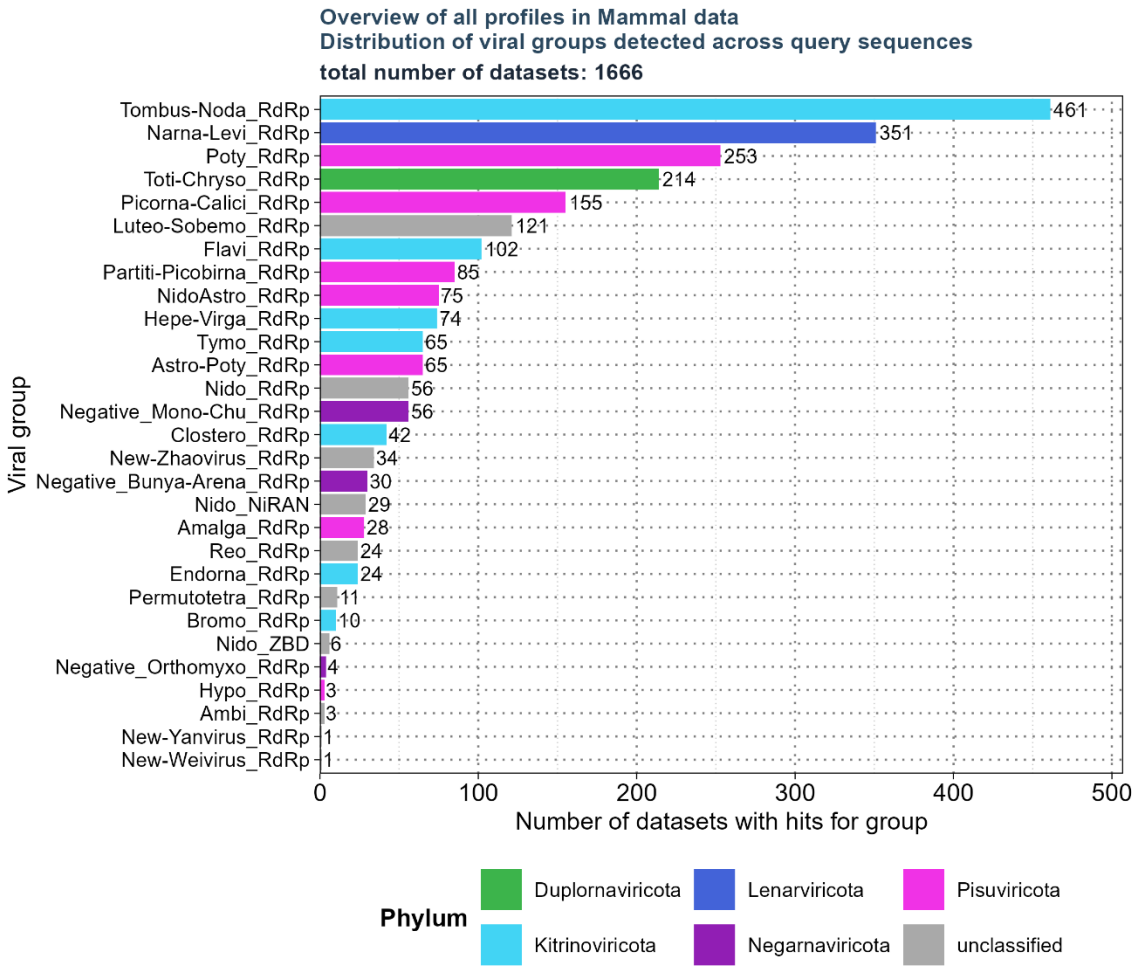

Figure 1: Distribution of viral groups detected across query sequences for all profiles identified after filtering the existing hit table. 1,666 unique SRA accessions with 29 profiles were found after applying filters for E-value < 1e-5, number of reads > 4, and viral sequence identity against the viral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

After filtering, 1,666 unique SRA accessions with 29 profiles remained (**see figure X**). To further limit the number of profiles required for analysis, eight profiles of interest were selected from the filtered Hittable. These included Flavi_RdRp, found in 102 out of 1,666 SRA accessions; NidoAstro_RdRp, found in 75; Hepe-Virga_RdRp, found in 74; Nido_RdRp, found in 56; Negative_Mono-Chu_RdRp, found in 56; Negative_Bunya-Arena_RdRp, found in 30; Nido_NiRAN, found in 29; and Negative_Orthomyxo_RdRp, found in 4. These profiles were then reanalyzed with VirusHunter, which was utilized specifically for RNA virus detection. Identified RNA viral micro-contigs were subsequently assembled into larger contigs with VirusGatherer for the first time in this master's thesis.

2.5 VirusHunter

At the first stage of the data-driven Virus discovery, VirusHunter was employed to identify viral sequences within raw sequencing datasets using a sensitive homology search with profile Hidden Markov Models. This subchapter provides an overview of how VirusHunter was implemented and utilized within the context of this master´s thesis.

2.5.1 SRA download

Both locally available FASTQ files and a list of SRA accessions are valid input files for VirusHunterGatherer. If a list containing SRA accession numbers is not provided, but a path with locally available FASTQ files is given, the locally available data will be used as input. Alternatively, when a list of SRA accession numbers is provided, the SRA Toolkit's prefetch and fastq-dump tools download sequencing data from the NCBI SRA database. Both types of input were used in this thesis: the data provided by Prof. Dr. med. Richard Taubert was available locally on the ‘Aeternitas’ server, while the datasets containing mammalian data were downloaded from the NCBI SRA database.

2.5.2 Preprocessing of FASTQ files for VirusHunter

Trimming of sequencing adapters and removal of low-quality bases occurred in parallel for each file using fastp. Auto-detection identified and removed adapter sequences. Reads shorter than 15 bases, containing more than 5 ‘N’ bases, or having more than 40% of bases with a Phred quality score of 15 were classified as low quality and discarded. For paired-end reads, a single FASTQ file combined both reads in an alternating order. Fastp generates report files in JSON and HTML format, along with a trimmed FASTQ file. The processed FASTQ files were transformed into FASTA format via seqtk, preserving only the sequence information. Thereafter, transeq translated the sequences into protein for each of the six reading frames.

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 tool, 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) domains. However, for hepatitis delta virus and nidoviruses, the profile included hepatitis delta antigen (HDAg) domains for hepatitis delta virus and RNA-dependent RNA-polymerase-associated Nucleotidyltransferase (NiRAN) domain for nidoviruses. For large DNA viruses, profiles mainly comprised major capsid protein (MCP) domains, whereas small DNA viruses queried mostly against the replication initiator protein (Rep) domains. A complete list of all profiles is provided **[Anhang]**. 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 results were considered significant. Score, E-value, and query profile for each significant hit were then exported. FASTA files with non-significant results 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, vsearch deduplicated hits by removing sequences that had identical length and nucleotide content. Deduplicated reads were then assembled to micro contigs via CAP3 assembler. 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 TBLASTX, and only matches with an E-Value lower than 1 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 **[Tabell emit typischen Aufbau für beide haben und drauf verweisen**].

Table 1: Columns of the VirusHunter 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). 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.

Hunter Hittable Structure

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 (QKU0 complete cds
ViralRefSeq_taxonomy	<chr>	taxid:1354498|Hepacivirus colobi|Hepacivirus|Flaviviridae|Amarillovirales|Flasuviricetes|Kitrinoviricota|Orthornavir taxid:64320|Orthoflavivirus zikaense|Orthoflavivirus|Flaviviridae|Amarillovirales|Flasuviricetes|Kitrinoviricota|Ortho taxid:31657|Pestivirus|Flaviviridae|Amarillovirales|Flasuviricetes|Kitrinoviricota|Orthorn
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, fastp 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 CAP3, 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-HMM. 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, BLASTN 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 BLASTX. For the BLASTX queries, the maximum number of reported high-scoring segment pairs (HSPs) – which are local alignments with the highest scores and no gaps – was set to 1. Additionally, the maximum number of target sequences returned per query was also set to 1.

2.6.3 Output of VirusGatherer

Table 2: 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 acc:YP_073558|RNA-dependent DNA polymerase [lymphocystis disease virus-China]
ViralRefSeq_taxonomy	<chr>	taxid:1076255|Centapoxvirus|Chordopoxvirinae|Poxviridae|Chitovirales|Pokkesviricetes| taxid:2027899|Myranavirus phabba|Myrnavirus|Ceeclamvirinae|Caudoviricetes|Urovirici taxid:256729|Lymphocystis disease virus 2|Lymphocystivirus|Alphairidovirinae|Iridoviridae|Pimascovirales|Megaviricetes|Nucleoc
date_analyzed	<date>	2024-06-26, 2024-06-26, 2024-06-30

VirusGatherer returned the sequences of the assembled viral contigs/ whole genomes. Furthermore, hittables similar to the VirusHunter Hittables were generated **[Bild von VirusGatherer hittable aufbau]**. 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.

2.7 Virusparies

One of the outcomes of this master´s thesis was the development of the R package Virusparies. 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. The relationship between viral reference sequence identity and the negative logarithm of viral E-values was depicted in scatter plots. When VirusGatherer were used as input, a bubble plot was generated instead, with contig lengths represented by the size of the bubbles. 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.

3 Results

4 Discussion

Vorteile von VirusHunterGatherer

- Hohe Sensitivität durch pHMM
- Nicht gebunden an kostenpflichtige Server wie Serratus.
- Map Host direkt zum gefundenen Virus. So können Surrogate gleich direkt zum Tier gemapped werden und das Tier als Model identifiziert werden.

Nachteile VirusHunterGatherer

- Sequenzen mit geringer Komplexität und hoher Ähnlichleit sorgen für Probleme bei Homology Search Approaches

-Lange Laufzeit von VirusHunter

-Kann nur Sequenzen assemblieren, die im pHMM ein Hit waren; Vien mit segmentierten Genomen wie Influenza haben das problem, dass nur 1/8 Segmenten das RdRp hat. Die anderen segmente brauchen ihr eigenes pHMM. Würde andere Segmente nicht erkennen. Bei neuen Viren sind nicht alle Segmente bekannt. Suche nach segmentierten Viren erschwert.

Virusparies:

- Vorteile von r packages hier gültig: Reproduciability of methods.
- Weiterhin hat es weiteren Nutzen für zukünftige Projekte.

5 Appendix

6 List of references