# ON BIOINFORMATICS OF THE HUMAN GUT VIROME

**Dissertation**

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Jaime Leonardo Moreno Gallego

aus Bogotá / Kolumbien

Tübingen

2022

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

| | |
|---|---|
| Tag der mündlichen Qualifikation: | 15.12.2022 |
| Stellvertretender Dekan: | Prof. Dr. Thilo Stehle |
| 1. Berichterstatterin: | Prof. Dr. Ruth E. Ley |
| 2. Berichterstatter: | Prof. Dr. Daniel Huson |

# Zusammenfassung

Die Shotgun-Metagenomik hat eine noch nie dagewesene Vielfalt von Viren enthüllt, die Wechselwirkungen mit Organismen aus allen Zweigen des Lebens aufweisen. Somit veränderte die Metagenomik den Blickwinkel auf Viren: von bloßen Krankheitserregern zu bedeutenden Akteuren innerhalb des ökologischen Gleichgewichts. Virale Metagenomik mariner Gewässer offenbarte die lokalen und globalen Auswirkungen von Virus-Wirt-Interaktionen durch ihren Einfluss auf biogeochemische und ökologische Prozesse. Im menschlichen Darm ist die Rolle und Bedeutung von Viren hingegen immer nicht ausreichend erforscht. Zwar sind Bakterien und Viren im menschlichen Darm die häufigsten Organismen, jedoch liegt die Beschreibung der viralen Fraktion weit hinter der Bakterienfraktion. Daher war es das Ziel dieser Forschungsarbeit, die bestehende Charakterisierung des menschlichen Darm-Viroms zu erweitern.

Zuerst analysierte ich Virome von eineiigen Zwillingen, zur Untersuchung, ob die virale Diversität die bakterielle Diversität im menschlichen Darm widerspiegelt - ein erwartetes ökologisches Muster, das jedoch nie zuvor bestätigt wurde. Die Analyse belegte ein einzigartiges, von Bakteriophagen dominiertes, menschliches Darmvirom. Durch den Vergleich Metriken viraler und bakterieller Vielfalt stellte ich fest, dass sich die Vielfalt des Darmmikrobioms innerhalb und zwischen den Probanden in ihren Viromen widerspiegelt. Darüber hinaus erwiesen sich die Häufigkeit und Vielfalt von Bakterien als Indikator für die Häufigkeit und Vielfalt des Viroms.

Zweitens wurde ich mit dem Mangel an annotierten Referenzgenomen konfrontiert, der die Analyse viraler Metagenome beeinträchtigte. Ich durchsuchte öffentliche Datenbanken nach transposablen Phagen des menschlichen Darms. Transposable Phagen sind gut beschrieben, da sie Mutationen, genomische Umstrukturierungen und horizontalen Gentransfer in ihren Wirten verursachen. Dennoch sind nur ein paar Dutzend Genome in öffentlichen Datenbanken vorhanden. Ich habe aus assemblierte Metagenom-Datenbanken 1.002 qualitativ hochwertige Assemblies von mutmaßlich transposablen Phagen identifiziert. Auf der Grundlage vergleichender Genomik und phylogenetischer Analysen stellte ich fest, dass transposable Phagen wider Erwarten keine monophyletische Gruppe sind. Schließlich befasste ich mich mit der Charakterisierung der Assemblies im Hinblick auf das bisher einzige bekannte Isolat einer transposablen Phagen aus dem menschlichen Darm: Mushu (NC_047913). Unter Verwendung von Mushu als Referenz und gemäß der taxonomischen Klassifizierungsrichtlinien habe ich die Mushu-ähnliche Familie definiert. Sie umfasst 9 Gattungen und 72 Arten von Phagen, die am horizontalen Gentransfer von auxiliären Stoffwechselgenen beteiligt sein können.

Diese Arbeit belegte die Korrelation zwischen bakterieller und viraler Vielfalt im menschlichen Darm; ein durch Bakteriophagen und nicht eukaryotischen Viren bestimmtes Muster. Ein wesentliches Hindernis bei der Beantwortung dieser Frage ist die spärliche Charakterisierung der Mehrzahl der viralen Metagenome. Eine Verbesserung der Charakterisierung viraler Komponenten des Mikrobioms ist daher von entscheidender Bedeutung.

Im Rahmen dieser Arbeit wurden Tausende von transposablen Phagen aus Datenbanken mit Metagenom-Zusammenstellungen ermittelt. Die gleichen Methoden können auch angewandt werden, um alle Genome der Familien von Bakteriophagen mit Schwanzstruktur zu finden. Ein solcher Ansatz hat das Potenzial, die Charakterisierung der viralen Fraktion des menschlichen Darms zu beschleunigen und künftige Untersuchungen über die Rolle der Phagen beim Aufbau und der Homöostase des Darmmikrobioms zu erleichtern.

# Abstract

Shotgun metagenomics has revealed an unprecedented diversity of viruses maintaining interactions with organisms from all divisions of life. Thus, metagenomics changed the perspective on viruses: from mere pathogens to significant players within ecological systems. Viral metagenomics of marine waters revealed the local and global scale impact of virus-host interactions by their influence on biogeochemical and ecological processes. In contrast, the description of viruses and their effect on the human gut is still poorly understood. Bacteria and viruses are the most abundant entities in the human gut, yet the description of the viral fraction lags behind the bacterial fraction. Thus, the objective of this research was to expand existing characterizations of the human gut virome.

First, I analyzed viromes from monozygotic twins to determine whether viral diversity mirrors bacterial diversity in the human gut, an expected ecological pattern but never before verified. The analysis confirmed a highly unique human gut virome dominated by bacteriophages. Comparing metrics of viral and bacterial diversity, I observed that gut microbiome diversity, within and between subjects, is mirrored in their viromes. Moreover, the abundance and diversity of bacteria proved to be indicative of the abundance and diversity of the virome.

Second, I faced the lack of annotated reference genomes impeding the analysis of viral metagenomes. I screened public databases in search of transposable phages from the human gut. Transposable phages are well-described as agents of mutation, genomic rearrangements, and horizontal gene transfer to their hosts. Despite this, only a few dozen genomes are available in public databases. I identified 1,002 good-quality assemblies of putative transposable phage in metagenomic assembly databases. Based on comparative genomics and phylogenetic analysis, I found that transposable phages are not a monophyletic group, contrary to expectations. Finally, I focused on the characterization of assemblies related to the only human intestine transposable phage isolated to date: Mushu (NC_047913). Using Mushu as a reference and following the taxonomic classification guidelines, I defined the Mushu-like family. It includes 9 genera and 72 species of phages that may be involved in the horizontal gene transfer of auxiliary metabolic genes.

This work demonstrated the correlation between bacterial and viral diversity in the human gut; a pattern driven by bacteriophages and not eukaryotic viruses. A major obstacle in studying the human gut virome is the sparse characterization of most viral metagenomes. Thus, improving the description of the viral component of the microbiome is critical. This work retrieved thousands of transposable phages from databases of metagenomic assemblies. The same methods can be applied to find genomes related to all families of tailed bacteriophages. Such an approach has the potential to catalyze the characterization of the viral fraction of the human gut and facilitate future investigation on the role of phages in the assembly and homeostasis of the gut microbiome.

# Acknowledgements

To my thesis advisory committee: Ruth Ley, Daniel Huson, and Alejandro Reyes. Thank you for the helpful discussions and guidance during the annual TAC meetings. Also, Ruth, thank you for providing me with the means and resources to develop my research, and for your fast-track English writing lessons. Daniel, thank you for your help in navigating the university system and for always being reachable. Alejandro, thank you for creating the "Viromas" group and introducing me to the fascinating world of viral metagenomics.

To Shao-Pei Chou and Ian Hewson: The viromes from monozygotic twin samples you selected, purified and sequenced paved the way towards my doctoral studies. Thank you for the fantastic dataset you created.

To all the members of the Ley lab: I never received anything but solidarity from you. Special thanks to Sophie, Silke, Jess, and Daphne: who helped me during the year I did wet lab. Hagay and Tony: who always motivated me with their questions about my project or with whom I enjoy chats about other random topics. Of course, thanks to Jacobo, Andrea, Tanja, Alban, Claudi, Hagay, Chris, Nick, Jill, and Ruth: who helped me with the "divide and conquer" approach I took to review this manuscript or any other work I submitted during my Ph.D. studies (TAC reports, abstracts, posters, and slides). Your help was invaluable.

To the viromas group: Alejandro, Guillermo, Gama, Ave, Lau Camelo, Lau Forero, Juanse, Lucho, Steven, and the most recent members: Michael, Natalia, and XXX. It has been very nice to share our troubles and successes in viromics. Thank you for sharing laughs and ideas through these years. Rest assured that you have helped to shape my project.

To my friends in Tübingen: Isabella, Aki, Nils, Laura J., Guille, Albane, Mirabeau, Jacobo, Aleja, Sara W., Adrian, Abiram, Roger, Diego, Cami, Eddy, Ian, and all the others I don't name but have enriched my life throughout these years as a Ph.D. student here in Tübingen. Thank you for the lab chats, the life chats, the football games, the biking tours, the ski lessons, the german jokes, the food, the mexican food. Thank you for your friendship.

To my friends that I left behind in Colombia: Nestor, Sebas, Cocar, Dani, Pipe Romero, Angie, Alzate, and "El profe" Eduin. Thank you for always being there for me, asking me how I am doing, and showing me your admiration to encourage me in this process.

To my parents and brother, whose sacrifices and efforts put me on the escalator it takes to reach the shoulders of giants. Without their sacrifices, I would never have been here sharing my research.

To my Andrea, for all your love and unconditional support.

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| VLP | Virus-like particles |
| UViGs | Uncultivated Viral Genomes |
| ICTV | International Committee of Taxonomy of Viruses |
| ViPhOGs | Eukaryotic Viruses and Phages Orthologous Groups |
| IGC | Integrated Gene Catalogs |
| HMM | Hidden Markov Model |

# Prologue

Viruses have caused several deadly pandemics throughout human history, the last one: Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) (2019-present). Because of that, and other historical reasons, viruses are commonly recognized as pathogens. However, viruses are more than merely pathogens. Viruses are the most abundant biological entities on earth, inhabit all kinds of environments, and maintain interactions with organisms from all divisions of life. Beyond infecting and reproducing, viruses are actors and builders of the web of life. They play an important role in determining community structure, modify the metabolism of their hosts, and facilitate gene transfer between and across species. During my doctorate, by examining viral metagenomes and databases derived from metagenomic data, I dedicated myself to the study of viruses inhabiting the human gut.

I organized this thesis as follows. In the first chapter, I introduce four topics that provide context and concepts to the reader. Chapters two and three, constitute my original research. Consequently, each chapter provides its motivation and scope, results, discussion, and methods. Finally, I close this thesis by highlighting the main outcomes of my research, clarifying my contribution to the field of viral metagenomics, and suggesting how this research might strength further research on the human gut virome.

# Chapter 1

# Introduction

## 1.1 Viral metagenomics

Microbial ecology is the study of the interactions of microorganisms with their environment, each other, and plant and animal species [1]. For almost 300 years, from the first report of microscopic organisms by Leeuwenhoek in 1676 [2] up until the invention of Sanger automated sequencing [3] and the use of ribosomal RNA genes as molecular markers for the classification of prokaryotes [4], the study of microorganisms was based entirely on morphology features, growth, and biochemical profiles. Without the use of genetic-based approaches, microbial ecology was restricted to the study of culturable microorganisms. The aforementioned advances paved the way for the development of metagenomics: the analysis of genomes of all microorganisms present in a specific environment [5]. Metagenomics opened the gate to an uncultured world of microbial communities.

Recently, metagenomic studies have moved beyond simply looking at communities of bacteria, but to other microorganisms as well, including viruses. "Viral Metagenomics" focuses on the study of viral genetic material in a particular habitat, with a 'viral metagenome' (a virome) being the collection of all viral genomes from a specific environment. Compared to the size of bacterial genomes (130 Kbp - 13 Mbp) [6, 7] viral genomes are incredibly small (324 bp - 2 Mbp) [8]. Therefore, the majority of genetic material that is found in a given sample is of non-viral origin. For this reason, it is often necessary to physically separate virus-like particles (VLP) from microbial cells to collect the virome. For solid samples with a high viral density, such as feces, a common approach is to resuspend the material in an osmotically neutral buffer, followed by several steps of centrifugation and filtration to remove large particles, including organic debris and larger microorganisms. Once the VLP are enriched, non-encapsulated free nucleic acids are removed by treatment with DNase and RNase, and DNA and RNA extraction methods can

then be applied to isolate the variety of genomes that can be found within a viral community. Finally, as the yield of DNA following extraction of nucleic acids from purified VLPs is often below the required minimum for sequencing, a step of random amplification of viral DNA is usually performed [9, 10].

Once virome DNA or RNA is extracted, samples are sequenced with next-generation sequencing technologies. The resulting data is then used to reconstruct the genomic content and characterize the community via computational methods. While this is transversal to all metagenomic studies, the analysis of viromes displays particular challenges and considerations. Virome data is characterized by high proportions of repeat regions within viral genomes, hypervariable genomic regions associated with host interaction, and high mutation rates, leading to increased metagenomic complexity and strain variation [11, 12, 13]. All these, result in fragmented virome assemblies, which impact the estimation of the underlying species and functional diversity [14, 15]. Additionally, viruses do not have a universal gene marker and are also characterized by a mosaic genome composition, with different genes having different evolutionary histories due to events of horizontal gene transfer [16, 17].

Two decades of viral metagenomics, since the first metagenomic evaluation of two marine viral communities [18], have left more than the recognition of the intrinsic challenges of viral metagenomics. Viral metagenomics has allowed us to uncover the vast array of genetic diversity found in viral communities. Consequently, the continuous analysis of viral genomes from a diverse range of biomes has promoted the development of computational tools and strategies that support their characterization [19]. The impact of these technologies has been so significant that, together with Metagenomic Assembled Genomes (MAGs), we now have Uncultivated Viral Genomes (UViGs), which are now officially recognized by the International Committee for the Taxonomy of Viruses (ICTV). UViGs make up more than 95% of the current diversity in public databases [20]. Classifying and characterizing them will facilitate taxonomic decomposition and functional characterization of viromes.

## 1.2 The human gut virome

The gastrointestinal (GI) tract is one of the most densely populated areas of the human body, providing a heterogeneous and expansive surface area ($> 200\ m^2$) for microbial life [21]. The human gut is estimated to contain between 30 and 400 trillion micro-organisms [22]. Although dominated by bacteria and their viruses (bacteriophages or more simply phages), there are also archaea as well as fungi, other eukaryotes, and their associated

viruses. This dense microbial ecosystem has been shown to perform a variety of essential functions, such as aiding digestion, conditioning our immune system, and protecting us from pathogens [23, 24].

In many environments phages have been found to outnumber their bacterial hosts by one order of magnitude [25, 26]. This is not the case in the human gut. Based on microscopy counts, the cellular fraction of the human gut microbiome appears typically at $10^{11}$ - $10^{12}$ cells per gram of feces [22], and VLP are found at an approximately equal proportion (between $10^9$ to $10^{12}$ VLP per gram of feces) [27, 28, 29]. However, the phage-to-bacteria ratio is not consistent throughout the GI tract. This ratio was observed to be increased on the mucosal surfaces, not only in humans but across other animal species [30]. Barr and collaborators proposed the 'bacteriophage adherence to mucus' model to explain this observation, and suggested that phages protect the host mucosa from bacterial infections and increase the frequency of interaction with their bacterial hosts.

In contrast to non-host-associated ecosystems such as aquatic environments, predator-prey interactions [25, 31] are not observed in the human gut. Instead, most of the dominant virotypes detected in the gut ecosystem show evidence of a temperate lifestyle as indicated by the frequency of integrase genes and other genetic features [32]. Temperate phages integrate into their host genome or exist as plasmids within their host cell for generations (lysogenic cycle) instead of directly killing their hosts (lytic cycle) [33]. Furthermore, both the phages and bacteria of the gut display some common patterns of diversity across hosts, such as high levels of interpersonal differences and relative stability over time, meaning low intrapersonal variation [32]. This pattern might be explained due to shared dietary habits, which drive similarity between microbiomes [34, 35], and also between viromes [36].

Human gut virome studies are still in a descriptive phase but this has still been a challenging task. The vast majority of virome reads cannot be annotated functionally or taxonomically, highlighting the vast level of novel gene content encoded by the virome. In consequence, most studies have ignored considerable proportions of the data [37]. Nonetheless, the amount of metagenomic information accumulated has allowed the creation of databases of metagenomic assemblies, such as the gut virome database [38] and the gut phage database [39]. Furthermore, key players of the human gut virome have been revealed; this is perhaps best exemplified by "CrAssphage", which was detected using a crossed-assembly strategy, hence its name. Then, it was identified as the most common phage of the human gut recruiting up to 90% of the reads [40]. Recently, CrAss-like phages were recognized as a class of bacteriophages that have potentially co-evolved with humans as they display local, and global clustering patterns [41]. While the role

of CrAss-like phages -and phages in general- is still unknown, phages of the human gut are expected to be important players in the assembly of the gut microbiome. In the gut, phages might modulate the bacterial community through their ability to lyse and kill host bacteria [42, 30], protect bacteria against the infection of other phages (superinfection exclusion) [43], facilitate gene transfer between strains and species (transduction) [44, 45], or by exerting an evolutionary pressure over their hosts (phage-bacteria arms race) [46]. All these events are expected to occur in the human gut as they have been observed in other environments. However, the description of how the sum of these events influences the assembly and maintenance of the gut microbiome remains an open question.

## 1.3 Viral taxonomy

The increasing interest in viral metagenomics together with the development of tools and computational methods to analyze them has revealed an astonishing diversity of viruses. Typically, the study of any biological entity begins with its description and, if possible, its classification in relation to other known biological organisms. The International Committee of Taxonomy of Viruses (ICTV) is charged with the task of developing, refining, and maintaining a universal virus taxonomy. This task encompasses the classification of virus species and higher-level taxa according to the genetic and biological properties of their members [47].

Until recently, the classification of viruses was based on molecular (the type of genetic material present), epidemiological (factors relating to host type, and transmission, among others), and morphological (virion type) characteristics. This classification initially covered family and genus, but then it was expanded into five taxonomic ranks, adding order, subfamily, and species [47]. Although this traditional classification was useful for studying viruses of epidemiological interest, and for understanding the "microevolution" of closely related species, the viral diversity discovered through metagenomics promoted the creation of an expanded classification framework that reflects the distant evolutionary relationships between viruses and their multiple origins [48, 49, 50]

To comprise the true extent of virus genomic diversity, the ICTV created a taxonomy of 15 hierarchical ranks that better aligns with the Linnaean taxonomic system [51]. The taxonomy is based on work that used phylogenetic analysis of virus hallmark genes combined with gene-sharing networks to establish the organization of the global virosphere [52]. The 15 ranks include eight primary ranks and seven derivative ranks. The primary ranks include four previously used (order, family, genus, and species) and four new: realm, kingdom, phylum, and class. All the new ranks are found above the order rank [51].

The new taxonomic system led to a change that will impact viral metagenomics of the human gut: the abolition of the order Caudovirales, the most abundant type of phages in the human gut [37]. The abolished Caudovirales group included all tailed bacteriophages and divided them into three families according to their morphologies: Myoviridae (phages with a long contractile tail), Siphoviridae (phages with a long non-contractile tail), and Podoviridae (phages with a short tail). With the advent of viral metagenomics in the early 2000s, the sequencing of phage genomes revealed a much higher diversity of tailed viruses than the one described at the moment. With the continuously increasing number of genomes, it became apparent that these tree families were not monophyletic [53]. Since then, several studies have illustrated the paraphyly of Caudovirales [54, 55, 56, 57, 58, 8]. After abolishing the order Caudovirales and its families, all underlying taxonomic ranks are kept in the class Caudoviricetes (Phylum: *Uroviricota*, Kingdom: *Heunggongvirae*, Realm: *Duplodnaviria*). At the time of writing, the class Caudoviricetes includes 4 orders, 47 families, 98 subfamilies, 1197 genera, and 3601 species.

Viral metagenomics has called for a change in the way that the evolutionary relationships of viruses are described. In response, the ICTV has developed a taxonomy that is not set in stone and is open to the classification UViGs from sequencing data. The continuous inclusion of UViGs in the taxonomy will improve the understanding of the global virosphere.

*Note*: Given that this change is recent, the second chapter of this thesis does not reflect the new ICTV taxonomy framework.

## 1.4 Orthologs Groups of Eukaryotic Viruses and Phages

Viruses are obligate parasites. Thus, all viruses associated with uncultured organisms are also unculturable. As a consequence, the viral diversity in public databases consisted only of a couple of thousand viruses that did not present much resemblance to new viruses sequenced via viral metagenomics [37]. On top of that, it is accepted that viruses are modular [16], mosaic [17], and have a fast evolutionary rate [59]. These intrinsic characteristics of viruses further complicate the comparison of the genomic diversity in the global virosphere, which has been referred to as "the viral dark matter" [60].

Orthologs are defined as genes in different species that have evolved through speciation events only. In contrast, paralogs arise by duplication events [61]. The analysis of shared orthologs is a key component of comparative genomic studies. In viral genomics, the analysis of orthologs has helped to overcome the challenges imposed by viral diversity. This approach has been applied to study the evolutionary relationships among phages,

allowing not only the reconstruction of phylogenies of particular phages but has also led to a proposed organization of the global virosphere, which has encouraged the transformation of the ICTV taxonomy [52]. The identification of orthologous genes serves two key purposes: delineating the genealogy of genes to investigate the forces and mechanisms of evolutionary processes, and collating groups of genes with the same biological functions [61].

A foundational work on the use of orthologs to study the evolution and molecular repertoire of dsDNA phages reported that the majority of phage genes have no paralogs in the same genome. It also gave rise to the first database of phage orthologous groups (POGs) [62]. This work has been continuously expanded and new databases of phage orthologs have emerged [63, 64]. During my master's studies, I also aimed to expand the search of orthologous groups to keep track of the increasing number of complete viral genomes in public databases. Importantly, I expanded the search to include not only phages but also eukaryotic viruses. My work resulted in a set of 31,150 ViPhOGs (Eukaryotic Viruses and Phages Orthologous Groups) [8]. Additionally, it confirmed that not a single orthologous group is present in all viral genomes but that there are some evolutionary links between eukaryotic and prokaryotic viruses. For example, between members of Herpesvirales and some tailed phages of the abolished Siphoviridae family [65], or between all ssRNA(+) viruses, which presumably co-evolved with their hosts before they split into eukaryotes [66].

The ViPhOGs database constitutes a "parts list" for all viruses available at the time. As each ViPhOG is a set of related genes, profile Hidden Markov Models (HMM) can be built from them. Profile HMMs seem to be the most effective to detect distantly related organisms, showing a higher precision than BLAST searches in metagenomic datasets, especially for more divergent viral sequences [67]. Thus, the set of ViPhOGs is a powerful tool to explore the viral dark matter.

# Chapter 2

# The virome of adult monozygotic twins with concordant or discordant gut microbiomes

## 2.1 Motivation

The temporal population dynamics of phages and their hosts is expected to be linked. Viruses play a key role in the regulation of bacterial populations in aquatic systems, where oscillations of both phages and their hosts have been described [26, 31, 68, 69]. In the human gut, patterns of predator/prey dynamics between virome and microbiome are not typical [36, 70, 46, 42]. Nonetheless, the virome and microbiome display common diversity patterns across hosts, such as high interpersonal differences and relative stability over time [32].

The degree to which the microbiome drives these common patterns has been difficult to evaluate due to confounding factors such as host relatedness. By studying the viromes of monozygotic (MZ) twin pairs, host genetic relatedness can be controlled. Based on beta-diversity distance metrics, a former member of the Ley lab purified and sequenced VLP from 21 twin pairs selected due to the low concordance (12 pairs, henceforth microbiome-discordant twins) or high concordance (9 pairs, henceforth microbiome-concordant twins) of the cellular fraction of their microbiome (Figure 2.1). The generation of a data set controlled by genetic relatedness resulted in microbiome diversity as the only factor explaining virome diversity. More details on the sample selection, purification, and sequencing can be found in the manuscript published together with Shao-Pei and the rest of the collaborators in 2019 [71].

In this chapter, I present the bioinformatic analysis of the aforementioned viromes

from microbiome-discordant and microbiome-concordant MZ twins. The results indicate that microbiome diversity and virome diversity measures are positively associated.

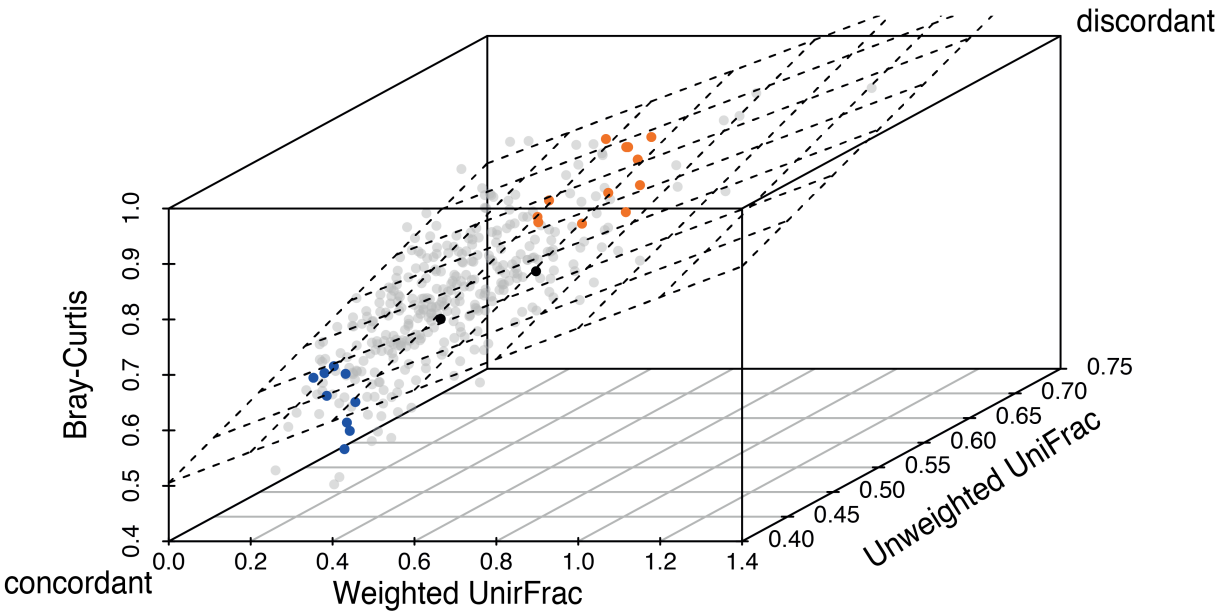


**Figure 2.1: Microbiome distance in MZ twins** The $\beta$-diversity measures of the microbiotas of 354 monozygotic twin pairs from a previous study [72] are shown. Each dot represents the $\beta$-diversity of a pair of twins, measured by the weighted UniFrac (x-axis), unweighted UniFrac (z-axis), and Bray-Curtis (y-axis) $\beta$-diversity metrics. The plane is the least squared-fitted plane Bray-Curtis  Weighted UniFrac + Unweighted UniFrac. A subset of twin pairs with concordant microbiotas (blue) and discordant microbiotas (orange) was chosen from the two edges. Black dots indicate the samples used for virome and whole fecal metagenome comparison.

## 2.2 Results

### 2.2.1 Bacterial DNA screening reveals complete bacterial genomes in viromes

Accurate quality control is essential for metagenomic studies. Biases and problems in sequencing data might affect further analysis, influencing the community's diversity and misleading the hypothesis assessment. Fortunately, methods for the quality control of metagenomic data are pretty mature, and several programs are available for the detection and correction of potential problems in data [73]. Nevertheless, there is no established procedure to deal with bacterial contamination in viromes. While there is not a clear

explanation of how bacteria might overcome the different filters included in the experimental methods for the purification of viral particles, it is clear that DNA extracted from VLP preparations carry bacterial DNA [74]. Thus, to accurately capture the viral diversity patterns, it is essential to assess the presence of bacterial contaminants in the viral metagenomes.

Here, I analyzed sequences from two separate libraries prepared with the DNA extracted and amplified from virus-like particles of fecal samples of microbiome-concordant and microbiome-discordant twins. The human gut microbiota includes a vast diversity of bacteria, archaea, eukaryotes, and viruses. For simplicity, I will refer to the (genetic material of the) cellular fraction of the microbiota as the microbiome, i.e. excluding the virome.

A first sequencing library ("large-insert-size library") was selected with an average insert size of 500 bp (34,325,116 paired reads in total; $817,265 \pm 249,550$ paired reads per sample after quality control; $average \pm std$ ) and used for de novo assembly of viral contigs. Smaller fragments with an average insert size of 300bp were purified in a second library ("small-insert-size library") and sequenced. I merged the resulting pair-end reads into 25,324,163 quality-filtered longer reads to increase mapping accuracy ($602,956 \pm 595,444$ merged reads per sample) (Appendix .1).

To assess bacterial DNA contamination, I mapped virome reads against a set of 8,163 fully assembled bacterial genomes and evaluated the coverage of each genome in bins of 100 Kb. Genomes with a median coverage greater than 100 were considered contaminants. Reads mapping to short regions were considered to be prophages or horizontally transferred genes and retained (Figure 2.2). Instead, reads mapping to potential contaminants were removed from further analyses. In total, 65 bacterial genomes were considered putative contaminants based on their coverage, with $1\% \pm 1.125$ reads per sample assigned to those bacterial genomes. The majority (37/68) belonged to the Firmicutes phylum; at the species level, *Bacteroides dorei*, *B. vulgatus*, *Ruminococcus bromii*, *Faecalibacterium prausnitzii*, *B. xylanisolvens*, *Odoribacter splanchnicus*, and *B. caecimuris* (in that order) were detectable in at least 50% of the samples. Assuming that the most abundant bacterial species in the microbiome are the most likely sources of contamination, their relative abundance should correspond with their relative abundance as contaminants. However, that was not the case (Figure 2.3). Attempting to identify patterns in putative bacterial contamination, I repeated the analysis in 66 samples from five additional publicly available studies on the human gut virome (Table 2.1). In total, I spotted 148 bacterial species as contaminants (including the ones from the viromes of MZ twins). Intriguingly, the species *R. bromii*, *Intestinibacter bartlettii*, *F. prausnitzii*, *B. uniformis*, *Prevotella*

*copri*, *Eubacterium siraeum*, *B. vulgatus*, *Romboutsia timonensis*, and some others were identified in at least 3 out of 5 studies (Figure 2.4).

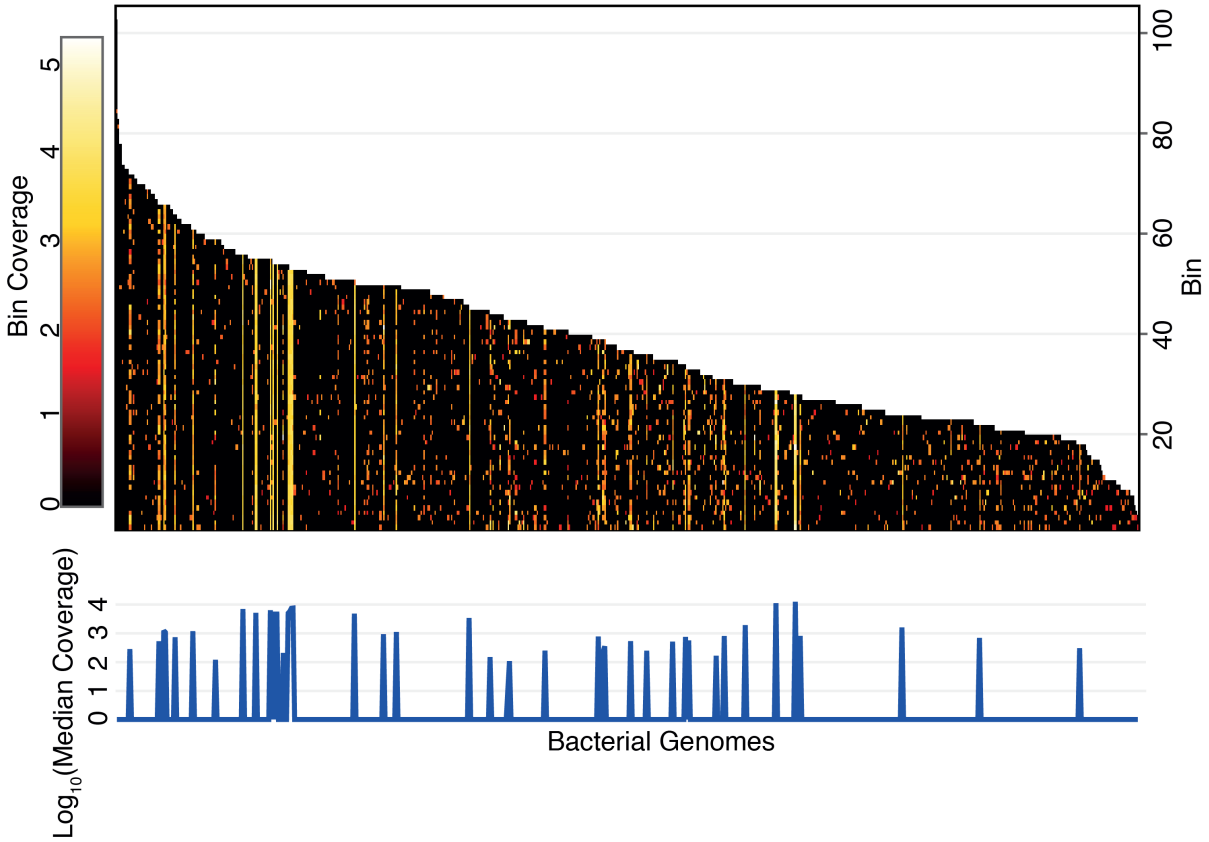


**Figure 2.2: Bacterial contamination in VLP preparations.** Heatmaps of VLP reads from a single sample (4A) mapping to bacterial genomes. Bacterial genomes are represented with vertical bars, sorted by length, and split into bins of 100,000 bp. Genomes with a median coverage greater than 100 were considered contaminants. The color scale to the left shows bin coverage and the line plot below shows the median bin coverage of each genome.


**Figure 2.3: Bacterial abundance doesn't explain contamination in viromes.** Cladogram based on the NCBI taxonomy of the 65 genomes identified as contaminants across all VLP extractions. Right: Spearman rank correlation coefficient (rho) between the abundance of the bacterial genomes in the VLP extractions and 16S rRNA gene profile from the microbiome. Left: total abundance of each bacterial genome added across all individuals.

| Study | Year | Samples (#) |
|---|---|---|
| Minot et al. [75] | 2013 | 24 |
| Lim et al. [76] | 2015 | 48 |
| Manrique et al. [77] | 2016 | 4 |
| Shkoporov et al. [78] | 2018 | 7 |
| Fernandez-Orth et al. [79] | 2018 | 7 |

**Table 2.1: Selected human gut virome studies.** References, publication year, and number of samples in each dataset from virome studies that used Illumina as sequencing platform.

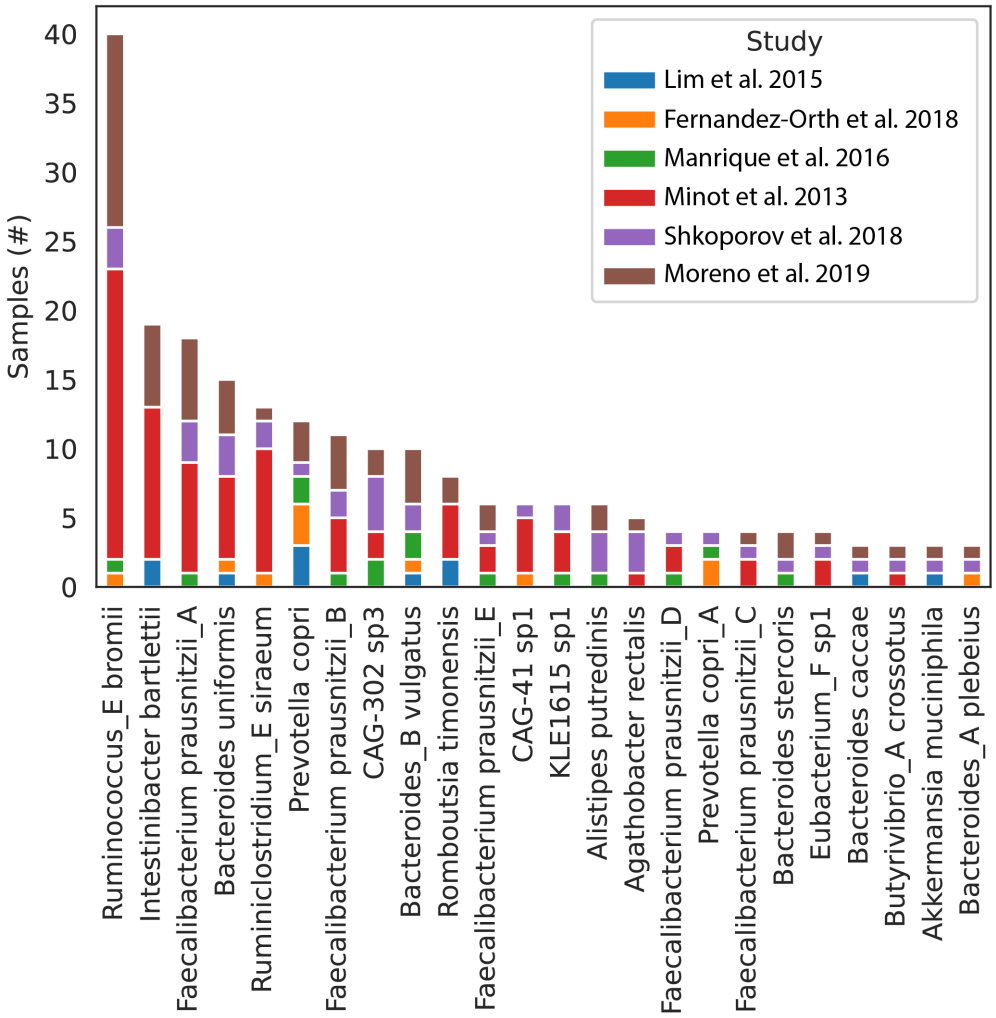


**Figure 2.4: Common bacterial contaminants in human gut virome studies.** Stacked bar plots showing the number of samples where a bacteria was identified as a contaminant. Only bacteria identified in at least 3 out of 6 virome studies are shown.

By mapping reads of viral metagenomes to an extensive dataset of bacterial genomes and differentiating putative contaminants from putative prophage regions, it was possible to deal with bacterial contamination in viromes. The reason for this putative contamination is unclear and requires further research. But so far, neither taxa relative abundance nor potential limitations of the experimental procedures provided the evidence to explain this observation.

### 2.2.2 Database-dependent and independent characterizations support viral enrichment in virus-like particles purifications

Viral metagenomics is a powerful tool for discovering uncultured viruses from different environments [80, 81]. Nevertheless, viral ecology analyses remain challenging since viruses do not have a universal molecular marker. Given phages' immense diversity and mosaic nature, viral genome assembly is affected, impacting the community profiles used to study the patterns of diversity in the environment of interest [14]. To overcome this particular challenge of viromes, I exploited the information in the two metagenomic libraries to produce three different layers of information: a functional layer, which is based on reads and is database-dependent; a virotype layer, independent of contig annotation -a virotype is an assembled contig that satisfies defined length and coverage thresholds-; and a taxonomy layer, which consists of the taxonomic annotation of all assembled contigs.

To assess the functional content of the viromes, I annotated the "short-insert-size library" raw reads using the KEGG annotation of the Integrated Gene Catalog (IGC) [82]. In line with previous reports [83, 36, 32], the majority of reads ($85.43 \pm 5.74\%$) from the viral metagenomes mapped to genes with unknown function (Figure 2.5). To further verify that sequences were derived from VLP and differentiated from the microbiome, metagenomes from VLP and bulk fecal DNA from 4 additional individuals (2 twin pairs) were provided. I analyzed these samples exactly as I analyzed the "short-insert-size library" (Figure 2.1). As expected, the functional profiles of viromes and microbiomes derived from the same samples were dissimilar. Virome reads that mapped to annotated genes were enriched in two categories: Genetic Information Process ($48.87\% \pm 12.12$) and Nucleotide Metabolism ($17.59\% \pm 8.81$), compared to $24.31\% \pm 1.28$ and $5.47\% \pm 0.4$ for the microbiome, respectively. Most of the functional categories present in the microbiome were essentially absent from the viromes. Furthermore, the functional annotations of the viromes showed higher between-sample variability than the ones from microbiomes. Also, a lower intraclass correlation coefficient (Figure 2.6).

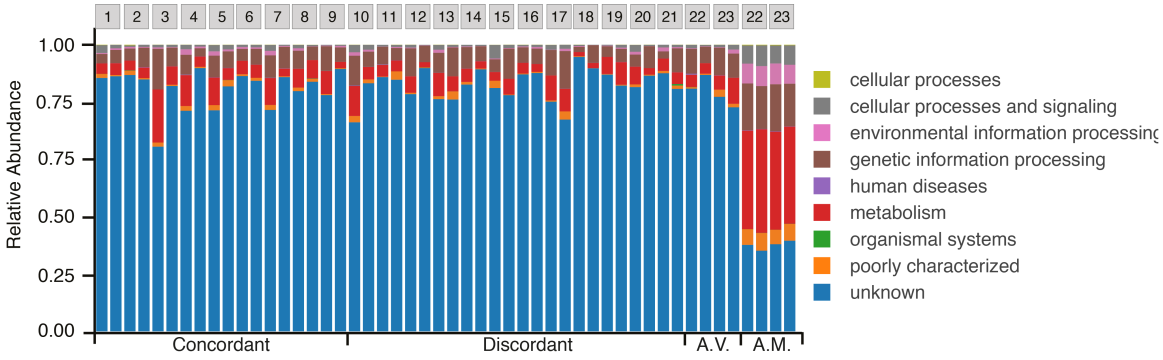


**Figure 2.5: Viromes are dominated by the unknown.** The relative abundance of KEGG categories in whole fecal metagenomes and viromes, including all hits to IGC genes, regardless of annotation.

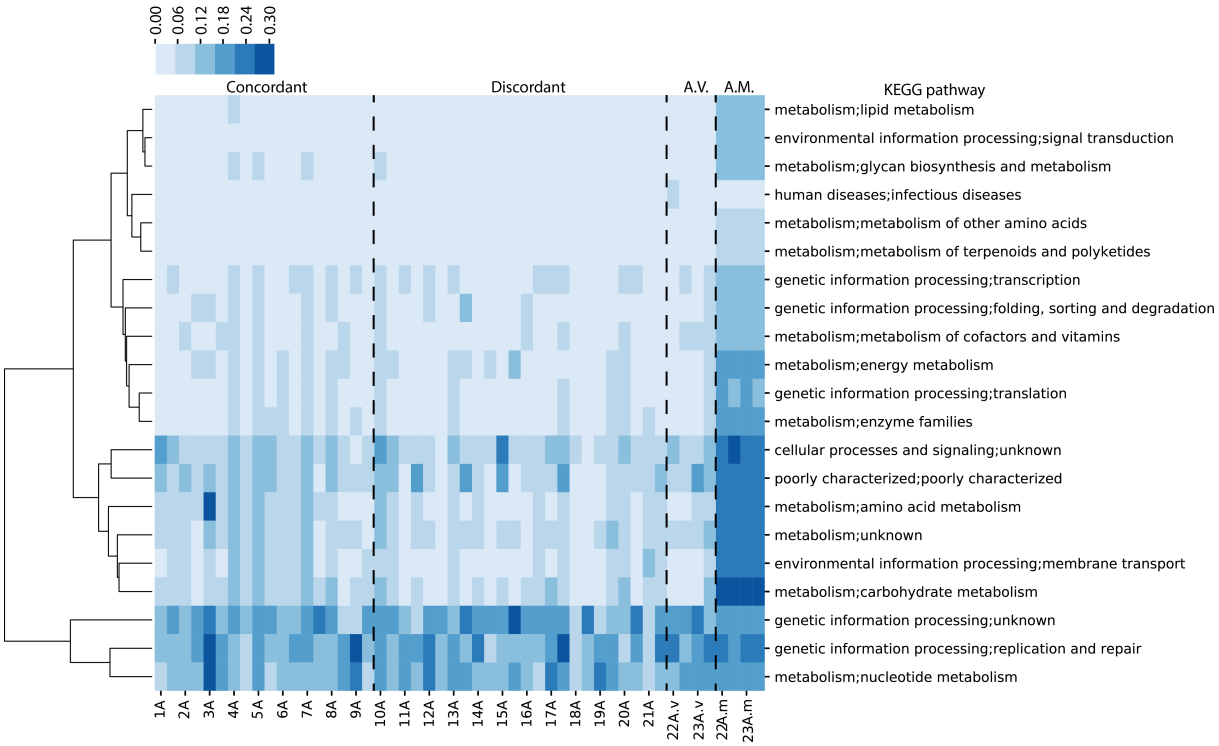


**Figure 2.6: Functional annotation of viromes.** Heatmap of the relative abundance of the second level of KEGG categories in whole fecal metagenomes and viromes, excluding the IGC genes with unknown annotation. The color scale shows the square root transformed relative abundances. A.V., additional viromes; A.M., additional microbiomes (whole-genome extractions). Intra-class coefficient (ICC) for A.M. = 0.99; ICC for A.V. = 0.85; ICC microbiome-concordant twins = 0.69; ICC microbiome-discordant twins = 0.68.

I used the "large-insert-size library" for metagenomic assembly. In total, 107,307 contigs $\geq$ 500 bp were assembled (max: 79,863 bp, average: $1,186bp \pm 1,741$) (Appendix

.1). To assess the viromes' structure and composition, I built a matrix of the recruitment of reads against dereplicated contigs. The recruitment matrix included 14,584 contigs that were both long ( $1,300bp$) and well covered ( $5X$); these are referred to as "virotypes". Analysis of the recruitment matrix showed that each individual harbored a unique set of virotypes: 3,415 virotypes (23.41%) were present in only one individual, 413 virotypes (2.83%) were present in at least 50% of the individuals, and only 18 virotypes (0.1%) were present in all individuals.

To build the taxonomic layer, I attempted to annotate all 66,446 dereplicated and well-covered contigs, using a voting system approach that exploited the information in both the assembled contigs and their encoding proteins. On top of the voting system, I used HMMs to identify the contigs of two viral groups commonly found in the gut virome: (i) CrAssphages [84, 85] and (ii) the *Microviridae* family [86]. For CrAssphage, I found 19 contigs: 11 contigs clustered with the original crAssphage, 3 contigs grouped with the reference Chlamydia phage, and 5 contigs grouped with the reference IAS virus (Figure 2.7). For the Microviridae family, only 11 contigs had a previous taxonomic assignment, all belonging to the Gokushovirinae. I confirmed these and 23 more as Gokushovirinae, 54 as the candidate subfamily Alpavirinae, and 1 (one) contig as the candidate subfamily Pichovirinae (Figure 2.8).

After collating the voting system annotation and the HMM annotation, a total of 12,751 contigs (29,62%) were taxonomically assigned. Viromes were dominated by bacteriophages with only 6.42% of contigs annotated as eukaryotic viruses. As expected, most of the contigs (96.98%) were annotated as dsDNA viruses, while only 2.43% of contigs were ssDNA viruses. *Caudovirales* was the most abundant order, with its three main families represented: *Myoviridae* ($20.22\% \pm 4.83$), *Podoviridae* ($10.54\% \pm 3.27$), and *Siphoviridae* ($35.25\% \pm 7.19$). The crAssphage family constituted on average $13.26\% \pm 12.24\%$ of the contigs, reaching a maximum contribution of 55.80% in one virome, and *Microviridae* represented $3.87\% \pm 2.57$ of the viromes. Interestingly, we observed that *Phycodnaviridae* exceeded 1% of average abundance (average: $1.77\% \pm 1.12$) (Figure 2.9) and that contigs related to any nucleocytoplasmic large DNA viruses had a mean relative contribution of $3.99\% \pm 2.22$. The 18 contigs detected in all samples included 10 crAssphages, 1 *Microviridae*, 2 annotated as "unclassified *Myoviridae*", 2 "unclassified *Caudovirales*", and 3 without any annotation. Within a defined taxonomic profile for each sample, we looked for differences in composition between viromes at all taxonomic levels for microbiome-concordant and microbiome-discordant twin pairs. There were no significant differences between groups for any taxa at the order and family levels, including CrAssphages (Figure 2.10).

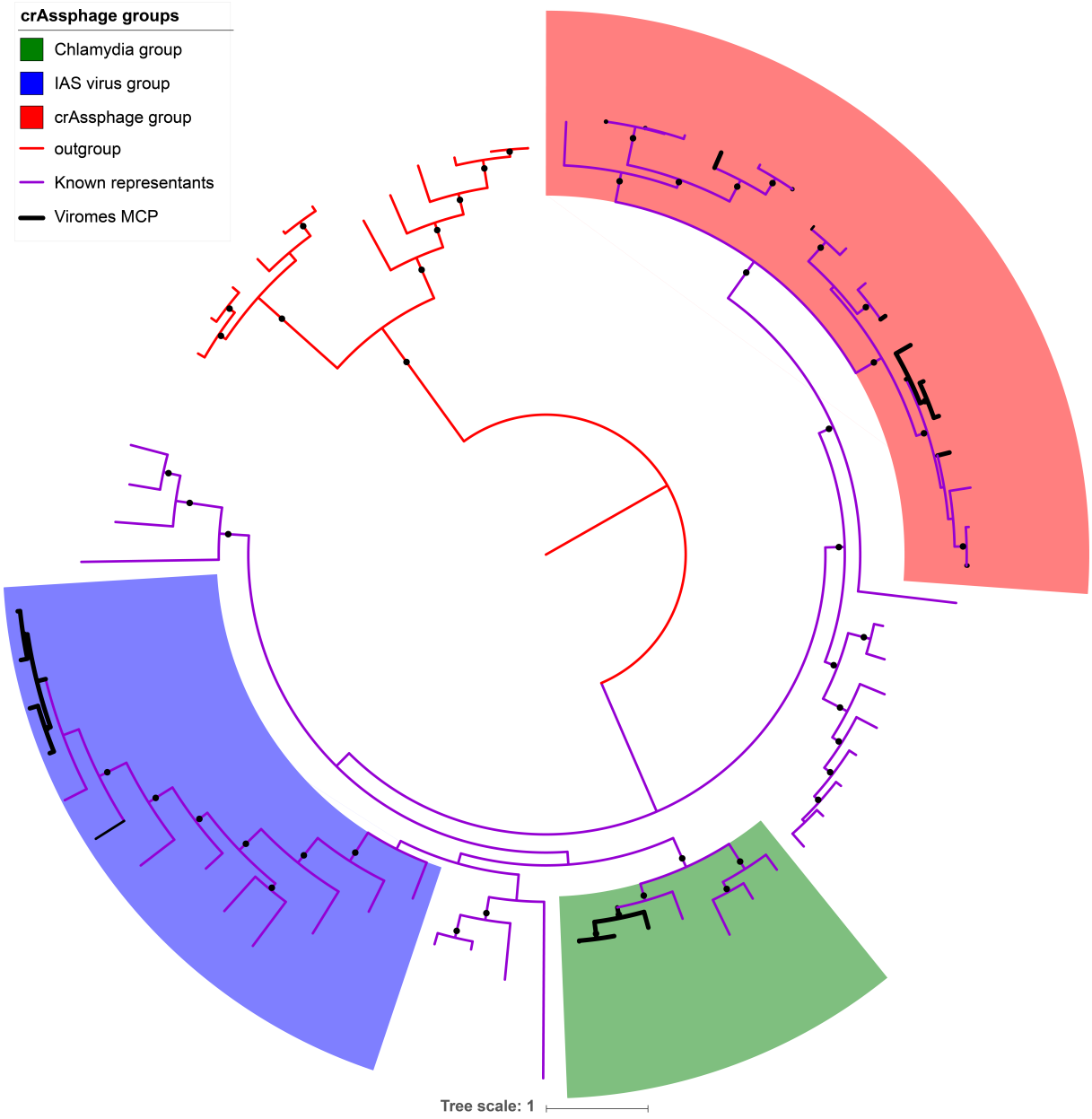


**Figure 2.7: CrAss-like phages in the human gut virome of MZ twins.** Maximum likelihood phylogenetic analysis of the MCP protein of crAss-like phages found in the 42 MZ viromes. Reference sequences are in purple, outgroup sequences are in red. The different MCP proteins found in this work are labeled in black. Circles indicate bootstrap values above 70%. Scale: Average substitutions per site.

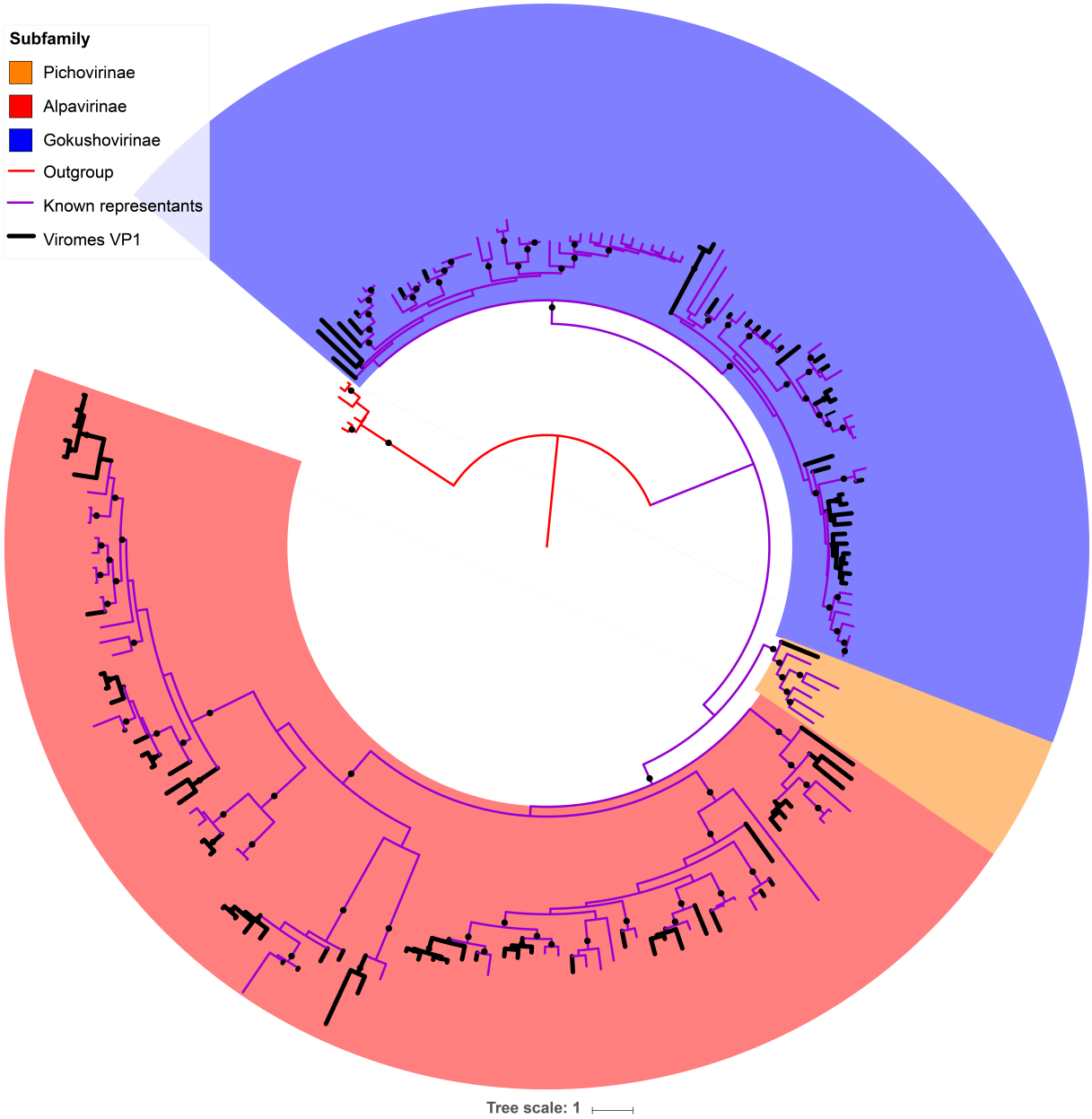


**Figure 2.8: *Microviridae* members in the human gut virome of MZ twins.** Maximum likelihood phylogenetic analysis of the VP1 protein of Microviridae phages found in the 42 MZ viromes. Reference sequences are in purple, outgroup sequences are in red. The different VP1 proteins found in this work are labeled in black. Circles indicate bootstrap values above 70%. Scale: Average substitutions per site.

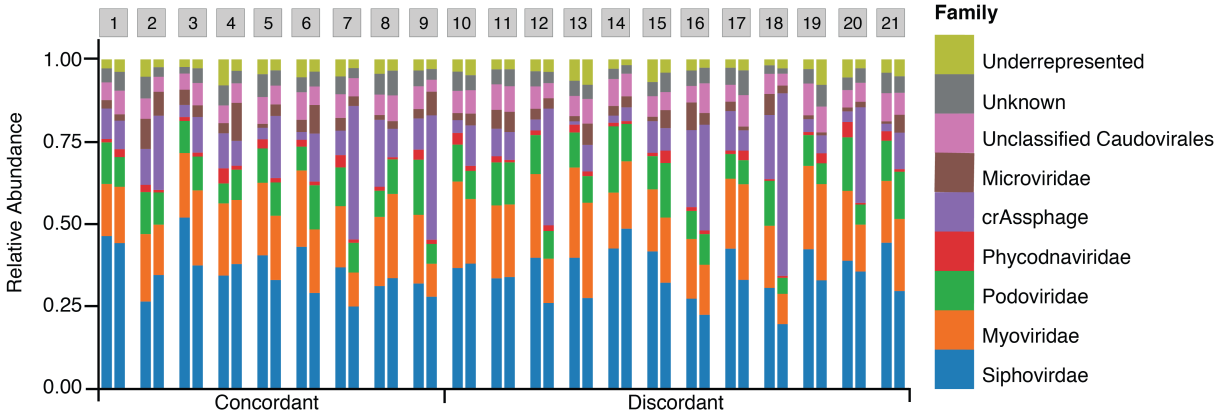


**Figure 2.9: Virome taxonomic composition** Comparison of the taxonomic profiles at the family level for the 21 MZ twin pairs. Microbiome-concordant twins: 1–9; Microbiome-discordant twins: 10-21.

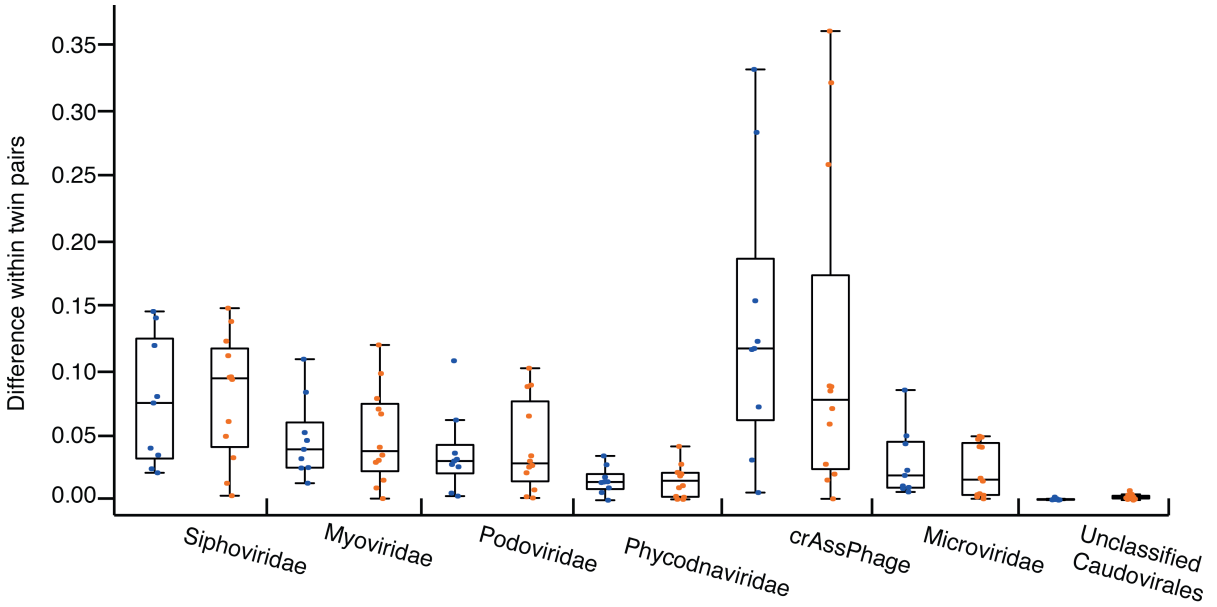


**Figure 2.10: Relative abundance comparison between microbiome-concordant/discordant twins.** Differences of the relative abundances of each viral family for microbiome-concordant (blue points, n = 9) and microbiome-discordant (orange points, n = 12) twin pairs.

In conclusion, the viral metagenomes of MZ twins presented a functional profile that differs from the general fecal microbiome -as shown in the functional layer-, are dominated by bacteriophages -taxonomy layer-, and are highly unique to each individual -virotypes layer-. Since the data supports that viromes were indeed enriched in VLP, I proceeded with calculating and comparing viral diversity patterns and bacterial diversity patterns.

### 2.2.3 Virome diversity correlates with microbiome diversity

The cellular fraction of the gut microbiota forms a very dense microbial ecosystem ($10^{11}$ - $10^{12}$ cells per gram of feces) [22] in which viruses are found in about equal proportion (between $10^{9}$ to $10^{12}$ VLP per gram of feces) [28, 29]. Also, the virome and microbiome display common diversity patterns across hosts, such as high interpersonal differences and relative stability over time [32]. This observation could be driven by host genetic relatedness, similar microbiomes, and other shared environmental factors. Here, I aimed at measuring the variance in human gut viral diversity due to the human gut microbiome. To control for the host genotype, I employed viral community profiles derived from viromes of microbiome-concordant and microbiome-discordant MZ twins.

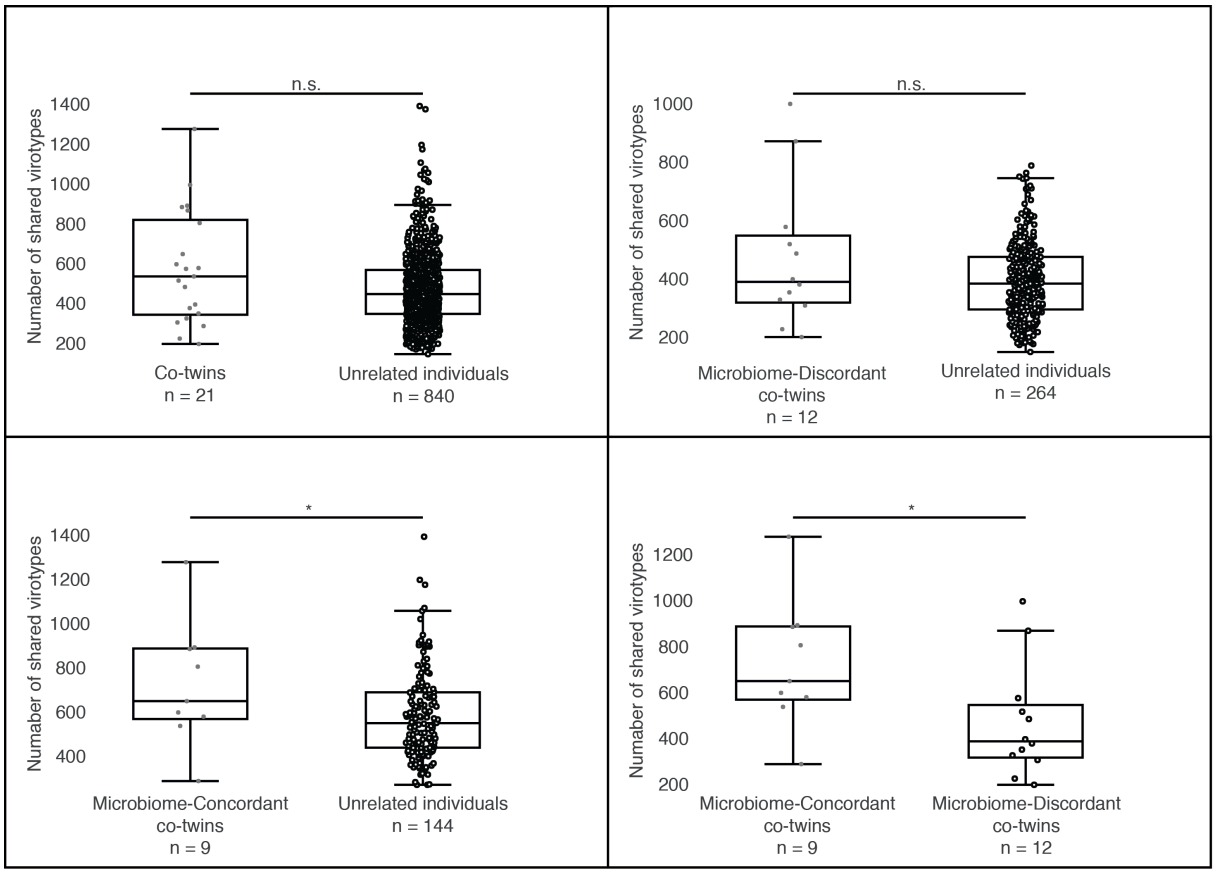


**Figure 2.11: Microbiome-concordant twins shared more virotypes.** Box plots showing the distribution of the number of shared virotypes between different groups made from the 21 MZ co-twins. Upper left: all co-twins vs unrelated individuals. Upper right: microbiome-discordant co-twins vs unrelated individuals in the same group. Lower left: microbiome-concordant co-twins vs unrelated individuals in the same group. Lower right: microbiome-concordant twins vs microbiome-discordant twins. Mann-Whitney's U test. "*": $p < 0.05$; n.s: no significant difference.

First, I tested whether co-twins share more virotypes than unrelated individuals and found they do not (Mann-Whitney U test, p = 0.074). Then, I assessed microbiome-concordant and microbiome-discordant twin pairs separately. I found that microbiome-concordant twins shared more virotypes than microbiome-discordant twins (Mann-Whitney U test, p = 0.015) (Figure 2.11). As expected, microbiome-concordant twins did share more virotypes than unrelated individuals (Mann-Whitney U test, p = 0.048). However, that was not the case for microbiome-discordant twins (Mann-Whitney U test, p = 0.254).

Second, I evaluated the relationship between the virome's diversity and the microbiome's diversity. I examined virome $\alpha$-diversity and $\beta$-diversity using the community profiles previously recovered based on the three annotation layers: i) virotypes, ii) taxonomically annotated contigs, and iii) annotated genes from short reads.

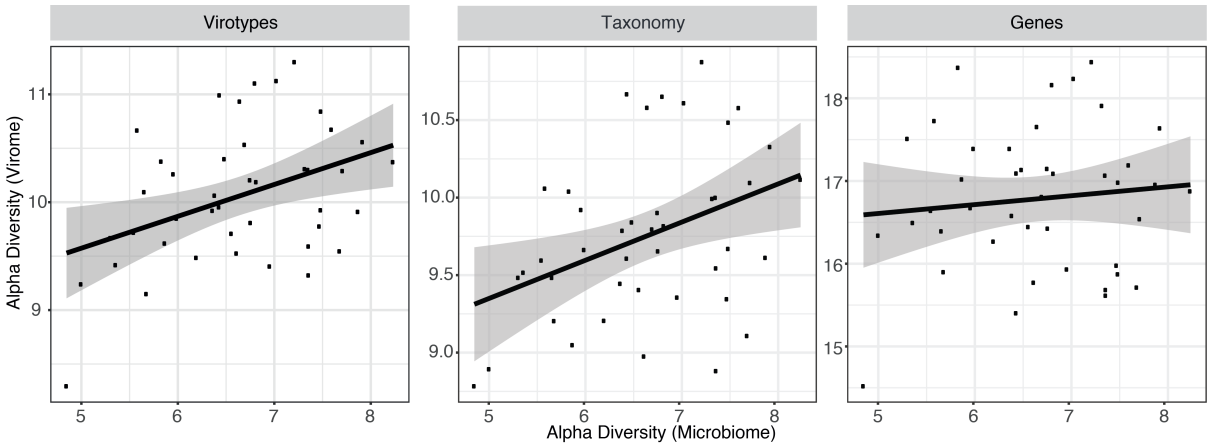


**Figure 2.12: Virome diversity correlates with microbiome diversity.** Correlation of Shannon $\alpha$-diversity of viromes to Shannon $\alpha$-diversity of microbiomes (n = 42). Best-fit lines with 95% confidence intervals from linear regression are plotted.

The $\alpha$-diversities of the virome and the microbiome were positively correlated for two of the three annotation layers (Virotypes: Pearson correlation coefficient = 0.406, m = 0.3, p = 0.007, R 2 = 0.165; Taxonomy: Pearson correlation coefficient = 0.389, m = 0.25, p = 0.010, R2 = 0.151; Genes: Pearson correlation coefficient = 0.105, m = 0.11, p = 0.506, R2 = 0.01) (Figure 2.12). Then, I used annotated contigs to explore how virome $\alpha$-diversity varies across subgroups of viruses. The subgroups were defined according to the type of viral genetic material (ss or dsDNA) and type of viral host (eukaryotes or prokaryotes). The results showed that the diversity of eukaryotic viruses does not correlate with the microbiome diversity. In contrast, bacteriophage diversity positively correlated with microbiome diversity. Both observations were independent of the genome's molecule (ssDNA eukaryotic viruses: Pearson correlation coefficient = 0.027, m = 0.034, p = 0.863,

R2 = 0.000751; ssDNA bacteriophages: Pearson correlation coefficient = 0.394, m = 0.35, p = 0.009, R2 = 0.155; dsDNA eukaryotic viruses: Pearson correlation coefficient = 0.143, m = 0.15, p = 0.368, R2 = 0.020; dsDNA bacteriophages: Pearson correlation coefficient = 0.400, m = 0.25, p = 0.008, R2 = 0.16) (Figure 2.13).

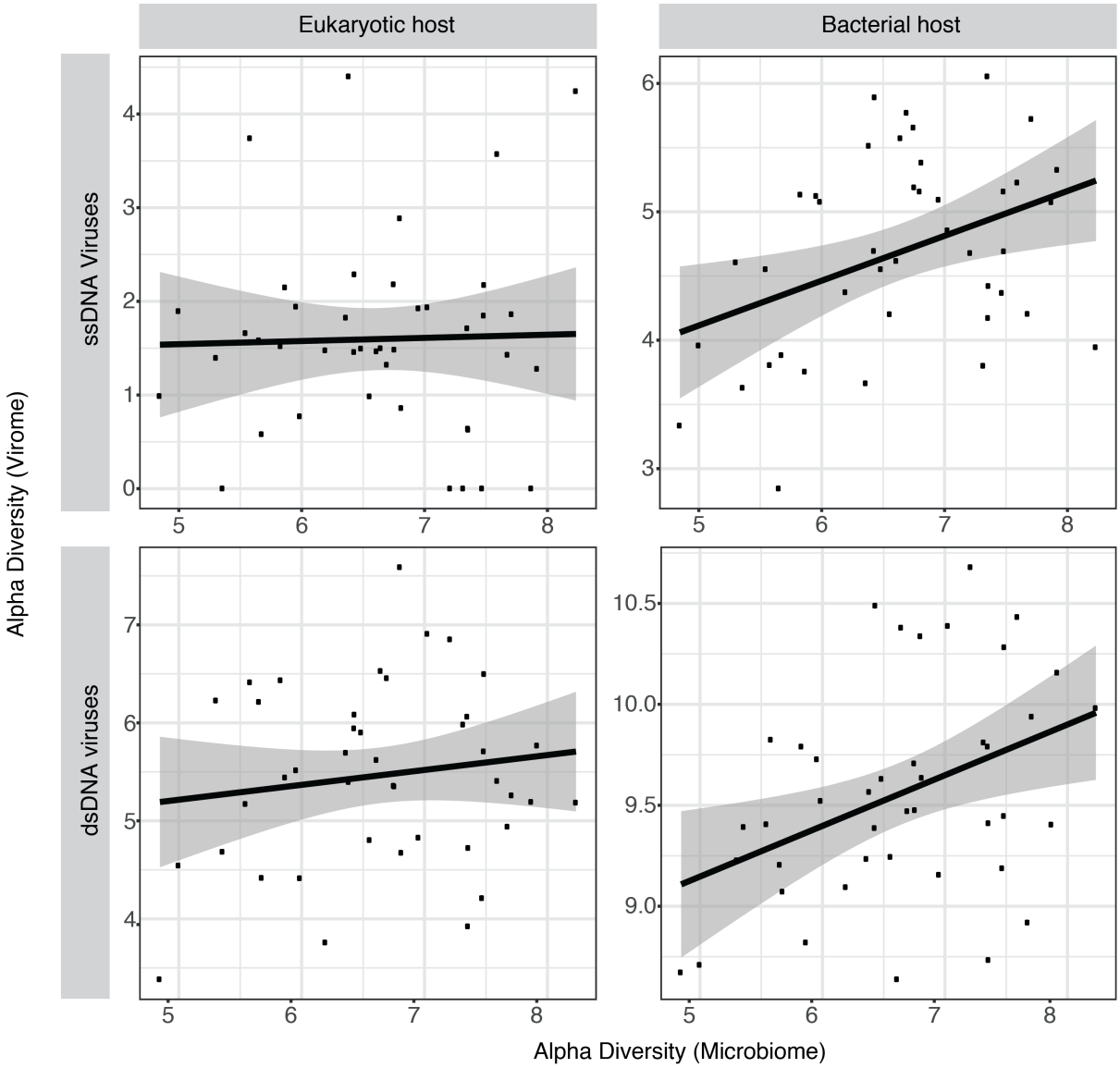


**Figure 2.13: Bacteriophages drive virome-microbiome diversity correlation** Correlation of the Shannon $\alpha$-diversity of the virome, calculated from contigs annotated as ssDNA eukaryotic viruses, ssDNA phages, dsDNA eukaryotic viruses, and dsDNA phages, to Shannon $\alpha$-diversity of the microbiome (n = 42). Best-fit lines with 95% confidence intervals from linear regression are plotted.

Regarding the $\beta$-diversity assessment, I observed that concordant twins had lower virome $\beta$-diversity compared to discordant twins using Hellinger distances (Virotypes:

Mann-Whitney U test, p = 0.04; Taxonomy: Mann-Whitney U test, p = 0.02; Genes: Mann-Whitney U test, p = 0.32) (Figure 2.14). Finally, I compared the virome and microbiome pairwise distances among twins and among all individuals. The pairwise distance matrices showed a positive correlation between virome and microbiome $\beta$-diversity, not only within twin pairs (Virotypes: Pearson correlation coefficient = 0.522, m = 0.188, p = 0.015, R2 = 0.1508; Taxonomy: Pearson correlation coefficient = 0.512, m = 0.186, p = 0.017, R2 = 0.224; Genes: Pearson correlation coefficient = 0.53, m = 0.182, p = 0.012, R2 = 0.248) but also across all individuals (Virotypes: Pearson correlation coefficient= 0.382, m = 0.167, p = 0.0005, R2 = 0.157; Taxonomy: Pearson correlation coefficient = 0.266, m = 0.140, p = 0.003, R2 = 0.0796; Genes: Pearson correlation coefficient = 0.344, m = 0.162, p = 0.0009, R2 = 0.123) (Figure 2.15). These results show that regardless of genetic relatedness between hosts, individuals with more similar microbiomes harbor more similar viromes.

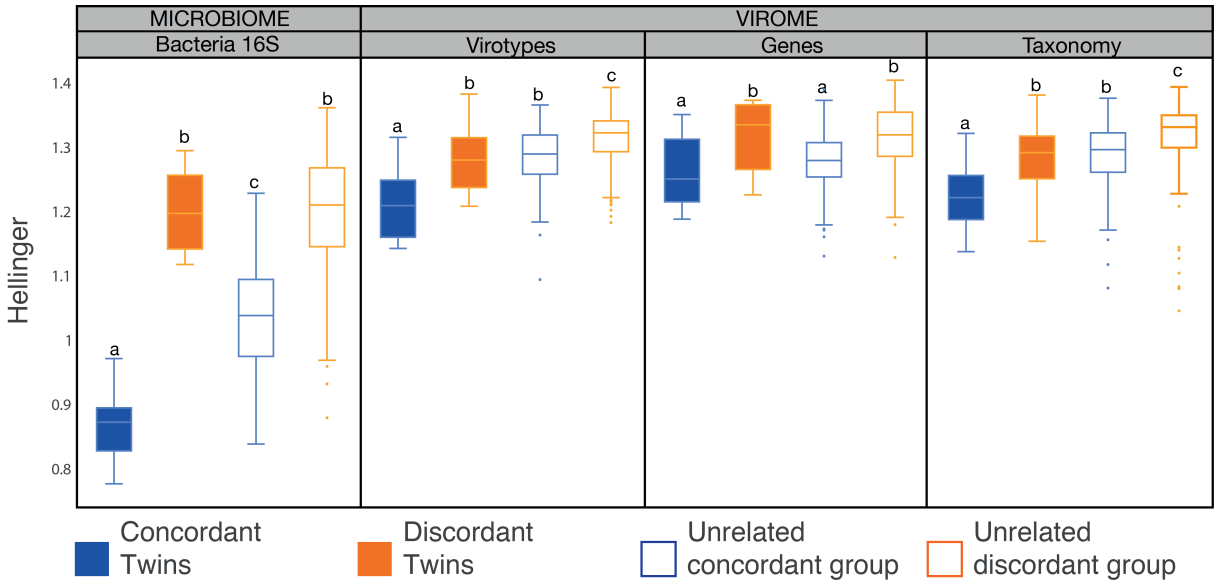


**Figure 2.14: Virome $\beta$-diversity patterns mirror microbiome $\beta$-diversity in MZ twins.** Boxplots show the distribution of Hellinger distances for microbiomes and viromes, according to the three different layers of information recovered (virotypes, genes, and taxonomy), for microbiome-concordant twins (solid blue, $n = 9$), microbiome-discordant twins (solid orange, $n = 12$), unrelated samples within the microbiome-concordant twins (blue outline, $n = 144$), and unrelated samples within the microbiome-discordant co-twins (orange outline, $n = 264$). Significant differences between means are denoted with different letters.

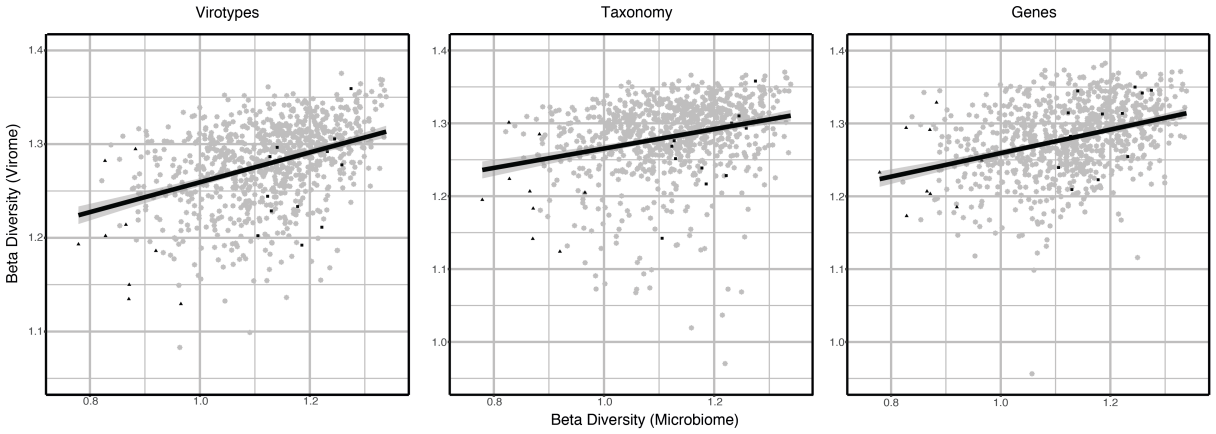


**Figure 2.15: Virome $\beta$-diversity correlates with microbiome $\beta$-diversity** Correlation between virome $\beta$-diversity and microbiome $\beta$-diversity (n=840) according to the three different layers of information recovered (virotypes, genes, and taxonomy). Lines describe linear regressions of pairwise distances among all individuals. Triangles indicate microbiome-concordant twins and squares indicate microbiome-discordant twins.

This work shows that gut microbiome richness and diversity correlate to virome richness and diversity. The mechanisms underlying this association remain to be resolved for the human gut. This relationship may be beneficial to consider when designing future studies on the human gut virome. For instance, the diversity of the reference microbiome may be meaningful for balancing between groups before assessing viromes' diversity.

## 2.3 Discusion

### 2.3.1 Putative contaminants are common in viromes

The analysis of viromes from microbiome-concordant and microbiome-discordant MZ twins showed that despite the high variation in the gut viromes between individuals, and regardless of host relatedness, the more dissimilar their microbiomes, the more dissimilar their viromes. By analyzing viromes from MZ twins, the gut microbiome was the single variable determining virome diversity. An important step in viromics is to ensure the quality of the VLP purifications and remove potential bacterial contaminants. Even though only a few reads mapped to putative contaminants, I was able to identify some taxa as putative contaminants. Bacterial contamination is a common issue in viral metagenomes. Zolfo and collaborators found that bacterial contamination is intrinsic to nearly all publicly available virome datasets, is independent of the VLP purification method, and varies between samples within the same study [87]. In addition, Roux et al. analyzed viromes

from various environments and identified 16S rRNA sequences within all human intestinal viromes. In contrast, only aquatic viromes lacked 16S rRNA sequences; suggesting that contamination could be related to the difficulty in purifying viral particles from complex matrices such as fecal samples. However, the identification of cellular DNA in viruses from several aquatic environments supports that contamination is not the only cause of microbial sequences in viromes [74]. Consistent with this observation, I found that putative contaminants in the virome of the MZ twins were not the most abundant members of their microbiome. In addition, I identified certain bacterial species as putative contaminants in unrelated studies of the human gut virome. While the exact reason for the high coverage of particular bacterial genomes with virome reads is still unclear, possible explanations are: vesicle production [88], transducing phages [89, 90], and gene transfer agents [91, 92].

Ensuring the remotion of possible sources of contamination is a mandatory step in metagenome analysis. This work and the discussed references constitute a call not to avoid bacterial contamination assessment in viral metagenomes. Further investigation of horizontal gene transfer mediated by any of the mentioned mechanisms and the characterization of bacterial mobilomes may help shed light on this question. Particularly for the taxa identified as putative contaminants in unrelated studies.

### 2.3.2 HMMs are promising tools for viromes' characterization

After the quality control assessment and the dedicated removal of putative bacterial contaminants, I characterized the virome of MZ twins. In general, the composition of the viromes described here was similar to what has been previously reported for adult fecal viromes: viromes are highly unique to each individual, are dominated by bacteriophages, and their functional profile is dominated by the unknown [36, 75, 32].

From the annotated fraction of the virome, the order *Caudovirales* and its families *Siphoviridae*, *Myoviridae*, and *Podoviridae*, along with CrAssphage, were the dominant phages in all samples. However, the *Caudovirales* order and its families were recently abolished after the introduction of the new virus taxonomy [51]. This change will help to address better the genetic diversity of tailed phages by increasing the achievable resolution of the human gut virome taxonomic profile. Besides tailed phages, I also recovered contigs of *Microviridae*, a family of ssDNA phages, and contigs that map to *Phycodnaviridae*, a family of nucleocytoplasmic large DNA viruses that infect marine or freshwater algae. While the formers are conspicuous members of the human gut virome [9, 93], the latter are not. Nonetheless, they have been increasingly reported as members of the human gut [94, 95, 96, 97].

To increase the proportion of annotated contigs, I used HMMs. Because each type of virus of interest requires its own HMM, and the families *Podo-*, *Myo-*, and *Siphoviridae* are not monophyletic groups [53], I applied this method to two groups of interest in the human gut: CrAssphages and *Microviridae*. When applied to crAssphages, the HMM retrieved only 9 different crAssphages (according to their MCP). Regardless, these few crAssphages accounted for more than half of the reads in one virome. In fact, CrAssphage is the most abundant viral group of the human gut [98]. CrAssphages, recognized since 2021 as the order *Crassvirales*, are a group of widespread human gut viruses that may have coevolved with humans. Furthermore, its phylogeography clusters within countries, cities, and individuals [41]. This explains why despite the high diversity and uniqueness of each virome described here, I found 10 crAssphages present in all samples. I also used HHMs to identify contigs of the family *Microviridae*. I confirmed the presence of diverse members of the subfamily *Gokushovirinae* and the proposed subfamily *Alpavirinae*. Although there is evidence that described *Alpavirinae* genomes constitute a third group of the *Microviridae* family [99, 100], they correspond to prophages, which are difficult to integrate into the viral taxonomy as they are difficult to isolate.

Using HMMs to annotate viral contigs proved to be a successful method in identifying contigs of interest. Several databases of HMMs of viral orthologs are currently available [63, 8, 64]. However, none of the databases provide a set of taxonomic markers, making them difficult to use for high-throughput characterization of viromes. Recently, viral taxonomy has changed to better address viral diversity [51]. In particular, tailed bacteriophages are no longer grouped into the *Myoviridae*, *Podoviridae*, and *Siphoviridae* families [53]. Instead, several orders and families have been defined under the class *Caudoviricetes*.

Contigs generated from the *de novo* assembly are often very fragmented and rarely longer than a few kilobases [101]. For this reason, I adopted a "multiple-layer annotation" to characterize the viromes. The functional characterization based on short reads allowed me to differentiate viromes from bulk fecal extractions. The definition of virotypes allowed me to recognize viruses shared between subjects. Finally, I retrieved the taxonomic annotation of contigs, which highlighted the predominance of tailed bacteriophages in the human gut. Furthermore, HMMs succeed in retrieving contigs that belong to crAssphage or *Microviridae*. The definition of new taxonomic groups and reference genomes may facilitate the identification of characteristic HMMs for different taxonomic groups and improve the resolution at which a virome can be characterized.

### 2.3.3 Virome diversity reflects microbiome diversity

Twins, like other siblings, generally have more similar gut microbiomes than unrelated individuals [102, 103, 104, 105]. Nonetheless, within a population of MZ twin pairs, the range of within-twin pair differences in microbiomes can be as great as for DZ twins [72]. I analyzed the viromes of microbiome-concordant and microbiome-discordant MZ twins. The results showed that despite the high variation in gut viromes between individuals, and regardless of host relatedness, the more dissimilar the microbiomes, the more dissimilar the viromes. Notably, the bacteriophage component of the virome drove this pattern.

Previous studies of the viromes of infant twin pairs showed that the viromes of twins were more similar than those of unrelated individuals, suggesting that shared host genotype and/or environment are key [76, 106]. In contrast, an earlier study of the virome of adult twins showed that adult twins did not have more similar viromes than unrelated individuals [32]. In light of my results, this was likely a power issue. Indeed, I observed that regardless of whether twins were concordant or discordant for their microbiomes, MZ twins had more similar viromes (virotypes and taxonomy) than unrelated individuals. In addition, regardless of genetic relatedness between hosts, individuals with more similar microbiomes harbored more similar viromes.

The previously reported higher virome similarity in infants compared to adult twins has been related to the fact that infants might have a greater shared environment compared to adults [76], particularly in terms of diet. Minot et al., have also shown that individuals on the same diet have more similar gut viromes than individuals on dissimilar diets [36]. Diet is a strong driver of daily microbiome fluctuation [107, 35, 108]. Thus, the effect of diet on the virome might be mediated by the microbiome. Unfortunately, the dataset I had access to did not include a control for diet. Accordingly, the microbiome discordance observed could be due to twins eating differently around the time of sampling. Regardless of what underlies the variance in microbiome diversity, virome diversity is strongly associated with it.

The relationship between virome and microbiome richness has not been addressed directly in adults before. I found that the $\alpha$-diversities of the microbiome and virome were positively correlated for two of the three layers of information describing virome diversity. Specifically, this pattern was observed for virotypes and taxonomy but not for genes. Since virome genes were specifically enriched in two categories, namely "genetic information processing" and "nucleotide metabolism", I would not expect differences in the diversity of virome genes between subjects. The taxonomic annotation layer showed that mainly bacteriophages and not eukaryotic viruses were driving this $\alpha$-diversity correlation pattern.

The positive relationship between virome and microbiome $\alpha$-diversity suggests that

greater availability of hosts results in a greater diversity of viruses. Nevertheless, it is unclear if the virome drives the microbiome diversification or the other way around. Longitudinal studies combining the collection of both virome and microbiome, together with a better understanding of phage-host interactions, and increased availability of characterized reference viral genomes, would help interpret the patterns observed in the human gut virome.

## 2.4 Methods

### 2.4.1 Assessment of Bacterial Contamination

A set of 8,163 finished bacterial genomes was retrieved from the NCBI FTP on 21 February 2017. Reads per sample were mapped against this bacterial genomes dataset using Bowtie2 v.2.2.8 [109] with the following parameters: –local –maxins 800 -k=3. Genome coverage per base was calculated considering only reads with a mapping quality above 20 using view and depth Samtools commands v.1.5 [110]. Next, genome coverage was averaged for 100Kbp bins. We observed that evenly covered genomes had a median bin coverage of at least 100; those genomes with a median bin coverage greater than 100 were considered as contaminants. The reads mapping to those genomes were removed. Bacterial genomes can have one or more prophage(s) in their genomes [111]; bursting events of those prophages can occur, generating several VLPs. As a conservative measure to avoid the loss of reads originating from prophages and not the bacterial genome per se, bins with a coverage over three standard deviations of the bacterial mean coverage were also identified and catalogued as prophages-like regions. Reads mapping to potential contaminant genomes were tagged as "contaminants" and removed from further analysis while reads mapping to high coverage bins were tagged as "possible prophages".

A matrix of the abundance of each potential contaminant per sample was built using an in-house Python script and normalized by RPKM. In parallel, from Goodrich et al. data [72], the relative abundance of each OTU was recovered and summarized at the species level using summarize_taxa.py qiime script. The Spearman rank order correlation between relative abundances of contaminants and their corresponding 16S rRNAs data was calculated for species in both sets.

### 2.4.2 Functional profiles

The joined and trimmed reads from the "short-insert-size library" were mapped onto IGC, an integrated catalog of reference genes in the human gut microbiome [82] by

BLASTX using DIAMOND v.0.7.5 [112] with maximum e-value cutoff 0.001, and maximum number of target sequences to report set to 25.

After the mapping onto IGC, an abundance matrix was generated using an in-house Python script. The matrix was then annotated according to the KEGG annotation of each gene provided by IGC. The annotated abundance matrix was rarefied (subsampling without replacement) to 2,000,000 read hits per sample. The KEGG functional profile was then generated using QIIME 1.9 [113] using the command summarize_taxa_through_plots.py. The Intraclass Correlation Coefficient of the functional profiles for each group (additional microbiomes, additional viromes, viromes of microbiome-concordant samples and viromes of microbiome-discordant samples) was calculated using the Psych R package.

### 2.4.3 De-novo assembly

Reads from the "large-insert-size library" that remain paired (forward and reverse) after the trimming step were assembled using the Integrated metagenomic assembly pipeline for short reads (InteMAP) [114] with insert size $325bp \pm 100$. Each sample was assembled separately. After the first run of assembly, all clean reads were mapped to the assembled contigs using Bowtie2 v.2.2.8 [109] with the following parameters: –local –maxins 800. The pairs of reads that aligned concordantly at least once were then submitted for the second run of assembly by InteMAP. Contigs larger than 500 bp from all samples were pooled together and compared all vs all, using an in-house Perl script. From this analysis, it was possible to identify potential circular genomes, and to dereplicate contigs that were contained in over 90% of their length within another contig.

The recruitment of reads to the dereplicated metagenomic assemblies was used to build an abundance matrix, applying a filter of coverage and length as recommended in Roux et al. [74]. Reads (not tagged as contaminants in the previous step) were mapped to dereplicated contigs using Rsubread v.1.28.0 [115]. Mapping outputs were parsed using an in-house Python script into an abundance matrix that was normalized by reads per kilobase of contig length per million sequenced reads per sample (RPKM) and transformed to $Log_{10}(x+1)$, $x$ being the normalized abundance. Contigs with a normalized coverage bellow 5X were excluded. Finally, a filter on contig length was applied to obtain virotypes. A length threshold was chosen as the elbow of the decay curve generated when plotting the number of contigs as a function of length, which occurred at a length of 1,300 bp.

### 2.4.4 HMM annotation

Independent HMM-profiles were built to identify crAss-like contigs and *Microviridae* contigs. To build the HMM-crAsslike profile, sequences for the Major Capsid Protein (MCP) of the proposed crAss-like family [85] were retrieved from `ftp.ncbi.nih.gov/pub/yutinn/crassphage_2017/`. Multiple sequence alignments (MSA) were done using MUSCLE v.3.8.31 [116] and inspected using UGENE v.1.31.0 [117]; positions with more than 30% of gaps were removed. Finally, the HMM-crAsslike profile was built using hmmbuild from the HMMER package v.3.1b2 (`hmmer.org`) [118]. For the *Microviridae* case, all HMM-profiles for the viral protein 1 (VP1) developed by Alves et al. [119] were adopted. Predicted proteins of the assembled contigs were queried for matching the HMM-profiles using hmmsearch [118]. Matching proteins with an e-value below $1X10^{-5}$ were considered as true homologs but only proteins between the size rank of the reference proteins (crAsslike MCP: 450-510 residues; *Microviridae*: 450-800 residues), a coverage of at least 50% and a percentage of identity of at least 40% to at least one reference sequence were used for further analysis. Coverage and identity percentages were determined with a BLASTp search of the true homologues against the reference sequences. True homologues passing the filters mentioned above were used in phylogenetic analysis. Reference and homologous sequences were aligned using MUSCLE v.3.8.31 [116]and sites with at least 30% of gaps were removed using UGENE v.1.31.0 [117]. A maximum-likelihood phylogenetic analysis was done using RAxML v.8.2.4 [120], the best model of evolution was obtained with prottest v.3.4.2 [121] and support for nodes in the ML trees were obtained by bootstrap with 100 pseudoreplicates.

### 2.4.5 Taxonomic profiles

To infer the taxonomic affiliation of the assembled VLPs, genes were predicted from all assembled contigs larger than 500 bp using GeneMarkS v.4.32 [122]. The amino acid sequence of the predicted genes was then used in a BLASTp search against the NR NCBI viral database using DIAMOND v.0.7.5 [112] with maximum e-value cutoff 0.001 and a maximum number of target sequences to report set to 25. Using the BLASTp results, the taxonomy of each gene was assigned by the lowest-common-ancestor algorithm in MEtaGenome ANalyzer (MEGAN5) v.5.11.3 [123] with the following parameters: Min Support: 1, Min Score: 40.0, Max Expected: 0.01, Top Percent: 10.0, Min-Complexity filter: 0.44. Independently, the taxonomy annotation of each contig was obtained using CENTRIFUGE v.1.0.4 [124] against the NT NCBI viral genomes database. The final taxonomic annotation of each contig was then assigned using a voting system where the

taxonomic annotation of each protein and the CENTRIFUGE annotation of the contig were considered as votes. With all the possible votes for a contig, an N-ary tree was built and the weight of each node was the number of votes including that node. The taxonomic annotation of a contig will be the result of traversing the tree passing through the heaviest nodes with one consideration: if all children-nodes of a node have the same weight the traversing must be stopped. The taxonomic profile was considered as a subset of the recruitment matrix containing all contigs annotated either by the voting system or annotated through the HMM profiles.

### 2.4.6 Diversity indexes

The Shannon diversity index within-samples ($\alpha$-diversity) and the Hellinger distance within twin pairs ($\beta$-diversity) were calculated using diversity and vegdist functions of Vegan R package for all three abundance matrices generated (function, taxonomy and read recruitment matrices). Correlations between virome $\alpha$-diversity and microbiome $\alpha$-diversity were measured using the Pearson correlation coefficient. Correlations between viromes $\beta$-diversity and the microbiomes $\beta$-diversity were computed with a the Mantel test using the Pearson correlation coefficient. Additionally, the $\beta$-diversity between concordant MZ co-twins was compared to the $\beta$-diversity between discordant MZ co-twins; p-values were calculated using a Mann-Whitney U test.

# Chapter 3

# An expanded diversity of transposable phages

## 3.1 Motivation

Annotated and characterized viruses constitute a small fraction of sequences stored in public databases. In contrast, UViGs make up more than 95% of the current diversity in public databases [20]. Characterized viruses do not have much resemblance to metagenomic assemblies [29], thus, the release of new viral metagenomes and UViGs is not translated into a better description of the viral community of interest. I decided to address this limitation by examining databases of metagenomic assemblies in the search for transposable phages.

As I showed in chapter 2, putative contaminants are common in viromes. Particularly, in viromes from complex matrices and high microbial diversity like the human gut [74]. Additionally, my results underscored how taxa relative abundance or limitations of the experimental procedures do not account for bacterial contamination in human gut viromes. One factor that might contribute to the detection of bacterial sequences in viromes is generalized transduction, i.e, the transfer of random fragments of the host's genomic DNA by phages from an infected bacterium to another [125]. Transposable phages are generalized transducing phages [126] and could be contributing to the transfer of bacterial DNA to viromes. Nevertheless, their known diversity is reduced to a few genomes [127]

In this chapter, I used comparative genomics and remote homologous searches to uncover the diversity of transposable phages in databases of UViGs. I characterized their genomic diversity based on phylogenetic analyses and delineated the first family of transposable phages of the human gut.

## 3.2 Results

### 3.2.1 Transposable phages are found by thousands in databases of metagenomic assemblies

Transposable phages, proposed family *Saltoviridae* [127], are dsDNA phages capable of generalized transduction, the transfer of any portion of the infected bacterium's DNA to another bacterium [126]. Additionally, transposable phages replicate their genome using replicative transposition, which leads to mutations and genomic rearrangements of their host's genome [128]. Despite their potential role in the evolution of their hosts and the impact that transposable phages have had on genetics and genetic engineering [129], the number of complete genomes of transposable phages is limited to a few dozen. The latest genomic analysis compared 26 complete genomes of transposable phages available at the time [127]. Although those 26 phages only infect members of $\alpha$, $\beta$, and $\gamma$-proteobacteria, studies based on predicted prophages extended the predicted host range of transposable phages to the classes $\delta$- and $\epsilon$-proteobacteria [130], and the phylum Firmicutes [89]. Given the importance of transposable phages, their suspected widespread nature, and the lack of sequenced genomes, here, I searched for new instances of complete genomes of transposable phages in databases of metagenomic assemblies.

The only complete genome sequence of a transposable phage infecting a Firmicutes is the one of Mushu phage (Genbank accession: MG711460.1 infecting *Faecalibacterium prausnitzii*) [131]. I used Mushu's genome as the query of a tblastx against all phages in NCBI's nucleotide collection and combined the best 100 hit genomes with the 26 genomes previously analyzed by Hulo and collaborators. To remove the probable redundancy inserted when both datasets were combined, I dereplicated the set of genomes at the species level: 95% average nucleotide identity (ANI). A total of 77 different genomes were further analyzed (Apendix .2).

To determine whether a genome was or not a transposable phage, the annotation and genomic organization of each genome were examined. For this purpose, I integrated the annotation provided in the GenBank files with protein annotations from eggNOG-Mapper [132] and ViPhOGs [8]. Working with these three sources of information (GenBank, eggNOG-Mapper, and ViPhOGs) I curated the annotation of proteins that share the same ViPhOG assignation, which facilitated the selection of transposable phage genomes as it significantly reduced the number of hypothetical proteins in the genomes (Mann-Whitney U test, $p = 0.00011$). A set of conserved genomic features is already defined for transposable phages. They correspond to proteins involved in replicative transposition, as well as proteins involved in the head, neck formation, and headful package mechanism

[127, 133]. I checked genomes to have those conserved features among their annotated proteins to select them as transposable phages. I identified 60 genomes. These transposable phages grouped into 25 genera according to their genomic identities, representing a small expansion compared to the last review by Hulo and collaborators (26 genomes; 14 genera; 25 species) [127]. The genome size of all transposable phages stored at NCBI ranged between 24,971 and 46,148 (median: 37,297, IQR: 36,696 - 38,301), and their mean gene density was 1.46 genes/kb $\pm$ 0.10. Regarding the phages' host, 58 out of 60 transposable phages infect members of Proteobacteria ($\alpha$-proteobacteria: 5, $\beta$-proteobacteria: 3, and $\gamma$-proteobacteria: 50). Only Mushu and Bal-Mu infect members of Firmicutes: Faecalibacterium and Bacillus, respectively (Appendix .2). These results suggest that the thriving diversity predicted by the analysis of prophages back in 2013 [89] still awaits to be isolated.

To exploit public metagenomic datasets in search of novel diversity of transposable phages, first, I established a set of marker-ViPhOGs of transposable phages and then, used those markers to explore two comprehensive databases of viral metagenomic assemblies, IMG/VR [134] and GPD [39]. As expected, the conserved proteins of transposable phages had at least one ViPhOG assigned to them. In the case of the portal protein, virion morphogenesis protein, GemA, and adapter protein from all NCBI transposable phages, each had only one ViPhOG assigned to them. That was not the case for the transposase A, transposase B, and minor head protein, where each protein had at least three ViPhOGs assigned to them, highlighting the modularity of phages. In addition to the known conserved proteins the major head protein, protease(I), and Mor were added to the list of marker-ViPhOGs since their prevalence across all NCBI transposable phages was high. In total, I used 22 marker-ViPhOGs associated with 10 conserved genomic features of transposable phages (Table 3.1) to screen IMG/VR [134] and GPD [39] databases. After searching target proteins for the marker ViPhOGs throughout the predicted coding sequences of metagenomic assemblies, 1,997 metagenomic assemblies coded for 6 out of 10 genomic features of transposable phages and were further analyzed. I re-annotated all candidates using Multiphate [135] for gene calling, and eggNOG-mapper and ViPhOGs for function annotation. Based on the re-annotation, I identified 4 additional ViPhOGs for transposase A (ids: 17094, 27076, 588, 6997), and 4 additional ViPhOGs for transposase B (ids: 30042, 16145, 19833, 289). After checking for the presence/absence of marker-ViPhOGs, I kept 1,002 genomes having marker-ViPhOGs for at least 8 out of 10 conserved features. These will be referred to as putative transposable phages.

The set of putative transposable phages outreaches the diversity of transposable phages stored in NCBI. According to their genome identity, the set of putative transposable

phages extends to 765 species and 440 genera. According to the information on the assemblies' quality, 589 genomes were high-quality assemblies (> 90% complete) while the other 413 assemblies were considered "genome-fragments" (< 90% complete). The genome size (median: 36,691 bp, IQR: 31,821 - 40,929) and the mean gene density (1.46 genes/kb ± 0.12) of the putative transposable phages did not differ from the values found for the set of transposable phages stored in NCBI (Mann-Whitney U test, $p = 0.08$, $p = 0.32$, respectively) supporting the validity of the putative transposable phages. According to the isolation source, putative transposable phages are widespread in nature. they can be found in free-living environments, from aquatic to terrestrial habitats, and host-associated environments, including plants and animal digestive systems. Finally, regarding the host range, putative transposable phages include instances infecting bacteria belonging to at least three phyla: Proteobacteria ($\alpha$-, $\beta$-, $\gamma$-, $\delta$-, and $\epsilon$-proteobacteria), Deinococcus-Thermus (*Thermus*), and Firmicutes (*Agathobaculum*, *Faecalibacterium*, *Intestinimonas*, *Clostridium*, *Hungatella*, *Roseburia*, and *Eubacterium*) (Appendix .3).

Starting with the identification of marker ViPhOGs in complete genomes of transposable phages stored in NCBI, I was able to identify more than a thousand new instances of transposable phages in databases of metagenomic assemblies. This set of putative transposable phages not only exceeds the number of genomes stored in NCBI but also constitutes a tangible set of genomes that represents the thriving diversity and widespread nature of transposable phages predicted a decade ago from the analysis of prophage sequences.

### 3.2.2 Transposable phages do not constitute a monophyletic group

Previously, it was suggested that transposable phages should be grouped into a single family, *Saltoviridae*, and split its members into two subfamilies, *Myosaltovirinae* and *Siphosaltovirinae*, according to the phage morphology [127]. Nevertheless, this proposal was based on the genomic analysis of 26 transposable phage isolates available at the time. The set of putative transposable phages exceeds the number and genomic diversity described by Hulo et al. Therefore, I aimed to determine how these putative transposable phages are related among themselves and to other dsDNA phages.

The repertoire of putative transposable phages comprised 765 species and 440 genera, based on 95% and 70% ANI, respectively. Nevertheless, these thresholds are proposed to be used on unclassified genomic isolates [53]. As shown in the previous section, not all putative transposable phages are complete genomes, which might affect their cluster-

ing by ANI [20]. To reduce the clustering variability expected from incomplete metagenomic assemblies, I calculated the similarity between genomes using the average amino acid identity of the proteins associated with the marker-ViPhOGs (markers-AAI) (Table 3.1). Markers-AAI is a more conservative measure than ANI, meaning that it might sub-estimate the actual number of genera but will not overestimate it. Using a 90% markers-AAI similarity threshold, the 1,002 transposable phages' genomes formed 423 clusters; fewer clusters than the number of genera obtained by the use of genomic identity (Figure 3.1). Next, from each of the 423 clusters, I selected a representative genome. Representative genomes are those with a minimal geometric distance to the mean genome length and mean gene density of the 60 transposable phages stored at NCBI. I further analyzed this set of representatives to understand the genomic diversity of all transposable phages.

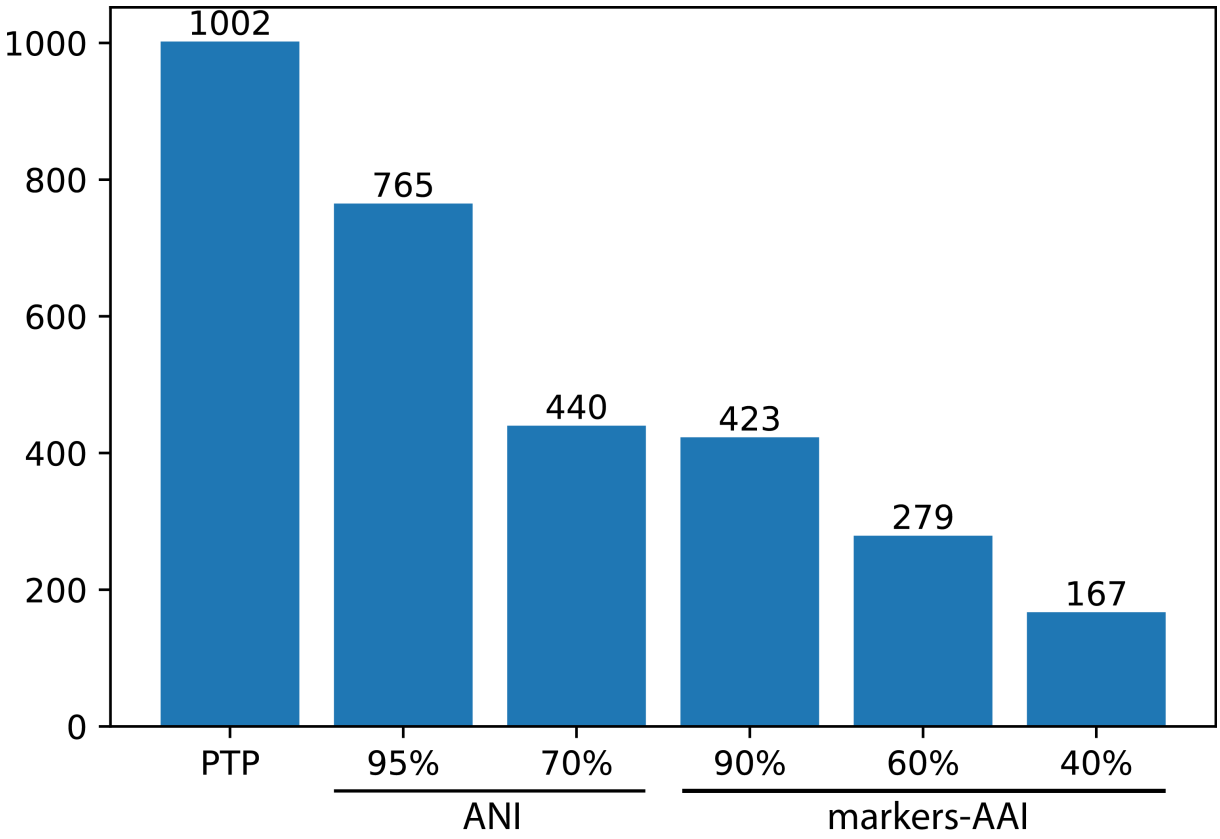


**Figure 3.1: Clustering of putative transposable phages** Number of clusters of putative transposable phages (PTP) using different thresholds for ANI or markers-AAI

To understand how transposable phages relate to dsDNA phages, I compared the representative genomes against each other and against 306 randomly chosen, and family-assigned dsDNA viruses. For that, I used GRAViTy, which combines gene homology

| Genomic Feature (Function) | ViPhOG (id) | Count |
|---|---|---|
| Major head (Capsid protein) | 16138 | 51 |
| | 17114 | 8 |
| Protease (I) (Capsid maturation) | 16105 | 51 |
| | 17112 | 9 |
| Mor (Transcriptional regulator) | 25407 | 4 |
| | 16136 | 30 |
| | 17088 | 23 |
| Adapter (Mu36) (Tail-capsid joining) | 16107 | 58 |
| GemA (Transcriptional regulator) | 16129 | 58 |
| virion morphogenesis (G) (Tail-capsid joining) | 5603 | 59 |
| minor head protein (F) (Capsid protein) | 16101 | 3 |
| | 16102 | 31 |
| | 17110 | 12 |
| | 17111 | 4 |
| | 4939 | 9 |
| Portal Genome packaging | 4837 | 60 |
| transposase B (Replicative transposition) | 16139 | 34 |
| | 17095 | 13 |
| | 25408 | 11 |
| transposase A (Replicative transposition) | 16144 | 35 |
| | 17098 | 13 |
| | 25398 | 9 |

**Table 3.1: Transposable phages' marker-ViPhOGs.** Prevalent proteins of transposable phages, their functional description, the different orthologous groups that matched those proteins, and their prevalence within the 60 NCBI transposable phages.

and genomic organization to calculate the genomic similarity between two genomes [56, 136]. According to the dendrogram, transposable phages' within-group similarity is higher than most other phage families (Figure 3.2). Only four families grouped in the same cluster with the transposable phage representatives: *Casjensviridae*, *Mesyanzhinovviridae*, *Peduoviridae*, and *Winoviridae*. The families *Casjensviridae* and *Mesyanzhinovviridae* were the most distant families, and none of their members encode for features related to transposable phages. For their part, the *Peduoviridae* and *Winoviridae* families were among transposable phage representatives, and their members include transposable and non-transposable phages. Suggesting that transposable phages are polyphyletic and not monophyletic.

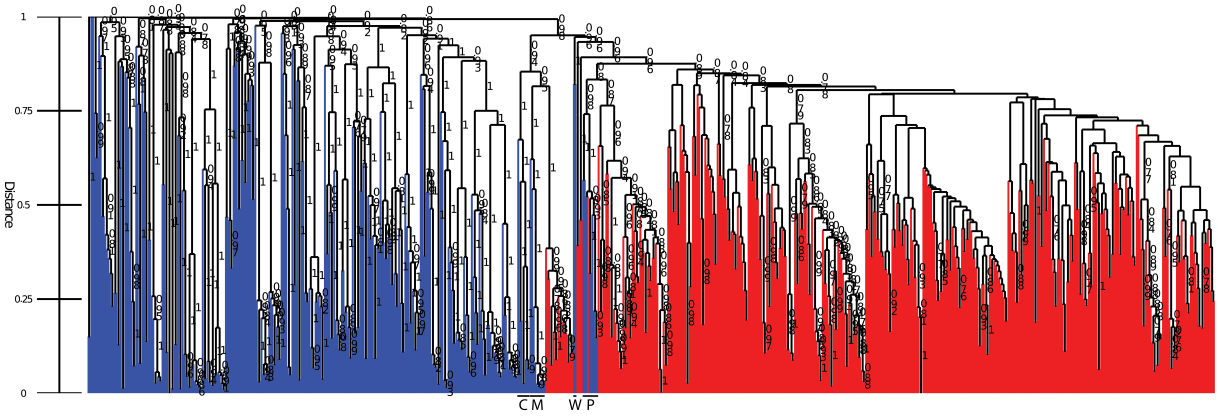


**Figure 3.2: Putative transposable phages within the dsDNA virus diversity.** Dendrogram of the composite generalized Jaccard (CGJ) distances among dsDNA phages (blue) and putative transposable phages (red) using GRAViTy. Bootstrap support values higher than 70% are shown above branches.

To address the evolutionary relationships among transposable phages, I reconstructed a maximum likelihood phylogeny based on their portal protein using RAxML [120]. The phylogenetic tree depicted transposable phages in five major clades, each including at least one NCBI isolate (Figure 3.3). Additionally, I showed, using the phylogenetic reconstruction, that phages with "Myo" and "Sipho" morphologies are present in each of the clades, meaning that transposable phages should not be divided into *Myosaltovirinae* and *Siphosaltovirinae*. Clade 1 includes several well-described transposable phages, such as Mu phage which infects *E. coli* and other *Enterobacteria*, Haemophilus SuMu phage, and several mu-like phages infecting *Pseudomonas* (e.g. MP22, DMS3, JBD24, and JBD88a). In Clade 2, I found phages infecting several Rhodobacterales, like *Pelagibacter*, *Thiobacimonas*, *Rodvoulum*, and *Rhodobacter*. Also, Vibrio phage 12B12 fell in this clade. In clade 3, the only isolate was Mushu. Mushu is the only transposable phage isolated

from a gram-positive bacterium (*Faecalibacterium prausnitzii*), and the only transposable phage isolated from feces [131]. In the tree, Mushu appears accompanied by other 65 transposable phage representatives, its closest relatives are predicted to infect other Firmicutes, e.g. members of the species *Intestinimonas*, *Roseburia*, and *Eubacterium*. Nonetheless, representatives infecting species from another phylum, e.g. *Thermus*, were also found in this clade. Finally, clade 4 includes Pseudomonas phages B3 and JBD67, Rizhobium phage RR1-B, and Rhodobacter phage RcapMuZZ4; while clade 5 includes phiE255 and KS10, both infecting members of *Burkholderia*.

Based on the genomic and phylogenetic analyses I presented here, I would like to suggest that transposable phages are a polyphyletic group of tailed phages, they should not be grouped in a single taxon, and certainly not grouped based on their morphology. Since transposable phages are an extremely diverse group, additional analyses of each major clade should be performed, to better define the taxonomy of this group.

### 3.2.3 Mushu is not alone. A whole family of Mushu-like phages is delimited from viral metagenomic assemblies

The transposable phages expert Arian Toussaint predict almost a decade ago that transposable phages inhabit the human gut microbiome [89]. However, only one isolate of a transposable phage from feces is found in public databases: Mushu phage (ACC: NC_047913.1). Mushu was identified as a prophage in the genome of a *Faecalibacterium prausnitzii* bacterium isolated from feces. It was isolated and sequenced, and the variable bacterial sequences on each side of the genome confirmed its capacity for replicative transposition, ensuring the designation of Mushu as a transposable phage [131]. The results I present here show that transposable phages are a polyphyletic, diverse, and ubiquitous group of phages. Transposable phages from the human gut include each of the five major clades of transposable phages. In particular, major clade 3 is dominated by assemblies from human gut samples and Mushu is the only isolate present in this clade. Here, I aim to delineate a Family for Mushu-like phages, and to dig into the functional capabilities of this kind of transposable phage inhabiting the human gut.

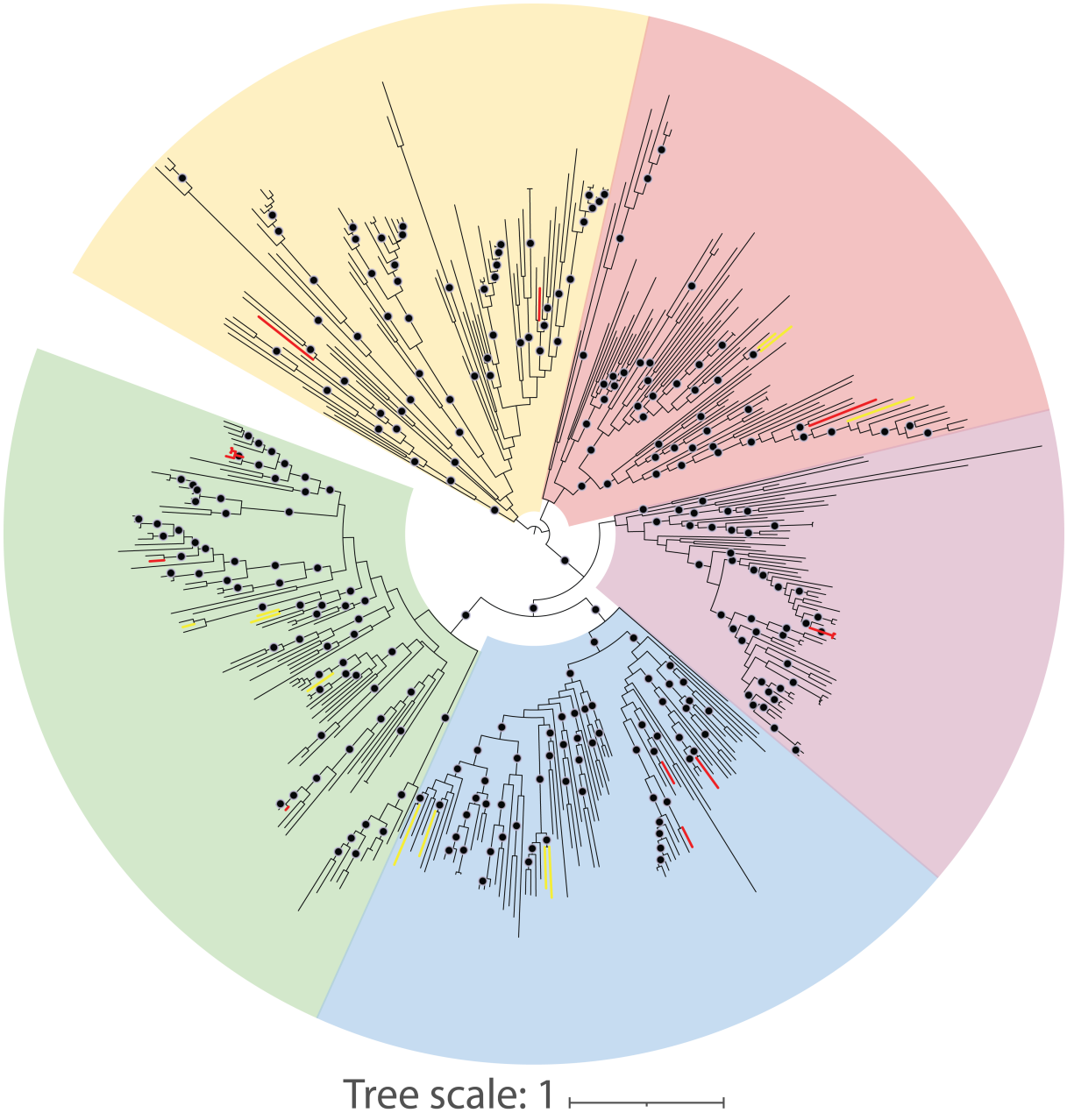


**Figure 3.3: Phylogenetic reconstruction of putative transposable phages** Maximum likelihood phylogenetic tree of the portal protein of putative transposable phages. Branches of representatives of the 60 NCBI transposable phages are shown in color yellow or red according to their morphology. Myovirus morphology or siphovirus morphology, respectively. Circles indicate bootstrap values above 70%. Scale: Average substitutions per site.

According to the latest ICTV guidelines, the taxonomy of complete sequenced isolates can be inferred -up to genus- based on genomic distances. Since the genomes presented here are products of metagenomic assemblies, designating a Family for Mushu required not only identifying a set of genomes similar to Mushu but dealing with any factor that

might interfere with the sequence identity calculation, i.e. flanking regions that belong to the host, sense of the assembly, and incompleteness. Using Clinker [137], I visually compared Mushu's to 95 "Mushu-like" metagenomic assemblies that presented a markers-AAI of at least 60%. In total, 34 assemblies were reversed, 45 assemblies were cut at the beginning or the end to remove their flanking regions, and 72 assemblies were confirmed to be complete (genome length  90% Mushu's length) (Figura 3.4).

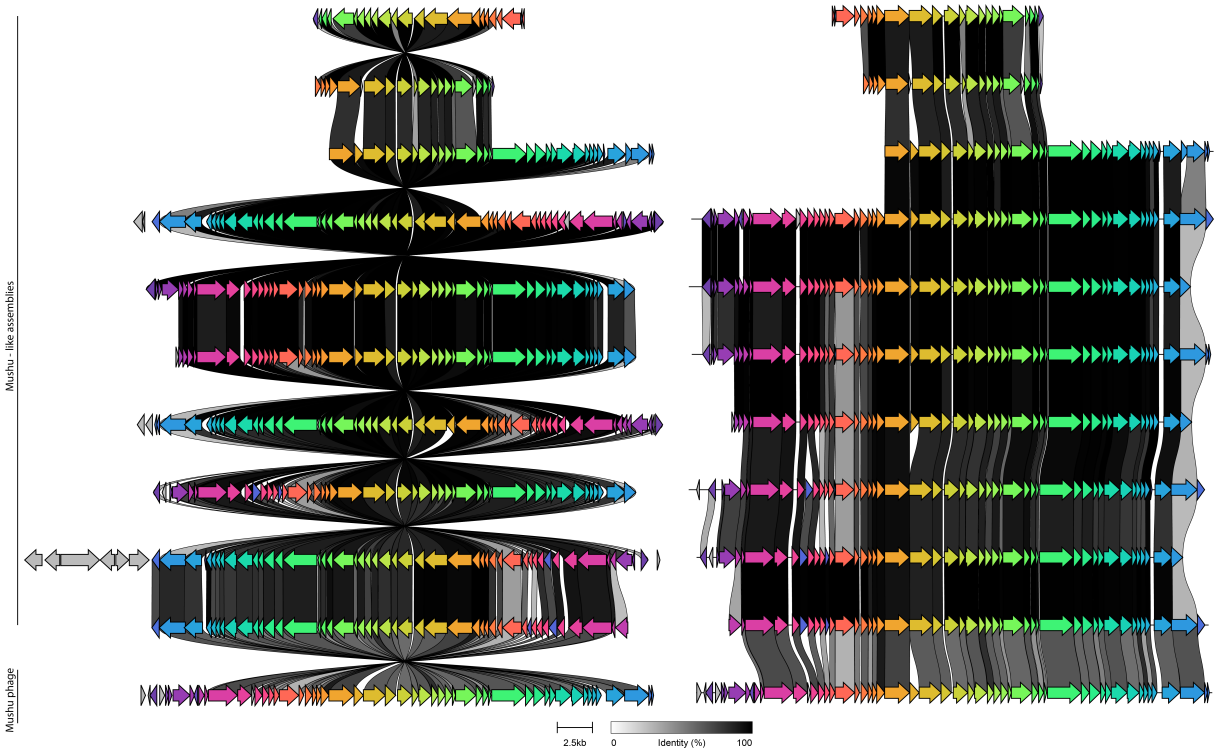


**Figure 3.4: Representation of Mushu-like metagenomic assemblies** Clinker comparison of Mushu and 10 Mushu-like metagenomic assemblies before (left) and after (right) manual curation. This shows how metagenomic assemblies can be incomplete (top tree), in counter-sense with the reference, or include flanking regions ($4^{th}$, $7^{th}$, and $9^{th}$ assemblies from top to down). Arrows are scaled proteins. Colored arrows represent homologous proteins connected by adjacent lines that reflect their identity percentage. Grey arrows are proteins with no homologs.

After calculating ANI among the set of complete genomes, it was clear that the set of Mushu-like genomes can be arranged in 9 genera and 34 species. However, it was unclear whether all genera constitute a single family, as they were arranged in 3 different clusters (Table 3.2). Interestingly, while some clusters showed no nucleotide similarity among each other, all genomes were at least 62.65% identical according to AAI (Figura 3.5). The mean length of Mushu-like genomes was $35,707bp \pm 982$, and their mean number of proteins

| Cluster | Genus (id) | Genomes (#) |
|---|---|---|
| A | 1 | 4 |
| | 2 | 1 |
| | 3 | 38 |
| | 4 | 2 |
| | 5 | 1 |
| B | 6 | 8 |
| | 7 | 16 |
| | 8 | 1 |
| C | 9 | 1 |

**Table 3.2: Organization of the Mushu-like genomes.** Clusters, genera, and number of genomes in each genus of the Mushu-like family according to ANI.

per genome was $50.49 \pm 2.04$. Given the low number of genomes in most of the genera, I only tested the difference in the mean number of proteins, length, and GC percentage between genera 3, 6, and 7. I only detected a significant difference in the GC% of genera 3 and 6 (Mann-Whitney U test, $p = 0.0005$), and genera 3 and 7 (Mann-Whitney U test, $p = 0.004$) (Figura 3.6).

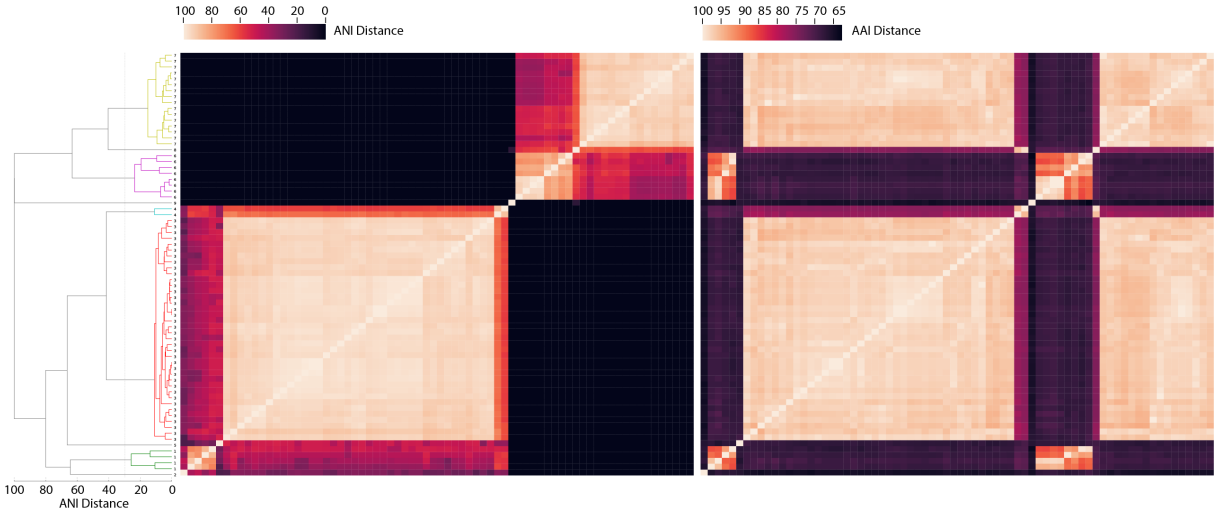


**Figure 3.5: ANI and AAI comparisons of the Mushu-like family.** Heatmaps of ANI (left) and AAI (right) between Mushu-like genomes. Both heatmaps are sorted according to the hierarchical clustering of the genomes based on ANI (left dendrogram), as shown in the dendrogram. In the dendrogram, each colored group is a genus.

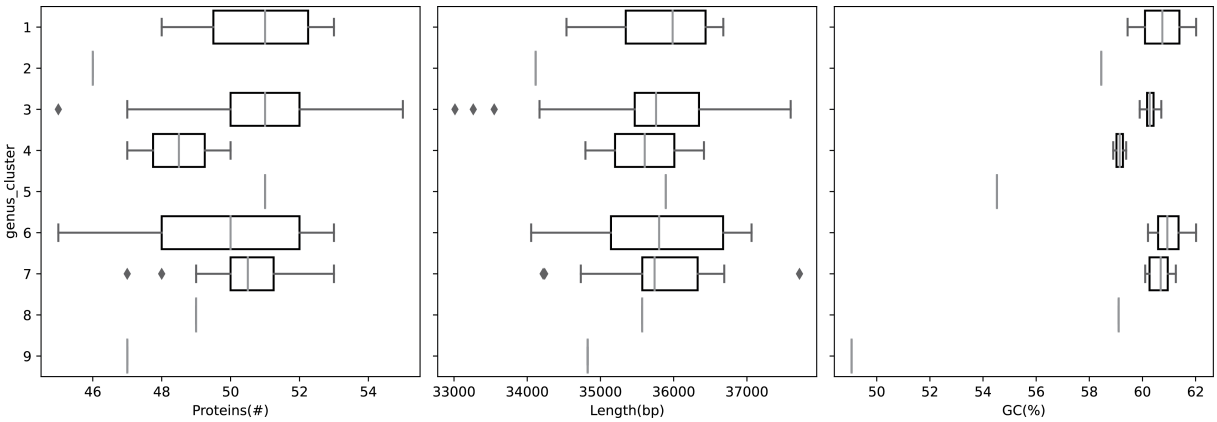


**Figure 3.6: Proteins, length, and GC percentage of the Mushu-like family.** Box plots showing the number of proteins, length, and GC percentage for each of the genus in the Mushu-like family. "*" indicates significative difference ($p < 0.05$)

According to the ICTV, at least 10% of the average number of genes in a family must be orthologs present in all members. To address this, I used PPanGGOLiN to characterize the super-pangenome of Mushu-like phages [138]. The core genome included 14 orthologs: 5 structural proteins, 3 proteins involved in the genome packaging, 2 in the virion assembly, 1 in the genome replication, 1 in evading the host defense system, and 1 hypothetical protein (Table 3.3). Since 14 orthologs correspond to 25% of the average number of genes, Mushu-like phages satisfied all the conditions to make a family proposal to the ICTV committee.

In addition to detecting the core genome, PPanGGOLiN splits the family's super-pangenome into three partitions: persistent, shell, and cloud. In total, the super-pangenome included 3,635 genes. The persistent partition contained 1,420 genes grouped in 20 orthologs; the shell partition, 367 genes grouped in 9 orthologs; and the cloud partition, 1,848 genes grouped in 1,775 orthologs. To understand the functional diversity within the Mushu-like family, I mapped the genes in each partition to the database of clusters of orthologous groups (COG) [139, 140]. Given that most of the genes were either not present in the COG database (35.85%), or assigned to the categories S (Function Unknown; 29.49%) or X (Mobilome: prophages and transposons; 7.95%), I re-evaluated the category assignation to describe the viral process in which the gene is involved based on the consensus annotation and following the bacterial virus ontology [141]. For example, the portal protein was assigned to the COG category S. Nevertheless, I have used it as a phylogenetic marker because it is a conserved protein involved in genome packaging. Accordingly, I re-evaluated the category assigned to the portal protein from "S" to the viral category "virus release from host cell" (vO), which includes the process of viral genome

| ViPhOG | Marker-ViPhOG | Protein | Description |
|---|---|---|---|
| 16144 | Yes | Transposase A | Viral replication |
| 16124 | Yes | Gam (Host nuclease inhibitor) | Host-Virus Interaction |
| 16138 | Yes | Major head subunit | Virion component |
| 16155 | No | Terminase large subunit | Virus release |
| 16157 | No | Hypothetical protein | NA |
| 5603 | Yes | Virion morphogenesis | Virus release |
| 4837 | Yes | Portal | Virus release |
| 17088 | Yes | Mor transcription activator | - |
| 16105 | Yes | Protease (I) | Virus release |
| 4594 | No | Baseplate methyltransferase | Virion component |
| 4711 | No | Tail tape measure protein | Virus release |
| 17108 | No | Terminase small subunit | Virus release |
| 16107 | Yes | Adapter (Mu36) | Virion component |
| 6081 | No | Baseplate J | Virion component |

**Table 3.3: Core genome of the Mushu-like family.** Orthologous groups found in the 72 complete genomes of the Mushu-like family, the protein associated with each of them, and the process in which the protein is involved according to the viral ontology. The column "Marker-ViPhOG" indicates if the orthologous group is a marker-ViPhOG of transposable phages.

packaging. By assigning functional categories according to the virus ontology, I reduced the number of genes assigned to the COG categories S, L, M, and X. In total, 100% of the genes in category X and 80.5% of the genes in category S were re-assigned. Unfortunately, almost all genes that were not present in the COG database also didn't have an annotation with PFAM or ViPhOGs databases, meaning that 1,296 genes (35.84%) of Mushu-like super-pangenome are orphan genes (Figure 3.7). The abundance of the functional categories in each partition showed that viral components/structural genes (vS), genes involved in the virus release from the host cell/lysis (vO), and genes involved in viral replication (vR or L) were found exclusively in the persistent partition, while genes related to the entry into the host (vI), and host-virus interaction (vHVI) were also found in the cloud partition. Categories M (cell wall/membrane) and G (carbohydrate metabolism), which include genes that might be associated with the entry into the host cell, were also found in the cloud partition (Figure 3.8).

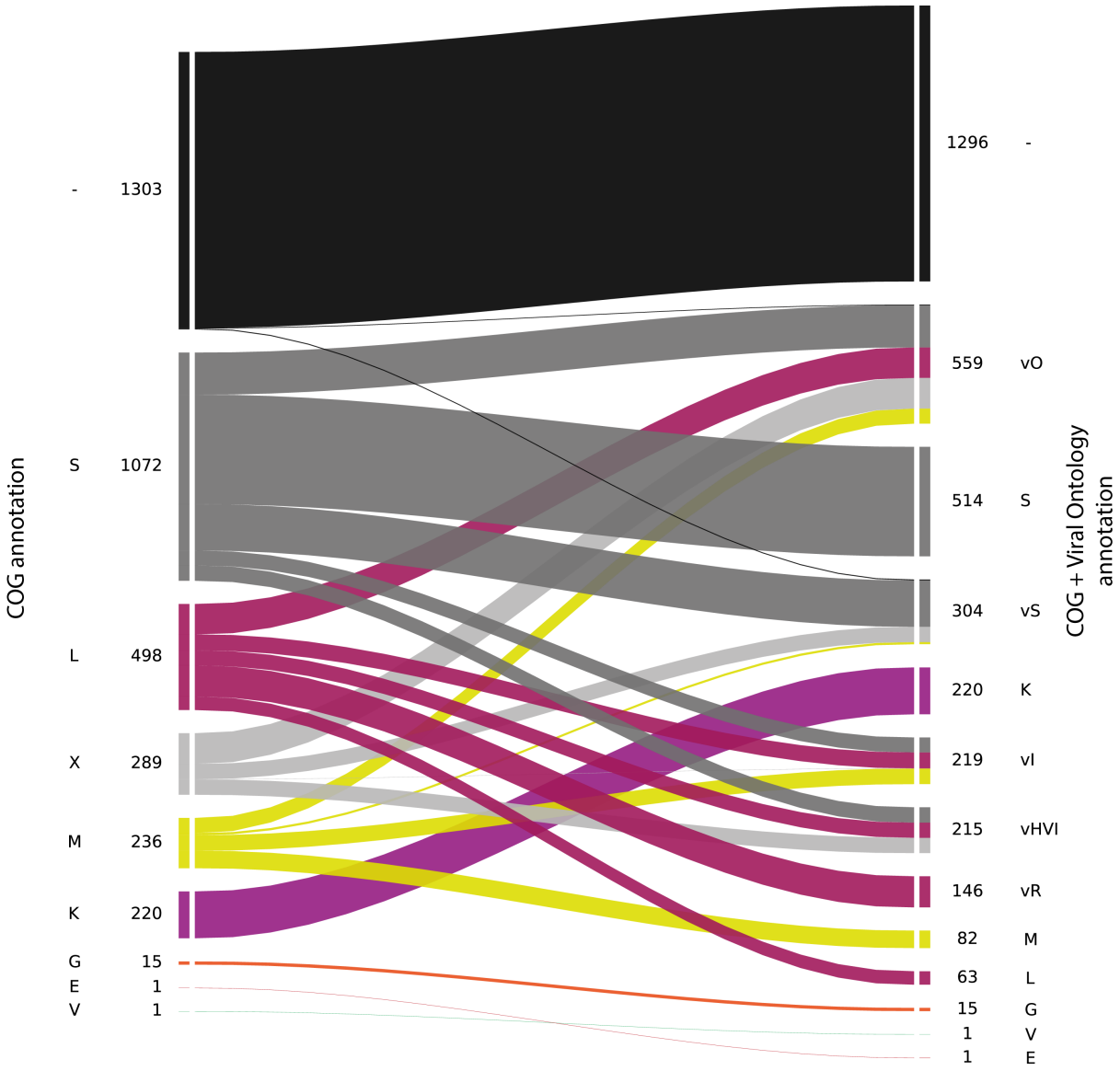


**Figure 3.7: Functional categories in the Mushu-like genomes** Alluvial plot showing how COG categories were re-assigned to categories created based on the viral ontology proposed by Hulo and collaborators [141]. Each letter represents a COG functional category as follows: S: Unknown function; L: Replication, recombination, and repair; X: Mobilom: prophages and transposons; M: Cell wall/membrane biogenesis; K: Transcription; G: Carbohydrate production and metabolism; E: Amino acid transport and metabolism; V: Defense mechanisms. Groups of letters starting with "v" are the viral categories I created based on the viral ontology; vS: Virion Component; vI: Virus entry into the host; vO: Virus release from host cell; vR: Viral replication; vHVI: host-virus interaction

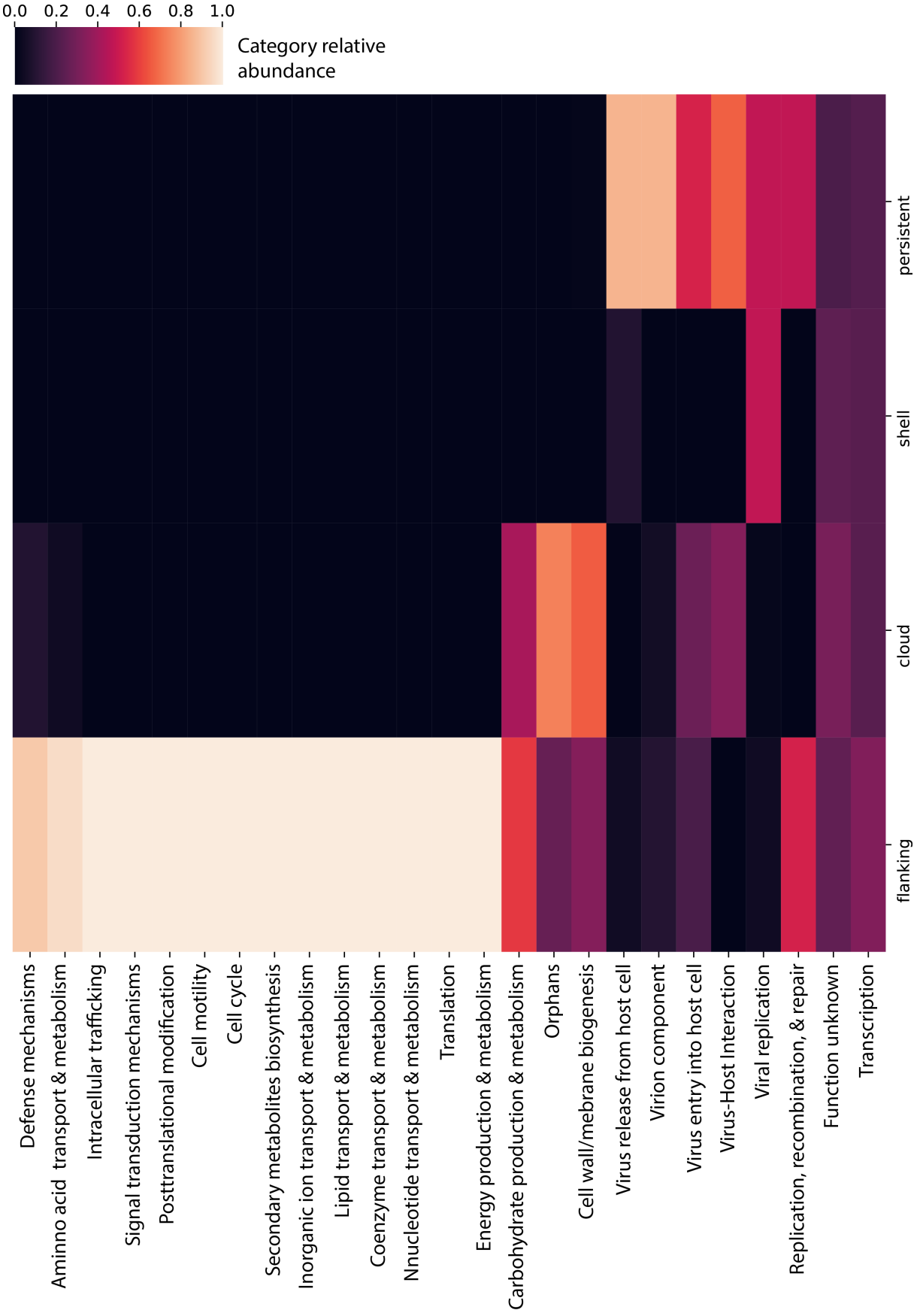


**Figure 3.8: Functional annotation of the Mushu-like family** Heatmap showing the relative abundance of the functional categories in each of the super-pangenome partitions and the flanking regions

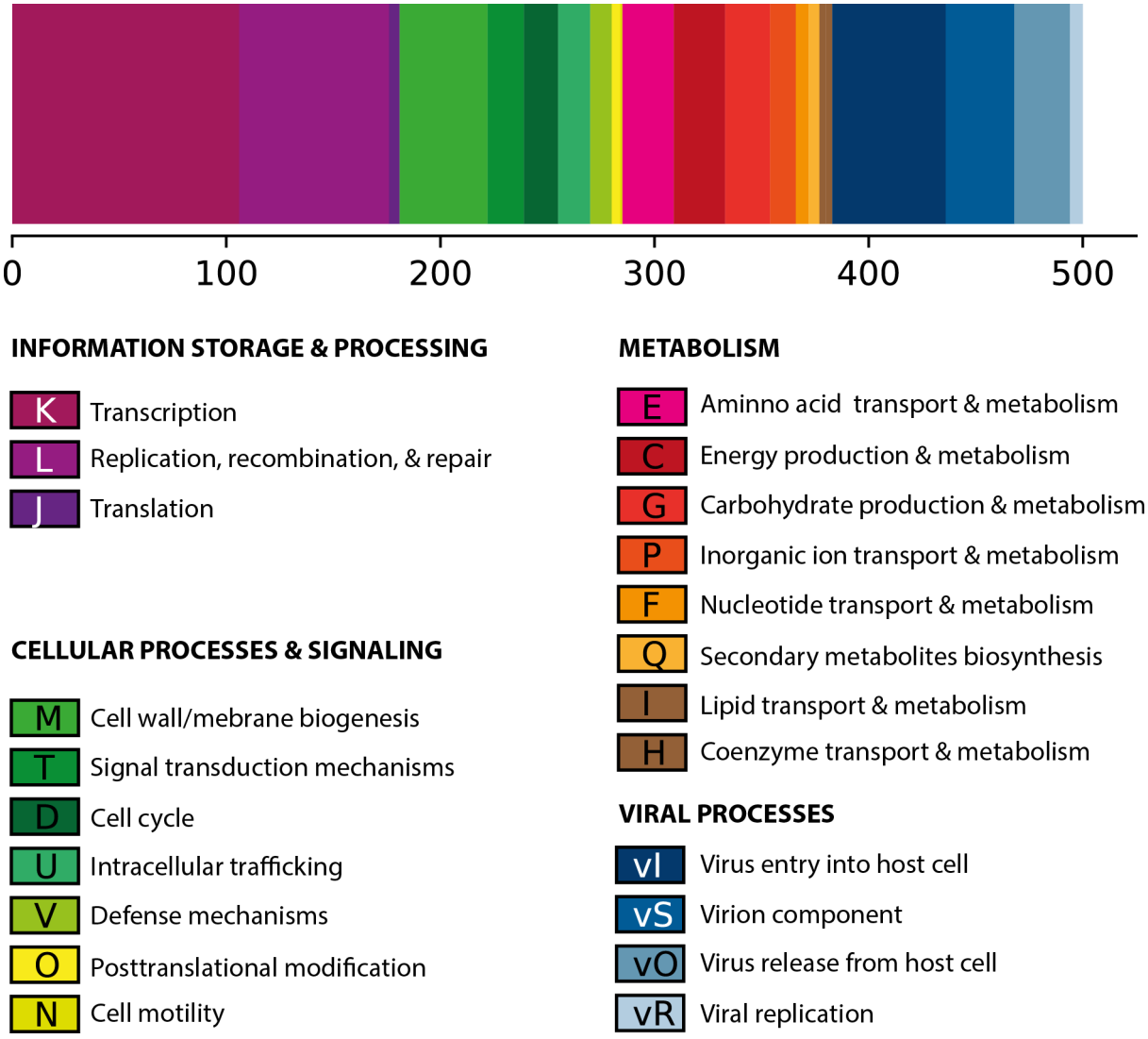


**Figure 3.9: Functional annotation of the flanking regions of Mushu-like phages.** Barplot showing the number of genes annotated from the flanking regions. Each color represents the indicated functional category.

Finally, to identify which genes are (potentially) transferred by the Mushu-like family in the human gut, I analyzed the functional profile of the flanking regions. Similarly to the cloud partition, flanking regions were dominated by unannotated genes but to a lesser extent. Additionally, a wide diversity of genes related to cellular signaling, cellular metabolism, and information processing might be subject to horizontal gene transfer (Figure 3.8). Among annotated genes, genes related to transcription regulation and replication corresponded to 35% of all annotated genes. Regarding genes involved in cellular signaling, categories M (Cell wall/membrane/envelope biogenesis) and T (Signal transduction mechanisms) account for more than half of the genes. Categories C (Energy

production and conversion) and E (amino acid metabolism) are the ones that cover more than half of the genes related to metabolism (Figure 3.9). This wide functional spectrum of the flanking regions might reflect the random insertion of transposable phages in their hosts and supports their ability of generalized transduction.

Using comparative genomics, I presented a family of phages that include Mushu, the only transposable phage isolated from fecal samples. In addition, I characterized the family's super-pangenome and exploited the annotation of the flanking regions to expose genes susceptible to horizontal gene transfer. Whether these genes are beneficial or detrimental to their hosts cannot be answered with these data. Overall, this work is an example of how to integrate metagenomic assemblies into the ICTV taxonomy, encourages the use of the bacterial virus ontology for the functional characterization of viral genomes, and constitutes the first step into the taxonomical and functional characterization of transposable phages of the human gut.

## 3.3 Discussion

### 3.3.1 The use of ViPhOGs and remote homologous search enhance the analysis of viral sequences

Here, I presented how from the identification of transposable phages stored in NCBI I was able to identify more than a thousand new instances of transposable phages in databases of metagenomic assemblies. The re-annotation of complete genomes of transposable phages led to the confirmation of features that have been known to be characteristic in transposable phages, these are their genome organization, and a set of 10 prevalent proteins (portal protein, virion morphogenesis protein, GemA, adapter protein, transposase A, transposase B, minor head protein, major head protein, protease, and Mor protein) [130, 127, 133]. Furthermore, the re-annotation of all proteins of the stored transposable phages led to the identification of one or several orthologous groups matching the prevalent genes. Notably, only the portal, virion morphogenesis, adapter proteins (all three involved in head formation and DNA packaging), and the gemA protein (an early regulatory protein) were associated with a single orthologous group, while the other six prevalent proteins were associated to multiple orthologs. This finding supports the fact that transposases are conserved only within transposable phages' groups, e.g B3-like and Mu-like transposable phages possess different transposases [133], and highlights the mosaic nature of phages, i.e phages can be described as modular entities able to share and shuffle their functional modules as long as the purpose of the shared module is fulfilled

[142].

Based on the prevalence of the characteristic proteins I designated their associated orthologous groups as markers. Thus, rather than using directly each signature gene to identify transposable phages from metagenomic assemblies by sequence comparison, I was able to perform a remote homologous search based on protein profiles. Remote homologous searches improve homology detection since it includes more information compared to pairwise methods [143, 144]. As expected, thanks to the remote homologous searches, I was able to identify 30 times more putative transposable than transposable phages' complete genomes stored in NCBI databases; 60 vs 1,002, respectively. Finding different marker orthologues for the characteristic functions of transposable phages, and using them in remote homologous searches tackle three intrinsic characteristics of viruses that have limited our capabilities to characterize viral metagenomic assemblies. First, viruses lack a universal gene that can be used as the basis for their classification [49]. Instead, marker genes for specific groups of phages have been used to describe their diversity, like for *Microviridae* [145, 100]. In this case, I presented a set of marker-ViPhOGs, none of them is a marker on their own, and together they represent a wide diversity of transposable phages. Second, the high rate of evolutionary change in viruses [11]. As mentioned before, remote homologous searches have been shown to be more sensitive than pair-wise comparisons to detect sequences originating from more distant relatives [143, 67, 144], overcoming this limitation. Lastly, the third limitation is that viruses are mosaic and modular, which might affect a genomic analysis since the sequences of a functional module differ among members [16, 17]. In this case, complementing the annotation of viral proteins with orthology information led to the recognition of the different functional modules accessible to the phages.

In summary, by using orthologs groups and remote homologous searches to explore metagenomic assemblies, I was able to vastly expand the diversity of transposable phages, providing a tangible set of genomes that represents their predicted thriving diversity and widespread nature. Furthermore, I provide a methodology that when applied to other groups of viruses, as it was used here for transposable phages, might facilitate the identification and classification of the viral dark matter that is continuously sequenced from metagenomic studies.

### 3.3.2 The vast diversity of transposable phages exposed from metagenomic assemblies reveals their polyphyletic origin

Addressing the diversity of viruses and elucidating the relationships between them has been a challenging task. Viruses lack a universal gene marker, have multiple origins [146], and are modular and mosaic [16, 17]. Until recently, virus classification was based on a phenotype-based characterization, however, the study of viral metagenomes revealed that this traditional classification does not reflect the intricate evolutionary relationships among viruses [48, 49, 50]. Recently, the organization of the virosphere was elucidated using protein domains encoded by viral hallmark genes [52]. This promoted the creation of a novel multi-rank taxonomy framework of viruses, which allows the delineation of novel taxa from sequence data and phylogenomics [51, 147]. In this chapter, I applied comparative genomics and phylogenetics to evaluate the relationships between hundreds of transposable phage representatives, in addition to their relationship to other bacterial and archaeal dsDNA viruses. The results support that transposable phages are not monophyletic and should be grouped in more than one taxonomic family.

The work of Hulo and collaborators [127], which defined the *Saltoviridae* family, is also based on comparative genomics of transposable phages and includes a comparison against other 450 phages. Nevertheless, the 26 genomes used in their study correspond to 25 species compared to (at least) 423 species that represent all putative transposable phages identified in IMG/VR and GPD. Furthermore, I found that 16 of 26 genomes belong to clade 1 of the phylogenetic tree of all putative transposable phage representatives. Thus, it appears that the low diversity of transposable phages available at the time was hiding their polyphyletic nature.

The comparison of transposable phages against other dsDNA phages revealed that some transposable phages grouped with the *Peduoviridae* and *Winoviridae* families. The *Peduoviridae* family was recently defined to elevate the subfamily *Peduovirinae* after the abolishment of the *Caudovirales* order and its families. Phages of the *Peduoviridae* family are in fact myoviruses related to enterobacteria phage P2, hence the family name. Since long ago, it is known that Mu phage is almost morphologically identical to P2 [148]. Even, some members, such as *Aresaunavirus* (e.g. NC_049432), *Arsyunavirus* (e.g. NC_025115), and *Baylovirus* (e.g. NC_047750, NC_028898), code for transposases. Nevertheless, the evolutionary relationship between *Peduoviridae* and Mu-like phages is still not clear. The *Winoviridae* family was recently defined after the isolation of viruses of *Flavobacteriia* [149]. It includes two genera: *Pippivirus* and *Peternellavirus*. Only members of the genera *Paternellavirus* but no members of *Pippivirus* show genomic features related to transposable phages. Based on this observation, three conclusions arise: First, since

phages of the *Winoviridae* family are not exclusively transposable, transposable phages are not a monophyletic group. Second, it is evident that a single family does not allow accurate classification of transposable phages. Third, it is expected to find other genomes related to transposable phages but without replicative transposition capacity. If found, their characterization and identification would help to understand the intricate network that underlies the diversity and evolutionary history of transposable phages.

Metagenomic studies have revealed the presence of a humongous variety of viruses in diverse environments [150]. This is changing our perspective on viruses and their impact. A better organization of this viral diversity might lead us to understand the origins and the forces that shape it. With this work, I have contributed to describing the relationships of transposable phages within the dsDNA phages, providing evidence of their polyphyletic origin. The description and delimitation of other families within the diversity of transposable phages will be important to understand their origins.

### 3.3.3 Metagenomic assemblies related to viral isolates can be integrated into the ICTV taxonomy framework

The number and diversity of viral sequences identified using viral metagenomics has exceeded by far the number and diversity of experimentally characterized viruses. The challenge has been to classify and incorporate this unprecedented diversity into the viral taxonomy [49]. To achieve this goal, the field has advanced on two fronts over the last five years. First, the criteria to ensure the quality and completeness of viral metagenomic assemblies were defined [20]. Importantly, tools to facilitate the criteria assessment have been developed [151, 152]. On the second front, given that the taxonomy was limited to cultured viruses and didn't reflect the true extent of the viral genomic diversity, the scientific community and the ICTV have collaborated to create a new virus taxonomic framework. The new framework delineates new ranks to accommodate the viral diversity, passing from a five-rank system to a fifteen-rank that better reflects a Linnaean taxonomy system [51]. Additionally, the bacterial and archaeal arm of the ICTV agreed on a genomic threshold for species (95%) and genus (70%) demarcation of coding complete genomes [53]. Given that related phages have similar lengths and proteins number [152], I used Mushu's coding complete sequence to delineate the Mushu-like family among the expanded diversity of transposable phages; setting an example of how the last advances in viral classification allow integrating complete metagenomic assemblies into the ICTV taxonomy.

The demarcation of genomic identity thresholds to define the lowest taxonomic groups

for phages is a game changer in viral ecology. It not only allows viruses to be named, but it will also facilitate the creation of higher taxonomic groupings and improve the catalog that enables the description of the viral diversity [153]. Although these taxonomic definitions are useful and will kickstart viral cataloging, the species and genus demarcation rules will probably change. Large pairwise differences in ANI are common in viruses. For example, in eukaryotic viruses, the demarcation of low taxonomic levels remains variable due to the difference in evolutionary rates, genome architectures, and replication strategies observed within some families [147]. In the case of the Mushu-like family, the nucleotide divergence within the group of phages was so high that some genomes didn't present similarities using Blast. Nonetheless, the AAI shows that all genomes are at least 65% similar. Interestingly, the clustering patterns of ANI and AAI are different. AAI shows genera 7 and 8 being more similar to genus 3, while ANI split genus 3 from genera 7 and 8. Maybe in the future, it will be needed to move towards a core-genome-based phylogeny. Protein-based methods have been shown to be robust enough even for viruses with highly divergent genomes and despite viruses' high mosaicism [56, 58]. Meanwhile, following the current ICTV guidelines, the Mushu-like family is the first family of human gut-inhabiting transposable phages to be defined. So far, it includes 72 genomes grouped into 34 species and 4 genera.

Family demarcation criteria requires that all family members share a significant number of orthologous genes [53]. I decided to evaluate not only the number of shared orthologs but to characterize the pangenome, or better, the super-pangenome of the Mushu-like family. A pangenome is defined as the entire set of genes present in a group of representative genomes from the same species. Khan et al have coined the term "super-pangenome" to address the need to develop a pangenome of pangenomes from different plant species that reflects the real diversity of plants, including crop wild relatives [154]. Given that high genomic divergence blurs taxa delimitation in viruses, I defined family as the taxa of interest to characterize the super-pangenome. One of the main factors that affects pangenome analysis is how the homology relationship between genes is defined [155]. Here, I used ViPhOGs as the unit for modeling the super-pangenome, i.e. two genes from different genomes were shared if they code for a protein that belongs to the same ViPhOG. Again, as a remote homology search is more sensitive than a pairwise comparison, my use of ViPhOGs enabled the identification of a family core genome. This highlights how ViPhOGs, or any other set of viral orthologs, are crucial to study the relationship among viruses. Once the super-pangenome was modeled, two main outcomes arose from its characterization. First, other members of the Major Clade 3 are expected to belong to the Mushu-like family. In other ICTV defined families, the number of orthologs shared among

all family members is approximately 10% of the average number of genes per genome [53]. In this case, the core genome comprises 14 orthologs, while the average number of genes per genome is 50. Therefore, other genomes in the major clade 3, more distant from Mushu, are expected to be part of the family. Second, the cloud partition of the super-pangenome revealed that those genes that change from member to member are involved in the phage entrance into the host cell and other phage-host interactions. Regarding their host, phages are species-specific or even strain-specific [156]. Phage's entrance into the host cell is dependent on the recognition between the phage's receptor binding protein and a host's receptor, e.g. outer membrane proteins, lipopolysaccharide receptors, or receptors located in capsule, pili, or flagella [157, 156]. The proposed Mushu-like family encompasses 72 transposable phage species, predicted host includes several members of the phylum Firmicutes, such as *Intestinimonas*, *Eubacterium*, and *Roseburia*. Given this diversity of phages and hosts, genes involved in those functional categories were expected to differ among family members.

One of the most interesting capabilities of transposable phages is that they are capable of generalized and specialized transduction [126]. In general, horizontal gene transfer has been considered essential for the evolution of prokaryotes given that its rate is comparable to the point mutation's rate, and surpasses the gene duplication rate [158]. There is extensive evidence of horizontal gene transfer in the human gut [159]. Nevertheless, it has been difficult to determine how much of the horizontal gene transfer is driven via transduction. Historically, conjugation and transformation were considered the major contributors to horizontal gene transfer. Indeed, the role of transduction may be underestimated. Metagenomics has uncovered that phages are the most abundant biological entities on the planet, and prophages have emerged as the major source of variation between bacterial strains [45]. The set of transposable phages revealed here from metagenomic datasets will be a valuable asset for determining what is the role and impact of transposable phages in horizontal gene transfer. The flanking regions of Mushu-like genomes included hundreds of genes involved in all types of metabolic and cellular processes. Genes related to transport and metabolism, for example, might improve bacterial growth by increasing the nutritional base or by allowing access to strong competed resources [160]. Like phages in the Mushu-like family, transposable phages in the other major clades can be characterized functionally and taxonomically based on a reference isolate. By defining more families, genera, and species it will be easier to compare and identify features of interest among different taxonomic groups. For example, it will be possible to ascertain which kinds of genes are potentially moved by transposable phages of different taxonomic groups, and if those genes are associated with the particular niche of their hosts within the human gut.

With this work, I participate and contribute to charting the structure of the virosphere. In particular, it defines and describes the first family of transposable phages from the human gut. Additionally, it lays the groundwork and provides references that will guide the experimental setup leading to the isolation of phage-host pairs required to assess their role and significance in microbiomes.

## 3.4 Methods

### 3.4.1 Search of transposable phages at NCBI

I used the genome of Mushu phage (Accession: MG711460.1) as a query in a tblastx-web-search `https://blast.ncbi.nlm.nih.gov/` against all bacterial viruses included in the nucleotide collection of the NCBI (Word size: 3; Expect value: $2X10^{-5}$; Gap Costs: 11,1; Matrix: BLOSUM62; Low Complexity Filter: Yes; Filter string: L; Genetic Code: 1; Window Size: 40; Threshold: 13. Database: nucleotide collection (nr/nt). Organism: Bacterial virus -taxid:28883-). In addition, genomes previously described by Hulo and collaborators [127] were added to the set of genomes retrieved by BLAST. To dereplicate the set of genomes I used the VIRIDIC web-service (`http://rhea.icbm.uni-oldenburg.de/VIRIDIC/`) [161]. As a representative of each species, I chose RefSeq accessions over Genbank accessions. If all entries of a species were Genbank accessions, I chose the longest genome as its representative. All species representatives' genomes were annotated (see section 3.4.2), and manually inspected to retain only genomes that encoded for the conserved features of transposable phages. According to Hulo and her collaborators, these are transposase A; transposase B; portal protein; minor head protein; virion morphogenesis protein; Adapter/neck protein; and GemA [127]. All ViPhOGs associated with those genomic features, or that were present in at least 50 out of 60 selected genomes were denominated as transposable phage marker-ViPhOGs. In total, 22 ViPhOGs associated to 10 genomic features (see table 3.1) were used as marker-ViPhOGs in further analysis.

### 3.4.2 Genome annotation

I annotated the coding sequences of each genome/metagenomic assembly using EggnOG mapper [132] and ViPhOGs [8]. From EggnOG mapper, I kept the annotations from COG and PFAM. To get the ViPhOG associated with each protein, I queried all HMM-ViPhOG-profiles against all predicted proteins of each genome using hmmsearch [118]. I chose the best hit among all matching ViPhOGs with an e-value below $1X10^{-5}$. After having the EggnOG and ViPhOG annotation for each coding sequence, I manually set a consensus

annotation for all proteins matching the same ViPhOG. In addition to coding sequences, I annotate tRNAs, tandem repeats, and inverted repeats (using ARAGORN [162], tandem repeats finder [163], and inverted repeat features [164], respectively). All the annotations were added to the genebank files following the "DDBJ/ENA/GenBank feature table definition" `https://www.insdc.org/submitting-standards/feature-table/` using Biopython [165].

### 3.4.3 Screaning of databases of metagenomic assemblies

I searched for transposable phages in two databases of UViGs, IMG/VR [134] and GPD [39]. In both cases, I used the set of transposable phages marker-ViPhOGs as a query of an hmmsearch against all predicted proteins of all UViGs stored in the database. For each protein, the marker-ViPhOG with the smallest e-value (if any) was associated with the protein (max. e-value: $1X10^{-5}$). Then, I built a presence/absence marker-ViPhOGs per genome matrix. Next, I grouped the marker-ViPhOGs by their associated genomic feature and kept those genomes that code for at least 6 features. One of them must be the portal protein -which I observed was associated with a single marker-ViPhOG in all 60 transposable phage genomes found on NCBI-. In total, 1,575 UViGs from IMG/VR and 422 UViGs from GPD satisfied the aforementioned criteria. These UViGs were combined with the 60 NCBI transposable phages and dereplicated at the species level (95% ANI) using a self-made Python script -since VIRIDIC's web server does not support thousands of genomes-. Then, I did gene calling for all transposable phage candidates using Multiphate [135] and re-annotated their genomes as it was done for the NCBI genomes (see section 3.4.2). After re-annotation, I built again a presence/absence marker-ViPhOGs per genome matrix and grouped the marker-ViPhOGs by their associated genomic feature. This time, I kept those UViGs that code for at least 8 transposable phages' genomic features. The 1,002 genomes that passed the threshold were denominated putative transposable phages.

### 3.4.4 Descriptive statistics

All the data regarding genomes length, number of proteins, and gene density was obtained using Biopython [165], analyzed using pandas [166, 167] and scipy [168], and plotted using seaborn [169] and matplotlib [170].

### 3.4.5 Markers-AAI clustering

I defined arkers-AAI between the genomes $A$ and $B$ as the follows: $\frac{\sum AI(a_{mi}, b_{mi}) * 2}{|A| + |B|}$ Where $AI(a_{mi}, b_{mi})$ is the amino acid identity between the proteins associated with the marker-ViPhOGs shared by the genomes A and B, $|A|$ is the number of marker-ViPhOGs in the genome A, and $|B|$ the number of marker-ViPhOGs in the genome B. I calculated the markers-AAI between all putative transposable phages. Then, I used single-linkage to cluster all putative transposable phages according to their markers-AAI distance. For each cluster formed using a distance threshold of 10%, I picked the genome with the least geometric distance to the mean genome length and mean gene density of the 60 transposable phages stored at NCBI as the representative genome of the cluster.

### 3.4.6 Comparison of putative transposable phages against the known diversity of dsDNA phages

The ICTV chooses an exemplar virus for each well-characterized virus species and collects all the information in the "Virus Metadata Resource" (VMR) `https://ictv.global/vmr`. I downloaded the VMR VMR_19-250422_MSL37.txt and keep only those entries of dsDNA bacterial and archeal viruses with a complete genome. From that subset of the VMR, I kept (at most) 10 random species' genomes per ICTV-defined family, for a total of 306 genomes from 70 different families. Then, I computed the sequence relatedness between the ICTV-defined families representatives and the putative transposable phages representatives using GRAViTy [56]. To assess the uncertainty of the observed sequence relatedness, I set to 100 the number of bootstraps.

### 3.4.7 Phylogenetic reconstruction

The portal protein of each putative transposable phage representative (proteins matching the ViPhOG 4837) were retrieved and aligned using MUSCLE v.3.8.31 [116]. Sites with at least 30% of gaps were removed using UGENE v.1.31.0 [117]. The best evolutionary model was obtained with modeltest-ng [171]. Maximum-likelihood phylogenetic analysis was done using RAxML v.8.2.4 [120]. Support values for nodes in the tree were obtained by bootstrap with 100 pseudoreplicates. The tree with supports was visualized and edited using iTOL [172]. The mentioned clades were visually selected based on the topology of the tree.

### 3.4.8 Mushu-like family delineation

All putative transposable phages having a markers-AAI distance below 40% to Mushu's genome were analyzed to delineate the Mushu-like family. First, I delimitated the beginning, end, and sense of all family candidates by comparing visually their genomes against Mushu's genome using Clinker [137]. Then, I processed with Python the JSON files generated by Clinker to retrieve the beginning and end coordinates of each genome. Only genomes with a length higher than 90% Mushus's genome length were further analyzed. I calculated ANI using the self-made python script used in section 3.4.3, and calculated AAI using the stand-alone AAI calculator from Kostas lab `http://enve-omics.ce.gatech.edu/aai/`. All candidates were hierarchically clustered based on their Euclidean distance using "complete linkage". Genera and species in the set of candidates were defined following the thresholds suggested by the ICTV guidelines (95% and 70%, respectively) [53].

The pangenome of all candidates was calculated with PPanGGOLiN [138]. To use the genome annotations I generated before, I parsed the GenBank files to GFF with Biopython. Also, I provided the gene to ViPhOG information to PPanGGOLiN to use ViPhOGs as gene clusters. The different gene clusters per partition were extracted from the "partitions" output folder and examined using Python. Finally, to provide a viral context to the gene annotations -since COG annotations were not informative-, I downloaded the "geneXannotation" output table and manually annotated each gene cluster to associate the annotations of their genes with the viral gene ontology [141].

# Conclusions and outlook

Two decades of viral metagenomics have changed the perspective from which viruses are observed, from mere pathogens to active builders of the web of life. Through viral metagenomics, it has been clear that viruses interact with organisms from all divisions of life and, together with microorganisms, they are present in all kinds of environments around the globe [173, 174, 175]. During my doctorate, I dedicated myself to the description of the diversity of viruses in the human gut, an environment in which microbial life performs essential functions but the viral players and their roles are not well understood.

First, I analyzed a dataset of viromes from monozygotic twins. Besides confirming that the human gut virome is highly unique and presents a high person-to-person dissimilarity [32, 36, 12], with this dataset, I was able to determine that the diversity of viruses correlates with the diversity of bacteria in the human gut by excluding host genetic relatedness. Additionally, I showed this pattern was driven by bacteriophages -which dominated the viral diversity over eukaryotic viruses-. Furthermore, I found that the viral-bacteria diversity mirroring pattern is generalizable. Meaning that, regardless of genetic relatedness, individuals with more similar microbiomes harbor more similar viromes. As shown, I have pushed forward our understanding of the microbial diversity of the human gut, by verifying these expected but unproven ecological patterns. I deposited the sequencing data of the viromes analyzed in chapter 2 in the European Nucleotide Archive under the study accession number PRJEB29491.

Although viral metagenomics is robust enough to study general patterns of diversity [9, 13], is the classification and characterization of UViGs that will allow us to improve our understanding of viruses, their relationship with other biological entities, and their roles in different ecosystems. In chapter 2, I presented two cases where I characterized UViGs using orthologs' profiles and remote homologous searches. Also, I exhibit how bacterial sequences -actually, genomes- are commonly found in human gut viromes. Thus, in chapter 3, I used orthologs to identify, compare, and describe putative transposable phages. As result, I have not simply found transposable phages in public databases, I expanded their diversity from 25 species [127] to at least 423 species -which represent 1,002 high-quality

metagenomic assemblies-, confirmed the predictions that there are transposable phages infecting members of Firmicutes [133], provide complete genome references of Firmicutes' transposable phages, proposed the Mushu-like family (a family of transposable phages inhabiting the human gut), and provided evidence on why transposable phages are not a monophyletic group.

With this work, I have provided the scientific community not only with new data but with new information to formulate or address new research questions. The main conclusion of chapter 2 is that viral diversity mirrors bacterial diversity in the human gut. It is expected that this pattern holds for other animmal guts dominated by bacteria. In such cases, it would be interesting to investigate animals having a less diverse microbiome to search for models to study the gut microbiome/virome assembly. In chapter 3, I provided the community with a methodology and a set of marker-ViPhOGs to identify and characterize thousands of transposable phages. The genomes that I identified here substantially increase the fraction of characterized phages. Thus, if this methodology is applied systematically, it might contribute to describing the viral dark matter. Furthermore, these reference genomes will be advantageous to guide culturomics studies [176]. Via guided culturomics, it would be possible to isolate transposable phages and their hosts. If achieved, based on the experience we (the scientific community) have using Mu-phage as a genetic tool [177], isolates of transposable phages from Firmicutes might serve as tools for the genetic engineering of Firmicutes inhabiting the human gut.

# Bibliography


[1] Larry L Barton and Diana E Northup. *Microbial Ecology*. John Wiley & Sons, October 2011.

[2] Sami Hamarneh. *Measuring the Invisible World*. the life and works of antoni van leeuwenhoek. a. schierbeek. Abelard-Schuman, new york, 1959. 223 pp. $4. *Science*, 132(3422):289–290, 1960.

[3] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, 74(12):5463–5467, December 1977.

[4] Carl R Woese and George E Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

[5] Alejandra Escobar-Zepeda, Arturo Vera-Ponce de León, and Alejandro Sanchez-Flores. The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. *Front. Genet.*, 6:348, December 2015.

[6] John P McCutcheon and Carol D von Dohlen. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Curr. Biol.*, 21(16):1366–1372, August 2011.

[7] Susanne Schneiker, Olena Perlova, Olaf Kaiser, Klaus Gerth, Aysel Alici, Matthias O Altmeyer, Daniela Bartels, Thomas Bekel, Stefan Beyer, Edna Bode, Helge B Bode, Christoph J Bolten, Jomuna V Choudhuri, Sabrina Doss, Yasser A Elnakady, Bettina Frank, Lars Gaigalat, Alexander Goesmann, Carolin Groeger, Frank Gross, Lars Jelsbak, Lotte Jelsbak, Jörn Kalinowski, Carsten Kegler, Tina Knauber, Sebastian Konietzny, Maren Kopp, Lutz Krause, Daniel Krug, Bukhard Linke, Taifo Mahmud, Rosa Martinez-Arias, Alice C McHardy, Michelle Merai, Folker Meyer, Sascha Mormann, Jose Muñoz-Dorado, Juana Perez, Silke Pradella, Shwan Rachid, Günter Raddatz, Frank Rosenau, Christian Rückert, Florenz Sasse, Maren Scharfe, Stephan C Schuster, Garret Suen, Anke Treuner-Lange, Gregory J Velicer, Frank-Jörg Vorhölter, Kira J Weissman, Roy D Welch, Silke C Wenzel, David E Whitworth, Susanne Wilhelm, Christoph Wittmann, Helmut Blöcker, Alfred Pühler, and Rolf Müller. Complete genome sequence of the myxobacterium sorangium cellulosum. *Nat. Biotechnol.*, 25(11):1281–1289, November 2007.

[8] Jaime Leonardo Moreno-Gallego and Alejandro Reyes. Informative regions in viral genomes. *Viruses*, 13(6):1164, June 2021.

[9] Alejandro Reyes, Nicholas P Semenkovich, Katrine Whiteson, Forest Rohwer, and Jeffrey I Gordon. Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.*, 10(9):607–617, September 2012.

[10] Nádia Conceição-Neto, Mark Zeller, Hanne Lefrère, Pieter De Bruyn, Leen Beller, Ward Deboutte, Claude Kwe Yinda, Rob Lavigne, Piet Maes, Marc Van Ranst, Elisabeth Heylen, and Jelle Matthijnssens. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci. Rep.*, 5:16532, November 2015.

[11] Siobain Duffy, Laura A Shackelton, and Edward C Holmes. Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.*, 9(4):267–276, March 2008.

[12] Samuel Minot, Stephanie Grunberg, Gary D Wu, James D Lewis, and Frederic D Bushman. Hypervariable loci in the human gut virome. *Proc. Natl. Acad. Sci. U. S. A.*, 109(10):3962–3966, March 2012.

[13] Simon Roux, Joanne B Emerson, Emiley A Eloe-Fadrosh, and Matthew B Sullivan. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ*, 5:e3817, September 2017.

[14] Rodrigo García-López, Jorge Francisco Vázquez-Castellanos, and Andrés Moya. Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. *Frontiers in Bioengineering and Biotechnology*, 3:141, 2015.

[15] Thomas D S Sutton, Adam G Clooney, Feargal J Ryan, R Paul Ross, and Colin Hill. Choice of assembly software has a critical impact on virome characterisation. *Microbiome*, 7(1):12, January 2019.

[16] Hans-W Ackermann. Tailed bacteriophages: The order caudovirales. In Karl Maramorosch, Frederick A Murphy, and Aaron J Shatkin, editors, *Advances in Virus Research*, volume 51, pages 135–201. Academic Press, January 1998.

[17] Graham F Hatfull. Bacteriophage genomics. *Curr. Opin. Microbiol.*, 11(5):447–453, October 2008.

[18] Mya Breitbart, Peter Salamon, Bjarne Andresen, Joseph M Mahaffy, Anca M Segall, David Mead, Farooq Azam, and Forest Rohwer. Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U. S. A.*, 99(22):14250–14255, October 2002.

[19] Juan Sebastián Andrade-Martínez, Laura Carolina Camelo Valera, Luis Alberto Chica Cárdenas, Laura Forero-Junco, Gamaliel López-Leal, J Leonardo Moreno-Gallego, Guillermo Rangel-Pineros, and Alejandro Reyes. Computational tools for the analysis of uncultivated phage genomes. *Microbiol. Mol. Biol. Rev.*, 86(2):e0000421, June 2022.

[20] Simon Roux, 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. Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.*, December 2018.

[21] Lora V Hooper and Andrew J Macpherson. Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nat. Rev. Immunol.*, 10(3):159–169, April 2015.

[22] Ron Sender, Shai Fuchs, and Ron Milo. Are we really vastly outnumbered? revisiting the ratio of bacterial to host cells in humans. *Cell*, 164(3):337–340, January 2016.

[23] Steven R Gill, Mihai Pop, Robert T Deboy, Paul B Eckburg, Peter J Turnbaugh, Buck S Samuel, Jeffrey I Gordon, David A Relman, Claire M Fraser-Liggett, and Karen E Nelson. Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778):1355–1359, June 2006.

[24] Seth Rakoff-Nahoum, Justin Paglino, Fatima Eslami-Varzaneh, Stephen Edberg, and Ruslan Medzhitov. Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell*, 118(2):229–241, July 2004.

[25] Harald Brüssow and Roger W Hendrix. Phage genomics: small is beautiful. *Cell*, 108(1):13–16, January 2002.

[26] Curtis A Suttle. Marine viruses–major players in the global ecosystem. *Nat. Rev. Microbiol.*, 5(10):801–812, October 2007.

[27] Lesley Hoyles, Anne L McCartney, Horst Neve, Glenn R Gibson, Jeremy D Sanderson, Knut J Heller, and Douwe van Sinderen. Characterization of virus-like particles associated with the human faecal and caecal microbiota. *Res. Microbiol.*, 165(10):803–812, December 2014.

[28] Josué L Castro-Mejía, Musemma K Muhammed, Witold Kot, Horst Neve, Charles M A P Franz, Lars H Hansen, Finn K Vogensen, and Dennis S Nielsen. Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. *Microbiome*, 3:64, November 2015.

[29] L A Ogilvie and B V Jones. The human gut virome: form and function. *Emerging Topics in Life Sciences*, 1(4):351–362, 2017.

[30] Jeremy J Barr, Rita Auro, Mike Furlan, Katrine L Whiteson, Marcella L Erb, Joe Pogliano, Aleksandr Stotland, Roland Wolkowicz, Andrew S Cutting, Kelly S Doran, Peter Salamon, Merry Youle, and Forest Rohwer. Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc. Natl. Acad. Sci. U. S. A.*, 110(26):10771–10776, June 2013.

[31] T F Thingstad. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.*, 45(6):1320–1328, 2000.

[32] Alejandro Reyes, Matthew Haynes, Nicole Hanson, Florent E Angly, Andrew C Heath, Forest Rohwer, and Jeffrey I Gordon. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, 466(7304):334–338, July 2010.

[33] Martha Rj Clokie, Andrew D Millard, Andrey V Letarov, and Shaun Heaphy. Phages in nature. *Bacteriophage*, 1(1):31–45, January 2011.

[34] Aurélie Cotillard, Sean P Kennedy, Ling Chun Kong, Edi Prifti, Nicolas Pons, Emmanuelle Le Chatelier, Mathieu Almeida, Benoit Quinquis, Florence Levenez, Nathalie Galleron, Sophie Gougis, Salwa Rizkalla, Jean-Michel Batto, Pierre Renault, ANR MicroObes consortium, Joel Doré, Jean-Daniel Zucker, Karine Clément, and Stanislav Dusko Ehrlich. Dietary intervention impact on gut microbial gene richness. *Nature*, 500(7464):585–588, August 2013.

[35] Lawrence A David, Corinne F Maurice, Rachel N Carmody, David B Gootenberg, Julie E Button, Benjamin E Wolfe, Alisha V Ling, A Sloan Devlin, Yug Varma, Michael A Fischbach, Sudha B Biddinger, Rachel J Dutton, and Peter J Turnbaugh. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563, January 2014.

[36] Samuel Minot, Rohini Sinha, Jun Chen, Hongzhe Li, Sue A Keilbaugh, Gary D Wu, James D Lewis, and Frederic D Bushman. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.*, 21(10):1616–1625, October 2011.

[37] Lesley A Ogilvie and Brian V Jones. The human gut virome: a multifaceted majority. *Front. Microbiol.*, 6:918, September 2015.

[38] Ann C Gregory, Olivier Zablocki, Ahmed A Zayed, Allison Howell, Benjamin Bolduc, and Matthew B Sullivan. The gut virome database reveals Age-Dependent patterns of virome diversity in the human gut. *Cell Host Microbe*, August 2020.

[39] Luis F Camarillo-Guerrero, Alexandre Almeida, Guillermo Rangel-Pineros, Robert D Finn, and Trevor D Lawley. Massive expansion of human gut bacteriophage diversity. *Cell*, 184(4):1098–1109.e, 2021.

[40] Bas E Dutilh, Robert Schmieder, Jim Nulton, Ben Felts, Peter Salamon, Robert A Edwards, and John L Mokili. Reference-independent comparative metagenomics using cross-assembly: crass. *Bioinformatics*, 28(24):3225–3231, December 2012.

[41] Robert A Edwards, Alejandro A Vega, Holly M Norman, Maria Ohaeri, Kyle Levi, Elizabeth A Dinsdale, Ondrej Cinek, Ramy K Aziz, Katelyn McNair, Jeremy J Barr, Kyle Bibby, Stan J J Brouns, Adrian Cazares, Patrick A de Jonge, Christelle Desnues, Samuel L Díaz Muñoz, Peter C Fineran, Alexander Kurilshikov, Rob Lavigne, Karla Mazankova, David T McCarthy, Franklin L Nobrega, Alejandro Reyes Muñoz, German Tapia, Nicole Trefault, Alexander V Tyakht, Pablo Vinuesa, Jeroen Wagemans, Alexandra Zhernakova, Frank M Aarestrup, Gunduz Ahmadov, Abeer Alassaf, Josefa Anton, Abigail Asangba, Emma K Billings, Vito Adrian Cantu, Jane M Carlton, Daniel Cazares, Gyu-Sung Cho, Tess Condeff, Pilar Cortés, Mike Cranfield, Daniel A Cuevas, Rodrigo De la Iglesia, Przemyslaw Decewicz, Michael P Doane, Nathaniel J Dominy, Lukasz Dziewit, Bashir Mukhtar Elwasila, A Murat Eren, Charles Franz, Jingyuan Fu, Cristina Garcia-Aljaro, Elodie Ghedin, Kristen M Gulino, John M Haggerty, Steven R Head, Rene S Hendriksen, Colin Hill, Heikki Hyöty, Elena N Ilina, Mitchell T Irwin, Thomas C Jeffries, Juan Jofre, Randall E Junge, Scott T Kelley, Mohammadali Khan Mirzaei, Martin Kowalewski, Deepak Kumaresan, Steven R Leigh, David Lipson, Eugenia S Lisitsyna, Montserrat Llagostera, Julia M Maritz, Linsey C Marr, Angela McCann, Shahar Molshanski-Mor, Silvia Monteiro, Benjamin Moreira-Grez, Megan Morris, Lawrence Mugisha, Maite Muniesa, Horst Neve, Nam-Phuong Nguyen, Olivia D Nigro, Anders S Nilsson, Taylor O'Connell, Rasha Odeh, Andrew Oliver, Mariana Piuri, Aaron J Prussin, Ii, Udi Qimron, Zhe-Xue Quan, Petra Rainetova, Adán Ramírez-Rojas, Raul Raya, Kim Reasor, Gillian A O Rice, Alessandro Rossi, Ricardo Santos, John Shimashita, Elyse N Stachler, Lars C Stene, Ronan Strain, Rebecca Stumpf, Pedro J Torres, Alan Twaddle, Maryann Ugochi Ibekwe, Nicolás Villagra, Stephen Wandro, Bryan White, Andy Whiteley, Katrine L Whiteson, Cisca Wijmenga, Maria M Zambrano, Henrike Zschach, and Bas E Dutilh. Global phylogeography and ancient evolution of the widespread human gut virus crassphage. *Nat Microbiol*, July 2019.

[42] Francisco Rodriguez-Valera, Ana-Belen Martin-Cuadrado, Beltran Rodriguez-Brito, Lejla Pasić, T Frede Thingstad, Forest Rohwer, and Alex Mira. Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.*, 7(11):828–836, November 2009.

[43] Jessica Ray, Michael Dondrup, Sejal Modha, Ida Helene Steen, Ruth-Anne Sandaa, and Martha Clokie. Finding a needle in the virus metagenome haystack–micrometagenome analysis captures a snapshot of the diversity of a bacteriophage armoire. *PLoS One*, 7(4):e34238, April 2012.

[44] José R Penadés, John Chen, Nuria Quiles-Puchalt, Nuria Carpena, and Richard P Novick. Bacteriophage-mediated spread of bacterial virulence genes. *Curr. Opin. Microbiol.*, 23:171–178, February 2015.

[45] Christine L Schneider. Bacteriophage-Mediated horizontal gene transfer: Transduction. In David Harper, Stephen Abedon, Benjamin Burrowes, and Malcolm McConville, editors, *Bacteriophages: Biology, Technology, Therapy*, pages 1–42. Springer International Publishing, Cham, 2017.

[46] Beltran Rodriguez-Brito, Linlin Li, Linda Wegley, Mike Furlan, Florent Angly, Mya Breitbart, John Buchanan, Christelle Desnues, Elizabeth Dinsdale, Robert Edwards, Ben Felts, Matthew Haynes, Hong Liu, David Lipson, Joseph Mahaffy, Anna Belen Martin-Cuadrado, Alex Mira, Jim Nulton, Lejla Pasić, Steve Rayhawk, Jennifer Rodriguez-Mueller, Francisco Rodriguez-Valera, Peter Salamon, Shailaja Srinagesh, Tron Frede Thingstad, Tuong Tran, Rebecca Vega Thurber, Dana Willner, Merry Youle, and Forest Rohwer. Viral and microbial community dynamics in four aquatic environments. *ISME J.*, 4(6):739–751, June 2010.

[47] Elliot J Lefkowitz, Donald M Dempsey, Robert Curtis Hendrickson, Richard J Orton, Stuart G Siddell, and Donald B Smith. Virus taxonomy: the database of the international committee on taxonomy of viruses (ICTV). *Nucleic Acids Res.*, 46(D1):D708–D717, January 2018.

[48] Peter Simmonds. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J. Gen. Virol.*, 96(Pt 6):1193–1206, June 2015.

[49] Peter Simmonds, Mike J Adams, Mária Benkő, Mya Breitbart, J Rodney Brister, Eric B Carstens, Andrew J Davison, Eric Delwart, Alexander E Gorbalenya, Balázs Harrach, Roger Hull, Andrew M Q King, Eugene V Koonin, Mart Krupovic, Jens H Kuhn, Elliot J Lefkowitz, Max L Nibert, Richard Orton, Marilyn J Roossinck, Sead Sabanadzovic, Matthew B Sullivan, Curtis A Suttle, Robert B Tesh, René A van der Vlugt, Arvind Varsani, and F Murilo Zerbini. Consensus statement: Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.*, 15(3):161–168, March 2017.

[50] A E Gorbalenya, C Lauber, and S Siddell. Taxonomy of viruses. In *Reference Module in Biomedical Sciences*. Elsevier, January 2019.

[51] International Committee on Taxonomy of Viruses Executive Committee. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat Microbiol*, 5(5):668–674, May 2020.

[52] Koonin Eugene V., Dolja Valerian V., Krupovic Mart, Varsani Arvind, Wolf Yuri I., Yutin Natalya, Zerbini F. Murilo, and Kuhn Jens H. Global organization and proposed megataxonomy of the virus world. *Microbiol. Mol. Biol. Rev.*, 84(2):e00061–19, 2020.

[53] Dann Turner, Andrew M Kropinski, and Evelien M Adriaenssens. A roadmap for Genome-Based phage taxonomy. *Viruses*, 13(3), March 2021.

[54] Gipsi Lima-Mendez, Jacques Van Helden, Ariane Toussaint, and Raphaël Leplae. Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.*, 25(4):762–777, April 2008.

[55] Jaime Iranzo, Mart Krupovic, and Eugene V Koonin. The Double-Stranded DNA virosphere as a modular hierarchical network of gene sharing. *MBio*, 7(4), August 2016.

[56] Pakorn Aiewsakun, Evelien M Adriaenssens, Rob Lavigne, Andrew M Kropinski, and Peter Simmonds. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy. *J. Gen. Virol.*, July 2018.

[57] Ho Bin Jang, 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. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.*, May 2019.

[58] Soo Jen Low, Mária Džunková, Pierre-Alain Chaumeil, Donovan H Parks, and Philip Hugenholtz. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order caudovirales. *Nat Microbiol*, May 2019.

[59] Steve Paterson, Tom Vogwill, Angus Buckling, Rebecca Benmayor, Andrew J Spiers, Nicholas R Thomson, Mike Quail, Frances Smith, Danielle Walker, Ben Libberton, Andrew Fenton, Neil Hall, and Michael A Brockhurst. Antagonistic coevolution accelerates molecular evolution. *Nature*, 464(7286):275–278, March 2010.

[60] Siddharth R Krishnamurthy and David Wang. Origins and challenges of viral dark matter. *Virus Res.*, February 2017.

[61] Gang Fang, Nitin Bhardwaj, Rebecca Robilotto, and Mark B Gerstein. Getting started in gene orthology and functional analysis. *PLoS Comput. Biol.*, 6(3):e1000703, 2010.

[62] David M Kristensen, Xixu Cai, and Arcady Mushegian. Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. *J. Bacteriol.*, 193(8):1806–1814, April 2011.

[63] Ana Laura Grazziotin, Eugene V Koonin, and David M Kristensen. Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.*, 45(D1):D491–D498, January 2017.

[64] Paul Terzian, Eric Olo Ndela, Clovis Galiez, Julien Lossouarn, Rubén Enrique Pérez Bucio, Robin Mom, Ariane Toussaint, Marie-Agnès Petit, and François Enault. PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genom Bioinform*, 3(3):lqab067, September 2021.

[65] Juan S Andrade-Martínez, J Leonardo Moreno-Gallego, and Alejandro Reyes. Defining a core genome for the herpesvirales and exploring their evolutionary relationship with the caudovirales. *Sci. Rep.*, 9(1):11342, August 2019.

[66] Yuri I Wolf, Darius Kazlauskas, Jaime Iranzo, Adriana Lucía-Sanz, Jens H Kuhn, Mart Krupovic, Valerian V Dolja, and Eugene V Koonin. Origins and evolution of the global RNA virome. *MBio*, 9(6), November 2018.

[67] Peter Skewes-Cox, Thomas J Sharpton, Katherine S Pollard, and Joseph L DeRisi. Profile hidden markov models for the detection of viruses within metagenomic sequence data. *PLoS One*, 9(8):e105067, August 2014.

[68] T Frede Thingstad, Selina Våge, Julia E Storesund, Ruth-Anne Sandaa, and Jarl Giske. A theoretical analysis of how strain-specific viruses can control microbial species diversity. *Proc. Natl. Acad. Sci. U. S. A.*, 111(21):7813–7818, May 2014.

[69] Joshua S Weitz and Jonathan Dushoff. Alternative stable states in host–phage dynamics. *Theor. Ecol.*, 1(1):13–19, March 2008.

[70] Alejandro Reyes, Meng Wu, Nathan P McNulty, Forest L Rohwer, and Jeffrey I Gordon. Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. *Proc. Natl. Acad. Sci. U. S. A.*, 110(50):20236–20241, December 2013.

[71] J Leonardo Moreno-Gallego, Shao-Pei Chou, Sara C Di Rienzi, Julia K Goodrich, Timothy D Spector, Jordana T Bell, Nicholas D Youngblut, Ian Hewson, Alejandro Reyes, and Ruth E Ley. Virome diversity correlates with intestinal microbiome diversity in adult monozygotic twins. *Cell Host Microbe*, 25(2):261–272.e5, February 2019.

[72] Julia K Goodrich, Jillian L Waters, Angela C Poole, Jessica L Sutter, Omry Koren, Ran Blekhman, Michelle Beaumont, William Van Treuren, Rob Knight, Jordana T Bell, Timothy D Spector, Andrew G Clark, and Ruth E Ley. Human genetics shape the gut microbiome. *Cell*, 159(4):789–799, November 2014.

[73] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, October 2016.

[74] Simon Roux, Mart Krupovic, Didier Debroas, Patrick Forterre, and François Enault. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.*, 3(12):130160, December 2013.

[75] Samuel Minot, Alexandra Bryson, Christel Chehoud, Gary D Wu, James D Lewis, and Frederic D Bushman. Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U. S. A.*, 110(30):12450–12455, July 2013.

[76] Efrem S Lim, Yanjiao Zhou, Guoyan Zhao, Irma K Bauer, Lindsay Droit, I Malick Ndao, Barbara B Warner, Phillip I Tarr, David Wang, and Lori R Holtz. Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.*, 21(10):1228–1234, October 2015.

[77] Pilar Manrique, Benjamin Bolduc, Seth T Walk, John van der Oost, Willem M de Vos, and Mark J Young. Healthy human gut phageome. *Proc. Natl. Acad. Sci. U. S. A.*, 113(37):10400–10405, September 2016.

[78] Andrey N Shkoporov, Feargal J Ryan, Lorraine A Draper, Amanda Forde, Stephen R Stockdale, Karen M Daly, Siobhan A McDonnell, James A Nolan, Thomas D S Sutton, Marion Dalmasso, Angela McCann, R Paul Ross, and Colin Hill. Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome*, 6(1):68, April 2018.

[79] Dietmar Fernández-Orth, Elisenda Miró, Maryury Brown-Jaque, Lorena Rodríguez-Rubio, Paula Espinal, Judith Rodriguez-Navarro, Juan José González-López, Maite Muniesa, and Ferran Navarro. Faecal phageome of healthy individuals: presence of antibiotic resistance genes and variations caused by ciprofloxacin treatment. *J. Antimicrob. Chemother.*, January 2019.

[80] David Paez-Espino, Emiley A Eloe-Fadrosh, Georgios A Pavlopoulos, Alex D Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N Ivanova, and Nikos C Kyrpides. Uncovering earth's virome. *Nature*, 536(7617):425–430, August 2016.

[81] Sean Benler, Natalya Yutin, Dmitry Antipov, Mikhail Rayko, Sergey Shmakov, Ayal B Gussow, Pavel Pevzner, and Eugene V Koonin. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome*, 9(1):78, March 2021.

[82] Junhua Li, Huijue Jia, Xianghang Cai, Huanzi Zhong, Qiang Feng, Shinichi Sunagawa, Manimozhiyan Arumugam, Jens Roat Kultima, Edi Prifti, Trine Nielsen, Agnieszka Sierakowska Juncker, Chaysavanh Manichanh, Bing Chen, Wenwei Zhang, Florence Levenez, Juan Wang, Xun Xu, Liang Xiao, Suisha Liang, Dongya Zhang, Zhaoxi Zhang, Weineng Chen, Hailong Zhao, Jumana Yousuf Al-Aama, Sherif Edris, Huanming Yang, Jian Wang, Torben Hansen, Henrik Bjørn Nielsen, Søren Brunak, Karsten Kristiansen, Francisco Guarner, Oluf Pedersen, Joel Doré, S Dusko Ehrlich, MetaHIT Consortium, Peer Bork, Jun Wang, and MetaHIT Consortium. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.*, 32(8):834–841, August 2014.

[83] Mya Breitbart, Matthew Haynes, Scott Kelley, Florent Angly, Robert A Edwards, Ben Felts, Joseph M Mahaffy, Jennifer Mueller, James Nulton, Steve Rayhawk, Beltran Rodriguez-Brito, Peter Salamon, and Forest Rohwer. Viral diversity and dynamics in an infant gut. *Res. Microbiol.*, 159(5):367–373, June 2008.

[84] Bas E Dutilh, Noriko Cassman, Katelyn McNair, Savannah E Sanchez, Genivaldo G Z Silva, Lance Boling, Jeremy J Barr, Daan R Speth, Victor Seguritan, Ramy K Aziz, Ben Felts, Elizabeth A Dinsdale, John L Mokili, and Robert A Edwards. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.*, 5:4498, July 2014.

[85] Natalya Yutin, Kira S Makarova, Ayal B Gussow, Mart Krupovic, Anca Segall, Robert A Edwards, and Eugene V Koonin. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol*, 3(1):38–46, January 2018.

[86] Anna J Székely and Mya Breitbart. Single-stranded DNA phages: from early molecular biology tools to recent revolutions in environmental microbiology. *FEMS Microbiol. Lett.*, 363(6), March 2016.

[87] Moreno Zolfo, Federica Pinto, Francesco Asnicar, Paolo Manghi, Adrian Tett, Frederic D Bushman, and Nicola Segata. Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.*, November 2019.

[88] Steven J Biller, Florence Schubotz, Sara E Roggensack, Anne W Thompson, Roger E Summons, and Sallie W Chisholm. Bacterial vesicles in marine ecosystems. *Science*, 343(6167):183–186, January 2014.

[89] Ariane Toussaint. Transposable mu-like phages in firmicutes: new instances of divergence generating retroelements. *Res. Microbiol.*, 164(4):281–287, May 2013.

[90] John Chen, Nuria Quiles-Puchalt, Yin Ning Chiang, Rodrigo Bacigalupe, Alfred Fillol-Salom, Melissa Su Juan Chee, J Ross Fitzgerald, and José R Penadés. Genome hypermobility by lateral transduction. *Science*, 362(6411):207–212, October 2018.

[91] B Marrs. Genetic recombination in rhodopseudomonas capsulata. *Proc. Natl. Acad. Sci. U. S. A.*, 71(3):971–973, March 1974.

[92] Andrew S Lang, Olga Zhaxybayeva, and J Thomas Beatty. Gene transfer agents: phage-like elements of genetic exchange. *Nat. Rev. Microbiol.*, 10(7):472–482, June 2012.

[93] Pilar Manrique, Michael Dills, and Mark J Young. The human gut phage community and its implications for health and disease. *Viruses*, 9(6):10.3390/v9060141, June 2017.

[94] Philippe Colson, Laura Fancello, Gregory Gimenez, Fabrice Armougom, Christelle Desnues, Ghislain Fournous, Niyaz Yoosuf, Matthieu Million, Bernard La Scola, and Didier Raoult. Evidence of the megavirome in humans. *J. Clin. Virol.*, 57(3):191–200, July 2013.

[95] S Halary, S Temmam, D Raoult, and C Desnues. Viral metagenomics: are we missing the giants? *Curr. Opin. Microbiol.*, 31:34–43, June 2016.

[96] Qiulong Yan, Yu Wang, Xiuli Chen, Hao Jin, Guangyang Wang, Kuiqing Guan, Yue Zhang, Pan Zhang, Taj Ayaz, Yanshan Liang, Junyi Wang, Guangyi Cui, Yuanyuan Sun, Manchun Xiao, Jian Kang, Wei Zhang, Aiqin Zhang, Peng Li, Xueyang Liu, Hayan Ulllah, Yufang Ma, Shenghui Li, and Tonghui Ma. Characterization of the gut DNA and RNA viromes in a cohort of chinese residents and visiting pakistanis. *Virus Evol*, 7(1):veab022, January 2021.

[97] Chuen Zhang Lee, Muhammad Zarul Hanifah Md Zoqratt, Maude E Phipps, Jeremy J Barr, Sunil K Lal, Qasim Ayub, and Sadequr Rahman. The gut virome in two indigenous populations from malaysia. *Sci. Rep.*, 12(1):1824, February 2022.

[98] Emma Guerin, Andrey Shkoporov, Stephen Stockdale, Adam G Clooney, Feargal J Ryan, Thomas D S Sutton, Lorraine A Draper, Enrique Gonzalez-Tortuero, R Paul Ross, and Colin Hill. Biology and taxonomy of crass-like bacteriophages, the most abundant virus in the human gut. April 2018.

[99] Mart Krupovic and Patrick Forterre. Microviridae goes temperate: microvirus-related proviruses reside in the genomes of bacteroidetes. *PLoS One*, 6(5):e19893, May 2011.

[100] Simon Roux, Mart Krupovic, Axel Poulet, Didier Debroas, and François Enault. Evolution and diversity of the microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS One*, 7(7):e40418, July 2012.

[101] Zachary Deng and Eric Delwart. ContigExtender: a new approach to improving de novo sequence assembly for viral metagenomics data. *BMC Bioinformatics*, 22(1):119, March 2021.

[102] Sunghee Lee, Joohon Sung, Jungeun Lee, and Gwangpyo Ko. Comparison of the gut microbiotas of healthy adult twins living in south korea and the united states. *Appl. Environ. Microbiol.*, 77(20):7433–7437, October 2011.

[103] Chana Palmer, Elisabeth M Bik, Daniel B DiGiulio, David A Relman, and Patrick O Brown. Development of the human infant intestinal microbiota. *PLoS Biol.*, 5(7):e177, July 2007.

[104] Sebastian Tims, Catherine Derom, Daisy M Jonkers, Robert Vlietinck, Wim H Saris, Michiel Kleerebezem, Willem M de Vos, and Erwin G Zoetendal. Microbiota conservation and BMI signatures in adult monozygotic twins. *ISME J.*, 7(4):707–717, April 2013.

[105] Tanya Yatsunenko, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N Baldassano, Andrey P Anokhin, Andrew C Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J Gregory Caporaso, Catherine A Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight, and Jeffrey I Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, May 2012.

[106] Alejandro Reyes, 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. Gut DNA viromes of malawian twins discordant for severe acute malnutrition. *Proc. Natl. Acad. Sci. U. S. A.*, 112(38):11941–11946, September 2015.

[107] Marcus J Claesson, Ian B Jeffery, Susana Conde, Susan E Power, Eibhlís M O'Connor, Siobhán Cusack, Hugh M B Harris, Mairead Coakley, Bhuvaneswari Lakshminarayanan, Orla O'Sullivan, Gerald F Fitzgerald, Jennifer Deane, Michael O'Connor, Norma Harnedy, Kieran O'Connor, Denis O'Mahony, Douwe van Sinderen, Martina Wallace, Lorraine Brennan, Catherine Stanton, Julian R Marchesi, Anthony P Fitzgerald, Fergus Shanahan, Colin Hill, R Paul Ross, and Paul W O'Toole. Gut microbiota composition correlates with diet and health in the elderly. *Nature*, 488(7410):178–184, August 2012.

[108] Carlotta De Filippo, Duccio Cavalieri, Monica Di Paola, Matteo Ramazzotti, Jean Baptiste Poullet, Sebastien Massart, Silvia Collini, Giuseppe Pieraccini, and Paolo Lionetti. Impact of diet in shaping gut microbiota revealed by a comparative study in children from europe and rural africa. *Proc. Natl. Acad. Sci. U. S. A.*, 107(33):14691–14696, August 2010.

[109] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359, March 2012.

[110] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[111] Jacob H Munson-McGee, Shengyun Peng, Samantha Dewerff, Ramunas Stepanauskas, Rachel J Whitaker, Joshua S Weitz, and Mark J Young. A virus or more in (nearly) every cell: ubiquitous networks of virus-host interactions in extreme environments. *ISME J.*, February 2018.

[112] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, 12(1):59–60, January 2015.

[113] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, 7(5):335–336, May 2010.

[114] Binbin Lai, Fumeng Wang, Xiaoqi Wang, Liping Duan, and Huaiqiu Zhu. InteMAP: Integrated metagenomic assembly pipeline for NGS short reads. *BMC Bioinformatics*, 16:244, August 2015.

[115] Yang Liao, Gordon K Smyth, and Wei Shi. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.*, 41(10):e108, May 2013.

[116] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, March 2004.

[117] Konstantin Okonechnikov, Olga Golosova, Mikhail Fursov, and UGENE team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, 28(8):1166–1167, April 2012.

[118] S R Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

[119] João M P Alves, André L de Oliveira, Tatiana O M Sandberg, Jaime L Moreno-Gallego, Marcelo A F de Toledo, Elisabeth M M de Moura, Liliane S Oliveira, Alan M Durham, Dolores U Mehnert, Paolo M de A Zanotto, Alejandro Reyes, and Arthur Gruber. GenSeed-HMM: A tool for progressive assembly using profile HMMs as seeds and its application in alpavirinae viral discovery from metagenomic data. *Front. Microbiol.*, 7:269, March 2016.

[120] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, May 2014.

[121] Diego Darriba, Guillermo L Taboada, Ramón Doallo, and David Posada. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–1165, April 2011.

[122] Mark Borodovsky and Alex Lomsadze. Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. *Curr. Protoc. Microbiol.*, 32:Unit 1E.7., February 2014.

[123] Daniel H Huson, Suparna Mitra, Hans-Joachim Ruscheweyh, Nico Weber, and Stephan C Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.*, 21(9):1552–1560, September 2011.

[124] Daehwan Kim, Li Song, Florian P Breitwieser, and Steven L Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.*, 26(12):1721–1729, December 2016.

[125] Hideo Ikeda and Jun-Ichi Tomizawa. Transducing fragments in generalized transduction by phage P1. *J. Mol. Biol.*, 14(1):85–109, November 1965.

[126] M M Howe. Transduction by bacteriophage MU-1. *Virology*, 55(1):103–117, September 1973.

[127] Chantal Hulo, Patrick Masson, Philippe Le Mercier, and Ariane Toussaint. A structured annotation frame for the transposable phages: a new proposed family "saltoviridae" within the caudovirales. *Virology*, 477:155–163, March 2015.

[128] M Faelen, O Huisman, and A Toussaint. Involvement of phage mu-1 early functions in mu-mediated chromosomal rearrangements. *Nature*, 271(5645):580–582, February 1978.

[129] Rasika M Harshey. The mu story: how a maverick phage moved the field forward. *Mob. DNA*, 3(1):21, December 2012.

[130] Gipsi Lima-Mendez, Ariane Toussaint, and Raphael Leplae. A modular view of the bacteriophage genomic space: identification of host and lifestyle marker modules. *Res. Microbiol.*, 162(8):737–746, October 2011.

[131] Jeffrey K Cornuault, Marie-Agnès Petit, Mahendra Mariadassou, Leandro Benevides, Elisabeth Moncaut, Philippe Langella, Harry Sokol, and Marianne De Paepe. Phages infecting faecalibacterium prausnitzii belong to novel viral genera that help to decipher intestinal viromes. *Microbiome*, 6(1):65, April 2018.

[132] Jaime Huerta-Cepas, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, and Peer Bork. Fast Genome-Wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.*, 34(8):2115–2122, August 2017.

[133] Ariane Toussaint and Phoebe A Rice. Transposable phages, DNA reorganization and transfer. *Curr. Opin. Microbiol.*, 38:88–94, August 2017.

[134] David Paez-Espino, Simon Roux, I-Min A Chen, Krishna Palaniappan, Anna Ratner, Ken Chu, Marcel Huntemann, T B K Reddy, Joan Carles Pons, Mercè Llabrés, Emiley A Eloe-Fadrosh, Natalia N Ivanova, and Nikos C Kyrpides. IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res.*, 47(D1):D678–D686, January 2019.

[135] Carol L Ecale Zhou, Stephanie Malfatti, Jeffrey Kimbrel, Casandra Philipson, Katelyn McNair, Theron Hamilton, Robert Edwards, and Brian Souza. multiPhATE: bioinformatics pipeline for functional annotation of phage isolates. *Bioinformatics*, 35(21):4402–4404, November 2019.

[136] Pakorn Aiewsakun and Peter Simmonds. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. *Microbiome*, 6(1):38, February 2018.

[137] Cameron L M Gilchrist and Yit-Heng Chooi. Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics*, January 2021.

[138] Guillaume Gautreau, Adelme Bazin, Mathieu Gachet, Rémi Planel, Laura Burlot, Mathieu Dubois, Amandine Perrin, Claudine Médigue, Alexandra Calteau, Stéphane Cruveiller, Catherine Matias, Christophe Ambroise, Eduardo P C Rocha, and David Vallenet. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput. Biol.*, 16(3):e1007732, March 2020.

[139] R L Tatusov, E V Koonin, and D J Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, October 1997.

[140] Michael Y Galperin, Kira S Makarova, Yuri I Wolf, and Eugene V Koonin. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, 43(Database issue):D261–9, January 2015.

[141] Chantal Hulo, Patrick Masson, Ariane Toussaint, David Osumi-Sutherland, Edouard de Castro, Andrea H Auchincloss, Sylvain Poux, Lydie Bougueleret, Ioannis Xenarios, and Philippe Le Mercier. Bacterial virus ontology; coordinating across databases. *Viruses*, 9(6), May 2017.

[142] D Botstein. A theory of modular evolution for bacteriophages. *Ann. N. Y. Acad. Sci.*, 354:484–490, 1980.

[143] J Park, K Karplus, C Barrett, R Hughey, D Haussler, T Hubbard, and C Chothia. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, 284(4):1201–1210, December 1998.

[144] Martin Steinegger, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J Haunsberger, and Johannes Söding. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, 20(1):1–15, September 2019.

[145] Mart Krupovic, David Prangishvili, Roger W Hendrix, and Dennis H Bamford. Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiol. Mol. Biol. Rev.*, 75(4):610–635, December 2011.

[146] Mart Krupovic, Valerian V Dolja, and Eugene V Koonin. Origin of viruses: primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.*, May 2019.

[147] Bas E Dutilh, Arvind Varsani, Yigang Tong, Peter Simmonds, Sead Sabanadzovic, Luisa Rubino, Simon Roux, Alejandro Reyes Muñoz, Cédric Lood, Elliot J Lefkowitz, Jens H Kuhn, Mart Krupovic, Robert A Edwards, J Rodney Brister, Evelien M Adriaenssens, and Matthew B Sullivan. Perspective on taxonomic classification of uncultivated viruses. *Curr. Opin. Virol.*, 51:207–215, December 2021.

[148] Rob Lavigne, Paul Darius, Elizabeth J Summer, Donald Seto, Padmanabhan Mahadevan, Anders S Nilsson, Hans W Ackermann, and Andrew M Kropinski. Classification of myoviridae bacteriophages using protein sequence similarity. *BMC Microbiol.*, 9:224, October 2009.

[149] Nina Bartlau, Antje Wichels, Georg Krohne, Evelien M Adriaenssens, Anneke Heins, Bernhard M Fuchs, Rudolf Amann, and Cristina Moraru. Highly diverse flavobacterial phages isolated from north sea spring blooms. *ISME J.*, September 2021.

[150] David Paez-Espino, Georgios A Pavlopoulos, Natalia N Ivanova, and Nikos C Kyrpides. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nat. Protoc.*, 12(8):1673–1682, August 2017.

[151] Stephen Nayfach, Antonio Pedro Camargo, Frederik Schulz, Emiley Eloe-Fadrosh, Simon Roux, and Nikos C Kyrpides. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.*, 39(5):578–585, May 2021.

[152] Dmitry Antipov, Mikhail Raiko, Alla Lapidus, and Pavel A Pevzner. Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics*, 36(14):4126–4129, August 2020.

[153] Peter Simmonds. A clash of ideas – the varying uses of the 'species' term in virology and their utility for classifying viruses in metagenomic datasets. *J. Gen. Virol.*, 99(3):277–287, February 2018.

[154] Aamir W Khan, Vanika Garg, Manish Roorkiwal, Agnieszka A Golicz, David Edwards, and Rajeev K Varshney. Super-Pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends Plant Sci.*, 25(2):148–158, February 2020.

[155] Emanuele Bosi, Renato Fani, and Marco Fondi. Defining orthologs and pangenome size metrics. In Alessio Mengoni, Marco Galardini, and Marco Fondi, editors, *Bacterial Pangenomics: Methods and Protocols*, pages 191–202. Springer New York, New York, NY, 2015.

[156] Edel Stone, Katrina Campbell, Irene Grant, and Olivia McAuliffe. Understanding and exploiting Phage-Host interactions. *Viruses*, 11(6), June 2019.

[157] Juliano Bertozzi Silva, Zachary Storms, and Dominic Sauvageau. Host receptors for bacteriophage adsorption. *FEMS Microbiol. Lett.*, 363(4), February 2016.

[158] Eugene V Koonin. Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Res.*, 5, July 2016.

[159] Eitan Yaffe and David A Relman. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. *Nat Microbiol*, 5(2):343–353, February 2020.

[160] Aaron Lerner, Torsten Matthias, and Rustam Aminov. Potential effects of horizontal gene exchange in the human gut. *Front. Immunol.*, 8:1630, November 2017.

[161] Cristina Moraru, Arvind Varsani, and Andrew M Kropinski. VIRIDIC – a novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. July 2020.

[162] Dean Laslett and Bjorn Canback. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.*, 32(1):11–16, January 2004.

[163] G Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, 27(2):573–580, January 1999.

[164] Peter E Warburton, Joti Giordano, Fanny Cheung, Yefgeniy Gelfand, and Gary Benson. Inverted repeat structure of the human genome: the x-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res.*, 14(10A):1861–1869, October 2004.

[165] Peter J A Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J L de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, June 2009.

[166] Wes McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*. SciPy, 2010.

[167] T Pandas development team. pandas-dev/pandas: Pandas. *Zenodo*.

[168] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R J Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E A Quintero, Charles R Harris, Anne M Archibald, Antônio H Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods*, 17(3):261–272, March 2020.

[169] Michael Waskom. seaborn: statistical data visualization. *J. Open Source Softw.*, 6(60):3021, April 2021.

[170] J D Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9:90–95, May 2007.

[171] Diego Darriba, David Posada, Alexey M Kozlov, Alexandros Stamatakis, Benoit Morel, and Tomas Flouri. ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.*, 37(1):291–294, August 2019.

[172] Ivica Letunic and Peer Bork. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, 49(W1):W293–W296, July 2021.

[173] Carolina Megumi Mizuno, Francisco Rodriguez-Valera, Nikole E Kimes, and Rohit Ghai. Expanding the marine virosphere using metagenomics. *PLoS Genet.*, 9(12):e1003987, December 2013.

[174] Yong-Zhen Zhang, Mang Shi, and Edward C Holmes. Using metagenomics to characterize an expanding virosphere. *Cell*, 172(6):1168–1172, March 2018.

[175] Uri Neri, Yuri I Wolf, Simon Roux, Antonio Pedro Camargo, Benjamin Lee, Darius Kazlauskas, I Min Chen, Natalia Ivanova, Lisa Zeigler Allen, David Paez-Espino, Donald A Bryant, Devaki Bhaya, RNA Virus Discovery Consortium, Mart Krupovic, Valerian V Dolja, Nikos C Kyrpides, Eugene V Koonin, and Uri Gophna. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell*, September 2022.

[176] Ami Diakite, Grégory Dubourg, Niokhor Dione, Pamela Afouda, Sara Bellali, Issa Isaac Ngom, Camille Valles, Mamadou Lamine Tall, Jean-Christophe Lagier, and Didier Raoult. Optimization and standardization of the culturomics technique for human microbiome exploration. *Sci. Rep.*, 10(1):9674, June 2020.

[177] Ariane Toussaint. Transposable bacteriophages as genetic tools. In Martha R J Clokie, Andrew M Kropinski, and Rob Lavigne, editors, *Bacteriophages: Methods and Protocols, Volume 3*, pages 263–278. Springer New York, New York, NY, 2018.

# Appendices

## .1 Viromes sample information

Table including additional information pertaining to the 21 selected MZ twin pairs (metadata), and counts of viromes reads and contigs per sample. https://docs.google.com/spreadsheets/d/1uKp2d8nizPZSUV7dVaN2QhvZFqatvpGz/edit?usp=sharing&ouid=112821969111568457201&rtpof=true&sd=true

## .2 Transposable phages in NCBI

Table inncluding all information about the transposable phages identified in NCBI. It includes all accession numbers, if the accession was inncluded in Hulo's studty [127] ("in Hulo" column), if the accession was included as reference in the set of putative transposable phage representatives ("TP-representative), the clade to which each accession belongs ("Clade" column), and other self-explained information. https://docs.google.com/spreadsheets/d/1jr479TNEEh54ROVyaqsVA3XecLKmRIPO/edit?usp=sharing&ouid=112821969111568&rtpof=true&sd=true

## .3 Putative transposable phages

Table inncluding all information about the putative transposable phages identified in NCBI, IMG/VR, and GPD. It includes all accession numbers, its origin database, the quality of the assembly, and the predicted host (if any). https://docs.google.com/spreadsheets/d/14wRryqe8Kp6lN_MJOf6vTt_pbOitaFc1/edit?usp=sharing&ouid=112821969111568&rtpof=true&sd=true