Technische Universität München
TUM School of Life Sciences

Computational Analysis of the Role of Bacteriophages in Environment and Health

Jinlong Ru

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Aurélien Tellier

Prüfer*innen der Dissertation:

1. Prof. Dr. Li Deng
2. Dr. Johannes Söding
3. Prof. Dr. Christian L. Müller

Die Dissertation wurde am 05.05.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 12.11.2023 angenommen.

Abstract

Viruses, particularly bacteriophages (or phages), are Earth's most abundant biological entities. They can infect nearly all organisms and serve as a global reservoir of genetic diversity. Bacteriophages play a crucial role in bacterial-phage co-evolution, shaping the genetic makeup of bacterial populations over time, regulating microbial communities in both the human body and natural environments, and potentially influencing the development of certain diseases. However, bacteriophage research has faced significant obstructions due to labor-intensive, cultivation-dependent isolation processes. Recently, viral metagenomics (viromics) has experienced unprecedented growth, driven by advancements in high-throughput sequencing technologies and computational methods, allowing researchers to bypass traditional wet-lab isolation steps. Novel software and databases have facilitated a deeper understanding of viral communities and their interactions with hosts. However, managing and ensuring the reproducibility of data analysis generated by these studies presents significant challenges, which arise due to the large-scale nature of the datasets, variations in computational platforms, differences in software and database versions, and the absence of an easy-to-use, comprehensive data analysis pipeline.

My Ph.D. projects aimed to address these challenges. First, I developed ViroProfiler, a containerized bioinformatics pipeline designed for scalable, reusable, and shareable analysis of viromic sequencing data. This user-friendly platform facilitates reproducible research by providing a standardized framework for data processing and analysis. It generates comprehensive results from raw sequencing data, enabling deeper insights into community structures, the behavior and characteristics of specific viral taxa, and the genetic elements within individual viral contigs.

Next, I demonstrated the power of this comprehensive virome analysis pipeline through a clinical study. Collaborating with wet lab scientists, we investigated the role of gut bacteriophages in Barrett's esophagus (BE) and esophageal adenocarcinoma (EAC). Our findings illuminate the complex interplay between viral entities and host bacteria, revealing distinct virome structures associated with different disease statuses and potential relationships between community structure shifts and disease development. Furthermore, we identified toxin-related auxiliary metabolic genes (AMGs) that were more abundant in disease groups, highlighting the potential role of phages in disease development.

Beyond human health, we also explored the role of bacteriophages in environmental contexts. Our primary focus was on identifying and characterizing virus-encoded hydrocarbon degradation genes (vHYDEGs). Given the unique nature of environmental viromic data and the research purpose, we used additional informatics tools to analyze metagenomic viral contigs obtained from a public database. Re-analyzing these viral contigs enabled us to discover previously unexplored viral functions related to the remediation of long-chain hydrocarbon-polluted environments. From the identified high quality vHYDEGs, six protein families were proved to be involved in the crucial steps of alkane degradation. Our findings provide further evidence of the unexplored contributions of bacteriophages to global carbon cycling.

In addition, this viromic data analysis pipeline also contributed to large-scale studies, including cohort studies. We characterized gut phages in allogeneic stem cell transplantation patients and identified an association between phage-encoded auxiliary metabolic genes (AMGs) and protective immuno-modulatory metabolites in the human gut. In a *Helicobacter pylori*-induced colorectal cancer (CRC) mouse model, we examined the gut virome and discovered strong phage-bacteria linkages in the initial stage of CRC. Additionally, we investigated the functional role of bacteriophages in childhood stunting. Building on these diverse applications of viromic analysis, I presented an extensive review of state-of-the-art bioinformatics methods employed in viromic studies, as well as my prospects.

Overall, this study presents an optimized pipeline, ViroProfiler, which integrates state-of-the-art workflow management system and software to offer a comprehensive range of options for viromic data analysis. By employing this pipeline, we were able to explore massive sequencing data obtained from both human and environmental viromic samples. Our collective findings provide valuable insights into the behavior and function of bacteriophages and their encoded genes in diseases and natural environments.

Zusammenfassung

Viren, insbesondere Bakteriophagen (oder Phagen), sind die häufigsten biologischen Einheiten auf der Erde. Sie können nahezu alle Organismen infizieren und dienen als globales Reservoir genetischer Vielfalt. Bakteriophagen spielen eine entscheidende Rolle in der Bakterien-Phagen-Koevolution und prägen im Laufe der Zeit das genetische Profil bakterieller Populationen. Sie regulieren mikrobielle Gemeinschaften sowohl im menschlichen Körper als auch in natürlichen Umgebungen und beeinflussen möglicherweise die Entstehung bestimmter Krankheiten. Die Bakteriophagenforschung ist jedoch aufgrund arbeitsintensiver, kultivierungsabhängiger Isolationsprozesse auf erhebliche Hindernisse gestoßen. In jüngster Zeit hat die virale Metagenomik (Viromik) dank Fortschritten in Hochdurchsatz-Sequenzierungstechnologien und rechnergestützten Verfahren ein beispielloses Wachstum erfahren, das es Forschern ermöglicht, traditionelle Nasslaborschritte zu umgehen. Neue Software und Datenbanken haben ein tieferes Verständnis von viralen Gemeinschaften und ihren Wechselwirkungen mit Wirten ermöglicht. Die Verwaltung und Gewährleistung der Reproduzierbarkeit von Datenanalysen, die von diesen Studien generiert werden, stellt jedoch erhebliche Herausforderungen dar, die aufgrund der groß angelegten Natur der Datensätze, Variationen in rechnergestützten Plattformen, Unterschieden in Software- und Datenbankversionen und dem Fehlen einer benutzerfreundlichen, umfassenden Datenanalysepipeline entstehen.

Meine Doktorarbeiten zielten darauf ab, diesen Herausforderungen zu begegnen. Zunächst entwickelte ich ViroProfiler, eine containerisierte Bioinformatik-Pipeline, die für skalierbare, wiederverwendbare und teilbare Analyse von Viromik-Sequenzdaten konzipiert ist. Diese benutzerfreundliche Plattform erleichtert reproduzierbare Forschung, indem sie einen standardisierten Rahmen für die Datenverarbeitung und -analyse bereitstellt. Sie erzeugt umfassende Ergebnisse aus Rohsequenzdaten und ermöglicht tiefere Einblicke in Gemeinschaftsstrukturen, das Verhalten und die Eigenschaften spezifischer viraler Taxa sowie die genetischen Elemente innerhalb einzelner viraler Contigs.

Als Nächstes demonstrierte ich die Leistungsfähigkeit dieser umfassenden Virom-Analysepipeline in einer klinischen Studie. In Zusammenarbeit mit Wissenschaftlern aus dem Nasslabor untersuchten wir die Rolle von Darmbakteriophagen bei Barrett-Ösophagus (BE) und Ösophagus-

Adenokarzinom (EAC). Unsere Ergebnisse beleuchten das komplexe Zusammenspiel zwischen viralen Entitäten und Wirtsbakterien und zeigen unterschiedliche Viromstrukturen, die mit verschiedenen Krankheitsstatus und möglichen Beziehungen zwischen Gemeinschaftsstrukturverschiebungen und Krankheitsentwicklung in Zusammenhang stehen. Darüber hinaus identifizierten wir toxinbezogene zusätzliche Stoffwechselgene (AMGs), die in Krankheitsgruppen häufiger vorkamen und die potenzielle Rolle von Phagen in der Krankheitsentwicklung hervorhoben.

Jenseits der menschlichen Gesundheit untersuchten wir auch die Rolle von Bakteriophagen in Umweltkontexten. Unser Hauptaugenmerk lag auf der Identifizierung und Charakterisierung von viruskodierten Hydrokohlenabbau-Genen (vHYDEGs). Angesichts der einzigartigen Beschaffenheit von Umweltviromik-Daten und des Forschungszwecks nutzten wir zusätzliche Informatikwerkzeuge, um metagenomische virale Contigs aus einer öffentlichen Datenbank zu analysieren. Durch die erneute Analyse dieser viralen Contigs konnten wir bisher unerforschte virale Funktionen im Zusammenhang mit der Sanierung von langkettigen Kohlenwasserstoff-verschmutzten Umgebungen entdecken. Aus den identifizierten hochwertigen vHYDEGs wurden sechs Protein-Familien als an den entscheidenden Schritten des Alkanabbaus beteiligt erwiesen. Unsere Ergebnisse liefern weitere Belege für die unerforschten Beiträge von Bakteriophagen zum globalen Kohlenstoffkreislauf.

Darüber hinaus trug diese Viromik-Datenanalysepipeline auch zu groß angelegten Studien bei, einschließlich Kohortenstudien. Wir charakterisierten Darmphagen bei Patienten mit allogener Stammzelltransplantation und identifizierten eine Assoziation zwischen phagenkodierten zusätzlichen Stoffwechselgenen (AMGs) und schützenden immuno-modulatorischen Metaboliten im menschlichen Darm. In einem Helicobacter pylori-induzierten kolorektalen Krebs (CRC) Mausmodell untersuchten wir das Darmvirom und entdeckten starke Phagen-Bakterien-Verbindungen im Anfangsstadium von CRC. Zusätzlich untersuchten wir die funktionelle Rolle von Bakteriophagen bei Wachstumsverzögerungen im Kindesalter. Aufbauend auf diesen vielfältigen Anwendungen der Viromik-Analyse präsentierte ich eine umfassende Übersicht über modernste bioinformatische Methoden, die in Viromik-Studien eingesetzt werden, sowie meine Aussichten.

Insgesamt stellt diese Studie eine optimierte Pipeline, ViroProfiler, vor, die ein modernes Workflow-Management-System und Software integriert, um eine umfassende Palette von Optionen für die Viromik-Datenanalyse zu bieten. Durch den Einsatz dieser Pipeline konnten wir massive Sequenzierungsdaten sowohl von menschlichen als auch von Umweltviromik-Proben untersuchen. Unsere gemeinsamen Ergebnisse liefern wertvolle Einblicke in das Verhalten und die Funktion von Bakteriophagen und ihren kodierten Genen bei Krankheiten und in natürlichen Umgebungen.

Acknowledgments

Pursuing a Ph.D. is a challenging and sometimes solitary journey. As I reach the end of this path, I would like to express my heartfelt gratitude and appreciation to everyone who has supported and encouraged me throughout this endeavor. Without their invaluable guidance, assistance, and encouragement, this thesis would not have been possible.

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Dr. Li Deng, for her support, expert guidance, and constructive feedback during my research. I am also grateful to Prof. Dr. Aurélien Tellier, who kindly agreed to be the chairman of the examination committee for my doctoral dissertation, and to Dr. Johannes Söding and Prof. Dr. Christian L. Müller, who kindly agreed to be the second examiner for my doctoral dissertation.

I would like to acknowledge my colleagues Dr. Tianli Ma, Dr. Jinling Xue, and Dr. Mohammadali Khan Mirzaei, for their active collaboration and assistance in my PhD projects. My gratitude also extends to all other members of the group, including Shiqi Luo, Wanqi Huang, Dr. Rita Costa, Magdalena Unterer, Kawtar Tiamani, Adrian Thaqi, and Sophie Smith, for their stimulating discussions and insightful suggestions that enhanced my research experience. I would like to thank Sarah-Irina Pfeifer-Nigisch, Judith Brehme, and Regina Geiger for helping with all the administrative tasks and providing additional support.

I would like to extend my appreciation to the colleagues from the "Biomedical Statistics and Data Science" group. Prof. Dr. Christian L. Müller offered me the opportunity to join the group, where I gained a deeper understanding of the fundamentals of statistics in biology. I am grateful to Daniele Pugno, Mara Stadler, Roberto Olayo Alarcón, Stefanie Peschel, Johannes Ostner, Oleg Vlasovets, Viet Tran, and Tong Wu for teaching me statistics and mathematics, as well as for engaging in fruitful discussions.

I would also like to extend my appreciation to the members of Dr. Johannes Söding's "Quantitative Biology and Bioinformatics" group: Annika Jochheim, Dr. Milot Mirdita, and Prof. Dr. Martin Steinegger, for their assistance in using their developed software, which saved me considerable time in almost all my projects.

Additionally, my gratitude extends to my friends and former colleagues, Dr. Jianfeng Sun, Prof. Dr. Xia Wang, and Dr. Busheng Xue, for their academic discussions, productive collaborations, and emotional support.

I gratefully acknowledge the financial support provided by the China Scholarship Council, which enabled me to pursue my research. I also appreciate the assistance and resources provided by Helmholtz Zentrum München and Technische Universität München, which significantly facilitated my work and enhanced my professional skills.

Finally, I would like to express my deepest gratitude to my parents for their unwavering love and support throughout my Ph.D. journey. Their belief in me has been a guiding light during the most challenging times, and their encouragement has been invaluable in helping me persevere.

List of Publications

This work constitutes a cumulative dissertation based on peer-reviewed publications. Chapters 3, 4, and 5 describe the published and submitted articles included in this dissertation (#Equal contributions):

- **Jinlong Ru**, Mohammadali Khan Mirzaei, Jinling Xue, Xue Peng, Li Deng, 2023. ViroProfiler: a containerized bioinformatics pipeline for viral metagenomic data analysis. *Gut Microbes* 15, 2192522.
- Tianli Ma#, **Jinlong Ru**#, Jinling Xue, Sarah Schulz, Klaus-Peter Janssen, Michael Quante, Li Deng, 2021. Differences in Gut Virome Related to Barrett Esophagus and Esophageal Adenocarcinoma. *Microorganisms* 9, 1701.
- **Jinlong Ru**#, Jinling Xue#, Jianfeng Sun, Linda Cova, Li Deng, 2023. Unveiling the hidden role of aquatic viruses in hydrocarbon pollution bioremediation. *Journal of Hazardous Materials* 459, 132299.

During the dissertation work, I was co-author of other peer-reviewed publications:

- Jianfeng Sun#, **Jinlong Ru**#, Lorenzo Ramos-Mucci, Fei Qi, Adam P. Cribbs, Li Deng, Xia Wang, 2023. DeepsmirUD: Prediction of Regulatory Effects on microRNA Expression Mediated by Small Molecules Using Deep Learning. *International Journal of Molecular Sciences* 24, 1878.
- Shiqi Luo, **Jinlong Ru**, Mohammadali Khan Mirzaei, Jinling Xue, Xue Peng, Anna Ralser, Raquel Mejías-Luque, Markus Gerhard, and Li Deng, 2023. Gut virome profiling identifies an association between temperate phages and colorectal cancer promoted by Helicobacter pylori infection. *Gut Microbes* 15, 2257291.
- Shiqi Luo, **Jinlong Ru**, Mohammadali Khan Mirzaei, Jinling Xue, Xue Peng, Anna Ralser, Joshua Lemuel Hadi, Raquel Mejías-Luque, Markus Gerhard, and Li Deng, 2023. Helicobacter pylori infection alters gut virome by expanding temperate phages linked to increased risk of colorectal cancer. *Gut* doi: 10.1136/gutjnl-2023-330362.
- Mohammadali Khan Mirzaei, Jinling Xue, Rita Costa, **Jinlong Ru**, Sarah Schulz, Zofia E. Taranu, Li Deng, 2021. Challenges of Studying the Human Virome – Relevant Emerging Technologies. *Trends in Microbiology* 29, 171–181.
- Mohammadali Khan Mirzaei, Md. Anik Ashfaq Khan, Prakash Ghosh, Zofia E. Taranu, Mariia Taguer, **Jinlong Ru**, Rajashree Chowdhury, Md. Mamun Kabir, Li Deng, Dinesh Mondal, Corinne F. Maurice, 2020. Bacteriophages Isolated from Stunted Children Can Regulate Gut Bacterial Communities in an Age-Specific Manner. *Cell Host & Microbe* 27, 199-212.e5.

- Elisabeth Entfellner, Ruibao Li, Yiming Jiang, **Jinlong Ru**, Jochen Blom, Li Deng, Rainer Kurmayer, 2022. Toxic/Bioactive Peptide Synthesis Genes Rearranged by Insertion Sequence Elements Among the Bloom-Forming Cyanobacteria Planktothrix. *Frontiers in Microbiology* 13:901762.

Contents

List of Figures

Acronyms

AMG	Auxiliary metabolic gene
Bp	base pairs
BE	Barrett's esophagus
EAC	Esophageal adenocarcinoma
HGT	Horizontal gene transfer
VC	Viral cluster
ORF	Open Reading Frame
RefSeq	Reference Sequence database
NCBI	National Center for Biotechnology Information
ICTV	International Committee on Taxonomy of Viruses
HMM	Hidden Markov Model
HMMER	Hidden Markov Model-based search tool
BLAST	Basic Local Alignment Search Tool
KEGG	Kyoto Encyclopedia of Genes and Genomes
vHYDEG	virus-encoded hydrocarbon degradation genes
eggNOG	Evolutionary genealogy of genes: Non-supervised Orthologous Groups
PHROG	Prokaryotic Virus Remote Homologous Groups
CANT-HYD	Calgary approach to ANnoTating HYDrocarbon degradation genes
IMG/VR	The Integrated Microbial Genome/Virus

1 Introduction

1.1 Viruses in nature and health

Viruses are ubiquitous in nature, comprising an estimated 10^{31} virus particles in the biosphere (Hendrix et al. 1999). This makes them the largest reservoir of genetic diversity on Earth and a driving force behind global geochemical cycles (Suttle 2005). Their omnipresence leads to numerous interactions with host organisms, significantly impacting ecosystems and playing a crucial role in the biochemical cycling of major elements. For instance, viruses can redirect the flow of carbon into particulate or dissolved organic matter through lysis of their bacterial hosts (Wilhelm and Suttle 1999). They have also been shown to directly encode enzymes involved in the metabolism of carbon, nitrogen, and sulfur (Kieft et al. 2021; Thompson et al. 2011).

Bacteriophages, or phages, are viruses that infect bacteria and constitute the majority of viruses on Earth. By influencing the mortality, diversity, and evolutionary trajectories of their bacterial hosts, phages can modulate microbial populations in various environments. This, in turn, shapes the ecology and impacts the homeostasis of microbiota (Chevallereau et al. 2021). Despite their immense ecological significance, phages have historically been understudied due to methodological limitations and difficulties in culturing and isolating them. Considering the vast number of phages in nature, only thousands of isolates with complete genome sequences exist to date, and current culture-independent approaches suggest we are only uncovering the tip of the phage iceberg (Perez Sepulveda et al. 2016).

The human virome, comprising the viral components of the human microbiome, consists of approximately 10^{13} particles per individual and is also dominated by bacteriophages, according to existing viral sequencing research (Liang and Bushman 2021). Though similarly vast and complex compared with human microbiome, the human virome remains largely unexplored, with majority sequence data in typical virome studies remaining unidentified and new viruses being discovered frequently. Nevertheless, increasing evidence associates viral community states with health or disease status (Liang et al. 2020; Ma et al. 2018; Norman et al. 2015; Reyes et al. 2015; Zhao et al. 2017). Most of these studies have produced primarily descriptive results, necessitating in-depth functional analyses of the virome and its behavior for a better understanding of its function and contribution to health and disease.

Nevertheless, the prosperous studies about bacteriophages and virome in the last decades have generated numerous results, revealed the previously neglected role that bacteriophages played in manipulating the microbiome, interacting with both bacteria hosts and human immune systems, and thus positively or negatively impact the human health or environments.

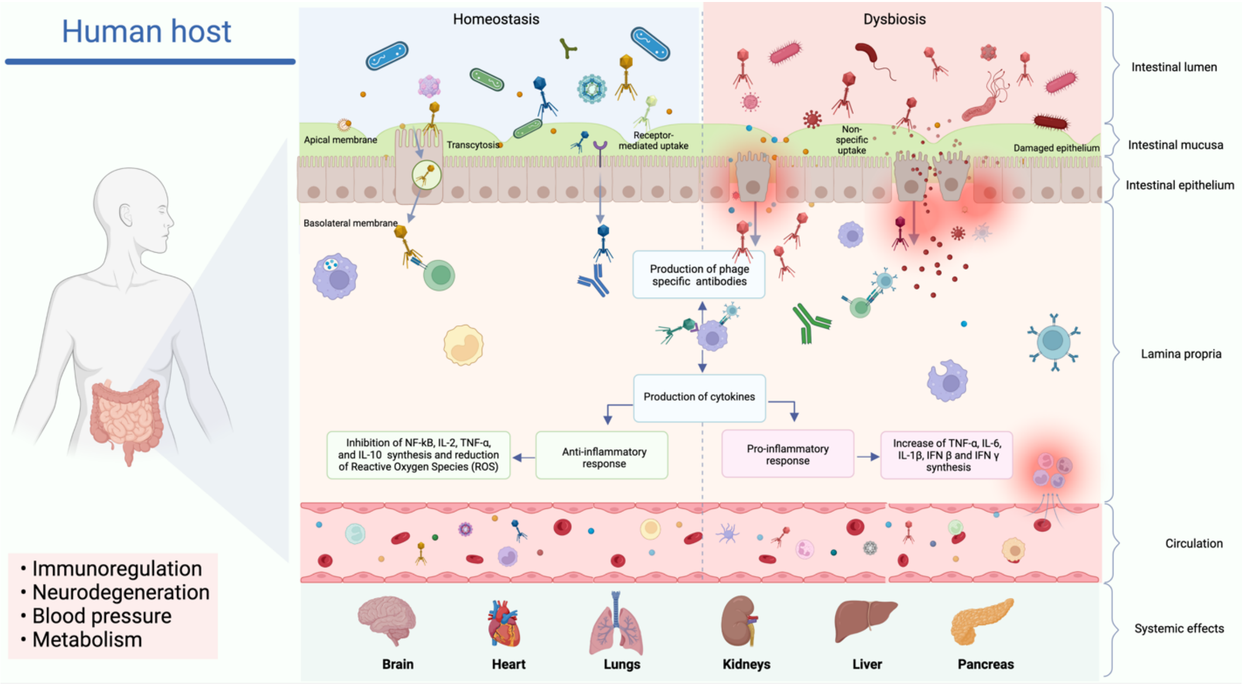


Figure 1. The schematic diagram of the diverse human body sites that phages have been reported directly and indirectly impact on health (Adapted from (Tiamani et al. 2022)).

1.2 Impact of bacteriophages on human health

The human body is a complex ecosystem, hosting trillions of microorganisms, including bacteria, viruses, fungi, and archaea. This vast microbial community, known as the human microbiome, plays a vital role in maintaining our health and well-being. Among these microorganisms, bacteriophages represent the most abundant and diverse group of viruses found within the human microbiome. They specifically target bacteria and are thought to impact the composition of our body's microbial environment considerably. In addition to directly affecting the ecology of their bacterial host, they may also operate through more indirect pathways, such as modulating metabolism or the immune system of the human body (Wahida, Tang, and Barr 2021). A lot of body sites that were previously believed to be sterile were later proved to be resident by phages, such as urinary tract, bladder and blood stream (Figure 1). While phages are not classical pathogens, they are able to bypass mammalian physical barriers, and it has become increasingly clear that they can influence the mammalian immune system (Barr 2017; Van Belleghem et al. 2019). Bacteriophages participate in the immune reaction of human hosts in different ways, and could offer the human host defence against pathogenic bacteria. The adherence of phages to the mucus layer of the gut might provide non-host immunity by protecting the epithelial cells from bacteria (Barr 2017). Inversely, some bacteriophages can encode ankyrins that, after bacterial expression, reduce the eukaryotic immune response and phagocytosis of bacteria (Jahn et al. 2019).

Recent research has shed light on the critical role that bacteriophages play in modulating the human microbiome and impacting overall health. Phages are known to influence bacterial populations by selectively infecting and lysing specific bacterial strains, thus contributing to bacterial diversity and competition. This dynamic relationship between phages and bacteria is essential for maintaining a balanced microbial community, which in turn affects various physiological processes, such as digestion, metabolism, and immunity. For example,

Phages and Antimicrobial Resistance. Phages may play a role in the dissemination of antimicrobial resistance genes among bacteria, potentially contributing to the global rise of antibiotic-resistant bacterial infections. Horizontal gene transfer (HGT) facilitated by phages can enable the spread of resistance genes between bacterial species. Provide benefit to their host bacteria under the environmental pressure such as antibiotic treatment, re-shape the microbiome structure and affecting the disease status via phage-mediated HGT (Mohan Raj and Karunasagar 2019).

Phages and Type 2 Diabetes. The potential role of bacteriophages in type 2 diabetes (T2D) has gained increasing attention in recent years. Dysbiosis of the gut microbiome has been linked to the development of T2D, with changes in phage populations possibly contributing to this imbalance. Virome in T2D patients are characterized by significantly altered viral taxonomic composition and weaken viral-bacterial correlations compared with lean controls (Yang et al. 2021). Phages may influence glucose metabolism and insulin sensitivity by modulating the composition and function of gut microbiota, which in turn can affect the production of short-chain fatty acids, inflammation, and gut barrier integrity. A consortium made up of eight phages were suggested to be a potential index of T2D (Chen et al. 2021).

Phages and Inflammatory Bowel Disease. Inflammatory bowel disease (IBD), including Crohn's disease and ulcerative colitis (UC), has been linked to alterations in the gut microbiome, with emerging evidence implicating phages in the pathogenesis of these disorders, interindividual dissimilarity between mucosal viromes was higher in UC than controls (Zuo et al. 2019). Escherichia phage and Enterobacteria phage were more abundant in the mucosa of UC than controls. Changes in phage populations may contribute to IBD by driving bacterial dysbiosis, triggering aberrant immune responses, and promoting chronic inflammation. Phage-based therapies, such as the targeted elimination of specific pathogenic bacteria or the restoration of a balanced gut microbiota through phage consortia treatment, could potentially offer novel approaches for managing IBD (Federici et al. 2022).

Phages and Respiratory Health. The respiratory microbiome, which includes the nasal passages, throat, and lungs, also contains a diverse community of phages that can influence respiratory health. The combination of serum cytokine and Propionibacterium phages could work as a strong predictor of acute respiratory tract infections (ARTIs), showing the tight relationship of the phage species and infections in respiratory tract (Li et al. 2019).

These findings highlight the complex interplay among phages, bacteria, and the human host. Early studies attempted to uncover the contributions of the entire virome community or phage consortia to disease development. However, these studies were limited by their focus on documenting changes in the community structure and calculating correlations between virome composition and specific disease types. Therefore, a deeper and more comprehensive understanding of phage biology and ecology, as well as their potential impact on human health, is necessary.

1.3 Impact of bacteriophages on environment

In addition to human health, bacteriophages have a significant impact on the environment, as they affect the ecology and evolution of bacteria and archaea which are important components of the biosphere and contribute to the geochemical cycling of key elements such as carbon (Figure 2). For example,

Implications for Aquatic Ecosystems. In aquatic ecosystems, bacteriophages play a significant role in shaping the dynamics of microbial communities. They can influence the population structure and activity of bacteria involved in processes such as nitrogen and sulfur cycling, which are critical for maintaining the health of aquatic environments. Additionally, phages can impact the food web by affecting the growth and survival of bacteria that serve as the primary food source for filter-feeding organisms like zooplankton. In marine, free phage particles were also thought to contribute to the dissolved organic matter (DOM) pool. Jover et al, suggested phages could significantly contribute to phosphorous reservoir in the phosphor limited environment, while Bonnain et al hypothesized that phages could serve as organic ligands of ion in ocean, acting as a reservoir of trace metals that frequently limit primary production (Bonnain, Breitbart, and Buck 2016; Jover et al. 2014).

Influence on Soil Fertility and Agriculture. Bacteriophages can affect soil fertility and agricultural productivity by interacting with the complex microbial communities present in the soil (Svircev, Roach, and Castle 2018). They can influence the abundance and activity of nitrogen-fixing bacteria, which play a crucial role in converting atmospheric nitrogen into a form that plants can utilize (Wang et al. 2022). Additionally, phages can impact the populations of bacteria involved in the decomposition of organic matter, thus influencing nutrient availability in the soil (Jansson and Wu 2023). By manipulating phage-host interactions, it may be possible to develop sustainable approaches to improve soil fertility and promote crop productivity.

Bioremediation and Pollution Control. Bacteriophages have the potential to contribute to bioremediation efforts, particularly in the context of pollution control. Certain bacteria are known to degrade pollutants like hydrocarbons, heavy metals, and pesticides. Bacteriophage replication related genes were found to be enriched in the deep-sea oil plume, suggesting the potential role of phages in the biodegradation if oil spills in deep-sea environments (Lu et al. 2012). Bacteriophages might also affect microbial oil degradation negatively; pollutants can induce prophages and the resulted phage particles caused lysis of bacterial cells (Head, Jones, and Röling 2006). Selectively targeting and promoting the growth of bacteria degraders through phage predation, it may be possible to enhance the efficiency of bioremediation strategies. Moreover, phages can be engineered to carry specific genes that enable the degradation of pollutants, further bolstering their potential application in environmental cleanup efforts.

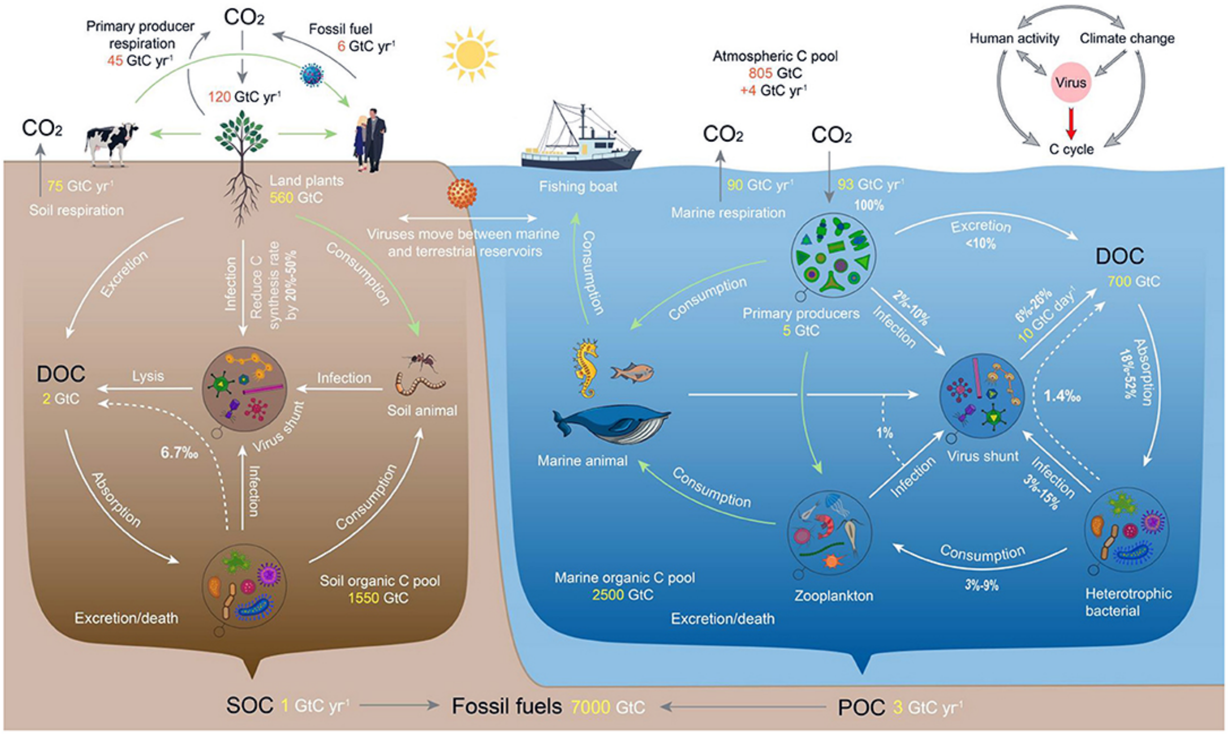


Figure 2. The potential contributions of viruses on carbon cycling in both the aquatic environment and soil. The arrows show the roles that viruses play in the traditional food web, the “microbial loop” and the C cycle network of ecosystems (Gao et al. 2022).

The environmental implications of phage-bacteria interactions are vast, affecting biogeochemical cycles, ecosystem dynamics, and even climate change, making them essential components of ecosystems.

1.4 Viral metagenomics (viromics)

The field of viromics has been revolutionized by the advent of next-generation sequencing (NGS) technology. This breakthrough enables researchers to bypass traditional culture-based methods and directly sequence entire viral communities, allowing for unbiased identification, characterization, and quantification of viruses without prior knowledge of their composition. The development of advanced algorithms and bioinformatics tools has further enhanced viromic research by facilitating the efficient analysis and interpretation of complex sequencing data.

A traditional viromic research workflow typically involves collecting samples from environmental, isolated culture, or host-associated sources. Subsequently, viral DNA or RNA is extracted and sequenced. The sequencing reads are then processed using a series of bioinformatics software tools to detect viral sequences. These viral sequences are annotated using various tools and databases to characterize their taxonomic and functional properties (Figure 3). In the following sections, we present a brief introduction of the steps involved in a viromic study.

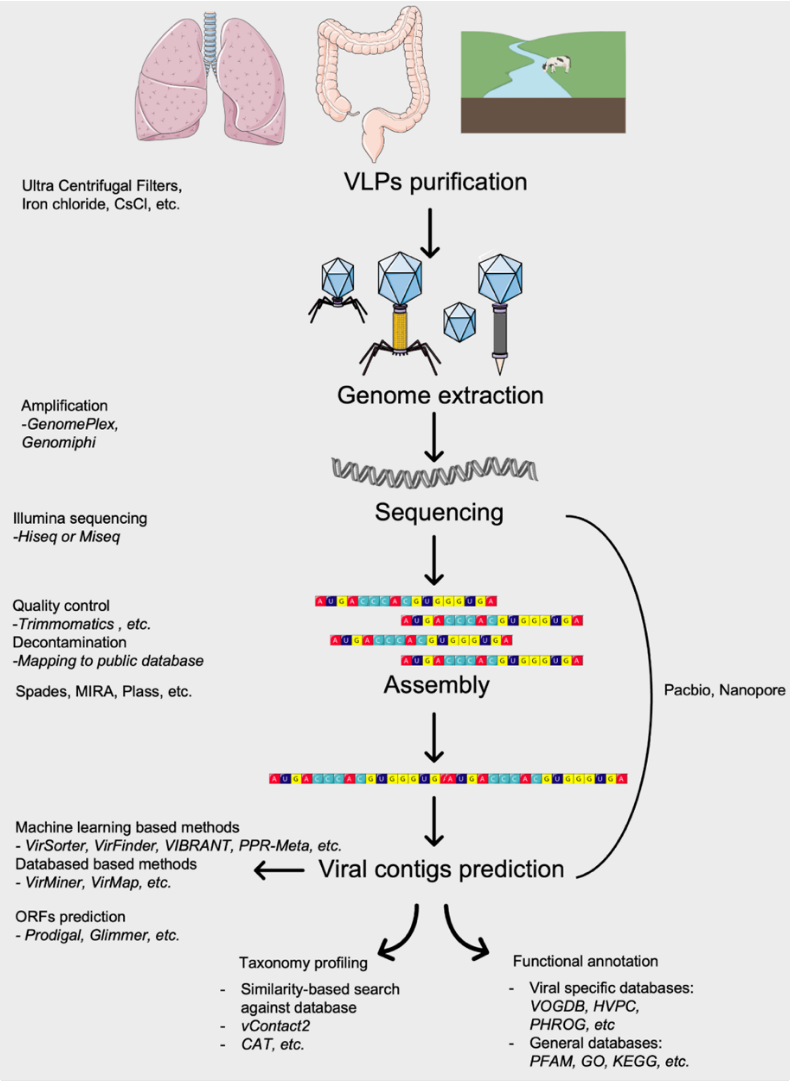

Ultra Centrifugal Filters,
Iron chloride, CsCl, etc.
VLPs purification
Amplification
-GenomePlex,
Genomiphi
Genome extraction
Illumina sequencing
-Hiseq or Miseq
Sequencing
Quality control
-Trimmomatics , etc.
Decontamination
-Mapping to public database
Spades, MIRA, Plass, etc.
Assembly
Pacbio, Nanopore
Machine learning based methods
- VirSorter, VirFinder, VIBRANT, PPR-Meta, etc.
Databased based methods
- VirMiner, VirMap, etc.
Viral contigs prediction
ORFs prediction
- Prodigal, Glimmer, etc.
Taxonomy profiling
- Similarity-based search against database
- vContact2
- CAT, etc.
Functional annotation
- Viral specific databases: VOGDB, HVPC, PHROG, etc
- General databases: PFAM, GO, KEGG, etc.

Figure 3. Workflow for identifying viral sequences in most common sample types (Adapted from (Mirzaei et al. 2021))

1.4.1 DNA extraction and amplification

DNA extraction techniques are essential for detecting viral DNA in biological samples. The typical viromic samples tend to be limited in quantity, have a low proportion of viral particles, and possess high levels of contamination from the host's microbiome (Liu et al. 2020; Sabatier et al. 2020). There are various methods available, but each has its own pros and cons that usually favor the recovery of more abundant organisms.

Common techniques involve the use of a 0.2 µm filter to remove larger particles like host cells and bacteria. However, this can lead to the elimination of large viruses and reduce the amount of viral DNA recovered. CsCl gradient ultracentrifugation purification can provide very pure samples but may be biased towards isolating certain types of phages. Various procedures for boosting the concentration of viral nucleic acids, such as random amplified shotgun library (RASL), linker-amplified shotgun library (LASL), and multiple displacement amplification (MDA), each have their own disadvantages and inaccuracies (Mirzaei et al. 2021). It is imperative to eliminate extraneous contamination using virus-like particle (VLP) purification techniques to gain a precise understanding of phage prevalence. Recently designed flow cytometry-based approaches can differentiate VLPs from other microorganisms through fluorescence-activated cell sorting (Deng et al. 2014; Gaudin and Barteneva 2015). Although these strategies reduce contaminations, they may still affect the sensitivity of viral detection.

Considering the constraints and biases posed by current sample processing methods, bioinformatics processes used in downstream analysis are essential for diminishing potential negative consequences caused by experimental techniques.

1.4.2 DNA sequencing

The most used NGS (Next-Generation Sequencing) platforms in viromics are Illumina's HiSeq and MiSeq systems. These platforms offer a broad selection of sequencing options, such as shotgun metagenomics, which enables the analysis of all genetic material present in a sample, and targeted metagenomics, which focuses on specific genes or genomic regions of interest. However, NGS technologies generate limited read length, which can hinder the accurate assembly and annotation of viral genomes, especially for those containing repeat regions or high genomic plasticity. To overcome these limitations, Third-Generation Sequencing (TGS) platforms, such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), have emerged as promising alternatives. These systems offer longer read lengths, with some reads exceeding 100 kilobases, greatly improving the resolution of complex viral communities and facilitating the assembly of complete viral genomes. However, the current error rates of TGS platforms are higher than those of NGS (Dohm et al. 2020).

Recognizing the complementary strengths of NGS and TGS platforms, some studies have begun to employ a hybrid sequencing approach, combining the advantages of both technologies. This strategy has proven to be particularly effective in generating high-quality viral genomes (Cook et al. 2023; Zaragoza-Solas et al. 2022).

1.4.3 Viral detection from viromic sequencing data

The primary objective of viromic studies is to identify viral contigs or genomes from sequencing data. One strategy is based on reads classification, where clean sequencing reads are aligned to a reference database of known viral genomes using sequence comparison and similarity search tools, such as BLAST (Camacho et al. 2009). More efficient algorithms such as k-mer-based methods use the frequency of short nucleotide sequences (k-mers) to classify reads. These methods are generally faster and more memory-efficient than alignment-based methods. Kraken, a k-mer-based taxonomic classifier, is a widely-used tool in this category. It constructs a compact k-mer database from reference viral genomes and classifies metagenomic reads by mapping their k-mers to the database (Wood and Salzberg 2014). Kraken's speed and accuracy have made it popular for large-scale metagenomic studies. With the growing interest in k-mer-based methods, other tools such as CLARK (Ounit et al. 2015) and Kaiju (Menzel, Ng, and Krogh 2016) have been developed. These tools also employ k-mer-based approaches but utilize distinct algorithms and optimizations to enhance classification accuracy and efficiency.

Another strategy is based on genome assembly. These methods involve assembling short sequencing reads into longer contigs, followed by identifying viral contigs using sequence similarity or other genomic features. One of the first widely used tools for genome assembly was Velvet (Zerbino 2010), a de Bruijn graph-based assembler. MetaVelvet (Namiki et al. 2012), the metagenome version of Velvet, was designed for metagenomic assembly and demonstrated improved performance in recovering viral genomes from complex environmental samples. Another de Bruijn graph-based assembler is IDBA-UD, developed as a metagenomic assembler capable of handling uneven sequencing depths and multiple genomes in a single dataset (Peng et al. 2012). SPAdes (Bankevich et al. 2012) and its metagenomic version, metaSPAdes (Nurk et al. 2017), were designed as general-purpose assemblers, which also proved useful for assembling viral genomes. Based on previous performance evaluations on viromic samples, metaSPAdes has shown the best performance for viral genome assembly in terms of contig length, assembly accuracy, and the recovery of low-abundance viruses, making it a good choice for viromic sequencing data assembly (Sutton et al. 2019).

After obtaining viral contigs from the assembly, multiple viral detection tools can be employed to identify viral contigs. These tools also rely on a virus reference database but utilize different strategies to distinguish viral contigs from non-viral contigs. For example, VirSorter (Roux et al. 2015) and MARVEL (Amgarten et al. 2018) identify viral contigs by comparing their proteins to viral proteins in the reference database. Some machine learning-based tools, such as VirFinder (Ren et al. 2017) and DeepVirFinder (Ren et al. 2020), employ k-mer profiles trained on viral reference databases. A more detailed introduction to various viral detection methods can be found in Section 7.2.

1.4.4 Viral contig binning

One challenge in analyzing viromic data is the short and fragmented nature of assembled contigs. This often leads to multiple contigs representing a single viral genome. Annotating these contigs separately can result in the loss of valuable information and incorrect annotation results due to the highly diverse and mosaic nature of viral genomes. To avoid this issue, researchers have turned to viral contig binning. This process groups contigs belonging to the same viral genome into a single bin, allowing for more comprehensive annotation of the viral genome.

Early efforts in viral contig binning involved using general-purpose metagenomic binning tools, such as CONCOCT (Alneberg et al. 2014), GroopM (Imelfort et al. 2014), and MaxBin (Wu et al. 2014), which were primarily developed for bacterial and archaeal genome recovery but could also be applied to viral datasets. However, these methods had limitations in accurately binning viral sequences due to the unique characteristics and high diversity of viral genomes. Later, specialized viral contig binning tools were developed, such as Phamb (Johansen et al. 2022) and vRhyme (Kieft et al. 2022), which consider the unique features of viral genomes.

Phamb enables the binning of thousands of viral genomes directly from bulk metagenomic data by combining a deep learning-based metagenomic binning algorithm with paired metagenome and metavirome datasets. However, one limitation of Phamb is that it requires more than 50,000 viral contigs as input to achieve better binning results. vRhyme is a versatile and fast tool for viral contig binning. It generates viral metagenome-assembled genomes (vMAGs) by comparing coverage variances and utilizing supervised machine learning for sequence characteristic classification. To improve binning, vRhyme exploits unique viral genome properties, such as the infrequency of duplicate genes in viruses. When evaluated on simulated data, vRhyme created more complete and less contaminated vMAGs than existing tools.

1.4.5 Viral contig and genome annotation

The annotation of viral contigs or bins involves identifying and characterizing functional elements, such as genes, regulatory elements, and functional non-coding RNA sequences. The first step in annotating viral genomes is to identify protein-coding genes within the assembled sequences. Several gene prediction tools have been developed, including Prodigal (Hyatt et al. 2010), GeneMarkS (Besemer, Lomsadze, and Borodovsky 2001), and Glimmer (Kelley et al. 2012). These tools use distinct algorithms and models to predict open reading frames (ORFs) and coding sequences (CDS) within viral genomes. Once the genes have been identified, their functions must be annotated. This can be done by comparing the predicted protein sequences to databases of known protein families, domains, or motifs using sequence comparison tools such as BLAST (Camacho et al. 2009), HMMER (Eddy 2011), MMseqs2 (Steinegger and Söding 2017), DIAMOND (Buchfink, Xie, and Huson 2015), InterProScan (Jones et al. 2014), and EggNOG-mapper (Cantalapiedra et al. 2021). Commonly utilized databases include InterPro (Paysan-Lafosse et al. 2022), EggNOG (Hernández-Plaza et al. 2022), PFAM (Mistry et al. 2021), KEGG (Kanehisa et al. 2021), UniProt (The UniProt Consortium 2019), and viral-specific databases such as NCBI viral RefSeq proteins (Li et al. 2021), VOGDB (https://vogdb.org), pVOG ((Grazziotin, Koonin, and Kristensen 2017)), and PHROG (Terzian et al. 2021) database. In addition to protein-coding genes, viral genomes may also contain functional non-coding RNA elements, such as transfer RNAs (tRNAs) and transfer-messenger RNAs (tmRNAs). Tools like tRNAscan-SE (Lowe and Eddy 1997) can be used to predict the presence of non-coding RNA elements within viral genomes. Several tools have been developed to integrate multiple steps of viral genome annotation into a single software. For instance, DRAM-v (Shaffer et al. 2020) annotates vMAGs using KEGG, UniRef90, PFAM, RefSeq viral proteins, VOGDB database, and custom databases. Pharokka (Bouras et al. 2022) identifies predicted CDS and annotates them using the PHROG database. A combination of computational tools, databases, and manual curation is often required to achieve accurate and comprehensive annotations of viral genomes.

1.4.6 Virus-host prediction

Identifying the host of viruses can offer valuable insights into viral ecology, evolution, transmission, and host interactions. This information can also contribute to developing strategies for controlling viral infections, such as vaccines or antiviral drugs. However, experimental methods for determining virus-host relationships can be challenging, time-consuming, and labor-intensive. Consequently, computational approaches have emerged as a beneficial method for predicting the host organisms a given virus can infect.

Various features can be utilized to predict the host of a given virus, particularly for bacteriophages (Edwards et al. 2016). (1) Sequence similarity-based methods: tools such as BLAST and tBLASTx can compare viral sequences against a database of known host genomes or host marker genes to identify potential hosts based on sequence similarity. CRISPR arrays in bacterial and archaeal genomes contain spacers originating from viral or plasmid sequences, providing a record of previous encounters with foreign genetic elements. Tools like SpacePHARER (Zhang et al. 2021) can predict viral hosts based on matches between viral sequences and CRISPR spacers in host genomes. (2) Machine learning-based methods: machine learning algorithms can be trained to predict viral hosts based on the k-mer frequency patterns in viral sequences. Tools like WIsH (Galiez et al. 2017), HostPhinder (Villarroel et al. 2016), and VirHostMatcher-Net (Wang et al. 2020) employ k-mer frequencies and machine learning techniques for viral host prediction.

Some tools integrate multiple features to enhance host prediction accuracy. For instance, iPHoP combines various computational approaches, including "host-based" tools and "phage-based" tools. Host-based tools leverage different levels and patterns of sequence similarity between phage and host genomes, while phage-based tools extract information from a database of reference phages and archaeoviruses with known hosts (Roux et al. 2023).

1.4.7 Virus taxonomy assignment

Traditional virus taxonomy, based primarily on morphological characteristics and host range, has its roots in the early days of virology when electron microscopy and cultivation were the primary methods for studying viruses. The order Caudovirales, for example, which encompasses all tailed bacteriophages, is well suited for classifying viruses isolated from cultivated bacteria or other hosts. However, the majority of viral sequences in viromic datasets lack corresponding host and morphological information, rendering traditional classification schemes inadequate. As a result, modern viral taxonomy relies heavily on sequence similarity between viral genomes and known viruses in public databases. However, the sheer volume and diversity of viral sequences uncovered through metagenomic studies often outpace the growth of these databases, resulting in many novel viruses lacking close relatives with which to compare. To address this challenge, the International Committee on Taxonomy of Viruses (ICTV) has begun to establish new nomenclature for viruses discovered through metagenomics based on their genomic content (Gorbalenya et al. 2020).

In parallel, the development of computational tools such as vConTACT (Bolduc et al. 2017), vConTACT2 (Bin Jang et al. 2019), CAT (von Meijenfeldt et al. 2019), and MMseqs2 taxonomy module (Mirdita et al. 2021) has facilitated the assignment of taxonomy to viral contigs in viromic datasets. These tools employ various strategies and reference databases, each offering unique advantages and limitations in terms of speed, accuracy, and compatibility with existing databases. A major hurdle in viral taxonomy assignment is the discrepancy in nomenclature between the ICTV and the NCBI taxonomy. This divergence complicates the process of converting between the two nomenclature systems and can lead to inconsistencies in viral classification. To overcome this issue, some tools, like the MMseqs2 taxonomy module, enable users to create customized reference databases tailored to their needs.

1.4.8 Virus lifestyle prediction

Virus lifestyle refers to the manner in which a virus interacts with its host. There are two primary types of viral lifestyles: virulent and temperate. Virulent viruses exhibit a strictly lytic life cycle, causing lysis (bursting) of the host cell after infection. In contrast, temperate viruses can integrate their DNA into the host chromosome and remain dormant (as proviruses) for many generations without causing host cell lysis. However, under specific conditions, such as exposure to stress factors, proviruses can be induced to enter a lytic cycle, leading to host cell lysis. Therefore, a virus's lifestyle determines its potential to undergo either a lytic or lysogenic life cycle.

Accurate prediction of phage lifestyles is crucial for several reasons. First, it aids in determining the potential use of phages in therapy. For instance, virulent phages are more suitable for phage therapy because they can rapidly lyse bacterial cells and decrease the bacterial population. In contrast, temperate phages may be less effective since they can remain dormant in the host cell without causing lysis. Second, predicting phage life cycles can help understand their biology and evolution. By examining the genetic and molecular mechanisms underlying lytic and lysogenic cycles, researchers can gain insights into how these viruses have evolved to interact with their bacterial hosts. Finally, predicting phage life cycles is essential for understanding their ecological role in natural environments such as soil and water. For example, temperate phages can transfer genes between bacteria through lysogeny, which can have significant implications for bacterial evolution and adaptation to changing environments.

Multiple features can be utilized to construct models for virus lifestyle prediction. BACPHLIP detects the presence of a set of lysogenic-related proteins in viral genomes and predicts lifestyle using a Random Forest classifier (Hockenberry and Wilke 2021). PhaTYP employs BERT (Bidirectional Encoder Representations from Transformers) to enhance the accuracy of lifestyle prediction on short contigs (Shang, Tang, and Sun 2023). PhageAI applies NLP (Natural Language Processing) techniques to encode phage genome sequences and predict lifestyle using an SVM (Support Vector Machine) model (Tynecki et al. 2020). DeePhage uses a "one-hot" encoding form to represent phage genome sequences and detects local features using a convolutional neural network (CNN) (Wu et al. 2021). Each of these tools has its advantages and limitations. For instance, PhageAI has the highest accuracy but can only be used from a web server with a limit on usage times. BACPHLIP is user-friendly but can only predict complete phage genomes. Our recently developed tool, Replidec, overcomes some limitations by detecting temperate-related protein families from all viral proteins. It exhibits similar performance compared to PhageAI and can be used on non-complete vMAGs.

1.4.9 Auxiliary metabolic genes

A viral-specific annotation is the identification and annotation of auxiliary metabolic genes (AMGs). Recent studies have shown that viruses can encode AMGs, which have important implications for the metabolism and ecology of the microbial communities they infect (Breitbart et al. 2018; Crummett et al. 2016; Gasper et al. 2017; Kieft et al. 2021; Thompson et al. 2011). These AMGs can alter host metabolism, enhance viral fitness, and contribute to biogeochemical cycling, as well as shaping microbial community structure and function, highlighting the critical impact of viral-encoded AMGs on microbial ecosystems. However, the identification and annotation of AMGs in viromic data is a challenging task (Pratama et al. 2021). AMGs can be highly diverse, and their sequences may be only distantly related to known reference genes, making it difficult to accurately detect and characterize them using traditional sequence comparison methods. Additionally, the fragmented and incomplete nature of viromic data can further complicate the identification and annotation of AMGs. To address these challenges, multiple tools have been developed recently. For example, VIBRANT (Kieft, Zhou, and Anantharaman 2020) and DRAM-v (Shaffer et al. 2020) use a ruleset for defining and annotating AMGs in viral genomes. These tools utilize advanced algorithms and curated reference databases to improve the detection and characterization of AMGs, even in cases where the AMGs have low similarity to known reference sequences.

1.5 Reproducible data analyses

In the fields of metagenomics and viromics, researchers often require complex data analysis strategies involving multiple steps, each with a range of tools to address similar issues. Researchers must not only choose the appropriate tools for each step but also install, compile, and run these tools on their own data. Often, multiple tools must be executed simultaneously, and their results combined, while ensuring that each subsequent step begins only after the completion of the preceding step. Some steps require more computational resources than others. In cases where analyses are performed on high-performance clusters (HPC) shared with other researchers, efficient management of computational resources is crucial for maintaining equitable access. Job scheduler software, such as PBS (Portable Batch System) and SLURM (Simple Linux Utility for Resource Management), can automatically manage and allocate resources based on pre-defined configuration files for each analysis script. However, manually creating configuration files and executing each step separately is impractical for large-scale data and projects.

Custom scripts can be used to chain all analysis steps and allow the entire workflow to be run using a single command. However, maintaining and sharing these custom pipelines can be challenging due to dependencies on specific software and platforms. This makes it difficult to reproduce analyses conducted by other researchers. To achieve reproducible analyses, it is essential to adhere to well-defined principles like the FAIR (Findable, Accessible, Interoperable, and Reusable) principles that are widely accepted within the scientific community (Barker et al. 2022). Workflow management systems, such as Snakemake (Mölder et al. 2021) and Nextflow (Di Tommaso et al. 2017), facilitate the development of FAIR-compliant workflows. Both of them are designed to manage and orchestrate complex data analysis pipelines, handling dependencies and parallelism automatically. They guarantee reproducibility, scalability, and maintainability of scientific workflows, and can operate on a variety of computational platforms, from single-core systems to high-performance computing clusters and cloud environments. Researchers only need to write analysis scripts for each step, define input and output data, specify the software used in each step, and provide pre-defined configuration files that define computational resources or other variables required by each analysis. The workflow management system automatically manages dependencies during execution. The choice between Snakemake and Nextflow largely depends on user preferences and programming background. Snakemake is based on Python and utilizes a domain-specific language resembling Python, allowing users to define rules for data processing tasks. Nextflow, on the other hand, is based on Groovy, which may be less familiar to researchers compared to Snakemake's Python-like syntax. Both support modular extension, but Nextflow has a more active community, including the nf-core initiative (Ewels et al. 2020; Yates et al. 2021), which offers an extensive range of well-documented and standardized modules and pipelines for various applications.

To ensure adherence to the FAIR principles, workflow code can be managed using Git and hosted on code-sharing platforms such as GitHub and GitLab. By hosting code on these online platforms, developers can access Continuous Integration/Continuous Deployment (CI/CD) services that streamline workflow development and maintenance. For example, GitHub Actions is a popular CI/CD tool that automates the testing and deployment of code changes, ensuring software remains reliable and up to date. It can also automatically create Docker containers and upload them to Docker Hub, a platform for sharing Docker images. Users can execute a single command to deploy the pipeline on a local computer, HPC, or cloud computing platform. Containerization technologies, such as Docker and Singularity (Kurtzer, Sochat, and Bauer 2017), further enhance the reproducibility and portability of scientific workflows. By encapsulating the software environment and dependencies into a single, unified container, researchers can ensure that their pipelines run consistently across different computational platforms. This eliminates the need for complex installation procedures and minimizes potential software conflicts, allowing users to focus on data analysis rather than troubleshooting.

In conclusion, the integration of FAIR principles, workflow management systems, containerization technologies, and CI/CD practices empowers researchers to produce transparent, reproducible, and easily extensible computational pipelines (Figure 4). The synergistic combination of these tools fosters a collaborative research environment, accelerates scientific progress, and ultimately improves the reproducibility and reliability of viromic studies.

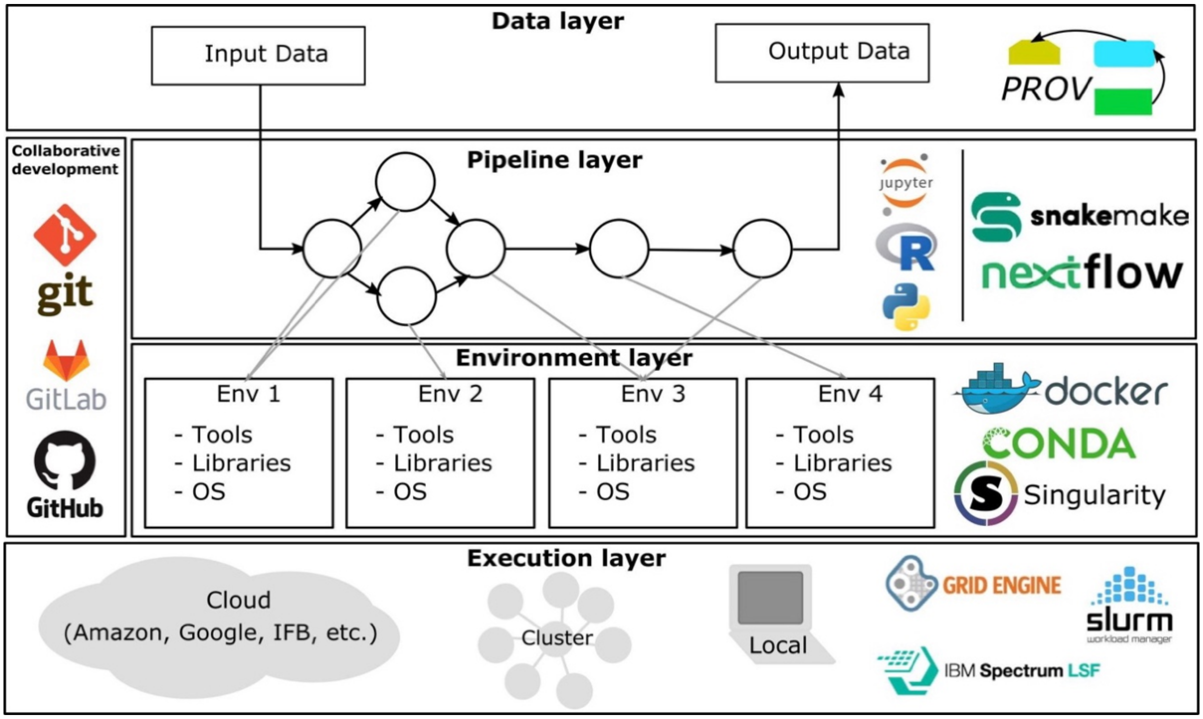


Figure 4. Illustration about Nextflow-based reproducible computational analyses pipeline. Four layers from Data, pipeline, software environment, and execution layers were included in the pipeline (Djaffardjy et al. 2023).

1.6 Motivation and overview of this work

As discussed before, increasing efforts were made to recognize the importance of the virome as it has been associated with diseases and energy flow in nature. However, there are still additional, unique challenges inherent to virome analysis that limited the more comprehensive understanding of the behavior and function of virome in human health and nature. In this thesis, the efforts were made from the following aspect to address the aforementioned questions. In chapter 3, I describe ViroProfiler, a computational pipeline for viromic data analysis. It integrates state-of-the-art bioinformatic tools of viromic data analysis via the modern workflow management framework Nextflow and ensures computational reproducibility using containerization techniques. An R package and Shiny APP were also provided for downstream analyses. In chapter 4, I present a study of the role of bacteriophages in Barrett esophagus and esophageal adenocarcinoma. We highlight a potential link between gut phages and esophageal diseases by identifying distinct gut phage communities and their disease specific AMGs in various stages of these diseases. Viral genes related to bacterial exotoxin and LPS biosynthesis proteins, have been associated with potentially affecting gut bacterial composition and inflammation. In chapter 5, I present a study of the role of viruses, mainly bacteriophages, in hydrocarbon pollution bioremediation. I show that viruses carry a variety of hydrocarbon degradation genes (vHYDEGs) that are involved in the crucial, rate-limiting step of alkane hydroxylation. Predictions of protein structures reveal their metabolic potential. These viruses exhibit a diverse range of taxa and evolutionary backgrounds and are associated with multiple hydrocarbon degraders, suggesting potential for engineering applications in hydrocarbon and crude oil bioremediation. In chapter 6, I describe other projects that I contributed to, including (1) a study of the role of bacteriophages in childhood stunting, we found distinct gut phages in stunted children compared to their non-stunted counterparts. *In vitro* experiments show these phages can regulate bacterial abundance and composition, suggesting their role in the pathophysiology of child stunting. (2) a study about the virome community and function in the *H. pylori*-promoted CRC mice model. We observed expanded temperate phages in the infected mice preceding the evident colonic tumor development, many temperate phages can infect probiotics with known antitumor effects such as butyrate producer *Clostridium butyricum*. (3) a study of the role of bacteriophages in allogeneic stem cell transplantation patients, we found a gene BCoAT in two phage contigs that is linked to high levels of butyric acid in the gut, suggesting that phage-encoded AMGs may contribute to the production of immune-modulating metabolites in the human gut. (4) a review article of the modern computational tools for viromic study. The final chapter is the summary of the dissertation and discusses future developments.

2 Materials and Methods

2.1 ViroProfiler workflow architecture

For the development of ViroProfiler, NextFlow (Di Tommaso et al. 2017) was utilized as the workflow framework. The modular design architecture and container support were incorporated by employing the nf-core template. Software dependencies were managed through Conda YAML configuration files, allowing for the efficient handling of complex dependency chains and versioning. The installation of these dependencies into the Docker container was facilitated by Micromamba, a lightweight and fast package manager. Customized Dockerfiles, based on the Micromamba image, were used to define the container content, ensuring the consistency and reproducibility of the computational environment across different platforms. The building and deployment of Docker containers were automated using GitHub Actions, a popular continuous integration and continuous deployment (CI/CD) service. This automation ensured that the latest updates and improvements to the workflow were seamlessly integrated and readily available to users through Docker Hub, a widely used container registry.

2.2 Standard analyses in ViroProfiler

Raw reads were subjected to cleaning using fastp (Chen et al. 2018). Clean reads were then assembled into contigs using metaSPAdes (Nurk et al. 2017). Contigs from multiple samples were merged into a single FASTA file. These contigs were clustered based on 95% sequence similarity and 85% coverage on the shorter contigs. From each cluster, the longest contig was selected as the representative contig, resulting in the creation of a non-redundant contig library (nrclib). Contig quality was assessed, and bacterial contamination regions from proviruses were removed using CheckV (Nayfach et al. 2021). Open reading frames (ORFs) and genes were identified with Prodigal (Hyatt et al. 2010), followed by clustering of gene and protein sequences using MMseqs2 (Steinegger and Söding 2017) to reduce redundancy. Annotation of the clustered sequences was performed with EggNOG-mapper (Cantalapiedra et al. 2021) and abricate. Clean reads were mapped to the nrclib using Bowtie2 (Langmead and Salzberg 2012), and abundance-related metrics, such as read counts, trimmed mean, covered fraction, and reads per base, were calculated using CoverM (https://github.com/wwood/CoverM). Optionally, tools such as Phamb (Johansen et al. 2022) and vRhyme (Kieft et al. 2022) were employed for binning the nrclib. Viral contigs from nrclib were detected using multiple software, including VirSorter2 (Guo et al. 2021), VIBRANT (Kieft et al. 2020), and DeepVirFinder (Ren et al. 2020). Identified viral contigs were annotated with DRAM-v (Shaffer et al. 2020), while auxiliary metabolic genes (AMGs) were identified and annotated using DRAM-v and VIBRANT. Viral contigs were clustered into genus clusters using vConTACT2 (Bin Jang et al. 2019). Taxonomy was assigned to viral contigs using both vConTACT2 and the MMseqs2 taxonomy module (Mirdita et al. 2021), which compared the contigs against reference viral genomes in the NCBI Virus RefSeq database. Hosts of the viruses were predicted using iPHoP (Roux et al. 2023). Replication lifestyles were predicted with either BACPHLIP (Hockenberry and Wilke 2021) or Replidec (Peng et al. 2022).

2.3 Virome DNA extraction and sequencing

In brief, a comprehensive protocol established in the lab was followed to prepare the samples for viral sequencing. Initially, a sample of less than 50 µL was mixed with 1/5 volume of chloroform and subjected to centrifugation at 14,000 g for 3 minutes. The upper phase was retained to remove proteins from the sample. Subsequently, DNaseI (1U/ µL, Invitrogen, USA, Lot No. 1158858) was introduced and an incubation period of 1 hour at 37°C was maintained to eliminate bacterial DNA fragments. Afterward, the sample was treated with lysis buffer (700 µL KOH stock (0.43g/10ml), 430 µL DDT stock (0.8g/10ml), 370 µL H2 O, pH=12) and incubated at room temperature for 10 minutes. This step was followed by freezing the sample at -80°C for 2 hours to ensure effective cell lysis. Next, the samples were incubated at 55°C for 5 minutes, and 1 µL Proteinase K (20mg/ml, Invitrogen, USA, Lot No. 1112907) was added, followed by another incubation for 30 minutes at 55°C to facilitate protein digestion. To purify the samples, AMPure beads (Agencourt, Beckman Coulter, USA) were utilized. The AMPure beads were added to the samples, and a co-incubation period of 15 minutes was allowed for DNA adsorption. The DNA was then eluted from the beads using 35 µL Tris buffer (10mM, pH=9.8) and stored at -80°C until further analysis. Finally, the prepared samples were subjected to viral sequencing on an Illumina Novoseq 6000 instrument, employing chemistry for 2 × 150 bp reads.

2.4 Additional bioinformatic analyses

In addition to the standard analyses previously mentioned, further analyses were conducted for Manuscript 2. In detail, contigs were annotated using the CAT tool (v5.0.4), which assigns taxonomy to each contig based on their protein content (von Meijenfeldt et al. 2019). For contigs that could not be assigned taxonomy by ViroProfiler and CAT, the Demovir software (https://github.com/feargalr/Demovir) was employed to assign family-level taxonomy. The number of reads mapped to genes was estimated using FeatureCounts (Liao, Smyth, and Shi 2014). The hosts of the viruses were predicted using the VirHostMatcher-Net (Wang et al. 2020).

2.5 Public viral protein and genome analyses

To investigate the role of bacteriophages in hydrocarbon pollution degradation, viral proteins, genomes, and metadata of sample sources were downloaded from the IMG/VR database (Camargo et al. 2022). Viral proteins were also downloaded from the PHROG database (Terzian et al. 2021). Hydrocarbon degradation genes in the viral genomes were annotated using the CANT-HYD database (Khot et al. 2022), which provides a comprehensive resource of genes and enzymes involved in hydrocarbon degradation pathways. InterProScan (Jones et al. 2014) was then utilized, searching against the InterPro database (Paysan-Lafosse et al. 2022), for further annotation of the vHYDEGs. The three-dimensional structure of the proteins were gained by employing ColabFold (Mirdita et al. 2022) in combination with the state-of-the-art protein structure prediction tool, AlphaFold2 (Jumper et al. 2021). The viral genomes were annotated using Pharokka (Bouras et al. 2022), a tool specifically designed for the annotation of bacteriophage genomes. Taxonomy was assigned to the viral sequences using vConTACT2 (Bin Jang et al. 2019) and the MMseqs2 taxonomy (Mirdita et al. 2021) based on sequence similarity between query and reference viral proteins. The genome structure of the analyzed viruses and respective bacteria were compared using Clinker (Gilchrist and Chooi 2021).

3 ViroProfiler: a containerized bioinformatics pipeline for viral metagenomic data analysis

Authors: **Jinlong Ru**, Mohammadali Khan Mirzaei, Jinling Xue, Xue Peng and Li Deng

Published in *Gut Microbes* 15, 2192522 (2023)

DOI: 10.1080/19490976.2023.2192522

3.1 Abstract

Bacteriophages play central roles in the maintenance and function of most ecosystems by regulating bacterial communities. Yet, our understanding of their diversity remains limited due to the lack of robust bioinformatics standards. Here we present ViroProfiler, an in-silico workflow for analyzing shotgun viral metagenomic data. ViroProfiler can be executed on a local Linux computer or cloud computing environments. It uses the containerization technique to ensure computational reproducibility and facilitate collaborative research. ViroProfiler is freely available at https://github.com/deng-lab/viroprofiler.

3.2 Contribution

J.R. developed the software. M.K.M. and J.R. drafted the manuscript. J.R and X.P. performed the analyses. J.X. wrote the documentation. M.K.M. and L.D. conceived and supervised the project. All authors reviewed and approved the manuscript.

4 Differences in Gut Virome Related to Barrett Esophagus and Esophageal Adenocarcinoma

Authors: Tianli Ma#, **Jinlong Ru**#, Jinling Xue, Sarah Schulz, Mohammadali Khan Mirzaei, Klaus-Peter Janssen, Michael Quante and Li Deng

Published in *Microorganisms* 9,1701 (2021)

DOI: 10.3390/microorganisms9081701

4.1 Abstract

The relationship between viruses (dominated by bacteriophages or phages) and lower gastrointestinal (GI) tract diseases has been investigated, whereas the relationship between gut bacteriophages and upper GI tract diseases, such as esophageal diseases, which mainly include Barrett’s esophagus (BE) and esophageal adenocarcinoma (EAC), remains poorly described. This study aimed to reveal the gut bacteriophage community and their behavior in the progression of esophageal diseases. In total, we analyzed the gut phage community of sixteen samples from patients with esophageal diseases (six BE patients and four EAC patients) as well as six healthy controls. Differences were found in the community composition of abundant and rare bacteriophages among the three groups. In addition, the auxiliary metabolic genes (AMGs) related to bacterial exotoxin and virulence factors such as lipopolysaccharides (LPS) biosynthesis proteins were found to be more abundant in the genome of rare phages from BE and EAC samples compared to the controls. These results suggest that the community composition of gut phages and functional traits encoded by them were different in two stages of esophageal diseases. However, the findings from this study need to be validated with larger sample sizes in the future.

4.2 Contribution

Conceptualization, K.-P.J., M.Q. and L.D.; methodology, T.M.; formal analysis, T.M., J.R., and M.K.M.; investigation, T.M. and J.X.; writing, T.M., J.R., S.S., J.X., M.K.M., K.-P.J., M.Q. and L.D. All authors have read and agreed to the published version of the manuscript.

5 Unveiling the hidden role of aquatic viruses in hydrocarbon pollution bioremediation

Authors: **Jinlong Ru**#, Jinling Xue#, Jianfeng Sun, Linda Cova and Li Deng

Published in *Journal of Hazardous Materials* 459, 132299.(2023)

DOI: 10.1016/j.jhazmat.2023.132299

5.1 Abstract

Hydrocarbon pollution poses substantial environmental risks to water and soil. Bioremediation, which utilizes microorganisms to manage pollutants, offers a cost-effective solution. However, the role of viruses, particularly bacteriophages (phages), in bioremediation remains unexplored. This study examines the diversity and activity of hydrocarbon-degradation genes encoded by environmental viruses, focusing on phages, within public databases. We identified 57 high-quality phage-encoded auxiliary metabolic genes (AMGs) related to hydrocarbon degradation, which we refer to as virus-encoded hydrocarbon degradation genes (vHYDEGs). These genes are encoded by taxonomically diverse aquatic phages and highlight the under-characterized global virosphere. Six protein families involved in the initial alkane hydroxylation steps were identified. Phylogenetic analyses revealed the diverse evolutionary trajectories of vHYDEGs across habitats, revealing previously unknown biodegraders linked evolutionarily with vHYDEGs. Our findings suggest phage AMGs may contribute to alkane and aromatic hydrocarbon degradation, participating in the initial, rate-limiting hydroxylation steps, thereby aiding hydrocarbon pollution bioremediation and promoting their propagation. To support future research, we developed vHyDeg, a database containing identified vHYDEGs with comprehensive annotations, facilitating the screening of hydrocarbon degradation AMGs and encouraging their bioremediation applications.

5.2 Contribution

Jinlong Ru: Methodology, Software, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. Jinling Xue: Conceptualization, Investigation, Data curation, Writing

- original draft, Writing - review & editing. Jianfeng Sun: Visualization. Linda Cova: Visualization. Li Deng: Conceptualization, Supervision, Funding acquisition.

6 Further contributions

6.1 Bacteriophages Isolated from Stunted Children Can Regulate Gut Bacterial Communities in an Age-Specific Manner

Publication

Mohammadali Khan Mirzaei, Md. Anik Ashfaq Khan, Prakash Ghosh, Zofia E. Taranu, Mariia Taguer, **Jinlong Ru**, Rajashree Chowdhury, Md. Mamun Kabir, Li Deng, Dinesh Mondal, Corinne F. Maurice, 2020. Bacteriophages Isolated from Stunted Children Can Regulate Gut Bacterial Communities in an Age-Specific Manner. *Cell Host & Microbe* 27, 199-212.e5. https://doi.org/10.1016/j.chom.2020.01.004

Manuscript abstract

Stunting, a severe and multigenerational growth impairment, globally affects 22% of children under the age of 5 years. Stunted children have altered gut bacterial communities with higher proportions of Proteobacteria, a phylum with several known human pathogens. Despite the links between an altered gut microbiota and stunting, the role of bacteriophages, highly abundant bacterial viruses, is unknown. Here, we describe the gut bacterial and bacteriophage communities of Bangladeshi stunted children younger than 38 months. We show that these children harbor distinct gut bacteriophages relative to their non-stunted counterparts. In vitro, these gut bacteriophages are infectious and can regulate bacterial abundance and composition in an age-specific manner, highlighting their possible role in the pathophysiology of child stunting. Specifically, Proteobacteria from non-stunted children increased in the presence of phages from younger stunted children, suggesting that phages could contribute to the bacterial community changes observed in child stunting.

Selected contributions

I performed phage genome annotation and contributed to the results interpretation.

6.2 Altered virome structure and function characterization in Helicobacter pylori-driven colorectal carcinogenesis and *H. pylori* eradication

Publication

Shiqi Luo, Jinling Xue, **Jinlong Ru**, Mohammadali Khan Mirzaei, Anna Ralser, Raquel Mejias Luque, Markus Gerhard, Li Deng, 2022. Altered virome structrue and function characterization in *Helicobacter pylori*-driven colorectal carcinogenesis and *H. pylori* eradication. *bioRxiv* 2022.07. 03.498559. https://doi.org/10.1101/2022.07.03.498559

Manuscript abstract

The understanding of gut virome and its role in *Helicobacter pylori*-driven colorectal cancer (CRC), as well as the long-term impact of *H. pylori* eradication via antibiotic treatment on it could contribute to better understanding the mechanisms of the disruption of gut bacteriome homeostasis involved in *H. pylori*-driven colorectal carcinogenesis and antibiotic therapy for *H. pylori* eradication. In the dynamic analysis of viral genome shotgun metagenomic of samples from lower gastrointestinal tract of the Apc+/1638N and C57BL/6 mice with *H. pylori* infection and eradication, stable viral abundance, and replacement of bursted unique viral contigs in infected and uninfected Apc+/1638N mice were observed. Temperate phages, which encoding comprehensive microbial functional genes and targeting various susceptible hosts, were expanded extremely prior to cancer exacerbation. In addition, short-term antibiotic exposure for *H. pylori* eradication was able to alter the gut virome and thrive the antibiotic resistance genes (ARGs) in the viral genome for at least 6 months. Collectively, these results point toward a potential role of the altered, but dynamically balanced gut virome, characterized by the expanded temperate phages, in contributing to the *H. pylori*-driven CRC, and indicate that viral genome may act as ARG reservoir for the antibiotic resistance of bacteria after the antibiotics therapy to *H. pylori* eradication.

Selected contributions

I performed the viromic data analyses and contributed to the results interpretation.

6.3 Bacterial and Bacteriophage Consortia are Associated with Protective Intestinal Immuno-modulatory Metabolites in Allogeneic Stem Cell Transplantation Patients

Publication

Erik Thiele Orberg, Elisabeth Meedt, Andreas Hiergeist, Jinling Xue, Paul Heinrich, **Jinlong Ru**, Sakhila Ghimire, Oriana Miltiadous, Sarah Lindner, Melanie Tiefgraber, Sophia Gölde, Tina Eismann, Alix Schwarz, Sascha Göttert, Sebastian Jarosch, Katja Steiger, Christian Schulz, Michael Gigl, Julius C. Fischer, Klaus-Peter Janssen, Michael Quante, Simon Heidegger, Peter Herhaus, Mareike Verbeek, Jürgen Ruland, Marcel RM van den Brink, Daniela Weber, Matthias Edinger, Daniel Wolff, Dirk H. Busch, Karin Kleigrewe, Wolfgang Herr, Florian Bassermann, André Gessner, Li Deng, Ernst Holler, Hendrik Poeck2. Bacterial and Bacteriophage Consortia are Associated with Protective Intestinal Immuno-modulatory Metabolites in Allogeneic Stem Cell Transplantation Patients. (Submitted)

Manuscript abstract

The human microbiome has a direct effect on clinical outcome in patients undergoing allogeneic hematopoietic stem cell transplantation (allo-SCT). Besides bacteria, fungi and viruses as well as intestinal microbiota-derived metabolites are involved, but it is still unclear how dynamic shifts in these three kingdoms contribute to the production of intestinal metabolites, how metabolites are impacted by GvHD or antibiotics and whether they are associated with clinical outcome.

To address this, we performed a prospective, longitudinal study that combined three-kingdom (bacteria, fungi, viruses) analysis of intestinal microbial communities with targeted metabolomics in allo-SCT patients (n=78) at two different transplantation centers. Using Multi-omics factor analysis (MOFA), we uncovered a microbiome signature of bacteria from the Lachnospiraceae and Oscillospiraceae families and their associated bacteriophages, which correlated with the production of immuno-modulatory metabolites including short-chain fatty acids (SCFAs), branched-chain fatty acids (BCFA), metabolites associated with induction of type-I IFN signaling (IIMs) and immuno-modulatory secondary bile acids. We established an Immuno-modulatory

Metabolite Risk Index (IMM-RI) consisting of five index immuno-modulatory metabolites (IMMs), which was associated with improved survival, less transplant-related mortality and reduced relapse rate. Onset of GI-GvHD and exposure to antibiotics significantly impacted intestinal levels of protective IMMs.

Using whole shotgun metagenomic sequencing, we observed that in IMM-RI low-risk patients, sustained production of protective IMMs was associated with a high abundance of SCFA biosynthesis pathways, specifically butyric acid via butyryl-CoA:acetate CoA-transferase (BCoAT). Through genome assembly from viral metagenomic sequencing data, we detected two bacteriophages which encoded BCoAT as an auxiliary metabolic gene. They were more abundant in IMM-RI low-risk patients and positively correlated with butyric acid concentration, suggesting that these bacteriophages may modulate bacterial metabolite biosynthesis.

Our study identifies a specific microbiome signature associated with protective IMMs that could improve fecal microbiota transplantation (FMT) donor selection and provides a rationale for the development of engineered metabolite-producing consortia and defined metabolite combination drugs as novel microbiome-based therapies for cancer patients.

Selected contributions

I performed the viromic data analyses and contributed to the results interpretation.

6.4 Challenges of Studying the Human Virome – Relevant Emerging Technologies

Publication

Mohammadali Khan Mirzaei, Jinling Xue, Rita Costa, **Jinlong Ru**, Sarah Schulz, Zofia E. Taranu, Li Deng, 2021. Challenges of Studying the Human Virome – Relevant Emerging Technologies. *Trends in Microbiology 29*, 171–181. https://doi.org/10.1016/j.tim.2020.05.021

Manuscript abstract

In this review we provide an overview of current challenges and advances in bacteriophage research within the growing field of viromics. In particular, we discuss, from a human virome study perspective, the current and emerging technologies available, their limitations in terms of de novo discoveries, and possible solutions to overcome present experimental and computational biases associated with low abundance of viral DNA or RNA. We summarize recent breakthroughs in metagenomics assembling tools and single-cell analysis, which have the potential to increase our understanding of phage biology, diversity, and interactions with both the microbial community and the human body. We expect that these recent and future advances in the field of viromics will have a strong impact on how we develop phage-based therapeutic approaches.

Selected contributions

I wrote the “Current Tools and Viral Databases” section and contributed to the “Unknown Viruses and Discovery” and “Identifying Unculturable Phages’ Host Range” section.

7 General Discussion

Extensive research has demonstrated the critical role of microbial communities in both natural environments and the human body, particularly regarding their impact on ecosystem functioning. For instance, studies have established the significant contribution of the microbiome to biogeochemical processes such as the global carbon and nitrogen cycles, as well as its involvement in food web dynamics (Fraterrigo, Balser, and Turner 2006; Grossart et al. 2020; Orland et al. 2019). Furthermore, the microbiome plays a vital role in the development of the human immune system, regulation of immune response, protection against pathogens, and maintenance of a healthy balance of microorganisms (Ling et al. 2020; Zheng, Liwinski, and Elinav 2020).

To date, most of this research has centered on the composition and function of bacteria or fungi, with relatively few studies exploring the virome component of the microbiome. It has only been in recent decades that researchers have begun to investigate the viral community within the microbiome. The complexity of the virome, its small genome size (particularly in comparison to bacterial counterparts), diverse host range, and the absence of a universal 16S ribosomal RNA equivalent present challenge for studying the virome. Consequently, our understanding of the virome remains limited, with only a small number of viral sequences and genomes identified and a significant proportion of sequenced data still cannot be annotated. Recent viromic studies employing advanced technologies for high-throughput, deep genome sequencing, and data analysis have started to illuminate the composition of virome and their impact on human health and environment.

Despite the rapid progress in viromic techniques, several challenges persist, including the development of a comprehensive and easy-to-use computational pipeline. To advance the study of viromic research and explore the role of viruses, particularly bacteriophages, in human health and environmental ecology, this thesis first presents a computational pipeline designed to streamline the data analysis workflow. Subsequently, we conducted two studies to reveal the role of viruses in human disease and bioremediation of hydrocarbon pollution. In addition, we discuss supplementary work examining the role of viruses in human health. This discussion encompasses all dissertation topics and offers an outlook on future research directions.

7.1 Distinct challenges in the analysis of environmental and human viromic data

Environmental and human virome samples exhibit distinct characteristics, resulting in unique challenges for their analysis. One such challenge arises from differences in contamination (Jurasz, Pawłowski, and Perlejewski 2021). Virome samples contain a substantial number of host and other contaminants, which can negatively impact the accuracy and reliability of metagenomic analyses. It is essential to remove host and contaminant sequences from virome samples for precise analysis, as incomplete removal, especially of contaminants with high sequence similarity to viral genomes, can introduce bias into metagenomic analyses. For human viromes, the process of removing host contaminants is relatively straightforward due to the availability of the human reference genome. However, environmental virome samples often contain a diverse range of host and contaminant sequences that are not as easily removed, owing to the absence of comprehensive reference genomes for the various organisms present in these environments.

Another notable difference between environmental and human viromes is the level of diversity observed within these viral communities. Environmental viromes typically exhibit a much higher degree of viral diversity compared to human viromes. This increased diversity presents challenges when attempting to assemble viral sequences into contigs. In environmental virome samples, the high diversity often results in shorter contigs due to the presence of a large number of distinct viral genomes or low sequencing depth. Consequently, this caused difficulties in accurately annotating and characterizing the viral species present within these samples. In contrast, human viromes generally display lower levels of diversity compared to environmental samples. This reduced complexity allows for the application of cross-assembly techniques, which can lead to the generation of longer contigs. These longer contigs provide a more comprehensive representation of the viral genomes present within the sample, thereby enhancing the accuracy and resolution of downstream metagenomic analyses.

To further improve the analysis of environmental and human viromes, distinct analysis strategies specifically tailored to address the challenges associated with these sample types should be employed. For instance, the reference-based contamination removal step is implemented in the ViroProfiler pipeline for human samples. For environmental samples, while there are currently no standard methods to remove contaminants, the development of advanced assembly algorithms for highly diverse environmental virome samples in the future may lead to more accurate contig generation and improved characterization of viral communities. In ViroProfiler, we utilized two approaches to remove contamination. The first one is for samples containing contamination of a large number of human or other organisms with known genomes. In this case, contaminant reads are removed by aligning them to the reference genome. While the second approach applicable to all samples, involves assembling reads into contigs and then identifying viral contigs using multiple computational methods. This process relies on the accuracy of viral detection methods, which were discussed in the next section.

7.2 Challenges in viral sequence identification from viromic data

Current computational methods for identifying viral contigs from viromic samples can be primarily classified into four categories. The first category includes sequence similarity-based approaches, which compare metagenomic sequence reads to reference databases of known viral genomes or genes using BLAST (Camacho et al. 2009). These methods, however, exhibit limited sensitivity in identifying novel or divergent viruses due to low sequence similarity. Their efficacy depends on the completeness and accuracy of reference databases, which may not encompass all known virus sequences or could contain misannotations.

The second category comprises gene content-based approaches, leveraging the presence of specific viral marker genes or functional gene profiles to identify viruses within metagenomic samples. Popular tools in this category include VirSorter (Roux et al. 2015) and MARVEL (Amgarten et al. 2018). These methods may fail to detect viruses that lack the selected marker genes or exhibit significant gene content variability. They may also produce false positives if the chosen marker genes are shared by other non-virus mobile genetic elements.

The third category encompasses k-mer frequency-based approaches, which identify viruses by analyzing the frequency of k-mers in metagenomic data. These methods operate under the assumption that the k-mer frequency distribution differs between viruses and non-viruses. Examples of such tools are Kraken (Lu et al. 2022; Wood, Lu, and Langmead 2019) and CLARK (Ounit et al. 2015). However, these methods are sensitive to sequencing errors and compositional biases, which may result in incorrect assignments, and they depend on the accuracy and completeness of reference k-mer databases.

The final category involves machine learning-based approaches, which train machine learning models to distinguish between phage and non-virus sequences based on sequence features or genomic properties. Examples of these tools are VirFinder (Ren et al. 2017) and DeepVirFinder (Ren et al. 2020). These methods also rely on the quality of training datasets, which may not encompass the full diversity of viral genomes, and they bear the potential for overfitting or poor generalization to novel or divergent viruses.

Recent benchmarking of virus identification tools suggests that each tool has its own advantages and limitations (Ho et al. 2023; Schackart et al. 2023), because they usually use one or a few virus-related features. Most of them strongly rely on known virus reference databases, thus failing to detect unknown viruses in the sample. To overcome this limitation, we combined multiple virus identification tools in ViroProfiler to improve sensitivity. Using this approach, we re-analyzed 10% samples of a published virome dataset and identified more new phages than the original study. Characterization of the newly identified phages revealed novel functions related to disease status (see Manuscript 1). Despite the availability of multiple virus identification tools, benchmarking different tools remains challenging. We anticipate that the establishment of standardized protocols and benchmark datasets for viral sequence identification will enable the comparison of various virus identification tools, helping researchers select and develop more advanced tools.

7.3 Studying the role of bacteriophages from the perspective of microbial community

With the identification of viral contigs/genomes from metagenomic samples, researchers can study the role of bacteriophages in the context of microbial communities. The most straightforward and intuitive way to understand the role of the phage community in a defined system is to analyze its composition, including documentation of virome richness and diversity, as well as taxonomy annotation using reference phage databases to investigate variations in community structure. Several studies have already found altered phage composition in diseased individuals compared to healthy ones, such as the significantly increased bacteriophage community diversity in colorectal cancer (CRC) patients, and clinical stage-dependent development of the enteric virome observed in partial redundancy analysis of virome species-level profiles (Nakatsu et al. 2018). Another example is childhood obesity and metabolic syndrome, where gut virome alterations in diversity and richness were observed (Bikel et al. 2021). In IBD, increased Caudovirales taxonomic richness in enteric virome were onserved ans thought to be associated with the disease (Norman et al. 2015). These early observations brought public attention to the importance of vieome in diseases. However, these studies often provide descriptive results for entire or partial bacteriophage communities, without in-depth functional analysis, in this period, researchers found community structure variations, and related these variations with disease status, but we still don't know who did what, and how they did these in the entire virome in short, no causality could be established.

Moreover, the traditional virus taxonomy classification system, which relies on morphology and virus host information, is not reliable since it requires virus cultivation in a lab. With the identification of more metagenomic-assembled viruses, it is necessary to consider virus genomes and use genome information to classify virus taxonomy. As a result, the International Committee on Taxonomy of Viruses (ICTV) has developed a new classification system for virus taxonomy. In ViroProfiler, we offer both the traditional taxonomy classification system from NCBI and the new taxonomy classification system from ICTV. Users can select the one that suits their study purpose. The accuracy of taxonomy classification will improve in the future as new reference virus genomes are added.

To overcome the limitations of recording only community structure variations, we conducted a study on the gut virome of patients with Barrett's esophagus (BE) and esophageal adenocarcinoma (EAC). We explored the virome structure variation and observed a disease-associated increase in diversity. To understand the contribution and function of the virome beyond the altered structure, we attempted to predict the hosts of viral contigs. We observed an expanded host range of rare viruses in the disease group, suggesting that this host-range expansion could facilitate horizontal gene transfer between hosts. Additionally, we discovered bacterial toxin-related auxiliary metabolic genes (AMGs) in phage contigs with higher relative abundance. This finding leads us to hypothesize that the spreading of these genes by phages might be one possible explanation for disease progression. Furthermore, we used BACPHLIP to predict the viral lifestyle, allowing us to further explore the behavior of bacteriophages and their potential role in disease development. Our results showed an increase in lysogenic replication cycles in the EAC group, indicating an enhanced likelihood of phage-mediated horizontal gene transfer. By combining variations in community structure, host range expansion, and altered lifestyle in disease groups, we provide more evidence on how phages could manipulate host bacteria and contribute to disease.

7.4 Studying the role of bacteriophages from the perspective of protein functions

Another perspective for studying the role of bacteriophages is at the molecular level. Understanding the molecular mechanisms by which viruses infect and replicate within their hosts can shed light on viral pathogenesis, host defense mechanisms, development of novel therapeutic strategies and environmental applications.

Phages have been shown to encode auxiliary metabolic genes (AMGs), which have been proven to be functional by experimental validation or metatranscriptomic data. One example is the photosynthesis gene *psbA* observed in cyanophages, which has a highly conserved amino acid sequence. Photosynthesis continues during infection despite the decline in expression of host photosynthesis genes, providing experimental evidence of the function of phage-encoded genes (Lindell et al. 2005). Another example is the *pmoC* gene discovered in large freshwater phages, which is potentially involved in methane oxidation. Transcriptional data show that the *pmoC* genes were highly expressed alongside genes encoding phage DNA packaging and particle assembly-related proteins (Chen et al. 2020). Our ViroProfiler pipeline can identify AMGs, providing a proxy for phage functional analyses at the molecular level. For instance, in Manuscript 2, we identified AMGs encoding bacterial toxins, such as *spyA*, *tccC*, *entB*, and *entD* genes. These were more abundant in the genomes of rare phages in BE and EAC patients and are involved in microbial cellular processes. This finding provides valuable evidence about the function of rare phages in disease. Although the relative abundances of these AMGs are low, and the statistical analysis is largely restrained due to the limited sample number, we did find some trends. Higher levels of these toxin-related AMGs were often associated with higher levels in the disease group, especially in EAC. In Manuscript 3, we identified hydrocarbon degradation AMGs in viral contigs from a public database. We predicted their protein structures and compared them with other experimentally confirmed bacterial protein structures, discovering high structural similarity and the existence of catalytic binding sites. This indicates maintained catalytic functions. In the phylogenetic analysis based on the protein sequences of these AMGs, we observed the evolutionary closeness between our vHYDEGs-encoded enzymes and those from identified bacterial degraders. Moreover, the function of these viral contigs was more related to their habitat rather than their taxonomy. This information provides insight into the potential role of viruses in hydrocarbon pollution bioremediation.

A limitation of AMG identification methods is that the function and activity of these identified genes are not guaranteed, and experimental work is usually required to further confirm their enzymatic potential. Another limitation is that current methods for AMG identification mainly rely on sequence-based similarity searches, such as BLAST analysis. This approach is limited by the fact that AMGs are subject to high evolutionary pressure and rapid diversification, often resulting in sequences that do not exhibit significant similarity to their homologues in prokaryotic hosts. For example, in Manuscript 3, we showed that viral proteins could have low sequence similarity with their prokaryotic homologs but retained high structural similarity. Since similar protein structures indicate similar protein functions, this finding suggests that we can integrate protein structure information into the detection and annotation of AMGs.

Fortunately, recent advancements in protein structure prediction, such as AlphaFold (Jumper et al. 2021) and RoseTTAFold (Baek et al. 2021), have provided a new approach to annotating protein functions using structural information. Related structural comparison tools like Foldseek (van Kempen et al. 2022) enable rapid searches for proteins with similar structures in reference protein structural databases. This method, known as structure-based functional annotation, allows for the identification of remote homologous proteins even when they do not share significant sequence similarity. Since protein structures are more conserved than sequences, as demonstrated in Manuscript 3, we expect that structure-based homologous detection methods will help us discover more virus-encoded functional auxiliary metabolic genes (AMGs). During the writing of this work, a preprint study utilized this idea and annotated more phage proteins using structural information than by relying solely on sequence information (Say et al. 2023). We anticipate that as more phage protein structures are predicted, an increasing number of functional genes, especially AMGs, will be discovered from phage genomes, thus expanding our understanding of the role of phages in the environment and human health. In addition, the detection and functional annotation of more novel phage genes and proteins will provide greater amounts of training data for designing advanced computational models to annotate phage genes.

7.5 Advantages and limitations of current viromic data analyses tools

Numerous tools for virome sequencing data analyses have been developed, with most of them addressing specific problems, such as detecting viruses from reads or contigs and annotating identified virus contigs or genomes. For users to gain a holistic understanding of viruses in their samples, they must decide which analyses to perform, which software to choose, and how to parse and integrate results from different analysis steps. In many cases, they also need to spend considerable time installing or even compiling the software. Even after analyzing their data using a combination of multiple software, reproducing these analyses can be challenging for other researchers, since computational platforms, software versions, and even database versions cannot be guaranteed to be the same as those used by the original author. Moreover, minor differences in parameter settings can lead to significant differences in results.

ViroProfiler is an attempt to address these problems. It is a well-designed workflow based on a modern computational analysis pipeline framework and well-established viromic analysis procedures. Reproducible analysis is ensured using version control and containerization techniques. The analysis steps are based on the requirements of the Minimum Information about an Uncultivated Virus Genome (MIUViG) standard (Roux et al. 2019). With ViroProfiler, manually installing software on high-performance computing (HPC) or other cloud computing platforms is no longer necessary. Furthermore, the output of the pipeline can be combined and visualized using multiple tools integrated into ViroProfiler or the companion R package, vpfkit.

Despite its advantages, ViroProfiler has certain limitations that must be addressed to improve its overall performance. For instance, databases used in multiple steps, such as DRAM-v and VIBRANT, have redundancies due to their reliance on VOGDB and PFAM databases. This overlap not only increases the size of the databases but also results in longer computational time and greater resource consumption. Moreover, conflicting versions of the same database within the pipeline can raise concerns about annotation consistency. Another limitation of our current pipeline is its primary focus on analyzing Illumina sequencing data. As the accuracy of third-generation sequencing technologies improves and their costs decrease, an increasing number of viromic studies are leveraging platforms such as PacBio and Nanopore long-read sequencing. These platforms offer several benefits, including the generation of longer reads and the ability to resolve complex genomic regions, making them particularly well-suited for viromic studies. To address this limitation, we plan to incorporate modules related to third-generation sequencing data analysis into a new sub-workflow in ViroProfiler. In the future, comprehensive and universal analysis of all sequencing platforms will be supported. The modular design of the Nextflow and nf-core framework facilitates this integration, enabling easier extension and customization of the pipeline to meet individuals' specific needs. The inclusion of these advanced sequencing technologies in future versions of ViroProfiler will significantly enhance its performance and utility.

7.6 Future work

We anticipate the widespread use of ViroProfiler in the future has the potential to significantly contribute to the development of a more comprehensive phage reference database. Phage detection and annotation have long been heavily dependent on these databases, as they provide a wealth of information that researchers rely on to accurately identify and classify phages. Although reference-independent methods, such as machine learning-based approaches, have made significant strides, they still require accurate phage annotation in reference databases for training purposes. As these studies uncover new viral sequences, databases can be updated with the latest information, thereby improving the accuracy and depth of phage detection. This updated knowledge can then be fed into phage identification algorithms to detect even more phage sequences from viromic samples. This creates a circular feedback loop that can rapidly speed up phage research. To facilitate the quick iteration of the feedback loop, one possible approach could be to create a cloud-based platform that automatically updates phage databases in real-time as new data becomes available. Furthermore, integrating machine learning models could further streamline the annotation process, detect novel phage sequences, and predict their properties, such as host range, virulence factors, and potential applications in biotechnology or medicine. These predictions could then be validated by experimental data, further refining the accuracy of the machine-learning models, and contributing to the expansion of the reference databases.

An area of future development for ViroProfiler is the creation of advanced visualization tools to present results in a more intuitive and accessible manner. Recognizing that the visualization of results should be easily executed on personal computers, we have separated this module from the main ViroProfiler pipeline and developed it as an R package. The R package includes essential functions for downstream analysis of outputs from the ViroProfiler pipeline and a Shiny App for user-friendly visualization of the results. The Shiny App is also packaged into a Docker container, allowing users to install and run it with a single command on any computer with Docker installed. The current implementation of the package and Shiny App provides basic functions and visualizations. Additionally, we plan to expand its functionality and visualization interfaces based on research requirements and user feedback in the future.

Another long-term plan for ViroProfiler is to integrate it with real-time sequencing techniques such as Nanopore sequencing. It is promising to upload the Nonapore sequencing results to a cloud computing platform, where ViroProfiler can be run automatically so that users can get results in real time. Since the ViroProfiler pipeline can be easily deployed on a cloud computation platform using a single command, it is very suitable for real-time analyses when users only need to upload raw sequencing results to the cloud computation platform with pre-defined configuration settings. It is especially suitable for studying viromes at the community level. With detailed metadata such as sample sources, the virome profiling results can be classified based on their origin, for example, based on habitat, ecosystem, or patient cohorts classified by ages, diseases, and other metadata. This classification of the viral genomes and their genomic product will facilitate more adjusted computational models for future research.

8 Conclusions and outlook

This study presents a standardized and reproducible computational pipeline for viromic data analysis. We applied this pipeline to human virome samples to investigate the role of bacteriophages in Barrett's esophagus and esophageal adenocarcinoma. We also conducted a study on virus-encoded hydrocarbon degradation genes and investigated their potential role in bioremediation. In other work, we applied NGS and viromic technology to potential phage therapy and provided a review of current technology for studying human virome.

In future research, one promising area is the use of bacteriophages in biotechnology and gene therapy. This rapidly developing field has many potential applications, such as using phages as vectors to deliver therapeutic genes to target cells, designing more efficient enzymes using phage-encoded proteins, and more. Moreover, phage therapy is gaining increasing attention as a potential alternative to traditional antibiotics, which could have significant implications for the treatment of bacterial infections. Further exploration of the mechanisms underlying the interactions between phages and host organisms could lead to new discoveries and applications in virology, medicine, and biotechnology. Thus, developing more advanced computational tools for analyzing viromic data and studying their interactions will be crucial in advancing our understanding of the role of phages in the environment and health, and pave the way for their vast applications.

References

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson, and Christopher Quince. 2014. "Binning Metagenomic Contigs by Coverage and Composition." *Nature Methods* 11(11):1144–46.

Amgarten, Deyvid, Lucas P. P. Braga, Aline M. da Silva, and João C. Setubal. 2018. "MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins." *Frontiers in Genetics* 9.

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. 2021. "Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network." *Science (New York, N.Y.)*.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. 2012. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology* 19(5):455–77.

Barker, Michelle, Neil P. Chue Hong, Daniel S. Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, Leyla Jael Castro, Morane Gruenpeter, Paula Andrea Martinez, and Tom Honeyman. 2022. "Introducing the FAIR Principles for Research Software." *Scientific Data* 9(1):622.

Barr, Jeremy J. 2017. "A Bacteriophages Journey through the Human Body." *Immunological Reviews* 279(1):106–22.

Besemer, John, Alexandre Lomsadze, and Mark Borodovsky. 2001. "GeneMarkS: A Self-Training Method for Prediction of Gene Starts in Microbial Genomes. Implications for Finding Sequence Motifs in Regulatory Regions." *Nucleic Acids Research* 29(12):2607–18.

Bikel, Shirley, Gamaliel López-Leal, Fernanda Cornejo-Granados, Luigui Gallardo-Becerra, Rodrigo García-López, Filiberto Sánchez, Edgar Equihua-Medina, Juan Pablo Ochoa-Romo, Blanca Estela López-Contreras, Samuel Canizales-Quinteros, Abigail Hernández-Reyna, Alfredo Mendoza-Vargas, and Adrian Ochoa-Leyva. 2021. "Gut DsDNA Virome Shows Diversity and Richness Alterations Associated to Childhood Obesity and Metabolic Syndrome." *IScience* 102900.

Bin Jang, Ho, Benjamin Bolduc, Olivier Zablocki, Jens H. Kuhn, Simon Roux, Evelien M. Adriaenssens, J. Rodney Brister, Andrew M. Kropinski, Mart Krupovic, Rob Lavigne, Dann Turner, and Matthew B. Sullivan. 2019. "Taxonomic Assignment of Uncultivated Prokaryotic Virus Genomes Is Enabled by Gene-Sharing Networks." *Nature Biotechnology* 37(6):632–39.

Bolduc, Benjamin, Ho Bin Jang, Guilhem Doulcier, Zhi-Qiang You, Simon Roux, and Matthew B. Sullivan. 2017. "VConTACT: An IVirus Tool to Classify Double-Stranded DNA Viruses That Infect Archaea and Bacteria." *PeerJ* 5:e3243.

Bonnain, Chelsea, Mya Breitbart, and Kristen N. Buck. 2016. "The Ferrojan Horse Hypothesis: Iron-Virus Interactions in the Ocean." *Frontiers in Marine Science* 3.

Bouras, George, Roshan Nepal, Ghais Houtak, Alkis James Psaltis, Peter-John Wormald, and Sarah Vreugde. 2022. "Pharokka: A Fast Scalable Bacteriophage Annotation Tool." *Bioinformatics (Oxford, England)* btac776.

Breitbart, Mya, Chelsea Bonnain, Kema Malki, and Natalie A. Sawaya. 2018. "Phage Puppet Masters of the Marine Microbial Realm." *Nature Microbiology* 3(7):754–66.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12(1):59–60.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10(1):421.

Camargo, Antonio Pedro, Stephen Nayfach, I-Min. A. Chen, Krishnaveni Palaniappan, Anna Ratner, Ken Chu, Stephan J. Ritter, T. B. K. Reddy, Supratim Mukherjee, Frederik Schulz, Lee Call, Russell Y. Neches, Tanja Woyke, Natalia N. Ivanova, Emiley A. Eloe-Fadrosh, Nikos C. Kyrpides, and Simon Roux. 2022. "IMG/vr V4: An Expanded Database of Uncultivated Virus Genomes within a Framework of Extensive Functional, Taxonomic, and Ecological Metadata." *Nucleic Acids Research* gkac1037.

Cantalapiedra, Carlos P., Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. 2021. "EggNOG-Mapper V2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale." *Molecular Biology and Evolution* (msab293).

Chen, Lin-Xing, Raphaël Méheust, Alexander Crits-Christoph, Katherine D. McMahon, Tara Colenbrander Nelson, Gregory F. Slater, Lesley A. Warren, and Jillian F. Banfield. 2020. "Large Freshwater Phages with the Potential to Augment Aerobic Methane Oxidation." *Nature Microbiology* 1–12.

Chen, Qian, Xiaojing Ma, Chong Li, Yun Shen, Wei Zhu, Yan Zhang, Xiaokui Guo, Jian Zhou, and Chang Liu. 2021. "Enteric Phageome Alterations in Patients with Type 2 Diabetes." *Frontiers in Cellular and Infection Microbiology* 10.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. "Fastp: An Ultra-Fast All-In-One FASTQ Preprocessor." *Bioinformatics (Oxford, England)* 34(17):i884–90.

Chevallereau, Anne, Benoît J. Pons, Stineke van Houte, and Edze R. Westra. 2021. "Interactions between Bacterial and Phage Communities in Natural Environments." *Nature Reviews Microbiology* 1–14.

Cook, Ryan, Nathan Brown, Branko Rihtman, Slawomir Michniewski, Tamsin Redgwell, Martha Clokie, Dov J. Stekel, Yin Chen, David J. Scanlan, Jon L. Hobman, Andrew Nelson, Michael A. Jones, Darren Smith, and Andrew Millard. 2023. "The Long and Short of It: Benchmarking Viromics Using Illumina, Nanopore and PacBio Sequencing Technologies."

Crummett, Lisa T., Richard J. Puxty, Claudia Weihe, Marcia F. Marston, and Jennifer B. H. Martiny. 2016. "The Genomic Content and Context of Auxiliary Metabolic Genes in Marine Cyanomyoviruses." *Virology* 499:219–29.

Deng, Li, J. Cesar Ignacio-Espinoza, Ann C. Gregory, Bonnie T. Poulos, Joshua S. Weitz, Philip Hugenholtz, and Matthew B. Sullivan. 2014. "Viral Tagging Reveals Discrete Populations in Synechococcus Viral Genome Sequence Space." *Nature* 513(7517):242–45.

Di Tommaso, Paolo, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. "Nextflow Enables Reproducible Computational Workflows." *Nature Biotechnology* 35(4):316–19.

Djaffardjy, Marine, George Marchment, Clémence Sebe, Raphaël Blanchet, Khalid Belhajjame, Alban Gaignard, Frédéric Lemoine, and Sarah Cohen-Boulakia. 2023. "Developing and Reusing Bioinformatics Data Analysis Pipelines Using Scientific Workflow Systems." *Computational and Structural Biotechnology Journal* 21:2075–85.

Dohm, Juliane C., Philipp Peters, Nancy Stralis-Pavese, and Heinz Himmelbauer. 2020. "Benchmarking of Long-Read Correction Methods." *NAR Genomics and Bioinformatics* 2(2):lqaa037.

Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLOS Computational Biology* 7(10):e1002195.

Edwards, Robert A., Katelyn McNair, Karoline Faust, Jeroen Raes, and Bas E. Dutilh. 2016. "Computational Approaches to Predict Bacteriophage–Host Relationships" edited by M. Smith. *FEMS Microbiology Reviews* 40(2):258–72.

Ewels, Philip A., Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2020. "The Nf-Core Framework for Community-Curated Bioinformatics Pipelines." *Nature Biotechnology* 38(3):276–78.

Federici, Sara, Sharon Kredo-Russo, Rafael Valdés-Mas, Denise Kviatcovsky, Eyal Weinstock, Yulia Matiuhin, Yael Silberberg, Koji Atarashi, Munehiro Furuichi, Akihiko Oka, Bo Liu, Morine Fibelman, Iddo Nadav Weiner, Efrat Khabra, Nyssa Cullin, Noa Ben-Yishai, Dana

Inbar, Hava Ben-David, Julian Nicenboim, Noga Kowalsman, Wolfgang Lieb, Edith Kario, Tal Cohen, Yael Friedman Geffen, Lior Zelcbuch, Ariel Cohen, Urania Rappo, Inbar Gahali-Sass, Myriam Golembo, Vered Lev, Mally Dori-Bachash, Hagit Shapiro, Claudia Moresi, Amanda Cuevas-Sierra, Gayatree Mohapatra, Lara Kern, Danping Zheng, Samuel Philip Nobs, Jotham Suez, Noa Stettner, Alon Harmelin, Naomi Zak, Sailaja Puttagunta, Merav Bassan, Kenya Honda, Harry Sokol, Corinna Bang, Andre Franke, Christoph Schramm, Nitsan Maharshak, Ryan Balfour Sartor, Rotem Sorek, and Eran Elinav. 2022. "Targeted Suppression of Human IBD-Associated Gut Microbiota Commensals by Phage Consortia for Treatment of Intestinal Inflammation." *Cell* 185(16):2879-2898.e24.

Fraterrigo, Jennifer M., Teri C. Balser, and Monica G. Turner. 2006. "Microbial Community Variation and Its Relationship with Nitrogen Mineralization in Historically Altered Forests." *Ecology* 87(3):570–79.

Galiez, Clovis, Matthias Siebert, François Enault, Jonathan Vincent, and Johannes Söding. 2017. "WIsH: Who Is the Host? Predicting Prokaryotic Hosts from Metagenomic Phage Contigs." *Bioinformatics (Oxford, England)* 33(19):3113–14.

Gao, Yang, Yao Lu, Jennifer A. J. Dungait, Jianbao Liu, Shunhe Lin, Junjie Jia, and Guirui Yu. 2022. "The 'Regulator' Function of Viruses on Ecosystem Carbon Cycling in the Anthropocene." *Frontiers in Public Health* 10.

Gasper, Raphael, Julia Schwach, Jana Hartmann, Andrea Holtkamp, Jessica Wiethaus, Natascha Riedel, Eckhard Hofmann, and Nicole Frankenberg-Dinkel. 2017. "Auxiliary Metabolic Genes- Distinct Features of Cyanophage-Encoded T-Type Phycobiliprotein Lyase ΘCpeT." *Journal of Biological Chemistry* jbc.M116.769703.

Gaudin, Raphaël, and Natasha S. Barteneva. 2015. "Sorting of Small Infectious Virus Particles by Flow Virometry Reveals Distinct Infectivity Profiles." *Nature Communications* 6(1):6022.

Gilchrist, Cameron L. M., and Yit-Heng Chooi. 2021. "Clinker & Clustermap.Js: Automatic Generation of Gene Cluster Comparison Figures." *Bioinformatics (Oxford, England)* 37(16):2473–75.

Gorbalenya, Alexander E., Mart Krupovic, Arcady Mushegian, Andrew M. Kropinski, Stuart G. Siddell, Arvind Varsani, Michael J. Adams, Andrew J. Davison, Bas E. Dutilh, Balázs Harrach, Robert L. Harrison, Sandra Junglen, Andrew M. Q. King, Nick J. Knowles, Elliot J. Lefkowitz, Max L. Nibert, Luisa Rubino, Sead Sabanadzovic, Hélène Sanfaçon, Peter Simmonds, Peter J. Walker, F. Murilo Zerbini, Jens H. Kuhn, and International Committee on Taxonomy of Viruses Executive Committee. 2020. "The New Scope of Virus Taxonomy: Partitioning the Virosphere into 15 Hierarchical Ranks." *Nature Microbiology* 5(5):668–74.

Grazziotin, Ana Laura, Eugene V. Koonin, and David M. Kristensen. 2017. "Prokaryotic Virus Orthologous Groups (PVOGs): A Resource for Comparative Genomics and Protein Family Annotation." *Nucleic Acids Research* 45(D1):D491–98.

Grossart, Hans-Peter, Ramon Massana, Katherine D. McMahon, and David A. Walsh. 2020. "Linking Metagenomics to Aquatic Microbial Ecology and Biogeochemical Cycles." *Limnology and Oceanography* 65(S1):S2–20.

Guo, Jiarong, Ben Bolduc, Ahmed A. Zayed, Arvind Varsani, Guillermo Dominguez-Huerta, Tom O. Delmont, Akbar Adjie Pratama, M. Consuelo Gazitúa, Dean Vik, Matthew B. Sullivan, and Simon Roux. 2021. "VirSorter2: A Multi-Classifier, Expert-Guided Approach to Detect Diverse DNA and RNA Viruses." *Microbiome* 9(1):37.

Head, Ian M., D. Martin Jones, and Wilfred F. M. Röling. 2006. "Marine Microorganisms Make a Meal of Oil." *Nature Reviews Microbiology* 4(3):173–82.

Hendrix, Roger W., Margaret C. M. Smith, R. Neil Burns, Michael E. Ford, and Graham F. Hatfull. 1999. "Evolutionary Relationships among Diverse Bacteriophages and Prophages: All the World's a Phage." *Proceedings of the National Academy of Sciences* 96(5):2192–97.

Hernández-Plaza, Ana, Damian Szklarczyk, Jorge Botas, Carlos P. Cantalapiedra, Joaquín Giner-Lamia, Daniel R. Mende, Rebecca Kirsch, Thomas Rattei, Ivica Letunic, Lars J. Jensen, Peer Bork, Christian von Mering, and Jaime Huerta-Cepas. 2022. "EggNOG 6.0: Enabling Comparative Genomics across 12 535 Organisms." *Nucleic Acids Research* gkac1022.

Ho, Siu Fung Stanley, Nicole E. Wheeler, Andrew D. Millard, and Willem van Schaik. 2023. "Gauge Your Phage: Benchmarking of Bacteriophage Identification Tools in Metagenomic Sequencing Data." *Microbiome* 11(1):84.

Hockenberry, Adam J., and Claus O. Wilke. 2021. "BACPHLIP: Predicting Bacteriophage Lifestyle from Conserved Protein Domains." *PeerJ* 9:e11396.

Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11:119.

Imelfort, Michael, Donovan Parks, Ben J. Woodcroft, Paul Dennis, Philip Hugenholtz, and Gene W. Tyson. 2014. "GroopM: An Automated Tool for the Recovery of Population Genomes from Related Metagenomes." *PeerJ* 2:e603.

Jahn, Martin T., Ksenia Arkhipova, Sebastian M. Markert, Christian Stigloher, Tim Lachnit, Lucia Pita, Anne Kupczok, Marta Ribes, Stephanie T. Stengel, Philip Rosenstiel, Bas E. Dutilh, and Ute Hentschel. 2019. "A Phage Protein Aids Bacterial Symbionts in Eukaryote Immune Evasion." *Cell Host & Microbe* 26(4):542-550.e5.

Jansson, Janet K., and Ruonan Wu. 2023. "Soil Viral Diversity, Ecology and Climate Change." *Nature Reviews Microbiology* 21(5):296–311.

Johansen, Joachim, Damian R. Plichta, Jakob Nybo Nissen, Marie Louise Jespersen, Shiraz A. Shah, Ling Deng, Jakob Stokholm, Hans Bisgaard, Dennis Sandris Nielsen, Søren J. Sørensen, and Simon Rasmussen. 2022. "Genome Binning of Viral Entities from Bulk Metagenomics Data." *Nature Communications* 13(1):965.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." *Bioinformatics (Oxford, England)* 30(9):1236–40.

Jover, Luis F., T. Chad Effler, Alison Buchan, Steven W. Wilhelm, and Joshua S. Weitz. 2014. "The Elemental Composition of Virus Particles: Implications for Marine Biogeochemical Cycles." *Nature Reviews Microbiology* 12(7):519–28.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 1–11.

Jurasz, Henryk, Tomasz Pawłowski, and Karol Perlejewski. 2021. "Contamination Issue in Viral Metagenomics: Problems, Solutions, and Clinical Perspectives." *Frontiers in Microbiology* 12.

Kanehisa, Minoru, Miho Furumichi, Yoko Sato, Mari Ishiguro-Watanabe, and Mao Tanabe. 2021. "KEGG: Integrating Viruses and Cellular Organisms." *Nucleic Acids Research* 49(D1):D545–51.

Kelley, David R., Bo Liu, Arthur L. Delcher, Mihai Pop, and Steven L. Salzberg. 2012. "Gene Prediction with Glimmer for Metagenomic Sequences Augmented by Classification and Clustering." *Nucleic Acids Research* 40(1):e9.

Khot, Varada, Jackie Zorz, Daniel A. Gittins, Anirban Chakraborty, Emma Bell, María A. Bautista, Alexandre J. Paquette, Alyse K. Hawley, Breda Novotnik, Casey R. J. Hubert, Marc Strous, and Srijak Bhatnagar. 2022. "CANT-HYD: A Curated Database of Phylogeny-Derived Hidden Markov Models for Annotation of Marker Genes Involved in Hydrocarbon Degradation." *Frontiers in Microbiology* 12.

Kieft, Kristopher, Alyssa Adams, Rauf Salamzade, Lindsay Kalan, and Karthik Anantharaman. 2022. "VRhyme Enables Binning of Viral Genomes from Metagenomes." *Nucleic Acids Research* 50(14):e83.

Kieft, Kristopher, Zhichao Zhou, and Karthik Anantharaman. 2020. "VIBRANT: Automated Recovery, Annotation and Curation of Microbial Viruses, and Evaluation of Viral Community Function from Genomic Sequences." *Microbiome* 8(1):90.

Kieft, Kristopher, Zhichao Zhou, Rika E. Anderson, Alison Buchan, Barbara J. Campbell, Steven J. Hallam, Matthias Hess, Matthew B. Sullivan, David A. Walsh, Simon Roux, and Karthik

Anantharaman. 2021. “Ecology of Inorganic Sulfur Auxiliary Metabolism in Widespread Bacteriophages.” *Nature Communications* 12(1):3503.

Kurtzer, Gregory M., Vanessa Sochat, and Michael W. Bauer. 2017. “Singularity: Scientific Containers for Mobility of Compute.” *PLOS ONE* 12(5):e0177459.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9(4):357–59.

Li, Wenjun, Kathleen R. O’Neill, Daniel H. Haft, Michael DiCuccio, Vyacheslav Chetvernin, Azat Badretdin, George Coulouris, Farideh Chitsaz, Myra K. Derbyshire, A. Scott Durkin, Noreen R. Gonzales, Marc Gwadz, Christopher J. Lanczycki, James S. Song, Narmada Thanki, Jiyao Wang, Roxanne A. Yamashita, Mingzhang Yang, Chanjuan Zheng, Aron Marchler-Bauer, and Françoise Thibaud-Nissen. 2021. “RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline Reach with Protein Family Model Curation.” *Nucleic Acids Research* 49(D1):D1020–28.

Li, Yanpeng, Xuemin Fu, Jinmin Ma, Jianhui Zhang, Yihong Hu, Wei Dong, Zhenzhou Wan, Qiongfang Li, Yi-Qun Kuang, Ke Lan, Xia Jin, Jian-Hua Wang, and Chiyu Zhang. 2019. “Altered Respiratory Virome and Serum Cytokine Profile Associated with Recurrent Respiratory Tract Infections in Children.” *Nature Communications* 10(1):2288.

Liang, Guanxiang, and Frederic D. Bushman. 2021. “The Human Virome: Assembly, Composition and Host Interactions.” *Nature Reviews Microbiology* 1–14.

Liang, Guanxiang, Maire A. Conrad, Judith R. Kelsen, Lyanna R. Kessler, Jessica Breton, Lindsey G. Albenberg, Sarah Marakos, Alissa Galgano, Nina Devas, Jessi Erlichman, Huanjia Zhang, Lisa Mattei, Kyle Bittinger, Robert N. Baldassano, and Frederic D. Bushman. 2020. “Dynamics of the Stool Virome in Very Early-Onset Inflammatory Bowel Disease.” *Journal of Crohn’s and Colitis*.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics (Oxford, England)* 30(7):923–30.

Lindell, Debbie, Jacob D. Jaffe, Zackary I. Johnson, George M. Church, and Sallie W. Chisholm. 2005. “Photosynthesis Genes in Marine Viruses Yield Proteins during Host Infection.” *Nature* 438(7064):86–89.

Ling, Fei, Natalie Steinel, Jesse Weber, Lei Ma, Chris Smith, Decio Correa, Bin Zhu, Daniel Bolnick, and Gaoxue Wang. 2020. “The Gut Microbiota Response to Helminth Infection Depends on Host Sex and Genotype.” *The ISME Journal* 14(5):1141–53.

Liu, Bo, Nan Shao, Jing Wang, SiYu Zhou, HaoXiang Su, Jie Dong, LiLian Sun, Li Li, Ting Zhang, and Fan Yang. 2020. “An Optimized Metagenomic Approach for Virome Detection of Clinical Pharyngeal Samples with Respiratory Infection.” *Frontiers in Microbiology* 11.

Lowe, Todd M., and Sean R. Eddy. 1997. "TRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence." *Nucleic Acids Research* 25(5):955–64.

Lu, Jennifer, Natalia Rincon, Derrick E. Wood, Florian P. Breitwieser, Christopher Pockrandt, Ben Langmead, Steven L. Salzberg, and Martin Steinegger. 2022. "Metagenome Analysis Using the Kraken Software Suite." *Nature Protocols* 1–25.

Lu, Zhenmei, Ye Deng, Joy D. Van Nostrand, Zhili He, James Voordeckers, Aifen Zhou, Yong-Jin Lee, Olivia U. Mason, Eric A. Dubinsky, Krystle L. Chavarria, Lauren M. Tom, Julian L. Fortney, Regina Lamendella, Janet K. Jansson, Patrik D'haeseleer, Terry C. Hazen, and Jizhong Zhou. 2012. "Microbial Gene Functions Enriched in the Deepwater Horizon Deep-Sea Oil Plume." *The ISME Journal* 6(2):451–60.

Ma, Yingfei, Xiaoyan You, Guoqin Mai, Taku Tokuyasu, and Chenli Liu. 2018. "A Human Gut Phage Catalog Correlates the Gut Phageome with Type 2 Diabetes." *Microbiome* 6(1):24.

Menzel, Peter, Kim Lee Ng, and Anders Krogh. 2016. "Fast and Sensitive Taxonomic Classification for Metagenomics with Kaiju." *Nature Communications* 7(1):11257.

Mirdita, M., M. Steinegger, F. Breitwieser, J. Söding, and E. Levy Karin. 2021. "Fast and Sensitive Taxonomic Assignment to Metagenomic Contigs." *Bioinformatics (Oxford, England)* (btab184).

Mirdita, Milot, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. 2022. "ColabFold: Making Protein Folding Accessible to All." *Nature Methods* 1–4.

Mirzaei, Mohammadali Khan, Jinling Xue, Rita Costa, Jinlong Ru, Sarah Schulz, Zofia E. Taranu, and Li Deng. 2021. "Challenges of Studying the Human Virome – Relevant Emerging Technologies." *Trends in Microbiology* 29(2):171–81.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, Lisanna Paladin, Shriya Raj, Lorna J. Richardson, Robert D. Finn, and Alex Bateman. 2021. "Pfam: The Protein Families Database in 2021." *Nucleic Acids Research* 49(D1):D412–19.

Mohan Raj, Juliet Roshini, and Indrani Karunasagar. 2019. "Phages amid Antimicrobial Resistance." *Critical Reviews in Microbiology* 45(5-6):701–11.

Mölder, Felix, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. 2021. *Sustainable Data Analysis with Snakemake*. 10:33. F1000Research.

Nakatsu, Geicho, Haokui Zhou, William Ka Kei Wu, Sunny Hei Wong, Olabisi Oluwabukola Coker, Zhenwei Dai, Xiangchun Li, Chun-Ho Szeto, Naoki Sugimura, Thomas Yuen-Tung Lam, Allen Chi-Shing Yu, Xiansong Wang, Zigui Chen, Martin Chi-Sang Wong, Siew Chien Ng, Matthew Tak Vai Chan, Paul Kay Sheung Chan, Francis Ka Leung Chan, Joseph

Jao-Yiu Sung, and Jun Yu. 2018. “Alterations in Enteric Virome Are Associated with Colorectal Cancer and Survival Outcomes.” *Gastroenterology* 155(2):529-541.e5.

Namiki, Toshiaki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. 2012. “MetaVelvet: An Extension of Velvet Assembler to de Novo Metagenome Assembly from Short Sequence Reads.” *Nucleic Acids Research* 40(20):e155.

Nayfach, Stephen, Antonio Pedro Camargo, Frederik Schulz, Emiley Eloe-Fadrosh, Simon Roux, and Nikos C. Kyrpides. 2021. “CheckV Assesses the Quality and Completeness of Metagenome-Assembled Viral Genomes.” *Nature Biotechnology* 39(5):578–85.

Norman, Jason M., Scott A. Handley, Megan T. Baldridge, Lindsay Droit, Catherine Y. Liu, Brian C. Keller, Amal Kambal, Cynthia L. Monaco, Guoyan Zhao, Phillip Fleshner, Thaddeus S. Stappenbeck, Dermot P. B. McGovern, Ali Keshavarzian, Ece A. Mutlu, Jenny Sauk, Dirk Gevers, Ramnik J. Xavier, David Wang, Miles Parkes, and Herbert W. Virgin. 2015. “Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease.” *Cell* 160(3):447–60.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. “MetaSPAdes: A New Versatile Metagenomic Assembler.” *Genome Research* 27(5):824–34.

Orland, Chloé, Erik J. S. Emilson, Nathan Basiliko, Nadia C. S. Mykytczuk, John M. Gunn, and Andrew J. Tanentzap. 2019. “Microbiome Functioning Depends on Individual and Interactive Effects of the Environment and Community Structure.” *The ISME Journal* 13(1):1–11.

Ounit, Rachid, Steve Wanamaker, Timothy J. Close, and Stefano Lonardi. 2015. “CLARK: Fast and Accurate Classification of Metagenomic and Genomic Sequences Using Discriminative K-Mers.” *BMC Genomics* 16(1):236.

Paysan-Lafosse, Typhaine, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A. Salazar, Maxwell L. Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, Julian Gough, Daniel H. Haft, Ivica Letunić, Aron Marchler-Bauer, Huaiyu Mi, Darren A. Natale, Christine A. Orengo, Arun P. Pandurangan, Catherine Rivoire, Christian J. A. Sigrist, Ian Sillitoe, Narmada Thanki, Paul D. Thomas, Silvio C. E. Tosatto, Cathy H. Wu, and Alex Bateman. 2022. “InterPro in 2022.” *Nucleic Acids Research* gkac993.

Peng, Xue, Jinlong Ru, Mohammadali Khan Mirzaei, and Li Deng. 2022. “Replidec - Use Naive Bayes Classifier to Identify Virus Lifecycle from Metagenomics Data.” *BioRxiv : The Preprint Server for Biology* 2022.07.18.500415.

Peng, Yu, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. 2012. “IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth.” *Bioinformatics (Oxford, England)* 28(11):1420–28.

Perez Sepulveda, Blanca, Tamsin Redgwell, Branko Rihtman, Frances Pitt, David J. Scanlan, and Andrew Millard. 2016. “Marine Phage Genomics: The Tip of the Iceberg.” *FEMS Microbiology Letters* 363(15):fnw158.

Pratama, Akbar Adjie, Benjamin Bolduc, Ahmed A. Zayed, Zhi-Ping Zhong, Jiarong Guo, Dean R. Vik, Maria Consuelo Gazitúa, James M. Wainaina, Simon Roux, and Matthew B. Sullivan. 2021. "Expanding Standards in Viromics: In Silico Evaluation of DsDNA Viral Genome Identification, Classification, and Auxiliary Metabolic Gene Curation." *PeerJ* 9:e11447.

Ren, Jie, Nathan A. Ahlgren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun. 2017. "VirFinder: A Novel K-Mer Based Tool for Identifying Viral Sequences from Assembled Metagenomic Data." *Microbiome* 5(1):69.

Ren, Jie, Kai Song, Chao Deng, Nathan A. Ahlgren, Jed A. Fuhrman, Yi Li, Xiaohui Xie, Ryan Poplin, and Fengzhu Sun. 2020. "Identifying Viruses from Metagenomic Data Using Deep Learning." *Quantitative Biology*.

Reyes, Alejandro, Laura V. Blanton, Song Cao, Guoyan Zhao, Mark Manary, Indi Trehan, Michelle I. Smith, David Wang, Herbert W. Virgin, Forest Rohwer, and Jeffrey I. Gordon. 2015. "Gut DNA Viromes of Malawian Twins Discordant for Severe Acute Malnutrition." *Proceedings of the National Academy of Sciences* 112(38):11941–46.

Roux, Simon, Evelien M. Adriaenssens, Bas E. Dutilh, Eugene V. Koonin, Andrew M. Kropinski, Mart Krupovic, Jens H. Kuhn, Rob Lavigne, J. Rodney Brister, Arvind Varsani, Clara Amid, Ramy K. Aziz, Seth R. Bordenstein, Peer Bork, Mya Breitbart, Guy R. Cochrane, Rebecca A. Daly, Christelle Desnues, Melissa B. Duhaime, Joanne B. Emerson, François Enault, Jed A. Fuhrman, Pascal Hingamp, Philip Hugenholtz, Bonnie L. Hurwitz, Natalia N. Ivanova, Jessica M. Labonté, Kyung-Bum Lee, Rex R. Malmstrom, Manuel Martinez-Garcia, Ilene Karsch Mizrachi, Hiroyuki Ogata, David Páez-Espino, Marie-Agnès Petit, Catherine Putonti, Thomas Rattei, Alejandro Reyes, Francisco Rodriguez-Valera, Karyna Rosario, Lynn Schriml, Frederik Schulz, Grieg F. Steward, Matthew B. Sullivan, Shinichi Sunagawa, Curtis A. Suttle, Ben Temperton, Susannah G. Tringe, Rebecca Vega Thurber, Nicole S. Webster, Katrine L. Whiteson, Steven W. Wilhelm, K. Eric Wommack, Tanja Woyke, Kelly C. Wrighton, Pelin Yilmaz, Takashi Yoshida, Mark J. Young, Natalya Yutin, Lisa Zeigler Allen, Nikos C. Kyrpides, and Emiley A. Eloe-Fadrosh. 2019. "Minimum Information about an Uncultivated Virus Genome (MIUViG)." *Nature Biotechnology* 37(1):29–37.

Roux, Simon, Antonio Pedro Camargo, Felipe H. Coutinho, Shareef M. Dabdoub, Bas E. Dutilh, Stephen Nayfach, and Andrew Tritt. 2023. "IPHoP: An Integrated Machine Learning Framework to Maximize Host Prediction for Metagenome-Derived Viruses of Archaea and Bacteria." *PLOS Biology* 21(4):e3002083.

Roux, Simon, Francois Enault, Bonnie L. Hurwitz, and Matthew B. Sullivan. 2015. "VirSorter: Mining Viral Signal from Microbial Genomic Data." *PeerJ* 3:e985.

Sabatier, Marina, Antonin Bal, Grégory Destras, Hadrien Regue, Grégory Quéromès, Valérie Cheynet, Bruno Lina, Claire Bardel, Karen Brengel-Pesce, Vincent Navratil, and Laurence Josset. 2020. "Comparison of Nucleic Acid Extraction Methods for a Viral Metagenomics Analysis of Respiratory Viruses." *Microorganisms* 8(10):1539.

Say, Henry, Ben Joris, Daniel Giguere, and Gregory B. Gloor. 2023. “Annotating Metagenomically Assembled Bacteriophage from a Unique Ecological System Using Protein Structure Prediction and Structure Homology Search.” *BioRxiv*.

Schackart, Kenneth E., Jessica B. Graham, Alise J. Ponsero, and Bonnie L. Hurwitz. 2023. “Evaluation of Computational Phage Detection Tools for Metagenomic Datasets.” *Frontiers in Microbiology* 14.

Shaffer, Michael, Mikayla A. Borton, Bridget B. McGivern, Ahmed A. Zayed, Sabina Leanti La Rosa, Lindsey M. Solden, Pengfei Liu, Adrienne B. Narrowe, Josué Rodríguez-Ramos, Benjamin Bolduc, M. Consuelo Gazitúa, Rebecca A. Daly, Garrett J. Smith, Dean R. Vik, Phil B. Pope, Matthew B. Sullivan, Simon Roux, and Kelly C. Wrighton. 2020. “DRAM for Distilling Microbial Metabolism to Automate the Curation of Microbiome Function.” *Nucleic Acids Research* 48(16):8883–8900.

Shang, Jiayu, Xubo Tang, and Yanni Sun. 2023. “PhaTYP: Predicting the Lifestyle for Bacteriophages Using BERT.” *Briefings in Bioinformatics* 24(1):bbac487.

Steinegger, Martin, and Johannes Söding. 2017. “MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets.” *Nature Biotechnology* 35(11):1026–28.

Suttle, Curtis A. 2005. “Viruses in the Sea.” *Nature* 437(7057):356–61.

Sutton, Thomas D. S., Adam G. Clooney, Feargal J. Ryan, R. Paul Ross, and Colin Hill. 2019. “Choice of Assembly Software Has a Critical Impact on Virome Characterisation.” *Microbiome* 7(1):12.

Svircev, Antonet, Dwayne Roach, and Alan Castle. 2018. “Framing the Future with Bacteriophages in Agriculture.” *Viruses* 10(5):218.

Terzian, Paul, Eric Olo Ndela, Clovis Galiez, Julien Lossouarn, Rubén Enrique Pérez Bucio, Robin Mom, Ariane Toussaint, Marie-Agnès Petit, and François Enault. 2021. “PHROG: Families of Prokaryotic Virus Proteins Clustered Using Remote Homology.” *NAR Genomics and Bioinformatics* 3(3):lqab067.

The UniProt Consortium. 2019. “UniProt: A Worldwide Hub of Protein Knowledge.” *Nucleic Acids Research* 47(D1):D506–15.

Thompson, Luke R., Qinglu Zeng, Libusha Kelly, Katherine H. Huang, Alexander U. Singer, JoAnne Stubbe, and Sallie W. Chisholm. 2011. “Phage Auxiliary Metabolic Genes and the Redirection of Cyanobacterial Host Carbon Metabolism.” *Proceedings of the National Academy of Sciences* 108(39):E757–64.

Tiamani, Kawtar, Shiqi Luo, Sarah Schulz, Jinling Xue, Rita Costa, Mohammadali Khan Mirzaei, and Li Deng. 2022. “The Role of Virome in the Gastrointestinal Tract and Beyond.” *FEMS Microbiology Reviews* 46(6):fuac027.

Tynecki, Piotr, Arkadiusz Guziński, Joanna Kazimierczak, Michał Jadczuk, Jarosław Dastych, and Agnieszka Onisko. 2020. "PhageAI - Bacteriophage Life Cycle Recognition with Machine Learning and Natural Language Processing." *BioRxiv : The Preprint Server for Biology* 2020.07.11.198606.

Van Belleghem, Jonas D., Krystyna Dąbrowska, Mario Vaneechoutte, Jeremy J. Barr, and Paul L. Bollyky. 2019. "Interactions between Bacteriophage, Bacteria, and the Mammalian Immune System." *Viruses* 11(1):10.

van Kempen, Michel, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Johannes Söding, and Martin Steinegger. 2022. "Foldseek: Fast and Accurate Protein Structure Search." *ArXiv* 2022.02.07.479398.

Villarroel, Julia, Kortine Annina Kleinheinz, Vanessa Isabell Jurtz, Henrike Zschach, Ole Lund, Morten Nielsen, and Mette Voldby Larsen. 2016. "HostPhinder: A Phage Host Prediction Tool." *Viruses* 8(5):116.

von Meijenfeldt, F. A. Bastiaan, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, and Bas E. Dutilh. 2019. "Robust Taxonomic Classification of Uncharted Microbial Sequences and Bins with CAT and BAT." *Genome Biology* 20(1):217.

Wahida, Adam, Fang Tang, and Jeremy J. Barr. 2021. "Rethinking Phage-Bacteria-Eukaryotic Relationships and Their Influence on Human Health." *Cell Host & Microbe*.

Wang, Weili, Jie Ren, Kujin Tang, Emily Dart, Julio Cesar Ignacio-Espinoza, Jed A. Fuhrman, Jonathan Braun, Fengzhu Sun, and Nathan A. Ahlgren. 2020. "A Network-Based Integrated Framework for Predicting Virus–Prokaryote Interactions." *NAR Genomics and Bioinformatics* 2(2).

Wang, Yajiao, Yu Liu, Yuxing Wu, Nan Wu, Wenwen Liu, and Xifeng Wang. 2022. "Heterogeneity of Soil Bacterial and Bacteriophage Communities in Three Rice Agroecosystems and Potential Impacts of Bacteriophage on Nutrient Cycling." *Environmental Microbiome* 17(1):17.

Wilhelm, Steven W., and Curtis A. Suttle. 1999. "Viruses and Nutrient Cycles in the Sea: Viruses Play Critical Roles in the Structure and Function of Aquatic Food Webs." *BioScience* 49(10):781–88.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. "Improved Metagenomic Analysis with Kraken 2." *Genome Biology* 20(1):257.

Wood, Derrick E., and Steven L. Salzberg. 2014. "Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments." *Genome Biology* 15(3):R46.

Wu, Shufang, Zhencheng Fang, Jie Tan, Mo Li, Chunhui Wang, Qian Guo, Congmin Xu, Xiaoqing Jiang, and Huaiqiu Zhu. 2021. "DeePhage: Distinguishing Virulent and Temperate Phage-Derived Sequences in Metavirome Data with a Deep Learning Approach." *GigaScience* 10(9).

Wu, Yu-Wei, Yung-Hsu Tang, Susannah G. Tringe, Blake A. Simmons, and Steven W. Singer. 2014. “MaxBin: An Automated Binning Method to Recover Individual Genomes from Metagenomes Using an Expectation-Maximization Algorithm.” *Microbiome* 2(1):26.

Yang, Keli, Junkun Niu, Tao Zuo, Yang Sun, Zhilu Xu, Whitney Tang, Qin Liu, Jingwan Zhang, Enders KW. Ng, Simon KH. Wong, Yun Kit Yeoh, Paul KS. Chan, Francis KL. Chan, Yinglei Miao, and Siew C. Ng. 2021. “Alterations in the Gut Virome in Obesity and Type 2 Diabetes Mellitus.” *Gastroenterology*.

Yates, James A. Fellows, Thiseas C. Lamnidis, Maxime Borry, Aida Andrades Valtueña, Zandra Fagernäs, Stephen Clayton, Maxime U. Garcia, Judith Neukamm, and Alexander Peltzer. 2021. “Reproducible, Portable, and Efficient Ancient Genome Reconstruction with Nf-Core/Eager.” *PeerJ* 9:e10947.

Zaragoza-Solas, Asier, Jose M. Haro-Moreno, Francisco Rodriguez-Valera, and Mario López-Pérez. 2022. “Long-Read Metagenomics Improves the Recovery of Viral Diversity from Complex Natural Marine Samples.” *MSystems* 7(3):e00192-22.

Zerbino, Daniel R. 2010. “Using the Velvet de Novo Assembler for Short-Read Sequencing Technologies.” *Current Protocols in Bioinformatics* 31(1):11.5.1–12.

Zhang, Ruoshi, Milot Mirdita, Eli Levy Karin, Clovis Norroy, Clovis Galiez, and Johannes Söding. 2021. “SpacePHARER: Sensitive Identification of Phages from CRISPR Spacers in Prokaryotic Hosts.” *Bioinformatics (Oxford, England)* 37(19):3364–66.

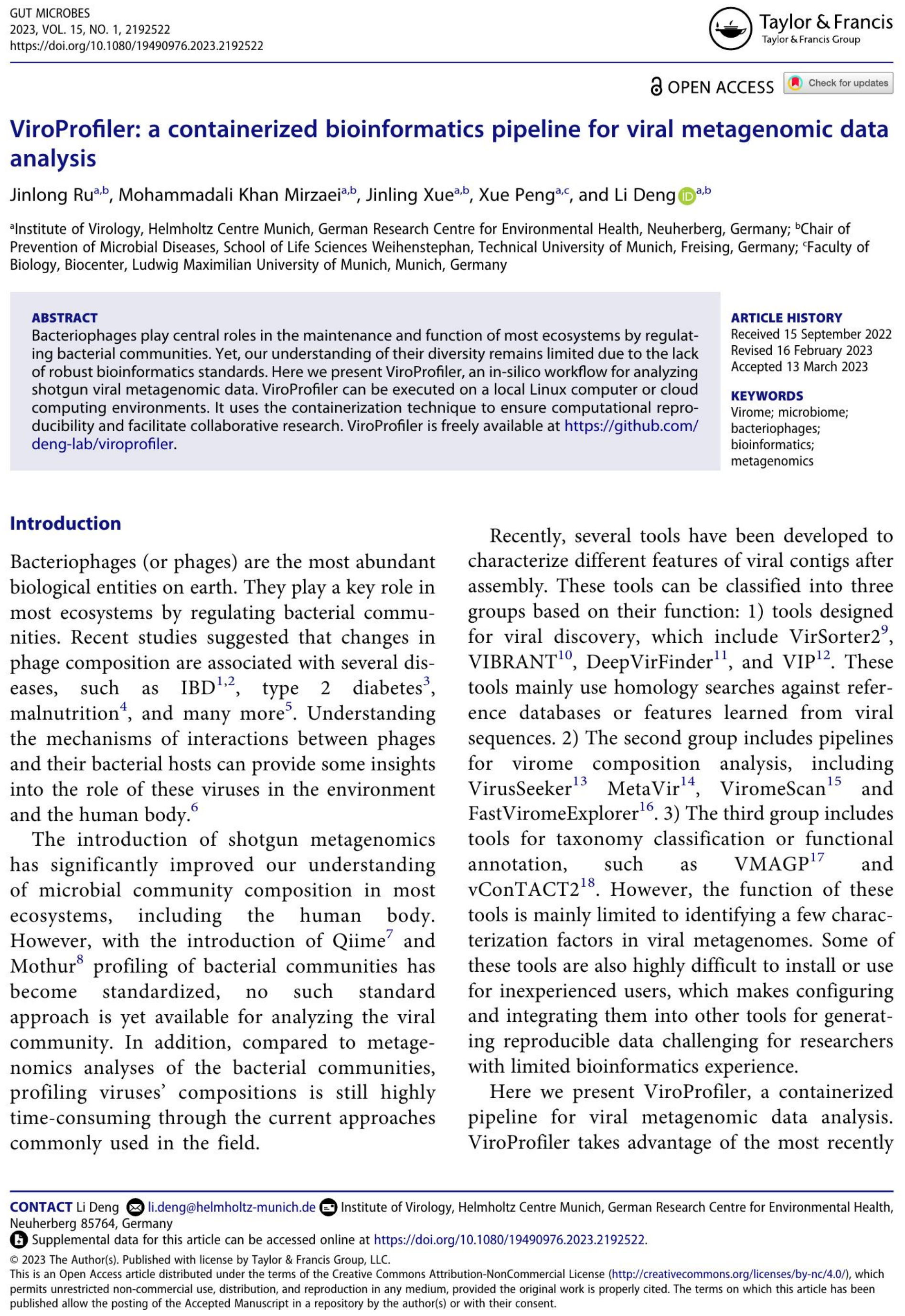
Zhao, Guoyan, Tommi Vatanen, Lindsay Droit, Arnold Park, Aleksandar D. Kostic, Tiffany W. Poon, Hera Vlamakis, Heli Siljander, Taina Härkönen, Anu-Maaria Hämäläinen, Aleksandr Peet, Vallo Tillmann, Jorma Ilonen, David Wang, Mikael Knip, Ramnik J. Xavier, and Herbert W. Virgin. 2017. “Intestinal Virome Changes Precede Autoimmunity in Type I Diabetes-Susceptible Children.” *Proceedings of the National Academy of Sciences* 114(30):E6166–75.

Zheng, Danping, Timur Liwinski, and Eran Elinav. 2020. “Interaction between Microbiota and Immunity in Health and Disease.” *Cell Research* 30(6):492–506.

Zuo, Tao, Xiao-Juan Lu, Yu Zhang, Chun Pan Cheung, Siu Lam, Fen Zhang, Whitney Tang, Jessica Y. L. Ching, Risheng Zhao, Paul K. S. Chan, Joseph J. Y. Sung, Jun Yu, Francis K. L. Chan, Qian Cao, Jian-Qiu Sheng, and Siew C. Ng. 2019. “Gut Mucosal Virome Alterations in Ulcerative Colitis.” *Gut* 68(7):1169–79.

Appendix

A1 Manuscript 1



GUT MICROBES
2023, VOL. 15, NO. 1, 2192522
https://doi.org/10.1080/19490976.2023.2192522

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

ViroProfiler: a containerized bioinformatics pipeline for viral metagenomic data analysis

Jinlong Ru[a,b], Mohammadali Khan Mirzaei[a,b], Jinling Xue[a,b], Xue Peng[a,c], and Li Deng[a,b]

[a]Institute of Virology, Helmholtz Centre Munich, German Research Centre for Environmental Health, Neuherberg, Germany; [b]Chair of Prevention of Microbial Diseases, School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany; [c]Faculty of Biology, Biocenter, Ludwig Maximilian University of Munich, Munich, Germany

ABSTRACT

Bacteriophages play central roles in the maintenance and function of most ecosystems by regulating bacterial communities. Yet, our understanding of their diversity remains limited due to the lack of robust bioinformatics standards. Here we present ViroProfiler, an in-silico workflow for analyzing shotgun viral metagenomic data. ViroProfiler can be executed on a local Linux computer or cloud computing environments. It uses the containerization technique to ensure computational reproducibility and facilitate collaborative research. ViroProfiler is freely available at https://github.com/deng-lab/viroprofiler.

ARTICLE HISTORY
Received 15 September 2022
Revised 16 February 2023
Accepted 13 March 2023

KEYWORDS
Virome; microbiome; bacteriophages; bioinformatics; metagenomics

Introduction

Bacteriophages (or phages) are the most abundant biological entities on earth. They play a key role in most ecosystems by regulating bacterial communities. Recent studies suggested that changes in phage composition are associated with several diseases, such as IBD[1,2], type 2 diabetes[3], malnutrition[4], and many more[5]. Understanding the mechanisms of interactions between phages and their bacterial hosts can provide some insights into the role of these viruses in the environment and the human body.[6]

The introduction of shotgun metagenomics has significantly improved our understanding of microbial community composition in most ecosystems, including the human body. However, with the introduction of Qiime[7] and Mothur[8] profiling of bacterial communities has become standardized, no such standard approach is yet available for analyzing the viral community. In addition, compared to metagenomics analyses of the bacterial communities, profiling viruses' compositions is still highly time-consuming through the current approaches commonly used in the field.

Recently, several tools have been developed to characterize different features of viral contigs after assembly. These tools can be classified into three groups based on their function: 1) tools designed for viral discovery, which include VirSorter2[9], VIBRANT[10], DeepVirFinder[11], and VIP[12]. These tools mainly use homology searches against reference databases or features learned from viral sequences. 2) The second group includes pipelines for virome composition analysis, including VirusSeeker[13] MetaVir[14], ViromeScan[15] and FastViromeExplorer[16]. 3) The third group includes tools for taxonomy classification or functional annotation, such as VMAGP[17] and vConTACT2[18]. However, the function of these tools is mainly limited to identifying a few characterization factors in viral metagenomes. Some of these tools are also highly difficult to install or use for inexperienced users, which makes configuring and integrating them into other tools for generating reproducible data challenging for researchers with limited bioinformatics experience.

Here we present ViroProfiler, a containerized pipeline for viral metagenomic data analysis. ViroProfiler takes advantage of the most recently

CONTACT Li Deng li.deng@helmholtz-munich.de Institute of Virology, Helmholtz Centre Munich, German Research Centre for Environmental Health, Neuherberg 85764, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19490976.2023.2192522.

developed viral metagenomic analysis tools and databases to improve the taxonomy and functional annotation of viruses and their gene products. In addition, ViroProfiler uses containerization to ensure computational reproducibility. ViroProfiler can be executed through a container platform such as Docker and Singularity[19] on Linux clusters or cloud computing environments. It can also be installed via the Conda recipe for high-performance computing clusters that don't support containers.

Results

Overview of the pipeline

Quality control, assembly, and viral discovery

We have included multiple quality control steps for generating an unbiased contig library for downstream analyses in ViroProfiler. These measures ensure to exclude redundancy in the contigs generated, identify prophages and dereplicate highly similar contigs of the same species. This provides a significant advantage to downstream analyses by accurately estimating the relative abundance of viral taxa and metabolic genes in samples. In addition, we included a binning option which enables construction of viral metagenome-assembled genomes (vMAGs) or bins, and provides a more realistic estimation of viral community compositions. After the non-redundant contig library (nrclib) or bins are built, we use VirSorter2[9], VIBRANT[10], DeepVirFinder[11] and CheckV[20] to detect putative viral sequences. VirSorter2, VIBRANT and CheckV identify viral sequences based on their homology to the reference databases, while DeepVirFinder uses a machine learning model to detect viral sequences. Therefore, it can detect novel viruses not showing homology to the public databases. ViroProfiler provides a scoring system for classifying viral contigs identified by multiple tools in this step (Figure 1).

Functional annotation and AMG prediction

In the annotation step, the pipeline provides two possible approaches. By default, ViroProfiler uses DRAM-v, the viral mode of DRAM[21], an automated pipeline for identifying microbial metabolism. DRAM-v can identify auxiliary metabolic genes (AMGs) in viral sequences and annotating their genomes using multiple publicly available databases. The downside of using DRAM-v for annotation is that it slows down the analyses. Therefore, to overcome this issue, we provide an alternative approach for gene annotation, which relies on

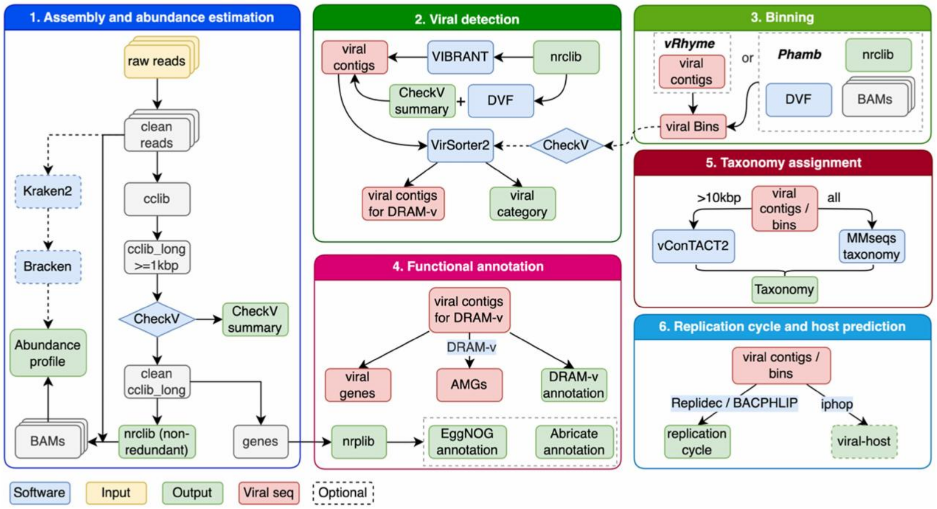


Figure 1. Schematic overview of the ViroProfiler pipeline. Optional steps are indicated with dashed boxes and arrows.

searching the EggNOG database[22] using eggNOG-mapper[23]. The latter is helpful if identifying AMGs in viral contigs is out of interest. For the taxonomy assignment, we combine vConTACT2[18] and MMseqs2 taxonomy[24] module searching against NCBI viral RefSeq database. Combining these two methods, we can significantly improve the accuracy of taxonomy assignment to viral sequences from metagenomics data (Figure 1).

Host prediction, and the assessment of replication cycle

The potential hosts of viral sequences are predicted using iPHoP[25], a recently developed tool which uses a two-step framework that integrates multiple methods for assigning hosts to different viruses based on their genomic signatures with a < 10% false-discovery rate. In addition, our pipeline allows predicting the replication cycle of viral sequences using BACPHLIP[26] and a newly developed in-house software Replidec[27], with a combined accuracy of more than 90%. These tools use the genetic signatures of viral sequences, which are associated with three different types of replication cycles in viruses, lytic, lysogenic, and chronic, to predict their replication cycles (Figure 1 and S1).

Visualization and downstream analyses

We developed an R package called vpfkit (short for “ViroProfiler Tookit”) for downstream analyses of ViroProfiler results in R. It contains functions for preprocessing data generated from multiple ViroProfiler steps, and a Shiny APP called ViroProfiler-viewer for visualizing and manipulating results interactively in a web page. ViroProfiler-viewer allows users to filter viral contigs based on their length, quality, and other annotations such as taxonomy, host, and replication type. In addition, a TreeSummarizedExperiment object file can be generated as inputs for downstream analyses in R. Intermediate files from ViroProfiler, such as genome sequences and BAM files, can be used in other software and pipelines, such as MetaPop[28] for micro- and macro-viral diversity analyses.

Metagenome analyses and validation of the pipeline

We used a simulated mock dataset[29] and an experimental dataset from previous studies to evaluate the performance of ViroProfiler. The mock dataset contains 14 simulated Illumina paired-end sequencing samples, each with 500–1000 viral genomes from the NCBI RefSeq database v69. We analyzed 13 out of the 14 samples using ViroProfiler (sample_12 had no reverse FASTQ file, so it was removed). We compared the viral detection precision and sensitivity of ViroProfiler with Kraken2[30], and abundance estimation performance with Bracken[31].

Specifically, the raw reads from the mock dataset were fed into ViroProfiler for preprocessing, assembly (without binning), annotation, and abundance estimation (”ViroProfiler” in Figure 2). For comparison, Kraken2 and its standard database were used to detect viruses from reads preprocessed by ViroProfiler. Bracken was then used to estimate the abundance of viruses identified by Kraken2 (”BrackenSTD” in Figure 2) and ViroProfiler (”BrackenVPF” in Figure 2), respectively. The taxonomy lineage of viruses was standardized using Taxonkit[32] on the NCBI taxonomy database (obtained on 2022-12-15). We compared the performance of these tools in virus identification using precision, sensitivity, and F1 score (harmonic mean of precision and sensitivity) on different taxonomic ranks and abundance thresholds. Our analyses show that ViroProfiler has the best performance (highest F1 score) at the phylum and order levels, especially at lower abundance thresholds, i.e., ViroProfiler can detect low-abundance viruses with high precision and sensitivity. While using Bracken with Kraken2 and its standard database (BrackenSTD) has the highest sensitivity, they showed a lower precision at the phylum and order levels. At the family level, ViroProfiler achieved performance comparable to BrackenSTD, while at the genus and species levels, the sensitivity of ViroProfiler dropped significantly.

This was expected, as in contrast to ViroProfiler, which uses lowest common ancestor (LCA) of all genes in viral contigs for taxonomy assignment, Kraken2 relies on LCA of exact

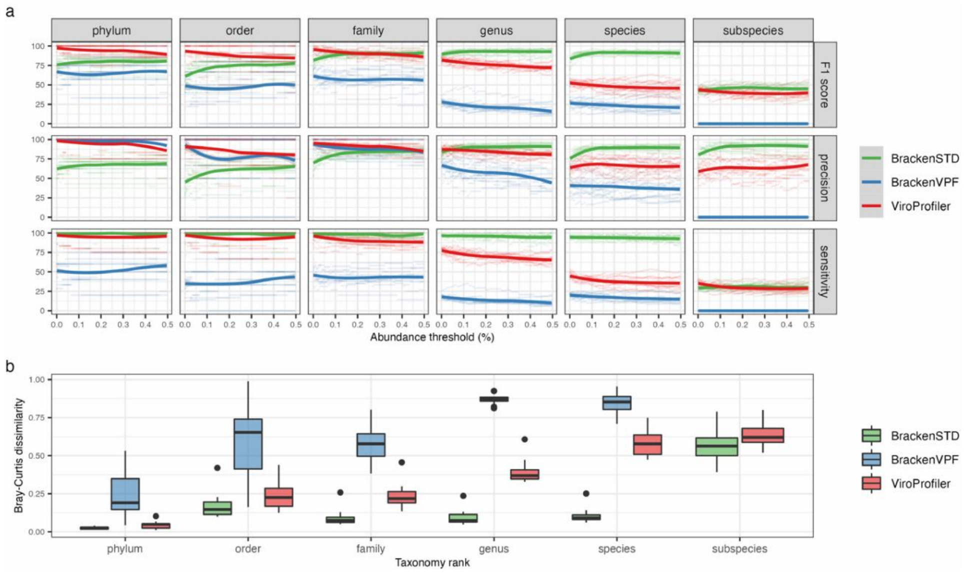


Figure 2. Benchmarking ViroProfiler on mock samples. a) Compares the performance of ViroProfiler with Kraken2 and Bracken in detecting viruses. b) Compares the performance of ViroProfiler and Bracken in providing estimations of viral abundance. BrackenSTD, when Bracken was used with the Kraken2 standard database. BrackenVPF, when Bracken was used with the custom database. Bracken was used for estimating the abundance of identified taxa. Smaller values indicate closer similarity to the true composition profile.

k-mer matches of partial genomes, which increases sensitivity when the viral sequences have representatives in the Kraken2 reference database. Since Kraken2 standard database and the mock dataset are highly similar, we created a custom database that only included viral contigs annotated by ViroProfiler to evaluate the performance of Kraken 2 when these two are less alike. Our results showed that BrackenVPF had the lowest sensitivity in all taxonomic ranks. Even at the phylum level, where ViroProfiler had>95% sensitivity and precision, BrackenVPF had only~50% sensitivity (BrackenVPF in Figure 2a). In addition, we compared the performance of BrackenSTD and BrackenVPF with ViroProfiler in estimating the viral abundances using the mock dataset. We compared the abundance profile generated by ViroProfiler, BrackenSTD, and BrackenVPF with the true composition profile from the original study using Bray-Curtis dissimilarity (Figure 2b). ViroProfiler and BrackenSTD showed similar performance at the phylum and order levels, while Kraken2 and Bracken with the standard database (BrackenSTD) performed better at the family, genus, and species levels. However, when Kraken2 and Bracken were used with the custom database (BrackenVPF), it showed the lowest performance in all taxonomic ranks.

Altogether, our analyses show that ViroProfiler can accurately classify viruses at phylum, order, and family levels. In addition, Viroprofiler provides a database-independent approach for viral classification, contrary to Kraken2. This is especially useful for metagenomic studies, as metagenomes usually include viruses with no homology to the reference database.

To evaluate the performance of ViroProfiler on real datasets, we randomly selected and analyzed 20 out of 266 samples from a previous study of viral community composition in fecal samples from ulcerative colitis (UC) patients and healthy individuals[2]. Using ViroProfiler, we significantly improved the viral discovery rate by identifying 761 viral contigs compared to 183 contigs assembled by the authors. We also observe differences in phage community composition identified by the earlier study compared to the ViroProfiler findings. For example, contrary to the initial analyses, we observed a higher proportion of Podoviridae in samples from healthy individuals than in UC patients (34.6% vs 12.3%). In addition, we did not observe significant differences in diversity scores, as seen in the initial analyses. Moreover, through ViroProfiler, we used DRAM-v, which with a higher accuracy, to strictly identify AMGs in viral contigs, contrary to the initial study that relied on the general functional capacity of the viral contigs, which could be misleading[2]. Finally, ViroProfiler assigned a host to each viral contig, showing that UC patients carry fewer phages that infect Bacteroidia than healthy individuals (Figure 3).

Computational requirements

ViroProfiler can be installed on most operating systems that support Conda and containerization techniques. However, it is recommended to run the pipeline on a High-Performance Computing (HPC) system. The minimum hard disk requirement for the databases and container images is~80GB. However, additional storage space is required if users want to run optional modules such as EggNOG annotation and PHAMB binning. A detailed storage space requirement for each module is available in supplementary table 1.

Our benchmarking analysis on 13 mock datasets using Helmholtz Munich's Scientific Computing HPC cluster (1 to 20 CPUs and 1 to 120 GB of RAM for each process) was finished in 12 hours. Host prediction was the most time-consuming and took 10 hours to complete. However, most analyses can be run in parallel; therefore, using more computational resources will decrease the running time. The execution times and the computational resources used for each step are provided in supplementary figure S1 and supplementary file 1, respectively.

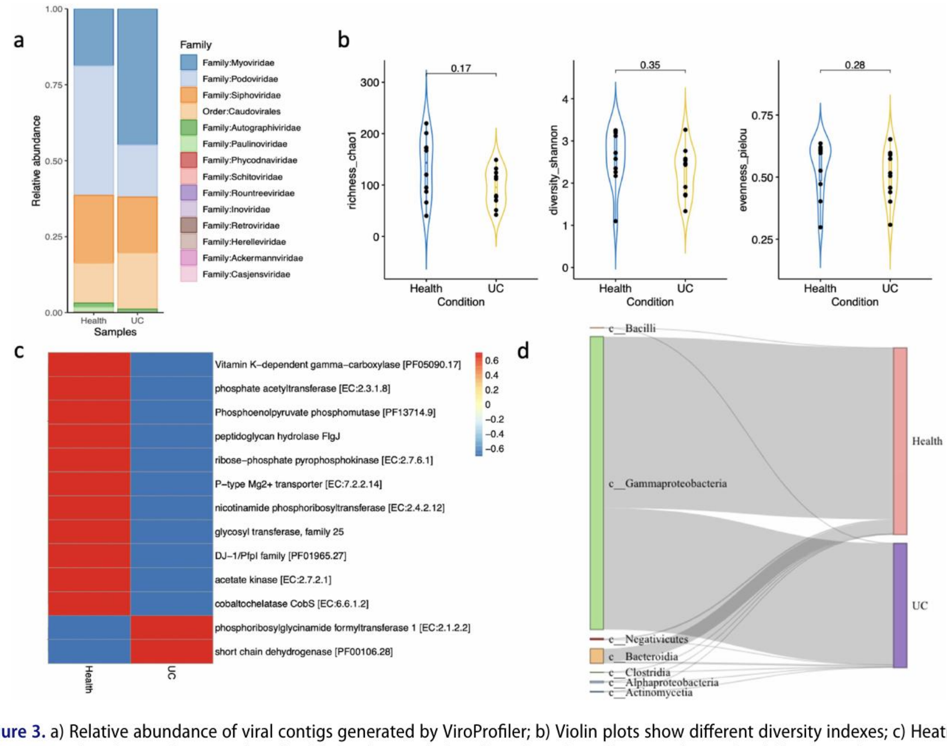


Figure 3. a) Relative abundance of viral contigs generated by ViroProfiler; b) Violin plots show different diversity indexes; c) Heatmap of AMGs predicted in viral contigs from healthy and UC samples; d) Sankey plot of host prediction for different viral contigs.

Discussion

Viral communities are central to the maintenance of most ecosystems, including the human body. The introduction of shotgun metagenomics has provided opportunities to study these communities. Yet, analyses of generated data require applying multiple bioinformatic tools and need relevant programming skills. We believe ViroProfiler, a containerized pipeline for virome data analysis, can address these issues. ViroProfiler combines stand-alone analytical tools and databases with a workflow management system which enables flexible and reproducible analyses of virome data in an interactive environment while significantly shortening the processing time.

We benchmarked ViroProfiler using mock datasets and compared its performance to the existing tools for classifying viruses. ViroProfiler showed high accuracy in classifying viruses at taxonomic ranks higher than genus. Moreover, it can detect viral replication cycles, predict hosts, and identify AMGs in viral sequences. We also used ViroProfiler for analyzing previously published experimental viral metagenome data as part of our validation step. We then compared our results with the original analyses, which showed significant improvement in multiple profiling steps, including viral discovery, taxonomy assignment, functional annotation, host and replication cycle predictions. This was achieved while less than ten percent of the published data were analyzed.

In conclusion, we believe that ViroProfiler can substantially improve the quality of data analyses in virome research and pave the ground for more standardized characterization of the viral communities from complex ecosystems. However, ViroProfiler is specifically designed for classifying viruses in samples with isolated viruses. Therefore, excessive environmental contaminations, usually found in metagenome sequences, could increase the running time of the pipeline and result in lower precision. Yet, this is a general issue with virome studies, and it is recommended to isolate the viral fractions before sequencing for an accurate estimation of viruses in the samples.

Methods

The pipeline

ViroProfiler integrates state-of-the-art bioinformatic tools via Conda environments and containerization techniques for processing viral metagenomic sequences in a nf-core[32] based Nextflow[33] pipeline (Figure 1). It executes series of standard viral metagenomics analysis subsequently or separately if part of the analysis has been done elsewhere. The installation process is described in detail at https://github.com/deng-lab/viroprofiler. For ensuring reproducible analyses, a specific version of the pipeline can always be run by using the version parameter in the command line (-r <version>). In addition, each container used in the workflow is tagged by the accompanying tool version, pre-build and stored on Docker Hub (https://hub.docker.com/u/denglab). The benefit of containers is that users don't need to install multiple software that may cause conflict. Each container contains one or more sub-workflows that is versioned, and Nextflow will automatically download and manage the containers used in each step. Core modules of ViroProfiler and integrated tools are listed in Table 1.

Quality control

The quality control of raw sequencing reads is performed using fastp[37]. The high-quality reads are generated by following five consecutive steps: 1) trimming adapters, 2) removing low-quality reads and 3) trimming the low-quality bases ($Q < 20$) at the end of reads, 4) removing the trimmed reads with length<30bp, and 5) if decontamination option is enabled, reads that show homology to mammalian host genomes will be removed[38]. This is specifically beneficial for identification of AMGs as the previous studies[20] have shown that the removal of host contamination substantially improves the accuracy of AMG identification and interpretation of viral-encoded functions.

Genome assembly and dereplication

Each sample was individually assembled using metaSPAdes[34]. The assembled contigs were then merged into a multi-FASTA file and contigs

Table 1. Core modules and integrated tools of ViroProfiler.

Software	Module	License	Reference
metaSPAdes	Assembly	NA	[35]
vRhyme	Binning	GPL v3	[36]
Phamb	Binning	MIT	[37]
CheckV	Virus detection and QC	BSD	[20]
VirSorter2	Virus detection	GPL v2	[9]
DeepVirFinder	Virus detection	USC-RL v1.0	[11]
VIBRANT	Virus detection and gene annotation	GPL v3	[10]
DRAM	Functional annotation	GPL v3	[21]
eggnog-mapper	Functional annotation	GPL v3	[22,23]
abricate	Functional annotation	GPL v2	https://github.com/tseemann/abricate
MMseqs2	Taxonomy assignment	GPL v3	[24]
vConTACT2	Taxonomy assignment	GPL v3	[18]
Bacphlip	Replication cycle prediction	MIT	[26]
Replidec	Replication cycle prediction	MIT	[27]
iPHoP	Host prediction	GPL v3	[25]
CoverM	Abundance estimation	GPL v3	https://github.com/wwood/CoverM
Kraken2	Virus detection	MIT	[30]
Bracken	Abundance estimation	GPL v3	[31]

shorter than a threshold (ex. 1kbp) were excluded from the further analyses. This step generated the long "complete contig library" (cclib_long). The quality of cclib_long was then evaluated using CheckV[20], which were assessed for their quality, completeness, and potential contamination. The host flanking region were also removed from the final contigs. To remove redundancy in the contig library, we dereplicated the cclib_long by clustering contigs following the MIUViG guidelines (95% ANI – Average Nucleotide Identity and 85% AF – Aligned Fraction)[39] using custom python script anical.py and aniclust.py from CheckV. This step generated a non-redundant contig library (nrclib) for downstream analyses.

Viral contig binning

Due to the limitation of assemblers, we usually get fragmented contigs of a viral genome. To overcome this limitation, ViroProfiler uses binning approach that relies on Phamb[36] and vRhyme[35] to identify contigs that belong to the same genome and classify them as a bin, or viral metagenome-assembled genome (vMAG). Phamb is a recently developed tool for binning phage genomes that relies on DeepVirFinder for viral contig discovery and a deep-learning algorithm for contig binning[40]. It requires>50,000 contigs as input, which sometimes can not be met. In that case, users can choose vRhyme for the binning step, which uses multi-sample coverage effect size comparisons between scaffolds, protein redundancy scoring mechanism, and machine learning model to detect bins. Viral quality, completeness and contamination ratio of bins were then assessed using CheckV. Binning is set as an optional step in ViroProfiler because the risk of false positive and the fact that contigs in a bin is connected randomly, which might not represent the actual viral genomes.

Viral contig identification

ViroProfiler integrates five different tools for identification of viral sequences: 1) VirSorter2[9], 2) MMseqs2 taxonomy assignment[24] based on NCBI viral RefSeq, 3) CheckV[20], 4) DeepVirFinder[11] and 5) VIBRANT[10]. Briefly, contigs or bins are identified as viruses when they satisfy one of the following criteria: 1) identified as viruses in category 1, 2, 4, or 5 by VioSorter2 with default parameters (–virome mode); 2) classified as viruses by Mmseqs2 taxonomy module; 3) classified as complete, high-quality, medium-quality and low-quality by CheckV; 4) have a score>0.9 and p-value<0.01 in the DeepVirFinder prediction; 5) identified as viruses by VIBRANT. Viral detection tools were selected based on their approach to identifying viral sequences. VirSorter2, VIBRANT, MMseqs taxonomy module, and CheckV identify viral sequences based on the homology of proteins in contigs to reference databases, which is more reliable than non-homology-based tools like DeepVirFinder. However, DeepVirFinder employs a machine-learning model trained on viral genomic signatures to distinguish viral sequences from non-viral sequences. Therefore, it can detect novel viruses with no homology to the reference databases. While homology-based tools like VirSorter2 and VIBRANT tend to have lower false positive rates on longer contigs (e.g.>3 kbp), non-homology-based tools like DeepVirFinder have shown higher sensitivity, making them more suitable for analyzing short contigs (e.g.<3 kbp) and detecting novel viruses[41–43].

ViroProfiler provides a confidence classification to the contigs or bins identified as viruses using the following criteria, 1) “high confident” is assigned if they are classified by VIBRANT, or as category 1,2 by VirSorter2, or as viruses by mmseqs2 taxonomy module, or have “Complete”, “High-quality”, “Medium-quality” annotation in CheckV; 2) “low confident” are rest contigs that predicted as viral sequences by DeepVirFinder, and “unclassified” by MMseqs2 taxonomy module or have “Low quality” annotation in CheckV.

Gene prediction and protein function annotation

To keep as many potential genes as possible, contigs in cclib_long are fed into Prodigal[44] for predicting protein-coding genes and translating them to proteins. To remove redundancy and improve annotation speed in downstream analysis, proteins are clustered using MMseqs2[45] using thresholds of minimum identity (0.7 by default) and coverage (0.9 by default). These thresholds can be modified in the params.yml config file before running the pipeline. Representative proteins of these clusters are used to make the non-redundant protein library (nrplib), which is assigned a computationally predicted function and gene ontology using eggNOG-mapper[23] searching against the EggNOG database[22]. This step will not be necessary in case prediction of AMGs is planned as DRAM-v also provides functional annotations. Functional annotations of viral contigs are annotated using DRAM-v, which searches viral genes against multiple databases, such as KEGG[46], PFAM[47], VOGDB (https://vogdb.org/) and NCBI viral RefSeq[48]. DRAM-v also detects auxiliary metabolic genes (AMGs) in viral genomes. In addition, antimicrobial resistance and virulence genes can be identified using Abricate (https://github.com/tseemann/abricate) to search genes against CARD[49], ResFinder[50] and VFDB[51, 52] databases.

Taxonomy assignment

Taxonomy assignment of viral contigs is performed using a combination of viral genome clustering and voting-based classification approaches. Briefly, for viral contigs longer than 10 kbp, their protein sequences are fed into vConTACT2[53] for virus clustering and taxonomy annotation. Since vConTACT2 does not report taxonomy names at the species and subspecies level, we combine vConTACT2 clustering with the MMseqs2 taxonomy module[24] using the NCBI viral RefSeq as references. MMseqs2 assigns taxonomy to viral sequences by comparing their proteins to reference databases and determining taxonomy using the lowest common ancestor. MMseqs2 was selected as it is fast and sensitive[24]. We combine the MMseqs2 results with viral clusters (VCs) generated by VConTACT2. When VCs contain multiple contigs with different taxonomies, we use LCA to assign the final taxonomy. However, users could manually check these VCs and determine taxonomy based on their domain knowledge. To be consistent with taxonomy assignment, names and lineages are standardized using taxonkit[32] and an in-house python script.

Host and replication cycle prediction

We used iPHoP to predict virus-host ranges[25], which integrates multiple methods to provide host predictions. This makes its predictions highly reliable compared to other tools available for host prediction. However, iPHoP has a big database (~200GB), thus we set host prediction as an optional step. Users can skip this step if they are not interested in the host predictions. The virus replication cycle is predicted using BACPHLIP[26] and Replidec[27].

Viral abundance estimation

ViroProfiler provides two approaches for viral abundance estimation. The first approach uses Bracken to estimate the abundance of each taxonomic category from the Kraken2 classification results. This provides accurate estimates of viral sequences with representatives in the Kraken2 reference database. However, Kraken2 fails to identify novel viruses with no homology to the databases. Therefore, the second approach estimates viral abundance based on mapping clean reads to ViroProfiler assembled viral contigs. Briefly, clean reads are mapped to contigs in nrclib using bowtie2[54] to create BAM files for each sample. Next, CoverM (https://github.com/wwood/CoverM) is used to remove spurious read mappings at less than 90% identity in BAM files and then calculate the number of reads (−m count), trimmed mean of coverage (-m trimmed_mean) and covered fraction (-m covered_fraction) of each contig across all samples. In the downstream analyses, the abundance of a viral contig in a sample is usually set to zero if reads from that contig cover less than a threshold percentage (ex. 50%) in the sample. This refinement of the abundance table can be generated in ViroProfiler-viewer in an interactive way. Finally, if the abundance of genes is of interest, featureCounts[55] is used to calculate number of reads mapped to each protein-coding gene. Altogether, these two approaches can accurately estimate viral abundance regardless of their homology to reference databases.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was funded by the German Research Foundation (D.F.G. Emmy Noether program, Project No. 273124240, SFB 1371, Project No. 395357507), Marie Sklodowska-Curie Actions Innovation Training Networks grant agreement no. 955974 (VIROINF), and the European Research Council Starting grant (ERC StG 803077) awarded to L.D. J.R was supported by the China Scholarship Council (CSC).

Author contributions

J.R. developed the software. M.K.M. and J.R. drafted the manuscript. J.R and X.P. performed the analyses. J.X. wrote the documentation. M.K.M. and L.D. conceived and supervised the project. All authors reviewed and approved the manuscript.

Data and software availability

ViroProfiler is available at https://github.com/deng-lab/viroprofiler. The development version of the pipeline will be updated once the dependent software are updated. The stable version will be updated yearly. The R package vpfkit is available at https://github.com/deng-lab/vpfkit. All data and reproducible analysis scripts used in this study are available as an R package at https://github.com/deng-lab/vpfpaper.

ORCID

Li Deng http://orcid.org/0000-0003-0225-0663

References

1. Clooney AG, Sutton TDS, Shkoporov AN, Holohan RK, Daly KM, O'regan O, Ryan FJ, Draper LA, Plevy SE, Ross RP, et al. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. Cell Host & Microbe. 2019;26:764–778.e5. doi:10.1016/j.chom.2019.10.009.
2. Zuo T, X-J L, Zhang Y, Cheung CP, Lam S, Zhang F, Tang W, Ching JYL, Zhao R, Chan PKS, et al. Gut mucosal virome alterations in ulcerative colitis. Gut. 2019;68:1169–1179. doi:10.1136/gutjnl-2018-318131.
3. Ma Y, You X, Mai G, Tokuyasu T, Liu C. A human gut phage catalog correlates the gut phageome with type 2 diabetes. Microbiome. 2018;6:24. doi:10.1186/s40168-018-0410-y.
4. Mirzaei MK, Khan MAA, Ghosh P, Taranu ZE, Taguer M, Ru J, Chowdhury R, Kabir MM, Deng L, Mondal D, et al. Bacteriophages isolated from stunted children can regulate gut bacterial communities in an age-specific manner. Cell Host & Microbe. 2020;27:199–212.e5. doi:10.1016/j.chom.2020.01.004.
5. Ma T, Ru J, Xue J, Schulz S, Mirzaei MK, Janssen K-P, Quante M, Deng L. Differences in gut virome related to Barrett esophagus and esophageal adenocarcinoma. Microorganisms. 2021;9:1701. doi:10.3390/microorganisms9081701.
6. Noble WS, Lewitter F. A quick guide to organizing computational biology projects. PLoS Comput Biol. 2009;5:e1000424. doi:10.1371/journal.pcbi.1000424.
7. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37:852–857. doi:10.1038/s41587-019-0209-9.
8. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ

Microbiol. 2009;75:7537–7541. doi:10.1128/AEM.01541-09.

9. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, Pratama AA, Gazitúa MC, Vik D, Sullivan MB, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome. 2021;9:37. doi:10.1186/s40168-020-00990-y.
10. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome. 2020;8:90. doi:10.1186/s40168-020-00867-0.
11. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Poplin R, Sun F. Identifying viruses from metagenomic data using deep learning. Quantitative Biology. 2020;8:64–77. doi:10.1007/s40484-019-0187-4.
12. Li Y, Wang H, Nie K, Zhang C, Zhang Y, Wang J, Niu P, Ma X. VIP: an integrated pipeline for metagenomics of virus identification and discovery. Sci Rep. 2016;6. doi:10.1038/srep23774.
13. Zhao G, Wu G, Lim ES, Droit L, Krishnamurthy S, Barouch DH, Virgin HW, Wang D. VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. Virology. 2017;503:21–30. doi:10.1016/j.virol.2017.01.005.
14. Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D, Enault F. Metavir: a web server dedicated to virome analysis. Bioinformatics. 2011;27:3074–3075. doi:10.1093/bioinformatics/btr519.
15. Rampelli S, Soverini M, Turroni S, Quercia S, Biagi E, Brigidi P, Candela M. ViromeScan: a new tool for metagenomic viral community profiling. BMC Genomics. 2016;17:165. doi:10.1186/s12864-016-2446-3.
16. Tithi SS, Aylward FO, Jensen RV, Zhang L. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. PeerJ. 2018;6:e4227. doi:10.7717/peerj.4227.
17. Lorenzi HA, Hoover J, Inman J, Safford T, Murphy S, Kagan L, Williamson SJ. The Viral MetaGenome Annotation Pipeline (VMGAP): an automated tool for the functional annotation of viral metagenomic shotgun sequencing data. Stand Genomic Sci. 2011;4:418–429. doi:10.4056/sigs.1694706.
18. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol. 2019;37:632–639. doi:10.1038/s41587-019-0100-8.
19. Kurtzer GM, Sochat V, Mw B, Gursoy A. Singularity: scientific containers for mobility of compute. PLoS One. 2017;12:e0177459. doi:10.1371/journal.pone.0177459.
20. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC . CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39(5): 578–585. doi:10.1038/s41587-020-00774-7.
21. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, Liu P, Narrowe AB, Rodríguez-Ramos J, Bolduc B, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res. 2020;48:8883–8900. doi:10.1093/nar/gkaa621.
22. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, et al. Ggnog 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2018;47:D309–14. doi:10.1093/nar/gky1085.
23. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J, Tamura K. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021;38:5825–5829. doi:10.1093/molbev/msab293.
24. Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E, Kelso J. Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics. 2021;37:3029–3031. doi:10.1093/bioinformatics/btab184.
25. Roux S, Camargo AP, Coutinho FH, Dabdoub SM, Dutilh BE, Nayfach S, Tritt A . iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes. bioRxiv. 2022. doi:10.1101/2022.07.28.501908.
26. Hockenberry AJ, Co W. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. PeerJ. 2021;9:e11396. doi:10.7717/peerj.11396.
27. Peng X, Ru J, Mirzaei MK, Deng L. Replidec – use I Bayes classifier to identify virus lifecycle from metagenomics data. bioRxiv. 2022. doi:10.1101/2022.07.18.500415.
28. Gregory AC, Gerhardt K, Zhong Z-P, Bolduc B, Temperton B, Konstantinidis KT, Sullivan MB. MetaPop: a pipeline for macro- and microdiversity analyses and visualization of microbial and viral metagenome-derived populations. Microbiome. 2022;10:49. doi:10.1186/s40168-022-01231-0.
29. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ. 2017;5:e3817. doi:10.7717/peerj.3817.
30. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257. doi:10.1186/s13059-019-1891-0.
31. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data.

PeerJ Computer Science. 2017;3:e104. doi:10.7717/peerj-cs.104.

32. Shen W, Ren H. TaxonKit: a practical and efficient NCBI taxonomy toolkit. Journal of Genetics and Genomics. 2021;48:844–850. doi:10.1016/j.jgg.2021.03.006.
33. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38:276–278. doi:10.1038/s41587-020-0439-x.
34. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–319. doi:10.1038/nbt.3820.
35. Nurk S, Meleshko D, Korobeynikov A, Pa P. metaSpades: a new versatile metagenomic assembler. Genome Res. 2017;27:824–834. doi:10.1101/gr.213959.116.
36. Kieft K, Adams A, Salamzade R, Kalan L, Anantharaman K. vRhyme enables binning of viral genomes from metagenomes. Nucleic Acids Res. 2022;50:e83. doi:10.1093/nar/gkac341.
37. Johansen J, Plichta DR, Nissen JN, Jespersen ML, Shah SA, Deng L, Stokholm J, Bisgaard H, Nielsen DS, Sørensen SJ, et al. Genome binning of viral entities from bulk metagenomics data. Nat Commun. 2022;13:965. doi:10.1038/s41467-022-28581-5.
38. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90. doi:10.1093/bioinformatics/bty560.
39. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell Host & Microbe. 2020;28(5):724–740. e8. doi:10.1016/j.chom.2020.08.003.
40. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A, et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). Nat Biotechnol. 2019;37:29–37. doi:10.1038/nbt.4306.
41. Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, Jensen LJ, Nielsen HB, Petersen TN, Winther O, et al. Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol. 2021;39:1–6. doi:10.1038/s41587-020-00777-4.
42. Schackart KE, Graham JB, Ponsero AJ, Hurwitz BL. Evaluation of computational phage detection tools for metagenomic datasets. Front Microbiol. 2023;14. doi:10.3389/fmicb.2023.1078760.
43. Pratama AA, Bolduc B, Zayed AA, Zhong Z-P, Guo J, Vik DR, Gazitúa MC, Wainaina JM, Roux S, Sullivan MB. Expanding standards in viromics: in silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation. PeerJ. 2021;9:e11447. doi:10.7717/peerj.11447.
44. Glickman C, Hendrix J, Strong M. Simulation study and comparative evaluation of viral contiguous sequence identification tools. BMC Bioinform. 2021;22:329. doi:10.1186/s12859-021-04242-0.
45. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Lj H. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 2010;11:119. doi:10.1186/1471-2105-11-119.
46. Steinegger M, Söding J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–1028. doi:10.1038/nbt.3988.
47. Kanehisa M, Goto SK. Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30. doi:10.1093/nar/28.1.27.
48. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49:D412–9. doi:10.1093/nar/gkaa913.
49. Li W, O'neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire MK, Durkin AS, et al. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. Nucleic Acids Res. 2021;49:D1020–8. doi:10.1093/nar/gkaa1105.
50. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen AL, Cheng AA, Liu S, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2020;48:D517–25. doi:10.1093/nar/gkz935.
51. Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FMY. ResFinder an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. Microbial Genomics. 2022;8:000748. doi:10.1099/mgen.0.000748.
52. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic Acids Res. 2019;47:D687–92. doi:10.1093/nar/gky1080.
53. Bolduc B, Jang HB, Doulcier G, You Z-Q, Roux S, Mb S. vContact: an iVirus tool to classify double-stranded DNA viruses that infect archaea and bacteria. PeerJ. 2017;5:e3243. doi:10.7717/peerj.3243.
54. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012 Apr;9(4):357–359. doi:10.1038/nmeth.1923.
55. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–930. doi:10.1093/bioinformatics/btt656.

A2 Manuscript 2

microorganisms

MDPI

Article

Differences in Gut Virome Related to Barrett Esophagus and Esophageal Adenocarcinoma

Tianli Ma [1,2,†], Jinlong Ru [1,2,†], Jinling Xue [1,2], Sarah Schulz [1,2], Mohammadali Khan Mirzaei [1,2], Klaus-Peter Janssen [3], Michael Quante [4,5,*] and Li Deng [1,2,*]

[1] Helmholtz Centre Munich—German Research Center for Environmental Health, Institute of Virology, 85764 Neuherberg, Germany; tianli.ma@helmholtz-muenchen.de (T.M.); jinlong.ru@helmholtz-muenchen.de (J.R.); jinling.xue@helmholtz-muenchen.de (J.X.); sarah.schulz@helmholtz-muenchen.de (S.S.); m.khanmirzaei@helmholtz-muenchen.de (M.K.M.)

[2] Institute of Virology, Technical University of Munich, 81675 Munich, Germany

[3] Department of Surgery, Klinikum Rechts der Isar, Technical University of Munich, 81675 Munich, Germany; klaus-peter.janssen@tum.de

[4] II. Medizinische Klinik, Klinikum Rechts der Isar, Technische Universität München, 81675 Munich, Germany

[5] Innere Medizin II, Universitätsklinik Freiburg, Universität Freiburg, 79106 Freiburg, Germany

* Correspondence: michael.quante@uniklinik-freiburg.de (M.Q.); li.deng@helmholtz-muenchen.de (L.D.)

† These authors contributed equally to this work.

Abstract: The relationship between viruses (dominated by bacteriophages or phages) and lower gastrointestinal (GI) tract diseases has been investigated, whereas the relationship between gut bacteriophages and upper GI tract diseases, such as esophageal diseases, which mainly include Barrett's esophagus (BE) and esophageal adenocarcinoma (EAC), remains poorly described. This study aimed to reveal the gut bacteriophage community and their behavior in the progression of esophageal diseases. In total, we analyzed the gut phage community of sixteen samples from patients with esophageal diseases (six BE patients and four EAC patients) as well as six healthy controls. Differences were found in the community composition of abundant and rare bacteriophages among three groups. In addition, the auxiliary metabolic genes (AMGs) related to bacterial exotoxin and virulence factors such as lipopolysaccharides (LPS) biosynthesis proteins were found to be more abundant in the genome of rare phages from BE and EAC samples compared to the controls. These results suggest that the community composition of gut phages and functional traits encoded by them were different in two stages of esophageal diseases. However, the findings from this study need to be validated with larger sample sizes in the future.

Keywords: esophageal diseases; esophageal carcinogenesis; gut bacteriophages; bacterial exotoxin; LPS biosynthesis proteins

Citation: Ma, T.; Ru, J.; Xue, J.; Schulz, S.; Mirzaei, M.K.; Janssen, K.-P.; Quante, M.; Deng, L. Differences in Gut Virome Related to Barrett Esophagus and Esophageal Adenocarcinoma. *Microorganisms* **2021**, *9*, 1701. https://doi.org/10.3390/microorganisms9081701

Academic Editor: Hanna M. Oksanen

Received: 28 May 2021
Accepted: 6 August 2021
Published: 10 August 2021

1. Introduction

Barrett's esophagus (BE) is the only known precursor for the development of esophageal adenocarcinoma (EAC) with a five-year survival rate of less than 20%. The incidence of these diseases is on the rise globally [1,2]. Early diagnosis of patients at risk could prevent the progression of BE to EAC, and effectively reduce the development of EAC. However, as only 0.3–0.5% of BE patients develop EAC, endoscopic biopsy surveillance, while linked to higher survival rates, is only recommended for at-risk patients [3]. In addition, endoscopies are often discomforting, and sometimes lead to inconclusive results [4]. Thus, noninvasive diagnostics with higher accuracy are sought after. The human gut is home to trillions of microorganisms, including bacteria, viruses, fungi, and protozoa. These microorganisms and their human host maintain a symbiotic relationship, in which the host provides a nutrient-rich habitat, and the microbiota supplies key metabolic capabilities, protects against pathogen invasion, and trains the immune system [5]. In addition, an imbalance in gut microbiota, termed dysbiosis, is associated with several human diseases or conditions, including inflammatory bowel disease (IBD), and colorectal cancer (CRC). These microbial communities have shown disease-specific community structure, suggesting that they can be used as signatures for diagnosing some dysbiosis-associated diseases [6–9].

Both BE and EAC biopsy samples have been found to harbor a unique bacterial community. Compared to the normal esophagus, Gram-positive bacteria (Firmicutes) were gradually replaced by Gram-negative bacteria (Bacteroidetes, Proteobacteria, Fusobacteria, and Spirochaetes) in BE [10]. As the disease progressed from BE to EAC, the Gram-negative bacteria *Escherichia coli* (*E.coli*) and *Fusobacterium nucleatum* became more dominant [11]. These changes are important as LPS, the outer membrane component of Gram-negative bacteria, could promote the secretion of pro-inflammatory cytokines through activating the Toll-like receptor (TLR) and the downstream NF- κB pathway in different cell types, contributing to the severity of esophageal diseases [11]. In human and mice models with BE, elevated levels of pro-inflammatory cytokines and activated TLR were observed in the gastroesophageal junction [12]. The resulting chronic inflammation could induce systemic immune responses, which further promote the development of GI tract diseases [13]. In the BE mouse model, the chemokines IL-1b and IL-8, secreted by epithelial cells in the esophagus and forestomach squamous epithelium, facilitated the progression of BE to EAC [6]. Moreover, the gut microbiome was associated with this process, as germ-free L2-1L1B mice did not develop dysplasia while the shift of the gut microbiome resulted in different speeds of developing esophageal dysplasia and tumor [6]. The above evidence further shows that these alterations of the bacterial community associated with inflammation can accelerate the development of esophageal diseases.

However, this is not limited to the gut bacteria as viruses, which outnumber bacterial cells by about tenfold in the gut, also contribute to human health and diseases [14–19]. In addition to the widely reported eukaryotic viruses [20–23], mounting data suggests that phages play a critical role in human health by affecting the bacterial community and function [19,24,25]. For example, bacterial-cell lysis caused by phage infection can lead to the release of nucleic acids, proteins, and lipids, which may trigger an inflammation response [26,27]. In addition, prophages that are integrated in bacterial genomes could supply them with virulence-associated genes that can increase their fitness under specific conditions [28]. Under stimulus (such as, DNA damage [29]), the prophages may switch to the lytic cycle [30], which can lead to gene exchange between bacteria, increasing their pathogenicity [31]. For example, the virulence gene that encodes the enterotoxin A was transferred to *Staphylococcus aureus* by phage-mediated horizontal gene transfer (HGT) [32,33]. Furthermore, phages can also obtain AMGs from bacteria to modulate bacterial metabolism [34]. These phage behaviors that regulate bacterial physiology could further indirectly influence human health, such as the occurrence of GI tract and non-GI tract diseases including IBD, CRC, Parkinson's disease, and Type I diabetes [27,35–37].

Former studies that investigated the role of phages in GI tract diseases have mainly focused on the phage community related to lower GI tract diseases, several studies have already described the disease-specific phage community that has been revealed in inflammation-induced diseases such as Crohn's disease and ulcerative colitis [27]. In a mouse model of intestinal colitis, it was reported that the bacteriophage community structure correlated with the disease status, and the presence of some phages during colitis was associated with an increase in pathobiontic host bacteria (*Escherichia-Shigella*, *Salmonella*, *Mycobacterium*) that was linked to the intestinal inflammation response [38]. However, the role of phages in the upper GI tract remains poorly described and limited to a few studies that have explored the viral community of the oral cavity [39,40]. The research related to the role of the phage community in esophageal diseases is also limited to one study that has used metagenomic data from the whole microbial community without isolating the viral like particles (VLPs) before sequencing [41]. Profiling the community composition of gut phages in esophageal diseases such as BE and EAC can provide some further insight into the role of phages in upper GI tract diseases.

This study aimed to investigate the alteration of gut phages in different stages of esophageal diseases. For this purpose, we (1) determined the composition of the isolated bacteriophage community in BE patients, EAC patients, and healthy controls (CT); (2) predicted the bacterial host ranges of the gut phages in all three groups; (3) identified the metabolic pathways encoded by these phages.

2. Materials and Methods

2.1. Sample Collection

Sixteen samples were selected from the German BarrettNET registry including six BE patients, four EAC patients, and six CT for virome analysis. The clinical data are shown in Table S1, and additional information can be found in a previous study [42]. Stool samples were collected using Stool Collection Tubes with Stool DNA Stabilizer (STRATEC Molecular GmbH, Berlin, Germany). The sampling procedure was conducted mostly at home or in the clinic if the patients were on outpatient visits. Samples were shipped to the clinic human sample biobank and stored at −80 °C until further virome DNA extraction.

2.2. Virome DNA Extraction

The stool samples were vortexed vigorously for 4 h at 4 °C, then centrifuged at 4000 g for 30 min to collect supernatant. The supernatant was passed through 0.22 µm filters (PES Membrane, Lot No. ROCB29300, Merck Millipore, Co., Cork, Ireland) to remove bacterial-associated particles, and the volume was subsequently concentrated to less than 50 µL by Amicon® Ultra Centrifugal Filters (10 kDA, Lot No. R9EA18187, Merck Millipore, Co., Cork, Ireland). Then 1/5 volume of chloroform was mixed with the samples and centrifuged at 14,000 g for 3 min, retaining the upper phase followed by a DNAse I (1 U/µL, Lot No. 1158858, Invitrogen, Carlsbad, CA, USA) treatment for 1 h at 37 °C to remove non-phage DNA. DNase I was inactivated by adding EDTA (0.1 M). Subsequently, lysis buffer (700 µL KOH stock (0.43 g/10 mL), 430 µL DDT stock (0.8 g/10 mL), and 370 µL H_2O, pH = 12) was added to the reaction and incubated at room temperature for 10 min followed by 2 h incubation at −80 °C, and 5 min at 55 °C. Lysed VLPs were then treated for 30 min at 55 °C with Proteinase K (20 mg/mL, Lot No. 1112907, Invitrogen, Carlsbad, CA, USA) to digest remaining viral capsid and extract the virome DNA. AMPure beads (Agencourt, Beckman Coulter, Brea, CA, USA) were added to the extracted DNA and incubated for 15 min at room temperatureF. DNA was eluted from beads by 35 µL Tris buffer (10 mM, pH = 9.8) and stored at −80 °C until it was sent for sequencing. Sequencing was performed on an Illumina HiSeq-PE150 platform.

2.3. Bioinformatic Analysis

On average, 9,358,935 ± 169,389 reads per samples were generated. Raw reads were processed with fastp (v0.20.1) [43] to remove adaptors and low-quality bases. Remaining reads were deduplicated using dedupe.sh from bbmap suite (v38.76) (https://sourceforge.net/projects/bbmap/; accessed on 29 January 2020). Then the obtained reads were assembled into contigs using metaSPAdes (v3.14.0) [44] with default parameters retaining only contigs longer than 1 kb. Redundant contigs were removed by dedupe.sh. Remaining contigs were used to predict viral sequences by the combination of VirSorter (v1.0.6) [45], CAT (v5.0.4) [46] and DeepVirFinder (v1.0) [47]. Contigs predicted as category 1 and 2 by Virsorter, or predicted as viruses by CAT, were classified as viruses. Contigs also were classified as viruses if they were predicted as category 3 by VirSorter or could not be classified to taxonomy by CAT but were predicted as a virus by DeepVirFinder with q value < 0.01. Predicted viral contigs were clustered using CD-HIT [48] if they shared >95% identity over 80% of the contig length, the longest contigs in each cluster were retained as a representative for downstream analysis.

For each representative viral contig, ORFs were predicted using Prodigal (v2.6.3) [49] and provided to vConTACT (v2.0) [50] for taxonomy annotation. For contigs that could not be assigned a taxonomy by vConTACT, CAT annotations were used. Otherwise, Order and Family level taxonomic annotations were predicted using Demovir script (https://github.com/feargalr/Demovir; accessed on 27 July 2019) with default parameters and database. To calculate the relative abundances of viruses in each sample, clean reads from each sample were mapped to viral contigs using bbmap.sh from bbmap suite (v38.76). CoverM (v0.4.0) (https://github.com/wwood/CoverM; accessed on 20 February 2020) was used to estimate contig coverage. Feature Counts (v2.0.0) [51] was then used to estimate the number of reads that mapped to each gene. Viral proteins predicted in the previous step were fed into VIBRANT (v1.2.1) [52] to identify lytic and lysogenic phages and the function was annotated using protein mode with default parameters. VIBRANT annotates viral proteins by searching viral proteins against KEGG [53], VOGDB and PFAM databases, which include function annotation of protein sequences and AMGs. The virus (phage)-bacteria (host) interactions were predicted by VirHostMatcher-Net, which is a method based on the combination of features: virus-virus similarity, virus-host alignment-free similarity, virus-host shared CRISPR spacers and virus-host alignment-based matches [54]. Bacterial hosts were predicted for contigs with a length greater than 10 kb and score higher than 95% according to VirHostMatcher-Net.

2.4. Statistics Analysis

Alpha diversity of phage community was measured using qiime2 (https://qiime2.org; accessed on 29 January 2020). Principal Coordinates Analysis (PCoA) based on "Bray-Curtis" similarities was performed using R (v3.2, package vegan, The R Foundation, Vienna, Austria, 2016). Permutational Multivariate Analysis of Variance (PERMANOVA) was used to test the significant difference. All data performed statistical analyses, which were conducted in Prism 9- GraphPad (v9.0.0, GraphPad Software, San Diego, CA, USA, 2020) for the two-way analysis of variance [ANOVA], Tukey's post hoc test, and R (v4.0.2, stats package, The R Foundation, Vienna, Austria, 2020) for the Kruskal–Wallis and Dunn's post hoc test. The Jonckheere trend test was conducted in IBM SPSS Statistics (v27.0, IBM Corporation, Armonk, NY, USA, 2020). Meanwhile, multiple testing correction were performed to adjust the p value based on the "Bonferroni Holm" method. Only significant differences were shown in figures. Graphs were generated using Prism 9- GraphPad (v9.0.0, GraphPad Software, San Diego, CA, USA, 2020), Origin (v2020b, OriginLab Corporation, Northampton, MA, USA, 2020), Microsoft Excel (v365, Microsoft Corporation, Redmond, WA, USA), and R (v3.3.3, ggplot2 package, The R Foundation, Vienna, Austria, 2017). The data in results are provided as average ± SE.

3. Results

3.1. Gut Bacteriophage Community Structure Differed for BE and EAC Compared to Their Healthy Counterparts

On average, 43 ± 2% of all reads generated through sequencing were from viruses. In total, 854 ± 50, 1136 ± 19, 920 ± 33 viral contigs were obtained from sequences identified as viruses for CT, BE, and EAC, respectively. On average, from these contigs, over 95% of sequences were assigned to phages. The order of *Caudovirales*, which included *Herelleviridae*, *Myoviridae*, *Podoviridae*, *Siphoviridae*, and *Unclassified Caudovirales*, were the most abundant phages, accounting for more than 50% of total sequences in all three groups (Figure 1a, Figure S1). Among those phage families, the relative abundance of *Herelleviridae* was lower than 1% in three groups, the relative abundance of *Myoviridae* (1.12–41.97% in CT, 7.19–18.61% in BE, 1.37–34.36% in EAC), *Podoviridae* (2.03–31.68% in CT, 5.72–18.44% in BE, 3.72–11.01% in EAC) and *Siphoviridae* (8.28–79.60% in CT, 36.89–57.19% in BE, 41.48–75.69% in EAC) showed great variation within each group ($p > 0.05$). Some viral contigs were assigned to other phage or viral families including *Inoviridae*, *Microviridae*, *Tectiviridae*, *Herpesvirales*, *Marseilleviridae*, and *Pithoviridae* with a relative abundance of less than 1%. Meanwhile, the large difference in specific viral taxa between individuals was observed in the same group, which may be attributable to multiple factors such as age, gender, diet, or drug usage (Table S1). We next determined the dominant phage replication cycle (lytic versus lysogenic cycle). On average, EAC samples had more temperate phages (lysogenic cycle) than BE and CT ($p > 0.05$), 11.97% ± 2.43% in CT, 13.47% ± 1.15% in BE, 19.13% ± 4.90% in EAC (Figure S2).

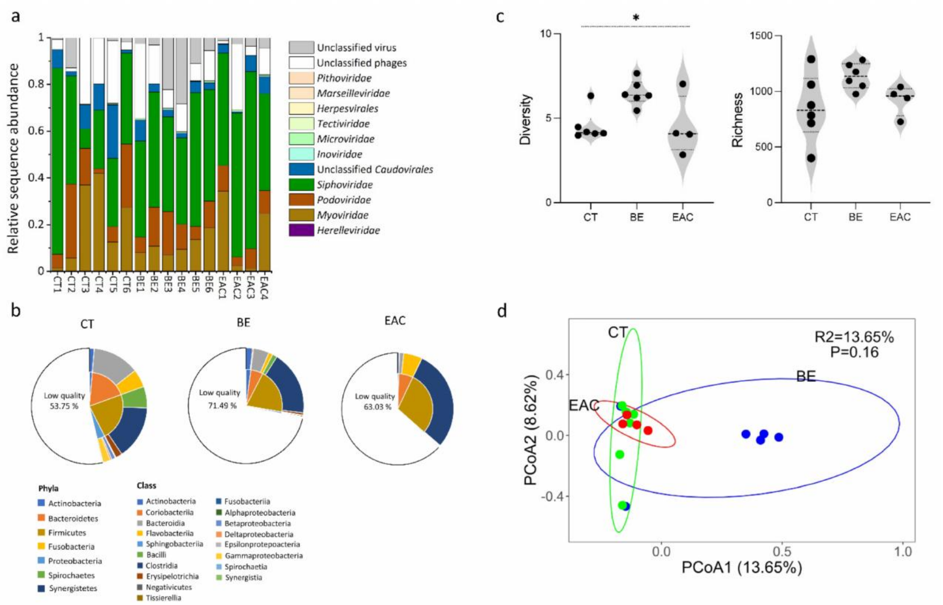


Figure 1. Composition of CT, BE, and EAC VLPs. (**a**) Relative abundance of viral families in CT, BE, and EAC; (**b**) The percentage of predicted bacterial hosts in CT, BE, and EAC. The inner cycle represents bacterial hosts at the phylum level, the outer cycle represents bacterial hosts at the class level. The low quality represents bacterial hosts predicted by contigs with a length lower than 10 kb and the score was lower than 95%; (**c**) Viral alpha diversity including richness (Ace) and diversity (Shannon) in samples from CT, BE, and EAC; (**d**) PCoA plot of the viral community composition based on the Bray–Curtis distances in CT, BE, and EAC samples. CT represents stool samples from healthy controls; BE represents stool samples from Barrett Esophagus patients; EAC represents stool samples from Esophageal Adenocarcinoma patients. Error bars indicate the average ± SE. Statistical significance was determined by Kruskal–Wallis, Dunn's post hoc test, asterisk indicates $p < 0.05$.

We next predicted the bacterial host range of the viral contigs from different groups in the study (Figure 1b). We observed that the bacterial hosts mainly spanned the phyla *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, and *Proteobacteria*, which were common across all three groups. In addition, we found that less than 0.1% of the phages were predicted to infect *Fusobacteria*, *Spirochaetes*, and *Synergistetes*. When the predicted bacterial host in class level was further compared, their relative abundance showed more obvious variation among the different groups, but these results were not statistically significant. For *Actinobacteria*, the relative abundance in CT (1.33% ± 0.28%) and BE (1.77% ± 0.21%) was higher than EAC (0.37% ± 0.11%) ($p > 0.05$). For *Flavobacteriia*, the relative abundance in CT (5.02% ± 1.45%) and EAC (5.38% ± 1.45%) was higher than BE (1.14% ± 0.16%) ($p > 0.05$). Notably, the classes *Betaproteobacteria*, *Deltaproteobacteria*, and *Gammaproteobacteria* were more abundant in CT compared with BE and EAC. Moreover, the relative abundance of *Bacteroidia* (13.08% ± 2.34% in CT, 4.38% ± 0.45% in BE, 1.25% ± 0.22% in EAC), *Bacilli* (5.97% ± 1.51% in CT,1.33% ± 0.14% in BE, <0.1% in EAC), and *Erysipelotrichia* (1.86% ± 0.54% in CT, 0.65% ± 0.093% in BE, <0.1% in EAC) were lower in BE and EAC compared to CT, while the relative abundance of *Clostridia* (15.06% ± 0.52% in CT, 18.04% ± 0.90% in BE, 29.20% ± 5.60% in EAC) was higher in BE and EAC compared to CT. However, there was no significant difference (Jonckheere trend test, $p > 0.05$). Furthermore, the remaining classes had a lower relative abundance (0.0001%–0.31%) across the three groups.

We further examined how the changes in phages community composition affected the overall diversity. For the alpha diversity, a significant difference in phage diversity (Shannon) was found among the three groups ($p = 0.036$), while no significant difference was observed in phage richness (Ace) ($p > 0.05$) (Figure 1c). Furthermore, the alpha diversity showed differences among BE and EAC compared to CT samples ($p > 0.05$). Specifically, in both BE (1136.17 ± 19.48) and EAC (920.50 ± 33.87), the richness (Ace) was higher compared with that in CT (854.00 ± 50.73). However, only in BE (6.50 ± 0.11), the diversity (Shannon) was higher compared with that in CT (4.53 ± 0.15). Furthermore, BE had a higher level of richness (Ace) and diversity (Shannon) than EAC. In addition, no significant difference was detected ($p > 0.05$) in beta diversity (PCoA) among the three groups (Figure 1d).

3.2. Abundant and Rare Phage Communities in the Gut May Contribute to the Progress of Esophageal Carcinogesis

We used a sorting approach commonly applied in ecological study that classifies microbes into three groups based on their abundance [55,56], aiming to explore the role of less abundant microbes in different ecosystems. Using this approach, the contribution of rare, less abundant, bacterial Operational Taxonomic Units (OTUs) to some of the key ecological functions was revealed in the environment [57], which was previously overlooked. We believe this approach can be beneficial for studying phages in the gut. To this end, we divided phage contigs into abundant phages (relative abundance was more than 1% in total viral contigs), moderate phages (relative abundance was more than 0.1% and less than 1% in total viral contigs), and rare phages (relative abundance was less than 0.1% in total viral contigs). At these three relative abundance levels, members of the order *Caudovirales* (*Myoviridae, Siphoviridae*, and *Podoviridae*) showed the highest relative abundance in all three groups (Figure 2a). Subsequently, we observed that abundant phages presented significantly higher relative abundance (79.54% ± 2.28% in CT, 54.28% ± 2.19% in BE, and 72.25% ± 4.06% in EAC) when compared with moderate (14.79% ± 1.83% in CT, 34.38% ± 1.68% in BE, and 21.19% ± 3.57% in EAC) and rare phages (4.51% ± 0.52% in CT, 11.34% ± 0.85% in BE, and 6.56% ± 0.52% in EAC) in all three groups (abundant vs. moderate $p < 0.001$, abundant vs rare $p < 0.001$) (Figure 2b,c), while the highest number of contigs was from rare phages (788 ± 48 in CT, 994 ± 18 in BE, and 836 ± 28 in EAC), exceeding abundant (13 ± 1 in CT, 17 ± 1 in BE, and 11 ± 1 in EAC) and moderate (54 ± 8 in CT, 126 ± 7 in BE, and 74 ± 15 in EAC) phages in all three groups (Figure 2b,c). The highest relative abundance of abundant phages and the highest number of contigs of rare phages may suggest their different behaviors in relation to the gut bacterial community and esophageal diseases. Moreover, a significant difference was observed in beta-diversity on abundant ($p = 0.004$) and rare phages ($p = 0.003$) (Figure S3), which may imply that these two groups of phages showed higher sensitivity to the changes in the upper GI tract through esophageal disease progression. In addition, we found that the abundance of temperate phages that displayed a lysogenic replication cycle increased with the development of esophageal diseases. This may suggest a higher occurrence of HGT in these samples.

To further evaluate the importance of rare phages in HGT, we compared these three groups of phages to the number of bacterial hosts they infect. On the class level, we observed small differences between phage groups from different health conditions, rare phages infected 18 different bacterial classes whereas abundant phages infected 14 (Figure 1c). However, when bacterial hosts were compared on the genus level, both diversity and abundance showed large differences, 84 for rare versus 46 for abundant phages (Table S2). In particular, contigs belonging to rare phages showed similar characteristics regarding the number of hosts they infect over three groups, showing a broader bacterial host range compared to moderate and abundant phages. For example, the contigs from rare phages were able to infect 6 or 7 different bacterial hosts at the genus level (Table S3), which was relatively higher than the bacterial hosts predicted for the contigs from abundant and moderate phages. The broader bacterial host range and higher number of contigs (Figure 2b, Tables S2 and S3) of rare phages could potentially lead to storing more AMGs in their genomes and, in turn, expand the frequency of HGT between gut bacteria.

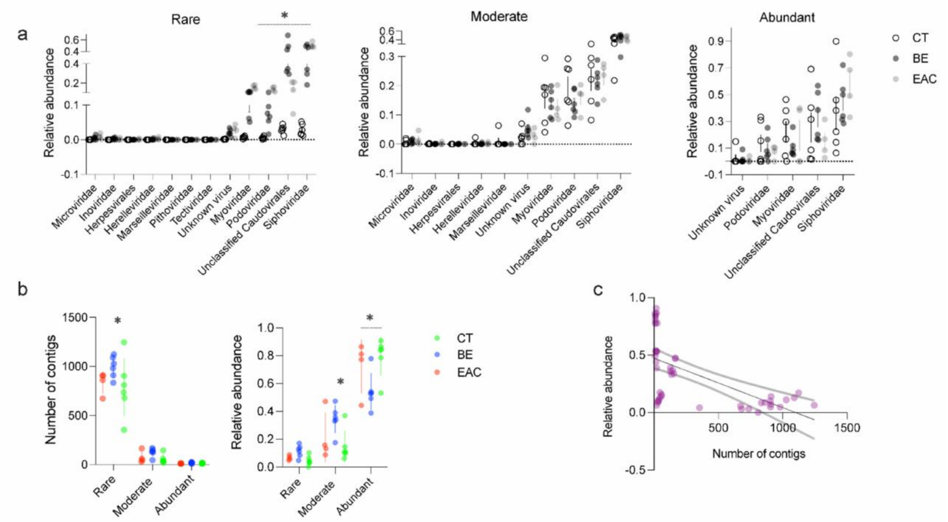


Figure 2. Composition of the rare, moderate, and abundant gut viruses in CT, BE and EAC samples. Rare, moderate, and abundant viruses were categorized based on the viral contig level. Abundant viruses represent viral contigs whose relative abundance was more than 1% in total contigs, moderate viruses represent viral contigs whose relative abundance was more than 0.1% and less than 1% in total contigs, and rare viruses represent viral contigs whose relative abundance was less than 0.1% in total contigs. (**a**) The relative abundance of viral families; (**b**) Number of contigs generated each viral contig category, rare, moderate, and abundant, on left and relative abundance of them on right. (**c**) Negative correlation between number of contigs, from rare, moderate, and abundant phages, and their relative abundance. CT represents stool samples from healthy controls; BE represents stool samples from Barrett Esophagus patients; EAC represents stool samples from Esophageal Adenocarcinoma patients. Statistical significance was determined by two–way analysis of variance [ANOVA], Tukey's post hoc test, asterisk indicates $p < 0.05$.

3.3. AMGs Found in Rare Bacteriophages Showed Increment in Esophageal Diseases

After annotation of the viral contigs, viruses were found to be involved in most of the microbial functions related to metabolism, cellular processes, genetic information processing, environment information processing, organismal system, and human disease (Figures 3a and S4). Significant differences were found for genes related to metabolism of cofactors and vitamins ($p = 0.0083$) and genes related to the prokaryotic defense system among the three groups ($p = 0.0202$) (Figure 3a). Genes involved in metabolism of cofactors and vitamins were found to be most abundant in CT phages, whereas genes related to the prokaryotic defense system were more abundant in EAC phages, suggesting a stronger arms race between phages and bacteria in this disease (Figure 3a). Notably, AMGs encoding bacterial toxins were found to be more abundant in the genome of rare bacteriophages including the *spyA* gene, *tccC* gene, *entB* gene and *entD* gene, which are involved in microbial cellular processes. The *spyA* gene, which encodes a C3 family ADP-ribosyltransferase (bacterial exotoxin) [58], showed a slightly higher level of relative abundance in BE and EAC ($p > 0.05$) compared with the other three AMGs (Figure 3b). Moreover, the *spyA* gene level was relatively higher in BE (0.00040 ± 0.00011) and EAC (0.0027 ± 0.0012) compared with CT (0.00031 ± 0.000012) ($p > 0.05$). Other AMGs that relate to LPS biosynthesis proteins were also found in the genome of rare phages including the *lpxD* gene, *kdsC* gene and *gmnB* gene, which are involved in microbial metabolism (Figure 3b). The *lpxD* gene only presented in BE with a relative abundance of 0.00031 ± 0.000113. The *kdsC* gene presented in BE (0.000089 ± 0.000036) and CT (0.0000024 ± 0.00000097). For the *gmnB* gene, it was relatively higher in EAC (0.00064 ± 0.00029) and BE (0.00024 ± 0.000094) compared with CT (0.00015 ± 0.000044) ($p > 0.05$). The higher abundance of these genes in phages from BE and EAC compared to CT may have resulted from the increase of pathogenic bacteria, mainly Gram-negatives, in the esophageal diseases, leading to a higher chance of obtaining AMGs, which are related to LPS biosynthesis proteins encoded by phages. We next explored the appearance of these genes in the Gut Phages Database (GPD) containing 142,809 non-redundant globally distributed phage genomes. We found many phages encoding these genes in GPD with one exception, *tccC*, showing these AMGs are ubiquitous in the human gut (Figure S5). Toxin complex (Tc) is a multisubunit toxin consisting of three components (TcA, B, and C) encoded by pathogenic bacteria infecting both insects and humans. TcAs that make functional pores combine with TcB-TcC subunits to create active chimeric holotoxins. Tc toxins are encoded by human pathogens like *Yersinia pestis*, *Y. pseudotuberculosis*, and *Morganella morganii* and are believed to significantly contribute to these bacteria's pathogenicity. Yet, their role in EAC remains to be revealed [59]. The increase of these genes in phages from BE and EAC may contribute to the severity of these diseases through exchanging genes that are involved in bacterial exotoxin production and LPS biosynthesis in esophageal carcinogenesis. This warrants further investigation.

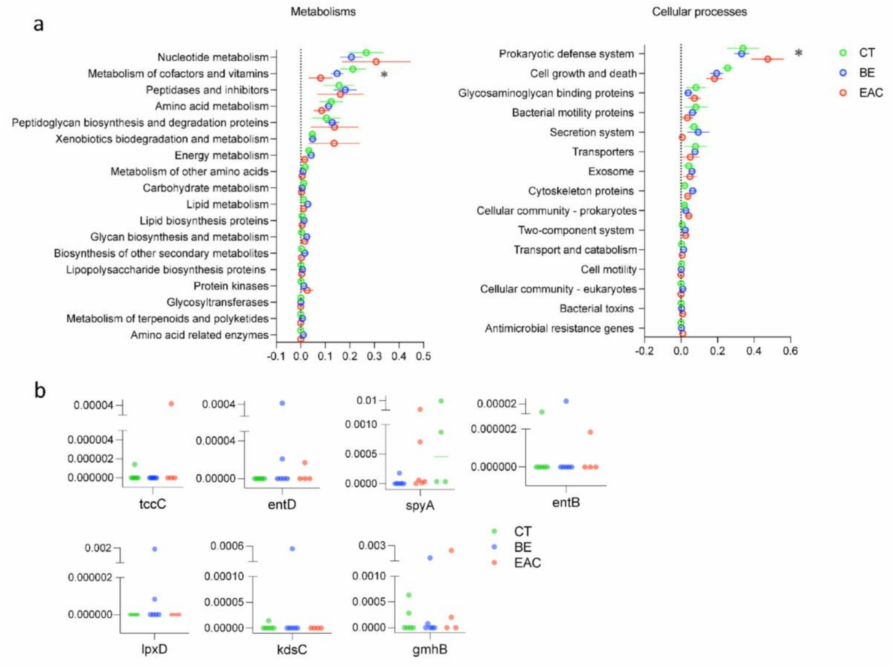


Figure 3. Viral functional traits. (**a**) The relative abundance of different functional traits in viral sequences; (**b**) The relative abundance of genes encoding four different bacterial toxins with higher abundance in BE and EAC samples compared with CT on the top, and genes encoding the LPS biosynthesis proteins on the bottom. Error bars indicate the average ± SE. Statistical significance was determined by two–way analysis of variance [ANOVA], Tukey's post hoc test, asterisk indicates $p < 0.05$.

4. Discussion

Barrett's esophagus (BE) is a condition caused by the metaplastic replacement of the normal squamous epithelium by columnar epithelium. BE is closely associated with the development of esophageal adenocarcinoma (EAC), a disease in which cancerous cells develop in the tissues of the esophagus with a high mortality rate [42]. It has been recently shown that gut dysbiosis can activate oncogenic signaling pathways, leading to the production of tumor-promoting metabolites, and further influence the esophageal mucosal inflammation and tumorigenesis [60]. For example, gut bacteria regulate bile acid (BA) metabolism. Under stimulation such as a high-fat diet, the gut bacteria changed, and the level of BA increased accordingly [61]. The reflux of BA to the esophagus caused esophageal damage, leading to BE and subsequent EAC. In an animal experiment simulating BA reflux, overexpression of the inflammatory cells, IL-6 and TNF- α, was found [62]. This indicated that gut bacterial alterations could indirectly induce the esophageal mucosal inflammation and carcinogenesis [62–64]. Despite a wealth of data on the role of gut bacteria in GI tract disease, we have only recently recognized the association of gut viruses with some GI tract diseases, including CRC in which the diversity of the gut viruses is significantly increased in stool samples from CRC patients, suggesting a disease-specific signature that can be used to differentiate CRC samples from controls [37]. The CRC-associated virome includes primarily temperate bacteriophages belonging to *Siphoviridae* and *Myoviridae* families [65]. The impact of phages on gut homeostasis is not restricted to their interactions with gut bacteria as phages can directly interact with the human host. In vitro studies have demonstrated that phages can cross the epithelial cell layer through transcytosis, thereby stimulating the underlying immune cells [22,66–69]. For example, the interaction between *E.coli* phages and the immune system has been associated with Type I Diabetes autoimmunity [36]. It has been reported that phages can activate IFN-γ produced by CD4+ T cells via the nucleotide-sensing receptor TLR9, which accelerates intestinal inflammation and colitis, leading to a systemic inflammation response [70]. The consistent disease-specific signature of gut viruses [27,37], suggests a potential association between gut viruses and human disease.

Studies that investigated the esophageal virome, using metagenomic data of whole microbial communities rather than profiling the isolated viral communities, have identified a range of phages, including *Streptococcus*, *Campylobacter*, *Lactococcus*, and γ-Proteobacteria phages [71]. The aforementioned and those that only explored the bacterial community of the esophagus have mainly used biopsy samples for virome and bacterium analysis [10,72,73]. Although, biopsies could directly reflect the disease-associated microbial signature at the lesion, the sampling procedure is invasive, time-consuming, costly, and may induce potential complications [74]. Moreover, biopsy samples often have limited microbial materials, with a lower probability of successful sequencing and downstream analysis [75]. Thus, an amplification step (e.g., whole genome amplification) is necessary, which might introduce biases to study results. On the contrary, stool samples collected by non-invasive methods often supply sufficient materials for research purposes [76].

Here we explored stool samples from BE, EAC, and CT phages community composition in esophageal diseases. Our in-depth gut virome analysis during esophageal carcinogenesis provided some evidence of gut phage community changes between different stages of esophageal diseases. Consistent with previous studies that have explored the gut viruses, mainly in the lower GI tract diseases such as IBD and CRC [27,65], phages from the order *Caudovirales* were the most dominant phages in the samples from esophageal diseases. Compared with CT, the alpha diversity has changed with the esophageal diseases progress, and a relatively higher alpha diversity was observed in BE samples compared to CT and EAC. This was not reflected in the beta diversity as no significant differences were observed among three groups. Using a common sorting approach in microbial ecology, we identified disease-associated differences in diversity and abundance of rare phages, suggesting a potential link between these phages and esophageal diseases. In addition, consistent with previous studies on diseases like IBD [77] and CRC [65], we observed changes in the proportion of lytic/lysogenic replication cycles of phages, and more temperate phages were observed in esophageal carcinogenesis. These results further support earlier studies that reported the dominance of virulent phages (lytic cycle) in the healthy human gut replaced by temperate phages in Crohn's disease and ulcerative colitis [23,24]. Furthermore, the relatively higher percentage of temperate phages in samples from esophageal diseases may imply more influence on the bacterial physiology through phage mediated HGT in those groups. However, we did not study the bacterial community of these samples, the community structure of the predicted bacterial hosts for the phages identified in the study may suggest a complex relationship between bacteria and bacteriophage community in esophageal diseases. Earlier studies on lower GI tract diseases such as CRC have observed that the effect of phages resulted from their interactions with the whole bacterial community, rather than the bacterial taxa directly contributing to the disease severity [65]. However, there was no direct correlation between bacterial diversity and phage diversity [27,37].

In addition, we found several AMGs in the genome of the rare phages, further emphasizing the potential role of phages in regulating bacterial physiology by supplying their host with beneficial genes. Specifically, a slightly higher abundance of *spyA* ($p > 0.05$) was observed in BE and EAC, potentially contributing to the production of bacterial exotoxins, which disrupt cytoskeletal structures and promote colonization of pathogenic bacteria [58]. The relatively higher abundance of AMGs related to LPS biosynthesis proteins were also found in BE and EAC, which may indicate the dominance of Gram-negative bacteria and the potential inflammatory effects of phage–bacteria interactions. Phages that carry these AMGs can introduce these genes to the genome of gut bacteria via integration, which may contribute to the severity of the esophageal diseases through lysogenic conversion. This could further induce gut inflammation through expression of the phage-derived virulence genes and deteriorate esophageal disease. Intestinal permeability caused by phage-mediated changes of gut microbiota could also lead to systemic inflammatory responses [78]. Given the high variability of the microbiome between individuals and the limited number of samples analyzed, it is difficult to identify significant differences in viral community structure between different groups in the current study. Thus, our findings should be further pursued with a larger sample size.

5. Conclusions

In summary, this study provides further evidence of potential relationship between gut phages and esophageal diseases. Interestingly, the distinct gut phage community structure was identified in two different stages of esophageal diseases, and these differences were mainly found in abundant and rare bacteriophages. Notably, rare phages and HGT mediated by them have been found to be more related to esophageal diseases. Specially, the rare phages contributed to enriching AMGs related to bacterial exotoxin and LPS biosynthesis proteins, and the possible upregulated level of these genes. These, in turn, may contribute to changes in the gut bacterial composition and inflammation, which lead to the development of esophageal diseases, as previously suggested [6]. However, given the small sample size in our study, the potential diagnostic importance of AMGs and disease-specific viral signature identified should be experimentally validated in further studies.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/microorganisms9081701/s1, Figure S1: Gene-sharing taxonomic network of viral sequences in this study, including viral RefSeq viruses v85. Figure S2: The proportion of lytic/lysogenic replication cycles predicted for the viral contigs from three groups. Figure S3: PCoA plot of the viral community composition based on the Bray-Curtis distances in CT, BE, and EAC samples. Figure S4: The relative abundance of different functional traits in viral sequences. Figure S5: The number of phages that contained the identified AMGs of this study in the Gut Phage Database (GPD). Table S1: Clinical information of individuals from three groups. Table S2: The relative abundance of bacterial host at genus level for abundant, moderate, and rare bacteriophages. Table S3: The percentage of contig relative abundance in different number of predicted bacterial genus types for abundant, moderate, and rare bacteriophages. Table S4: The relative abundance of identified AMGs.

Author Contributions: Conceptualization, K.-P.J., M.Q. and L.D.; methodology, T.M.; formal analysis, T.M., J.R., and M.K.M.; investigation, T.M. and J.X.; writing, T.M., J.R., S.S., J.X., M.K.M., K.-P.J., M.Q. and L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by German Research Foundation (DFG Emmy Noether program, Project No. 273124240; and SFB/CRC 1371, Project No. 395357507, awarded to L.D., M.Q. and K.-P.J.) and European Research Council starting grant (ERC StG 803077) awarded to L.D., as well as the Deutsche Krebshilfe Max Eder Program (no grant number available) awarded to M.Q.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in NCBI Sequence Read Archive (SRA) under accession number SUB8621833.

Acknowledgments: We would like to thank the groups of Dirk Haller, Klaus-Peter Janssen and Markus List at the Technical University of Munich, as well as the member of Deng and Quante Lab for their support in sampling, logistics, and constructive discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, Y.; Li, D.; Shan, B.; Liang, D.; Shi, J.; Chen, W.; He, J. Incidence and mortality of esophagus cancer in China, 2008–2012. *Chin. J. Cancer Res.* **2019**, *31*, 426–434. [CrossRef] [PubMed]
2. Pennathur, A.; Gibson, M.K.; Jobe, B.A.; Luketich, J.D. Oesophageal carcinoma. *Lancet* **2013**, *381*, 400–412. [CrossRef]
3. Martinucci, I.; De Bortoli, N.; Russo, S.; Bertani, L.; Furnari, M.; Mokrowiecka, A.; Malecka-Panas, E.; Savarino, V.; Savarino, E.; Marchi, S. Barrett's esophagus in 2016: From pathophysiology to treatment. *World J. Gastrointest. Pharmacol. Ther.* **2016**, *7*, 190–206. [CrossRef]
4. Pohl, H.; Koch, M.; Khalifa, A.; Papanikolaou, I.; Scheiner, K.; Wiedenmann, B.; Rösch, T. Evaluation of endocytoscopy in the surveillance of patients with Barrett's esophagus. *Endoscopy* **2007**, *39*, 492–496. [CrossRef]
5. Hooper, L.V.; Gordon, J.I. Commensal host-bacterial relationships in the gut. *Science* **2001**, *292*, 1115–1118. [CrossRef]
6. Münch, N.S.; Fang, H.-Y.; Ingermann, J.; Maurer, H.C.; Anand, A.; Kellner, V.; Sahm, V.; Wiethaler, M.; Baumeister, T.; Wein, F.; et al. High-fat diet accelerates carcinogenesis in a mouse model of barrett's esophagus via interleukin 8 and alterations to the gut microbiome. *Gastroenterology* **2019**, *157*, 492–506. [CrossRef]
7. Snider, E.J.; Compres, G.; Freedberg, D.E.; Khiabanian, H.; Nobel, Y.R.; Stump, S.; Uhlemann, A.-C.; Lightdale, C.J.; Abrams, J.A. Alterations to the esophageal microbiome associated with progression from Barrett's esophagus to esophageal adenocarcinoma. *Cancer Epidemiol. Prev. Biomark.* **2019**, *28*, 1687–1693. [CrossRef]
8. Ren, Z.; Li, A.; Jiang, J.; Zhou, L.; Yu, Z.; Lu, H.; Xie, H.; Chen, X.; Shao, L.; Zhang, R.; et al. Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early hepatocellular carcinoma. *Gut* **2019**, *68*, 1014–1023. [CrossRef]
9. Manor, O.; Dai, C.L.; Kornilov, S.A.; Smith, B.; Price, N.D.; Lovejoy, J.C.; Gibbons, S.M.; Magis, A.T. Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* **2020**, *11*, 5206. [CrossRef] [PubMed]
10. Yang, L.; Lu, X.; Nossa, C.W.; Francois, F.; Peek, R.M.; Pei, Z. Inflammation and intestinal metaplasia of the distal esophagus are associated with alterations in the microbiome. *Gastroenterology* **2009**, *137*, 588–597. [CrossRef]
11. Yang, L.; Francois, F.; Pei, Z. Molecular pathways: Pathogenesis and clinical implications of microbiome alteration in esophagitis and Barrett esophagus. *Clin. Cancer Res.* **2012**, *18*, 2138–2144. [CrossRef]
12. Zaidi, A.H.; Kelly, L.A.; Kreft, R.E.; Barlek, M.; Omstead, A.N.; Matsui, D.; Boyd, N.H.; Gazarik, K.E.; Heit, M.I.; Nistico, L.; et al. Associations of microbiota and toll-like receptor signaling pathway in esophageal adenocarcinoma. *BMC Cancer* **2016**, *16*, 52. [CrossRef] [PubMed]
13. Okereke, I.; Hamilton, C.; Wenholz, A.; Jala, V.; Giang, T.; Reynolds, S.; Miller, A.; Pyles, R. Associations of the microbiome and esophageal disease. *J. Thorac. Dis.* **2019**, *11*, S1588–S1593. [CrossRef]
14. Wylie, K.M.; Weinstock, G.M.; Storch, G.A. Emerging view of the human virome. *Transl. Res.* **2012**, *160*, 283–290. [CrossRef] [PubMed]
15. Carding, S.R.; Davis, N.; Hoyles, L. The human intestinal virome in health and disease. *Aliment. Pharmacol. Ther.* **2017**, *46*, 800–815. [CrossRef] [PubMed]
16. Cadwell, K. The virome in host health and disease. *Immunity* **2015**, *42*, 805–813. [CrossRef] [PubMed]
17. Lepage, P.; Leclerc, M.C.; Joossens, M.; Mondot, S.; Blottière, H.M.; Raes, J.; Ehrlich, D.; Doré, J. A metagenomic insight into our gut's microbiome. *Gut* **2013**, *62*, 146–158. [CrossRef] [PubMed]

18. Mills, S.; Shanahan, F.; Stanton, C.; Hill, C.; Coffey, A.; Ross, R.P. Movers and shakers: Influence of bacteriophages in shaping the mammalian gut microbiota. *Gut Microbes* **2013**, *4*, 4–16. [CrossRef] [PubMed]
19. Dalmasso, M.; Hill, C.; Ross, R.P. Exploiting gut bacteriophages for human health. *Trends Microbiol.* **2014**, *22*, 399–405. [CrossRef]
20. Ungaro, F.; Massimino, L.; Furfaro, F.; Rimoldi, V.; Peyrin-Biroulet, L.; D'alessio, S.; Danese, S. Metagenomic analysis of intestinal mucosa revealed a specific eukaryotic gut virome signature in early-diagnosed inflammatory bowel disease. *Gut Microbes* **2019**, *10*, 149–158. [CrossRef]
21. Tetz, G.; Tetz, V. Prion-like domains in eukaryotic viruses. *Sci. Rep.* **2018**, *8*, 8931. [CrossRef] [PubMed]
22. Breitbart, M.; Hewson, I.; Felts, B.; Mahaffy, J.M.; Nulton, J.; Salamon, P.; Rohwer, F. Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **2003**, *185*, 6220–6223. [CrossRef] [PubMed]
23. Reyes, A.; Haynes, M.; Hanson, N.; Angly, F.E.; Heath, A.C.; Rohwer, F.; Gordon, J.I. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **2010**, *466*, 334–338. [CrossRef]
24. Sabino, J.; Hirten, R.P.; Colombel, J.F. bacteriophages in gastroenterology—From biology to clinical applications. *Aliment. Pharmacol. Ther.* **2020**, *51*, 53–63. [CrossRef]
25. De Sordi, L.; Lourenço, M.; Debarbieux, L. The battle within: Interactions of bacteriophages and bacteria in the gastrointestinal tract. *Cell Host Microbe* **2019**, *25*, 210–218. [CrossRef]
26. Łusiak-Szelachowska, M.; Weber-Dąbrowska, B.; Jończyk-Matysiak, E.; Wojciechowska, R.; Górski, A. Bacteriophages in the gastrointestinal tract and their implications. *Gut Pathog.* **2017**, *9*, 44. [CrossRef]
27. Norman, J.M.; Handley, S.A.; Baldridge, M.T.; Droit, L.; Liu, C.Y.; Keller, B.C.; Kambal, A.; Monaco, C.L.; Zhao, G.; Fleshner, P.; et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **2015**, *160*, 447–460. [CrossRef] [PubMed]
28. Mirzaei, M.K.; Xue, J.; Costa, R.; Ru, J.; Schulz, S.; Taranu, Z.E.; Deng, L. Challenges of studying the human virome–relevant emerging technologies. *Trends Microbiol.* **2021**, *29*, 171–181. [CrossRef]
29. Lin, E.C.; Lynch, A.S. *Regulation of Gene Expression in Escherichia coli*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
30. Feiner, R.; Argov, T.; Rabinovich, L.; Sigal, N.; Borovok, I.; Herskovits, A.A. A new perspective on lysogeny: Prophages as active regulatory switches of bacteria. *Nat. Rev. Microbiol.* **2015**, *13*, 641–650. [CrossRef] [PubMed]
31. Brüssow, H.; Canchaya, C.; Hardt, W.-D. Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **2004**, *68*, 560–602. [CrossRef] [PubMed]
32. Saunders, J.R.; Allison, H.; James, C.E.; McCarthy, A.J.; Sharp, R. Phage-mediated transfer of virulence genes. *J. Chem. Technol. Biotechnol.* **2001**, *76*, 662–666. [CrossRef]
33. Coleman, D.C.; Sullivan, D.J.; Russel, R.J.; Arbuthnott, J.P.; Carey, B.F.; Pomeroy, H.M. Staphylococcus aureus bacteriophages mediating the simultaneous lysogenic conversion of β-lysin, staphylokinase and enterotoxin A: Molecular mechanism of triple conversion. *Microbiology* **1989**, *135*, 1679–1697. [CrossRef]
34. Crummett, L.T.; Puxty, R.J.; Weihe, C.; Marston, M.F.; Martiny, J.B. The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. *Virology* **2016**, *499*, 219–229. [CrossRef]
35. Tetz, G.; Brown, S.M.; Hao, Y.; Tetz, V. Parkinson's disease and bacteriophages as its overlooked contributors. *Sci. Rep.* **2018**, *8*, 10812. [CrossRef]
36. Tetz, G.; Brown, S.M.; Hao, Y.; Tetz, V. Type 1 diabetes: An association between autoimmunity, the dynamics of gut amyloid-producing E. coli and their phages. *Sci. Rep.* **2019**, *9*, 9685. [CrossRef]
37. Nakatsu, G.; Zhou, H.; Wu, W.K.K.; Wong, S.H.; Coker, O.O.; Dai, Z.; Li, X.; Szeto, C.-H.; Sugimura, N.; Lam, T.Y.-T.; et al. Alterations in enteric virome are associated with colorectal cancer and survival outcomes. *Gastroenterology* **2018**, *155*, 529–541. [CrossRef] [PubMed]
38. Duerkop, B.A.; Kleiner, M.; Paez-Espino, D.; Zhu, W.; Bushnell, B.; Hassell, B.; Winter, S.E.; Kyrpides, N.C.; Hooper, L.V. Murine colitis reveals a disease-associated bacteriophage community. *Nat. Microbiol.* **2018**, *3*, 1023–1031. [CrossRef]
39. Ly, M.; Abeles, S.R.; Boehm, T.K.; Robles-Sikisaka, R.; Naidu, M.; Santiago-Rodriguez, T.; Pride, D.T. Altered oral viral ecology in association with periodontal disease. *MBio* **2014**, *5*, e01133-14. [CrossRef]
40. Abeles, S.R.; Robles-Sikisaka, R.; Ly, M.; Lum, A.G.; Salzman, J.; Boehm, T.K.; Pride, D.T. Human oral viruses are personal, persistent and gender-consistent. *ISME J.* **2014**, *8*, 1753–1767. [CrossRef]
41. Deshpande, N.P.; Riordan, S.M.; Castaño-Rodríguez, N.; Wilkins, M.R.; Kaakoush, N.O. Signatures within the esophageal microbiome are associated with host genetics, age, and disease. *Microbiome* **2018**, *6*, 227. [CrossRef] [PubMed]
42. Wiethaler, M.; Slotta-Huspenina, J.; Brandtner, A.; Horstmann, J.; Wein, F.; Baumeister, T.; Radani, N.; Gerland, S.; Anand, A.; Lange, S.; et al. BarrettNET—A prospective registry for risk estimation of patients with Barrett's esophagus to progress to adenocarcinoma. *Dis. Esophagus* **2019**, *32*, doz024. [CrossRef]
43. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [CrossRef] [PubMed]
44. Nurk, S.; Meleshko, D.; Korobeynikov, A.; Pevzner, P.A. metaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **2017**, *27*, 824–834. [CrossRef] [PubMed]
45. Roux, S.; Enault, F.; Hurwitz, B.L.; Sullivan, M.B. VirSorter: Mining viral signal from microbial genomic data. *PeerJ* **2015**, *3*, e985. [CrossRef] [PubMed]

46. von Meijenfeldt, F.B.; Arkhipova, K.; Cambuy, D.D.; Coutinho, F.H.; Dutilh, B.E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **2019**, *20*, 217. [CrossRef]
47. Ren, J.; Song, K.; Deng, C.; Ahlgren, N.A.; Fuhrman, J.A.; Li, Y.; Xie, X.; Poplin, R.; Sun, F. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* **2020**, *8*, 64–77. [CrossRef] [PubMed]
48. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef]
49. Hyatt, D.; Chen, G.-L.; LoCascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **2010**, *11*, 119. [CrossRef]
50. Jang, H.B.; Bolduc, B.; Zablocki, O.; Kuhn, J.H.; Roux, S.; Adriaenssens, E.M.; Brister, J.R.; Kropinski, A.M.; Krupovic, M.; Lavigne, R.; et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **2019**, *37*, 632–639. [CrossRef]
51. Liao, Y.; Smyth, G.K.; Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923–930. [CrossRef] [PubMed]
52. Kieft, K.; Zhou, Z.; Anantharaman, K. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **2020**, *8*, 90. [CrossRef]
53. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **2017**, *45*, D353–D361. [CrossRef]
54. Wang, W.; Ren, J.; Tang, K.; Dart, E.; Ignacio-Espinoza, J.C.; Fuhrman, J.A.; Braun, J.; Sun, F.; Ahlgren, N.A. A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genom. Bioinform.* **2020**, *2*, lqaa044. [CrossRef]
55. Ji, M.; Kong, W.; Stegen, J.; Yue, L.; Wang, F.; Dong, X.; Cowan, D.A.; Ferrari, B.C. Distinct assembly mechanisms underlie similar biogeographical patterns of rare and abundant bacteria in Tibetan Plateau grassland soils. *Environ. Microbiol.* **2020**, *22*, 2261–2272. [CrossRef]
56. Sjöstedt, J.; Koch-Schmidt, P.; Pontarp, M.; Canbäck, B.; Tunlid, A.; Lundberg, P.; Hagström, Å.; Riemann, L. Recruitment of members from the rare biosphere of marine bacterioplankton communities after an environmental disturbance. *Appl. Environ. Microbiol.* **2012**, *78*, 1361–1369. [CrossRef]
57. Vigneron, A.; Cruaud, P.; Alsop, E.; de Rezende, J.R.; Head, I.M.; Tsesmetzis, N. Beyond the tip of the iceberg; a new view of the diversity of sulfite-and sulfate-reducing microorganisms. *ISME J.* **2018**, *12*, 2096–2099. [CrossRef]
58. Coye, L.H.; Collins, C.M. Identification of SpyA, a novel ADP-ribosyltransferase of Streptococcus pyogenes. *Mol. Microbiol.* **2004**, *54*, 89–98. [CrossRef] [PubMed]
59. Leidreiter, F.; Roderer, D.; Meusch, D.; Gatsogiannis, C.; Benz, R.; Raunser, S. Common architecture of Tc toxins from human and insect pathogenic bacteria. *Sci. Adv.* **2019**, *5*, eaax6497. [CrossRef]
60. Deng, Y.; Tang, D.; Hou, P.; Shen, W.; Li, H.; Wang, T.; Liu, R. Dysbiosis of gut microbiota in patients with esophageal cancer. *Microb. Pathog.* **2021**, *150*, 104709. [CrossRef] [PubMed]
61. Schwabe, R.F.; Jobin, C. The microbiome and cancer. *Nat. Rev. Cancer* **2013**, *13*, 800–812. [CrossRef] [PubMed]
62. Sun, D.; Wang, X.; Gai, Z.; Song, X.; Jia, X.; Tian, H. Bile acids but not acidic acids induce Barrett's esophagus. *Int. J. Clin. Exp. Pathol.* **2015**, *8*, 1384. [PubMed]
63. Schmidt, M.; Ankerst, D.P.; Chen, Y.; Wiethaler, M.; Slotta-Huspenina, J.; Becker, K.-F.; Horstmann, J.; Kohlmayer, F.; Lehmann, A.; Linkohr, B.; et al. Epidemiologic Risk Factors in a Comparison of a Barrett Esophagus Registry (BarrettNET) and a Case–Control Population in Germany. *Cancer Prev. Res.* **2020**, *13*, 377–384. [CrossRef]
64. Elliott, D.R.F.; Perner, J.; Li, X.; Symmons, M.F.; Verstak, B.; Eldridge, M.; Bower, L.; O'Donovan, M.; Gay, N.J.; Fitzgerald, R.C. Impact of mutations in Toll-like receptor pathway genes on esophageal carcinogenesis. *PLoS Genet.* **2017**, *13*, e1006808.
65. Hannigan, G.D.; Duhaime, M.B.; Ruffin, M.T.; Koumpouras, C.C.; Schloss, P.D. Diagnostic potential and interactive dynamics of the colorectal cancer virome. *MBio* **2018**, *9*, e02248-18. [CrossRef]
66. Sinha, A.; Maurice, C.F. Bacteriophages: Uncharacterized and dynamic regulators of the immune system. *Mediat. Inflamm.* **2019**, *2019*, 3730519. [CrossRef] [PubMed]
67. Nguyen, S.; Baker, K.; Padman, B.S.; Patwa, R.; Dunstan, R.A.; Weston, T.A.; Schlosser, K.; Bailey, B.; Lithgow, T.; Lazarou, M.; et al. Bacteriophage transcytosis provides a mechanism to cross epithelial cell layers. *MBio* **2017**, *8*, e01874. [CrossRef] [PubMed]
68. Bichet, M.C.; Chin, W.H.; Richards, W.; Lin, Y.-W.; Avellaneda-Franco, L.; Hernandez, C.A.; Oddo, A.; Chernyavskiy, O.; Hilsenstein, V.; Neild, A. Bacteriophage uptake by mammalian cell layers represents a potential sink that may impact phage therapy. *Iscience* **2021**, *24*, 102287. [CrossRef]
69. Wahida, A.; Tang, F.; Barr, J.J. Rethinking phage-bacteria-eukaryotic relationships and their influence on human health. *Cell Host Microbe* **2021**, *29*, 681–688. [CrossRef]
70. Gogokhia, L.; Buhrke, K.; Bell, R.; Hoffman, B.; Brown, D.G.; Hanke-Gogokhia, C.; Ajami, N.J.; Wong, M.C.; Ghazaryan, A.; Valentine, J.F.; et al. Expansion of bacteriophages is linked to aggravated intestinal inflammation and colitis. *Cell Host Microbe* **2019**, *25*, 285–299. [CrossRef]
71. Moonen, A.; Annese, V.; Belmans, A.; Bredenoord, A.J.; Des Varannes, S.B.; Costantini, M.; Dousset, B.; Elizalde, J.I.; Fumagalli, U.; Gaudric, M.; et al. Long-term results of the European achalasia trial: A multicentre randomised controlled trial comparing pneumatic dilation versus laparoscopic Heller myotomy. *Gut* **2016**, *65*, 732–739. [CrossRef]

72. Pei, Z.; Bini, E.J.; Yang, L.; Zhou, M.; Francois, F.; Blaser, M.J. Bacterial biota in the human distal esophagus. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 4250–4255. [CrossRef]
73. Pei, Z.; Yang, L. Bacterial biota in reflux esophagitis and Barrett's esophagus. *World J. Gastroenterol.* **2005**, *11*, 7277–7283. [CrossRef] [PubMed]
74. Fillon, S.A.; Harris, J.K.; Wagner, B.D.; Kelly, C.J.; Stevens, M.J.; Moore, W.; Fang, R.; Schroeder, S.; Masterson, J.C.; Robertson, C.E.; et al. Novel device to sample the esophageal microbiome—The esophageal string test. *PLoS ONE* **2012**, *7*, e42938. [CrossRef]
75. Lim, E.H.; Zhang, S.-L.; Li, J.-L.; Yap, W.-S.; Howe, T.-C.; Tan, B.-P.; Lee, Y.-S.; Wong, D.; Khoo, K.-L.; Seto, K.-Y.; et al. Using whole genome amplification (WGA) of low-volume biopsies to assess the prognostic role of EGFR, KRAS, p53, and CMET mutations in advanced-stage non-small cell lung cancer (NSCLC). *J. Thorac. Oncol.* **2009**, *4*, 12–21. [CrossRef] [PubMed]
76. Bull, M.J.; Plummer, N.T. Part 1: The human gut microbiome in health and disease. *Integr. Med. A Clin. J.* **2014**, *13*, 17.
77. Clooney, A.G.; Sutton, T.D.; Shkoporov, A.N.; Holohan, R.K.; Daly, K.M.; O'Regan, O.; Ryan, F.J.; Draper, L.A.; Plevy, S.E.; Ross, R.P.; et al. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. *Cell Host Microbe* **2019**, *26*, 764–778. [CrossRef] [PubMed]
78. Tetz, G.; Tetz, V. Bacteriophage infections of microbiota can lead to leaky gut in an experimental rodent model. *Gut Pathog.* **2016**, *8*, 33. [CrossRef]

A3 Manuscript 3

Journal of Hazardous Materials 459 (2023) 132299

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Research Paper

Unveiling the hidden role of aquatic viruses in hydrocarbon pollution bioremediation

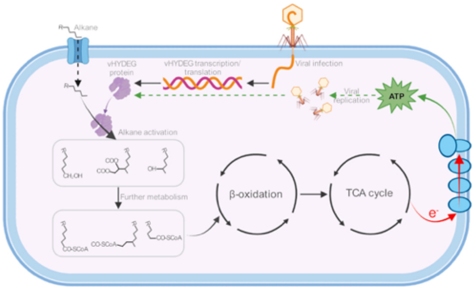
Jinlong Ru[a,b,1], Jinling Xue[a,b,1], Jianfeng Sun[c], Linda Cova[a], Li Deng[a,b,*]

[a] *Institute of Virology, Helmholtz Centre Munich - German Research Centre for Environmental Health, Neuherberg 85764, Germany*
[b] *Chair of Prevention for Microbial Infectious Disease, Central Institute of Disease Prevention and School of Life Sciences, Technical University of Munich, Freising 85354, Germany*
[c] *Botnar Research Centre, University of Oxford, Oxford OX3 7LD, UK*

HIGHLIGHTS

- Aquatic viruses encoding various hydrocarbon degradation genes (vHYDEGs).
- vHYDEGs are involved in the initial and rate-limiting steps of alkane hydroxylation.
- Protein structure prediction shows their identity and the catalytic activity sites.
- The viruses have diverse taxa, evolution history, related to multiple oil degraders.
- These vHYDEGs have potential engineering values for crude oil bioremediation.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Shaily Mahendra

Keywords:
Bioremediation
Bacteriophages
Virus-encoded hydrocarbon degradation genes
Alkane hydroxylases
Auxiliary metabolic gene

ABSTRACT

Hydrocarbon pollution poses substantial environmental risks to water and soil. Bioremediation, which utilizes microorganisms to manage pollutants, offers a cost-effective solution. However, the role of viruses, particularly bacteriophages (phages), in bioremediation remains unexplored. This study examines the diversity and activity of hydrocarbon-degradation genes encoded by environmental viruses, focusing on phages, within public databases. We identified 57 high-quality phage-encoded auxiliary metabolic genes (AMGs) related to hydrocarbon degradation, which we refer to as virus-encoded hydrocarbon degradation genes (vHYDEGs). These genes are encoded by taxonomically diverse aquatic phages and highlight the under-characterized global virosphere. Six protein families involved in the initial alkane hydroxylation steps were identified. Phylogenetic analyses revealed the diverse evolutionary trajectories of vHYDEGs across habitats, revealing previously unknown biodegraders linked evolutionarily with vHYDEGs. Our findings suggest phage AMGs may contribute to alkane and aromatic hydrocarbon degradation, participating in the initial, rate-limiting hydroxylation steps, thereby aiding hydrocarbon pollution bioremediation and promoting their propagation. To support future research, we developed

[*] Corresponding author at: Institute of Virology, Helmholtz Centre Munich - German Research Centre for Environmental Health, Neuherberg 85764, Germany.
E-mail address: li.deng@helmholtz-munich.de (L. Deng).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jhazmat.2023.132299
Received 1 June 2023; Received in revised form 28 July 2023; Accepted 12 August 2023
Available online 15 August 2023

vHyDeg, a database containing identified vHYDEGs with comprehensive annotations, facilitating the screening of hydrocarbon degradation AMGs and encouraging their bioremediation applications.

1. Introduction

The mining and processing of crude oil require vast amounts of water resources. Despite efforts to recycle water, a significant amount of the industry's wastewater still ends up in tailing ponds. In addition, oil product transportation is mainly through waterways, either by subwater pipelines or shipping. Although oil spillage is rare in this process, it can still cause significant environmental disasters, leading to adverse effects on both the environment and human health [62]. Every year, more than two million tons of oil enter the marine ecosystem, and over 85 % of it is from human activities [51]. Some hydrocarbons are carcinogenic, neurotoxic, and genotoxic to both humans and other organisms in the environment. In aquatic organisms, crude oil causes DNA damage, defects in cardiac function, and oxidative stress, resulting in reduced abundance and diversity of fish, ultimately disrupting ecosystems [1]. Bioremediation, which employs microorganisms to degrade hydrocarbons, has emerged as a promising and eco-friendly strategy for mitigating hydrocarbon pollution [37,87]. Despite the extensive research on the role of bacteria and fungi in hydrocarbon biodegradation [25,61,89,97,98], the potential contribution of viruses, especially bacteriophages, to this process remains unexplored.

Viruses, including (bacterio)phages, impact nearly all organisms on Earth, including microbial communities and their associated biogeochemical processes. Highly diverse viral communities have been identified in both soil and water ecosystems. Soil ecosystems are estimated to contain between 10^7 and 10^{10} viruses per gram of soil. Over 200,000 virus species have been discovered in the ocean, a number that is two orders of magnitude greater than earlier records [28]. Notably, viruses in the ocean are responsible for killing 20 % of microbial biomass daily, thereby playing a pivotal role in nutrient and energy cycles [78]. The global distribution of these viruses seems to be driven by a combination of multiple biotic and abiotic factors. In addition, recent studies have demonstrated that phages often contain auxiliary metabolic genes (AMGs), which contribute to the adaptive stress resistance of their bacterial hosts [71], augment host metabolism [40], and enhance fitness [47].

These findings suggest that phages may contribute to hydrocarbon degradation through undiscovered AMGs. However, unlike the numerous bacterial and non-methanogenic archaeal phyla known for sulfur reduction, fewer bacteria have been reported to degrade crude oil. Currently, anaerobic alkane degraders have been isolated or enriched from only two phyla: *Proteobacteria* and *Firmicutes* [76], which may limit the range of oil degradation-related genes acquired by phages. Additionally, some AMGs are frequently gained and lost from phage genomes due to variable natural selection pressures [17]. This suggests that while AMGs are ubiquitous in phages, their specific context and function in phage genomes are subject to natural selection. These factors present significant challenges in identifying oil-degradation-related AMGs within phage genomes.

In this study, we aimed to investigate the potential role of viruses in hydrocarbon biodegradation by identifying and characterizing virus-encoded hydrocarbon-degradation genes (vHYDEGs). We performed a systematic search for hydrocarbon-degradation genes in the Integrated Microbial Genomes/Viruses (IMG/VR) database [15] and identified 595 putative vHYDEGs, of which 57 were classified as high quality according to the defined trust index (see Methods). These vHYDEGs could be grouped into 15 protein families involved in the initial key steps of hydrocarbon degradation, such as hydroxylation, carboxylation, and fumarate addition. Phylogenetic analyses revealed a diverse evolutionary history of these proteins while also demonstrating their close relationships with various known/unknown oil degraders. Protein structural analyses showed high similarity of these phage-coding enzymes to experimentally validated protein references, some of them also have identical activity sites that act as the binding site with their own substrates, suggesting that they possess the same metabolic potential as their homologous proteins in bacteria. All identified vHYDEGs and their annotations were integrated into the vHyDeg database to facilitate future research. Overall, our study provides novel insights into the contribution of bacteriophages to bioremediation and highlights the importance of considering the role of bacteriophages in microbial ecology and bioremediation strategies, while also have important implication on the future development of eco-friendly and sustainable bioremediation enzymes and methods.

2. Materials and methods

2.1. vHYDEG identification and classification

To identify viral proteins that participate in the initial activation step of hydrocarbon degradation, we downloaded viral proteins from the Integrated Microbial Genomes/Viruses (IMG/VR v4) [15] and PHROG (v4) [80] database. Only clustered PHROG proteins and proteins from high-confidence genomes in IMG/VR were included. Protein sequences longer than 10,000 amino acids were removed. In total, more than 115 million proteins from ~5.6 million viral genomes and metagenomic viral contigs (mVCs) were obtained. Protein sequences were first annotated using hmmscan [19] to search against the curated hydrocarbon degradation gene database, CANT-HYD [38]. Each HYDEG family was represented by a profile hidden Markov model (hydHMM), which has a "trusted" cutoff (Table S1) to determine if a query sequence can be confidently annotated for the function. Query sequences above the trusted cutoff can be confidently assigned a function. The trusted cutoff of each hydHMM was determined using full-length protein sequences of that gene from non-viral genomes. However, using these cutoffs may underestimate the number of vHYDEGs because viruses often encode truncated versions of metabolic genes that can still perform functions [16]. To overcome this limitation, we defined a relaxed cutoff for each hydHMM using the following strategy. First, we downloaded all profile HMM of viral genes (vHMMs) and their protein sequences from the PHROG database, which contains viral protein families with remote homologous. Second, we annotated all protein sequences of each vHMM using hydHMMs. As protein families in the PHROG database were high-quality viral proteins, we assume they were not homologous of any hydrocarbon-degradation genes and hydHMMs. Therefore, for each targeted hydHMM, the highest bit-score of the PHROG viral protein (named "bitscore_vmax") was set as its relaxed cutoff. To further reduce false positive rates, the relaxed cutoff was set to 150 if the respective bitscore_vmax is smaller than 150. In fact, all relaxed cutoffs were smaller than 150 and the original trusted cutoffs, indicating that there are no full-length hydrocarbon-degradation proteins in the PHROG database. Hits of IMG/VR proteins with a bit-score greater than the relaxed cutoff, e-value$<$ 0.00001, and genome length $>$ 5 kbp were considered vHYDEGs. For consistent genome annotation, viral genomes were annotated using Pharokka [11]. To increase the accuracy of protein domain annotation, all vHYDEGs were further annotated using the Pfam database [57]. To validate the identified vHYDEGs were AMGs, both DRAM-v [72] and VIBRANT [39] were used to annotate viral genomes to detect AMGs. Since DRAM-v and VIBRANT take FASTA format genome sequence as input and output protein headers differently, we utilized the mmseqs easy-rbh module [77] to find the reciprocal-best hit (RBH) among proteins annotated by DRAM-v, VIBRANT, and Pharokka. If a vHYDEG was annotated as an AMG by either DRAM-v or VIBRANT, it was marked as "true" in the "AMG" column of Table S1. Otherwise, it was marked as "unknown". All *pmoA* and *pmoC*, along with four out of six *alkB*, were identified as AMGs by either DRAM-v or VIBRANT. However, *almA*, CYP153 gene, and *ladAα* were not recognized as AMGs by either tool. This discrepancy can be attributed to the fact that both DRAM-v and VIBRANT use a curated list of known AMGs that does not currently include *almA*, CYP153 gene, and *ladAα*. However, a manual inspection of their genomic content suggests these are metabolic genes in viral genomes.

vHYDEGs were further classified into three groups based on the following criteria.

(1) A trust index value (trust_idx) was determined for each vHYDEG by dividing its bit-score by the trusted cutoff of the corresponding hydHMM.

(2) A vHYDEG was classified as high quality (HQ) if the trust_idx $>$ 0.5, and its respective genomes are not prophage and have an estimated contamination ratio of zero. The trust_idx threshold is set as 0.5 because the confirmed phage-encoded particulate methane monooxygenase gene *pmoC* [16] has a trust_idx equal to 0.56 using our annotation method. Please note that the threshold “0.5” is selected based only on *pmoC* gene, this threshold is a rough setting and can be adjusted for the specific protein of interest in the future.

(3) A vHYDEG was classified as medium quality (MQ) if the trust_idx $>$ 0.5, but respective genomes were classified as prophage or had an estimated contamination ratio greater than zero.

(4) All other vHYDEGs were classified as low quality (LQ).

For each vHYDEG family, the one with the highest trust index value in HQ was selected as the representative sequence (HQrep). In total, we got 362 LQ, 176 MQ and 57 HQ vHYDEGs. Six HQreps were used for downstream structural, phylogenetic and genome analyses.

2.2. vHYDEG protein structure prediction and comparison

Protein structures of six HQreps were predicted using ColabFold [55] with default settings, which combines MMseqs2 [77] for fast homology search and AlphaFold2 [35] for accurate structure prediction. Predicted structures were then compared with protein structures in UniProt [81] and RCSB PDB database [7] using Foldseek [85] with default settings. Structural alignments of the Foldseek results were evaluated via TM-score (Template Modeling score) normalized by the length of the reference protein. TM-score is a measure used to assess the structural similarity between two protein structures. It is less sensitive to the local error and more focused on the global topology similarity. The TM-score is a number between 0 and 1, where 1 indicates a perfect match between two protein structures. Typically, a TM-score $>$ 0.5 indicates a model of correct topology and a TM-score $<$ 0.17 means random similarity [99]. Structural comparison results were manually checked to select the top hit that was annotated as relative function. The top hit in PDB database was chosen if its annotation is the same function as the query protein, otherwise, the top hit in UniProt database was chosen and the related AlphaFold-predicted structure was selected from the AlphaFold Protein Structure Database (AlphaFold DB) [86] as target structure. Binding sites and domain features were extracted from UniProt database and visualized on the structural alignment of the query and target structures using PyMOL (https://github.com/schrodinger/pymol-open-source).

2.3. vHYDEG protein phylogenetic tree reconstruction

Phylogenetic analysis was used to investigate the evolutionary origin of the representative vHYDEGs. Protein sequences of each HQreps were searched against the NCBI refseq_protein database [45] using BLASTp with bit-score $>$ 50 and e-value $<$ 0.001 to recruit closely related sequences and to add non-viral context to the phylogenetic analysis. Non-viral hits in the top 100 BLAST hits were clustered by 90 % sequence similarity using MMseqs2 [77] to reduce redundancy. Viral hits were combined with the representative sequences of each cluster and corresponding HQ vHYDEGs, and used to build multiple sequence alignment (MSA) using MUSCLE [20]. Phylogenetic tree was then created using MSA by IQ-TREE [58] with the “LG + G4” model. Visualization of the phylogenetic tree was created using ggtree [94] with the mVC metadata from IMG/VR database and taxonomy annotations.

2.4. mVC genome annotation and taxonomy assignment

Genome sequences of 593 vHYDEG-viruses were downloaded from the IMG/VR database (v4). For consistent annotations, genomes were annotated by Pharokka pipeline [11], which uses PHANOTATE [53] to predict open reading frames, and PHROG database [80] to annotate protein functions. BACPHLIP [30] was used to predict whether viruses were temperate or virulent. BACPHLIP determines the presence or absence of conserved protein domains associated with a temperate lifestyle, attributing a probability of being temperate to each virus. The identified protein sequences were further annotated using InterProScan [34] to get more broad function annotation. Since some viruses use sulfur or nitrogen as electron donors during hydrocarbon degradation, we also annotated proteins using nitrogen metabolism genes [24] from KEGG database [36], and sulfur metabolism genes from a custom HMM database [5,92]. Genome annotations were visualized using pyCircos v0.3.0 (https://github.com/ponnhide/pyCircos).

Taxonomy of mVCs was assigned using two approaches that were implemented in the ViroProfiler pipeline [69]. The first one was the protein-sharing network analysis using vConTACT2 [8], which clustered mVCs with reference viral genomes based on their shared protein clusters. NCBI viral RefSeq (v211) was selected as the reference database. The second one was a protein-voting-based method using MMseqs2 taxonomy module [56], which searched proteins of mVCs against proteins in NCBI viral RefSeq database, add then assign a lowest-common ancestor (LCA) taxonomy based on majority voting of protein taxonomy in that mVC. Conflict taxonomic assignments of these two methods were manually checked to select the LCA taxonomy. Finally, the protein-sharing network and corresponding annotation table were imported into Cytoscape [73] for network visualization.

2.5. Comparative genomic analysis

To compare the genomic context of a vHYDEG with related genomes, we downloaded the genome of the top bacterial hit in BLASTp results. We then compared this genome with corresponding viral genomes using Clinker [26]. Clinker used MMseqs2 to cluster proteins in these genomes, generating genome maps and linking genes belonging to the same protein family through links. The annotation of bacterial genomes was downloaded from NCBI in GenBank format. If no annotation was available, the FASTA file of the genome was downloaded and annotated using the Bakta annotation pipeline [70].

2.6. Database construction

To support future vHYDEG research, we have created the vHyDeg database. This database contains all the identified vHYDEGs in this study, and integrates information on vHYDEG classification, comprehensive functional annotations, and genomic and host information. Functional annotations are linked to corresponding databases through hyperlinks, and genome annotations can be accessed via IMG/VR database links. The vHyDeg database can be accessed at https://deng-lab.github.io/vhydeg.

3. Results

3.1. Viruses encode diverse hydrocarbon degradation genes

We identified 595 vHYDEG proteins from 593 of 5.6 million mVCs in IMG/VR database, representing 15 of 37 hydrocarbon degradation gene families in the CANT-HYD database. Based on the trust index of each hydHMM (see Materials and Methods), 362 vHYDEGs were classified as low quality (LQ), 176 as medium quality (MQ) and 57 as high quality (HQ). Detailed annotation of all vHYDEGs can be found in the vHyDeg database. Distribution of mVCs based on the encoded vHYDEGs and their quality is shown in Fig. 1A. The fifty-seven HQ vHYDEGs include thirty-three *pmoC*, thirteen CYP153 gene, six *alkB*, two *almA*, one *ladA*α, and one *pmoA* (Table. S1). Among these genes, *alkB*, CYP153 gene, *ladA*α, and *almA* encode alkane hydroxylases, initiating aerobic long-chain alkane degradation. *pmoC* and *pmoA* encode particulate methane monooxygenase (pMMO) subunits, aiding methane-to-methanol oxidation. The *pmoC* gene has been previously identified in bacteriophage genomes [16]. The remaining nine MQ- and LQ-vHYDEG protein families include various hydrolases and alkane monooxygenases, targeting middle- and long-chain alkanes and polycyclic aromatic hydrocarbons (PAHs).

The trust index value distribution of vHYDEGs was skewed towards low values (Fig. 1B), with only 17 vHYDEGs (2.85 %) above one, and 78 (13.1 %) above 0.5. This suggests most vHYDEGs have low sequence similarity to known hydrocarbon degradation genes, or their encoded proteins contain domains for other metabolic functions. To identify these domains that were not included in the CANT-HYD proteins, we further annotated all vHYDEG proteins using the Pfam database. The annotation results showed that many of them are involved in hydrocarbon degradation pathways (Fig. 1C). For example, in addition to the most common fatty acid desaturase and cytochrome P450, other annotations such as ring hydroxylating, molybdopterin oxidoreductase, and pyruvate formate-lyase-like domains are all related to either aerobic or anaerobic hydrocarbon degradation [49,66,75]. Besides, among all the viral contigs that encode HQ and MQ vHYDEGs, four of them have complete genome sequences which are determined by direct terminal repeat (DTR) or inverted terminal repeat (ITR), 29 prophages, eight giant virus metagenome-assembled genomes (GVMAG), and 228 linear genomes. The divergent taxonomic origins of these viruses, coupled with the identification of both virulent (92 %) and temperate (8 %) lifestyles, underline the widespread distribution of vHYDEGs. This indicates a broad range of diverse viruses have the potential to encode vHYDEGs, contributing to the process of hydrocarbon degradation (Fig. 1D). To some extent, it also highlights the under-characterization of the global virosphere.

3.2. Viruses-encoded alkane hydroxylases are involved in key steps of alkane degradation

Previous studies have shown that bacteria adapt to challenging conditions like saline crude oil or engage in interspecies collaboration for crude oil breakdown by developing related functional proteins or metabolic systems via horizontal gene transfer [29,43,60]. Although several bacteria were found to contain multiple types or copies of alkane hydroxylases (AH) [59,88], mainly plasmid-transferred genes were reported previously while few phage-contributed HGT were discovered that contribute to long-chain (LC) alkane degradation. Among the HQ vHYDEGs identified in this study, *alkB*, CYP153 gene, *ladA*a and *almA* encode AHs which indicates the potential involvement of viruses in the initial steps of aerobic alkane degradation. The AHs in bacteria that catalyze the first step of alkane biodegradation, from alkane to iso-ethanol, are key enzymes of aerobic degradation of alkanes. AlkB and CYP450 family proteins are two common proteins of AH in bacteria, besides, the flavin-binding LC-alkane hydroxylase (AlmA) and thermophilic soluble LC-alkane hydroxylase (LadA) were also proved to be involved in the hydroxylation of alkanes [74,93].

A complete AlkB-virus UViG_3300035703_002088 has an unusual giant genome of 719 kbp (with ITR detected) was chosen as an example to show the genome assembling and AMGs' distribution. It contains majority genes with unknown functions and abnormal genome assembling patterns, such as the massive genes responsible for nucleotide metabolism located throughout the genome, and dispersed tail fiber genes (Fig. 2A). Giant viruses have been previously reported using alternative codes and reassigning some of the stop codons to be translated as amino acids [2,10]. Interestingly, we observed *alkB* genes frequently located in the beginning or ending part of the genome scaffolds (Fig. 2B, Fig. S1). In addition, almost all the AH-encoded genes in this study were located where switch strands happened (Fig. S2, Fig. S3), showing the mosaicisms of the viral genome that related to the different origins of the genes [18,53]. The sequences of AHs were annotated and compared within each vHYDEG. After screening the full set of phage fragments that encode AlkB, we observed the amino-acid identity of AlkB is between 30–50 % when compared with those of the reference bacteria (Fig. 2B). When comparing the genomic content of AlkB-viruses, it is obvious that AlkB is the only protein specific to all six AlkB-viruses (Fig. 2B). This suggests that these viruses encoded AlkBs were either arising independently or acquired via horizontal gene transfer.

The CYP153-viruses have relatively higher amino acid identities between 50 % and 95 % within the CYP153-virus group and the bacterial reference protein (Fig. 2D). The similarity between different contigs could be observed in scaffold alignment, UViG_3300020463_000001 and UViG_3300020438_000013 were highly similar in the compared fragments of genome, same for UViG_3300002092_000034 and UViG_3300032239_000241 (Fig. 2D). The environmental origins of the former two contigs were marine ecosystems, while the latter two were from freshwater (Fig. 4B). These divergences in amino acid identities implied the possibility that two independent evolutionary events existed when CYP153 gene was acquired by viruses. The complete viral genome of UViG_3300020463_000001 (200 kbp, DTR detected) included several tRNA copies in the genome was shown as an example of CYP153-virus (Fig. 2C). Notably, four other CYP153-viruses also encode multiple tRNAs (Fig. S2). These tRNAs might contribute to avoiding bacterial defense and compensate for the differences in codon usage between phages and their bacteria hosts [6,64]. These tRNA-abundant viruses indicate their potential active life cycles as the translation of the viral genes is under self-regulation.

Although the genome length is relatively short in the case of LadA-phage (~20 kb, Fig. 2E), the amino acid identity of LadAα in phage is 35 % when compared with bacterial reference protein (Fig. 2F). In addition, this contig encoded another AMG that was annotated by PHROG database as acyl-CoA N-acyltransferase, which is important for lipid metabolism [23]. The HQ *ladA*α originated from a bioreactor wastewater sample (Table. S1). Despite the rare incidence, *ladA*α in phage was with high trust index value and when comparing the phage genome with a bacteria reference, LadAα is the only protein shared between the phage and reference (Fig. 2F). LadA homologues were previously identified in *Geobacillus thermodenitrificans* NG80–2 as a single copy on the plasmid pLW1071 [22], suggesting that it is an alien gene. Later in *Geobacillus thermoleovorans* B23 chromosome, a "*ladAB* gene island" was observed that consisted of three *ladA* homologues [9], and all these three homologues are proved to be functional.

Two contigs with similar genomes were screened from the database as AlmA-viruses (Fig. 2H), both originated from river water and with DTR detected. They were identified as complete genomes, with a size of 364,247 bp (Table. S1). Notably, like the large CYP153-viruses, AlmA-viruses encoded multiple copies of tRNA as well as other AMGs (Fig. 2G). In summary, for all high-quality vHYDEG-viruses in this study, multiple copies of AMGs could be observed in the genome, and high-quality AHs are of great interest (Fig. S1, Fig. S2).

3.3. vHYDEGs have conserved protein structures and active catalytic sites

Phage proteins usually have low amino acid identity when compared with homologies from their host, thus the sequence-based comparison usually fails to detect remote homologous genes. Protein structures of

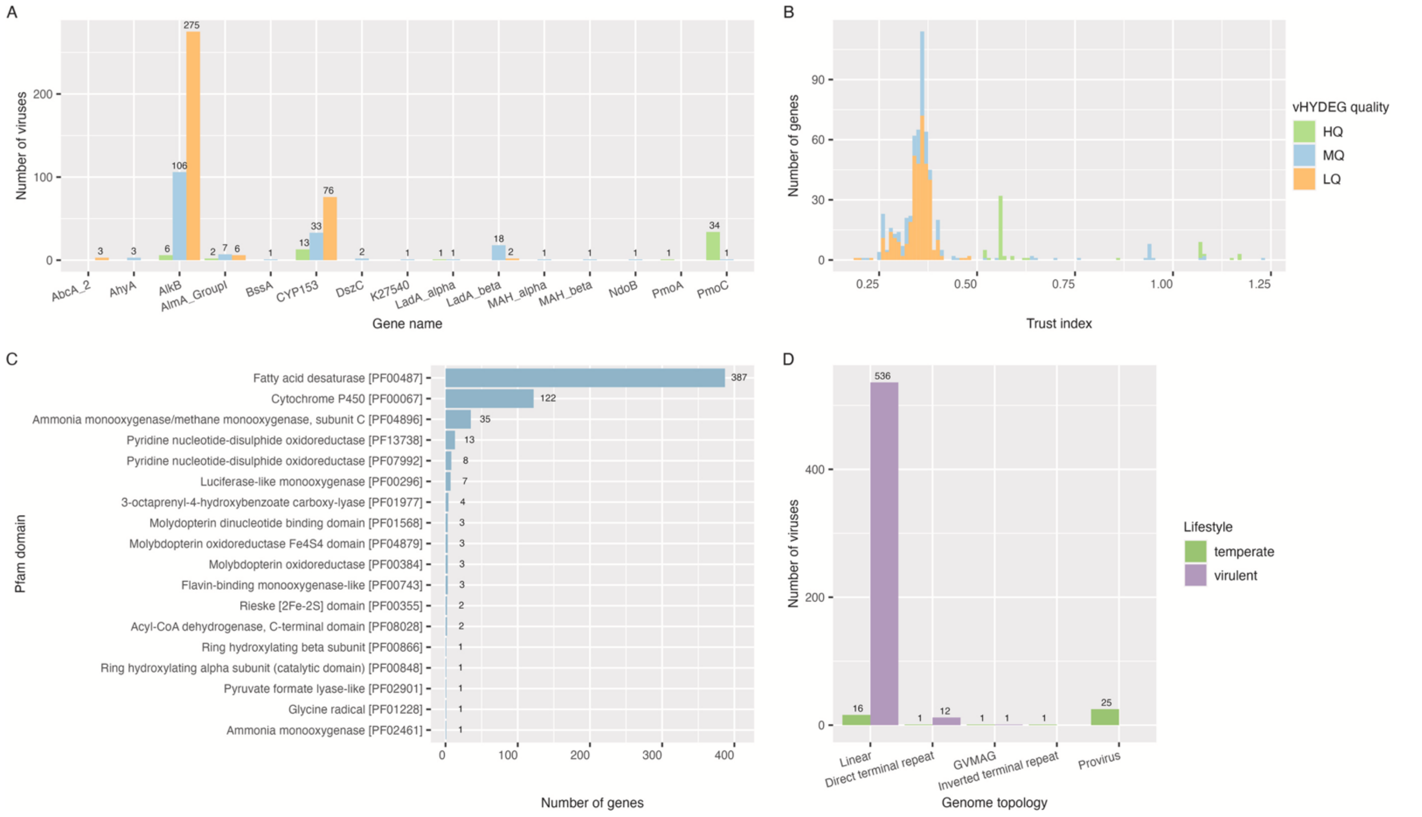


Fig. 1. The overview of identified vHYDEGs from IMG/VR database. (A). Distribution of mVCs based on the encoded vHYDEGs and their quality. (B) Distribution of vHYDEGs based on quality and trust index. (C). Pfam annotation results for all vHYDEG proteins. (D). Distribution of the number of mVCs based on their lifestyle and genome topology.

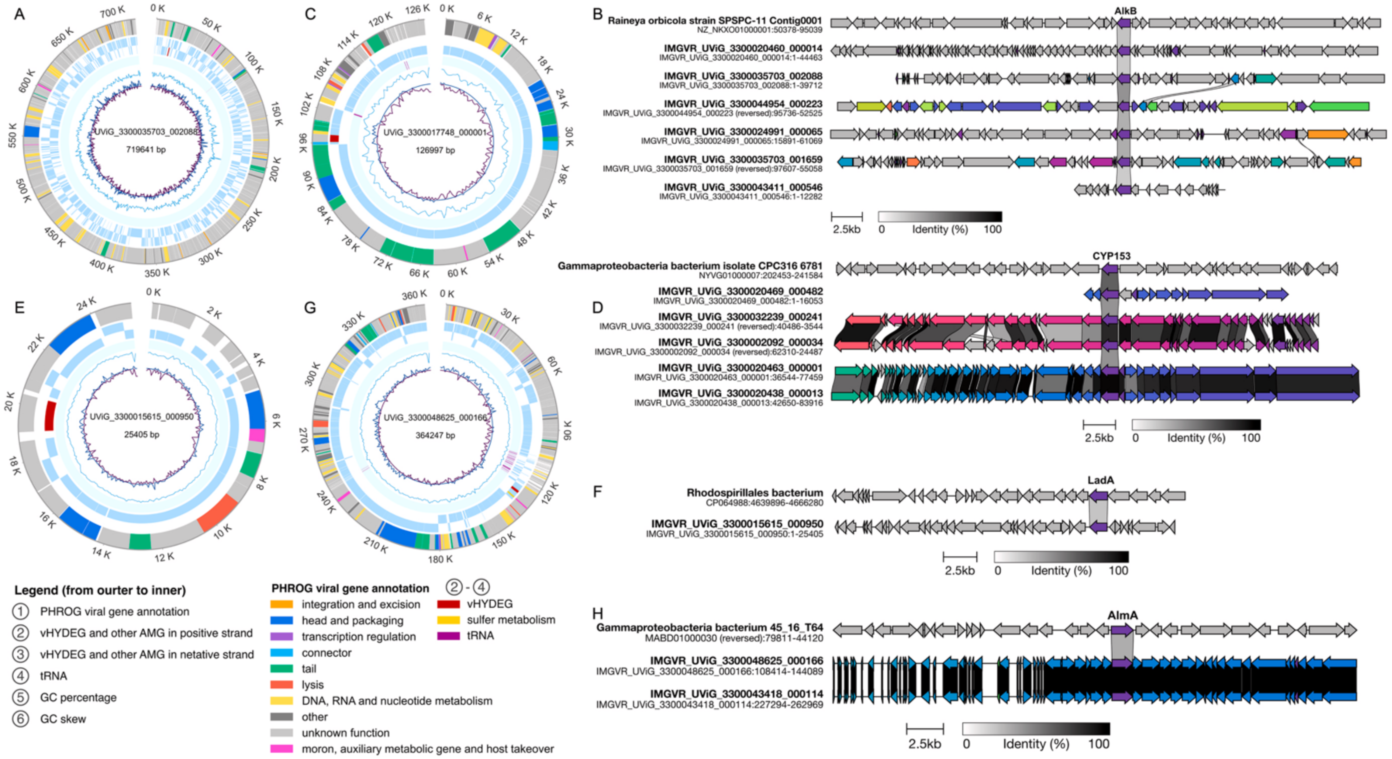


Fig. 2. Scaffold alignments and representative genome of the four alkane hydroxylase-encoding mVCs. (A). a complete genome of UViG_3300035703_002088 that encoding *alkB*. (C). a complete genome of UViG_3300020463_000001 that encoding CYP153 gene. (E). a non-complete genome of *ladAα*-encoding mVC. (G). A complete genome of UViG_3300048625_000166 that encoding *almA*. (B), (D), (F), (H). genome alignment of high-quality AlkB/CYP153/LadA/AlmA-encoding mVCs with a reference bacteria genome. The reference genomes are on the top of each group of alignments.

homologous genes, however, are usually more conserved than their sequences. To validate that vHYDEGs have the same metabolic potential as their host's homologues, we used a structural-based comparison approach. Comparing the 3D structures of vHYDEG proteins ensures the validation of viruses-encoded proteins have the necessary domain for enzymatic reactions. The state-of-art computational methods were applied to predict protein structures of vHYDEGs, and comparison with those experimentally determined reference structures from either RCSB PDB database or Alphafold2 predicted structures from UniProt database were performed, to further confirm the most similar structures. Binding sites were highlighted on the 3D structures if they can be obtained from the UniProt feature annotations. Results showed that phage-encoded vHYDEGs have high structural similarity with those reference structures (Fig. 3). Binding sites regions of AlkB, AlmA and LadAα in our study overlapped with those in the reference proteins (Fig. 3A, B, C). While CYP153, PmoA and PmoC have no binding sites annotation in the reference database, their TM-scores are from 0.83 to 0.94, suggesting they have the same structures as their reference proteins (Fig. 3D, E, F). These observations indicate the functional domains were maintained by vHYDEGs and might be active when infecting the host. While AlkB, AlmA, CYP153 and LadAα are full-length proteins, PmoA and PmoC are only subunits of the pMMO. This finding suggests that phages sometimes encode part of the gene with essential functions from their host, which is in line with previous studies that showed phages only encode core domains of large enzymes [16]. Although structural comparisons suggest vHYDEGs maintain conserved structures of their reference proteins, further experimental evidence is needed to validate their hydrocarbon-degradation potential.

3.4. Diversity and evolution of the vHYDEGs

Phylogenetic analysis of these vHYDEGs showed they are related to diverse bacterial species (Fig. 4), but the majority of which belong to phylum *Proteobacteria*. For the AlkB-viruses, the amino acid sequences from the mVCs formed four relatively distant phylogenetic groups, closely related to *Bacteroidota*, *Proteobacteria*, *Actinobacteria*, and *Candidatus Blackallbacteria*, respectively (Fig. 4A). Many bacterial species reported as oil degraders belong to the first two phyla [50,95], while the recently discovered uncultivated phyla *Candidatus Blackallbacteria*, which were identified from CO_2-derived geyser, have not yet been reported to be associated with long-chain alkane degradation [65].

Phage-encoded CYP153 were grouped into two separate clusters in the phylogenetic tree (Fig. 4B), which resemble their genome scaffold alignment (Fig. 2D). The two clusters were closely related to *Actinobacteria* and *Proteobacteria*, respectively. Most of these related bacteria were able to survive in oil-contaminated habitats, such as *Pseudomonas* and *Rhodococcus*, which have been observed in Arabian gulf sediment [3]. Additionally, *Acinetobacter baumannii* from the Acinetobacter calcoaceticus-Acinetobacter baumannii complex (ACB complex) has been reported to co-metabolite crude oil with *Talaromyces sp* [98]. This evolutionary closeness revealed the possibilities of HGT between these oil degradation bacteria and their phages, although only in rare cases. The mVCs with a complete protein length of CYP153, however, are more abundant compared with other three AHs (Fig. 4B). The potential explanation might be both *alkB* and *ladA* are discovered to be located in the plasmid of bacteria [4,75], thus the horizontal transfer might more frequently happen via plasmid.

LadA and AlmA are similar in that they are flavin-dependent monooxygenases belonging to the family of bacterial luciferases. However, the structure of LadA is unique in that it has three functional domains that are segregated. One domain acts as a monooxygenase, while the other two are for NADPH oxidation. This structure enables LadA to hydroxylate alkanes without the need for rubredoxin and rubredoxin reductase, which are required for other AHs enzyme systems [44]. In the phylogenetic tree of LadAα (Fig. 4C), only one of the LadAα homologues in phage was shown which was closely related to *Rhodospirillales* and other *Proteobacteria*. It is interesting as in the previous studies on bacteria, *LadAs* were mainly found in *Geobacillus* and *Aeribacillus* [9,83], which belong to phylum *Firmicutes*. LadA homologues have also been discovered in fungi [63], and the metabolic gene clusters in fungi were hypothesized to be a result of HGT as well [67]. Previous work also showed that the KilA-N domain, which was first identified in bacteriophage P1, was later found as endogenized viral genes in the genome of bacteria and fungi. All these evidence highlighting the HGT as a mechanism of genetic innovation in eukaryotes as well [27,54].

Two AlmA-phages are phylogenetically clustered together and the amino acid is 42 % identical with bacterial reference (Fig. 2H). The two most popular genera that encode *almA* gene are *Alcanivorax* and *Marinobacteria* [90], and the latter has a close phylogenetic relationship with the AlmA-phages in our results (Fig. 4D).

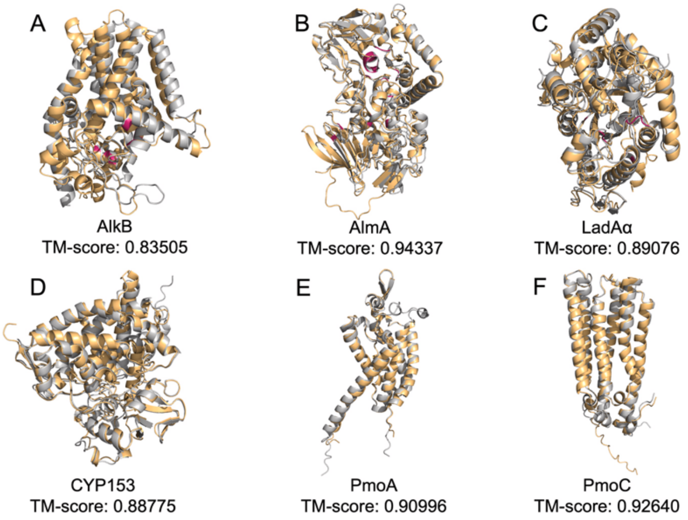


Fig. 3. The predicted protein structures of four AHs and two subunits of PMMO. The experimentally validated proteins are labeled gold as reference, while the virus encoded enzymes in this study are labeled silver, activity sites were labeled pink. (A). structure comparison and active catalytic site of AlkB. (B). structure comparison and active catalytic site of AlmA. (C). structure comparison and active catalytic site of LadAα. (D). structure comparison of CYP153. (E). structure comparison of PmoA. (F). structure comparison of PmoC.

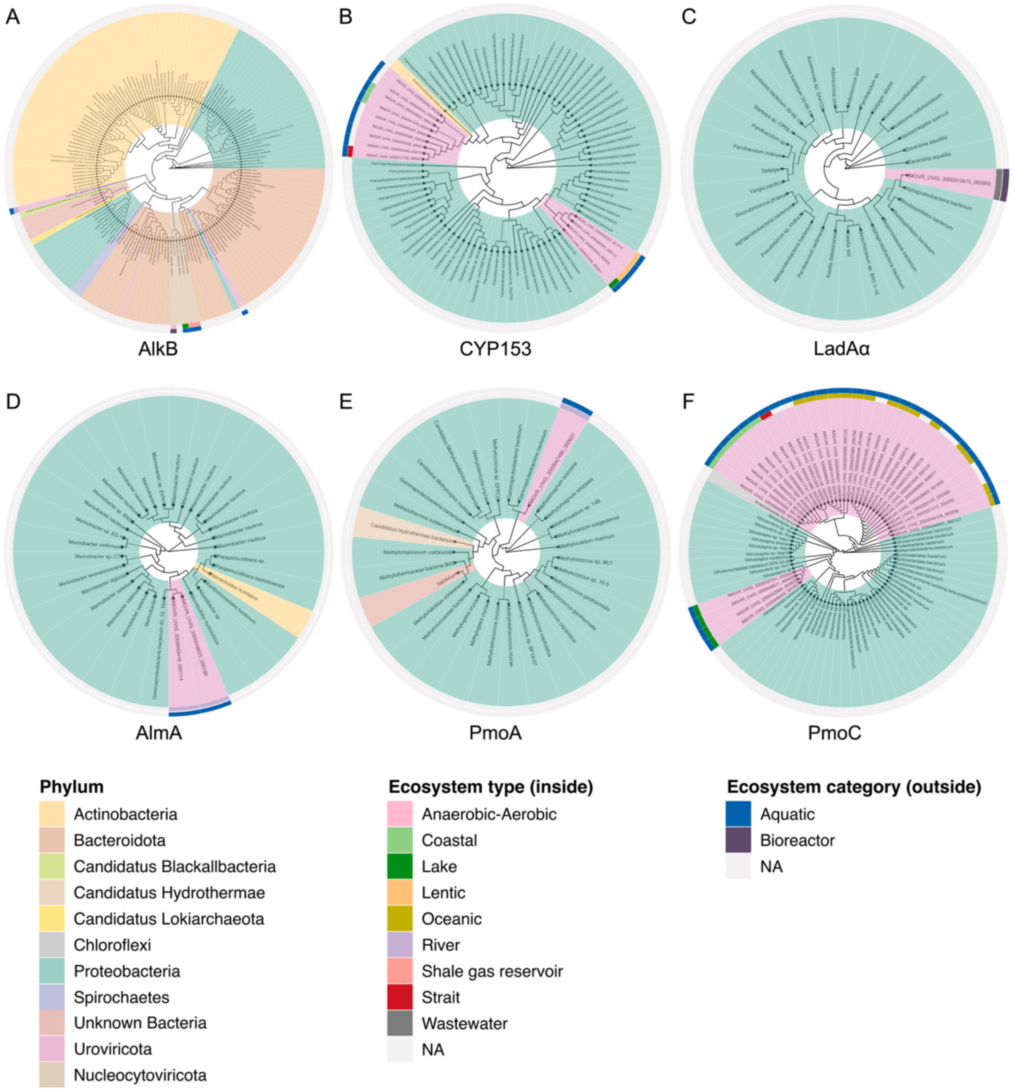


Fig. 4. Phylogenetic trees of high-quality vHYDEG proteins. The ecological distributions of the proteins are labeled with different colors based on the ecosystem type (inner circle) and category (outer circle), and blank in each ring are for microbial references from RefSeq database. The background color indicating the phylum of mVCs or references, only the proteins from HQ vHYDEGs were included in the trees. (A), (B), (C), (D), (E), (F), The phylogenetic tree of AlkB, CYP153, LadAα, AlmA, PmoA and PmoC, respectively.

Both PmoC and PmoA are subunits of pMMO, which is involved in methane oxidation [16]. We observed that PmoA is phylogenetically related to methanotrophs such as *Methylomagum*, *Methylocadum* and *Methylococcus* (Fig. 4E), indicating the potential co-evolution of phage and their host bacteria. However, the PmoC is closely related to nitrogen-cycling bacteria in *Proteobacteria* (Fig. 4F). This result resembles the phage-encoded *amoC* genes from marine samples [24], and we also observed that the protein structure of AmoC in that study is quite similar to our PmoC (Fig. 3F), this is in line with a former study that the pMMO and AMO are evolutionarily related despite their different physiological role [31]. These interesting results showed the potential multiple functions of *pmoC/amoC* genes in phages, which suggests that phages can potentially provide new metabolic capabilities to a microbial community.

3.5. vHYDEG-viruses are taxonomically diverse

vHYDEG-viruses were annotated and clustered with NCBI reference viral genomes based on shared proteins. Similar viruses are connected and grouped closer within the network (Fig. 5). The protein-sharing

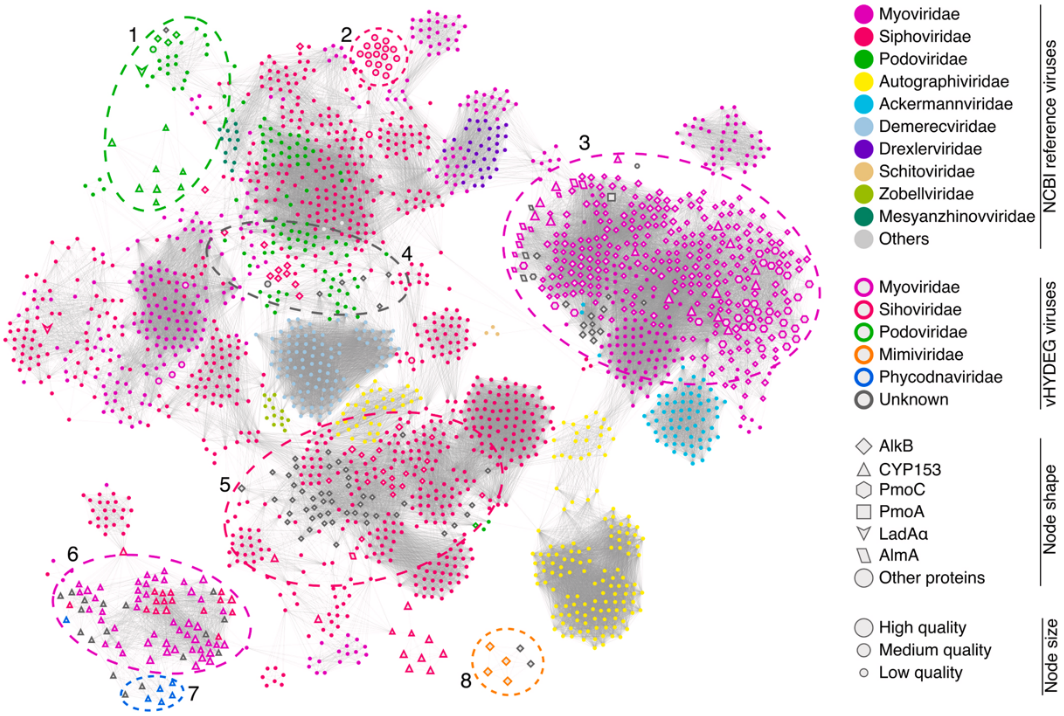


Fig. 5. Taxonomy and protein-sharing network clustering with reference viruses. Only mVCs with their first and second neighbor (neighbor of neighbor) reference viruses in the network are shown. Each dot represents a mVC (dots with outlines) from IMG/VR database or a reference virus (dots without outlines) from NCBI viral RefSeq. The shape of the dots indicates the type of vHYDEGs encoded in the genome, and the size of the dots varies based on the vHYDEG quality. Similar viral genomes are connected by lines based on their shared proteins. Genomes that are more similar are connected with closer proximity. Clusters containing more than five vHYDEG-viruses are marked using dashed circles with numbers. Cluster four is a pseudo-cluster that represents the overlapping region of multiple viral families.

network shows that vHYDEG-viruses are mainly clustered with the family *Podoviridae* (cluster 1), *Siphoviridae* (clusters 2 and 5), and *Myoviridae* (cluster 3). Cluster 6 contains both *Myoviridae* and *Siphoviridae*. Cluster 4 is an overlapping region of multiple families, including *Myoviridae*, *Siphoviridae*, and *Podoviridae*, and is not a true cluster. All these families are bacteriophages that belong to the order *Caudovirales*. Cluster 7 and cluster 8 were classified as *Phycodnaviridae* and *Mimiviridae*, respectively, and both belong to the phylum *Nucleocytoviricota*, also known as nucleocytoplasmic large DNA viruses (NCLDV). NCLDV are a group of viruses that infect a wide range of eukaryotic hosts, including animals, plants, and protists. These viruses are characterized by their large genomes and complex replication cycles [42]. Cluster 7 is composed of five *Phycodnaviridae* and two unknow ones, all of them encode CYP153, while cluster 8 contains four *Mimiviridae* and two unknown viruses, all of which encode AlkB. vHYDEG-viruses with no taxonomic annotation (unknown viruses) were prevalent in all above-mentioned clusters, except cluster 2, indicating the novelty and taxonomic diversity of vHYDEG-viruses. Although they have no taxonomy annotation from either the protein-sharing or protein-voting-based taxonomy assignment approach, clustering results suggest that they are sharing high similarity to *Myoviridae, Siphoviridae,* and *Podoviridae*, as well as NCLDV, respectively.

As shown in Fig. 5, vHYDEG proteins are not constrained by their taxonomy, for example, *Podoviridae* in cluster 1 encodes four different vHYDEGs, and same in cluster 5 of *Siphoviridae*. Especially in cluster 3 of *Myoviridae*, viruses in this family could encode almost all AHs except for LadAα, as well as PmoC and PmoA. Based on these findings, we hypothesize that functions of vHYDEGs were more associated with specific hydrocarbon degradation pathways than viral taxonomy. This means that different viruses might obtain different HYDEGs independently, or from horizontal gene transfer (HGT). This observation also coincides with the phylogenetic analysis, that the mVCs with same vHYDEG have diverse and independent evolutionary origins of these genes, and they are closely related with bacteria from distinct phylum (Fig. 4).

3.6. vHYDEG-viruses inhabit in multiple ecological niches

To study the environmental sources of vHYDEG-viruses, we obtained environmental source metadata of mVCs from IMG/VR database. Our findings revealed that these viruses could be found in a variety of ecological categories, including aquatic, terrestrial, bioreactor, plant, and host-associated samples. While most of these viruses were discovered in freshwater and marine samples, there were a few cases found in extreme environments, such as shale gas reservoirs, saline, and thiocyanate environments (Fig. 6). For example, two viruses that encode HQ *alkB* were isolated from shale gas reservoirs (Fig. 4A), and one encodes *alkB* from an anaerobic-aerobic thiocyanate-remediating bioreactor. The only *ladA*α was from a wastewater bioreactor. All other viruses that encoding HQ CYP153 gene, *almA*, *pmoA* and *pmoC* were from aquatic samples. One of the possible explanations for the rare cases identified from extreme environments could be attributed to the fact that there are fewer public datasets available from these extreme environments as

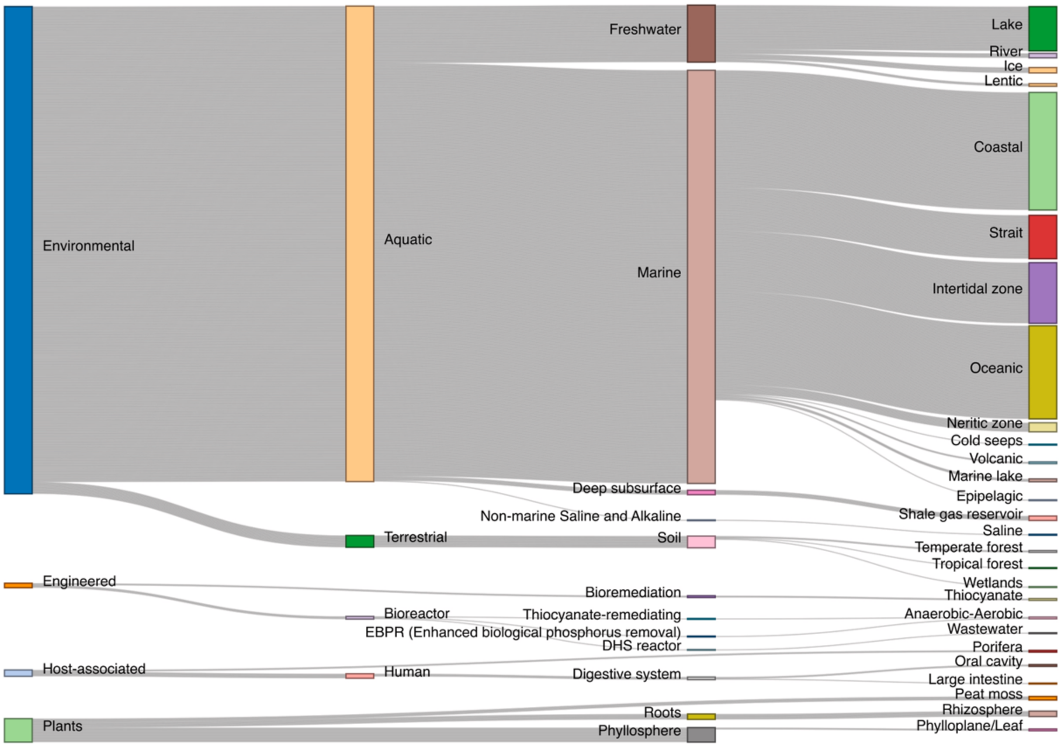


Fig. 6. Ecological classification and origins of all 593 mVCs. Ecosystem data was obtained from the IMG/VR database and classified into four levels. The majority of viruses originated from marine environments. However, due to incomplete metadata in the original data source, some viruses lack detailed ecosystem annotations, leading to discrepancies in the number of inflows and outflows. For instance, approximately one-fourth of marine viruses and most soil viruses lack annotations at the fourth level of classification.

compared to freshwater and marine samples. This indicates that the current database provides limited view of the actual diversity of viruses that encode vHYDEGs. We anticipate that more viral metagenomics samples from diverse environments be sequenced in the future would help to expand current knowledge on vHYDEGs.

4. Discussion

Viruses encoded a vast range of genomic content that can profoundly influence other organisms [18]. In addition to viral structural genes, recent studies have shown that viruses can also encode various metabolic genes that may affect their host's metabolism [40,82]. As a result, viruses play key roles in microbial evolution, marine nutrient cycling, and human disease. We are just at the beginning of documenting the diversity and host range of aquatic viruses, as well as their potential impact on microbial communities and marine biochemistry. Since viruses can act as vectors of horizontal gene transfer, these genes can be transferred to other organisms, influencing their metabolism, community, and ultimately, the entire biological ecosystem. Therefore, studying the function of viral-encoded genes, particularly their metabolic functions, is crucial not only for the environment but also for potential biotechnological innovations.

In this study, we identified plenty of vHYDEGs, which can participate in the initiation of degradation of various hydrocarbons, including methane, long-chain alkanes, and aromatic hydrocarbons. The genomic contents, evolutionary relationships, and protein structural analyses validated their putative hydrocarbon-degradation capability. Our findings are consistent with the previous hypothesis that phage-encoded AMGs are largely involved in the key-step of host metabolism pathways. An explanation for this is that vHYDEGs participate in the initiation of hydrocarbon activation and may provide more evolution advantage for phages than those encode enzymes involved in the middle or end of the degradation pathway, as AMGs that contribute to critical, rate-limiting steps were thought to help phages to boost the host metabolism during infection, which in turn benefit to phage propagation [14]. The high proportion of virulent viruses aligns with previous research, which suggested that AMGs primarily offer advantages to phages during short-term, active lytic infections [40]. For temperate phages, they integrate their genomes into the host chromosome during the lysogenic cycle, lying dormant for extended periods. During this dormant stage, they can still express AMGs, which may enhance the host bacterium's ability to adapt to its environment and increase its chances of survival. In our study, HQ vHYDEGs encode six protein families: AlkB, CYP153, AlmA, LadAα, PmoA and PmoC. The former four protiens are alkane hydroxylases that initiate the first step of aerobic long-chain alkane degradation, and the rest two are subunits of PMMO, a key enzyme of methane degradation. Previous study identified nine glycoside hydrolase families encoded by viruses from permafrost samples, which had capacities for pectin, hemicellulose, starch, and cellulose cleavage [21]. In both studies, carbon degradation genes in viruses are responsible for key-steps of substrate hydroxylation, the same is for the sulfur dissimilatory metabolic genes and the photosynthesis genes that have been found in phages [12,40]. All these AMGs, despite their varied functions and virus habitats, provide fitness benefits to their host bacteria, allowing them to better adapt to changing environmental conditions. Some bacteria, such as *Alcanivorax* isolates, have multiple copies of the CYP153 gene and alkB in their genome [48]. Other bacteria, including *Dietzia*, *Pseudomonas aeruginosa* PAO1, and *Rhodococcus erythropolis*, also have multiple AHs [46,52,84,91], which are thought to provide the bacterium with a wider substrate range and better environmental adaptation. It is reasonable to hypothesize that vHYDEGs could work similarly by providing an extra copy of AH-coding genes.

The *pmoC* gene in phages have already been discovered before [16], and thus not extensively discussed here. However, in our study, the identified *pmoC*-phage is closely related with a *Nitrosomonadaceae bacterium* (Fig. S4), which is different from the large fresh water *pmoC* that has high similarity with methanotrophs. The possible explanation might be the multi-function of and evolutionary homologies of pMMO and AMO, or due to the wide host range of the *pmoC*-encoding viruses. Another difference between our study and previous study is, our pmoC-viruses could be clustered into two big clusters. The larger cluster are mVCs from marine ecosystem instead of freshwater (Fig. 4F), indicating that marine ecosystem contains abundant methane degradation related AMGs that were not discovered before. Besides, we identified another subunit PmoA here in viral genomes. As discovered in the large freshwater phages that PmoC is the substrate binding domain of the enzyme, PmoA is also considered essential in methane oxidation and has been regarded as a functional marker of anaerobic methanotrophs [79]. Similar to PmoC, the PmoA is also 28–30 kD, both of them are almost entirely embedded in the membrane and comprise seven and five transmembrane helices [41]. It's the first identification of *pmoA* in the phage genome. Considering that phages encode partial genes with vital functional sites and recombined genes, we hypothesize that vHYDEGs may be involved in hydrocarbon degradation in ways divergent from those in bacteria. However, further research is needed to validate this hypothesis.

Our study reveals frequent independent evolutionary origins of vHYDEGs, such as AlkB, CYP153, and PmoC, across various geographic locations (Fig. 4). These vHYDEGs are acquired and retained in distinct ecological niches (e.g., CYP153 in marine and freshwater environments), highlighting their presumed importance in evolution and retention within diverse phage taxonomies (Fig. 5) across different niches (Fig. 6A). Given the potential costs of maintaining extra genetic material for small phage genomes, only essential AMGs are likely conserved evolutionarily [13]. This suggests that vHYDEGs might provide significant advantages to their hosts.

Protein function annotation can be challenging due to inconsistencies in accuracy and sensitivity across databases and methods, particularly for viral-encoded proteins. To address this issue, we first used the CANT-HYD database [38] with a relaxed threshold to enhance detection sensitivity and lower computational costs. Next, we used multiple databases to obtain a thorough functional annotation for detected proteins. Annotation results were classified into different quality levels based on bit-score value and mVC quality. A trust index was created for consistent annotation comparison. This approach significantly improved detection sensitivity and reduced false positive rates. Furthermore, we predicted protein structures of high-quality vHYDEGs to assess their metabolic potential. Despite divergent amino acid sequences, we observed high structural similarities between viral-encoded enzymes and bacterial reference proteins for six key enzymes (Fig. 3). These structural differences in vHYDEGs provide insights into enzyme catalytic mechanisms and valuable information for de novo engineering of more effective enzymes through cutting-edge machine learning technologies [96].

Previous studies have found a wealth of AMGs involved in carbon metabolism, including energy production via the pentose phosphate pathway (PPP) and tricarboxylic acid (TCA) cycle [32,33,68], as well as glycoside hydrolase families capable of degrading pectin, hemicellulose, and starch [21]. The discovery of high-quality vHYDEGs in phages now adds one previously missing piece to the whole map of carbon degradation, suggesting that phages can potentially provide new degradation capabilities to microbial communities.

To facilitate further investigation on the topic, we created a database containing all identified vHYDEGs with their comprehensive annotations, as well as links to external resources such as IMG/VR and PHROG database. Researchers can screen their interested vHYDEGs in this database, and get an overview of genomic content, as well as information abount the viruses' sources and their host. Most vHYDEG-viruses were not similar to those in the NCBI viral RefSeq database in terms of protein similarity, indicating that vHYDEGs-viruses are quite novel regarding their unknown taxonomy. Notably, most viruses in the NCBI viral RefSeq database were isolated from known host, this finding indicates there are more diverse viruses in the environment that have yet to be identified, and their role and contribution to the entire ecosystem are yet to be understood.

5. Conclusions

In this study, we analyzed approximately 115 million viral proteins from 5.6 million mVCs in the IMG/VR databases. Although less than 0.07 % (593) mVCs encode at least one vHYDEG, they are encoded by taxonomically divergent phages including both major lifestyles of virulent and temperate, and inhabit multiple ecological niches. We chose 57 mVCs that encoding high quality vHYDEGs, distinguishing them from prophages, and providing genome structures and complete annotation of the full set of genes. These 57 mVCs were proven to encode key step enzymes of hydrocarbon degradation, as well as other important genes that are involved in e.g., protein translation, sulfur and lipid metabolism. To further validate the metabolic potential of the observed vHYDEG proteins, their 3D structures were predicted, and catalytic sites were evaluated, showing a high level of structural identity when compared with experimental validated homologous.

The habitat might contribute to the independent evolutionary events of vHYDEGs. Phages acquired these genes originating from different environments and ecosystems, primarily aquatic settings. With diverse taxonomy, vHYDEGs-mVCs are evolutionarily related to a collective of bacteria from varied phylum, and clustered into distinct clades. Together with the facts that numerous unexplored environmental samples that are out of our current study, it is reasonable to hypothesize that there are novel enzymes that participate in the hydrocarbon degradation pathway, however, have not been recorded in the public protein database. Finally, the vHyDeg database provides a comprehensive repository of vHYDEGs, which could serve as a valuable resource for the development of novel enzymes, thereby contributing to bioremediation research targeting hydrocarbon and crude oil contamination.

CRediT authorship contribution statement

Jinlong Ru: Methodology, Software, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Jinling Xue:** Conceptualization, Investigation, Data curation, Writing - original draft, Writing - review & editing. **Jianfeng Sun:** Visualization. **Linda Cova:** Visualization. **Li Deng:** Conceptualization, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

All data and custom analysis scripts used in this study are available at https://github.com/deng-lab/vhydeg. The vHyDeg database is available at https://deng-lab.github.io/vhydeg.

Acknowledgments

This work was funded by the German Research Foundation (DFG DE2360/2-1 and Project No. 391644373). JR was supported by the China Scholarship Council (CSC).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jhazmat.2023.132299.

References

[1] Afzal, M., Rehman, K., Shabir, G., Tahseen, R., Ijaz, A., Hashmat, A.J., et al., 2019. Large-scale remediation of oil-contaminated water using floating treatment wetlands. npj Clean Water 2, 3. https://doi.org/10.1038/s41545-018-0025-7.

[2] Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., et al., 2020. Clades of huge phages from across Earth's ecosystems. Nature 578, 425–431. https://doi.org/10.1038/s41586-020-2007-4.

[3] Al-Thukair, A.A., Malik, K., Nzila, A., 2020. Biodegradation of selected hydrocarbons by novel bacterial strains isolated from contaminated Arabian Gulf sediment. Sci Rep 10, 21846. https://doi.org/10.1038/s41598-020-78733-0.

[4] Alonso-Gutiérrez, J., Teramoto, M., Yamazoe, A., Harayama, S., Figueras, A., Novoa, B., 2011. Alkane-degrading properties of Dietzia sp. H0B, a key player in the Prestige oil spill biodegradation (NW Spain). J Appl Microbiol 111, 800–810. https://doi.org/10.1111/j.1365-2672.2011.05104.x.

[5] Anantharaman, K., Hausmann, B., Jungbluth, S.P., Kantor, R.S., Lavy, A., Warren, L.A., et al., 2018. Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. ISME J 12, 1715–1728. https://doi.org/10.1038/s41396-018-0078-0.

[6] Bailly-Bechet, M., Vergassola, M., Rocha, E., 2007. Causes for the intriguing presence of tRNAs in phages. Genome Res 17, 1486–1495. https://doi.org/10.1101/gr.6649807.

[7] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., et al., 2000. The protein data bank. Nucleic Acids Res 28, 235–242. https://doi.org/10.1093/nar/28.1.235.

[8] Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., et al., 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol 37, 632–639. https://doi.org/10.1038/s41587-019-0100-8.

[9] Boonmak, C., Takahashi, Y., Morikawa, M., 2014. Cloning and expression of three ladA-type alkane monooxygenase genes from an extremely thermophilic alkane-degrading bacterium Geobacillus thermoleovorans B23. Extrem: life Extrem Cond 18, 515–523. https://doi.org/10.1007/s00792-014-0636-y.

[10] Borges, A.L., Lou, Y.C., Sachdeva, R., Al-Shayeb, B., Penev, P.I., Jaffe, A.L., et al., 2022. Widespread stop-codon recoding in bacteriophages may regulate translation of lytic genes. Nat Microbiol 1–10. https://doi.org/10.1038/s41564-022-01128-6.

[11] Bouras, G., Nepal, R., Houtak, G., Psaltis, A.J., Wormald, P.-J., Vreugde, S., 2022. Pharokka: a fast scalable bacteriophage annotation tool. Bioinformatics btac776. https://doi.org/10.1093/bioinformatics/btac776.

[12] Bragg, J.G., Chisholm, S.W., 2008. Modeling the fitness consequences of a cyanophage-encoded photosynthesis gene. PLoS One 3, e3550. https://doi.org/10.1371/journal.pone.0003550.

[13] Breitbart, M., Bonnain, C., Malki, K., Sawaya, N.A., 2018. Phage puppet masters of the marine microbial realm. Nat Microbiol 3, 754–766. https://doi.org/10.1038/s41564-018-0166-y.

[14] Breitbart, M., Thompson, L.R., Suttle, C.A., Sullivan, M.B., 2007. Exploring the vast diversity of marine viruses. Oceanography 20, 135–139. https://doi.org/10.5670/oceanog.2007.58.

[15] Camargo, A.P., Nayfach, S., Chen, I.-MinA., Palaniappan, K., Ratner, A., Chu, K., et al., 2022. IMG/VR v4: An expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res gkac1037. https://doi.org/10.1093/nar/gkac1037.

[16] Chen, L.-X., Méheust, R., Crits-Christoph, A., McMahon, K.D., Nelson, T.C., Slater, G.F., et al., 2020. Large freshwater phages with the potential to augment aerobic methane oxidation. Nat Microbiol 1–12. https://doi.org/10.1038/s41564-020-0779-9.

[17] Crummett, L.T., Puxty, R.J., Weihe, C., Marston, M.F., Martiny, J.B.H., 2016. The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. Virology 499, 219–229. https://doi.org/10.1016/j.virol.2016.09.016.

[18] Dion, M.B., Oechslin, F., Moineau, S., 2020. Phage diversity, genomics and phylogeny. Nat Rev Microbiol 18, 125–138. https://doi.org/10.1038/s41579-019-0311-5.

[19] Eddy, S.R., 1998. Profile hidden Markov models. Bioinforma (Oxf, Engl) 14, 755–763. https://doi.org/10.1093/bioinformatics/14.9.755.

[20] Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792–1797. https://doi.org/10.1093/nar/gkh340.

[21] Emerson, J.B., Roux, S., Brum, J.R., Bolduc, B., Woodcroft, B.J., Jang, H.B., et al., 2018. Host-linked soil viral ecology along a permafrost thaw gradient. Nat Microbiol 3, 870–880. https://doi.org/10.1038/s41564-018-0190-y.

[22] Feng, L., Wang, W., Cheng, J., Ren, Y., Zhao, G., Gao, C., et al., 2007. Genome and proteome of long-chain alkane degrading Geobacillus thermodenitrificans NG80-2 isolated from a deep-subsurface oil reservoir. Proc Natl Acad Sci USA 104, 5602–5607. https://doi.org/10.1073/pnas.0609650104.

[23] Fu, W., Shen, Y., Hao, J., Wu, J., Ke, L., Wu, C., et al., 2015. Acyl-CoA N-acyltransferase influences fertility by regulating lipid metabolism and jasmonic acid biogenesis in cotton. Sci Rep 5, 11790. https://doi.org/10.1038/srep11790.

[24] Gazitúa, M.C., Vik, D.R., Roux, S., Gregory, A.C., Bolduc, B., Widner, B., et al., 2020. Potential virus-mediated nitrogen cycling in oxygen-depleted oceanic waters. ISME J 1–18. https://doi.org/10.1038/s41396-020-00825-6.

[25] Ghorbannezhad, H., Moghimi, H., Dastgheib, S.M.M., 2022. Biodegradation of high molecular weight hydrocarbons under saline condition by halotolerant Bacillus subtilis and its mixed cultures with Pseudomonas species. Sci Rep 12, 13227. https://doi.org/10.1038/s41598-022-17001-9.

[26] Gilchrist, C.L.M., Chooi, Y.-H., 2021. Clinker & Clustermap.Js: Automatic generation of gene cluster comparison figures. Bioinformatics 37, 2473–2475. https://doi.org/10.1093/bioinformatics/btab007.

[27] Gladyshev, E.A., Meselson, M., Arkhipova, I.R., 2008. Massive horizontal gene transfer in Bdelloid Rotifers. Sciences 320, 1210–1213. https://doi.org/10.1126/science.1156407.

[28] Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., et al., 2019. Marine DNA viral macro- and microdiversity from pole to pole. Cell 177, 1109–1123.e14. https://doi.org/10.1016/j.cell.2019.03.040.

[29] Heyerhoff, B., Engelen, B., Bunse, C., 2022. Auxiliary metabolic gene functions in pelagic and benthic viruses of the Baltic Sea. Front Microbiol 13. https://doi.org/10.3389/fmicb.2022.863620.

[30] Hockenberry, A.J., Wilke, C.O., 2021. BACPHLIP: Predicting bacteriophage lifestyle from conserved protein domains. PeerJ 9, e11396. https://doi.org/10.7717/peerj.11396.

[31] Holmes, A.J., Costello, A., Lidstrom, M.E., Murrell, J.C., 1995. Evidence that participate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. FEMS Microbiol Lett 132, 203–208. https://doi.org/10.1111/j.1574-6968.1995.tb07834.x.

[32] Hurwitz, B.L., Hallam, S.J., Sullivan, M.B., 2013. Metabolic reprogramming by viruses in the sunlit and dark ocean. Genome Biol 14, R123. https://doi.org/10.1186/gb-2013-14-11-r123.

[33] Hurwitz, B.L., U'Ren, J.M., 2016. Viral metabolic reprogramming in marine ecosystems. Current Opinion in Microbiology. Environ Microbiol * Spec Sect: Megaviromes 31, 161–168. https://doi.org/10.1016/j.mib.2016.04.002.

[34] Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al., 2014. InterProScan 5: Genome-scale protein function classification. Bioinformatics 30, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

[35] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 1–11. https://doi.org/10.1038/s41586-021-03819-2.

[36] Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., Tanabe, M., 2021. KEGG: Integrating viruses and cellular organisms. Nucleic Acids Res 49, D545–D551. https://doi.org/10.1093/nar/gkaa970.

[37] Khanafer, M., Al-Awadhi, H., Radwan, S., 2017. Coliform bacteria for bioremediation of waste hydrocarbons. BioMed Res Int 2017, e1838072. https://doi.org/10.1155/2017/1838072.

[38] Khot, V., Zorz, J., Gittins, D.A., Chakraborty, A., Bell, E., Bautista, M.A., et al., 2022. CANT-HYD: a curated database of phylogeny-derived hidden markov models for annotation of marker genes involved in hydrocarbon degradation. Front Microbiol 12. https://doi.org/10.3389/fmicb.2021.764058.

[39] Kieft, K., Zhou, Z., Anantharaman, K., 2020. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome 8, 90. https://doi.org/10.1186/s40168-020-00867-0.

[40] Kieft, K., Zhou, Z., Anderson, R.E., Buchan, A., Campbell, B.J., Hallam, S.J., et al., 2021. Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. Nat Commun 12, 3503. https://doi.org/10.1038/s41467-021-23698-5.

[41] Koo, C.W., Rosenzweig, A.C., 2020. Particulate methane monooxygenase and the PmoD protein. In: Encyclopedia of inorganic and bioinorganic chemistry. John Wiley & Sons, Ltd, pp. 1–8.

[42] Koonin, E.V., Yutin, N., 2019. Evolution of the large nucleocytoplasmic DNA viruses of eukaryotes and convergent origins of viral gigantism. Adv Virus Res 103, 167–202. https://doi.org/10.1016/bs.aivir.2018.09.002.

[43] Lee, I.P.A., Eldakar, O.T., Gogarten, J.P., Andam, C.P., 2022. Bacterial cooperation through horizontal gene transfer. Trends Ecol Evol 37, 223–232. https://doi.org/10.1016/j.tree.2021.11.006.

[44] Li, L., Liu, X., Yang, W., Xu, F., Wang, W., Feng, L., et al., 2008. Crystal structure of long-chain alkane monooxygenase (LadA) in complex with coenzyme FMN: unveiling the long-chain alkane hydroxylase. J Mol Biol 376, 453–465. https://doi.org/10.1016/j.jmb.2007.11.069.

[45] Li, W., O'Neill, K.R., Haft, D.H., DiCuccio, M., Chetvernin, V., Badretdin, A., et al., 2021. RefSeq: expanding the prokaryotic genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res 49, D1020–D1028. https://doi.org/10.1093/nar/gkaa1105.

[46] Likhoshvay, A., Lomakina, A., Grachev, M., 2014. The complete alk sequences of Rhodococcus erythropolis from Lake Baikal. SpringerPlus 3, 621. https://doi.org/10.1186/2193-1801-3-621.

[47] Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., Chisholm, S. W., 2004. Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proc Natl Acad Sci USA 101, 11013–11018. https://doi.org/10.1073/pnas.0401526101.

[48] Liu, C., Wang, W., Wu, Y., Zhou, Z., Lai, Q., Shao, Z., 2011. Multiple alkane hydroxylase systems in a marine alkane degrader, Alcanivorax dieselolei B-5. Environ Microbiol 13, 1168–1178. https://doi.org/10.1111/j.1462-2920.2010.02416.x.

[49] Liu, Y.-F., Qi, Z.-Z., Shou, L.-B., Liu, J.-F., Yang, S.-Z., Gu, J.-D., et al., 2019. Anaerobic hydrocarbon degradation in candidate phylum `Atribacteria' (JS1) inferred from genomics. ISME J 13, 2377–2390. https://doi.org/10.1038/s41396-019-0448-2.

[50] Lyu, L., Li, J., Chen, Y., Mai, Z., Wang, L., Li, Q., et al., 2022. Degradation potential of alkanes by diverse oil-degrading bacteria from deep-sea sediments of Haima cold seep areas, South China Sea. Front Microbiol 13. https://doi.org/10.3389/fmicb.2022.920067.

[51] Marigómez, I., 2014. Oil, crude. In: Wexler, P. (Ed.), Encyclopedia of toxicology, Third Ed. Academic Press, Oxford, pp. 663–669.

[52] Marín, M.M., Yuste, L., Rojo, F., 2003. Differential expression of the components of the two alkane hydroxylases from Pseudomonas aeruginosa. J Bacteriol 185, 3232–3237. https://doi.org/10.1128/JB.185.10.3232-3237.2003.

[53] McNair, K., Zhou, C., Dinsdale, E.A., Souza, B., Edwards, R.A., 2019. PHANOTATE: A novel approach to gene identification in phage genomes. Bioinformatics 35, 4537–4542. https://doi.org/10.1093/bioinformatics/btz265.

[54] Medina, E.M., Walsh, E., Buchler, N.E., 2019. Evolutionary innovation, fungal cell biology, and the lateral gene transfer of a viral KilA-N domain. Curr Opin Genet Dev Evolut Genet 58–59, 103–110. https://doi.org/10.1016/j.gde.2019.08.004.

[55] Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., Steinegger, M., 2022. ColabFold: Making protein folding accessible to all. Nat Methods 1–4. https://doi.org/10.1038/s41592-022-01488-1.

[56] Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J., Levy Karin, E., 2021. Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics. https://doi.org/10.1093/bioinformatics/btab184.

[57] Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., et al., 2021. Pfam: The protein families database in 2021. Nucleic Acids Res 49, D412–D419. https://doi.org/10.1093/nar/gkaa913.

[58] Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol 32, 268–274. https://doi.org/10.1093/molbev/msu300.

[59] Nie, Y., Chi, C.-Q., Fang, H., Liang, J.-L., Lu, S.-L., Lai, G.-L., et al., 2014. Diverse alkane hydroxylase genes in microorganisms and environments. Sci Rep 4, 4968. https://doi.org/10.1038/srep04968.

[60] Pang, T.Y., Lercher, M.J., 2019. Each of 3,323 metabolic innovations in the evolution of E. Coli arose through the horizontal transfer of a single DNA segment. Proc Natl Acad Sci USA 116, 187–192. https://doi.org/10.1073/pnas.1718997115.

[61] Patel, S., Homaei, A., Patil, S., Daverey, A., 2019. Microbial biosurfactants for oil spill remediation: Pitfalls and potentials. Appl Microbiol Biotechnol 103, 27–37. https://doi.org/10.1007/s00253-018-9434-2.

[62] Pavlenko, L.F., Barabashin, T.O., Zhukova, S.V., Korablina, I.V., Anohina, N.S., Klimenko, T.L., et al., 2022. Components of oil pollution in water and bottom sediments of the northeastern Part of the Russian Black Sea Region. Oceanology 62, 59–67. https://doi.org/10.1134/S0001437022010118.

[63] Perera, M., Wijesundera, S., Wijayarathna, C.D., Seneviratne, G., Jayasena, S., 2022. Identification of long-chain alkane-degrading (LadA) monooxygenases in Aspergillus flavus via in silico analysis. Front Microbiol 13. https://doi.org/10.3389/fmicb.2022.898456.

[64] Prabhakaran, R., Chithambaram, S., Xia, X., 2014. Aeromonas phages encode tRNAs for their overused codons. Int J Comput Biol Drug Des 7, 168–182. https://doi.org/10.1504/IJCBDD.2014.061645.

[65] Probst, A.J., Ladd, B., Jarett, J.K., Geller-McGrath, D.E., Sieber, C.M.K., Emerson, J. B., et al., 2018. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. Nat Microbiol 3, 328–336. https://doi.org/10.1038/s41564-017-0098-y.

[66] Rabus, R., Boll, M., Heider, J., Meckenstock, R.U., Buckel, W., Einsle, O., et al., 2016. Anaerobic microbial degradation of hydrocarbons: from enzymatic reactions to the environment. Microb Physiol 26, 5–28. https://doi.org/10.1159/000443997.

[67] Rokas, A., Wisecaver, J.H., Lind, A.L., 2018. The birth, evolution and death of metabolic gene clusters in fungi. Nat Rev Microbiol 16, 731–744. https://doi.org/10.1038/s41579-018-0075-3.

[68] Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., et al., 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537, 689–693. https://doi.org/10.1038/nature19366.

[69] Ru, J., Khan Mirzaei, M., Xue, J., Peng, X., Deng, L., 2023. ViroProfiler: a containerized bioinformatics pipeline for viral metagenomic data analysis. Gut Microbes 15, 2192522. https://doi.org/10.1080/19490976.2023.2192522.

[70] Schwengers, O., Jelonek, L., Dieckmann, M.A., Beyvers, S., Blom, J., Goesmann, A., 2021. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genom 7. https://doi.org/10.1099/mgen.0.000685.

[71] Secor, P.R., Sweere, J.M., Michaels, L.A., Malkovskiy, A.V., Lazzareschi, D., Katznelson, E., et al., 2015. Filamentous bacteriophage promote biofilm assembly and function. Cell Host Microbe 18, 549–559. https://doi.org/10.1016/j.chom.2015.10.013.

[72] Shaffer, M., Borton, M.A., McGivern, B.B., Zayed, A.A., La Rosa, S.L., Solden, L.M., et al., 2020. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res 48, 8883–8900. https://doi.org/10.1093/nar/gkaa621.

[73] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13, 2498–2504. https://doi.org/10.1101/gr.1239303.

[74] Shao, Z., Wang, W., 2013. Enzymes and genes involved in aerobic alkane degradation. Front Microbiol 4. https://doi.org/10.3389/fmicb.2013.00116.

[75] Somee, M.R., Amoozegar, M.A., Dastgheib, S.M.M., Shavandi, M., Maman, L.G., Bertilsson, S., et al., 2022. Genome-resolved analyses show an extensive diversification in key aerobic hydrocarbon-degrading enzymes across bacteria and archaea. BMC Genom 23, 690. https://doi.org/10.1186/s12864-022-08906-w.

[76] Stagars, M.H., Ruff, S.E., Amann, R., Knittel, K., 2016. High diversity of anaerobic alkane-degrading microbial communities in marine seep sediments based on (1-methylalkyl)succinate synthase genes. Front Microbiol 6. https://doi.org/10.3389/fmicb.2015.01511.

[77] Steinegger, M., Söding, J., 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35, 1026–1028. https://doi.org/10.1038/nbt.3988.

[78] Suttle, C.A., 2007. Marine viruses - major players in the global ecosystem. Nat Rev Microbiol 5, 801–812. https://doi.org/10.1038/nrmicro1750.

[79] Tentori, E.F., Richardson, R.E., 2020. Methane monooxygenase gene transcripts as quantitative biomarkers of methanotrophic activity in methylosinus trichosporium OB3b. Appl Environ Microbiol 86, e01048-20. https://doi.org/10.1128/AEM.01048-20.

[80] Terzian, P., Olo Ndela, E., Galiez, C., Lossouarn, J., Pérez Bucio, R.E., Mom, R., et al., 2021. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinforma 3, lqab067. https://doi.org/10.1093/nargab/lqab067.

[81] The UniProt Consortium, 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47, D506–D515. https://doi.org/10.1093/nar/gky1049.

[82] Thompson, L.R., Zeng, Q., Kelly, L., Huang, K.H., Singer, A.U., Stubbe, J., et al., 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci USA 108, E757–E764. https://doi.org/10.1073/pnas.1102164108.

[83] Tourova, T.P., Sokolova, D.S., Semenova, E.M., Shumkova, E.S., Korshunova, A.V., Babich, T.L., et al., 2016. Detection of N-alkane biodegradation genes alkB and ladA in thermophilic hydrocarbon-oxidizing bacteria of the genera Aeribacillus and Geobacillus. Microbiol (Read, Engl) 85, 693–707. https://doi.org/10.1134/S0026261716060199.

[84] van Beilen, J.B., Funhoff, E.G., 2007. Alkane hydroxylases involved in microbial alkane degradation. Appl Microbiol Biotechnol 74, 13–21. https://doi.org/10.1007/s00253-006-0748-0.

[85] van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Söding, J., Steinegger, M., 2022. Foldseek: Fast and accurate protein structure search. arXiv 2022 02 07 479398. https://doi.org/10.1101/2022.02.07.479398.

[86] Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., et al., 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50, D439–D444. https://doi.org/10.1093/nar/gkab1061.

[87] Vidal-Verdú, À., Gómez-Martínez, D., Latorre-Pérez, A., Peretó, J., Porcar, M., 2022. The car tank lid bacteriome: a reservoir of bacteria with potential in bioremediation of fuel. npj Biofilms Micro 8, 1–12. https://doi.org/10.1038/s41522-022-00299-8.

[88] Viggor, S., Jõesaar, M., Vedler, E., Kiiker, R., Pärnpuu, L., Heinaru, A., 2015. Occurrence of diverse alkane hydroxylase alkB genes in indigenous oil-degrading bacteria of Baltic Sea surface water. Mar Pollut Bull 101, 507–516. https://doi.org/10.1016/j.marpolbul.2015.10.064.

[89] Wang, W., Cai, B., Shao, Z., 2014. Oil degradation and biosurfactant production by the deep sea bacterium Dietzia maris As-13-3. Front Microbiol 5. https://doi.org/10.3389/fmicb.2014.00711.

[90] Wang, W., Shao, Z., 2012. Diversity of flavin-binding monooxygenase genes (almA) in marine bacteria capable of degradation long-chain alkanes. FEMS Microbiol Ecol 80, 523–533. https://doi.org/10.1111/j.1574-6941.2012.01322.x.

[91] Wang, X.-B., Chi, C.-Q., Nie, Y., Tang, Y.-Q., Tan, Y., Wu, G., et al., 2011. Degradation of petroleum hydrocarbons (C6–C40) and crude oil by a novel Dietzia strain. Bioresour Technol 102, 7755–7761. https://doi.org/10.1016/j.biortech.2011.06.009.

[92] Wolf, P.G., Cowley, E.S., Breister, A., Matatov, S., Lucio, L., Polak, P., et al., 2022. Diversity and distribution of sulfur metabolic genes in the human gut microbiome and their association with colorectal cancer. Microbiome 10, 64. https://doi.org/10.1186/s40168-022-01242-x.

[93] Xu, A., Wang, D., Ding, Y., Zheng, Y., Wang, B., Wei, Q., et al., 2020. Integrated comparative genomic analysis and phenotypic profiling of pseudomonas aeruginosa isolates from crude oil. Front Microbiol 11.

[94] Xu, S., Li, L., Luo, X., Chen, M., Tang, W., Zhan, L., et al., 2022. Ggtree: a serialized data object for visualization of a phylogenetic tree and annotation data. iMeta. https://doi.org/10.1002/imt2.56.

[95] Xu, X., Liu, W., Tian, S., Wang, W., Qi, Q., Jiang, P., et al., 2018. Petroleum hydrocarbon-degrading bacteria for the remediation of oil pollution under aerobic

conditions: a perspective analysis. Front Microbiol 9. https://doi.org/10.3389/fmicb.2018.02885.

[96] Yeh, A.H.-W., Norn, C., Kipnis, Y., Tischer, D., Pellock, S.J., Evans, D., et al., 2023. De novo design of luciferases using deep learning. Nature 614, 774–780. https://doi.org/10.1038/s41586-023-05696-3.

[97] Zhang, S., Hu, Z., Wang, H., 2019. Metagenomic analysis exhibited the co-metabolism of polycyclic aromatic hydrocarbons by bacterial community from estuarine sediment. Environ Int 129, 308–319. https://doi.org/10.1016/j.envint.2019.05.028.

[98] Zhang, X., Kong, D., Liu, X., Xie, H., Lou, X., Zeng, C., 2021. Combined microbial degradation of crude oil under alkaline conditions by Acinetobacter baumannii and Talaromyces sp. Chemosphere 273, 129666. https://doi.org/10.1016/j.chemosphere.2021.129666.

[99] Zhang, Y., Skolnick, J., 2004. Scoring function for automated assessment of protein structure template quality. Protein: Struct Funct Bioinforma 57, 702–710. https://doi.org/10.1002/prot.20264.