DISSERTATION | DOCTORAL THESIS

Titel | Title

Viral bioinformatics and phage-bacteria interactions

verfasst von | submitted by

Lovro Trgovec-Greif bacc.biol.mol. MSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 794 685 437
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Biologie
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Thomas Rattei Shawna McCallin PhD

Contents

1 Abstract **5**
1.1 Summary 5
1.2 Zusammenfassung 5

2 Introduction **7**
2.1 Viruses and bacteriophages 7
2.2 Virus evolution 8
2.3 Viral genomics and metagenomics 9
2.4 Bacteriophages in phage therapy 10
2.5 Research questions and the scope 13

3 VOGDB - Virus Orthologous Groups Database **14**
3.1 Context 14
3.2 Contribution 14

4 Phage cocktail content analysis **16**
4.1 Introduction 16
4.2 Methods 17
4.2.1 Phage cocktails 17
4.2.2 Raw read processing and assembly 17
4.2.3 Removal of unwanted contigs 18
4.2.4 Mapping to orthologous groups 18
4.2.5 Comparison of phage cocktails 18
4.3 Results 18
4.3.1 Sequencing data overview 18
4.3.2 Taxonomic composition of the phage cocktails 18
4.3.3 Comparison in terms of the vFAM content 20
4.3.4 Agreement of vFAM content overlap and the overlap of family composition . 21
4.4 Discussion 21
4.4.1 VRPYO1997 and VRPYO2014 21
4.4.2 PR15 and PG pyophage and Smono 21
4.4.3 ColiProteus (CP12, CP15) cocktails 21
4.4.4 Benefits of using VOGDB for the analysis of phage cocktails 22
4.4.5 Shortcommings of using VOGDB for cocktail analysis 22

5 Viruses of Great Apes **23**
5.1 Introduction 23
5.2 Methods 24
5.2.1 Input data and material 24
5.2.2 Preprocessing of the sequencing reads 24
5.2.3 Filtering of reconstructed proteins 24
5.2.4 Software 24
5.3 Results 25

5.4 Discussion . . . 26
5.4.1 Dental calculus samples . . . 26
5.4.2 Opportunities and challenges . . . 27
5.4.3 Detected viral families . . . 28
5.5 Conclusion . . . 28

6 Phage-host interactions 30
6.1 Introduction . . . 30
6.2 Materials and methods . . . 31
6.2.1 Determination of the phage host range . . . 31
6.2.2 Determining traits responsible for protection against a phage . . . 31
6.3 Results . . . 33
6.3.1 Data . . . 33
6.3.2 Defense systems . . . 33
6.3.3 Defense systems responsible for protection . . . 35
6.4 Discussion . . . 37
6.4.1 Limitations . . . 38
6.4.2 Opportunities . . . 39
6.5 Conclusion . . . 39

7 Other involvements 40

8 Conclusion 42

References 44

A VOGDB manuscript preprint 52

B Viruses of Great Apes data report 66

Chapter 1

Abstract

1.1 Summary

The thesis is oriented towards developing viral bioinformatics and is divided into two distinct parts. In the first part a new resource for viral genomics is introduced – VOGDB. VOGDB is a database where viral proteins from RefSeq are grouped into homologous groups of increasing evolutionary distance. First, challenges for detecting orthologs and homologs in viruses are discussed followed by the presentation of methodology for the creation of VOGDB. The results of the clustering are presented and cluster quality metrics are compared with the clusters from other databases grouping homologous proteins. Second, application of VOGDB in two scenarios is demonstrated. In the first scenario I analyzed the dental calculus of museum samples of great apes where I show it is possible to detect traces of past viral infections using automated pipeline based on VOGDB. A second example is an analysis of complex phage cocktails where VOGDB is used to assess the similarity and taxonomic composition of cocktails. The second part of the thesis is about understanding interactions between phages and their bacterial host with the aim of improving the phage selection procedure for phage therapy. The focus is placed on the anti-phage defense systems. I analyze the defense systems present in a collection of clinically relevant E. coli strains isolated from the human urine and compare the results with a published study on a bigger set of genomes. Later, I analyze the interactions between the bacterial strains and a collection of phages determined by spot tests and search for the defense systems that prevent individual phages from lysing certain strains.

1.2 Zusammenfassung

Die Dissertation zielt darauf ab, die bioinformatische Forschung im Bereich der Viren voranzutreiben und ist in zwei deutlich voneinander abgegrenzte Teile unterteilt. Im ersten Teil wird eine neue Ressource für die virale Genomik eingeführt - VOGDB. VOGDB ist eine Datenbank, in der virale Proteine aus RefSeq in homologe Gruppen mit zunehmender evolutionärer Distanz eingeteilt werden. Zunächst werden die Herausforderungen bei der Erkennung von Orthologen und Homologen in Viren erörtert, gefolgt von der Vorstellung der Methodik zur Erstellung von VOGDB. Die Ergebnisse der Clusterbildung werden präsentiert und die Metriken zur Qualitätsbewertung der Cluster werden mit den Clustern aus anderen Datenbanken verglichen, die homologe Proteine gruppieren. Zweitens wird die Anwendung von VOGDB in zwei Szenarien demonstriert. Im ersten Szenario habe ich die Zahnsteinproben von Museumsstücken von Menschenaffen analysiert, bei denen ich zeige, dass es möglich ist, Spuren vergangener viraler Infektionen mithilfe eines automatisierten Workflows auf der Grundlage von VOGDB zu erkennen. Ein zweites Beispiel ist die Analyse komplexer Phagen-Cocktails, bei der VOGDB zur Bewertung der Ähnlichkeit und taxonomischen Zusammensetzung der Cocktails verwendet wird. Der zweite Teil der Dissertation befasst sich mit dem Verständnis der Wechselwirkungen zwischen Phagen und ihren bakteriellen Wirtsorganismen mit dem Ziel, das Auswahlverfahren für Phagentherapien zu verbessern. Der Fokus liegt auf den Abwehrsystemen gegen Phagen. Ich analysiere die Abwehrsysteme in einer Sammlung klinisch rel-

evanter E. coli-Stämme, die aus menschlichem Urin isoliert wurden, und vergleiche die Ergebnisse mit einer veröffentlichten Studie an einem größeren Satz von Genomen. Später analysiere ich die Wechselwirkungen zwischen den bakteriellen Stämmen und einer Sammlung von Phagen, die durch Spot-Tests bestimmt wurden, und suche nach den Abwehrsystemen, die einzelne Phagen daran hindern, bestimmte Stämme zu lysieren.

Chapter 2

Introduction

2.1 Viruses and bacteriophages

Viruses are an interesting biological entity for which it was previously considered that they are between the living and the nonliving world. This was due to their nature of being obligate molecular parasites completely dependent on the host metabolism to proliferate [1] and without their own metabolic capacity. In the current times, it is more accepted to say viruses are alive because they replicate and evolve [2] and as an example supporting the statement, dormant bacteria are given which also don't show metabolic activity. Some authors have suggested that the "real" viral organism is an infected cell, called a virocell, while the viral particle or the virion is only a transient form of a viral organism [3]. What defines viruses is their infectious nucleic acid (either DNA or RNA) and the protein capsid which contains it. The protein capsid is what makes viruses different from other infectious or autonomous nucleic acids like plasmids, viroids and transposable elements [4].

Viruses are known to infect organisms from the entire tree of life including animals, plants, fungi, protozoans, archaea and bacteria. As disease agents they are a major problem for humans. As an example, influenza, a virus causing respiratory infections causes an economic loss measured in billions of dollars [5] while plant viruses threat the agricultural production and cause another few billion dollars of economic damage annually [6]. However, even though viruses are mostly associated with a disease and have negative connotation, they have important ecological roles. Viruses are an important factor in nutrient cycling [7], especially bacteriophages which account for the lysis of around 50% of the marine microbial biomass per day and infect around 10^{23} microbes per second [8]. Moreover, bacteriophages are part of the human microbiome (phageome) and it seems that the phageome composition is associated with the health status of an individual [9]. In terms of the biomass, the total biomass of viruses on earth is estimated to be only 10 times smaller than the total biomass of animals [10].

Even though viruses are so important and abundant, due to their small size which makes them invisible to the light microscope, it was not easy to discover their true nature in the early days of microbiology. The first idea that something like a virus exists came from the experiment with the Tobacco mosaic virus, where a filtrate was still infectious even though the filter size would prevent bacteria from passing through [11]. It was also learned that this new infectious agent is composed of proteins and nucleic acids. Today, we have a vast array of methods to explore viruses including electron microscopy which makes it possible to visualize the viral particle and the nucleic acid sequencing technologies that allow us to read the viral genome with ease. The accumulated knowledge of virus biology allowed for the development of different antiviral strategies that prevent the virus form entry into the host cell or inhibit various steps essential for the virus life cycle [12]. However, except for a few of the viruses problematic for humans, we don't know how most of the viruses function and we can't even predict the function of the most of the viral genes [13]

Today, viruses are in the focus of many different research disciplines. In the medicinal context, the exploration of phageomes (the entirety of bacteriophages in the holobiont) in relation to health is getting in focus more often as well as the use of phages to treat infections and modulate the microbiome. The field of evolutionary biology explores viruses to better understand the origin of life and the origin of genes. Ancient DNA research is focusing on viruses to better understand the previous disease dynamics and the evolution of viral pathogens. Most importantly for the society in general, research is being done to learn more about the contemporary pathogenic viruses to prevent the future damage to human, animal and crop health.

2.2 Virus evolution

When discussing evolution, whether it is the evolution of cellular organisms or viruses, we need the concept of a species and species names. A species is an abstract concept to which we assign biological entities when they meet certain inclusion criteria [14], and it is important to keep in mind the distinction between the virus and the virus species in discussions and writing [15]. Historically, viruses and viral species were named by the descriptive names describing the disease they were causing or the symptoms of the infection [16] as well as by the mode of transmission and replication [17]. Naming of bacteriophages is more diverse with names being greek letters (for example the phage lambda), combinations of letters and numbers (e.g. coliphages T4, T7) or the name of the bacterial host species (e.g. Pseudomonas phage PhiPA3) [18]. However, there are guidelines to make the bacteriophages naming consistent and informative putting the emphasis on the bacterial host in the name [19]. On higher taxonomic levels, bacteriophages were further classified based on the virion morphology with the most abundant group being tailed bacteriophages (former order Caudovirales) and further subdivided based on the tail length and shape (former families Myoviridae, Siphoviridae and Podoviridae). However, it was shown that the groups were polyphyletic and therefore do not represent evolutionary relatedness which signaled the need to reconsider viral taxonomy [20].

In modern microbiology, the tendency is to define species (and higher taxonomic levels) operationally based on the degree of genome sequence similarity [21] and the same principles are applied to the viral taxonomy [22]. Although classifying any virus based only on its nucleotide sequence is still not possible, for uncultivated DNA viral genomes of bacterial and archaeal viruses retrieved from metagenomes there are accepted threshold for demarcating taxonomic categories of different ranks [23]. In dsDNA viruses the threshold for species demarcation is 95% nucleotide identity over the length of 85% of the length of the shorter genome. For ssDNA viruses the percentage of identity is in the range 69-78% [23]. It is clear the species definition is a more complex problem than declaring fixed threshold for the nucleotide identity and those approaches are only useful for the operational and not the conceptual definition of taxa.

Regarding the evolution, viruses are especially interesting as by what is the current opinion, they are a polyphyletic group and have emerged independently on multiple occasions [24]. Some authors suggest that the viruses emerged as RNA molecular parasites using the primordial replicators as hosts in the early days of biological evolution and that viruses as biological form precede the cellular life [25]. However, conflicting views suggest that viruses could have also emerged by some genes escaping the cellular genome and becoming parasitic or by transformation of a cellular organism into a virus [24]. The emergence of molecular parasites, however, might not be just a result of chance. A thought experiment by Koonin et al. [26] demonstrates that parasite free states in simple self-sustainable replicators are evolutionarily unstable. The conclusion is based on the fact that in a population of replicators there is always a chance that an entity that is replicated, but does not produce the replicase emerges and the fact that indefinitely maintaining an absolute defense against such parasites in their absence (if the defense is so efficient) is evolutionarily unfavorable.

Viruses are polyphyletic group and don't share a common ancestor which as a consequence means that there are no genes that are present in all viruses like ribosomal genes are present in cellular organisms. Viral realms are groups of viruses that have common evolutionary origin and share common genes called viral hallmark genes which are usually genes involved in virus replication or are important for the virion formation [27]. Famous examples of viral hallmark genes are the jelly roll capsid protein, RNA-dependent RNA polymerase and the reverse transcriptase [25]. Viral hallmark genes are specific to viruses and there are only very distant homologs among the cellular genes [25]. Except for grouping the viruses into realms and understanding virus evolution, viral hallmark genes are used in bioinformatic tools which try to find sequences of viral origin in a genome or metagenome. An example of such tool is VirSorter2 [28] which aligns genes of a potentially viral sequence to the Hidden Markov Models of the viral hallmark genes to classify the sequence as either viral or not. Databases storing clusters of homologous proteins of a broad set of viruses are essential for studying viral hallmark genes as they serve as a place where already grouped genes can be retrieved and offer a single standardized reference for a wide group of researchers. One such database is VOGDB which will be described in more details later.

Viral genomic sequences change and evolve quickly, but not only due to mutation, but also due to the recombination which is especially present among phages and makes their genomes mosaic [29]. Phage genes in the bacterial genomes are in many cases responsible for the genomic and phenotypic differences between the strains and are often encoding virulence factors causing the bacteria to become pathogenic [30]. Interestingly, many genes that defend bacteria from phage infections are also coming from phages [31]. Genetic exchange is also present in the domain of eukaryotic viruses. It is known that viruses from the family Poxviridae acquire genes from their host which help the virus evade the host antiviral defense [32]. Perhaps the most interesting from the human perspective are the mammalian genes of viral origin. The gene Syncitin 1, a gene important for the human placenta physiology is derived from an endogenous retrovirus [33] together with the a total of around 8% of the human DNA that has retroviral origin [34].

2.3 Viral genomics and metagenomics

Metagenomics is the study of all the genomes present in a (environmental) sample and the first objective of metagenomics was to find novel functional biomolecules in soil and in most part included cloning of the environmental genes into E. coli for analysis [35]. The modern metagenomics is all about sequencing and generating vast amounts of sequence data and a typical workflow includes sampling, extraction of the DNA, sequencing and the analysis of the sequencing data. In the beginning of metagenomics, the focus was on studying bacteria and due to the higher sequencing cost, it was common to sequence the universal marker genes (16S rRNA gene) instead of the entire collection of nucleic acid in the sample [36]. Obviously, sequencing of the 16S rRNA gene would not carry any information about the bacteriophages since they don't carry that gene. Moreover, bacteriophages don't have a conserved set of genes which could be used as marker genes for sequencing. Only when the shotgun metagenome sequencing became more common was it possible to study bacteriophages in the metagenomic datasets.

With a phage genome being much smaller than the bacterial genome (169kb for the E. coli phage T4, NC_000866 as compared to the 4.6mb for the E. coli K12, NC_007779), a typical metagenome will contain a majority of bacterial sequencing reads. To tackle the problem when sequencing the phageome (genomes of all the phages in a sample), an option is to enrich the sample for virus-like particles (VLPs) before sequencing [37]. The enrichment is typically done by filtration where small particles that represent the viral portion of the sample are retained. The choice of the VLP enrichment method will however influence what viruses will be retained in what proportions [38]. It is interesting to note that some bacteriophages cannot be sequenced by using the standard DNA sequencing technologies. Alternative and hypermodified bases are more common in phages than in bacteria and with such DNA molecules the library preparation and sequencing procedures don't work [39, 40]. Tricks like transcriptome sequencing could be used to sequence some of those phages [39], but the efforts like that would not work well for all of such phages in a sample at the same time. Due to this property, entire phage groups are probably missed from the catalog of the global phage diversity.

In the bioinformatic sequence analysis of viruses in metagenomes, multiple challenges are present. The assembly of viral genomes is often fragmented and contigs are not easily matched with the viral genomes from the databases and we need to use separate tools to identify contigs of viral origin [41]. If we estimate the global number of phage species to be 100 million [42] and genomic databases containing not even a million of phage sequences (and not all the nonredundant sequences are from different species) [43] it is clear that the database representation of the phage diversity is low. With the low database representation, it is often difficult to find homologs of the newly discovered viral genes and give them a functional annotation. Due to all the challenges, it is said that viruses in metagenomes are the biological dark matter [44].

Even though virus (phage) oriented metagenomics poses many challenges, in the recent years there has been a lot of bioinformatic development for the analysis of viral metagenomes. Metaviralspades [45] is a version of the short read assembled spades that is tuned to increase the retrieval rate of (full length) viral contigs from a metagenome. CheckV [46] is a tool created after the bacterial counterpart checkM [47] which gives the quality and completeness estimate of a phage contig. Tools like VirSorter2 [28], geNomad [48] and PHASTEST [49] use statistical learning to predict whether a given contig is viral or not. Phanotate [50] is a gene prediction tool specifically designed for phages which takes into account the tendency of phages to have compact genomes with overlapping genes. For binning contigs into genomic bins there are several tools specialized for viral genomes like PHAMB [51] and vRhyme [52] that are based on machine learning. The basis for most of the new tools designed for the analysis of viral genomes and metagenomics, especially those based on machine learning is the quality of the viral databases.

To bioinformatically characterize newly discovered viral genes and genomes, it is crucial to have both databases with the raw genomic information of known viruses and the databases with derived information (for example viral protein families). As more viruses are sequenced and identified, the general purpose genomic and protein databases like NCBI nr and nt [53] will naturally be richer with viral data. However, there are databases specialized for viral information. An example of a database with the primary sequence information of the vast amount of environmental viruses is the IMG/VR [54]. The databases with protein family models are especially interesting since they often provide Hidden Markov Models (HMMs) of the protein family which is often used as input for other tools, for example checkV and VirSorter2. Databases grouping viral proteins into families are pVOG [55] (bacteriophages), eggNOG [56] (bacteriophages and eukaryotic viruses separated) and PHROGs [57] (bacteriophages). It is interesting to note that there is no database where genes from bacteriophages and eukaryotic viruses would be grouped together into shared families. VOGDB (described later) among others aims to fill that gap in the available resources.

Learning and understanding more about the viruses is not only interesting to better understand the origin of life, but there are also practical and medical applications of being able to understand the nature of a phage or a virus from its sequence retrieved from a metagenome. If we were able to predict properties of an unknown virus from its sequence, we could have another tool in fighting the spread of new infectious diseases without doing risky experiments with a live pathogen. On the other hand, knowing about a phage from its sequence could make it possible to efficiently mine the large amount of metagenomes for phages that could for example be used in the phage therapy to treat bacterial infections, an approach that is getting more attention every day given the issue with the spread of the antibiotic resistance among the bacterial pathogens.

2.4 Bacteriophages in phage therapy

The problem of the increased resistance of pathogenic bacteria to antibiotics is given a lot of attention and dark prognoses are made about the economic loss and the loss of lives in the future due to the complications of the infections [58]. So far, bacteria have developed resistance to all known antibiotics and there is no reason resistance would not emerge for any newly developed antibiotic [59]. Bacteriophages could be used in place or in combination with the antibiotics as a form of a biological control of the pathogenic bacteria [60]. There are so many bacteriophages in the world that it is very unlikely that the resistance to bacteriophages would pose a serious problem in case phage therapy would be widely applied.

Besides the benefit of a low probability of a resistance problem similar to the one with the antibiotics, there are several benefits that therapeutic phages offer compared to the antibiotics. First, bacteriophages seem to be safe and they don't cause damage to the patient per se [61]. Second, phages are specific to the bacterial species or strain that they recognize as a host and would not unspecifically destroy the microbiome. Third, when lysing a bacterial cell, phages release progeny which has an effect of locally increasing the concentration of phages at the site of infection. However, there are also challenges associated with the usage of phages for therapy.

Even though bacteriophages are not harmful for humans, they interact with them via the immune system. Antibodies that target phages could be formed which would reduce the efficacy of the administered therapy [62]. During the course of the treatment, bacterial resistance to the specific phage could emerge which would require adaptation of the therapeutic preparation [63]. Third, *in vivo* conditions are different from what phages encounter in the laboratory or in the wild and phages could be inhibited by the host body, for example by plasma or synovial fluid [64]. Lastly, the strong specificity of a phage to a strain could pose a problem when finding an appropriate phage for therapy, especially since a very related bacterial strains could have different phages that infect them. The immune sensitization could be mitigated by special formulation of phage products [65]. The challenge of specificity and the emergence of the resistance seems solvable when we learn more about what makes a certain bacterium a good host for a phage.

When a phage is able to use a bacterium as a host, there are a few obstacles before it can hijack the bacterial cellular machinery to produce viral offspring. The first step of a successful phage infection is the entry of a phage nucleic acid into the host and avoidance of the bacterial anti-phage defense systems [66]. Figure 2.1 show a simplified schematic of obstacles for a phage when infecting a cell. For phage therapy it is important to consider both the barriers for entry and the possibility of neutralization by the host. To become resistant to phages, bacteria can either prevent the phage entry by modifying or blocking the surface receptors [67] or acquire defense systems that disturb other steps of the phage infection process. In cases of phage therapy failures, it is not clear whether it is due to the spread of the defense systems or due to the changes in the receptors. As receptors could be down-regulated without their genetic sequence being changed, simple DNA sequencing would not be enough to answer this question.

Phage therapeutics come in different forms. One way to classify phage therapeutics would be those that are pre-made and available for sale in pharmacies (for example in Georgia and Russia) and personalized preparations made and tested for strains of a specific patient. Pre-made preparations are most often in a form of a complex phage cocktails that have a broad host range and are periodically adapted for the new pathogenic strains circulating in the environment. Personalized products are made for a specific patient and are produced on a small scale, mostly by research laboratories [68]. Both approaches have benefits and drawbacks. Pre-made cocktails can be produced on a larger scale and are readily available, but might not work against a pathogen for which it is applied. Personalized products are designed with a specific strain in mind and it is proven in the laboratory that the phages are able to lyse the bacteria. However, production is slow, labor-intensive and has a small yield. Laboratory producing the product is limited to the phages available in its collection. It is not unusual that phages are requested over mailing lists in hope someone in the world has an appropriate phage in the collection. Phages received in such a way have to be tested with the host strain first to see if they will cause lysis. Even if everything looks promising in the lab, it still often happens that the course of a phage therapy is a failure, often without an adequate explanation. It is therefore crucial to better understand the interactions between phages and their (potential) hosts in order to be able to predict the best phages for the treatment of specific strains.

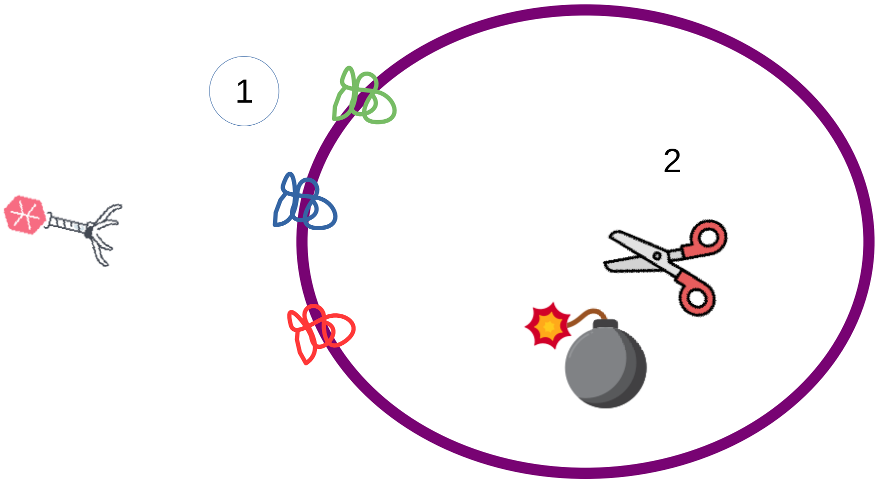


Figure 2.1: Illustration of the phage entry into a bacterial cell. On a high level of abstraction, for a phage to successfully replicate in a competent cell, there are two obstacles it has to solve. The first one is the entry or the injection of the nucleic acid into the host cell. To fulfill that step, an interaction has to be established between the phage receptors and the bacterial surface receptors (1). After the phage nucleic acid is in the cell, it has to survive a plethora of anti-phage defense systems that can cut the foreign nucleic acid or somehow else prevent the replication, including the cell suicide (2).

Phage therapy implies the use of virulent bacteriophages, however it is not necessary to use the complete phages to achieve the therapeutic effect. Phage derived lytic proteins are promising candidates as future antimicrobials as well. Currently, researched are polysaccharide depolymerases and lysins. Compared to the therapy with phages, they offer an additional benefit as they are able to destroy even metabolically inactive cells [69]. Research of lytic enzymes is ongoing and there are efforts to release lysin based medications. On the other side, there is work showing conceptually that temperate phages also have a potential use in phage therapy. By using sub inhibitory concentrations of antibiotics, temperate phages could be activated and they would eradicate the bacteria [70] Another way is to use temperate phages to deliver genes that change the bacterial phenotype to make them less virulent or to resensitize them to antibiotics [71]. As a benefit compared to lytic phages, absence of lysis removes to risk of toxic shock due to endotoxin release [71]. The possibility to use temperate phages for therapy would greatly increase the repertoire of potential phages used to formulate a therapeutic cocktail.

Future development of phage therapy depends on how well can we understand the interactions between the phage and it's host, phage and the human host and the interaction of the phage and it's host inside the human host. One of the obstacles is the lack of understanding of what makes a certain cell an appropriate host to a phage. A study by Gaborieau et al. [72] showed that it is possible to predict phage-host interactions between E. coli and phages based on the genomic data, but only to a certain extent (AUC = 0.85) and the interactions were explained by the adsorption factors. Another study bacteria from the genus Vibrio and phages showed that the defense systems are what defines the phage host range [73]. Another thing to keep in mind is that a parasitic strategy where the host would be quickly exterminated is not a particularly successful one. Phages parasitize on bacteria and have co-evolved for a long time and their interplay shows different dynamics, but it doesn't normally happen that the bacteria disappear. As knowledge of phage biology and ecology is increasing, there is more chance that we find a way to "hack" the phage-bacteria interaction and create phage based therapeutics to tackle the one of the most pressing problems for the future, the bacterial antimicrobial resistance.

2.5 Research questions and the scope

In this dissertation I will describe the work done on four projects I have worked on and mention other projects in which I were involved, but which are not yet mature enough for a chapter or a publication.

In the first part, which consists of three chapters (chapters three, four and five) I work on the development of a new resource for virus bioinformatics (VOGDB) and apply in to the analysis of two types of metagenomes. For the development of VOGDB, one of the main questions was how to properly represent homologous groups of viral proteins. Related to the question is how to validate such grouping. On top of the fundamental questions, another challenge is in the organization and implementation of a large computational pipeline that has to re-create the database with every RefSeq release. Chapter three with the manuscript preprint in the appendix describe the VOGDB, main decisions made during the development and the validation procedure and results.

Chapter four shows the application of VOGDB to phage cocktail metagenomes. The analyses of such metagenomes are rarely present in the literature and are not standardized with authors using custom solutions. This has a consequence of making it more difficult to compare the results across different studies. My goal was to propose an analysis pipeline based on VOGDB that requires minimal user intervention. The pipeline is supposed to output the cocktail composition and the degree of similarity of different phage cocktail metagenomes.

In chapter five, I work on the analysis of metagenomes of the dental calculus from the museum samples of great apes. The goal was to detect signals of the past viral infections. The challenge lies in the very low abundance of the reads representing genomes of interest. I had to find an optimal assembly strategy to deal with the rare reads of interest and find a suitable database searching strategy to filter the assemblies and get the taxonomic information of the mammalian viruses present in the samples. VOGDB was used as one of the databases for searching.

Chapter six is the second part of the dissertation. I wanted to explore the role of the anti-phage defense systems in the *E.coli* - bacteriophage interactions and how can that knowledge be used in improving phage selection for phage therapy. From resources I had an interaction matrix of dozens of bacterial strains with dozens of phages together with the genomic sequences of both bacteria and phages and the information on the receptors phages use to attach to the bacterium. My main goal was to confidently detect anti-phage defense system and based on their presence or absence in certain strain try to deduce which phage are neutralized by what systems. On top of obtaining the desired information, I wanted to create an easily expandable analysis framework which could be used by different labs involved in phage therapy to improve the phage selection process.

Chapter seven lists involvements in other projects and a short description of my role there.

Chapter 3

VOGDB - Virus Orthologous Groups Database

3.1 Context

The initial version of VOGDB was developed as part of another study [74], but its potential usefulness for other virus bioinformatics work was recognized and the decision was made to continue the development and offer it as a stand-alone resource. By the time I started my doctoral studies and got involved into the VOGDB project, the pipeline prototype already existed and gave as the output groups of viral proteins showing signs of remote homology based on the HMM-HMM alignments.

Multiple goals still had to be achieved before the VOGDB publication would be ready and during further development, the original pipeline was improved and in some aspects significantly expanded. For example, after the appearance of Alphafold2 [75], it became possible to predict the structures of proteins from VOGDB. As we had access to the appropriate computational resources for the structure prediction with Alphafold2 (compute cluster with GPUs), a decision was made to expand VOGDB with clusters based on structural similarities. While we were implementing the structural prediction and clustering, we decided to extend what VOGDB offers even more by making available also the protein clusters based on the recent homology (detected by direct sequence comparison) resulting in a layered structure grouping together proteins with increasingly remote homology.

Other important work left to be done was the validation of the VOGDB clusters, stable numbering of the clusters between releases, implementation in a workflow management system and bug fixes.

3.2 Contribution

The original VOGDB pipeline was created with a mixture of bash and python scripts and to make it more readable and maintainable we wanted to use a workflow management system. At first, I tried to re-implement the pipeline using Nextflow [76], but I quickly encountered limitations that made me stop with the further efforts. Nextflow creates multiple files in the Nextflow working directory for every individual process and as there are many steps in the VOGDB pipeline in which proteins or clusters are processed separately, the number of files created by Nextflow became unreasonable (multiple millions). Another workflow management system called Snakemake [77] allowed for much more control of the number of files created in the runs and was therefore used in the further development. As the VOGDB pipeline was already quite big, re-implementation of the complete workflow in Snakemake was deprioritized and only the new functionality (structure prediction and clustering) was implemented in Snakemake and integrated into the main pipeline. I have been involved into the design and implementation of the structural clustering step from the beginning.

Other improvements to the original prototype where I contributed is the design and implementation of the stable numbering of the clusters between releases based on the cluster protein content overlap. I also noticed and fixed a bug related to PSI-BLAST [78] search that caused many of the proteins to not be processed. When doing a PSI-BLAST search with multiple queries in a single file, if it happens that a protein sequence is entirely masked as a low complexity region, the search will stop with a silent error and the proteins coming after the offending sequence will not be processed. I have implemented the pre-filter that removes such problematic proteins before the start of the PSI-BLAST search.

The idea to validate the clustering based on the homogeneity of functional and structural annotations of the protein members existed before, however, I have implemented the step as well as I came up with a randomization test to obtain the statistical significance. As part of the validation, I have made the comparisons with other similar databases.

Regarding the publication, I have created all of the figures and have written most of the text. The preprint of the VOGDB manuscript is in the appendix A of this dissertation. It has been submitted for publication to *MDPI Viruses* in June 2024.

Chapter 4

Phage cocktail content analysis

4.1 Introduction

Phage cocktails are mixtures of phages designed for use in bacteriophage (phage) therapy for treatment of bacterial infections. Phage therapy originated from the time before the discovery and widespread use of antibiotics and is still widely present in some countries where Russia and Georgia are the most notable and there phage product are commercially available [79]. The most popular cocktails are designed for the treatment of skin infections or for gastrointestinal infections. The exact way cocktails are produced is partially a company secret, but we know new phages are periodically added to the cocktails according to the circulating pathogens [79] resulting in cocktails being a complex phage mixture with cocktails of the same name coming from a different time point being different. Production of phage cocktails sometimes employs experimental evolution of phages which is supposed to increase the range of species being lysed by the phage preparation. The experimental evolution procedure is known as the Applemans protocol [80] which as a result increases the genomic diversity of the final phage product.

The use of phages and phage cocktails in medicine is generally not approved by the regulatory agencies in the western world [81]. If phage products are used, in most cases it is for compassionate use [82], while in some countries like Belgium, the use of phages is more accepted and the regulatory framework is more relaxed [83]. The lack of understanding of the exact composition of the commercial phage cocktails is problematic for western medical agencies [84]. With the advancement of the genome sequencing technologies, some authors have tried to better understand the composition of phage cocktails by using metagenomics [85–88]. They focused on reconstructing the full length phage genomes, getting an overview of the taxonomy of reconstructed phage genomes and searching for genes that could compromise the cocktail safety, for example antibiotic resistance genes. Virulence genes, or genes that facilitate genetic exchange [89]. However, commercial phage cocktails contain many phages of different abundances and similarity which makes it challenging for metagenomic methods to provide a detailed insight into their composition.

Metagenomic analysis of commercial phage cocktails is interesting and relevant for multiple reasons. Detailed genomic characterization of such complex phage mixtures is relevant for the regulating agencies that have the power to approve such product for use. By (meta)genomic analysis it is possible to detect potential threats in phage cocktails such as antibiotic resistant genes and virulence factors [90]. Moreover, having better knowledge of the genes and genomes present in the cocktails would increase the understanding of the results of the clinical trials including phage cocktails [91]. Besides practical reasons, metagenomes containing a large number of phage species with different similarities are challenging to analyze and those samples are good candidates for testing current bioinformatic methods and develop new ones.

In this work the aim was to use VOGDB to create an automated pipeline for quick and easy coarse-grained analysis of phage cocktail metagenomes. As the database representation of phages is low compared to the global phage diversity (6151 phage genomes used to create VOGDB version 221 compared to millions of phage species [92]) using virus orthologous groups models is one approach to increase mappings of genes from metagenomes to the known genes. We employed VOGDB to get an estimate of the taxonomic composition of phage cocktails and their composition in terms of the vFAMs from VOGDB. The pairwise similarity between cocktails was estimated from the relative abundance of viral families and the overlap of the present vFAMs. When available, the composition was compared to the already published results obtained using various custom analysis pipelines. For cocktails with the same name from a same producer, but from a different time point we observed the reduction of similarity both in terms of vFAMs and the taxonomic composition.

4.2 Methods

4.2.1 Phage cocktails

Metagenomic sequencing data of 13 phage cocktails created with Illumina paired-end sequencing technology were available to the author. Some of the cocktails were already published while others represent unpublished data. The cocktails originate either from the Eliava institute in Georgia (9 cocktails) or from the Russian company Microgen (4 cocktails). The list of cocktails with their short description is shown in the table 4.1. The cocktails named VRPYO1997 and VRPYO2014 were sequenced by Villarroel et al [87] and the sequences were retrieved from ENA database. Sequencing data of the other cocktails were provided by Dr. Shawna McCallin.

Table 4.1: Phage cocktails used in the study. Cocktail descriptions were retrieved from the producers' webpage (Microgen from Russia and Eliava Institute from Georgia) by the cocktail name.

Cocktail	Origin	Description
CP12	Russia	ColiProteus from 2012
CP15	Russia	ColiProteus from 2015
ENKO	Georgia	Active against Shigella, Salmonella, E.coli and Staphylococcus
Fers	Georgia	Fersis cocktail active against Staphylococcus and Streptococcus
IntiG	Georgia	Intesti cocktail active againss Shigella, Salmonella and E.coli
PG	Georgia	For treatment of purulent and enteric infections
PR15	Russia	For treatment of purulent and enteric infections
Pyo14	Georgia	For treatment of purulent and enteric infections
SES	Georgia	Against Staphylococcus, Streptococcus and Enteropathogenic E. coli
Smono	Georgia	Treatment of S. aureus infections
Sxta	Russia	Polyvalent bacteriophage mixture for inflammatory and enteric infections
VRPYO1997	Georgia	For treatment of purulent and enteric infections
VRPYO2014	Georgia	For treatment of purulent and enteric infections

4.2.2 Raw read processing and assembly

The trimming and quality filtering was done using trimmomatic (v0.39) [93], but was kept minimal. Only the sequencing adapters were removed (TruSeq3-PE-2.fa) and reads shorter than 50bp were removed.

Assembly was performed in two steps to maximize the length of retrieved contigs. First, metaviralspades (v3.15.5) [45] was used on the filtered reads and reads were mapped back to the contigs using bwa-mem (v0.7.17) [94]. Unmapped reads were extracted with samtools (v1.19.2) [95] with the options -f 12 -F 256 and they were assembled using the basic metaspades algorithm [96] and the reads were again mapped to the contigs. The assembly and mapping results from the two steps were merged into a single bam and fasta file.

4.2.3 Removal of unwanted contigs

Due to the way phage cocktails are produced (unfiltered lysate), it was expected that bacterial reads make a significant part of the sequenced metagenome. The function easy-taxonomy [97] from the mmseqs2 software (v15-6f452) [98] was used with the nr database (version 1707752721) [53] and the assembled contigs as queries to get an estimate of the taxonomic composition of the metagenomes. The obtained taxonomic annotation of the contigs was also used to remove the unwanted contigs before further processing. All of the contigs that were classified as bacterial were removed from further processing while those classified as viral or unclassified were kept. Seqtk (v1.3-r122) (https://github.com/lh3/seqtk) was used for subsetting the fasta files.

4.2.4 Mapping to orthologous groups

Prodigal (v2.6.3) [99] in the metagenomic mode was used to predict genes on the contigs that passed the filtering and the predicted genes were mapped to the hidden markov models (HMMs) of vFAMs from the VOGDB version 221 using hmmscan from hmmer (v3.4) [100]. Proteins were assigned to the best scoring vFAM if the e-value of the alignment was lower than 0.001. The relative abundance of the genes was estimated from the number of reads mapping to the genes.

Mappings to vFAMs was used to get the taxonomic annotation of the genes and contigs on which they were located using the information of the lowest common ancestor provided with every vFAM. When genes were located on the same contig, taxonomic annotation of the genes with a more detailed annotation was given to the genes with less detailed annotation if no contradiction would occur as described in the first chapter.

4.2.5 Comparison of phage cocktails

Phage cocktails were compared on two levels. One is the taxonomic composition in terms of the content of phage families and the other is the overlap in the vFAM content. Overlap of the vFAM content is defined as the proportion of genes that are mapped to the same vFAM taking into account the relative abundance.

4.3 Results

4.3.1 Sequencing data overview

Cocktails were sequenced to a different depth and from each cocktail I was able to retrieve different number of contigs. The summary is provided in the table 4.2 . The proportion of reads coming from the bacterial genomes ranges from 26% in the IntiG cocktail to 86% in the CP15 with most of the cocktails (11) having more than 50% of bacterial reads.

4.3.2 Taxonomic composition of the phage cocktails

Taxonomic composition in terms of phage families was determined for the phage cocktails to compare the diversity between and within cocktails. Plot 4.1 shows the comparison of relative abundance of phage families between different cocktails. In terms of number of phage families, Smono cocktails shows the least amount of diversity containing only one family (Herelleviridae) while the cocktail

Table 4.2: Sequencing depth and the number of contigs. Cocktails were sequenced on various ocassions and have different sequencing depth and the number of retrieved viral genes.

Cocktail	Nr seqs	Nr contigs	Nr viral contigs	Nr viral genes
CP12	50720347	18547	2018	4347
CP15	64361748	24564	2431	5270
ENKO	36110187	5792	2116	8257
Fers	44540894	9802	1922	3476
IntiG	42847550	3031	1842	5902
PG	72556813	9225	2051	6890
PR15	54603075	22825	4031	9003
Pyo14	32399296	19384	3950	8200
SES	33761458	6449	2105	5918
Smono	36460138	10079	1396	1445
Sxta	29972690	20866	2304	6423
VRPYO1997	4885499	3127	958	4638
VRPYO2014	16429958	3708	1052	4369

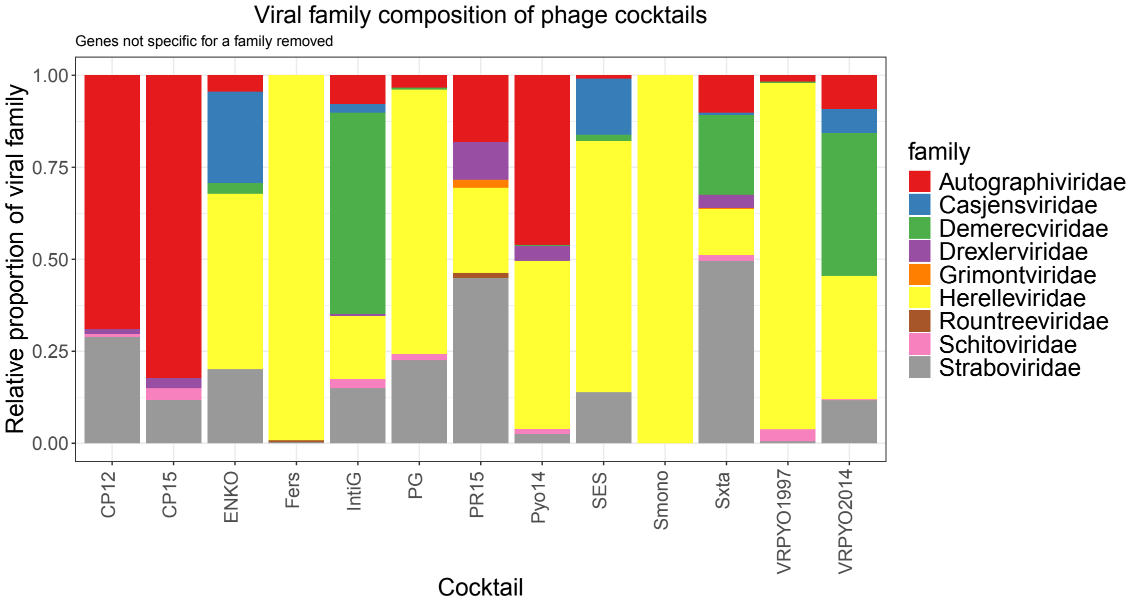


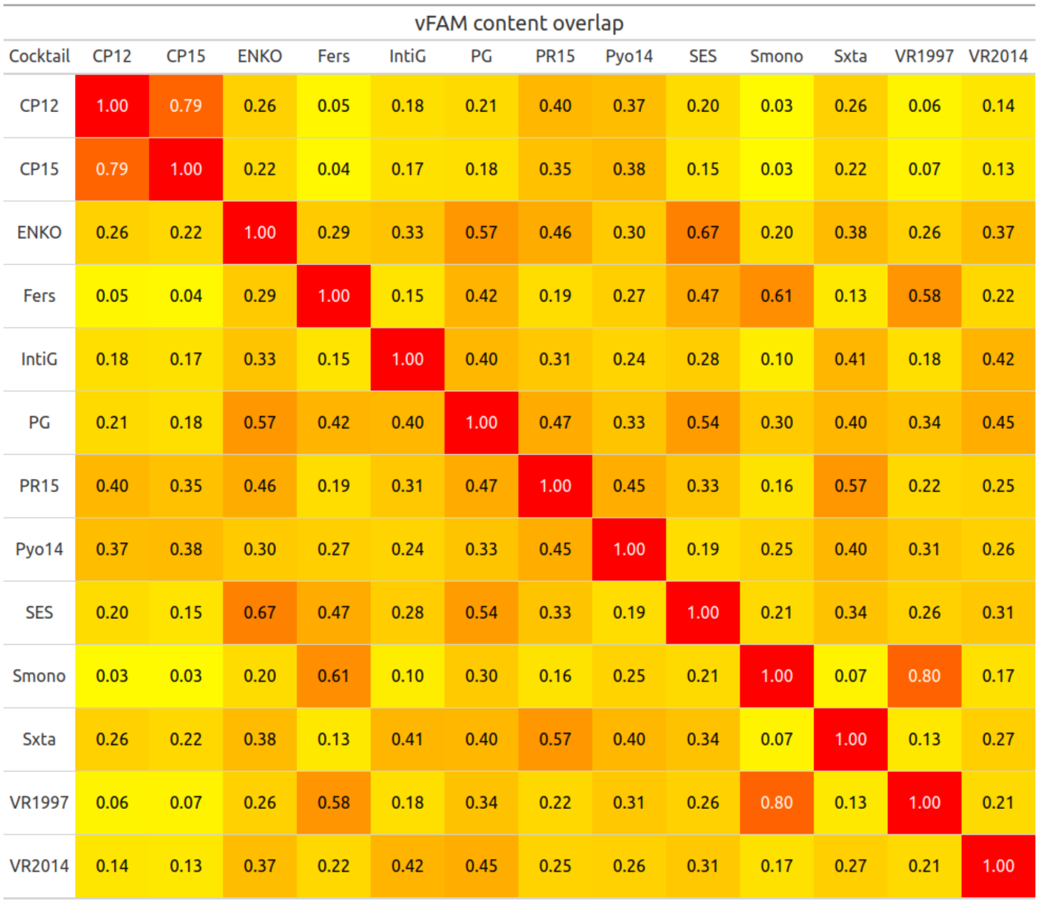
Figure 4.1: Taxonomic composition of the phage cocktails in terms of phage families. Genes for which the taxonomic annotation could not have been determined to the level of family were not included. Cocktails CP12 and CP15 are similar in the taxonomic composition with some differences in the relative abundance of the present families which is expected given they are the same cocktail, but from a different time point. Smono contains only one phage and therefore is it expected to see only one family. It is interesting that cocktails VRPYO1997 and VRPYO2014 contain different families in different relative abundances as they were both sold under the same name, but were created 17 years apart.

IntiG contains representatives from 7 different families out of the total of 9 families present in all cocktails.

Except for the difference in the number of families, different cocktails show different relative abundance of families they contain. Some cocktails (Fers, Smono, VRPYO1999) show strong domination of a single family while in the other cocktails (VRPYO2014, PR15) the distribution is more even.

4.3.3 Comparison in terms of the vFAM content

Another way to compare the content of phage cocktails is to directly compare the vFAM content of the cocktails. A pairwise comparison of the vFAM content overlap of all cocktails is shown in figure 4.2 . According to the heatmap, cocktails showing the most similarity are VRPYO1999 - Smono and CP12 - CP15. The similarity of CP12 and CP15 cocktails is expected as they are the same cocktail from the same producer produced three years apart. Cocktails showing the least similarity are CP12, CP15 - Smono and Sxta - Smono.



vFAM content overlap

Cocktail	CP12	CP15	ENKO	Fers	IntiG	PG	PR15	Pyo14	SES	Smono	Sxta	VR1997	VR2014
CP12	1.00	0.79	0.26	0.05	0.18	0.21	0.40	0.37	0.20	0.03	0.26	0.06	0.14
CP15	0.79	1.00	0.22	0.04	0.17	0.18	0.35	0.38	0.15	0.03	0.22	0.07	0.13
ENKO	0.26	0.22	1.00	0.29	0.33	0.57	0.46	0.30	0.67	0.20	0.38	0.26	0.37
Fers	0.05	0.04	0.29	1.00	0.15	0.42	0.19	0.27	0.47	0.61	0.13	0.58	0.22
IntiG	0.18	0.17	0.33	0.15	1.00	0.40	0.31	0.24	0.28	0.10	0.41	0.18	0.42
PG	0.21	0.18	0.57	0.42	0.40	1.00	0.47	0.33	0.54	0.30	0.40	0.34	0.45
PR15	0.40	0.35	0.46	0.19	0.31	0.47	1.00	0.45	0.33	0.16	0.57	0.22	0.25
Pyo14	0.37	0.38	0.30	0.27	0.24	0.33	0.45	1.00	0.19	0.25	0.40	0.31	0.26
SES	0.20	0.15	0.67	0.47	0.28	0.54	0.33	0.19	1.00	0.21	0.34	0.26	0.31
Smono	0.03	0.03	0.20	0.61	0.10	0.30	0.16	0.25	0.21	1.00	0.07	0.80	0.17
Sxta	0.26	0.22	0.38	0.13	0.41	0.40	0.57	0.40	0.34	0.07	1.00	0.13	0.27
VR1997	0.06	0.07	0.26	0.58	0.18	0.34	0.22	0.31	0.26	0.80	0.13	1.00	0.21
VR2014	0.14	0.13	0.37	0.22	0.42	0.45	0.25	0.26	0.31	0.17	0.27	0.21	1.00

Figure 4.2: Overlap of the cocktail content in terms of vFAMs for all pairs of cocktails. Relative abundance of a vFAM in a cocktail was used to calculate the overlap according to the following formula Overlap = $\sum_{vFAM \in all\_vFAM} \min(sample1\_abundance(vFAM), sample2\_abundance(vFAM))$. Cocktails CP12 and CP15 have the overlap of 0.79 and they are the same cocktail from the same producer produced three years apart. In that light it is interesting to note the overlap of only 0.21 between VR1997 and VR2014 as they are the same cocktail from the same producer produced 17 years apart.

4.3.4 Agreement of vFAM content overlap and the overlap of family composition

Cocktails that show the most similar taxonomic composition are Smono, Fers and VRPYO1999 as they are strongly dominated by the phages of the family Herelleviridae. This is also reflected in the overlap of vFAM content as they show the most overlap to each other (figure 4.2). They also show the least overlap with the cocktails CP12 and CP15 which don't contain phages from the Herelleviridae family. Other cocktails that show similar taxonomic composition and high vFAM content overlap are SES - ENKO and VRPYO2014 - IntiG.

4.4 Discussion

The phage metagenome analysis with VOGDB is fast and requires little user intervention, but this comes with a cost of a reduced resolution and the ability to only see a high level overview. The composition of some of the cocktails used in this study has already been described in the literature and it is interesting to compare the results obtained by using VOGDB with what has already been described.

4.4.1 VRPYO1997 and VRPYO2014

The composition of the cocktails VRPYO1997 and VRPYO2014 has been described in the paper from Villaroel et al [87]. Even though the authors use different methodology based on BLAST [101] searches and the families from the outdated taxonomy (myoviridae, siphoviridae, podoviridae) [20], we reach a similar outcome regarding the taxonomic composition of the cocktails where VRPYO1997 is dominated by one family and in VRPYO2014 more families have more equal relative abundance. Direct comparison in terms of taxonomy is difficult to make since the viral taxonomy is an evolving field and the family Herelleviridae which we find to be dominant in the VRPYO1997 did not exist as family in the viral taxonomy in 2017 when authors published their analysis. [102].

4.4.2 PR15 and PG pyophage and Smono

The metegenome analysis of the cocktails PG, PR15 and Smono has been published by McCallin et al [88]. Again, it is difficult to directly compare taxonomic composition determined by the authors and by this study due to the changes in the virus taxonomy. However, the PR15 cocktail contains phages against one pathogen that is not targeted in the PG cocktail (Klebsiella) which could be reflected in the bigger richness of phage families in the PR15 cocktail compared to PG. Authors report the overlap of proteins in the two cocktails to be around 20%, while we report the overlap of vFAM content of 45%. Higher similarity in terms of orthologous groups compared to the proteins is expected as related proteins would map to the same HMM representing an orthologous group while they would be recognized as separate proteins by clustering based on direct sequence comparison.

4.4.3 ColiProteus (CP12, CP15) cocktails

Metagenomic analysis of a ColiProteus cocktail from Russia has been published by McCallin et al [86], however, the metagenomes from this study (CP12 and CP15) are not the ones describes by the authors, but CP12 and CP15 are cocktails of the same name and from the same producer. Due to the old taxonomy used by the authors [86] we can't directly compare the taxonomic composition of the cocktails, but we can detect a level of disagreement between what we find and what has been previously described. Authors [86] report domination by what was formerly known as myoviridae and around 25% of what was formerly known as podoviridae. We however see domination by Autographviridae which was previouslly placed in the podoviridae family.

4.4.4 Benefits of using VOGDB for the analysis of phage cocktails

Using VOGDB for the analysis of metagenomes dominated by viruses offers benefits compared to the methodology described in previous publications. First, mapping of genes to vFAMs allows for more sensitive homology detection compared to direct sequence mappings. Even though the diversity of phages in the RefSeq database [103] (number of phage genomes in the release 221) is small compared to the global diversity of phages which is measured in millions of species [92], sequence profiles represent common patterns of the gene family which could be detected in the genomes not present in the database. Second, VOGDB is a continuously updated resource which makes it easy to rerun the analysis when new version is available which incorporates important changes like the updated phage taxonomy. Third, in case of a fragmented assembly, the annotation of the vFAM lca allows for easy taxonomic annotation of genes or gene fragments.

4.4.5 Shortcommings of using VOGDB for cocktail analysis

Using VOGDB for the analysis of viral metagenomes brings some challenges as well. VOGDB is calculated from the viral portion of RefSeq and the orthologous group are calculated from a tiny subset of viruses present in the biosphere. Therefore, genes from a viral genome drastically different from what is present in RefSeq will not get mapped to the orthologous groups. Other issue is with potential over-annotation of taxonomy as lca of vFAMs are not fixed. In future releases, when more phage genomes are added to RefSeq it can happen that lca of some vFAMs becomes less specific (if lca was at the family level it could change to a class). The described shortcomings of using VOGDB come from the database representation of the viruses and the same problems arise whenever database information is used for analysis.

Chapter 5

Viruses of Great Apes

5.1 Introduction

Dental calculus is a calcified material forming around the teeth. As it forms, it entraps host cells, bacteria, viruses and food particles [104]. Being mostly inorganic matter, dental calculus does not decompose as the organism dies which makes it possible to extract the DNA from the ancient samples of dental calculus. DNA from the ancient samples of dental calculus has been used to explore the ancient microbiomes of various animals [105], including the great apes, the closest human relative. While studies have made high resolution analysis of the microbial DNA from calculus samples, there are no studies that focus on the viruses that infected great apes in the past and whose remnants might be found in the calculus.

Analysis of the ancient DNA is challenging due to the fragmentation of the molecule and the specific age-dependent degradation of the sequence. State of the art for the analysis of the degraded samples is mapping of the sequencing reads to the reference genomes [106]. The challenges lie in the need for the adequate reference genomes which might not always be present in the databases and in the difficulty to automate the analysis procedure. Another option is to enrich the samples for the genomes of interest before the sequencing, but it is still necessary to know in advance what are the target genomes in order design the capture probes.

Analysis of the ancient viruses from the dental calculus comes with additional challenges compared to the analysis of the prokaryotes. Excluding bacteriophages, viruses are not common residents in a healthy mouth and are present only during an infection [107]. A small portion of the viruses in an infected individual would end up incorporated into the dental calculus. Compared to the abundance of the bacteria and bacteriophages, eukaryotic viruses form only a miniscule portion of the total DNA in the calculus and consequently only a minority of reads in the calculus metagenome comes from eukaryotic viruses. Moreover, sequence databases for viral genomes are not as developed as for bacterial genomes and it might be difficult to find the right reference genome [108].

In this study I aimed to develop an analysis procedure for detecting great ape viruses in the ancient dental calculus samples in an automatic way. The viruses of interest are scarce and the sample metagenomes are complex which caused the nucleotide assembly to leave many reads unassembled and almost no reconstructed contigs belonged to the eukaryotic viruses. As there is no standard accepted assembly based way to analyze such samples, I hypothesized that more proteins could be retrieved by a protein oriented assembly [109] on which two step mapping to the nr database [53] and the hidden markov models of virus orthologous groups from VOGDB could be applied to remove the non-viral and phage genes and identify eukaryotic viruses. However, the fragmented assembly and rare viral genes would not allow for a detailed overview of the past infections, but could be used to estimate the taxonomic composition of the eukaryotic viruses in the dental calculus on the family level. I was able to get such an estimate without the prior selection of reference genomes with a fully automatic pipeline that does not require user intervention.

5.2 Methods

The detection pipeline has been implemented in snakemake [110] to make it easy to extend the pipeline and to ensure reproducibility.

5.2.1 Input data and material

The analysis was performed on 40 samples out of which in 4 (L1946, L1947, L1948amp10, L1949amp10) the presence of Monkeypox virus has been proven [111]. The sequencing data was created by sequencing the metagenome of the dental calculus from the museum samples of the great apes (around 100 years old). Details of the sample collection and laboratory processing are not known to the author (except for the 4 published samples).

The data is generated by sequencing the extracted DNA with the Illumina sequencer in a paired end fashion for 35 samples and single end for 5 samples (210269, L1946, L1947, L1948amp19, L1949amp10). Along the dental calculus samples, swabs, extraction blank and library blanks were sequenced as well. Swabs were taken from the museum shelves and the ape skulls and represent contamination.

All of the sequencing data were provided by Martin Kuhlwilm from the Computational Admix Lab at the University of Vienna.

5.2.2 Preprocessing of the sequencing reads

The raw sequencing reads had to be preprocessed in order to detect sequences of viral origin. As a first step, I performed trimming with trimmomatic (v0.39) [93] to remove the sequencing adapters and removed all of the reads with length smaller than 50bp.

The preprocessed reads were assembled using the software Plass [109]. Plass differs from other assemblers as it uses six frame translations of the sequencing reads to reconstruct proteins from a metagenome. It was shown [109] Plass can recover up to 10 times more proteins from a complex metagenome and given the nature of our samples (complex metagenome and low abundance of the target sequences) and the subsequent focus on protein sequences, Plass was selected as the most appropriate tool for the job. Plass was run with the default settings except for the parameter –min-length which was set to 20.

5.2.3 Filtering of reconstructed proteins

The set of proteins obtained by Plass had to be filtered to remove sequences that did not come from eukaryotic viruses. For filtering, I used mmseqs easy-taxonomy [97, 98] with the NCBI nr database (24.10.2022) [53] and DIAMOND (v.2.1.9) [112] with the nr database. First, the taxonomy of the proteins was determined by easy-taxonomy and proteins being classified as coming from a virus as well as the proteins that were unclassified were additionally processed by DIAMOND blastp. Proteins that showed hits (best hits) to the nonviral proteins were filtered out. Unclassified proteins and those without hits to the nr database were mapped to the vFAMs (VOGDB version 221) using easy-search from mmseqs2 with the e-value cutoff 10^{-3}. Seqtk was used for subsetting fasta files (https://github.com/lh3/seqtk).

5.2.4 Software

The processing pipeline was implemented in snakemake with bash and python (https://www.python.org/). Analysis and plots were created in R v(4.3.2) [113] using packages ggplot2 [114], reshape2 (v1.4.4) [115], dplyr (v1.1.4) [116], stringr (v1.5.1) [117] and gtools (v3.9.5) [118].

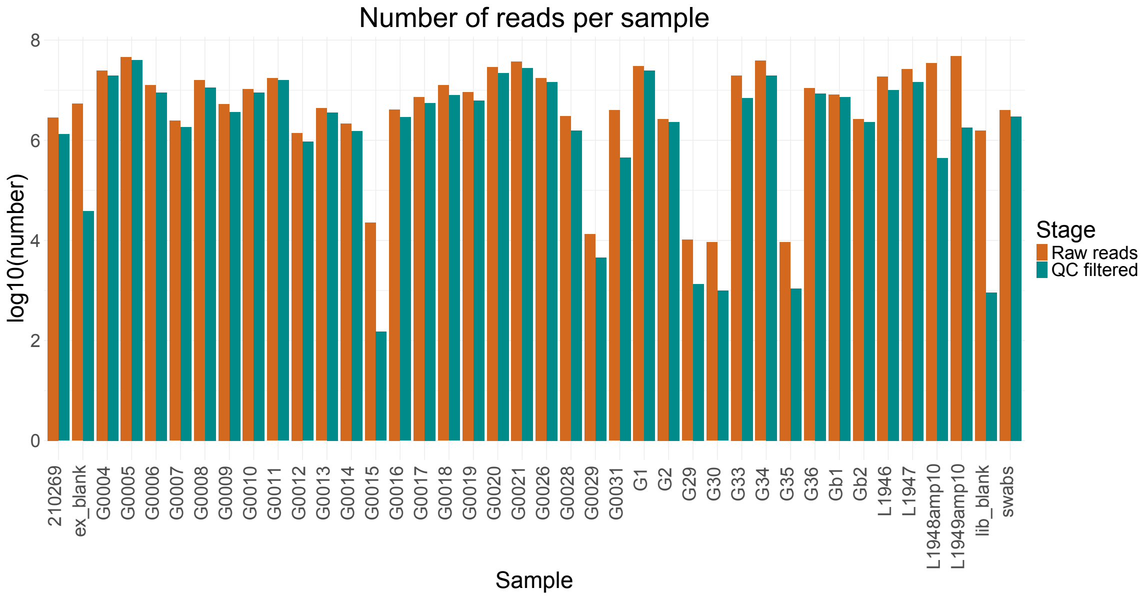


Figure 5.1: Number of reads per sample on a logarithmic scale. The brown column represents the raw number of reads while the green column represents the number of reads without the sequencing adapters and longer than 50bp after processing with trimmomatic. The sequencing depth differs greatly among samples with almost 4 orders of magnitude difference between the sample with the least number of filtered reads (G0015) and the sample with the most number of filtered reads (G0005).

5.3 Results

I have created a computational pipeline for the detection of eukaryotic viral sequences in the short read data of the semi-ancient museum samples. Due to the nature of the samples, the sequencing depth is unequal and the amount of reads that pass the quality filtering differs between samples. The comparison of the number of the raw reads and the number of reads that pass the filtering is shown in the figure 5.1 . The amount of raw reads varies more than thousand-fold between samples and in some of the samples (especially G0015) the filtering removed around 90% of the reads.

Detected proteins

As Plass was used for the assembly, the output are protein sequences. Again, different number of proteins was retrieved from different samples. Proteins from viruses in families that infect mammals are scarce with most of the samples having no proteins or only one protein that got classified as belonging to a protein from the families of interest (Adenoviridae, Circoviridae, Hepadnaviridae, Orthoherpesviridae, Papillomaviridae, Parvoviridae, Poxviridae, Retroviridae). The number of proteins retrieved by plass and the number of mammalian virus proteins are shown in the figure 5.2 . Out of 37 samples (excluding controls experimental blank, library blank and contamination swabs), 18 samples contain one or more of the mammalian viral proteins with the sample L1946 having the most proteins (74 proteins)

Viral families

A few samples have more than a few proteins from the desired families (G1, G34, G36, L1946, L1947, L1949amp10) and the families that dominate are Retroviridae, Poxviridae and Papillomaviridae. The relative proportion of mammalian viral proteins between samples and the relative proportion of the viral families within samples is visualized in the figure 5.3 .

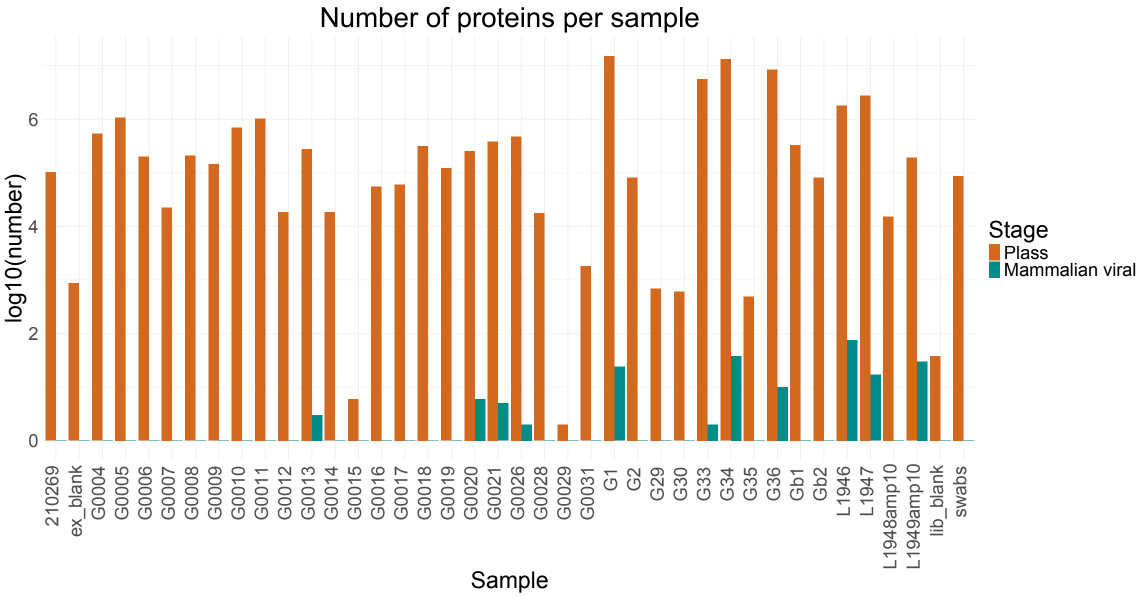


Figure 5.2: Number of genes per sample at different stages of processing on a logarithmic scale. The brown column represents the number of proteins or protein fragments retrieved by the assembly with Plass and the green column is the subset that was determined as coming from a mammalian virus. Many samples don't contain any mammalian viral proteins and in those that do contain them, they represent only a minor fraction of all proteins or protein fragments (4 orders of magnitude difference in the sample with the most mammalian viral proteins - L1946)

Retroviridae In the whole dataset, 153 proteins are classified as belonging to the family Retroviridae, however, at least 79 of those are from endogenous retroviruses. More interesting are viruses of the order Lentivirus as they include human and simian immunodeficiency viruses. In the dataset we found 14 proteins that seem to come from a lentivirus. 9 in the sample G1, 1 in the sample G34 and 4 in the sample L1947 and they all were classified by mmseqs easy-taxonomy. 8 out of 14 proteins are taxonomically classified as HIV1.

Poxviridae In some of the samples (L1946, L1947, L1948amp10 and L1949amp10) the presence of monkeypox virus has already been proven [111] and in this work we also detect proteins coming from poxviruses. Monkeypox virus (NC_003310.1) has 180 proteins, but we are only able to detect part of it. In the sample L1949amp10, we detected 21 proteins from the Orthopoxvirus genus and out of those 12 were classified as coming from Monkeypox virus. In the sample L1946, 5 proteins coming from a genus Orthopoxvirus have been detected and two are classified as coming from the Monkeypox virus.

Papillomaviridae Among all the samples, 43 proteins were classified as coming from Papillomaviruses, the most coming from the sample G34 (19 proteins). All together, 39 proteins are classified as coming from human papillomavirus.

5.4 Discussion

5.4.1 Dental calculus samples

The DNA for the study comes from the dental calculus of the museum samples of the great apes of around 100 years old. The skulls that contain the teeth from which the dental calculus was scraped come from various locations and it is not known how the skulls were treated before they reached the museum. As they were sitting in the museum for 100 years it is reasonable to expect that human

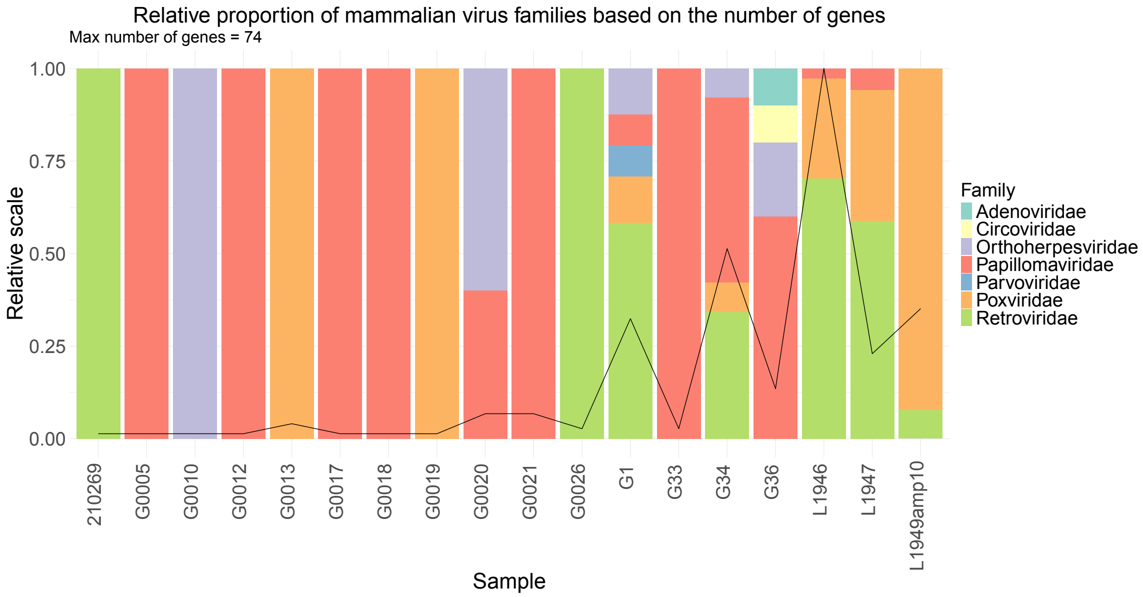


Figure 5.3: Relative proportion of mammalian virus families and relative amount of mammalian virus reads per sample. The line represents the relative amount of reads belonging to the mammalian virus families with the sample with the most reads (G34) having 644180 reads is scaled to 1. The more reads there are, the more confidence there is that the detected proportion of families resembles the real proportion of families. In almost all samples, Herpesviridae are the dominant family, followed by Poxviridae and Adenoviridae on the third place.

viruses might have also reached the skulls and since humans and great apes are close relatives it is not always possible to tell simian and human viruses apart from gene fragments [119]. A way to control for contamination is to sample the surrounding museum shelves and the skulls and consider the viruses found there as the environmental contamination when found in the dental calculus. We didn't find any viral genes in the skull and shelve swabs which increases the confidence that the detected genes originate from the time before museum samples reached the museum.

Even with the complete control over how the samples were handled before the DNA extraction, we would not expect to find a lot of DNA belonging to the viruses that infected great apes. The viral DNA that ends up in the calculus comes from the virions that got incorporated during the mineralization and since the calculus buildup is slow compared to the duration of the viral infection, only a very few of the virions would end up trapped and their DNA would be a minority compared to the bacteria, bacteriophage and food DNA. Taken all of the sample limitations into consideration, it is not reasonable to expect to extract detailed information of the past infections, but an overview of the higher taxonomic levels that were present in great apes.

5.4.2 Opportunities and challenges

The presented results show the limits and opportunities for detection of ancient viruses in semi-ancient samples from the illumina reads using a combination of assembly and mapping of the predicted genes to the viral proteins and viral orthologous groups. The state of the art for the analysis of ancient samples is mapping of the sequencing reads to the reference genomes. However, mapping to a reference genomes means that one must know in the advance what one is looking for. Moreover, adequate reference genomes could often not be available and the collection of reference genomes together with the analysis of the results require a lot of manual work. The analysis approach suggested in this paper removes the need for reference genomes and makes it possible to create an automatic processing and analysis pipeline.

Another common approach in ancient DNA analysis is the targeted capture of the desired fragments [111]. When analyzing novel samples it is useful to know what is expected to be found to guide the design of the probes. Reference-less approach as in this study makes it possible to quickly and easily get an estimate of the sample composition on different taxonomic levels which could be used in designing the probes.

The automatic analysis pipeline based on the mappings to the orthologous groups from VOGDB also comes with limitations. Firstly, the taxonomy determination of the gene mapped to a vFAM is based on the lowest common ancestor (LCA) of the vFAM. As many of the vFAMs contain proteins from diverse viral taxonomic group they have a LCA label high in the taxonomy and in the worst case, the LCA is just "Virus" which makes in completely uninformative. This is more of a limitation for samples where the genes of interest are found on the short contigs as in case of this study where viral DNA is rare and the assemblies of viral genomes produce contigs mostly containing only a single gene preventing us to use the taxonomic information of the vFAM to which a neighboring gene would be mapped. Secondly, there is a possibility of genes which are not of viral origin and are mapped to vFAMs. The VOGDB release 221 contains more than 4000 vFAMs which are not specific to viruses (https://vogdb.org/reports/release_stats) and homologs are found in in cellular organisms. We can't be sure that the protein mapping to the virus-unspecific vFAM comes from a virus unless we know the context of that gene. A slight improvement could be made in the filtering step where the bacterial genes are removed, but that mostly depends on the improvements of the databases with bacterial genomes. Using profile HMMs of bacterial gene families would be only of limited use since a viral gene homologous to a bacterial gene would map to both a VOG and a bacterial HMM.

As an alternative to VOGDB, the viral portion of eggNOG [56] could have been used as it provided the sequence models of orthologous groups of eukaryotic viruses. VOGDB was selected as it is based on more viral genomes since it is updated every two months and I expected to get more hits compared to the groups from eggNOG.

Another limitation comes from unreliable information in the databases. By manually BLAST-ing [101] a protein classified as coming from a HIV1 against the nr database, the list of hits contained a single partial HIV1 protein for which I couldn't find homologs in other HIV genomes. This is a strange situation given HIV1 is a much researched virus resequenced many times. However, setting custom confidence thresholds for deciding whether a hit is trustworthy or not would contradict the idea of an automatic analysis pipeline. We therefore need to accept that the output of this and any similar pipeline is not completely correct and the user would manually explore the hits if they seem unusual or to confirm the desired result.

5.4.3 Detected viral families

There are no similar studies exploring the ancient great ape viruses from the dental calculus samples. However, in a study exploring viromes of the great apes and humans from fecal samples [120], the authors detected many of the viral families discovered in this study. Those are adenoviridae, hepadnaviridae, herpesviridae, papillomaviridae, parvoviridae and poxviridae. Another study [121] exploring viruses in saliva of the sanctuary chimpanzees in Congo detected viruses of circoviridae, herpesviridae and papillomaviridae families. Comparison with other studies show that the viral families found in the museum samples are the same viral families as in the contemporary great apes. However, dental calculus samples should in theory contain information about the past infections during the whole life of an ape, while the fecal samples or saliva should only contain viruses of the acute infections. Therefore, we expected that viromes of dental calculus are more diverse.

5.5 Conclusion

This work explores a novel approach to analyzing viromes of semi-ancient samples. The main benefits of assembly and mapping to orthologous groups compared to the read mapping to the reference genomes are the opportunity to automate the processing pipeline and no need for the viral reference genomes. Due to the nature of the data (scarce viral reads), the results have limited resolution, but are nevertheless useful for getting an overview of the composition of the ancient virome which could be especially useful for analyzing novel samples where we would not have a prior expectation of the virome composition. A publication from the collaboration partners which would include the work presented in this chapter is expected in the future. A data report has been submitted to *Scientific Data* on 24.4.2024 and the preprint version is at the end of this document in the appendix B.

Chapter 6

Phage-host interactions

6.1 Introduction

Phages are the natural predators of bacteria. They base their existence on using the bacterial metabolism and resources to propagate. During the process they can destroy the host cell, but also duplicate together with the host genome or chronically release progeny [122]. Their ability to kill bacteria can be exploited by using them as therapeutics for bacterial infections, especially against antibiotic resistant bacteria [123]. As it usually happens in predator-pray interactions, bacteria have acquired various ways to protect themselves against a phage attack in form of defense system genes found in their genomes or mobile elements [124].

It has been shown that anti-phage defense systems are part of the accessory genome and are often found on mobile genetic elements [125], including prophages [126]. Presence on mobile elements makes it easy to spread in the bacterial population and change the phage susceptibility pattern. In case of the defense systems that integrate into the host chromosome, the integration is not necessarily random as there are hotspots where the mobile elements tend to preferentially integrate. It was shown for E. coli where the hotspots are and if there is a tendency for defense systems to integrate into a specific hotspot [127].

Phage therapy is an instance of application of phages in medicine to treat bacterial infections [128] and even though the approach seems promising, phages often fail to cure the infection for an unknown reason [129, 130]. As we have only recently discovered the vast repertoire of anti-phage defense systems [131, 132], we hypothesize that the defense systems are to blame for the failure of the phage treatments and we have explored the possibility to find a list of defense systems that are likely to offer protection against a set of well characterized phages. The phage-host interaction matrix on which the study was based was made with 55 E. coli strains isolated from human bladder and is an adequate dataset to explore the interactions for human pathogens that could potentially be treated with phages. Such interaction matrices where phages are well characterized are difficult to obtain and are not readily available, but offer an opportunity to get valuable insights to direct future efforts in phage therapy. Besides exploring the interaction of defense systems and the phages, we also compared the hotspots and their associated defense genes of our medically relevant collection of E. coli strains with the previously published study on the large collection of E. coli strains from the IMG/M database [133].

The motivation for the study was to develop an analysis framework that would make the selection of phages for phage therapy faster and cheaper. The main hypothesis is that by comparing the defense systems in the strains that are lysed by specific phages and those that are protected, we can find the important protective systems by looking at the difference of the two sets. An alternative approach would be to use the whole genomes of bacteria and phages together with the knowledge of the phage host range to come up with a machine learning model to predict the strains resistant to phages. However, since E. coli genome has around 5000 genes with a huge pangenome of more than 120000 genes [134], we would need much bigger interaction matrix to develop a confident model.

6.2 Materials and methods

The experimental determination of the phage host range on clinically relevant E.coli strains was done at the Balgrist University Hospital. Genomes of the E. coli strains were sequenced by the Institute of Medical Microbiology at the University of Zürich.

6.2.1 Determination of the phage host range

Bacterial strains and phages Collection of 312 clinically relevant E. coli strains was obtained from patients visiting the Balgrist University Hospital in Zürich. A subset of 55 strains was used for the phage-host range determination. Patients from which the bacteria were sampled had their urinary tract colonized by the bacteria, but in most of the cases they did not show symptoms of infection. The bacteria were cultured from the patient urine, grown on agar plates and frozen on -80°C. Phages used in the study are from a published and characterized collection of 69 E.coli phages called BASEL [135]. Phages from the BASEL collection are all double stranded DNA phages from the class Caudovirales. Genomes of the phages as well as receptors they use for binding to the bacterial surface and inject their DNA are provided with the publication. For experiments, phages were propagated on the E. coli K-12 strain.

Host range testing Host range of phages on the E. coli collection was determined by spot assays on the double agar. When plaque was visible after incubation at 37°C, the phage was labeled as being able to lyse the bacterial strain. Host range testing was done for all 69 phages on a subset of 55 strains.

6.2.2 Determining traits responsible for protection against a phage

Preprocessing of bacterial sequencing data Raw fastq files of the bacterial genomes were processed with trimmomatic [93] to remove the sequencing adapters and reads that are shorter than 50 base pairs. Assembly into scaffolds was done with spades (v3.15.5) [136] using the option –careful and genes were predicted by prokka (v1.14.16) [137]. Completeness and contamination of the assemblies was estimated using CheckM2 (v1.0.2) [47].Prophage sequences on bacterial assemblies were predicted by the PHASTER web server [138].

Defense systems Bacterial anti-phage defense systems were predicted using defense-finder (v1.2.0) [139, 140]. We aimed to describe the localization of the defense systems in terms of the integration hotspots as described in [127]. Integration hotspots are described by the two genes that border the hotspot. We retrieved the amino acid sequences of the bordering genes from [127] and used BLAST [101] to find the genes in the assemblies of the clinical strains. When a defense system is surrounded by the genes defining a hotspot, it gets the label of the hotspot affiliation. Hotspots are named by numbers from 1 to 41, the same way as they are named in the publication where they were described [127]. It is not possible to reliably determine the hotspot for all of the defense systems due to the fragmented nature of the assemblies. We were only confident to the hotspot assignment when both of the bordering genes defining a hotspot were found on the same contig. For all of the other cases, it was impossible to precisely tell what exactly belongs to the hotspot. To explore how often does the same defense system appear in the same hotspot, we used the Shannon information entropy as a measure of a tendency of a defense system to consistently appear in the same hotspot across strains.

Defense systems variability Genes within a single defense system subtype are not necessary the same in different strains. To estimate the diversity of genes forming a single defense system subtype we used Roary [141] to find the pangenome of all of the E. coli strains and then counted how many groups of genes are found in a single system subtype across all of the strains available.

Receptor compatibility As we know what surface receptors phages from the collection use to attach to the bacterial cell (BtuB, FhuA, LamB, LptD, NfrA, TolC and YncD), we extracted the appropriate receptor sequence from a list of predicted genes. To determine whether the different strains from the collection have the surface receptors compatible with the phage receptors we aligned the the receptor sequences using mafft (v7.520, –auto) [142] and in a group of receptors with identical sequence one strain is lysed by a phage in question, the receptor is labeled as compatible with the phage. When we don't see lysis with a phage in a group of receptors, we don't have enough information tell whether the receptors are compatible or not. Illustration is shown in the figure 6.1

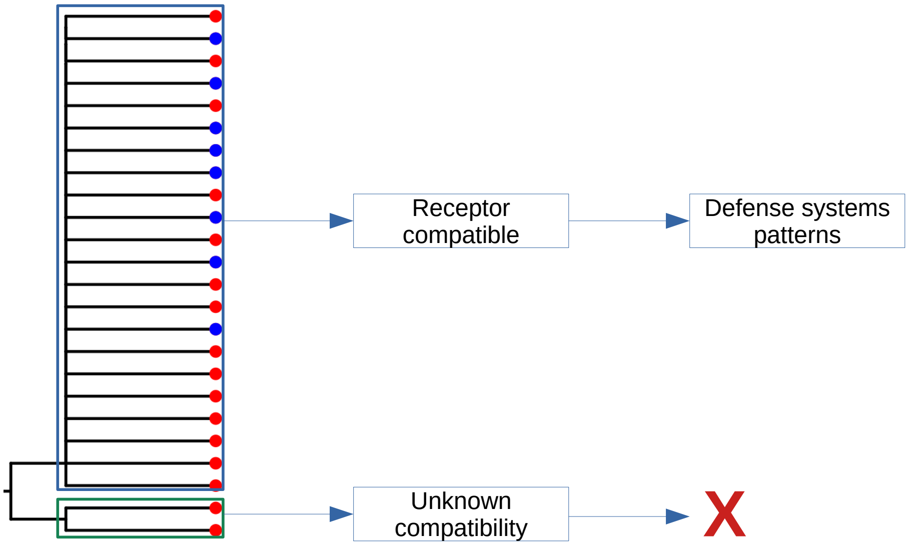


Figure 6.1: Bacterial surface receptors compatible with the phage terminal receptors. The tree represents amino acid sequences of different bacterial surface receptors. Red leaf means that the bacterium did not get lysed by a phage using the shown receptor as the terminal receptor and the blue leaf means that the bacterium did get lysed. When the receptor sequences are identical and there are strains that are lysed by phages, as in case of the receptors inside the blue box, we can conclude that the bacterial and phage receptors are compatible and the reasons for the differential lysis pattern lysed in the difference in the defense system content. When none of the bacteria carrying a specific variant of a receptor is lysed as in the case of the receptors in the green box, it is not known whether phage and bacteria receptors are incompatible or all of the strains get defended from the lysis by the defense systems.

.

Responsible phage defense candidates We suggest a simplified model for a successful phage infection as a two-step process. The first step is the successful attachment of a phage to the bacterial receptor and injection of the phage DNA. For the step to succeed it is essential that the phage terminal receptor and the bacterial surface receptor are compatible [143]. The second step is phage DNA survival of the bacterial anti-phage defense systems. We searched for the defense systems responsible for defending bacteria from phages in our collection. Per phage, we found, using the phage-bacteria interaction matrix and the method described above, all of the strains having the compatible surface receptors. We divide the strains with compatible receptors into two groups, lysed and nonlysed and look at the defense systems in those two groups. Defense systems that are exclusively found in the group that is not lysed are labeled as the candidates responsible for the bacteria surviving exposure to the particular phage. Some phages don't have a terminal receptor determined and those were not used in the analysis.

Comparison of the strain collection with a reference collection To determine the degree of similarity between our bacterial collection with the collection from which the integration hotspots were defined [127], we created a phylogenetic tree from the genomes of our strain collection and the E. coli genomes from IMG/M [133] downloaded on 13.12.2023. We used anvi'o (v8) [144] to extract the amino acid sequences of 71 single copy genes from the genomes (set Bacteria_71) that were subsequently aligned by muscle (part of anvi'o) [145] and a tree was calculated using FastTree (v2.1.11) [146]. The phylogenetic tree was visualized using R (v4.3.2) [113] with packages ggplot2 [114] and ggtree [147].

6.3 Results

6.3.1 Data

The completeness as estimated by CheckM2 is more than 97% for 280 strains with the rest (32 strains) having completeness between 6.66% and 82%. For the analysis of the defense systems contents, only the 280 strains with completeness larger than 97% were used. Out of the 55 strains that were used in spot assays, 52 have the estimated completeness of 100% and three have completeness between 38% and 62%. Out of the 280 strains with almost complete genomes, 19 don't have identified intact prophages 44 strains have 1, 94 strains have 2 and 123 have 3 or more predicted intact prophages. Comparison of the relatedness of the strains from our collection with the strains from IMG/M that were used in [127] with which we compare the defense system content is shown on the phylogenetic tree in the figure 6.2 . The strains from our collection mostly come from a subset of the clades and that partially explains the difference in the defense system content and the localization of those systems in the hotspots.

6.3.2 Defense systems

Even though all of the strains used in this study come from the same hospital and are isolated from the same habitat (human urine), they differ in the amount of detected defense systems in their genomes. The distribution of the amount of defense systems per genome is shown in the figure 6.3 . Within the collection in some strains we were not able to detect any defense systems while on the other side some of the strains had as much as 17 or 15 different defense systems. In total, 1792 defense systems have been observed in the 280 E. coli strains with the majority of 920 belonging to the Restriction-Modification class of all types (I, II, III and IV) out of 57 distinct classes as defined by the defense-finder.

Localization of defense systems Defense systems are often present on mobile genetic elements and don't appear in a random location within a genome, but have a tendency to integrate into so-called hotspots [125]. We compared the hotspot occupancy with defense systems in our strain collection with the E coli strains from IMG/M as done by [127]. The hotspots that were confidently detected with the frequency of occupancy is shown in the figure 6.4 . The frequency of occupancy of some of the hotspots is similar to the published work. For example hotspots 7, 36 and 37 were previously found to carry a defense system most often. Our data reproduces the finding they are the hotspot often carrying a defense system, but there are also hotspots 17 and 28 which often carry a defense system while in the previous work they seldom have one.

Preference for the integration site Some of the defense systems could show a preference for integration in a certain hotspot. To explore the tendency for a defense system to preferably integrate into a hotspot, we used the shannon entropy with the probability of a defense system to be

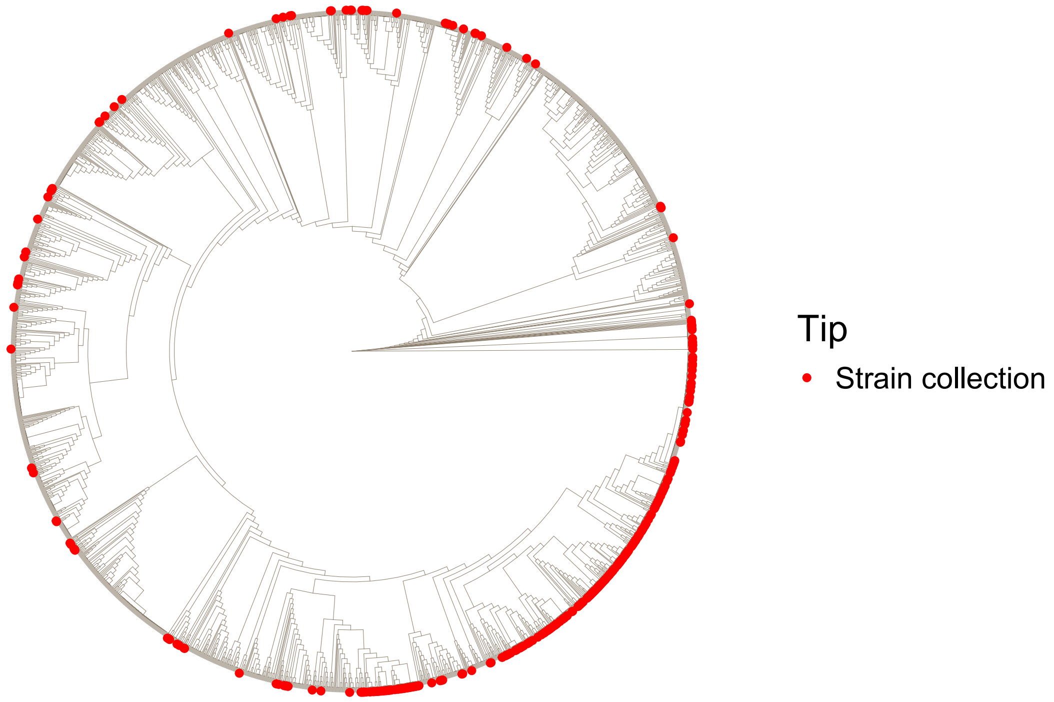


Figure 6.2: Relatedness of strains from our collection and all of the E. coli strains from the IMG/M database. The strains from our collection (marked with the red point) are mostly clustered on the one side of the phylogenetic tree, but many of the clades have a member from our collection of strains.

found in a certain hotspot as presented in the figure 6.5 . Shannon entropy index is a more reliable measure of the integration preference the more times a defense system is observed. Therefore, the systems that seem to have a strong preference to be integrated into a specific hotspot are those that have minimal entropy and are present more than a few times. Those systems are AVAST_IV, CAS_Class1_Subtype-I-E, CAS_Class1_Subtype-I-F and Thoeris_II. On the other side, some systems look like they tend to integrate at a random location, judging by their high value for entropy. These are Restriction modification systems of all classes. Not all defense systems of the same subtype are necessarily the same. They could be composed of different number of components and the genes performing a similar role in the system could be different. We explored the number of different protein families that are found within the above-mentioned systems and compared the number with the number of components. The results are summarized in the table 6.1 . The results show that the subsystems Avast_IV and Thoeris_II are always composed of the same genes in our sample collection and the have a strong preference to integrate into a specific genomic position. Cas_Class1_Subtype-I-F also has a strong preference, but the subsystem is not always composed of the same genes in different strains while Cas_Class1_Subtype-I-E also has a preference for the integration site, but the set of genes forming the subsystem is more diverse between the analyzed

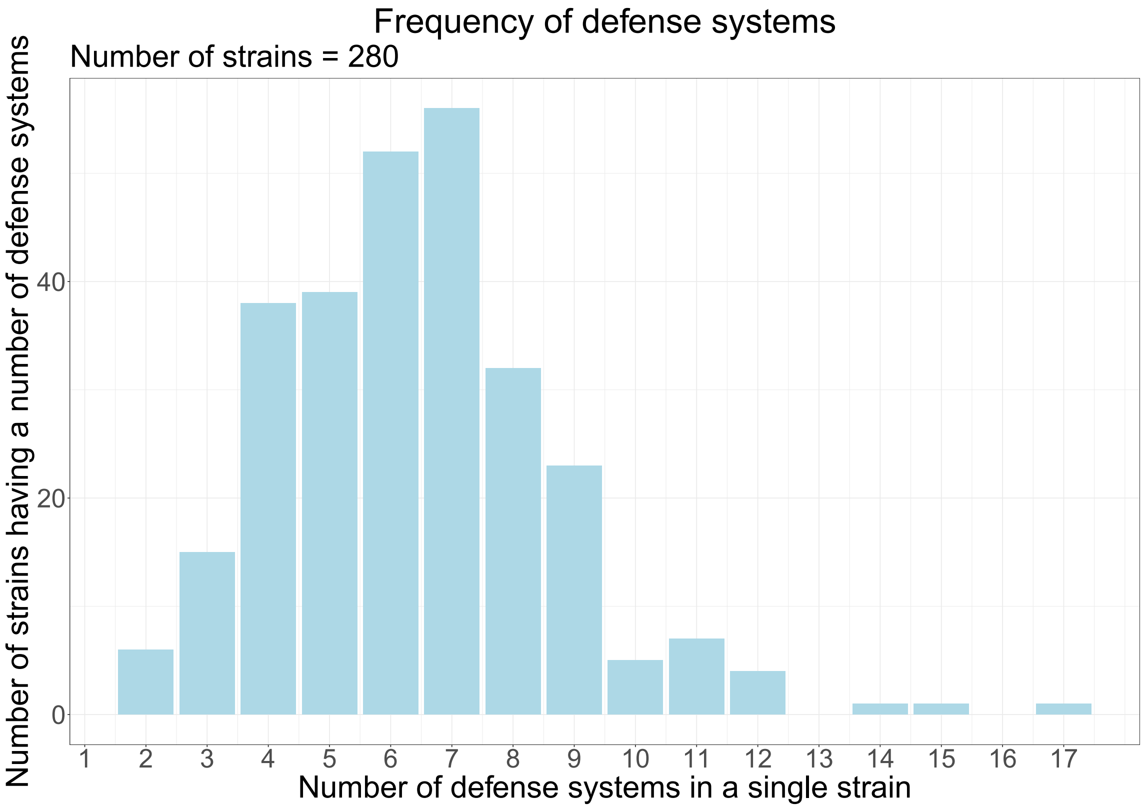


Figure 6.3: Number of defense systems in the E. coli strains from the collection. Most commonly a strain from our collection would have 7 defense systems encoded in its genome. The largest number of detected systems in a strain is 17.

Table 6.1: Number of components and distinct genes defense system subtypes. Number of components depicts the range of the number of components for a specific defense system subtype as found in the dataset. The number of distinct genes is the number of distinct gene groups into which all of the genes belonging to a subtype from all of the analyzed strains cluster.

Subtype	**Number of components**	**Number of distinct genes**
Avast_IV	1	1
Cas-I-E	7-9	27
Cas-I-F	6-7	8
RM_class 1	3-5	38
RM_class 2	2-3	14
RM_class 3	2	8
RM_class 4	1-3	17
Thoeris_II	3	3

strains. Defense systems of the Restriction-Modification class don't show a strong preference to integrate into a specific position and are composed of diverse genes in different strains.

6.3.3 Defense systems responsible for protection

Phages from the BASEL collection [135] use 7 different bacterial surface proteins as their terminal receptors. The bacterial receptors are not identical in sequence and some show more diversity than the others as shown in 6.2 . The YncD receptor shows the highest diversity with 46 unique sequences

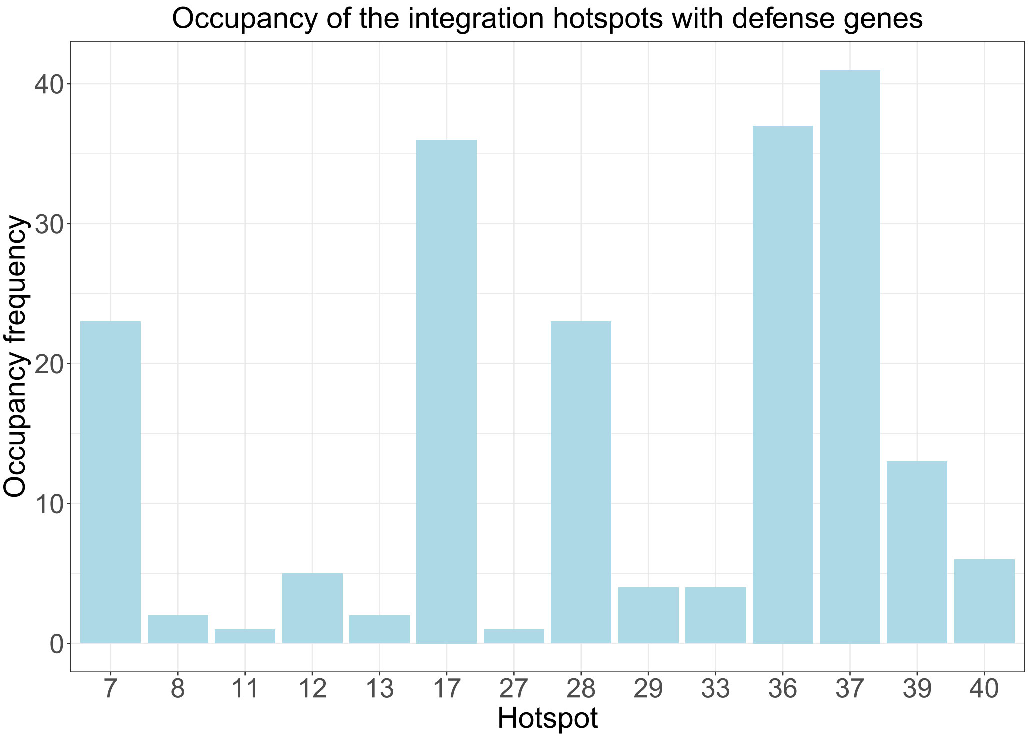


Figure 6.4: Occupancy of the confidently detected hotspots. The hotspots that contain defense systems most often are 7, 17, 28, 36 and 37. We were able to confidently detect only 14 hotspots while there are in total 41. Due to the fragmented assembly of the illumina reads, for many hotspots the bordering genes are detected on two contigs which makes it impossible to precisely tell the content of the hotspot. Those hotspots are not shown here which results in only a subset of hotspots being presented.

Table 6.2: Number of unique amino acid sequences of the terminal receptors used by the phages from the BASEL collection.

Receptor	**Unique sequences**
BtuB	44
FhuA	30
LptD	18
TolC	6
YncD	46
LamB	14
NfrA	36

among the 55 strains for which the phage host range was determined while there are only 6 unique sequences for TolC.

The list of defense systems potentially responsible for the protection against the lysis by a phage is created for 34 phages for which the terminal receptor is known. Depending on what systems are present in the strains with the compatible receptors, for some phages the list of systems potentially preventing the successful lysis is shorter than for the others. Phage Bas08 does not lyse any of the tested strains and phage Bas16 lyses only one strain and for those two phages, no list of defense systems potentially responsible for the protection could be determined which leaves us with 32 phages for which we could determine defense systems to which they are sensitive. Phages with the shortest lists of defense systems protecting against them are Bas05 ad Bas07 with 14 defense system subtypes

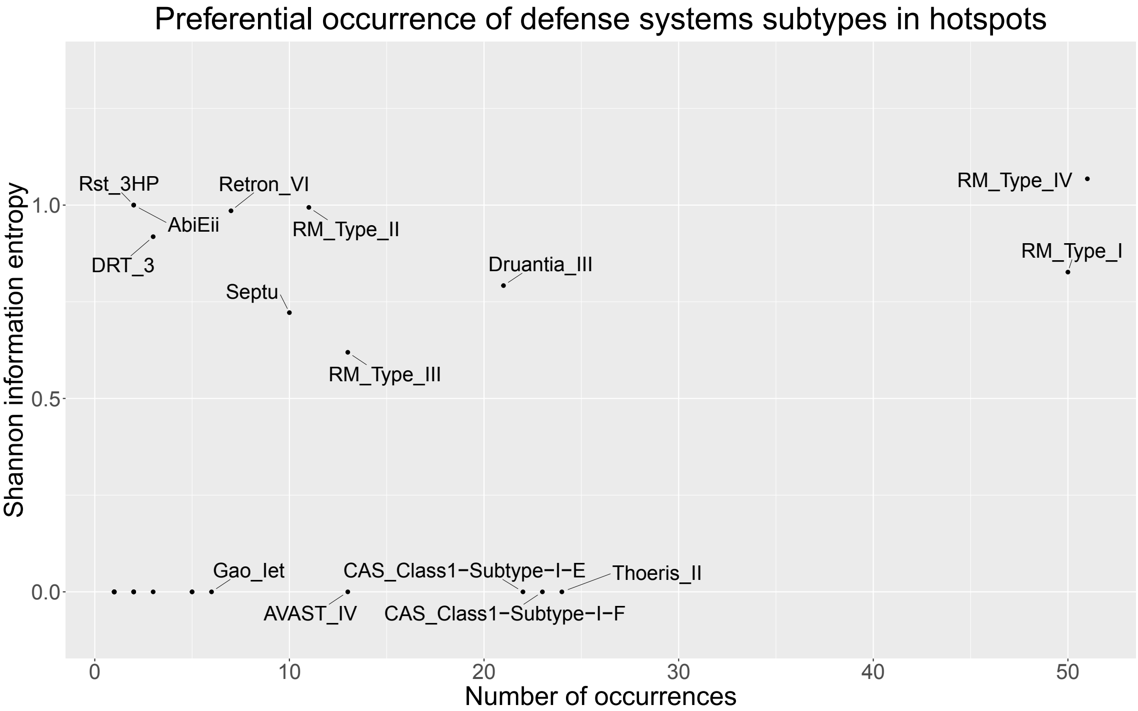


Figure 6.5: Preference of defense systems to integrate into specific hotspots measured by the Shannon information entropy index. Some defense systems show preference to integrate into the same hotspot, while other seem to integrate into random hotspots. For defense systems present only a few times, the index is unreliable, but for those that present more often it suggests which ones have a preference and which not. This plot suggests that systems AVAST_IV, CAS_Class1_Subtype-I-E, CAS_Class1_Subtype-I-F and Thoeris_II have a strong preference for one hotspot as indicated by their minimal entropy and the high entropy of Restriction-Modification systems suggests they integrate randomly. Some labels are excluded for clarity.

that are exclusively found in strains that are not lysed. Interestingly, the list of the defense system subtypes are the same and consists of the following members: AVAST_II, AVAST_IV,BREX_I, dCTPdeaminase, DRT_1, DRT_3, Gao_Iet, Gao_Ppl, Hachiman, Pif, Retron_IV, Rst_3HP, SspBCDE, Wadjet_III. When looking at the detected defense systems for all phages, interesting to observe is that the systems Pif and Gao_Ppl are found to be present in the list of potentially protective systems for all phages (32 times), followed by the Rst_3HP (30 times), Wadjet_III, SspBCDE, Hachiman, Gap_Iet and dCTPdeaminase (29 times). However, with a samples size this small (small number of tested bacterial strains and the big diversity of bacterial surface receptors) it is difficult to tell without additional experimental work whether the detected systems are really responsible for the protection against the phages.

6.4 Discussion

Knowledge of the anti-phage defense systems that offer protection against lysis by a phage could be valuable for practical purposes. One such case is phage therapy [148], usage of lytic phages to treat bacterial infections in combination with or instead of antibiotics [149]. Since defense systems are often found on mobile genetic elements and can easily spread in a population [125], it is important to know what defense systems are present in the infecting bacterial population even if they are absent from most of the individual cells. The approach outlined in this work aims to find for each phage from a collection a set of defense systems which should not be present in the bacterial population for a phage to be selected as a candidate for phage therapy. With a fixed collection of well characterized phages, additional pathogenic strains collected could be tested against the phages to get more precise defense system candidates to avoid during therapy.

For some phages, we identified defense systems that potentially neutralize them and we found two systems that appear to neutralize all of the tested phages. To confirm the findings, additional laboratory experiments are needed. Ideally, a set of strains with compatible receptors and systems which were found to be potentially protective should be tested in order to add evidence to the computational finding. The initial discoveries suggest that the proposed analysis framework is a promising and simple method for shortlisting phage candidates for the phage therapy, but additional experimentation is needed for the refinement and final confirmation.

6.4.1 Limitations

The proposed approach has some inherent limitations together with the limitations related to the small sample size. The first limitation comes from the sequencing data of the bacterial genomes. Assembly of the short reads results in a couple of hundreds of contigs. We don't know the ordering of the contigs which is important for the defense-finder and the underlying MacSysFinder [140] to find the defense systems as the ordering of the genes that are part of the defense systems is important for detection of the functional systems. Unless all the components of the defense systems are on a single contig, the defense systems could be missed or over-predicted. For the same reason, we can't determine the hotspot content if the bordering genes are not on the same contig. The issues related to the assembly could be solved by sequencing the genomes with the long read technology like Nanopore or PacBio.

In this study we assume that the bacterial surface receptors are identical if they have the same translated gene sequence and if a strain has a gene for a receptor that is determined to be compatible with a phage, then the phage can enter the cell. This assumption does not always hold. A phage could be unable to enter to cell for various reasons even if the receptors are compatible at the sequence level. For example, Cor protein from a prophage inactivates the receptor FhuA [150] making entry of the phages using it impossible as well as a gp15 protein from a phage HK97 which prevents other phages from entering without influencing the receptor [151]. Given that most of the strains from our collection are predicted to carry prophages, it is not unlikely that there might be interference with the entry of the phages even when the receptors are compatible.

The interaction matrix between the BASEL phages and the E. coli collection were determined based on the spot assays in the LB medium. Not all plaques formed after spotting phages on a bacterial lawn always look the same. Some plaques are turbid and it is up to the person noting the results to decide whether the plaque represents the lysis or not. The spot assays are not the only way to determine if a phage is able to lyse a bacterium. Other options include assays in the liquid medium and turbidity reduction assays where bacteria and phages are mixed together and the turbidity of the culture (representing bacterial growth) is compared to the turbidity of a pure bacterial culture. It is important to keep in mind that the results obtained in the laboratory do not necessarily translate to the ability of a phage to infect bacterium in a natural environment or in a human body in case of the phage therapy [152].

When searching for the defense systems, we rely on the tools that find known defense systems based on the specific genes. However, new systems and new classes of defense systems are constantly being discovered [132, 153]. Therefore, we know that we don't know the full anti-phage arsenal in the analyzed strains. Except for the presence of undiscovered defense systems that could influence the phage-bacteria interaction another problem is the interaction between defense systems. Defense systems don't operate in isolation from each other [154, 155]. Defense provided by individual systems could be "leaky" which means that they don't offer full protection against a phage by themselves, but partially contribute to the elimination of a threat and the order of activation could be hierarchical. For example, activation of retrons causes cell death and the trigger is interference of phage proteins with the RecBCD complex [156]. Production of phage proteins can only happen if the invading DNA was not completely eliminated by nucleases. In this work we don't account for the interaction between the systems due to the limited data which is effectively even smaller because for every phage only the strains for which we know the receptor compatibility are taken into consideration for analysis.

6.4.2 Opportunities

Besides limitations, the proposed analysis approach also offers opportunities. First of all, the methodology can be applied to any phage-bacteria systems as defense-finder is not specific to E. coli, but is able to find defense systems in any bacterial genome. Moreover, the analysis pipeline is automated and is simple to use even for labs without bioinformatic expertise. The methodology is also extendable and could be adapted to search for other exclusion criteria of phage and host pairs on top of the defense systems. It requires easily obtainable data like genomic sequences and the interaction matrix. The information most challenging to obtain is the phage terminal receptor, but by using well characterized phages like to ones from the BASEL collection can mitigate the problem.

6.5 Conclusion

In this work we present an analysis approach to identify defense systems causing certain E. coli strains to be resistant to an infection of specific phages. Even with the limited data available, we were able to detect systems that offer protection against a range of phages. A phage therapy lab implementing the presented analysis framework in the practice would be able to better guide the selection of therapeutic phages based on the pathogen genome and by creating a bigger interaction matrix, make it possible to improve the conclusions.

Chapter 7

Other involvements

Besides the projects described in the previous chapters, I have been involved in various other projects during my doctoral studies which are not mature enough for a chapter or publication.

Co-infection of bacteriophages

I collaborate with Jacob Bobonis from the group of Martin Polz in a project about phage coinfections in different species from the *Vibrio* genus. My contribution involves sequence analysis of the bacterial genomes. The main aims are finding defense systems, describing their genomic environment, describing the overlap of the defense system content across different species and strains and analyzing the diversity of the bacterial surface proteins involved in the phage attachment. The project is about exploring infections with multiple phages at the same time with a goal of better understanding the interaction between phage proteins and defense systems and how can a simultaneous presence of one phage influence the infection outcome of the other phage.

Prophage detection

With a PhD student from the Balgrist University Hospital in Zürich, Lindah Prossie Nankya, I work on a project focusing on the prophages in different *E. coli* strains isolated from the human bladder. I contribute by creating an automated pipeline for detection of prophage sequences within bacterial genomes, describing their genomic environment and by managing the sequencing data.

Phage-training experiment

As part of my training, I have spent 9 weeks working in the lab at the Balgrist University Hospital in Zürich, even though my PhD study is focused on bioinformatics. I have designed an experiment where I serially propagated mixtures of phages on different *E. coli* strains in two types of media, rich medium and the synthetic human urine. I have extracted the phage DNA from the every step of the experiment, sent it for sequencing and analyzed the resulting data. The hypothesis was that phages from the mixture would infect the host at the same time and recombine which would lead to the emergence of novel phage genomes with different host range. The sequencing data showed the outgrowth and the absolute domination of a single phage which was opposite of what was desired and the project was not continued.

Skin metabolome and aging

I have spent 6 weeks at the Medical University in Greifswald working with the human skin metabolome. The goal was to find a relationship between the metabolome profile and the age of the people donating the samples in order to guide the research in cosmetic industry. I have done the exploratory analysis of the metabolome and have started applying machine learning algorithms to predict the age based on the metabolites. I managed to show that there is correlation between the metabolome and the age and that different algorithms (XGBOOST and support vector machines) are able to predict the age better than random. After I have left Greifswald, another person took over the project due data protection as the data was not allowed to leave the institute.

Chapter 8

Conclusion

In this dissertation I had a dual focus. In the first part I introduce VOGDB - the newly developed resource for virus bioinformatics and viral genome analysis. Subsequently, I show the application of VOGDB in practice by analyzing two types of metagenomes, phage cocktails and dental calculus from museum samples. In the second part I focus on the phage-host interactions via the anti-phage defense systems and try to find which defense systems are responsible for offering *E. coli* the resistance against specific phages.

VOGDB as a resource is not completely unique as there are other similar databases grouping homologous viral proteins (pVOG [157], PHROG [158], eggNOG [56]), but has a unique selling proposition of being regularly updated and grouping phage and non-phage proteins together. On top, it has explicitly layered structure which allows users to select the desired homology "remoteness" for the analysis. The most remote homology is detected by clustering proteins structures predicted by AlphaFold2, however, due to the limitations in the compute power, we were not able to predict the structure of all the proteins, but had to select a subset of representatives for which the structure was predicted. We are aware that this approach is suboptimal, but it could be easily fixed in the future if viral proteins will be added into the AlphaFold DB [159].

We showed that quality assessment metrics of VOGDB are comparable to the other similar databases and therefore we expect VOGDB to be used and accepted by the virus bioinformatics community. It is already included in the widely used bioinformatics tools and workflows (Virsorter2 [160], CheckV [161]). As we plan the long-term maintenance, novel genomes will be regularly added, tools will be updated and new features will be added (better functional annotation is planned).

In the next two chapters, VOGDB was applied in the analysis of metagenomes. First, an automatic analysis pipeline was created that gives the estimate of the taxonomic composition of the phage cocktails and to numerically express similarity between them. In the scarce literature on the phage cocktail metagenomes, bioinformatic procedures are not standardized and are implemented as custom solutions by the authors. With the suggested workflow based on VOGDB we make one small step towards standardization and make it easier to compare the results of different studies. The VOGDB based pipeline makes it easy to reanalyze the samples and get the results with the newest viral taxonomy which is important if we want to compare the taxonomic composition as virus taxonomy is an evolving field [20].

When prior knowledge of the cocktails is taken into account (the producer, the indications and the production year), the results meet our expectation. Cocktails from the same producer and designed for the same indications are more similar to each other than to other cocktails and cocktails that are very different according to their descriptions are also dissimilar on the level of vFAM content and the taxonomy.

The analysis of the dental calculus metagenome was different for two reasons. The viral reads were rare and we focused on eukaryotic viruses. The goal was again to create an automated analysis pipeline to get an estimate of the viral families present in the samples in order to guide future experimental efforts. The challenge lied in the need to remove the non-viral proteins as they could be falsely assigned to vFAMS. Besides removing non-viral proteins, a problem with the direct protein assembly is that we don't know what proteins come together on one contig and therefore we can't use the neighboring genes to more precisely determine the taxonomic annotation of all the genes from the contig. However, we managed to get the estimate of the mammalian viral families from the dental calculus metagenome and the results are in the agreement with the literature.

The final chapter is about phage-host interaction with a focus on phage therapy setting. The partner lab (led by Dr. Shawna McCallin at the Balgrist University Hospital in Zürich) collected hundreds of *E. coli* strains from the human bladder and tested which phages from a published collection [135] can lyse which strains. I hypothesized that different defense systems are to be blamed for a different susceptibility of different strains to the phages from the collection. The defense system content in lysed and resistant strains was compared and I tried to detect a few of the defense systems that are the reason why a specific strain is resistant. A few systems that are present in the resistant strains could be the answer, but experimental confirmation is needed. Hopefully, research in this direction will help in selecting phages for the phage therapy by informing us which phages to avoid if we find certain defense systems in the pathogen population.

References

1. López-García, P. The Place of Viruses in Biology in Light of the Metabolism-versus-replication-first Debate. *History and Philosophy of the Life Sciences.* JSTOR: 43831419 (3, 2012).
2. Koonin, E. V. & Starokadomskyy, P. Are Viruses Alive? The Replicator Paradigm Sheds Decisive Light on an Old but Misguided Question. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* (2016).
3. Forterre, P. Manipulation of Cellular Syntheses and the Nature of Viruses: The Virocell Concept. *Comptes Rendus. Chimie* (2010).
4. Koonin, E. V., Dolja, V. V., Krupovic, M. & Kuhn, J. H. Viruses Defined by the Position of the Virosphere within the Replicator Space. *Microbiology and Molecular Biology Reviews* (2021).
5. De Courville, C., Cadarette, S. M., Wissinger, E. & Alvarez, F. P. The Economic Burden of Influenza among Adults Aged 18 to 64: A Systematic Literature Review. *Influenza and Other Respiratory Viruses.* pmid: 35122389 (2022).
6. Hilaire, J. *et al.* Risk Perception Associated with an Emerging Agri-Food Risk in Europe: Plant Viruses in Agriculture. *Agriculture & Food Security.* pmid: 35310134 (2022).
7. Jover, L. F., Effler, T. C., Buchan, A., Wilhelm, S. W. & Weitz, J. S. The Elemental Composition of Virus Particles: Implications for Marine Biogeochemical Cycles. *Nature Reviews Microbiology* (2014).
8. Sandaa, R.-A. Burden or Benefit? Virus–Host Interactions in the Marine Environment. *Research in Microbiology. Exploring the Prokaryotic Virosphere* (2008).
9. Manrique, P. *et al.* Healthy Human Gut Phageome. *Proceedings of the National Academy of Sciences* (2016).
10. Bar-On, Y. M., Phillips, R. & Milo, R. The Biomass Distribution on Earth. *Proceedings of the National Academy of Sciences* (2018).
11. Harrison, B. D. & Wilson, T. M. A. Milestones in Research on Tobacco Mosaic Virus. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* (1999).
12. Lou, Z., Sun, Y. & Rao, Z. Current Progress in Antiviral Strategies. *Trends in Pharmacological Sciences.* pmid: 24439476 (2014).
13. Krishnamurthy, S. R. & Wang, D. Origins and Challenges of Viral Dark Matter. *Virus Research. Deep Sequencing in Virology* (2017).
14. Van Regenmortel, M. Viruses Are Real, Virus Species Are Man-Made, Taxonomic Constructions. *Archives of Virology* (2003).
15. Zerbini, F. M. *et al.* Differentiating between Viruses and Virus Species by Writing Their Names Correctly. *Archives of Virology* (2022).
16. Fauquet, C. Taxonomy, Classification and Nomenclature of Viruses. *Encyclopedia of Virology.* pmid: null (2008).
17. FENNER, F. R. A. N. K. *et al.* Classification and Nomenclature of Viruses. *Veterinary Virology.* pmid: null (1987).

18. Ackermann, H.-W. & Abedon, S. T. *Bacteriophage Names 2000*

19. Adriaenssens, E. M. & Brister, J. R. How to Name and Classify Your Phage: An Informal Guide. *Viruses.* pmid: 28368359 (2017).

20. Turner, D. *et al.* Abolishment of Morphology-Based Taxa and Change to Binomial Species Names: 2022 Taxonomy Update of the ICTV Bacterial Viruses Subcommittee. *Archives of Virology* (2023).

21. Zong, Z. Genome-Based Taxonomy for Bacteria: A Recent Advance. *Trends in Microbiology.* pmid: 32980201 (2020).

22. Eloe-Fadrosh, E. A. Towards a Genome-Based Virus Taxonomy. *Nature Microbiology* (2019).

23. Dutilh, B. E. *et al.* Perspective on Taxonomic Classification of Uncultivated Viruses. *Current Opinion in Virology* (2021).

24. Forterre, P. The Origin of Viruses and Their Possible Roles in Major Evolutionary Transitions. *Virus Research. Comparative Genomics and Evolution of Complex Viruses* (2006).

25. Koonin, E. V., Senkevich, T. G. & Dolja, V. V. The Ancient Virus World and Evolution of Cells. *Biology Direct* (2006).

26. Koonin, E. V., Wolf, Y. I. & Katsnelson, M. I. Inevitability of the Emergence and Persistence of Genetic Parasites Caused by Evolutionary Instability of Parasite-Free States. *Biology Direct.* pmid: 29202832 (2017).

27. Simmonds, P. *et al.* Four Principles to Establish a Universal Virus Taxonomy. *PLOS Biology* (2023).

28. Guo, J. *et al.* VirSorter2: A Multi-Classifier, Expert-Guided Approach to Detect Diverse DNA and RNA Viruses. *Microbiome* (2021).

29. Hendrix, R. W. Bacteriophage Genomics. *Current Opinion in Microbiology* (2003).

30. Brüssow, H., Canchaya, C. & Hardt, W.-D. Phages and the Evolution of Bacterial Pathogens: From Genomic Rearrangements to Lysogenic Conversion. *Microbiology and Molecular Biology Reviews.* pmid: 15353570 (2004).

31. Patel, P. H. & Maxwell, K. L. Prophages Provide a Rich Source of Antiphage Defense Systems. *Current Opinion in Microbiology* (2023).

32. Hughes, A. L. & Friedman, R. Poxvirus Genome Evolution by Gene Gain and Loss. *Molecular Phylogenetics and Evolution* (2005).

33. Cáceres, M., Thomas, J. W. & NISC Comparative Sequencing Program. The Gene of Retroviral Origin Syncytin 1 Is Specific to Hominoids and Is Inactive in Old World Monkeys. *Journal of Heredity* (2006).

34. Stein, R. A. & DePaola, R. V. Human Endogenous Retroviruses: Our Genomic Fossils and Companions. *Physiological Genomics* (2023).

35. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products. *Chemistry & Biology* (1998).

36. New, F. N. & Brito, I. L. What Is Metagenomics Teaching Us, and What Is Missed? *Annual Review of Microbiology* (2020).

37. Conceição-Neto, N. *et al.* Modular Approach to Customise Sample Preparation Procedures for Viral Metagenomics: A Reproducible Protocol for Virome Analysis. *Scientific Reports* (2015).

38. Soria-Villalba, A. *et al.* Comparison of Experimental Methodologies Based on Bulk-Metagenome and Virus-like Particle Enrichment: Pros and Cons for Representativeness and Reproducibility in the Study of the Fecal Human Virome. *Microorganisms* (1 2024).

39. Leskinen, K. *et al.* YerA41, a Yersinia Ruckeri Bacteriophage: Determination of a Non-Sequencable DNA Bacteriophage Genome via RNA-Sequencing. *Viruses* (6 2020).

40. Rihtman, B. *et al.* A New Family of Globally Distributed Lytic Roseophages with Unusual Deoxythymidine to Deoxyuridine Substitution. *Current Biology.* pmid: 34033748 (2021).

41. Rose, R., Constantinides, B., Tapinos, A., Robertson, D. L. & Prosperi, M. Challenges in the Analysis of Viral Metagenomes. *Virus Evolution* (2016).

42. Rohwer, F. Global Phage Diversity. *Cell* (2003).

43. Wang, R. H. *et al.* PhageScope: A Well-Annotated Bacteriophage Database with Automatic Analyses and Visualizations. *Nucleic Acids Research* (2024).

44. Aevarsson, A. *et al.* Going to Extremes – a Metagenomic Journey into the Dark Matter of Life. *FEMS Microbiology Letters* (2021).

45. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. Metaviral SPAdes : Assembly of Viruses from Metagenomic Data. *Bioinformatics* (2020).

46. Nayfach, S. *et al.* CheckV Assesses the Quality and Completeness of Metagenome-Assembled Viral Genomes. *Nature Biotechnology* (2021).

47. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: A Rapid, Scalable and Accurate Tool for Assessing Microbial Genome Quality Using Machine Learning. *Nature Methods* (8 2023).

48. Camargo, A. P. *et al.* Identification of Mobile Genetic Elements with geNomad. *Nature Biotechnology* (2023).

49. Wishart, D. S. *et al.* PHASTEST: Faster than PHASTER, Better than PHAST. *Nucleic Acids Research* (2023).

50. McNair, K., Zhou, C., Dinsdale, E. A., Souza, B. & Edwards, R. A. PHANOTATE: A Novel Approach to Gene Identification in Phage Genomes. *Bioinformatics* (2019).

51. Johansen, J. *et al.* Genome Binning of Viral Entities from Bulk Metagenomics Data. *Nature Communications* (2022).

52. Kieft, K., Adams, A., Salamzade, R., Kalan, L. & Anantharaman, K. vRhyme Enables Binning of Viral Genomes from Metagenomes. *Nucleic Acids Research* (2022).

53. Sayers, E. W. *et al.* Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* (2022).

54. Camargo, A. P. *et al.* IMG/VR v4: An Expanded Database of Uncultivated Virus Genomes within a Framework of Extensive Functional, Taxonomic, and Ecological Metadata. *Nucleic Acids Research* (2023).

55. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): A Resource for Comparative Genomics and Protein Family Annotation. *Nucleic Acids Research* (2017).

56. Huerta-Cepas, J. *et al.* eggNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences. *Nucleic Acids Research* (2016).

57. Terzian, P. *et al.* PHROG: Families of Prokaryotic Virus Proteins Clustered Using Remote Homology. *NAR Genomics and Bioinformatics* (2021).

58. Jim O'Neill. Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. *The review on antimicrobial resistance* (2016).

59. Wright, G. D. Q&A: Antibiotic Resistance: Where Does It Come from and What Can We Do about It? *BMC Biology* (2010).

60. Diallo, K. & Dublanchet, A. Benefits of Combined Phage–Antibiotic Therapy for the Control of Antibiotic-Resistant Bacteria: A Literature Review. *Antibiotics.* pmid: 35884092 (2022).

61. Liu, D. *et al.* The Safety and Toxicity of Phage Therapy: A Review of Animal and Clinical Studies. *Viruses.* pmid: 34209836 (2021).

62. Berkson, J. D. *et al.* Phage-Specific Immunity Impairs Efficacy of Bacteriophage Targeting Vancomycin Resistant Enterococcus in a Murine Model. *Nature Communications* (2024).

63. Oechslin, F. Resistance Development to Bacteriophages Occurring during Bacteriophage Therapy. *Viruses* (7 2018).

64. Mutti, M. *et al.* Phage Activity against Staphylococcus Aureus Is Impaired in Plasma and Synovial Fluid. *Scientific Reports* (2023).

65. Malik, D. J. *et al.* Formulation, Stabilisation and Encapsulation of Bacteriophage for Phage Therapy. *Advances in Colloid and Interface Science. Recent Nanotechnology and Colloid Science Development for Biomedical Applications* (2017).

66. Seed, K. D. Battling Phages: How Bacteria Defend against Viral Attack. *PLOS Pathogens* (2015).

67. Labrie, S. J., Samson, J. E. & Moineau, S. Bacteriophage Resistance Mechanisms. *Nature Reviews Microbiology* (2010).

68. Pirnay, J.-P. & Verbeken, G. Magistral Phage Preparations: Is This the Model for Everyone? *Clinical Infectious Diseases* (Supplement_5 2023).

69. Danis-Wlodarczyk, K. M., Wozniak, D. J. & Abedon, S. T. Treating Bacterial Infections with Bacteriophage-Based Enzybiotics: In Vitro, In Vivo and Clinical Application. *Antibiotics* (12 2021).

70. Al-Anany, A. M., Fatima, R. & Hynes, A. P. Temperate Phage-Antibiotic Synergy Eradicates Bacteria through Depletion of Lysogens. *Cell Reports* (2021).

71. Monteiro, R., Pires, D. P., Costa, A. R. & Azeredo, J. Phage Therapy: Going Temperate? *Trends in Microbiology. Special Issue: Antimicrobial Resistance and Novel Therapeutics* (2019).

72. Gaborieau, B. *et al. Predicting Phage-Bacteria Interactions at the Strain Level from Genomes* `http://biorxiv.org/lookup/doi/10.1101/2023.11.22.567924` (2024). preprint.

73. Hussain, F. A. *et al.* Rapid Evolutionary Turnover of Mobile Genetic Elements Drives Bacterial Resistance to Phages. *Science* (2021).

74. Kiening, M. *et al.* Conserved Secondary Structures in Viral mRNAs. *Viruses* (5 2019).

75. Jumper, J. *et al.* Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* (2021).

76. Di Tommaso, P. *et al.* Nextflow Enables Reproducible Computational Workflows. *Nature Biotechnology* (2017).

77. Mölder, F. *et al.* Sustainable Data Analysis with Snakemake. *F1000Research* (2021).

78. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research* (1997).

79. Kutter, E. *et al.* Phage Therapy in Clinical Practice: Treatment of Human Infections. *Current Pharmaceutical Biotechnology* (2010).

80. Burrowes, B., Molineux, I. & Fralick, J. Directed in Vitro Evolution of Therapeutic Bacteriophages: The Appelmans Protocol. *Viruses* (2019).

81. Yang, Q., Le, S., Zhu, T. & Wu, N. Regulations of Phage Therapy across the World. *Frontiers in Microbiology* (2023).

82. McCallin, S., Sacher, J. C., Zheng, J. & Chan, B. K. Current State of Compassionate Phage Therapy. *Viruses* (4 2019).

83. Verbeken, G. & Pirnay, J.-P. European Regulatory Aspects of Phage Therapy: Magistral Phage Preparations. *Current Opinion in Virology* (2022).

84. Vázquez, R. *et al.* Essential Topics for the Regulatory Consideration of Phages as Clinically Valuable Therapeutic Agents: A Perspective from Spain. *Microorganisms.* pmid: `35456768` (2022).

85. Zschach, H. *et al.* What Can We Learn from a Metagenomic Analysis of a Georgian Bacteriophage Cocktail? *Viruses* (2015).

86. McCallin, S. *et al.* Safety Analysis of a Russian Phage Cocktail: From MetaGenomic Analysis to Oral Application in Healthy Human Subjects. *Virology* (2013).

87. Villarroel, J., Larsen, M., Kilstrup, M. & Nielsen, M. Metagenomic Analysis of Therapeutic PYO Phage Cocktails from 1997 to 2014. *Viruses* (2017).

88. McCallin, S., Sarker, S. A., Sultana, S., Oechslin, F. & Brüssow, H. Metagenome Analysis of Russian and Georgian Pyophage Cocktails and a Placebo-controlled Safety Trial of Single Phage versus Phage Cocktail in Healthy *Staphylococcus Aureus* Carriers. *Environmental Microbiology* (2018).

89. Pirnay, J.-P. *et al.* Quality and Safety Requirements for Sustainable Phage Therapy Products. *Pharmaceutical Research* (2015).

90. Sarker, S. A. *et al.* Oral T4-like Phage Cocktail Application to Healthy Adult Volunteers from Bangladesh. *Virology. Special Issue: Viruses of Microbes* (2012).

91. Leitner, L. *et al.* Intravesical Bacteriophages for Treating Urinary Tract Infections in Patients Undergoing Transurethral Resection of the Prostate: A Randomised, Placebo-Controlled, Double-Blind Clinical Trial. *The Lancet Infectious Diseases.* pmid: 32949500 (2021).

92. Güemes, A. G. C. *et al.* Viruses as Winners in the Game of Life. *Annual Review of Virology* (Volume 3, 2016 2016).

93. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* (2014).

94. Li, H. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM* arXiv: 1303.3997 [q-bio]. http://arxiv.org/abs/1303.3997 (2024). preprint.

95. Danecek, P. *et al.* Twelve Years of SAMtools and BCFtools. *GigaScience* (2021).

96. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: A New Versatile Metagenomic Assembler. *Genome Research* (2017).

97. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J. & Levy Karin, E. Fast and Sensitive Taxonomic Assignment to Metagenomic Contigs. *Bioinformatics* (2021).

98. Steinegger, M. & Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nature Biotechnology* (2017).

99. Hyatt, D. *et al.* Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinformatics* (2010).

100. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* (2011).

101. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *Journal of Molecular Biology* (1990).

102. Barylski, J., Kropinski, A. M., Alikhan, N.-F., Adriaenssens, E. M. & ICTV Report Consortium. ICTV Virus Taxonomy Profile: Herelleviridae. *Journal of General Virology* (2020).

103. O'Leary, N. A. *et al.* Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. *Nucleic Acids Research* (2016).

104. Li, Q. *et al.* Dental Calculus: A Repository of Bioinformation Indicating Diseases and Human Evolution. *Frontiers in Cellular and Infection Microbiology* (2022).

105. Brealey, J. C. *et al.* Dental Calculus as a Tool to Study the Evolution of the Mammalian Oral Microbiome. *Molecular Biology and Evolution* (2020).

106. Oliva, A., Tobler, R., Cooper, A., Llamas, B. & Souilmi, Y. Systematic Benchmark of Ancient DNA Read Mapping. *Briefings in Bioinformatics* (2021).

107. Baker, J. L., Mark Welch, J. L., Kauffman, K. M., McLean, J. S. & He, X. The Oral Microbiome: Diversity, Biogeography and Human Health. *Nature Reviews Microbiology* (2024).

108. Krishnamurthy, S. R. & Wang, D. Origins and Challenges of Viral Dark Matter. *Virus Research. Deep Sequencing in Virology* (2017).

109. Steinegger, M., Mirdita, M. & Söding, J. Protein-Level Assembly Increases Protein Sequence Recovery from Metagenomic Samples Manyfold. *Nature Methods* (2019).

110. Köster, J. & Rahmann, S. Snakemake—a Scalable Bioinformatics Workflow Engine. *Bioinformatics* (2012).

111. Hämmerle, M. *et al.* Link between Monkeypox Virus Genomes from Museum Specimens and 1965 Zoo Outbreak. *Emerging Infectious Diseases.* pmid: 38526306 (2024).

112. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND. *Nature Methods* (2021).

113. Team, R. C. *R: A Language and Environment for Statistical Computing* 2023.

114. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis* (2016).

115. Wickham, H. Reshaping Data with the **Reshape** Package. *Journal of Statistical Software* (2007).

116. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. *Dplyr: A Grammar of Data Manipulation* 2023.

117. Wickham, H. *Stringr: Simple, Consistent Wrappers for Common String Operations* 2023.

118. Warnes, G. R. *et al. Gtools: Various R Programming Tools* 2023.

119. Duncan, M. *et al.* Adenoviruses Isolated from Wild Gorillas Are Closely Related to Human Species C Viruses. *Virology* (2013).

120. Narat, V. *et al.* Higher Convergence of Human-Great Ape Enteric Eukaryotic Viromes in Central African Forest than in a European Zoo: A One Health Analysis. *Nature Communications* (2023).

121. Dunay, E. *et al.* Viruses in Saliva from Sanctuary Chimpanzees (Pan Troglodytes) in Republic of Congo and Uganda. *PLOS ONE* (2023).

122. Hobbs, Z. & Abedon, S. T. Diversity of Phage Infection Types and Associated Terminology: The Problem with 'Lytic or Lysogenic'. *FEMS Microbiology Letters* (2016).

123. Hatfull, G. F., Dedrick, R. M. & Schooley, R. T. Phage Therapy for Antibiotic-Resistant Bacterial Infections. *Annual Review of Medicine* (2022).

124. Hampton, H. G., Watson, B. N. J. & Fineran, P. C. The Arms Race between Bacteria and Their Phage Foes. *Nature* (2020).

125. Hussain, F. A. *et al.* Rapid Evolutionary Turnover of Mobile Genetic Elements Drives Bacterial Resistance to Phages. *Science* (2021).

126. Vassallo, C. N., Doering, C. R., Littlehale, M. L., Teodoro, G. I. C. & Laub, M. T. A Functional Selection Reveals Previously Undetected Anti-Phage Defence Systems in the E. Coli Pangenome. *Nature Microbiology* (2022).

127. Hochhauser, D., Millman, A. & Sorek, R. The Defense Island Repertoire of the Escherichia Coli Pan-Genome. *PLOS Genetics* (2023).

128. Strathdee, S. A., Hatfull, G. F., Mutalik, V. K. & Schooley, R. T. Phage Therapy: From Biological Mechanisms to Future Directions. *Cell* (2023).

129. Leitner, L. *et al.* Intravesical Bacteriophages for Treating Urinary Tract Infections in Patients Undergoing Transurethral Resection of the Prostate: A Randomised, Placebo-Controlled, Double-Blind Clinical Trial. *The Lancet Infectious Diseases* (2021).

130. Sarker, S. A. *et al.* Oral Phage Therapy of Acute Bacterial Diarrhea With Two Coliphage Preparations: A Randomized Trial in Children From Bangladesh. *EBioMedicine* (2016).

131. Doron, S. *et al.* Systematic Discovery of Antiphage Defense Systems in the Microbial Pangenome. *Science* (2018).

132. Millman, A. *et al.* An Expanded Arsenal of Immune Systems That Protect Bacteria from Phages. *Cell Host & Microbe* (2022).

133. Chen, I.-M. A. *et al.* The IMG/M Data Management and Analysis System v.7: Content Updates and New Features. *Nucleic Acids Research* (2023).

134. Park, S.-C., Lee, K., Kim, Y. O., Won, S. & Chun, J. Large-Scale Genomics Reveals the Genetic Characteristics of Seven Species and Importance of Phylogenetic Distance for Estimating Pan-Genome Size. *Frontiers in Microbiology* (2019).

135. Maffei, E. *et al.* Systematic Exploration of Escherichia Coli Phage–Host Interactions with the BASEL Phage Collection. *PLOS Biology* (2021).

136. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics* (2020).

137. Seemann, T. Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics* (2014).

138. Arndt, D. *et al.* PHASTER: A Better, Faster Version of the PHAST Phage Search Tool. *Nucleic Acids Research* (2016).

139. Tesson, F. *et al.* Systematic and Quantitative View of the Antiviral Arsenal of Prokaryotes. *Nature Communications* (1 2022).

140. Abby, S. S., Néron, B., Ménager, H., Touchon, M. & Rocha, E. P. C. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLOS ONE* (2014).

141. Page, A. J. *et al.* Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis. *Bioinformatics* (2015).

142. Katoh, K., Misawa, K., Kuma, K.-i. & Miyata, T. MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Research.* pmid: 12136088 (2002).

143. Hyman, P. & Abedon, S. T. in *Advances in Applied Microbiology* (Elsevier, 2010).

144. Eren, A. M. *et al.* Community-Led, Integrated, Reproducible Multi-Omics with Anvi'o. *Nature Microbiology* (2020).

145. Edgar, R. C. MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity. *BMC Bioinformatics* (2004).

146. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* (2010).

147. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. Ggtree: An r Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data. *Methods in Ecology and Evolution* (2017).

148. Egido, J. E., Costa, A. R., Aparicio-Maldonado, C., Haas, P.-J. & Brouns, S. J. J. Mechanisms and Clinical Importance of Bacteriophage Resistance. *FEMS Microbiology Reviews* (2022).

149. Liu, C., Hong, Q., Chang, R. Y. K., Kwok, P. C. L. & Chan, H.-K. Phage–Antibiotic Therapy as a Promising Strategy to Combat Multidrug-Resistant Infections and to Enhance Antimicrobial Efficiency. *Antibiotics* (2022).

150. Arguijo-Hernández, E. S. *et al.* Cor Interacts with Outer Membrane Proteins to Exclude FhuA-dependent Phages. *Archives of Virology* (2018).

151. Cumby, N., Edwards, A. M., Davidson, A. R. & Maxwell, K. L. The Bacteriophage HK97 Gp15 Moron Element Encodes a Novel Superinfection Exclusion Protein. *Journal of Bacteriology* (2012).

152. Hyman, P. Phages for Phage Therapy: Isolation, Characterization, and Host Range Breadth. *Pharmaceuticals* (2019).

153. Duan, N., Hand, E., Pheko, M., Sharma, S. & Emiola, A. Structure-Guided Discovery of Anti-CRISPR and Anti-Phage Defense Proteins. *Nature Communications* (2024).

154. Tesson, F. & Bernheim, A. Synergy and Regulation of Antiphage Systems: Toward the Existence of a Bacterial Immune System? *Current Opinion in Microbiology* (2023).

155. Wu, Y. *et al. Synergistic Anti-Phage Activity of Bacterial Defence Systems* preprint (Microbiology, 2022).

156. Millman, A. *et al.* Bacterial Retrons Function In Anti-Phage Defense. *Cell* (2020).

157. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): A Resource for Comparative Genomics and Protein Family Annotation. *Nucleic Acids Research* (2017).

158. Terzian, P. *et al.* PHROG: Families of Prokaryotic Virus Proteins Clustered Using Remote Homology. *NAR Genomics and Bioinformatics* (2021).

159. Varadi, M. *et al.* AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Research* (2022).

160. Guo, J. *et al.* VirSorter2: A Multi-Classifier, Expert-Guided Approach to Detect Diverse DNA and RNA Viruses. *Microbiome* (2021).

161. Nayfach, S. *et al.* CheckV Assesses the Quality and Completeness of Metagenome-Assembled Viral Genomes. *Nature Biotechnology* (2021).

Appendix A

VOGDB manuscript preprint

VOGDB - Database of Virus Orthologous Groups

Lovro Trgovec-Greif[1,2], Hans-Jörg Hellinger[2], Jean Mainguy[1], Alexander Pfundner[1,2], Dmitrij Frishman[1], Michael Kiening[1], Nicole Webster[1], Patrick Laffy[1], Michael Feichtinger[2], Thomas Rattei[2]

[1]Doctoral School of Microbiology and Environmental Systems Science, University of Vienna

[2]Center for Microbiology and Environemtnal Systems Science, University of Vienna

Abstract

Computational models of homologous protein groups are essential in sequence bioinformatics. Due to the diversity and rapid evolution of viruses, the grouping of protein sequences from virus genomes is particularly challenging. The low sequence similarities of homologous genes in viruses require specific approaches for sequence- and structure-based clustering. Furthermore, the annotation of virus genomes in public databases is not as consistent and up-to-date as for many cellular genomes. To tackle these problems, we have developed VOGDB, a database of Virus Orthologous Groups. VOGDB is a multi-layer database that progressively groups viral genes into groups connected by increasingly remote homology. The first layer is based on pairwise sequence similarities, the second layer is based on the sequence profile alignments and the third layer uses predicted protein structures to find the most remote homology. VOGDB groups allow for more sensitive homology searches of novel genes and increase the chance of predicting annotations or inferring phylogeny. VOGDB uses all virus genomes from RefSeq and partially re-annotates them. VOGDB is updated with every RefSeq release. The unique feature of VOGDB is inclusion of both prokaryotic and eukaryotic viruses in the same clustering process which makes it possible to explore old evolutionary relationships of the two groups. VOGDB is freely available at `https://vogdb.org` **under the CC BY 4.0 license.**

Introduction

Viruses are a diverse group of biological entities that share the property of being obligate cellular parasites. Unlike in cellular organisms, no common genes or gene families are shared between all viruses [1]. This raises fundamental questions about virus ancestry and evolution. Moreover, the number of viruses on earth is huge (more than 10^{31} particles) [2, 3] and it is estimated they carry between 10^8 and 10^{10} unique genes [4]. Most of the viral diversity is currently unexplored and for the most sequenced viral genes, little is known about their function [5].

Viral genes not only encode a high number of different functions, which leads to a huge diversity of viral genomes, but also form heterogeneous groups of genes having similar function [6]. Due to the nature of viral lifestyle and their quick replication, mutations and selection, viruses explore the sequence space of genes in less evolutionary time than cellular organisms do [7, 8]. Because of the heterogeneity of viral proteins, it is often difficult to find homologs in databases by traditional bioinformatics, such as pairwise sequence alignments.

The computational inference of gene homology is valuable for annotating genes that are known from their sequence, but have not been experimentally characterized. Homologous genes have diverged from a common ancestral gene and are likely to have same or similar functions in different organisms. A particularly informative computational observation is gene orthology. Orthologous genes have diverged from a common ancestor by a process of speciation (as opposed to the gene duplication in paralogy). Orthologous genes are more likely to keep the ancestral function [9]. Orthologous genes from multiple organisms form orthologous groups. Homologous relationships are deduced from sequence comparisons due to the assumption that important sequence motifs will stay conserved during the evolution [10]. However, due to the absence of universal phylogenetic markers for all viruses and frequent horizontal gene transfers between viruses and viruses as well viruses and hosts, no universal concepts for the orthology of viral protein families are so far available in bioinformatics.

Due to the quick viral evolution, it is often impossible to detect homology by pairwise alignment of two protein sequences, especially for proteins that diverged longer time ago. However, by building a sequence model based on the group of easily detectable homologs, a conserved pattern becomes clear and it could be used to connect more distant groups [11]. This approach is widely used by databases that cluster together viral proteins, including pVOG [12], which focuses on prokaryotic viruses, as well as the viral sequences of eggNOG [13]. The PHROGs database [14] clusters phage genomes in two steps, first by grouping them based on the direct sequence comparison and later by clustering group Hidden Markov Models (HMM) to capture the remote homology. However, none of these databases represents the high number and wide diversity of virus genome sequences available to date.

We therefore introduce VOGDB, a database of virus orthologous groups, virus protein families and virus protein structural folds. VOGDB provides these three layers of homologous groups for all viral proteins from RefSeq genomes [15]. The layers are intended to gather proteins with the increasing evolutionary distance reflected in the bigger sequence divergence. Contrary to the prokaryotic genomes from RefSeq, where PGAP [16] is used for the consistent re-annotation of genomes, virus genomes in RefSeq keep their annotation from their GenBank

[17] submission. VOGDB, making use of all virus genomes from RefSeq, addresses the problem of inconsistent and outdated annotation by filtering and partial re-annotation, in order to ensure higher quality of final clusters. The first layer VOGDB is constructed by all against all pair-wise sequence comparison and represents the easily detectable homologs. The second layer is created by clustering sequence models (HMMs) from the first layer to capture homology of proteins that diverged beyond the point where the homology can be detected by pair-wise alignments. In the third layer we group together families from the second layer by their shared features within predicted 3D structures. This layer represents remotely homologous groups whose members diverged so much that sequence comparison methods can't detect their similarity anymore. As there is no standard way to validate viral orthologous groups, we suggest an approach based on the homogeneity of functional and structural annotations in terms of SwissProt [18] keywords and SCOPe [19] superfamily labels. The calculation of homogeneity was also applied to other similar databases (pVOG, PHROGs and COG), to compare if VOGDB shows similar homogeneity despite its higher number and wider diversity of genome sequences. pVOG and PHROGs are databases with viral proteins and are directly comparable to the first and second layers from VOGDB. The COG database contains prokaryotic proteins grouped by orthology and was included as a control.

Material and methods

Preprocessing of input data

Input from RefSeq The input data are all of the complete viral genomes from RefSeq [15] which have at least one protein annotated. Around 98% of records from RefSeq enter the VOGDB pipeline meaning VOGDB represents almost the entire viral portion of RefSeq. All sequence records with the same taxonomy ID, strain and isolate are considered one genome in VOGDB, further called VOGDB genomes.

Polyproteins Polyproteins are present in DNA viruses and almost all RNA and Retroviruses. A polyprotein is translated as a large polypeptide from a single ORF and is later cleaved into functional proteins [20]. At the moment, no general computational strategy exists that would predict the cleavage sites in the polyprotein and find the borders of the individual peptides. We have developed a strategy to annotate the individual peptides from the polyprotein sequence. First, individual peptides originating from a polyprotein or from RefSeq records that have been validated by the VOGDB team are collected in a peptide reference database. Second, non-annotated or incompletely annotated polyproteins are then reannotated by the best non-overlapping pairwise sequence alignments against the peptide reference database. Within VOGDB, annotated or reannotated peptides replace the respective segments of their initial polyprotein records and together with the rest of the proteins are called VOGDB proteins.

Creation of the first layer clusters - VOGs

VOGDB proteins are used as the input to the COGSoft pipeline with the aim of constructing clusters of recently diverged proteins [21]. In short, all against all psiblast [22] search is done followed by the COGtriangles [21] procedure to find the orthologous groups. We use the strict clustering which doesn't allow for a single protein to be a member of multiple clusters. For each orthologous group, a multiple sequence alignment of all member proteins is calculated using Clustal Omega [23]. From the multiple alignment we calculate Hidden Markov Model (HMM) using hmmbuild from hmmer [24]. The resulting groups are called VOGs to reflect they are viral equivalent to orthologous groups.

Functional annotation Annotations of VOGDB clusters are made with the aim of describing most of the cluster members as specifically as possible and therefore we are taking a consensus of the annotations of the individual proteins as the cluster annotation. vVGs are functionally annotated, if possible, by deriving functional annotations from hits to the most recent SwissProt [18] database or from the annotations as provided by RefSeq. To retrieve the annotation from SwissProt we used BLAST [25] to search the SwissProt database with the members of a VOG. For individual protein from a VOG we retained the functional annotation of a maximum of 5 hits if the e-value was less than 10^{-10} and the alignment coverage was more than 90%. All annotations of all proteins in a VOG are collected and the most common annotation string found for a VOG is used as the annotation for that VOG. In cases when it is not possible to get the annotation from SwissProt, we collect annotations of proteins in a VOG as they are in RefSeq and use the most common annotation string as the annotation of the VOG.

As an additional step in the annotation, we maintain a list of SwissProt keywords with which we associate a functional category. Every functional annotation of VOGs belongs to one or more functional categories: virus replication (Xr), virus structure (Xs), viral protein beneficial for the host (Xh), viral protein beneficial for the virus(Xp) and unknown function(Xu).

Naming VOG are named with a prefix "VOG" and a number padded with zeroes. To facilitate the comparison of the results between releases, we implemented a stable numbering scheme. VOGs from the older release are compared to the VOGs from the newer release and the newer VOG get the name of the largest older VOG for which 50% or more of the proteins are found in the new VOG. For VOGs that don't get the name from the previous release, a new number is created.

Creation of the second layer clusters - VFAMs

Clustering using MCL To create second layer clusters (VFAMs), we first need to align HMMs of VOGs. The alignment is done using hhalign function from hhsuite [26]. The alignments are used as input to the MCL clustering algorithm [27] where VOGs are clustered with the inflation value of 2. Clustered sequences are aligned with Clustal Omega [23] and HMM of the alignment is calculated by the function hmmbuild from hmmer [24]. The functional annotation of VFAMs are obtained in the same way as for VOGs. Naming works the same as for VOGs, but with a prefix ”VFAM”.

Creation of the third layer clusters - VFOLDs

Third layer of VOGDB consists of VFOLDs, clusters of VFAMs grouped based on the shared structural features. A few experimentally resolved structures of viral proteins are available in the public databases like pdb [28]. Therefore, we used alphafold 2 [29] to predict structures of viral proteins in VFAMs. Since there are more than 500000 proteins in VFAMs, predicting this number of structures would not be feasible. The strategy was to select one representative for every VFAM and cluster the representatives instead of the whole VFAMs. To select a representative, we have aligned all members of VFAM to the HMM of that VFAM and selected the highest scoring member as a candidate for which the structure would be predicted by alphafold 2. After obtaining structure predictions for all representatives, we did the clustering using the FoldSeek tool [30] with the default settings. Functional annotations of VFOLDs are obtained in the same way as for VOGs and VFAMs.

Quality assessment of the clustering results

The quality of the clustering was assessed by the homogeneity of functional annotations of cluster members and the structural superfamily membership of cluster members. Homogeneity of clusters from VOGDB was compared to the homogeneity of a random model. We obtained the random model by randomly scrambling functional annotation keywords or structure superfamily labels between annotated proteins and calculated the homogeneity. The randomization step was repeated 1000 times. For functional annotations we searched SwissProt with all protein members of a group, used the keyword of the top-level functional annotation of the hits and calculated the relative frequency of the most common annotation compared to all retrieved annotations. To assess the homogeneity of structural patterns, we used a similar approach, but instead of searching SwissProt, we searched astral95 database (v2.08) [19] using cd-hit [31]. Hits were associated with protein structural superfamilies as described in the SCOPe database [19]. The homogeneity for superfamilies was calculated as relative frequency of the most common superfamily per group. Comparison of the homogeneity to the random model was done using the Kolmogorov-Smirnov test.

Results

Database As the RefSeq database is updated bi-montly, VOGDB is updated with every RefSeq release and the new release is made available shortly after the newest version of RefSeq is released. The release number of VOGDB is the same as the release number of RefSeq which was used to build it. As an example in the text, the VOGDB version 221 based on the RefSeq 221 will be used and it contains 14974 VOGDB genomes. The polyprotein reannotation step predicted 5499 additional peptides from 995 polyproteins.

Content In the VOGDB release 221, 606019 of viral proteins were clustered and produced 59196 of VOGs, 38576 of VFAMs and 30516 of VFOLDs (figure 1) . Due to the clustering, 352350 of proteins have functional annotation, compared to 333379 of the initial proteins from RefSeq that were not annotated as hypothetical proteins. The size distribution of the groups from all three layers shows the expected pattern observed in the similar databases where there are many of the smaller groups and a few of the larger groups. The distribution of the VOGs, VFAMs and VFOLDs divided in size bins ranging from very small to extremely large is visualized in the figure 2. A feature of VOGs, VFAMs and VFOLDs is the information on the lowest common ancestor (LCA) of the viruses contributing proteins to the groups. Interesting groups are those with LCA "Viruses" which means that proteins from different viral realms got clustered together. There are 2441 such VOGs (4.1%) and 1443 VFAMs (3.7%) and

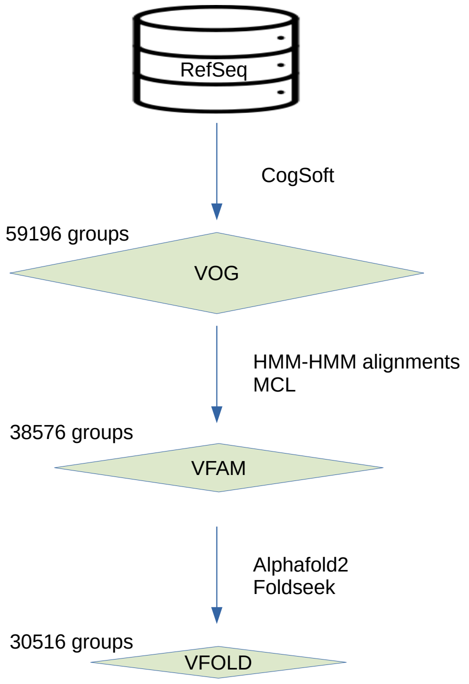


Figure 1: Schema of the layered structure of the database. For each layer different tools were used to create clusters. Clusters from every next layer are built from the clusters of the previous layer and are connected by more remote homology.

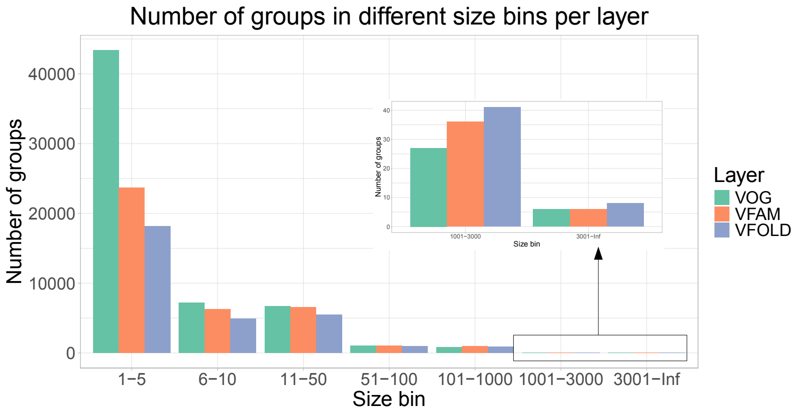


Figure 2: Number of groups per layer in different size bins. Size bins represent the range of the number of proteins for groups in a certain bin. The distribution with many smaller clusters and fewer of the larger ones is what is also observed in the similar databases.

1515 VFOLDs (4.8%).

Quality assessment

As there is not yet a universal standard procedure to evaluate the clustering of the viral proteins into orthologous groups, we assessed the quality of the VOGDB clusters using the homogeneity of functional annotation and structural classification. If the clustering would perfectly group the proteins by structure and function, all proteins in one cluster would have the same and unique functional and structural annotation. To achieve the best result, the level of granularity has to be carefully selected so that it could be applied to all of the databases and to be maximally informative as too coarse of the granularity would overestimate the homogeneity and too fine would underestimate it. We selected the SwissProt keywords and the SCOPe superfamilies as the granularity level at which we calculate the homogeneity. Because there is a limited number of keywords describing the function, we estimated the baseline of the homogeneity from the random model described earlier. Quality assessment based on the homogeneity (figure 3) shows that both functional and structural homogeneity of groups from different layers of VOGDB are significantly larger than the baseline for all of the size bins (Kolmogorov-Smirnov test, p-value $<10^{-5}$).

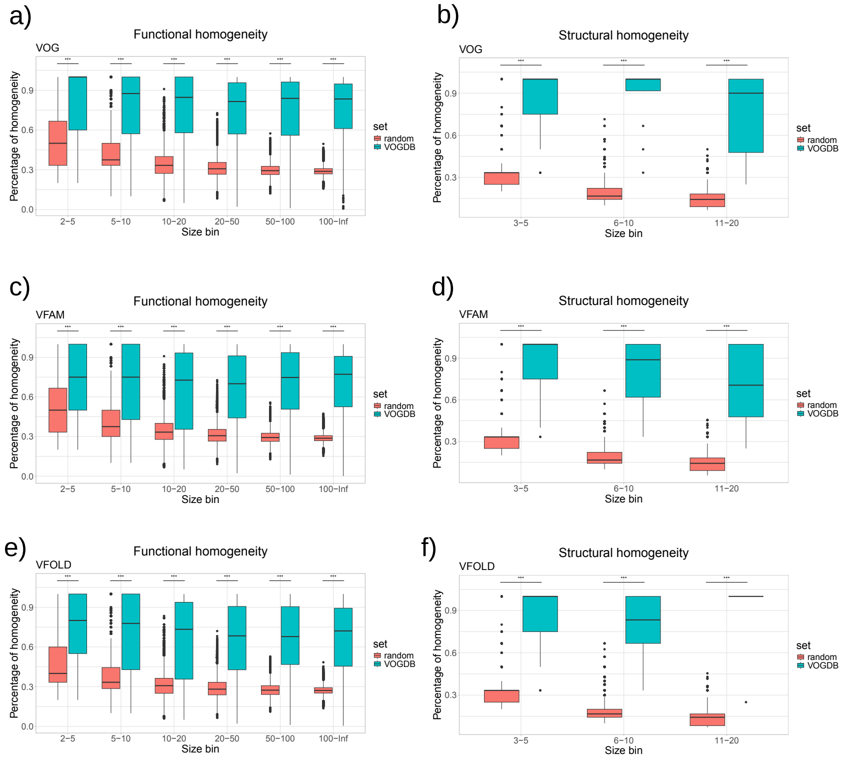


Figure 3: Homogeneity of functional annotations and protein structure classifications in VOGDB layers compared to the random model. The groups from each layer are put into size bins based on the number of proteins with functional and structural annotation. The random model is created by randomly redistributing the functional and structural annotation labels between the proteins with respective annotation for 1000 times and calculating the overall homogeneity. The results show that groups from VOGDB layers are significantly more homogeneous in terms of SwissProt keywords and structural classifications based on the SCOPe superfamilies (Kolmogorov-Smirnov test, p $<10^{-5}$).

Comparison with similar databases To evaluate the homogeneity of functional annotations and structural features, we calculated the homogeneity of the COG database (the release form 2020) [32], the PHROG database (v3) [14] and the pVOG database (May 2016) [12] in the same way as for the VOGDB layers . Clusters in the pVOG database are created similarly as VOGs and the PHROGs clusters are similar to VFAMs. However, VOGDB has a bigger scope than pVOG and PHROG by including both phages and eukaryotic viruses and is therefore faced with a harder clustering task. The COG database was included as a control as it was creating using similar clustering methodology and is manually curated. The figure 4 shows that the homogeneity of VOGDB layers is in the same range as the homogeneity of databases grouping prokaryotic (COG)

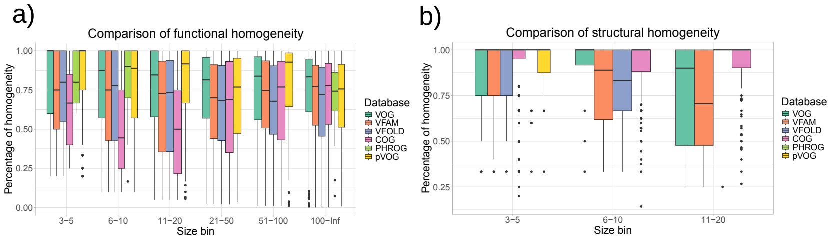


Figure 4: Homogeneity of SwissProt keywords (a) and SCOPe superfamilies (b) for layers from VOGDB and the other databases with orthologous/homologous groups: pVOG (phage orthologous groups), PHROG (phage remote orthologous groups) and COG (prokaryotic orthologous groups). The databases are split into size bins according to the number of proteins with a functional or structural annotation and if the bin has less than 3 members it is not shown. The results show that the function and structure based homogeneity of the layers from VOGDB are in the same range as in other similar databases.

orthologs, phage orthologs (pVOG) and phage remote homologs (PHROG). Homogeneity of clusters from pVOG and VOGs are very similar which is expected as both are created using COGSoft [21].

Availability

Webpage VOGDB is accessible online at the https://vogdb.org where it is possible to browse the clusters and see the statistics of the latest release. The webpage is updated regularly as new version of VOGDB is calculated.

Available files Apart from being accessible via the webpage, we offer all of the resulting files for download. The files offered are formatted similarly to the files offered by EggNOG database [33]. Most important files offered are HMMs of the clusters and multiple sequence alignments, file with the lowest common ancestry, files with functional annotation of clusters and predicted structures of VFAM representatives.

Discussion and application examples

Limitations VOGDB is so far the most complete database for virus orthologous groups, virus protein families and virus protein structural similarities. However, it is based on the annotations provided by the underlying RefSeq database. So far, several annotation quality filters and the re-annotation of polyproteins are the only means that VOGDB uses to ensure high accuracy of its input data. A consistent re-annotation of all virus genomes is not in the scope of VOGDB. Nevertheless, such re-annotation will become increasingly important to sustain the value of comparative genomics of viruses. The VOGDB groups can be a valuable tool towards this aim, e.g. by predicting protein-coding genes that were missed in the original genome annotations.

Support for bioinformatic workflows Viral hallmark genes [34] could be defined as genes that are found in diverse viruses, but have no or only few homologs in cellular organisms and are therefore indicative of the viral origin of a sequence. HMMs of VOGs and VFAMs that represent viral hallmark genes can be used to predict viral sequences from unknown genomes [35] and to estimate the contamination of a viral sequence with bacterial genes [36]. HMMs of groups of viral proteins (either hallmark or not) could be used as input for various other tools. For example, the tool HMM-GraspX [37] uses protein family HMMs to guide the assembly which is useful if the aim of the analysis is to analyze viruses in samples with low abundance of viral reads or when the focus on specific families are needed [38]. For VOGDB clusters, we calculate the virus specificity based on the number of hits to the cellular organisms based on the HMM-HMM search to the most recent eggNOG database [33]. The virus specificity information can be used to identify clusters representing the viral hallmark genes. The table 1 shows the number of virus specific VOGs and VFAMs at different stringency criteria, accepting few cellular homologs as expected e.g. from proviruses.

Table 1: Virus specificity of vFAMs. Virus specific vFAMs are useful for identifying the viral hallmark genes, the genes definitive for the viral state and with only very remote homology to cellular genes. In VOGDB, viral specificity is defined with three stringency levels. Strict, medium and low with hits to maximally two, three or four cellular genomes with e-value up to 10^{-4}, 10^{-10} and 10^{-15}.

Layer	**Strict**	**Medium**	**Low**
vOG	38562	45613	48627
vFAM	28500	32546	33951

Usage for metagenome analysis VOGDB is useful for analyzing metagenomic datasets that intentionally or accidentally contain virus DNA. When pair-wise sequence database searches fail to reveal hits, homology searches with databases containing HMMs, as from VOGDB, are more sensitive and allow for more proteins to be annotated. Besides the functional annotation, lineage information of the genome carrying the gene can be inferred. By mapping all genes of a viral contig to VFAMs and using the information about the lowest common ancestor of VFAMs, one can estimate the lineage of the whole contig (figure 5). When comparing viromes, the overlaps of the contents of the groups from VOGDB can be used to estimate relatedness and similarity.

Conclusion

VOGDB is a novel resource in the field of virus bioinformatics and it offers unique features compared to the similar databases and will complement the

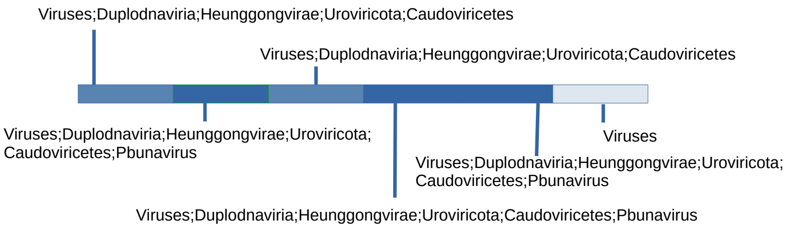

Viruses;Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes
Viruses;Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes
Viruses;Duplodnaviria;Heunggongvirae;Uroviricota;
Caudoviricetes;Pbunavirus
Viruses
Viruses;Duplodnaviria;Heunggongvirae;Uroviricota;
Caudoviricetes;Pbunavirus
Viruses;Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes;Pbunavirus

Figure 5: Taxonomic classification of a viral contig with 6 identified genes retrieved from a phage cocktail metagenome [39] (ERS1989570). If genes from a viral contig can be mapped to VFAMs, we can use the LCA of the VFAMs to deduce the possible taxonomic classification of a contig. The lowest level that could be reach is the lowest level of all LCAs that don't contradict each other. In this example, the lowest taxonomic level that could be reached is the genus Pbunavirus since other genes are mapped to VFAMs that have LCA compatible with the genus Pbunavirus.

current toolbox for studying viral genomes. By including both phages and eukaryotic viruses from RefSeq, VOGDB has the biggest scope of all virus orthology databases and it still ranks similarly with them in terms of the homogeneity of functional annotations and structural classes. The three layers of grouping give the opportunity to analyze the gene clusters connected by the increasingly remote homology. Downloadable files, including functional annotations of clusters and HMMs, as well as bi-monthly updates that follow the RefSeq releases, make VOGDB a universal tool for downstream workflows in virus bioinformatics.

VOGDB is in constant development and new knowledge about viruses is quickly implemented (for example the new phage taxonomy [40]). On the other hand, the stable naming of clusters allows for the comparability of the results obtained by different releases of the database.

References

1. Villarreal, L. in *Encyclopedia of Virology* (Elsevier, 2008).
2. Hendrix, R. W., Smith, M. C. M., Burns, R. N., Ford, M. E. & Hatfull, G. F. Evolutionary Relationships among Diverse Bacteriophages and Prophages: All the World's a Phage. *Proceedings of the National Academy of Sciences* (1999).
3. Mushegian, A. R. Are There 10 ^{31} Virus Particles on Earth, or More, or Fewer? *Journal of Bacteriology* (2020).

4. Koonin, E. V., Krupovic, M. & Dolja, V. V. The Global Virome: How Much Diversity and How Many Independent Origins? *Environmental Microbiology* (2023).

5. Krishnamurthy, S. R. & Wang, D. Origins and Challenges of Viral Dark Matter. *Virus Research* (2017).

6. Kuchibhatla, D. B. *et al.* Powerful Sequence Similarity Search Methods and In-Depth Manual Analyses Can Identify Remote Homologs in Many Apparently "Orphan" Viral Proteins. *Journal of Virology* (2014).

7. Stern, A. & Andino, R. in *Viral Pathogenesis* (Elsevier, 2016).

8. Koonin, E. V., Dolja, V. V. & Krupovic, M. The Logic of Virus Evolution. *Cell Host & Microbe* (2022).

9. Koonin, E. V. Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics* (2005).

10. Pearson, W. R. An Introduction to Sequence Similarity ("Homology") Searching. *Current Protocols in Bioinformatics* (2013).

11. Yoon, B.-J. Hidden Markov Models and Their Applications in Biological Sequence Analysis. *Current Genomics* (2009).

12. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): A Resource for Comparative Genomics and Protein Family Annotation. *Nucleic Acids Research* (2017).

13. Huerta-Cepas, J. *et al.* eggNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences. *Nucleic Acids Research* (2016).

14. Terzian, P. *et al.* PHROG: Families of Prokaryotic Virus Proteins Clustered Using Remote Homology. *NAR Genomics and Bioinformatics* (2021).

15. Haft, D. H. *et al.* RefSeq and the Prokaryotic Genome Annotation Pipeline in the Age of Metagenomes. *Nucleic Acids Research* (2024).

16. Li, W. *et al.* RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline Reach with Protein Family Model Curation. *Nucleic Acids Research* (2021).

17. Benson, D. A. *et al.* GenBank. *Nucleic Acids Research* (2018).

18. Boutet, E. *et al.* UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods in Molecular Biology (Clifton, N.J.)* pmid: 26519399 (2016).

19. Chandonia, J.-M. *et al.* SCOPe: Improvements to the Structural Classification of Proteins – Extended Database to Facilitate Variant Interpretation and Machine Learning. *Nucleic Acids Research* (2022).

20. Yost, S. A. & Marcotrigiano, J. Viral Precursor Polyproteins: Keys of Regulation from Replication to Maturation. *Current Opinion in Virology.* pmid: 23602469 (2013).

21. Kristensen, D. M. *et al.* A Low-Polynomial Algorithm for Assembling Clusters of Orthologous Groups from Intergenomic Symmetric Best Matches.

22. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research* (1997).

23. Sievers, F. *et al.* Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Molecular Systems Biology.* pmid: 21988835 (2011).

24. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* (2011).

25. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *Journal of Molecular Biology* (1990).

26. Steinegger, M. *et al.* HH-suite3 for Fast Remote Homology Detection and Deep Protein Annotation. *BMC Bioinformatics* (2019).

27. Van Dongen, S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications* (2008).

28. Burley, S. K. *et al.* RCSB Protein Data Bank (RCSB.Org): Delivery of Experimentally-Determined PDB Structures alongside One Million Computed Structure Models of Proteins from Artificial Intelligence/Machine Learning. *Nucleic Acids Research* (2023).

29. Jumper, J. *et al.* Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* (2021).

30. Van Kempen, M. *et al.* Fast and Accurate Protein Structure Search with Foldseek. *Nature Biotechnology* (2023).

31. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* (2012).

32. Galperin, M. Y. *et al.* COG Database Update: Focus on Microbial Diversity, Model Organisms, and Widespread Pathogens. *Nucleic Acids Research* (2021).

33. Hernández-Plaza, A. *et al.* eggNOG 6.0: Enabling Comparative Genomics across 12 535 Organisms. *Nucleic Acids Research* (2023).

34. Koonin, E. V., Senkevich, T. G. & Dolja, V. V. The Ancient Virus World and Evolution of Cells. *Biology Direct* (2006).

35. Guo, J. *et al.* VirSorter2: A Multi-Classifier, Expert-Guided Approach to Detect Diverse DNA and RNA Viruses. *Microbiome* (2021).

36. Nayfach, S. *et al.* CheckV Assesses the Quality and Completeness of Metagenome-Assembled Viral Genomes. *Nature Biotechnology* (2021).

37. Zhong, C., Edlund, A., Yang, Y., McLean, J. S. & Yooseph, S. Metagenome and Metatranscriptome Analyses Using Protein Family Profiles. *PLOS Computational Biology* (2016).

38. Laffy, P. W. *et al.* Reef Invertebrate Viromics: Diversity, Host Specificity and Functional Capacity. *Environmental Microbiology* (2018).

39. Villarroel, J., Larsen, M., Kilstrup, M. & Nielsen, M. Metagenomic Analysis of Therapeutic PYO Phage Cocktails from 1997 to 2014. *Viruses* (2017).

40. Turner, D. *et al.* Abolishment of Morphology-Based Taxa and Change to Binomial Species Names: 2022 Taxonomy Update of the ICTV Bacterial Viruses Subcommittee. *Archives of Virology* (2023).

Appendix B

Viruses of Great Apes data report

Screening great ape museum specimens for DNA viruses

Authors

Michelle Hämmerle[1,2], Meriam Guellil[1,2], Olivia Cheronet[1,2], Susanna Sawyer[1,2], Irune Ruiz-Gartzia[3], Esther Lizano[4,3], Aigerim Rymbekova[1,2], Pere Gelabert[1,2], Paolo Bernardi[1], Sojung Han[1,2], Lovro Trgovec-Greif[5,6], Thomas Rattei[5], Verena J. Schuenemann[1,2,7,8], Tomas Marques-Bonet[3,4,9,10], Katerina Guschanski[11,12], Sebastien Calvignac-Spencer[13,14], Ron Pinhasi[1,2], Martin Kuhlwilm[1,2]

Affiliations

1. Department of Evolutionary Anthropology, University of Vienna, Djerassiplatz 1, 1030 Vienna, Austria
2. Human Evolution and Archaeological Sciences (HEAS), University of Vienna, Austria
3. Institute of Evolutionary Biology (UPF-CSIC), PRBB, Dr. Aiguader 88, 08003 Barcelona, Spain
4. Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain
5. Centre for Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria
6. Doctoral School of Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria
7. Institute of Evolutionary Medicine, University of Zurich, Zurich, Switzerland
8. Department of Environmental Sciences, University of Basel, Basel, Switzerland
9. National Center for Genomic Analysis (CNAG), Baldiri i Reixac 4, 08028 Barcelona, Spain
10. Institució Catalana de Recerca i Estudis Avançats (ICREA) and Universitat Pompeu Fabra. Pg. Luís Companys 23, 08010, Barcelona, Spain
11. Institute of Ecology and Evolution, School of Biological Sciences, University of Edinburgh, Edinburgh, UK
12. Department of Ecology and Genetics, Animal Ecology, Uppsala University, SE-75236 Uppsala, Sweden
13. Helmholtz Institute for One Health, Helmholtz-Centre for Infection Research (HZI), 17489 Greifswald, Germany
14. Faculty of Mathematics and Natural Sciences, University of Greifswald, 17489 Greifswald, Germany

corresponding author: Martin Kuhlwilm (martin.kuhlwilm@univie.ac.at)

Abstract

Natural history museum collections harbour a record of wild species from the past centuries, providing a unique opportunity to study animals as well as their infectious agents. Thousands of great ape specimens are kept in these collections, and could become an important resource to study the evolution of DNA viruses, whose genetic material is likely to be preserved in dry museum specimens. Here, we screened 209 great ape museum specimens for 99 different DNA viruses, using hybridization capture coupled with short-read high-throughput sequencing. We report a capture design for great ape DNA viruses, sequencing data obtained using this approach, as well as findings regarding the presence of viruses, and several viral genomes obtained from historical specimens.

Background & Summary

The extensive collections of fossils and specimens preserved by natural history museums document the diversity of life forms, ecosystem dynamics, and the transformative processes that have shaped our planet over millions of years. They are a critical resource for scientific research. By employing high-throughput DNA sequencing techniques tailored to ancient or historical specimens, scientists can reconstruct ancient genomes[1] opening a new dimension in the field of museomics. The integration of genomic technologies with traditional museum collections presents unprecedented opportunities to address consequential questions in ecology and evolutionary biology[2,3]. Comparing genomic data from museum specimens with modern populations allows for the reconstruction of a more comprehensive tree of life and facilitates assessments of environmental change impacts on genetic diversity and population structure. This information is essential not only for understanding biodiversity and the forces that have shaped it in the past but also for developing effective conservation strategies to preserve genetic diversity and mitigate anthropogenic disturbances to natural ecosystems[4].

An intriguing avenue of investigation for museomics is the study of infectious diseases[5,6]. Natural history museum specimens have rarely been used for this purpose, but archaeological specimens have demonstrated the immense potential of ancient microbial genetic material in shedding light on the evolution and spread of infectious agents[7]. This is notably true for research on viruses with a DNA genome (DNA viruses). Hundreds to thousands of years old DNA viral genomes reconstructed from archeological specimens have unveiled key aspects of their recent evolutionary histories. For example, researchers have shown complete lineage turn-over of hepatitis B viruses (HBV; family *Hepadnaviridae*) in European human populations around the Bronze Age[8,9]. Genomic pathways to increased host adaptation and virulence have also been clarified for variola viruses (*Poxviridae*) in medieval human populations[10], for Marek's disease virus (*Herpesviridae*) in 19th/20th century poultry[11], or myxoma virus (*Poxviridae*) in rabbits[12].

In parallel, efforts geared toward a better understanding of the determinants of health in our closest relatives, the nonhuman great apes (orangutans – *Pongo pygmaeus*, *P. tapanuliensis* and *P. abelli*, gorillas – *Gorilla gorilla* and *G. beringei*, chimpanzees – *Pan troglodytes* – and bonobos – *Pan paniscus*), have revealed that these species host a broad range of DNA viruses. This includes enzootic viruses belonging to the families *Adenoviridae*, *Anelloviridae*, *Circoviridae*, *Hepadnaviridae*, *Herpesviridae*, *Papillomaviridae*, *Parvoviridae* and *Polyomaviridae*, as well as emerging viruses such as members of the family *Poxviridae*[13,14]. Co-phylogenetic analyses have shown that DNA viruses have often been stably associated with their hominid hosts for extensive periods of time. While co-divergence with their hosts is frequent, rare cross-species transmission events between hominids have contributed to shaping all hominid DNA viromes. For example, the herpes simplex virus 2 (*Herpesviridae*) that causes genital herpes in humans likely arose from the cross-species transmission of a virus infecting members of the gorilla lineage, several million years ago[15]. Conversely, patterns of genomic variation suggest that HBV was transmitted from humans to nonhuman great apes over the last few thousand years[9]. As we learn more about nonhuman great ape viromes, we will uncover more of the complex origins of their DNA viruses and those infecting humans.

Explorations of non-human great ape viromes are ongoing and should proceed in the framework of long-term studies of wild populations, but they could be efficiently complemented by analyses of the thousands of specimens kept in natural history museums around the world. Not only do these collections provide immediate access to animal tissues (as opposed to the noninvasive samples that constitute the bulk of biological sampling from modern populations), they also open a direct window on 200 years of increased anthropogenic disturbance characterized by dwindling, increasingly fragmented populations of these animals[16]. These processes have likely altered nonhuman DNA viromes significantly, and possibly resulted in the extinction of viral lineages[17]. Moreover, emerging viruses constitute a threat to wild great ape populations[14,18], and they are often of human origin[19]. Therefore, studying the DNA viromes of nonhuman great ape museum specimens will provide both information on viruses that still circulate in contemporary populations and document hidden evolutionary trajectories taken by extinct viruses. This may also allow to monitor changes in viromes of wild populations with the potential to intervene should the populations be exposed to novel viruses (e.g. from humans or other species) they have not co-evolved with.

Here, we present high-throughput sequencing data from 209 nonhuman great ape museum specimens which were sampled and analysed in search for DNA viruses. These specimens cover all nonhuman great ape species, most subspecies and large parts of their recent geographical distribution. A majority originated in the wild, but our sample set also includes captive individuals from European zoos (13%, n=28). Obtaining data from ancient specimens can be challenging due to DNA degradation and damage, as well as potential contamination during storage and handling[1]. While museum specimens are younger than typical archaeological remains, high variability in DNA quality has been observed[20], and handling in specific clean laboratory facilities, following strict protocols is necessary[21], and was applied for this study.

In addition, we also developed and implemented an efficient customised enrichment strategy, in-solution hybridization capture with RNA baits[22] to account for the extremely low abundance of viral genetic material. Although metagenomic detection of viruses is possible, it is particularly challenging for DNA viruses that, with a few exceptions, usually do not replicate as intensely as RNA viruses[23]. Hybridization capture offers considerable flexibility, because it allows us to multiplex baits targeting many different viruses, as well as enrich even significantly divergent targets (up to 58% divergence tolerance according to some authors[24]). However, probably many DNA viruses evolve relatively slowly[25]. We designed and used a bait set that covered the genomes of 99 viral lineages representing 13 families. For capture, 8-10 samples were pooled to allow for a more cost-effective screening and sequenced on an Illumina platform.

Analyses of the resulting sequences hints at the presence of multiple DNA viruses in this sample set, including the detection of >500 reads of monkeypox virus (MPXV; family *Poxviridae*) in three orangutans, as well as HBV in two chimpanzees and one gorilla, respectively. For these specimens, the reconstruction of near complete viral genomes was possible. Our work demonstrates the feasibility of viral DNA enrichment and detection from museum specimens of great apes.

Methods

Samples

For this project, we collected a total of 209 great ape specimens (Fig. 1), from which 214 sequencing libraries were produced. Of these libraries, 66 were from gorillas, 84 from chimpanzees (10 of these as *Pan* sp., but most likely *Pan troglodytes* ssp.), 8 from bonobos, and 56 from orangutans, including different subspecies. We note that for five specimens two separate extracts and sequencing libraries were prepared. The samples were obtained from specimens housed in European natural history museums, namely in Germany in Berlin (n = 28), Bonn (n = 28), Dresden (n = 26), Frankfurt (n = 63) and Stuttgart (n = 6), in the Czech Republic in Prague (n = 24), and in Austria in Salzburg (n = 22) and Vienna (n = 12). Approximately 92% of libraries were obtained from teeth (n = 196), 17 libraries from soft tissues (Table S1), and two specimens were phalanges. According to the museum metadata, the oldest specimen was from 1838, and the most recent (a captive individual) was from 2014. Some individuals were held in captivity, mostly in zoos, whereas others were wild-caught. More detailed information concerning the individuals and the libraries can be found in Table S1.

The museum identifiers of the specimens are as follows: Senckenberg Forschungsinstitut und Naturmuseum Frankfurt/M.: 10325, 1110, 1111, 1112, 1113, 1114, 1115, 1118, 1119, 1120, 1121, 1126, 1132, 1134, 1576, 1579, 15792, 15817, 16180, 17826, 17961, 24510, 2495, 2538, 2638, 2639, 2654, 3221, 4103, 4104, 4106, 4107, 4108, 4109, 45713, 5277, 5532, 59140, 59147, 59158, 59296, 59297, 59298, 59299, 59301, 59303, 59304, 6716, 6779, 6782, 6785, 6992, 89780, 89781, 92953, 94796, 94797, 94799, 96255, 96550, 97029, 97143, ZIH9; Naturhistorisches Museum Vienna: NMW 1779, NMW 20516, NMW 25124, NMW 3081, NMW 3105/ST 663, NMW 3106, NMW 3107, NMW 3111/ST 665, NMW 3119, NMW 3948, NMW 7136, NMW 793/ST 1647; Přírodovědecké muzeum Prague: NMP 09605, NMP 10588, NMP 10784, NMP 22891, NMP 22892, NMP 22893, NMP 23283, NMP 23284, NMP 23295, NMP 23296, NMP 23297, NMP 24474, NMP 24475, NMP 46815, NMP 46816, NMP 46816-b, NMP 47007, NMP 47656, NMP 49711, NMP 50432, NMP 94205, NMP 94564, NMP 94957, NMP 95098; Senckenberg Naturhistorische Sammlungen Dresden: MTD B11877, MTD B12034, MTD B12062, MTD B12099, MTD B12101, MTD B12177, MTD B12178, MTD B1384-A.S. 1289, MTD B14244, MTD B15789, MTD B1607-A.S. 1690, MTD B247-A.S. 216, MTD B249-A.S. 214, MTD B251-A.S. 231, MTD B253-A.S. 221, MTD B266-A.S. 211, MTD B281-A.S. 239, MTD B287-A.S. 200, MTD B288-A.S. 198, MTD B3686, MTD B4188, MTD B4786, MTD B4788, MTD B4789, MTD B4793, MTD B61-A.S. 244; Museum für Naturkunde Berlin: ZMB_Mam_108652, ZMB_Mam_11637, ZMB_Mam_11638, ZMB_Mam_12799, ZMB_Mam_14644, ZMB_Mam_17011, ZMB_Mam_24838, ZMB_Mam_30755, ZMB_Mam_31617, ZMB_Mam_31621, ZMB_Mam_37523, ZMB_Mam_45130, ZMB_Mam_48173, ZMB_Mam_83519, ZMB_Mam_83522, ZMB_Mam_83547, ZMB_Mam_83606, ZMB_Mam_83607, ZMB_Mam_83617, ZMB_Mam_83642, ZMB_Mam_83643, ZMB_Mam_83647, ZMB_Mam_83648, ZMB_Mam_83653, ZMB_Mam_83675, ZMB_Mam_83681, ZMB_Mam_83682, ZMB_Mam_83685; Museum Koenig Bonn: ZFMK_MAM1938-0136, ZFMK_MAM1957-0003, ZFMK_MAM1957-0004, ZFMK_MAM1962-0131, ZFMK_MAM1963-0660, ZFMK_MAM1965-0544, ZFMK_MAM1965-0545, ZFMK_MAM1965-0546, ZFMK_MAM1965-0547, ZFMK_MAM1965-0550, ZFMK_MAM1976-0410, ZFMK_MAM1994-0482, ZFMK_MAM1997-0070, ZFMK_MAM1997-0076, ZFMK_MAM2012-0036, ZFMK_MAM2015-0479, ZFMK_MAM2019-0404, ZFMK_MAM2019-0405, ZFMK_MAM2019-0407, ZFMK_MAM2019-0408, MAM2019-0410, ZFMK_MAM2019-0415, ZFMK_MAM2019-0416, ZFMK_MAM2019-0417, ZFMK_MAM2019-0418, ZFMK_MAM2019-0419, ZFMK_MAM2019-0420, ZFMK_MAM2019-0421; Haus der Natur Salzburg: HNS-Mam-S-0073, HNS-Mam-S-0075, HNS-Mam-S-0076, HNS-Mam-S-0077, HNS-Mam-S-0078, HNS-Mam-S-0079, HNS-Mam-S-0082, HNS-Mam-S-0084, HNS-Mam-S-0085, HNS-Mam-S-0086, HNS-Mam-S-0519, HNS-Mam-S-0524, HNS-Mam-S-0525, HNS-Mam-S-0530, HNS-Mam-S-0531, HNS-Mam-S-0532, HNS-Mam-S-0533, HNS-Mam-S-0534, HNS-Mam-S-0535, HNS-Mam-S-0536, HNS-Mam-S-0550, HNS-Mam-S-0742; Staatliches Museum für Naturkunde Stuttgart: SMNS-Z-MAM-001687, SMNS-Z-MAM-001750, SMNS-Z-MAM-002012, SMNS-Z-MAM-045995, SMNS-Z-MAM-046000, SMNS-Z-MAM-048948.

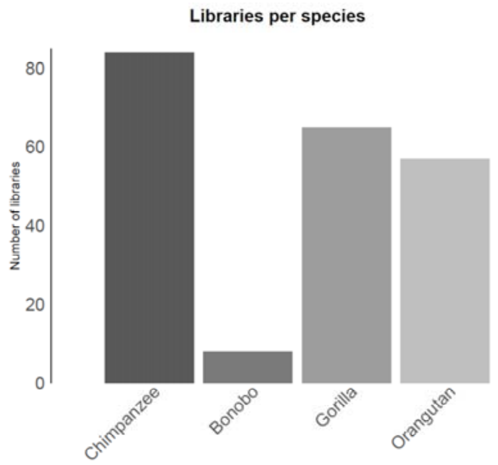


Figure 1. Number of libraries per great ape species included in this study.

DNA extraction and library preparation

All steps from grinding to indexing except the qPCR were performed in laboratories designed and dedicated only to ancient DNA (aDNA) research while wearing protective clothing and following aDNA laboratory best practices. DNA was directly extracted from soft tissue. Bone and teeth were treated with a sandblaster and ground to bone powder using a *MixerMill* (Retsch). For each specimen, 50 mg of powder was collected. DNA was extracted using an established protocol used for aDNA[26]. Single-stranded DNA libraries were prepared[27], followed by a clean-up with the *QIAGEN MinElute PCR Purification Kit*. We performed a quantitative PCR for calculating the cycle number in the indexing PCR. Indexing was performed in quadruplicates using *NEBNext Q5U*, followed by a clean-up using the *NucleoMag® NGS Clean-up and Size Select* kit. Indexes are listed in Table S1. We assessed quantity and quality using an *Invitrogen QubitTM 4 Fluorometer* and an *Agilent 4150 TapeStation*. The grinding step for all samples was performed in the Vienna Ancient DNA laboratory of the University of Vienna, while the DNA extraction, library and QC steps were performed for a subset of libraries (n=97, Table S1) in the ancient DNA laboratory at the Universitat Pompeu Fabra in Barcelona, following the exact same protocols and best practices.

To maximise economic efficiency, we constructed pools of 8-10 libraries, except of four, where the specimen condition might have been influenced by a pathological condition of the individual. If DNA concentration was below 5ng/µl per sample, we performed another amplification, to preserve sufficient amounts of library. If the concentration was above 25 ng/µl, a dilution was required. Between 8 and 10 libraries were pooled by equal concentration, whereby it was crucial to avoid pooling those with overlapping P5 or P7 adapters. In total, 210 libraries were pooled in 24 pools. We used the same concentration threshold for un-pooled libraries.

Hybridization capture

As aDNA or historical DNA does not only contain host and host-associated microbiome DNA, but often an overwhelming abundance of bacterial and environmental DNA, shotgun sequencing is usually economically infeasible for viruses[28], and target-specific approaches can help in enrichment of sequences of interest[21]. We designed a capture set containing 99 different viruses from 13 families, whereby 49 are human-infecting, 18 great ape-infecting, and the remaining were isolated from other primate species. The design was based on reference genomes available in NCBI, and commercially produced by *myBaits®*, where a BLAST search against the human reference genome was performed to exclude sequences with any hits to it; baits were designed to be 80 nucleotides long with a 2X tiling. The viruses and the NCBI reference sequence are reported in Table S2. We followed the protocol for the capture provided by the manufacturer (Version 5.03, as ordered in July 2021). Briefly, after adding the blockers and the hybridization mix with the RNA baits, the libraries were incubated for approximately 40 hours at 60°C, to increase the efficiency of the capture reaction and to allow for higher sequence deviation from the baits. DNA was eluted from the beads in 30 µl Buffer E, and the supernatant was kept. A qPCR of the capture product was performed in order to estimate the yield, and another PCR to amplify the capture product. Libraries were pooled to 20ng total DNA, and single-end sequencing was performed on the *Illumina NovaSeq 6000 (SP SR100 XP workflow)* at the Vienna BioCenter. The targeted amount of sequencing reads per library was around one million reads.

Bioinformatic processing

Adapters were trimmed from the fastq files with trimmomatic[29] (version 0.39)**,** and BBmap (version 39.01) clumpify was used to remove duplicates introduced by PCR amplification[30]. To determine the metagenomic composition of the sequenced libraries, taxonomic classification via Kraken2 using the standard database was performed[31]. This database contains the strains included in the capture design. Heatmaps were plotted via a customized python3 script for the target taxa at species and genus levels.

Where Kraken2 assigned more than 500 reads to one of the reference genomes included in the capture kit, we performed a mapping with BWA[32] (version 0.7.17), using bwa aln (with parameters "-n 0.04 -l 1000"). Mapping coverage along the reference genome was visualized using aDNA-BAMplotter[33], and inspected individually. In case of unequal coverage along the genome, we performed a literature search for alternative genomes obtained from great apes. Summary statistics were calculated using these best reference genomes, including edit distance, mapping quality, mapping quality ratio, and the percentage for 1-, 2- , and 10-fold coverage, using a customized python3 script. Other figures were created with the R package ggplot2[34].

Data Records

A custom hybridization capture design for great ape DNA viruses is reported in this publication. We report the bait design for 99 virus strains of potential relevance to great apes. The final design can be found as Supplementary Material SM2, the underlying NCBI identifiers are reported in Table S2, and the NCBI sequences used in fasta format as Supplementary Material SM1. Raw sequencing data after capture for the 214 individual libraries has been uploaded to the Short Read Archive under the accession ID PRJEB75038.

Technical Validation

Sequencing data

The median number of raw reads per library was 1,122,746 reads, with a high variability of yield (SD=3,556,238). After adapter trimming, a median of 956,296 reads was retained, with a reduction due to removal of adapter-multimers between 0.11 and 93.38%. The rates of duplicated reads ranged between 21.35% and 88.67%, leaving a median of 338,560 high-quality unique reads across the 214 libraries (SD=971,276; Fig. 2A-B).

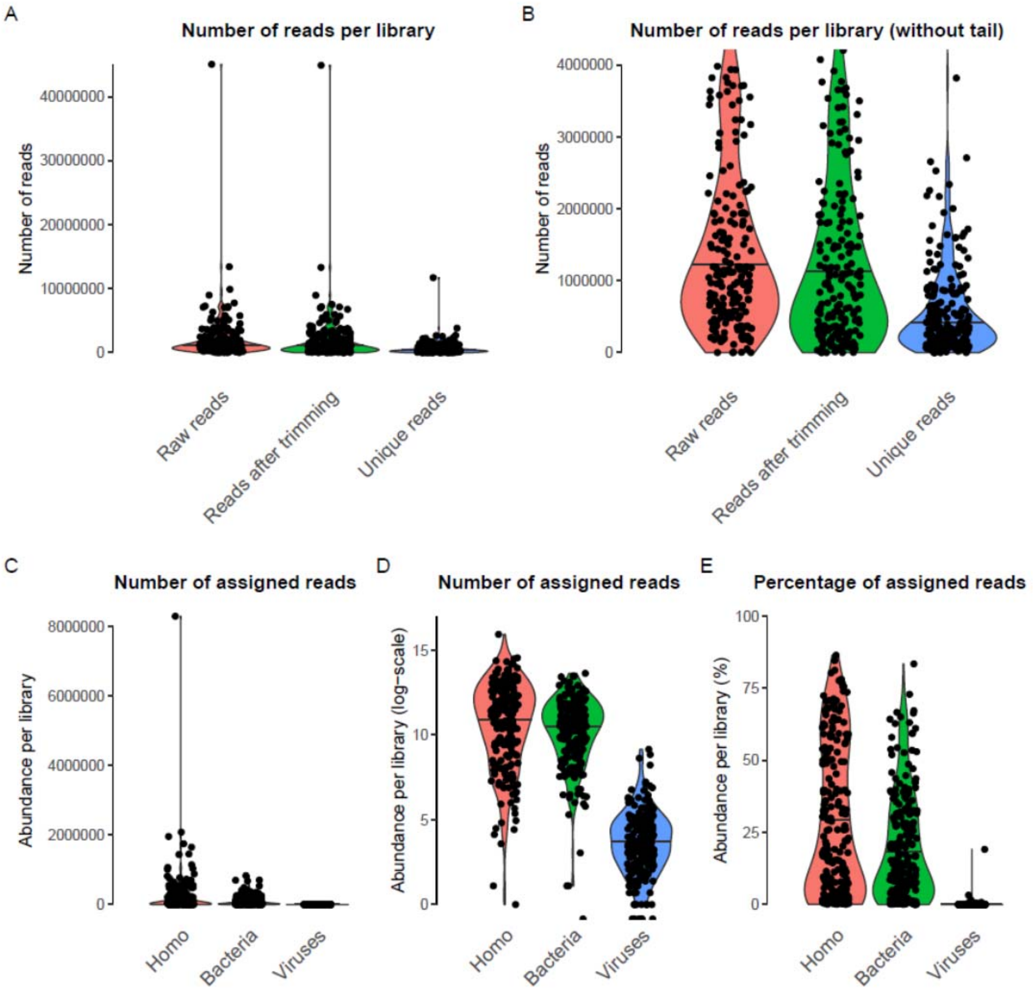


Figure 2: Distribution of sequencing reads per library. A) Numbers of raw reads, reads after trimming and unique reads (after BBmap clumpify) across 214 libraries. B) Same as in A) cut at 4 million reads (without tail). C) Number of reads assigned to *Homo*, Bacteria, or Viruses using Kraken2 per library. D) Same as C), but number of reads on log-scale. E) Percentage of assigned reads per library.

The sequence length of the raw reads was 100 bp, as defined by the sequencing setup, and after filtering and trimming, the median fragment length per library ranged from 40 to 100 bp. This suggests that fragmentation typical for ancient DNA[1] may not have been present across all libraries, as might be expected for DNA extracted from relatively recent museum specimens, and after the enrichment performed here.

Results from virus capture

Fragments from targeted virus fragments are likely absent or very rare in the libraries. Given the low expected abundance of any viral DNA, a large degree of amplification led to high duplication rates (see above). Furthermore, as we chose a lower hybridization temperature and a prolonged incubation time for the already highly amplified libraries, more unspecific binding resulted from our approach. We observed a median assignment of 21.95% to *Homo sapiens*, 13.91% to bacterial sequences and solely 0.1% to viruses (Fig. 2C-E). Mapping to human DNA likely reflected endogenous great ape DNA (and possibly human contamination), and bacterial DNA likely resulted from post-mortem colonization of the specimen. While many viral reads were assigned to bacteriophages (viruses infecting bacteria, not animals), many libraries contained at least one read assigned to a viral family known to infect animals (Table S3, Fig. 3A-B). Numerous libraries contained at least a small number of reads (25 or more) assigned to a viral family (Fig. 3C). A surprising amount of libraries apparently comprised poxvirus reads, which likely represented wrong assignments. When restricting to virus strains that were included in the capture design, we observe fewer instances (n=30) of libraries with potential virus fragments (Fig. 4), which likely reflects are more accurate picture.

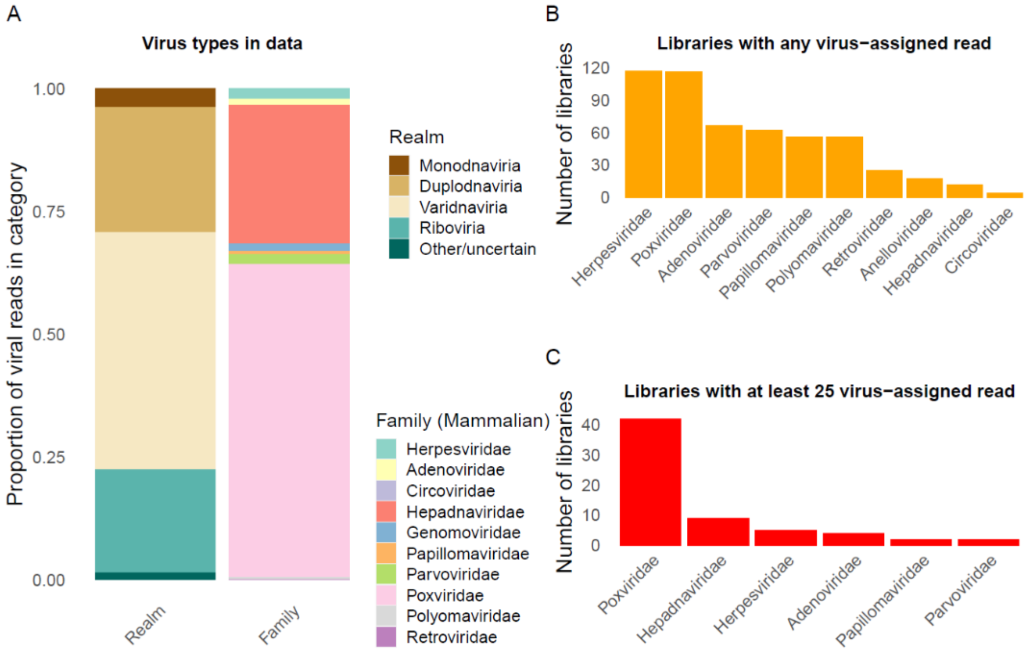


Figure 3. Summaries of reads assigned to different virus domains and families. A) Proportion of virus-assigned reads based on kraken2 across 214 libraries, stratified by realm and by family (for DNA viruses and *Retroviridae*). B) Number of libraries with any read assigned to virus families. C) Number of libraries with at least 25 reads assigned to virus families.

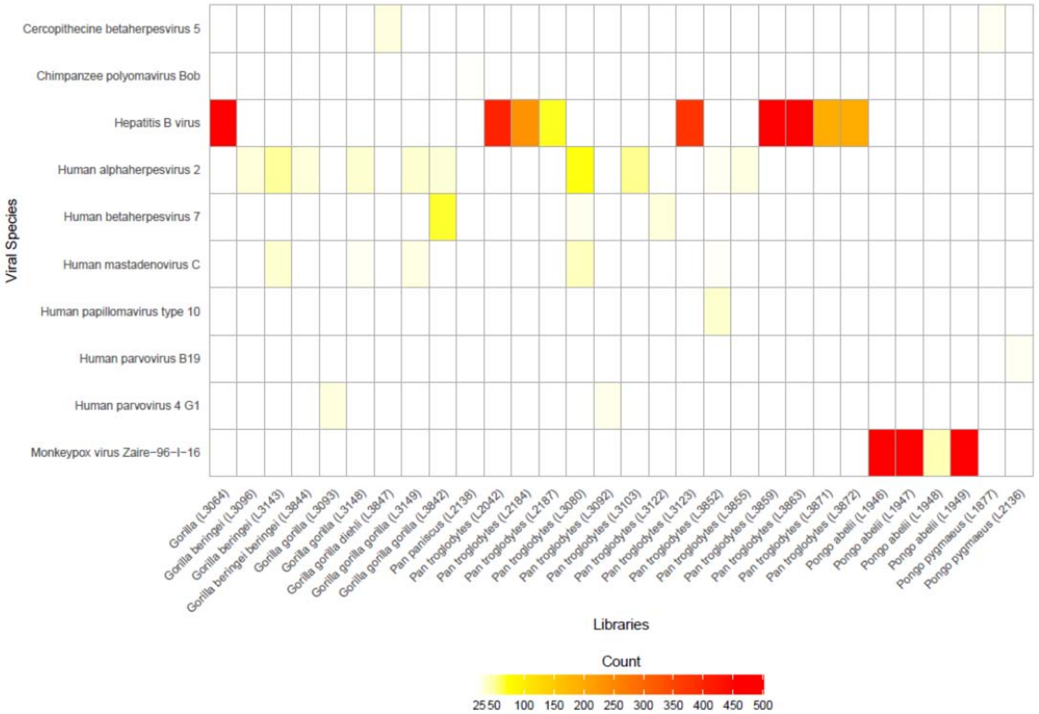


Figure 4: Heatmap of positive reads per viral species. The figure depicts all libraries that have at least 25 hits assigned to one of the 99 viruses in the capture kit (with merged numbers for Cytomegalovirus strains).

Henceforth, we focused on libraries with more than 500 assigned reads, which are the most likely to reflect true viral infection. We found six such instances, with assignments either to the family *Poxviridae* (n=3; all orangutan specimens) or the family *Hepadnaviridae* (n=3; 1 gorilla and 2 chimpanzee specimens). The only library with more viral than bacterial or human-mapping sequences is L1949 (19.13% viral reads), obtained from an orangutan specimen. We note that further rounds of enrichment capture might have provided higher on-target coverage[1,35], but at substantially higher costs.

Positive specimens

To validate the results from classifying reads, we attempted to reconstruct the corresponding viral genomes. Using reference-based mapping, we managed to obtain low-coverage monkeypox virus genomes from three libraries (up to 3.6-fold coverage on the reference genome sequence KJ642614) (Table 1). Further investigation of these specimens, including deeper sequencing to obtain high-coverage genomes, was published in another study[36], where we present their origin from a zoo outbreak in 1965. Similar efforts aimed at assembling HBV genomes from the three most promising HBV-positive specimens resulted in one high-coverage genome from a wild gorilla, and two from wild chimpanzees (Table 1). In other samples, putative viral sequences were at much lower abundance (Table S3).

Our data shows that a sizeable fraction of museum great ape specimens is amenable to the recovery of whole DNA virus genomes. This resource could contribute to a much better understanding of recent viral evolution in our closest living relatives.

Library ID (genus)	Individual ID (museum)	Reference	Mapped Reads	Mean ED	Mean DOC	1X	5X	10X
L1946	ZFMK_MAM1965-	KJ642614	7,931	0.51	3.83	93.04	33.25	2.58

(*Pongo*)	0546 (Bonn)	(monkeypox virus)						
L1947 (*Pongo*)	ZFMK_MAM1965-0545 (Bonn)	KJ642614 (monkeypox virus)	4,659	0.5	2.29	84.45	12.04	0.09
L1949 (*Pongo*)	ZFMK_MAM1965-0546 (Bonn)	KJ642614 (monkeypox virus)	2,352	0.61	1.13	63.83	1.03	0
L3064 (*Gorilla*)	96550 (Frankfurt)	AJ131567 (hepatitis B virus)	2,969	0.82	97.07	99.62	99.56	99.43
L3863 (*Pan*)	ZMB_Mam_83617 (Berlin)	ON706349 (hepatitis B virus)	517	1.82	18.19	94.44	73.60	53.80
L3859 (*Pan*)	ZMB_Mam_11638 (Berlin)	AF305327 (hepatitis B virus)	919	1.98	31.27	92.36	75.83	61.56

Table 1. Statistics for virus genomes identified in this study. Mapping metrics for the six viral genomes discovered in this study, including proportion of the reference genome covered by at least 1, 5 or 10 reads. Mapped reads = unique reads with MQ >30. DOC = Depth of coverage. ED = Edit distance.

Usage Notes

Data can be reprocessed using the tools described in the Methods section, namely kraken2 for metagenomic classification, bwa for mapping to reference genomes, following the documentation in the associated repository. Custom code for processing and visualisation is provided below.

Code Availability

The code used is available under https://github.com/admixVIE/Great-Ape-DNA-Virome.

Acknowledgements

The computational results of this work have been achieved using the Life Science Compute Cluster (LiSC) of the University of Vienna. We are grateful to the Zoologisches Forschungsmuseum A. Koenig, Leibniz-Institut zur Analyse des Biodiversitätswandels in Bonn, in particular Eva Bärmann, Jan Decher, and Christian Montermann at the Section Theriology; to Irina Ruf and Katrin Krohmann at the Mammalogy collection at Senckenberg Forschungsinstitut und Naturmuseum Frankfurt/M.; to Frank Zachos and Alexander Bibl at Naturhistorisches Museum Wien; to Stefan Merker at Zoology at Staatliches Museum für Naturkunde Stuttgart; to Petr Benda at Department of Zoology at National Museum (Natural History) Prague; to Clara Stefen and Jens Jakobitz at Mammalogie at Senckenberg Naturhistorische Sammlungen Dresden; to Robert Lindner at Haus der Natur – Museum für Natur und Technik in Salzburg; to Christiane Funk and Frieder Mayer at the mammalian collection at Museum für Naturkunde/Leibniz Institute for Evolution and Biodiversity research in Berlin.

This project has been funded by the Vienna Science and Technology Fund (WWTF) [10.47379/VRG20001] and by the Austrian Science Fund (FWF) [FW547002] to M.K. K.G. received support from the Swedish Research Council (VR) through grant 2020-03398.

Author contributions

M.K. conceived the topic of the study, supervised data analysis and wrote the manuscript. M.H. performed experiments and data analysis and wrote the manuscript. R.P. conceived the study and supervised the experimental work. M.G. supervised data analysis. S.C.-S. wrote the manuscript. S.S., E.L. and O.C. supervised experimental work. I. R.-G. performed experiments and contributed to the design of the dataset. P.B. performed experiments. A.R., P.B. and L.G. provided help in analysing data. P.G., T.R., V.J.S., T.M.-B. and K.G. provided help in writing the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Orlando, L. *et al.* Ancient DNA analysis. *Nat. Rev. Methods Prim.* **1**, 14 (2021).
2. Raxworthy, C. J. & Smith, B. T. Mining museums for historical DNA: advances and challenges in museomics. *Trends Ecol. Evol.* **36**, 1049–1060 (2021).
3. Holmes, M. W. *et al.* Natural history collections as windows on evolutionary processes. *Mol. Ecol.* **25**, 864–881 (2016).
4. Blair, M. E. Conservation museomics. *Conserv. Biol.* **n/a**, (2023).
5. Cook, J. A. *et al.* Integrating Biodiversity Infrastructure into Pathogen Discovery and Mitigation of Emerging Infectious Diseases. *Bioscience* **70**, 531–534 (2020).
6. DiEuliis, D., Johnson, K. R., Morse, S. S. & Schindel, D. E. Specimen collections should have a much bigger role in infectious disease research and response. *Proc. Natl. Acad. Sci.* **113**, 4–7 (2016).
7. Duchêne, S., Ho, S. Y. W., Carmichael, A. G., Holmes, E. C. & Poinar, H. The Recovery, Interpretation and Use of Ancient Pathogen Genomes. *Curr. Biol.* **30**, R1215–R1231 (2020).
8. Mühlemann, B. *et al.* Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature* **557**, 418–423 (2018).
9. Arthur, K. *et al.* Ten millennia of hepatitis B virus evolution. *Science (80-.).* **374**, 182–188 (2021).
10. Mühlemann, B. *et al.* Diverse variola virus (smallpox) strains were widespread in northern Europe in the Viking Age. *Science (80-.).* **369**, eaaw8977 (2020).
11. Fiddaman, S. R. *et al.* Ancient chicken remains reveal the origins of virulence in Marek's disease virus. *Science (80-.).* **382**, 1276–1281 (2023).
12. Alves, J. M. *et al.* Parallel adaptation of rabbit populations to myxoma virus. *Science (80-.).* **363**, 1319–1326 (2019).
13. Calvignac-Spencer, S., Düx, A., Gogarten, J. F., Leendertz, F. H. & Patrono, L. V. Chapter One - A great ape perspective on the origins and evolution of human viruses. in (eds. Kielian, M., Mettenleiter, T. C. & Roossinck, M. J. B. T.-A. in V. R.) vol. 110 1–26 (Academic Press, 2021).
14. Calvignac-Spencer, S., Leendertz, S. A. J., Gillespie, T. R. & Leendertz, F. H. Wild great apes as sentinels and sources of infectious disease. *Clin. Microbiol. Infect.* **18**, 521–527 (2012).
15. Wertheim, J. O. *et al.* Discovery of Novel Herpes Simplexviruses in Wild Gorillas,

Bonobos, and Chimpanzees Supports Zoonotic Origin of HSV-2. *Mol. Biol. Evol.* **38**, 2818–2830 (2021).

16. van der Valk, T., Díez-del-Molino, D., Marques-Bonet, T., Guschanski, K. & Dalén, L. Historical Genomes Reveal the Genomic Consequences of Recent Population Decline in Eastern Gorillas. *Curr. Biol.* **29**, 165-170.e6 (2019).

17. Beatrix, K. *et al.* Local Virus Extinctions following a Host Population Bottleneck. *J. Virol.* **89**, 8152–8161 (2015).

18. Fontsere, C. *et al.* The genetic impact of an Ebola outbreak on a wild gorilla population. *BMC Genomics* **22**, 735 (2021).

19. Tan, C. C. S., van Dorp, L. & Balloux, F. The evolutionary drivers and correlates of viral host jumps. *Nat. Ecol. Evol.* (2024) doi:10.1038/s41559-024-02353-4.

20. van der Valk, T., Lona Durazo, F., Dalén, L. & Guschanski, K. Whole mitochondrial genome capture from faecal samples and museum-preserved specimens. *Mol. Ecol. Resour.* **17**, e111–e121 (2017).

21. Spyrou, M. A., Bos, K. I., Herbig, A. & Krause, J. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat. Rev. Genet.* **20**, 323–340 (2019).

22. Furtwängler, A. *et al.* Comparison of target enrichment strategies for ancient pathogen DNA. *Biotechniques* **69**, 455–459 (2020).

23. Moustafa, A. *et al.* The blood DNA virome in 8,000 humans. *PLOS Pathog.* **13**, e1006292 (2017).

24. Wylie, T. N., Wylie, K. M., Herter, B. N. & Storch, G. A. Enhanced virome sequencing using targeted sequence capture. *Genome Res.* **25**, 1910–1920 (2015).

25. Patterson Ross, Z. *et al.* The paradox of HBV evolution as revealed from a 16th century mummy. *PLOS Pathog.* **14**, e1006750 (2018).

26. Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci.* **110**, 15758–15763 (2013).

27. Kapp, J. D., Green, R. E. & Shapiro, B. A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. *J. Hered.* **112**, 241–249 (2021).

28. Gaudin, M. & Desnues, C. Hybrid Capture-Based Next Generation Sequencing and Its Application to Human Infectious Diseases. *Front. Microbiol.* **9**, (2018).

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

30. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner.

31. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

32. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

33. Guellil, M. MeriamGuellil/aDNA-BAMPlotter: aDNA-BAMPlotter. at https://doi.org/10.5281/zenodo.5676093 (2021).

34. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2009).

35. Fontsere, C. *et al.* Maximizing the acquisition of unique reads in non-invasive capture sequencing experiments. *Mol. Ecol. Resour.* 1755–0998.13300 (2020) doi:10.1111/1755-0998.13300.

36. Hämmerle, M. *et al.* Link between Monkeypox Virus Genomes from Museum Specimens and 1965 Zoo Outbreak. *Emerg. Infect. Dis. J.* **30**, 815 (2024).