UNIVERSITY OF TRENTO - Italy

Department CIBIO

International PhD Program in Biomolecular Sciences

Department of Cellular, Computational and Integrative Biology – CIBIO

XXXII Cycle

Metagenomics-based discovery of unknown bacteriophages In the human microbiome

Tutor

Prof. Nicola Segata

CIBIO, University of Trento

Ph.D. Thesis of

Moreno Zolfo

Department of Cellular, Computational and Integrative Biology

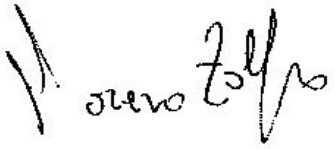
University of Trento

Academic Year 2018-2019

Declaration

I, **Moreno Zolfo**, confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Trento, 28th June 2020



Abstract

Viruses, and particularly bacteriophages, are key players in many microbial ecosystems and can profoundly influence the human microbiome and its impact on human health. While the bacterial and archaeal fraction of the human microbiome can now be profiled at an unprecedented resolution via cultivation-free metagenomics, viral metagenomics is still extremely challenging. The lack of universal viral genetic markers limits the de-novo discovery of viral entities, and the low number of available viral reference genomes from cultivation studies does not cover well the phage diversity in human microbiome samples

Viral-like particle (VLP) purification has been proposed as a set of experimental tools to concentrate viruses in samples prior to sequencing, but it remains unclear how efficient and reproducible such tools are in practice. In this thesis we aim to address some of these challenges and better exploit the potential of viral metagenomics in the context of the human microbiome. First, we performed and studied the performance of VLP procedures on freshwater and sediment samples. We found that bacteria can still be abundant at the end of the filtration process, thus lowering the efficiency of the enrichment. Analyzing samples with a low enrichment may lead to inconsistent conclusions, as the residual bacterial contamination might misdirect the computational analysis. To better quantify the extent of non-viral contamination in VLP sequencing, we designed ViromeQC, a novel open-source tool able to assess and rank viromes by their viral purity directly from the raw reads. In ViromeQC, rRNA genes and bacterial single-copy proteins are used as a proxy to estimate non-viral contamination. With the ViromeQC, we conducted the largest meta-analysis on the degree of enrichment of thousands of viral metagenomes, and concluded that the vast majority of them are three-fold less enriched than a standard metagenome. ViromeQC was then used to select the human gut viromes that had the highest enrichment as a starting point for a novel reference-free pipeline for the discovery of previously uncharacterized viral entities. The approach included metagenomic assembly of the enriched viromes as well as extensive mining of many thousands of assembled metagenomes, and led to a catalog of 162,876 sequences of highly-trusted viral origin. Most of these predicted viral sequences had no match against any known virus in RefSeq even though some of them showed a prevalence in gut metagenomes of up to 70%. Our analyses and publicly available tools and resources are helping to uncover the still hidden virome diversity and improve the support for current and future investigations of the human virome.

Index

List of Abbreviations

VLP: Viral Like Particle
ITS: Internal Transcribed Spacer
OTU: Operational Taxonomic Unit
MDS: Multi Dimensional Scaling
NMDS: Non-metric Multi Dimensional Scaling
SSU-rRNA: large ribosomal subunit rRNA (gene)
LSU-rRNA: small ribosomal subunit rRNA (gene)
CPR: Candidate phyla radiation
VSC: Viral Sequence Cluster
HEVC: Highly Enriched Viral Contig
MAG: Metagenome-Assembled Genome
SGB: Species-level Genome Bin
SNV: Single nucleotide variant
MLST: Multi Locus Sequence Typing
ST: Sequence Type
MST: Minimum Spanning tree
IQR: Inter Quantile Range
PCA: Principal Component Analysis

Chapter 1

Introduction and overall summary

Chapter 1 | Introduction and overall summary

Viruses are obligate parasites that are able to replicate uniquely within living cells of other organisms. Viral particles are composed of a protein shell protecting their genetic information, stored in single- or double-stranded DNA or RNA molecules, and can be surrounded by a phospholipid bilayer. While they are not considered life-forms, viruses are the most abundant biological entity on the planet, with 10^{31} viruses estimated to be present on Earth at any given time (1, 2), far surpassing the abundance of any other life-form. Natural waters, for example, can harbor as many as 10^{8} viruses per milliliter (2–4), and soil is an enormous reservoir of viral genetic diversity, with up to 10^{9} viruses per dry gram (5). Viruses can remain viable after long-range atmospheric transport through the upper troposphere (6) and can infect a great variety of life-forms, including animals, plants, bacteria, archaea, and humans.

Bacteriophages (or simply phages) are viruses that infect bacteria and are key players in microbial ecology. For example, lateral-gene transfer, an important driver of bacterial evolution, is often mediated by bacteriophages, which are capable of transporting genetic elements across bacterial strains and species (7). Indeed, bacteriophage tropism towards bacteria is often so specific that phages that infect one species or strain may be incapable of infecting others. As a result, the possibility to exploit phages to treat infections from antibiotic-resistant bacteria in clinical contexts has been investigated multiple times (8, 9) and remains one of the most promising approaches to fight antibiotic resistance and to avoid the side effects of wide-spectrum antibiotics.

Numerous studies addressed how phages influence the equilibrium of bacterial populations within microbial communities. In some models, phages rapidly infect their target species in an attempt to replicate and propagate fast, and engage an evolutionary “arms race” against their bacterial counterparts (10). It was estimated that this continuous predatory cycle is responsible for the daily loss of as much as 20% of the biomass in oceans (2). In other models of bacterial-host interactions, temperate phages tend to integrate into their target’s genome as pro-phages and replicate only at a later stage. Temperate phages can contribute to maintaining the genetic diversity in species through transduction-mediated lateral gene transfer, and prevent other phages to super-infect their targets. (11–13).

Despite their role as a major source of health concerns, most viruses are non-pathogenic. Indeed, viruses are an essential component of the human microbiome, which is the collection of all micro-organisms that are associated with humans. Although research on these aspects

is at its infancy, it is already widely accepted that a substantial part of the human microbiome is composed by viral particles, and those account for more than 5% of the total DNA in the gut, with more than 10^9 viral particles per grams of feces (14, 15).

The human microbiome has been strongly associated with many conditions and diseases in human health. In the last few decades, bacterial (16–18), fungal (19, 20), eukaryotic (21, 22), as well as viral (23) components of the microbiome have been linked to numerous aspects of human health and lifestyle. For instance, specific microbial signatures are found in patients with diabetes (24), colorectal cancer (25, 26), psoriasis (19) and periodontitis (27). Important microbiome changes were associated with dietary habits (28), age and development (29) and the microbiome composition is crucial in patients' responses to some drugs and therapies (30, 31).

Metagenomics is the sequencing of the overall genetic content of a sample and allows to qualify and quantify all the microbes within a sample (32–34). Metagenomics unlocked the possibility to study the microbiome, particularly uncultivable or extremely low abundant bacteria, without the need to isolate every single species. With the exponential decrease of sequencing costs and the development of novel computational approaches to analyze microbial sequences, metagenomics now allows to reach an unprecedented resolution and throughput in the study of microbial genomics. To date, microbes in the microbiome can be followed at the resolution of single strains (35, 36) and prevalent uncultivated species can be discovered directly from metagenomic samples (37). However, while metagenomics has been extensively used to characterize the human microbiome, the viral fraction of the human microbiome remains widely understudied.

Several limitations arise when metagenomics is used to study viruses. First, there is no universal marker for viruses. The lack of markers such as the 16S rRNA gene in bacteria or the Internal transcribed spacer (ITS) region between the large and small subunits of the rRNA genes in fungi, as well as the absence of conserved proteins in viruses, makes it difficult to identify viral sequences from raw DNA reads (38). This means that viruses can be identified only by direct or indirect homology relation to other similar and previously observed viral genomes. In fact, to date only 12,194 viral genomes are available in RefSeq (39), which is at least one order of magnitude lower than the number of bacterial and micro-eukaryotes. Moreover, the available genomes are biased towards pathogens of clinical relevance, with only a few thousands of phage reference genomes available. Because computational tools to identify microbes rely on sequence similarity to known genomes and protein, with so few representative references and such a high diversity, viruses that diverge substantially from already known genomes can easily be transparent to the computational eye (12, 40).

In the last few years numerous studies mined large metagenomic datasets to retrieve novel microbial genomes. Through a technique known as metagenomic assembly, short reads are assembled into longer sequences called contigs, and then binned into microbial full genomes. While this has led to expanded and high-quality atlases of hundreds of thousands of previously unknown microbial species (37, 41, 42), such endeavors were however not focused on viruses.

Metagenomics has nonetheless been applied to viruses as well, and while most of the research has been targeted towards double-stranded DNA viruses, also RNA viruses are starting to be studied with metatranscriptomics (43). However, the quality and confidence of the viral genomes discovered with metagenomics do not match those of metagenomically assembled bacterial genomes. Many studies described the viral communities in the human gut - referred to as the “virome” - of healthy individuals (44, 45), explored the development of the human virome at birth (46, 47), and identified links between the virome and human diseases (48, 49). Novel viruses have also been characterized by metagenomic approaches (50, 51) and were later confirmed independently with cultivation based methods. Indeed, although discovering new viral species through metagenomics is possible, the identification and validation of such species from the millions of contigs often contained in a metagenome is extremely difficult. This is greatly due to the lack of common features to define what a virus looks like and it greatly limits the application of viral profiling in metagenomics.

When metagenomics is used to study viruses in complex communities, sequencing libraries are often enriched for viruses prior to sequencing. This is known as Viral Like Particle (VLP) enrichment, and aims to mitigate the problem of the low viral biomass in metagenomes through the sequential filtration and concentration of the smaller particles in a sample. The basis of VLP-enrichment is the mechanical filtration of particles smaller than a certain size (typically 0.22 to 0.45 microns). Some approaches also concentrate viral particles through tangential-flow filtration or filter them through a gradient of cesium chloride (23, 52). Numerous protocols exist, each with their own peculiarities and biases, and each approach is specific to certain viruses (e.g. certain sizes or densities, single- or double-stranded DNA viruses, and so on). VLP enrichment has been extensively used as a starting point for several virome studies, many of which culminated in metagenomic assembly and viral characterization of both known and unknown viruses.

An important aspect of VLP enrichment is the degree of purity of the final sequencing library (i.e. the concentration of nucleic acids of viral origin). This is because the higher is the purity of the virome, the lower will be the amount of bacterial, archaeal, fungal, and eukaryotic sequences that can be considered “contaminants”. In a highly pure virome, the

raw reads can be more confidently labeled as viral, therefore increasing the reliability of any viral-discovery approach. However, the efficiency of enrichment is seldomly taken into consideration after sequencing (53, 54). This can result in potentially wrong conclusions when highly contaminated viromes are analyzed under the assumption that they contain only viral sequences (55). Some studies may test for known contaminants with ad-hoc PCRs on the extracted nucleic acids (e.g. 16S rRNA to estimate bacterial contamination), but contamination may still be present in the sequenced metagenomic library.

Currently, there is no general consensus on how to assess the enrichment of a virome, despite the importance of this aspect in the downstream computational analysis.

1.1 Aim of the thesis

In this thesis, we examine viromes with the ultimate goal of discovering and cataloguing at a large scale novel viral diversity present in the human microbiome. We aim to do that by incorporating a novel concept of viral enrichment quantification and by exploiting via variants of metagenomic assembly and binning the large collection of several thousands of virome and microbiome metagenomic samples available. Specifically, we focused on the following aims:

1. In the scope of an ongoing study on the viral communities of alpine lakes, we wanted to analyze the outcome of the enrichment process in a typical virome study. The goal is to understand what and to what extent non-viral particles can bypass the filters used in the enrichment and to set up operative guidelines to perform such experiments and appropriately interpret the sequencing data they produce.

2. Currently, there is no consensus regarding the evaluation of VLP enrichment efficiency. Therefore, we aimed to develop a computational, easy to use, and open-source software to quantify viral enrichment directly from raw metagenomic reads, and use it to comprehensively assess the quality and the thousands of available VLP-enriched metagenomic samples.

3. Metagenomic assembly is a powerful technique to uncover new strains and species. We aimed to incorporate the concept of viral-enrichment into a de-novo discovery pipeline for new viruses in the human microbiome, start the characterization of some of the most prevalent newly discovered viruses, and to release a curated collection of novel viral sequences as a resource for other studies and researchers.

1.2 Summary of the Results

In this thesis, I explored different aspects of Viral Like Particles (VLP) enrichment applied to virome metagenomics, with the overall goal to reliably retrieve new viral sequences from publicly available viromes.

In a pilot study conducted in collaboration with Dr. Federica Pinto for the experimental part, we analyzed the nature of non-viral contamination in VLP preparations of water and sediment samples from an alpine lake. Shotgun metagenomics on the final VLP preparations confirmed that bacterial DNA was still present even after careful viral enrichment and DNAse treatment. In lake sediments the enrichment was much lower, suggesting that the intrinsic characteristics of a sample, such as the presence of particles and minerals in the sample matrix, can greatly influence the efficiency of enrichment. Through 16S rRNA gene amplicon sequencing on the filters used to enrich viromes, we were able to track the bacterial taxa that were retained in each step of the sequential filtration protocol. We showed that each filter size was associated with a unique community with distinct ecological attributes. Moreover, bacterial taxa characterized by cell sizes close to the smallest pore size could be detected in the final virome. Overall these results highlight that the source of non-viral contamination in VLP viromes is compatible with the hypothesis of bacterial cells bypassing the enrichment filters. Our findings call for careful considerations on the final VLP viromes, as not all the sequences retrieved in those samples may be of viral origin.

Non-viral contamination appears to be an unavoidable problem of VLP preparations. Moreover, the last available meta-analysis on the phenomenon only included a few samples (53). To assess the degree of contamination in published viromes, we developed ViromeQC, a computational tool to quantify contamination via the number of reads mapping against bacterial and archaeal universal markers. In our meta-analysis of >2000 viromes, we showed that most samples were less than three times enriched if compared to a standard (i.e. unenriched) metagenome. As a low enrichment can induce spurious conclusions, we urge the community to exercise caution when interpreting virome sequencing results and to include contamination assessment into bioinformatic analyses.

Finally, we exploited the published human gut viromes that were highly enriched to drive the discovery of previously unobserved viral sequences. Metagenomic assembly of the 199 most enriched samples allowed us to recover 162,876 high-confidence viral sequences that we then grouped into 3,944 viral sequence clusters (VSCs). We reconstructed the phylogeny

of many prevalent VSCs, and we also retrieved previously observed viruses such as crAssphage. The VSCs were organized in a public release that can be a useful resource to identify and track gut viruses in future studies.

1.3 Materials and Methods of the overall work

In this section, we analyze the material and methods used in the thesis and provide the context and rationale of the most important experimental choices. References to each chapter point the reader to further information, when available.

Detection of the contaminant bacterial species at each step of the filtration

In our initial evaluation of viral enrichment from freshwater and sediments, we used viral-like particles (VLP) enrichment protocols aimed to selectively enrich for viruses. These purification techniques rely on various combinations of filtrations, concentrations, and selective gradient-purification techniques to produce a sequencing library composed mainly of viral sequences (i.e. a virome). Different techniques enrich for different groups of viruses and experimental procedures are tailored both to the viral target of interest and to the specific sample type. In our study, we used a combination of sequential filtration through filters of decreasing size, to prevent obstructing the smallest filters with the biggest sediments. Small particles were then concentrated with iron chloride flocculation and sequenced.

Since our goal was to quantify and characterize not only the end-product (i.e. the purified virome), but also the microbes able to pass each filtration step, we sequenced each filter as well. Sequences were analyzed with QIIME, a software to identify bacteria from 16S rRNA sequencing data. Hence, we obtained the taxonomic label of the phyla and genera of bacteria detected at each step of the filtration. We also used MetaPhlAn2 and Kraken, two softwares for metagenomics-driven taxonomic profiling, to quantify the fraction of unassignable reads (i.e. the “dark matter”) in water and sediments. More details on this aspect can be found in **Chapter 2**.

Assessment of the level of contamination in viromes

While 16S rRNA sequencing can provide information on the contaminants (i.e. the non-viral sequences) in samples, it is not directly applicable to all studies. Moreover, publicly available datasets are constantly used in larger meta-analyses, but samples with extremely different enrichments cannot be directly compared. Before the work presented in this thesis,

there was no tool to measure the true viral enrichment from raw-reads, although some studies evaluated the presence of universal markers for bacteria (e.g. the 16S rRNA gene) through qPCR or sequence-mapping. We then developed ViromeQC, a software that can be used on shotgun metagenomes to estimate the enrichment basing on the percentage of sequencing reads that align to universal bacterial markers. We developed the software in Python and carefully validated the thresholds to allow for precise identification of contaminants. I conducted part of the validation on the universal bacterial markers during a research visit in Prof. Frederic Bushman's Lab, at the Perelman School of Medicine (University of Pennsylvania, PA U.S.A.). ViromeQC was used to conduct the largest meta-analysis on virome enrichment since 2013 (see **Chapter 3**), and it is the starting point for our enrichment-aware de-novo genome discovery approach, described in **Chapter 4**.

Retrieval of new viral sequences from viromes

Finally, we wanted to assess the potential contribution of highly enriched viromes in the retrieval of new viral genomes. We used ViromeQC (described above) to select the viromes that were enriched more than 50 times with respect to a standard metagenome. In total, 199 out of a total of 3,044 samples retrieved from NCBI-SRA matched this criterion. Through metagenomic assembly, we organized the reads of each sample into longer fragments (contigs). To filter out any further source of bacterial contamination, we mapped the sequences against a set composed both of reference bacterial genomes (from NCBI) and of bacterial genomes that were reconstructed from metagenomes (Metagenome Assembled Genomes - or MAGs). The remaining sequences were thus reliably of viral origin and were used as a bait to retrieve even more homologous sequences from public datasets. In total, we curated a set of more than 150,000 potentially viral contigs. We quantified the prevalence of each sequence by mapping them against several thousands of metagenomes. Through multiple-sequences alignment and maximum-likelihood phylogeny we were also able to study the phylogenetic relationships between similar viruses. Further information on this aspect is available in **Chapter 4**.

1.4 Structure of the thesis

This thesis is structured in seven chapters. **Chapter 1** presents a general introduction to the topics and exposes the scientific questions and aims of the thesis. Section 1.3 recapitulates the methods developed and the strategies used to address the scientific questions.

Chapter 2 contains the manuscript "*A step-by-step sequence-based analysis of virome enrichment protocols for freshwater and sediment samples"* (F. Pinto*, M. Zolfo* *et al.*, * denotes equal contribution), which is currently being finalized for submission to a scientific journal.

Chapter 3 contains the manuscript "*Detecting contamination in viromes using ViromeQC*" (M. Zolfo *et al.,* Nature Biotechnology - 2019), and illustrates the development and applications of ViromeQC, the software I developed to estimate viral enrichment from raw metagenomic reads.

Chapter 4 illustrates the results achieved by applying ViromeQC and metagenomic assembly to highly enriched viromes, in the context of *de-novo* virome discovery. The chapter contains the manuscript: "*Discovering and exploring the hidden diversity of the human gut virome"*, which is currently being finalized in preparation for submission to a scientific journal.

Chapter 5 recapitulates the overall work described in this thesis and illustrates the future steps to be carried out in the context of viral dark matter characterization and discovery.

Chapters 6-7 illustrate other contributions to published research articles in which I was involved. Although the papers included here are not always directly related to the main subject of the thesis, many of them relate to software tools or analyses that were used in the other chapters. A brief paragraph explaining the context of the research carried out in the papers is provided before each manuscript abstract.

1.5 References

Below are reported the references to **Chapter 1**. Please note that the references to the manuscripts in **Chapters 2-4 and 6** are provided after each manuscript and with independent numbering.

1. A. G. Cobián Güemes, *et al.*, Viruses as Winners in the Game of Life. *Annu Rev Virol* **3**, 197–214 (2016).
2. C. A. Suttle, Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801 (2007).
3. O. Bergh, K. Y. Børsheim, G. Bratbak, M. Heldal, High abundance of viruses found in aquatic environments. *Nature* **340**, 467–468 (1989).
4. S. Dávila-Ramos, *et al.*, A Review on Viral Metagenomics in Extreme Environments. *Front. Microbiol.* **10**, 2403 (2019).
5. K. E. Williamson, J. J. Fuhrmann, K. E. Wommack, M. Radosevich, Viruses in Soil Ecosystems: An Unknown Quantity Within an Unexplored Territory. *Annu Rev Virol* **4**, 201–219 (2017).
6. I. Reche, G. D'Orta, N. Mladenov, D. M. Winget, C. A. Suttle, Deposition rates of viruses and bacteria above the atmospheric boundary layer. *ISME J.* **12**, 1154–1162 (2018).
7. X. Wang, *et al.*, Cryptic prophages help bacteria cope with adverse environments. *Nat. Commun.* **1**, 147 (2010).
8. A. P. Fabijan, *et al.*, Safety of bacteriophage therapy in severe Staphylococcus aureus infection. *Nature Microbiology* **5**, 465–472 (2020).
9. D. P. Pires, A. R. Costa, G. Pinto, L. Meneses, J. Azeredo, Current challenges and future opportunities of phage therapy. *FEMS Microbiol. Rev.* (2020) https:/doi.org/10.1093/femsre/fuaa017.
10. S. T. Abedon, *Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses* (Cambridge University Press, 2008).
11. A. Reyes, N. P. Semenkovich, K. Whiteson, F. Rohwer, J. I. Gordon, Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* **10**, 607–617 (2012).
12. A. N. Shkoporov, C. Hill, Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell Host Microbe* **25**, 195–209 (2019).
13. C. B. Silveira, F. L. Rohwer, Piggyback-the-Winner in host-associated microbial communities. *npj Biofilms and Microbiomes* **2** (2016).
14. L. Hoyles, *et al.*, Characterization of virus-like particles associated with the human faecal and caecal microbiota. *Res. Microbiol.* **165**, 803–812 (2014).
15. M.-S. Kim, E.-J. Park, S. W. Roh, J.-W. Bae, Diversity and abundance of single-stranded DNA viruses in human feces. *Appl. Environ. Microbiol.* **77**, 8062–8070 (2011).
16. Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

17. J. Li, *et al.*, An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).

18. C. R. Armour, S. Nayfach, K. S. Pollard, T. J. Sharpton, A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome. *mSystems* **4** (2019).

19. A. Tett, *et al.*, Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis. *NPJ Biofilms Microbiomes* **3**, 14 (2017).

20. H. Sokol, *et al.*, Fungal microbiota dysbiosis in IBD. *Gut* **66**, 1039–1048 (2017).

21. F. Beghini, *et al.*, Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. *ISME J.* **11**, 2848–2863 (2017).

22. M. R. Olm, *et al.*, Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome* **7**, 26 (2019).

23. M. Breitbart, *et al.*, Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**, 6220–6223 (2003).

24. A. D. Kostic, *et al.*, The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17**, 260–273 (2015).

25. Q. Feng, *et al.*, Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).

26. A. M. Thomas, *et al.*, Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).

27. G. Hajishengallis, *et al.*, Low-abundance biofilm species orchestrates inflammatory periodontal disease through the commensal microbiota and complement. *Cell Host Microbe* **10**, 497–506 (2011).

28. E. D. Sonnenburg, *et al.*, Diet-induced extinctions in the gut microbiota compound over generations. *Nature* **529**, 212–215 (2016).

29. T. Yatsunenko, *et al.*, Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).

30. N. R. Klatt, *et al.*, Vaginal bacteria modify HIV tenofovir microbicide efficacy in African women. *Science* **356**, 938–945 (2017).

31. N. Koppel, J. E. Bisanz, M.-E. Pandelia, P. J. Turnbaugh, E. P. Balskus, Discovery and characterization of a prevalent human gut bacterial enzyme sufficient for the inactivation of a family of plant toxins. *Elife* **7** (2018).

32. G. W. Tyson, *et al.*, Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).

33. J. C. Venter, *et al.*, Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).

34. C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, N. Segata, Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

35. C. Quince, *et al.*, DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18**, 181 (2017).

36. D. T. Truong, A. Tett, E. Pasolli, C. Huttenhower, N. Segata, Microbial strain-level population

structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).

37. E. Pasolli, *et al.*, Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019).

38. J. R. Brum, M. B. Sullivan, Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13**, 147–159 (2015).

39. J. R. Brister, D. Ako-Adjei, Y. Bao, O. Blinkova, NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–7 (2015).

40. A. M. Thomas, N. Segata, Multiple levels of the unknown in microbiome research. *BMC Biol.* **17**, 48 (2019).

41. A. Almeida, *et al.*, A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).

42. S. Nayfach, Z. J. Shi, R. Seshadri, K. S. Pollard, N. C. Kyrpides, New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).

43. Y.-Z. Zhang, M. Shi, E. C. Holmes, Using Metagenomics to Characterize an Expanding Virosphere. *Cell* **172**, 1168–1172 (2018).

44. P. Manrique, *et al.*, Healthy human gut phageome. *Proceedings of the National Academy of Sciences* **113**, 201601060 (2016).

45. M. Ly, *et al.*, Transmission of viruses via our microbiomes. *Microbiome* **4**, 64 (2016).

46. G. Liang, *et al.*, The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* (2020) https:/doi.org/10.1038/s41586-020-2192-1.

47. A. McCann, *et al.*, Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* **6**, e4694 (2018).

48. J. M. Norman, *et al.*, Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447–460 (2015).

49. G. Zhao, *et al.*, Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proceedings of the National Academy of Sciences*, 201706359 (2017).

50. A. A. Abbas, *et al.*, Redondoviridae, a Family of Small, Circular DNA Viruses of the Human Oro-Respiratory Tract Associated with Periodontitis and Critical Illness. *Cell Host Microbe* **25**, 719–729.e4 (2019).

51. B. E. Dutilh, *et al.*, A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).

52. R. V. Thurber, M. Haynes, M. Breitbart, L. Wegley, F. Rohwer, Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).

53. S. Roux, M. Krupovic, D. Debroas, P. Forterre, F. Enault, Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* **3**, 130160 (2013).

54. M. Zolfo, *et al.*, Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.* **37**, 1408–1412 (2019).

55. F. Enault, *et al.*, Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J.* **11**, 237–247 (2017).

Chapter 2

Efficiency and variability of viral-like particles filtration in environmental samples

2.1 | Contribution and Context

In this paper we explored the viral communities of two sites of the pre-alpine lake Caldonazzo, in North-Eastern Italy. Samples of water and sediments were collected and Viral-Like Particle (VLP) enrichment was performed. The main goal of this work was to understand how the extraction procedure impacts the performance of VLP enrichment. We hence profiled the microbial communities of the starting sample, as well as those of the final enriched product and of all the filtration steps. 16S rRNA gene amplicon sequencing was used to profile filter-associated communities, while shotgun metagenomics was applied to assess the overall final enrichment. The results of this research contribute to explain what is the entity of microbial contaminants in enriched viromes, and can serve as a starting point on future analyses on the improvement of VLP protocols.

In this project, I took care of the bioinformatic analysis to estimate viral enrichment. In the scope of the data analysis conducted for this project, I developed the basic experimental techniques that were then refined and used to design the ViromeQC software, better described in **Chapter 3**. I also performed the shotgun metagenomics analyses and quantified the amount of unassignable metagenomic reads for each of the samples.

2.2 | Manuscript

A step-by-step sequence-based analysis of virome enrichment protocols for freshwater and sediment samples

Federica Pinto [1,*], Moreno Zolfo [1,*], Francesco Beghini [1], Federica Armanini [1], Francesco Asnicar [1], Andrea Silverj [1], Adriano Boscaini [2], Nico Salmaso [2], Nicola Segata [1,**]

[1] Department CIBIO, University of Trento, Trento, Italy

[2] Hydrobiology Unit, Edmund Mach Foundation, San Michele all'Adige, Italy

[*] Equal contribution

[**] Corresponding authors: Federica Pinto (federica.pinto@unitn.it) and Nicola Segata (nicola.segata@unitn.it).

[In finalization for the submission to a scientific journal]

Abstract

Cultivation-free metagenomic analysis afforded unprecedented details on the diversity, structure and potential functions of microbial communities in different environments. When employed to study the viral fraction of the community that is recalcitrant to cultivation, metagenomics can shed light on the diversity of viruses and their role in natural ecosystems. However, despite the increasing interest in virome metagenomic sequencing, methodological issues still hinder the proper interpretation and comparison of results across studies. Virome enrichment experimental protocols are key multi-step processes needed for separating and concentrating the viral fraction from the whole microbial community prior to sequencing. However, there is little information on their efficiency and their potential biases. To fill this gap, we used metagenomic and amplicon sequencing to examine the microbial community composition through the serial filtration and concentration steps commonly used to produce viral-enriched metagenomes. The analyses were performed on water and sediment samples from an alpine lake. We found that, although the diversity of the retained microbial communities declined progressively during the serial filtration, the final viral fraction contained a large proportion of non-viral taxa, and that the efficacy of filtration showed strong biases both based on taxonomy and supposed morphology. Our results indicate which bacterial taxa are expected to be found after filtration at different filter pore sizes. Moreover, since viral-enriched samples contained a significant portion of microbial taxa, computational sequence analysis should account for such biases in the downstream interpretation pipeline.

Introduction

Viruses populate all kinds of ecosystems from natural environments to human-associated ecosystems (e.g. the gut). Their ecological importance derives not only from their astounding abundance - being the most abundant biological entities on Earth (1) - but also from the key role they play within microbial communities. In aquatic systems, viruses can regulate the microbial community influencing biogeochemical cycles and driving the exchange of genes between prokaryotic cells (2, 3). Water in the environment can comprise up to 10^4-10^8 viral-like particles (VLP) per milliliter, but such viral diversity is still largely uncharacterised and unexplored (1). Next-generation sequencing of the environment genetic material (metagenomics) has allowed the exploration of microbial diversity to an unprecedented level of detail (4–7). However, some key methodological limitations greatly hinder the quantification of viral diversity and the characterization of their function within the microbial community. The typically limited concentration of viral particles (<1 ng μl^{-1}) in the community, biases in nucleic acid extraction, and the lack of universally conserved genomic regions in viral genomes, such as the 16S for bacteria, are common issues faced during virome analysis, particularly for environmental samples of complex matrix such as soil or sediment (8). To date the number of viral genomes available in public databases such as RefSeq is greatly lower compared to other microbes (9, 10). While human, plant and animal viruses of clinical and industrial relevance have been widely studied and sequenced, many other viruses including most bacteriophages and many ssDNA or RNA viruses are still greatly underrepresented. Since most computational viral-discovery approaches are driven by homology-based searches against known genomes, this lack of references limits our understanding of the viral realm, and our capability to detect them.

The separation between viral particles and the solid phase can be difficult because of their strong interactions, which depend on the physico-chemical characteristics of the particulate matter (11, 12). In order to separate viral-like particles (VLP) from particulate matter and microorganisms and to increase viruses concentration (and thus viral genetic material), VLP enrichment protocols are employed. Current VLP enrichment protocols use several steps, such as dissolution, centrifugation, filtration, and purification/concentration, which can vary from study to study.

Benchmark investigations have revealed that different viral enrichment protocols could generate different biases on the final virome product, mostly related to microbial contamination and biases against specific viruses (13–17). These findings call for strong

caution on profiling and detection of VLPs in viromes, specifically when associations between pathologies and samples/microorganisms are claimed (18).

Filtration is a size-based procedure that is commonly used as a separation step for virome enrichment analysis. The filters pore size of 0.45 μm and/or 0.22 μm are normally adopted, assuming that only particles smaller than their pore size would pass through the filter and that the resulting filtrate would be therefore free of microbial cells, and enriched with viruses (19, 20). However, several investigations of aquatic ecosystems revealed bacteria able to pass through 0.22 μm filters (21). The presence of microbial genetic material has been broadly confirmed in a recent meta-analysis of viromes studies from human, animal and environmental samples, highlighting how enrichment protocols can hinder the correct analyses of viral community because most of the viromes were contaminated by bacterial, archaeal and fungal genetic material (16).

Despite many studies on the comparison and optimization of enrichment protocols are publicly available, a detailed examination of the efficacy of filtration and the effects at the microbial community level (microbial community composition) is lacking.

The aim of this study was to understand the effect of filtration on virome preparation. In particular, we tested its effectiveness in removing microbial cells from viral-enriched filtrate of particulate-associate and aquatic-based samples. We run a combination of serial filtration and concentration steps commonly used to produce a viral-enriched metagenome. In order to examine the composition of the microbial fraction progressively retained in the filters, amplicon sequencing of DNA recovered at each step was performed paired with shotgun sequencing of viromes. The experiment was performed with sediment and water samples of an Alpine lake.

Results

To test the effect of multiple filtration steps on the composition of the input microbial community and on the induced relative abundance of the viral fraction, we performed multiple consecutive filtration steps on freshwater and sediment samples and sequenced the retained material at each step. Samples of sediment and water were collected at the deepest point (X) and along the coastline (Y) of a pre-alpine lake in Northern Italy. Water was sampled from the epilimnion (WE), thermocline (WT), and hypolimnion (WI). After filtering the input material through three filters of decreasing pores size (see **Methods**), amplicon 16S rRNA gene sequencing was performed on each filtering to assess the richness and composition of the bacterial component. Final enriched samples were also analysed by means of shotgun metagenomic sequencing (see **Methods**).

The amplicon sequencing data included 33 samples (4 sediment microbiomes, 25 filters, and 4 viromes) with a total number of operational taxonomic units (OTUs, clustered at 97% similarity) of 15,047. Shotgun sequencing obtained 322,414,388 quality filtered pair-end reads, which were on average 92% of the initial reads (**Table 2.1**). Taxonomic profiling of the shotgun reads with MetaPhlAn and Kraken (22, 23) showed that 99% of the reads in viromes and 60% in sediment metagenomes were not assignable (i.e. “genetic dark matter”, see **Methods**). No major differences were identified between particulate-associate and aquatic-based samples.

Viral enrichment scores are sample-type dependent

We evaluated the amount of bacterial contamination in the enriched samples obtained at the end of the filtration steps. Specifically, we compared the relative abundance of DNA fragments from ribosomal genes (16S/18S rRNA and 23S/28S rRNA genes) and universal bacterial markers (31 markers) in the initial unfiltered samples with the final viral enriched filtrates (viromes) using the ViromeQC tool (16). For each sample, ViromeQC reports an enrichment score of 1X when the sample is as enriched as a normal (i.e. non-VLP-enriched) metagenome. Hence, values lower than 1X indicate a higher abundance of non-viral contaminants (and thus, a lower viral enrichment) with respect to the reference compendium of environmental metagenomes used by ViromeQC. Values lower than 1X may be explained by the presence of small bacterial cells with multiple copies of 16S and 23S rRNA genes (i.e. more copies than what observed in the ViromeQC training set). Conversely, values greater than 1X indicate progressively higher viral enrichments.

Water and sediment viral enriched filtrates (viromes) contained microbial genetic materials, as shown by the number of OTUs detected by the amplicon sequencing (**Table 2.1**). We found a modest viral enrichment score for the sediment samples with less than 50% bacterial depletion compared to ViromeQC compendium of water unenriched metagenomes, and an enrichment compared to metagenomes from the sample specimen smaller than one order of magnitude (6.5X for site Y and 3.75X for site X) (**Table 2.1**).

Filtration performed better in the water samples, with an enrichment score ranging between 28.7X and 71.1X compared to the ViromeQC reference unenriched water samples. Nonetheless, a total of 417 reads (0.001% of the total) still mapped against 16S/18S rRNA or 23S/28S rRNA genes even for the most enriched samples (WE_CaY_V, enrichment score = 71x). DNA extraction of water unenriched metagenomes did not yield enough DNA for library preparation and sequencing, and thus only water viromes were analyzed.

				16S	Shotgun sequencing				
	Sample_ID	Site	Filtration step	Reads/ sample	Starting reads	Post QC, Human DNA & PhiX Removal	% of retained reads	MetaPhlAn 3 unknown %	ViromeQC enrichment score
Sediment	SED_CaX1-S	X	Sediment	45,649	47,447,505	44,285,662	93	88	0.4
	SED_CaX1-V	X	Virome	143,498	28,607,320	25,460,033	89	100	1.5
	SED_CaY1-S	Y	Sediment	78,635	41,680,115	38,866,734	93	76	0.2
	SED_CaY1-V	Y	Virome	208,931	35,839,636	33,694,683	93	100	1.3
Water	WE_CaX-V	X_Epilimnion	Virome	103,041	7,703,530	7,288,583	95	100	28.7
	WE_CaY-V	Y_Epilimnion	Virome	77,751	46,649,912	43,758,137	94	100	71.1
	WI_CaX-V	X_Hypolimnion	Virome	85,225	50,124,568	45,908,161	92	100	52.8

Table 2.1. 16S and shotgun sequencing reads statistics. *SED: sediment, WE: water epilimnion (upper layer of the water column), WI: water hypolimnion (lower level of the water column). Site X is the deepest point of the lake, while site Y is located by the coastline. Post QC reads and percentage of retained reads refer to reads that were retained after the quality-control step (see* ***Methods****). The ViromeQC score indicates the relative enrichment with respect to an unenriched environmental virome. Specifically, a score of 1X equals to no viral enrichment (i.e. same concentration of contaminants of an unenriched reference metagenome); values lower than 1X indicate samples less enriched for viruses than the medians observed in the reference environmental metagenomes used to train ViromeQC. Extended statistics are reported in* ***Supplementary Table 2.1****.*

Effect of filtration on microbial diversity and on the detection of rare taxa

We next sought to examine how the different filtration steps impacted bacterial richness and diversity. While the number of total OTUs at the initial filtration step was higher in sediment compared to water (respectively 3780 ± 1638 and 1058 ± 963, **Fig. 2.1A**), a similar number of OTUs were present in the final enriched samples (1062 ± 243 and 759 ± 79). This implies that the filtration performed differently between water and sediment matrix (28% and 72% decreases in OTUs, **Fig. 2.1A**). In sediment, Shannon diversity decreased along with the filtration steps, indicating a progressive elimination of the less abundant bacteria (Linear model, $p<0.01$. See **Methods**). Abundant OTUs were still present at the last step of filtration, hiding the detection of low abundant OTUs (detection level 0.2%) (**Fig. 2.1D**). Conversely, in water samples, filtration removed bacterial OTUs more homogeneously, with diversity and evenness indices remaining almost stable along the filtration process (Linear model, $p>0.05$, **Fig. 2.1B-C**).

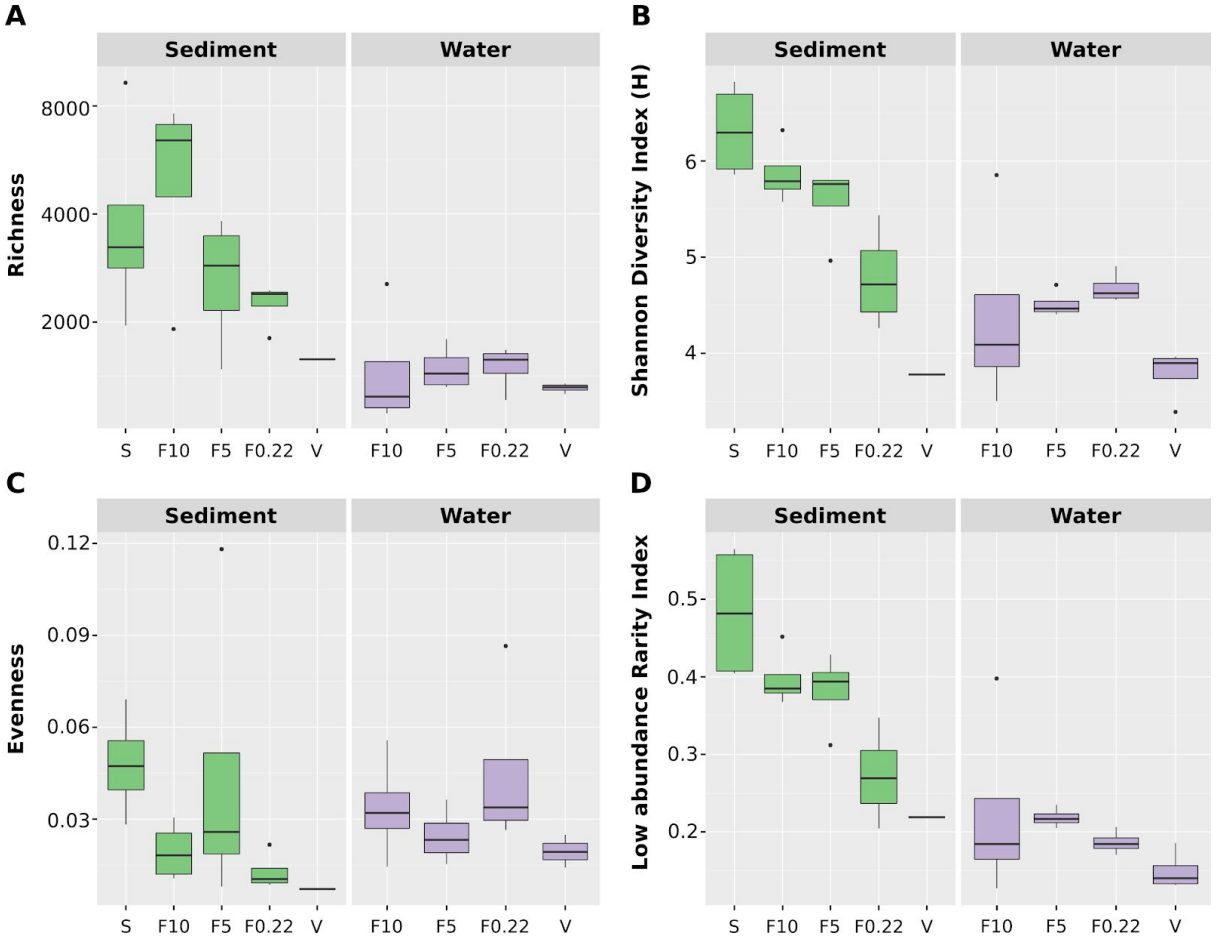


Figure 2.1 (previous page). Alpha Diversity indexes of metagenomes, filters, and viromes of water and sediments. *Boxplots represent the alpha diversity calculated on OTUs from sediments (green) and water (purple).* ***A)*** *Richness (i.e. the number of distinct OTUs in each sample),* ***B)*** *Shannon diversity index (accounting for richness and evenness),* ***C)*** *Evenness (i.e. the similarity of frequencies of OTUs in a pair of samples) and* ***D)*** *Low abundance rarity index (i.e. relative proportion of OTUs below 0.2%, regardless of their prevalence). Boxes encompass the quartiles of the distribution, while the median is indicated as a horizontal line in each box. Whiskers extend to show 1.5 Interquartile range. X-axis: filtration categories (S = raw sediment, F10 = filter 10 µm, F5 = filter 5 µm, F0.22 = filter 0.22 µm, V = virome).*

Accordingly, sediment samples displayed a significant decrease of rare species (defined by the rarity low abundance index that measures the relative proportion of species with detection level below 0.2%, regardless of their prevalence) (Linear model, $p<0.01$. **Fig. 2.2D**), whereas in water samples the relative proportion of rare species remained stable along with filtration (**Fig. 2D**).

Filtration effects on microbial composition and virome contamination

After assessing bacterial contamination in the enriched samples, we examined bacterial compositional changes induced by filtration steps using 16S rRNA gene amplicon sequencing. Microbial communities differed significantly between water and sediment samples (ADONIS test, $p<0.01$). Multi-Dimensional Scaling (MDS) of each experimental replicate based on Bray-Curtis dissimilarity showed that samples' clustering was coherent with filter pore sizes (0.22 µm, 5 µm, 10 µm), which were well sorted along the first MDS axis both for sediment and water (**Fig. 2.2**)

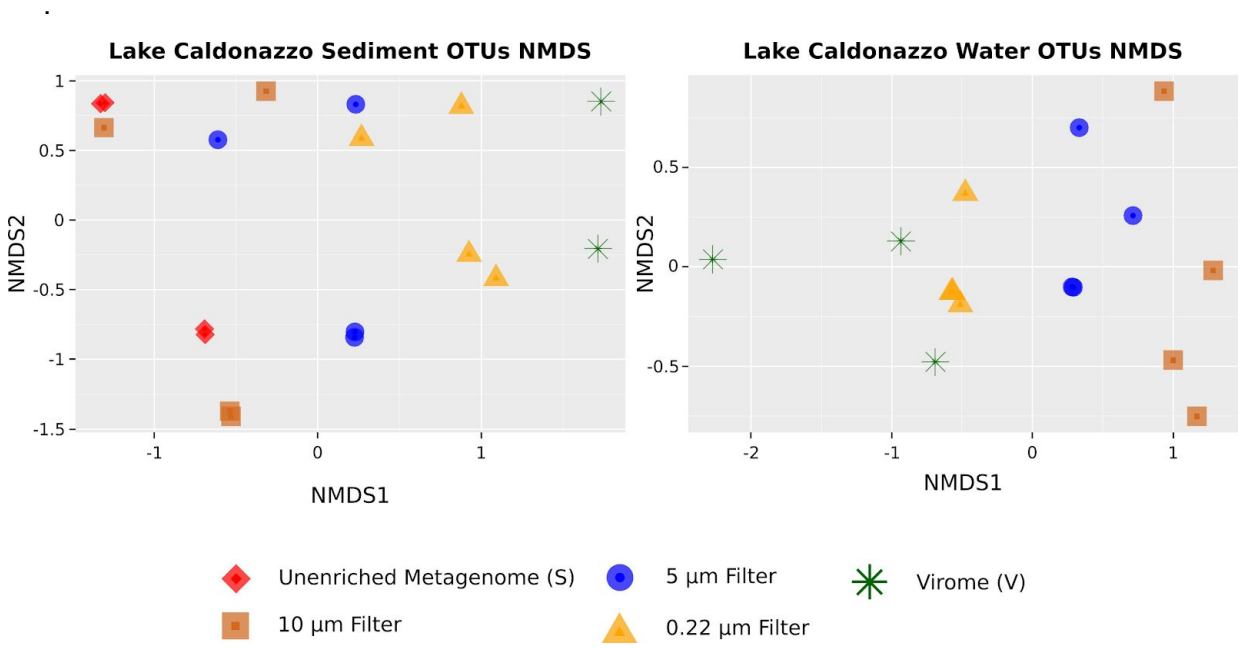


Figure 2.2. MDS Analysis on OTUs from metagenomes, filters, and viromes of water and sediments. *Non-metric multidimensional scaling (NMDS) of distances of metagenomes extracted from different pore size filters and the viromes of sediment and water samples. OTUs relative abundances were used to calculate the Bray Curtis distance between each couple of samples. Filtration categories: Unenriched metagenomes (red diamonds), 10 µm filters (orange squares), 5µm filters (blue circles), 0.22 µm filters (yellow triangles), and enriched viromes (green stars).*

We analyzed in greater detail the number of OTUs shared among filters of decreasing pore sizes. Overall, less than one-tenth of the OTUs present in the original sediment sample was retrieved after the 0.22 µm filtering step (minimum 2% maximum 9%, **Fig. 2.3B**) with higher retention rates for the water samples (minimum 11% maximum 28%, **Fig. 2.3D**). Conversely, the most enriched samples included between 142 (sediment) and 507 (water) unique OTUs that were not detected in the starting (i.e unenriched) samples, and were hence specific of the enriched viromes. While few of these virome-specific OTUs could still be the result of contamination in such low-biomass samples not detected by our computational contamination detection (see **Methods**), the majority of these OTUs likely represent taxa that were below the limit of detection in the unenriched samples.

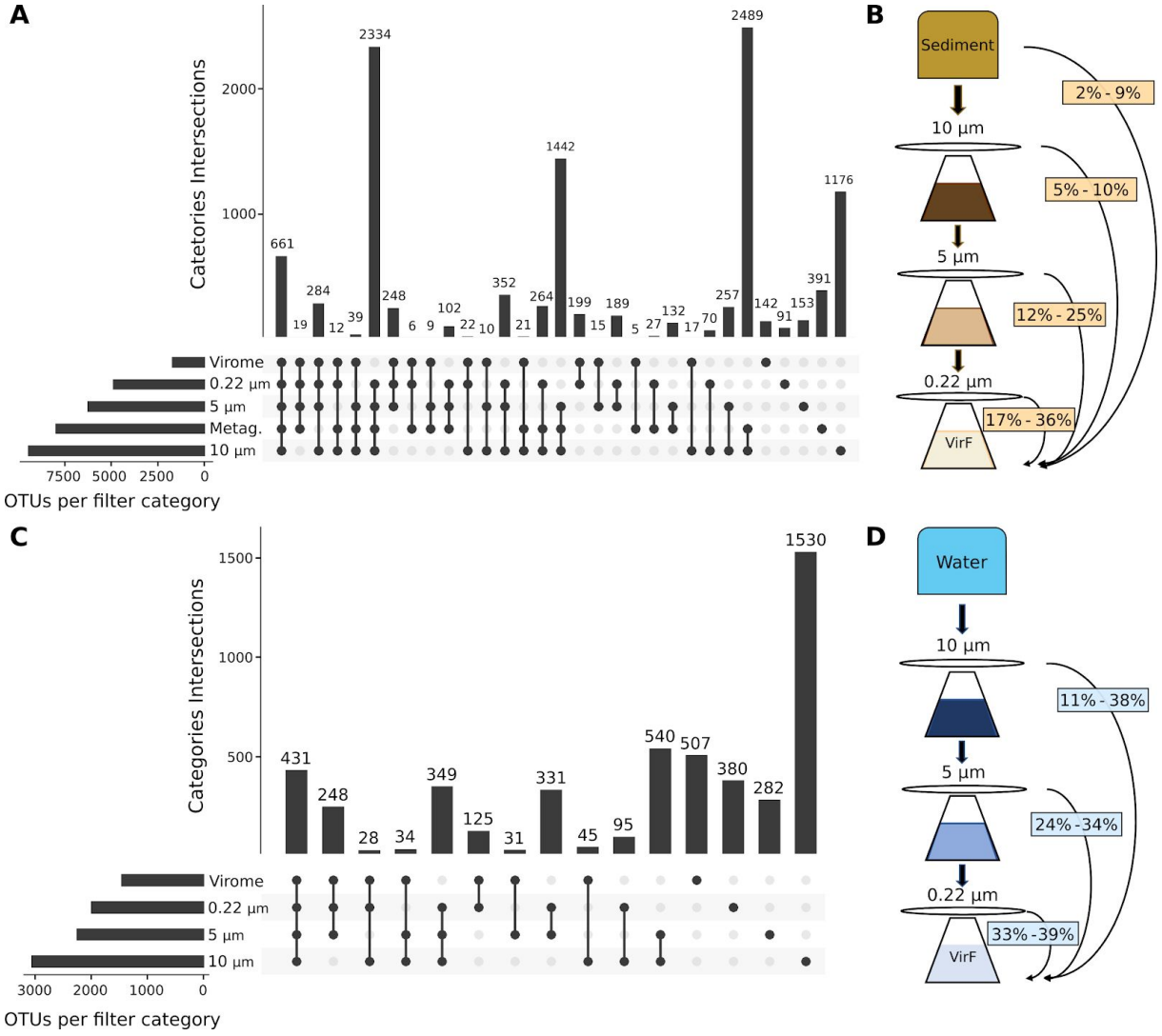


Figure 2.3 Shared OTUs at different steps of sequential filtration. A-B) *Sediment samples.* ***A)*** *Shared OTUs among and between filters and viromes.* ***B)*** *Pairwise percentage of shared OTUs, between viromes and each filter. Minimum and maximum values are reported.* ***C-D)*** *Water samples.* ***C)*** *Shared OTUs among and between filters and viromes.* ***D)*** *Pairwise percentage of shared OTUs, between viromes and each filter. The upset plots show the cardinality of each intersection between filter sizes, indicating how many OTUs are detected in each subset of filters and metagenomes.*

An illustration of the compositional changes occurring at phylum and family level along the filtration step is given in **Fig. 2.4**. Overall, the taxonomic composition of water and sediments differed at Phylum and more evidently at Family level (**Fig. 2.4**) in the final enriched filtrates. At the initial stage of filtration, Proteobacteria, Cyanobacteria, Bacteroidetes and Planctomycetes are the most abundant phyla. Viromes were still characterised mostly by these phyla, with members of Firmicutes at higher relative abundances (from 5 to 10%). The virome-specific OTUs clearly differed between sediment and water. Sediments were still

dominated by Proteobacteria (90%) and Firmicutes (from 5 to 8%), whereas in water a more diverse community was retrieved, including Actinobacteria (from 4 to 8%) and members of the candidate phyla radiation (CPR) (from 0.1 to 3%) such as TM6, Microgenomates, Parcubacteria, Peregrinibacteria, Saccharibacteria, and Omnitrophica.

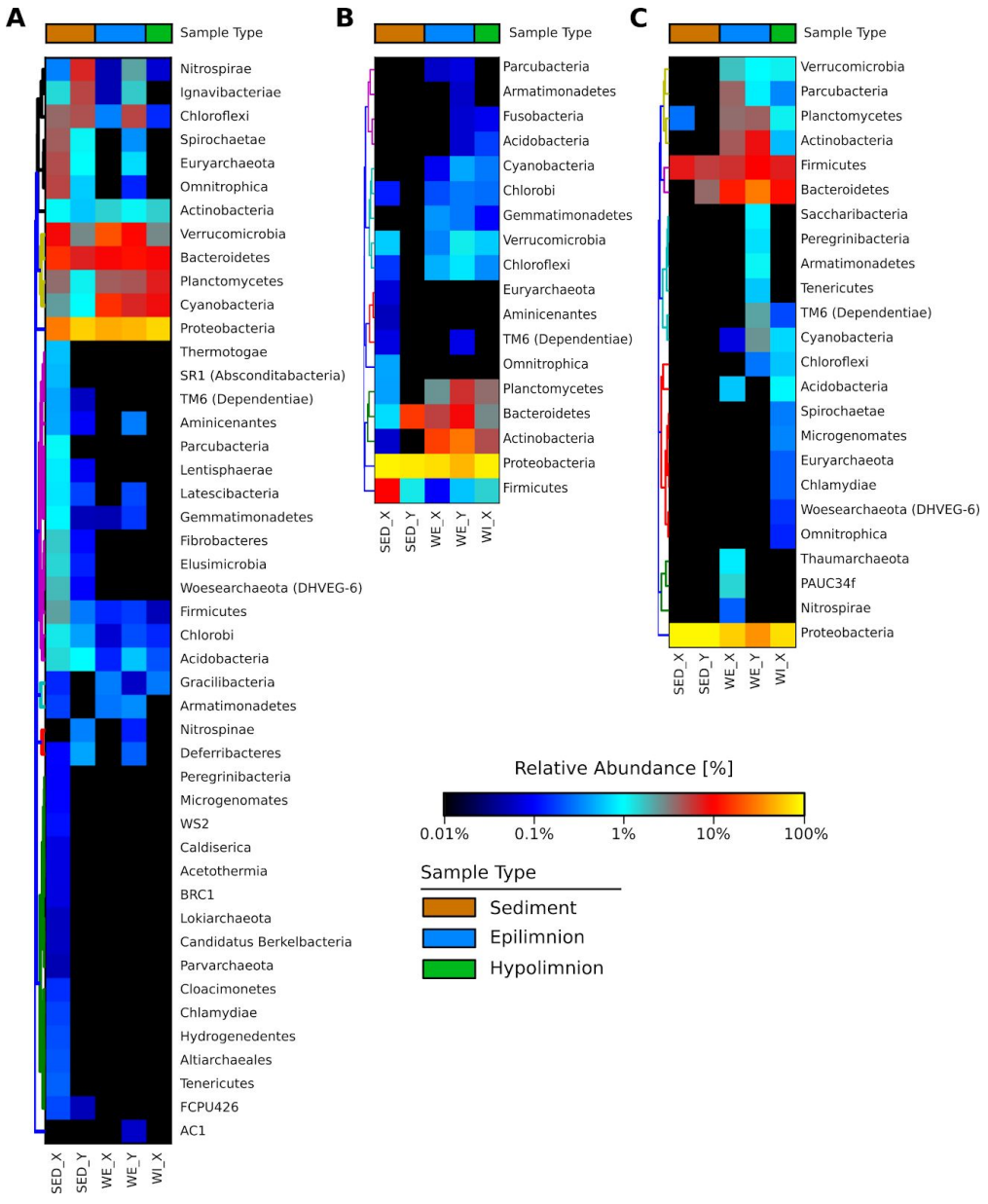


Figure 2.4. Heatmap of OTUs relative abundance at Phylum level. *The relative abundance of each OTU is shown for* ***A)*** *OTUs at the starting point of filtration. Sediment samples refer to the sediment metagenomes, while water samples refer to samples taken from the 10µm filters.* ***B)*** *As in A), but OTUs from enriched viromes are shown.* ***C)*** *OTUs present only in enriched viromes. OTUs are hierarchically clustered using Bray-Curtis distance. The sample type is indicated in the upper metadata colored bar. Family level OTUs relative abundances are shown in* ***Supplementary Fig. 2.1.***

Discussion

Assigning sequences to viruses poses considerable computational challenges in the study of viromes. The elimination of the genetic material of non-viral origin from the samples is thus fundamental to simplify analyses and avoid biased interpretations. Our study is among the first to specifically test the efficacy of the filtration steps used to eliminate microbial cells from environmental samples during viral enrichment protocols. Particularly, we examined changes in microbial composition occurring throughout the filtration process until the final viral enriched filtrate.

We found that the efficacy of filtration differed between water and sediment samples. The filtration was more effective in water-based samples where it appeared to gradually trim the larger and most abundant species. As such, filtration in water produced viromes that were orders of magnitude more enriched than sediment, as indicated by the ViromeQC enrichment score (16). This result further indicates that the presence of particles and minerals in the sample matrix, such as those occurring in sediments (24, 25) and, more specifically, sediment characteristics such as porosity and organic matter content, can profoundly influence the efficacy of viral enrichment (26, 27). Consequently, different approaches have been used to account for the retention properties of similar matrices (e.g. soil/sediment, faeces, respiratory samples), such as sample homogenization, as employed here (19, 28, 29). However, our results indicate that additional investigations are needed to develop better laboratory protocols.

Water samples produced much higher enrichment scores, and yet we still detected substantial microbial genetic material in the final enriched samples. This calls for caution when downstream sequence analyses are performed, even after apparent successful enrichment, as it cannot be assumed that the sample contains only viral particles.

Examination of changes in microbial diversity during the filtration provided additional details on how the process differed between sample types. The progressive decline in the Shannon diversity and the low abundance rarity index along the filtration steps in sediment sample implies that the efficacy of filtration primarily reflected the relative abundance of taxa, whereby common taxa were more likely to pass through. Conversely, filtration of taxa in water samples appeared to be less dependent on their relative abundance, with both common and rare taxa equally likely to be retained, as indicated by the more stable diversity values. This suggests that filtration in sediments might be relatively more stochastic compared to water samples, where taxa were presumably retained according to their cell size, rather than to their abundance. As previously mentioned, the presence of particle

aggregates in the sediment matrix might explain these results and the lower efficacy of the enrichment.

Although the enrichment differed between sample matrices, filtration steps produced consistent compositional changes across replicate filters in both water and sediment samples, with the first MDS axis mirroring the distribution of pore sizes (**Fig. 2.2**). This indicates that, regardless of the overall efficacy, filtration procedures can produce consistent and reproducible outcomes within a given sample matrix.

The key assumption of the enrichment protocols is that only particles smaller than the minimum filter pore size (0.45 µm and/or 0.22 µm) are able to pass through (20). However, in line with other recent studies (15, 16, 30, 31), results from our experiments indicate that viral enriched samples still contained microbial genetic material. This could have practical implications in many research fields. A recent meta-analysis of viromes studies from human, animal and environmental samples, highlighted how commonly used enrichment protocols can hinder the correct analyses of viral communities because of contamination by bacterial, archeal or fungal genetic material (16). Besides contamination occurring during the experimental procedures, the detection of microbial genetic material in the enriched filtrates could be associated to i) changes in cell size and shape due to external factors; and ii) presence of very small bacteria, such as those belonging to the newly discovered Candidate Phyla Radiation (CPR) (32–34).

For instance, the presence of rod-shaped bacteria such as Oxalobacteraceae, anaerobic purple sulfur Chromatiaceae (35, 36), and Bryobacter (Acidobacteria) (37) in the final enriched samples could be due to the physical effect of vacuum during filtration. These are relatively small bacteria of 0.3 - 0.5 µm cell width, whose shape and size could be modified by mechanical forces in ways that could increase their passability (filterability) through the smallest filters (38).

Together with the presence of Planococcaceae, Pseudomonadaceae and Sphingobacteriaceae that are commonly found in aquatic and terrestrial habitats (39, 40) examination of virome-specific bacterial OTUs revealed also the presence of rod-shaped cells with a width of 0.3 µm - 0.5 µm such as Oxalobacteraceae, anaerobic purple sulfur Chromatiaceae (35, 36) in sediment and Bryobacter (Acidobacteria) (37) in water, which could pass through the smallest pore size filter.

In water virome unique OTUs, Phyla belonging to the candidate phyla radiation (CPR) were also retrieved. TM6, Microgenomates, Parcubacteria, Peregrini bacteria, Saccharibacteria and Omnitrophica, characterised by small sizes (0.009±0.002 µm) and limited biosynthetic capabilities due to their short genome length (34, 41), were detected in

the enriched final filtrates but were under the limit of detection level at the starting point of filtration. Thus, efficient filtration might enrich not only viral particles but also low abundant microbial species. Hence, this kind of enriched samples may be useful to study low abundant ultra-small bacteria such as those belonging to CPR, whose sequences could be retrieved at a greater depth and coverage. Although the enrichment protocol is specifically designed to enrich viruses, additional research may provide further insight in the characterization of such bacterial species.

16S rRNA amplicon sequencing of filters shed light on the archaeal and bacterial contamination at each filtration step, while shotgun metagenomics allowed to study the VLP enrichment efficacy in its entirety by applying ViromeQC. While the main scope of this paper was to highlight the nature of the contaminants along the filtration step, future studies could apply shotgun metagenomics and ViromeQC to each individual filter, in order to better explore how the enrichment efficacy changes at each step of the enrichment.

Considering our results on the microbial community during protocols to enrich viruses, we conclude that non-viral particles can be highly abundant in environmental viromes. Hence, we hence call for caution when analysing and profiling VLPs, as it cannot be assumed that metagenomes sequenced from VLP samples contain only viral sequences.

Materials and Methods

Study site and sampling

Caldonazzo Lake is a meso-eutrophic lake located at an elevation of 449 m in Trentino, Italy. Sampling occurred in March 2017 during the lake stratification period in two sites, the deepest point (site X, 49 m depth) and close to the coastline (site Y, 7 m depth). Specifically, the first 2 cm of four sediment cores were collected in duplicates and pooled together to collect in total 200 g of sediment. Water samples (2 L) were collected from the epilimnion (WE), thermocline (WT), and hypolimnion (WI) of the stratified lake. All bottles and devices were acid rinsed and sterilised before use.

Microbial and viral DNA extraction

Sediments (100 g) were treated with sodium pyrophosphate (final concentration 5 mM), sonicated, and centrifuged in order to separate and collect the sediment pore water. Samples (2 L of water and 100 ml of sediment pore water) were then serially filtered through 10 µm, 5 µm and 0.22 µm filter pore size (**Fig. 2.5**) (Whatman filter, Merck KGaA, Darmstadt, Germany) using sterilised filtration units (Nalgene, Thermo Fisher Scientific, USA) mounted on sterile glass bottles. Filters were stored at -20 C. Virus-like particles (VLPs) in the final filtrate, defined here as the viral fraction, were then concentrated using the iron chloride precipitation protocol (42) and Amicon Ultra filters (100KDa), reaching a final volume of 1-2 ml. Samples were stored at -80 C.

DNA was extracted from both filters and viral fractions using different protocols. The 10, 5 and 0.22 µm pore-size filters were processed using the DNeasy PowerWater Kit (QIAGEN, Hilden, Germany) following the manufacture instructions. Viral fractions, instead, were first treated with DNase I (15U ml^{-1}) for 1 h at 37C; then DNA was extracted using QIAamp DNA Mini Kit (QIAGEN, Hilden, Germany). DNA was extracted using DNeasy Power Soil Kit (QIAGEN, Hilden, Germany) directly from sediment (250 mg) following the manufacturer instructions. The DNA was quantified using the Qubit™ dsDNA HS Assay Kit (Life Technologies, Carlsbad, CA).

The extraction was also performed with the unfiltered lake water, but the retrieved genetic material was under the detection limit. Therefore, water metagenomes could not be sequenced.

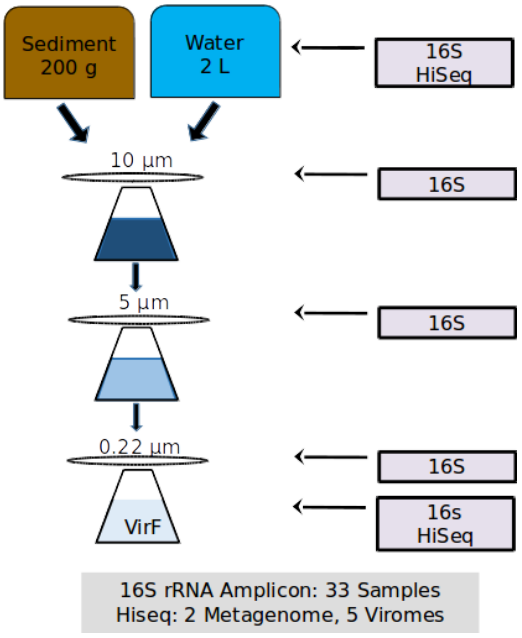


Figure 2.5. Overview of the extraction procedure. *Overall, 33 16S rRNA gene amplicon libraries and 7 shotgun libraries were extracted from freshwater and sediments. The type of library is indicated in the gray boxes on the right. VirF stands for "Viral Fraction". Pore sizes are indicated above each filter.*

16S rRNA Amplicon and shotgun sequencing

To characterise the microbial community along filtration, DNA from filters, from sediments, and from viral filtrates were subjected to PCR amplification of the 16S rRNA variable regions V3-V4. Amplicons were pooled and sequenced on an Illumina MiSeq platform.

Shotgun sequencing was applied to the DNA extracted directly from sediment (metagenomes) and to viral fractions (viromes). Libraries, prepared using Nextera XT DNA Library Prep Kit (Illumina) according to the manufacturer's instructions, were quality checked by the Perkin Elmer LabChip GX (Perkin Elmer) and sequenced on a HiSeq 2500 platform (Illumina).

Bioinformatics and statistical analysis.

16S rRNA gene analysis was performed with QIIME (43). Operational taxonomic units (OTUs) were picked with the open-reference approach and the SILVA database release 128 at 97% clustering (44). In R, data was processed using phyloseq (45), vegan (46) and Upset packages. Archaea, Chloroplast and Mitochondria were removed from the dataset. For the

non-metric multidimensional scaling (NMDS) community analysis, Bray-Curtis dissimilarity was used after removing rare OTUs (<5 occurrences). Adonis analysis was performed using the R Vegan package to determine the differences between habitats, filters and sampling location. Differences in bacterial diversity indexes over the filtration process were tested using a linear regression model, setting as base level the first step of filtration (raw sediment and filter 10 µm for water). Bacterial richness was log-transformed. To determine and represent shared OTUs among categories (filters and viromes) and unique OTUs, upsetR was applied.

Raw metagenomic reads were preprocessed with Trim Galore (47) to remove low quality (i.e. Phred score < 20) and short (i.e. length < 75 bp) reads (parameters: --stringency 5 --length 75 --quality 20 --max_n 2 --trim-n). Metagenomes were analyzed with MetaPhlAn (22) v. 3.0 with the --unknown_estimation option and Kraken2 (23), version 2.0.8 and Braken (48) to quantify the percentage of unknown sequences. Percentages are reported in **Supplementary Table 2.1**.

Viral enrichment was calculated with ViromeQC, a computational tool that estimates the efficacy of VLP enrichment by quantifying the abundance of unwanted microbial contaminants. ViromeQC version 1.0 (16) was run on the metagenomic reads with the --environmental option.

Data Availability

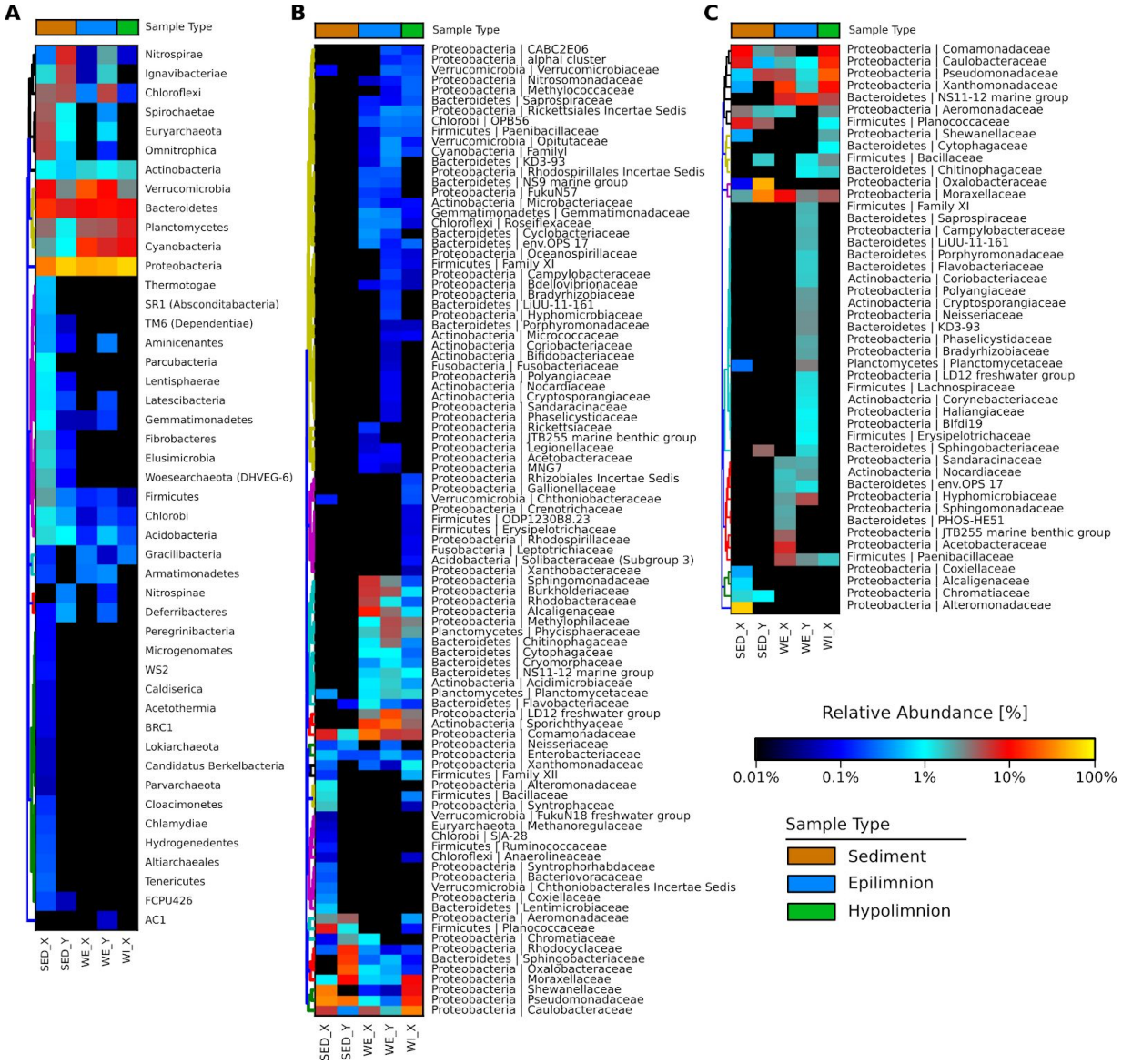
The raw sequencing reads of the 16S rRNA amplicon sequencing and shotgun metagenomics were submitted to the NCBI-SRA archive and are available under the BioProject PRJNA658338.

Acknowledgements

This project received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 704603 to FP and the European Research Council (ERC-STG project MetaPG No. 716575) to NS. We would like to thank Stefano Larsen and Nicolai Karcher for their feedback and technical support.

Supplementary Material

Supplementary Table 2.1. Amplicon and shotgun sequencing reads statistics. SED: sediment, WE: water epilimnion, WT: water thermocline, WI: water hypolimnion. Site X refers to the deepest point of the lake, site Y refers to the coastline. Table available at: http://segatalab.cibio.unitn.it/data/metaviralp/Mvir_Suppl_material_T1.docx



Supplementary Figure 2.1. Heatmap of OTUs relative abundance at Family level. *The relative abundance of each OTU is shown for* ***A)*** *OTUs at the starting point of filtration. Sediment samples refer to the sediment metagenomes, while water samples refer to samples taken from the 10µm filters.* ***B)*** *As in A), but OTUs from enriched viromes are shown.* ***C)*** *OTUs present only in enriched viromes. OTUs are hierarchically clustered using Bray-Curtis distance. The sample type is indicated in the upper metadata colored bar.*

References

1. C. A. Suttle, Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801 (2007).

2. C. Canchaya, G. Fournous, S. Chibani-Chennoufi, M. L. Dillmann, H. Brüssow, Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.* **6**, 417–424 (2003).

3. X. Wang, *et al.*, Cryptic prophages help bacteria cope with adverse environments. *Nat. Commun.* **1**, 147 (2010).

4. G. W. Tyson, *et al.*, Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).

5. J. C. Venter, *et al.*, Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).

6. D. M. S, S. De Mandal, A. K. Panda, Microbial Ecology in the Era of Next Generation Sequencing. *Journal of Next Generation Sequencing & Applications* **01** (2015).

7. A. S. Hahn, K. M. Konwar, S. Louca, N. W. Hanson, S. J. Hallam, The information science of microbial ecology. *Curr. Opin. Microbiol.* **31**, 209–216 (2016).

8. D. Pan, Y. Morono, F. Inagaki, K. Takai, An Improved Method for Extracting Viruses From Sediment: Detection of Far More Viruses in the Subseafloor Than Previously Reported. *Front. Microbiol.* **10**, 878 (2019).

9. C. Schmidt, The virome hunters. *Nat. Biotechnol.* **36**, 916–919 (2018).

10. J. R. Brister, D. Ako-Adjei, Y. Bao, O. Blinkova, NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–7 (2015).

11. M. G. Weinbauer, J. R. Dolan, K. Šimek, A population of giant tailed virus-like particles associated with heterotrophic flagellates in a lake-type reservoir. *Aquat. Microb. Ecol.* **76**, 111–116 (2015).

12. F. Hassard, *et al.*, Abundance and Distribution of Enteric Bacteria and Viruses in Coastal and Estuarine Sediments—a Review. *Frontiers in Microbiology* **7** (2016).

13. R. V. Thurber, M. Haynes, M. Breitbart, L. Wegley, F. Rohwer, Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).

14. A. N. Shkoporov, *et al.*, Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* **6**, 68 (2018).

15. S. Roux, M. Krupovic, D. Debroas, P. Forterre, F. Enault, Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* **3**, 130160 (2013).

16. M. Zolfo, *et al.*, Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.* **37**, 1408–1412 (2019).

17. A. C. Gregory, O. Zablocki, A. Howell, B. Bolduc, The human gut virome database. *BioRxiv* (2019).

18. M. Asplund, *et al.*, Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. *Clin. Microbiol. Infect.* **25**, 1277–1285 (2019).

19. A. Reyes, N. P. Semenkovich, K. Whiteson, F. Rohwer, J. I. Gordon, Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* **10**, 607–617

(2012).

20. C. d'Humières, *et al.*, A simple, reproducible and cost-effective procedure to analyse gut phageome: from phage isolation to bioinformatic approach. *Sci. Rep.* **9**, 11331 (2019).

21. L.-A. J. Ghuneim, D. L. Jones, P. N. Golyshin, O. V. Golyshina, Nano-Sized and Filterable Bacteria and Archaea: Biodiversity and Function. *Front. Microbiol.* **9**, 1971 (2018).

22. D. T. Truong, *et al.*, MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).

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

24. D. A. Bazylinski, R. B. Frankel, Magnetosome formation in prokaryotes. *Nat. Rev. Microbiol.* **2**, 217–230 (2004).

25. M. P. Taylor, K. A. Hudson-Edwards, The dispersal and storage of sediment-associated metals in an arid river system: the Leichhardt River, Mount Isa, Queensland, Australia. *Environ. Pollut.* **152**, 193–204 (2008).

26. J. L. Castro-Mejía, *et al.*, Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. *Microbiome* **3**, 64 (2015).

27. R. R. Helton, L. Liu, K. E. Wommack, Assessment of factors influencing direct enumeration of viruses within estuarine sediments. *Appl. Environ. Microbiol.* **72**, 4767–4774 (2006).

28. M. Breitbart, *et al.*, Viral diversity and dynamics in an infant gut. *Res. Microbiol.* **159**, 367–373 (2008).

29. C. Kohl, *et al.*, Protocol for Metagenomic Virus Detection in Clinical Specimens1. *Emerging Infectious Diseases* **21** (2015).

30. Y. Wang, F. Hammes, M. Düggelin, T. Egli, Influence of size, shape, and flexibility on bacterial passage through micropore membrane filters. *Environ. Sci. Technol.* **42**, 6749–6754 (2008).

31. M. Bekliz, J. Brandani, M. Bourquin, T. J. Battin, H. Peter, Benchmarking protocols for the metagenomic analysis of stream biofilm viromes. *PeerJ* **7**, e8187 (2019).

32. A. V. Fedotova, Y. M. Serkebaeva, V. V. Sorokin, S. N. Dedysh, Filterable microbial forms in the Rybinsk water reservoir. *Microbiology* **82**, 728–734 (2013).

33. R. S. Kantor, *et al.*, Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *MBio* **4**, e00708–13 (2013).

34. B. Luef, *et al.*, Diverse uncultivated ultra-small bacterial cells in groundwater. *Nat. Commun.* **6**, 6372 (2015).

35. E. Rosenberg, E. F. DeLong, S. Lory, S. Stackebrandt, F. Thompson, "The Family Chromatiaceae" in *The Prokaryotes. Gammaproteobacteria*, E. Rosenberg, E. F. DeLong, S. Lory, S. Stackebrandt, F. Thompson, Eds. (Springer, 2014), pp. 151–178.

36. J. I. Baldani, *et al.*, "The Family Oxalobacteraceae" in *The Prokaryotes: Alphaproteobacteria and Betaproteobacteria*, E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, F. Thompson, Eds. (Springer Berlin Heidelberg, 2014), pp. 919–974.

37. I. S. Kulichevskaya, N. E. Suzina, W. Liesack, S. N. Dedysh, Bryobacter aggregatus gen. nov., sp. nov., a peat-inhabiting, aerobic chemo-organotroph from subdivision 3 of the Acidobacteria. *Int. J. Syst. Evol. Microbiol.* **60**, 301–306 (2010).

38. S. Cesar, K. C. Huang, Thinking big: the tunability of bacterial cell size. *FEMS Microbiol. Rev.* **41**,

672–678 (2017).

39. S. Shivaji, T. N. R. Srinivas, G. S. N. Reddy, “The family planococcaceae” in (Berlin: Springer-Verlag, 2014).

40. A. Lambiase, The family sphingobacteriaceae. *The Prokaryotes* **4**, 907–9014 (2014).

41. R. E. Danczak, *et al.*, Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. *Microbiome* **5** (2017).

42. S. G. John, *et al.*, A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.* **3**, 195–202 (2011).

43. J. G. Caporaso, *et al.*, QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).

44. C. Quast, *et al.*, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).

45. P. J. McMurdie, S. Holmes, phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217 (2013).

46. F. G. B. Jari Oksanen, *et al.*, Vegan: community ecology package. *R package version* **2** (2018).

47. F. Krueger, Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* (2015).

48. J. Lu, F. P. Breitwieser, P. Thielen, S. L. Salzberg, Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).

Chapter 3

Detecting contamination in viromes using ViromeQC

3.1 | Contribution and Context

In this paper, we focused on the development and application of ViromeQC, a computational approach that can be used on shotgun viromes to evaluate the enrichment efficiency. The method is based on the quantification of the number of reads that align against a set of universal markers for bacteria, archaea, and fungi (i.e. the large and small subunits of the rRNA gene plus 31 single copy bacterial markers). The tool is public and freely available at http://segatalab.cibio.unitn.it/tools/viromeqc/ and represents a novel resource to evaluate viromes directly from the raw sequencing reads. We used ViromeQC to perform the largest meta-analysis on the available viromes to date and concluded that caution is needed when referring to viromes as if they contained only (or predominantly) viral sequences.

In this research, I retrieved and preprocessed all the samples considered in the meta-analysis, performed the validation experiments, and developed and tested the software ViromeQC. I conducted the validation of single-copy markers during my abroad visit at the University of Pennsylvania, in Prof. Frederic Bushman's Laboratory.

3.2 | Manuscript

Detecting contamination in viromes using ViromeQC

Moreno Zolfo [1], Federica Pinto [1], Francesco Asnicar [1], Paolo Manghi [1], Adrian Tett [1], Frederic D. Bushman [2] and Nicola Segata [1,*]

[1] Department CIBIO, University of Trento, Trento, Italy

[2] Department of Microbiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

[*] Corresponding author: Nicola Segata (nicola.segata@unitn.it)

Published: Nature Biotechnology 37, 1408–1412, 2019
https://doi.org/10.1038/s41587-019-0334-5

***Note:** This manuscript is the pre-print version of the manuscript, prior to editorial edits.*

Abstract

Eukaryotic viruses and bacteriophages are key players in the ecology of microbial communities, but the characterization of viromes with cultivation-based and metagenomic tools is still a difficult task. Viral-like particle (VLP) purification techniques allow concentration of viruses from complex microbiome samples prior to sequencing. However, different forms of starting material can constrain the accessible degree of purity, so that in some cases unexpected infiltration of non-viral entities causes a high abundance of cellular nucleic acids and can lead to misleading conclusions. An analysis of 2,050 VLP samples highlights how viromes can be affected by prokaryotic contamination, with the majority of samples showing only modest viral enrichment, and also having variable enrichment levels within the same study. We provide a simple software tool to assess prokaryotic contamination in untargeted virome samples and recommend strict contamination quality control to help guide downstream virome sequence analysis.

Introduction and Motivation

Viruses are key members of human and environmental microbial communities that influence both overall microbiome ecology and the genetic makeup of microbial strains (1–4). Bacteriophages, viruses of bacteria and archaea, are extremely abundant and diverse, and influence microbial populations in diverse ways. Transduction, for example, is an important mechanism of lateral gene transfer (5, 6). Bacteriophage therapy is an attractive antimicrobial strategy when dealing with antibiotic resistance in clinical settings (7). Metagenomics offers the opportunity to uncover and survey phage populations, but a complication is that phage sequences are so diverse, and evolve so rapidly, that they are poorly represented in existing databases. In addition, there are no universal viral genetic markers and commonly the overall biomass of viruses with respect to other microbes in samples is low. Phage sequences are thus typically hard to recognize in metagenomic data. Nevertheless, several methods based in part on sequence characteristics of known phages have been developed (8–10).

Virus-like particle (VLP) purification can be used to enrich microbiome samples for viral nucleic acids (11, 12), thereby improving detection of viruses. Virome protocols have widely varying goals, ranging from careful untargeted analysis of highly purified phage populations, enabling deep ecological studies of phage communities, all the way to targeted identification of rare sequences of viral pathogens in complex mixtures from diagnostic samples. In cases where the goal is to use metagenomics to detect viral pathogens, a low purity sample may suffice for their identification by alignment of sequence reads to viral databases. However, in cases where previously unknown viruses are sought, or where whole viral populations are of interest, designation of sequence reads as virus-derived is reliant on the quality of the VLP purification procedure. Methods for untargeted investigations of the viral diversity in complex microbial communities typically include filtration through small pore size filters that retain bacteria, cesium chloride gradient purification, treatment with chloroform to disrupt membranes, and exposure to nucleases to reduce free DNA and RNA concentration. When coupled with untargeted shotgun sequencing (13), VLP enrichment techniques aim to increase the proportion of sequence reads of viral origin, and have allowed numerous studies in human (4, 14–17), environmental (18, 19), and built-environment settings (20).

Standardized validated experimental VLP enrichment protocols for untargeted virome sequencing are lacking, and no single protocol is likely to be optimal for all sample types. Regardless of the specific VLP protocol employed, non-viral genetic material is commonly present after the enrichment steps (12, 21, 22) as both prokaryotic and eukaryotic cells can

infiltrate into the final VLP preparations. Given the objective of VLP purification, such unwanted nucleic acids can be considered as contaminants, and their presence is problematic, especially for the task of *de novo* discovery of phages in untargeted virome sequencing from complex microbial communities. With a non-contaminated VLP virome, one can assemble reads into longer sequences and consider them viral genomes without the need for computational prediction approaches that are unavoidably affected to some extent by low-confidence calls and false negatives for particularly divergent viral sequences (8, 9). The level of concentration of viruses in the starting VLP sample is thus an important factor in recovering new viral diversity, but methods for evaluating the purity of VLP samples have not been explored systematically.

Some studies assessed microbial contamination of VLP-preparations before sequencing by targeting with PCR the 16S rRNA gene as a proxy for prokaryotic presence and abundance (23–29). Others used the same gene or another single marker to test contamination in the downstream sequenced samples by sequence mapping (30–35). However, these studies did not extend to use of additional ribosomal or protein-coding genes and they did not provide a validated pipeline to quantify viral enrichment in viromes and unenriched samples. They also typically focused on each study separately. While efforts have been performed towards VLP-protocol optimization (12, 35–37), the largest meta-analysis of post-sequencing non-viral quantification performed so far considered only 67 viromes (21). The use of VLP enrichment for virome sequencing is rapidly increasing, so there is a need for methods to evaluate non-viral contamination in the full set of >2,000 samples available, which we address here.

Results and Discussion

We have devised and validated an approach to evaluate bacterial contamination in VLP metagenomic samples for untargeted virome analysis of complex microbial communities. To assess the enrichment rates of publicly available viromes, we applied the method on a collection of 2,050 human, animal and environmental samples from 35 metagenomics virome sequencing studies that performed one of the available VLP enrichment techniques (**Supplementary Table 3.1**). As controls, we studied 2,189 metagenomes that were not enriched for viruses from the curatedMetagenomicData (38) and NCBI-SRA (39) repositories, as well as 108 publicly accessible synthetic metagenomes (40, 41) and one mock community (42). After uniform preprocessing to remove low-quality reads (see **Methods**), we first computed the percentage of raw reads in each sample aligning to the small subunit ribosomal RNA gene (SSU rRNA), which has never been found to be encoded by a virus. This provides a proxy for non-viral microbial sequence abundance (21). We hence estimated the abundance of the bacterial and archaeal 16S and micro-Eukaryotic 18S ribosomal genes across all the viromes and metagenomes. Unenriched metagenomes provided a baseline estimation of the environment-specific rRNA abundance from which we calculated the relative enrichment of viromes with respect to the metagenomes. Environmental and human/animal unenriched metagenomes had a median rRNA gene abundance of 0.08% (n=320, interquartile-range=0.07%) and 0.25% (n=1,551, interquartile-range=0.1%) (**Fig. 3.1**).

Prokaryotic and micro-Eukaryotic contamination of viromes estimated by the quantification of the SSU-rRNA showed a wide range of enrichment efficiencies, with a large fraction of samples (n=567, 28.7%) showing no enrichment at all, and more than 50% (n=990) with less than 3x enrichment. A substantially smaller fraction of samples (n=339, 17.15%) showed high enrichment (>100x). The differences in enrichment rates were poorly linked to the VLP-purification procedure employed, although the heterogeneity of protocols makes it difficult to provide statistical support to this observation. According to the taxonomic annotations of the rRNA sequences retrieved in viromes, the highest source of contamination had a bacterial origin for most of the samples, while 199 samples showed instead higher abundances of eukaryotic associated SSU rRNAs (**Supplementary Table 3.3**). The rRNA gene abundance variability was higher in viromes than in metagenomes (Mann–Whitney U test p-value = 7.5e-8, **Supplementary Fig. 3.1**), highlighting that not only many viromes are poorly enriched, but also that the level of prokaryotic infiltration is rather unpredictable, and not preferentially associated with any VLP enrichment technique.

The intra-dataset enrichment efficiencies were extremely variable, spanning more than two orders of magnitude in almost half of the studies (48.7%), indicating that even the same virome-enrichment protocol applied to samples from the same study can still display vastly different levels of contamination. For example, the 91 stool samples from the dataset of Ly *et al*.(27) had a 16S rRNA abundance standard deviation equal to 4.6 times the average (**Fig. 3.1**, ref. 38). This suggests that quality-benchmarking viromes after sequencing is a crucial step to evaluate the presence of contaminants and that the intra-dataset variability should be carefully considered in downstream analyses of untargeted viromes.

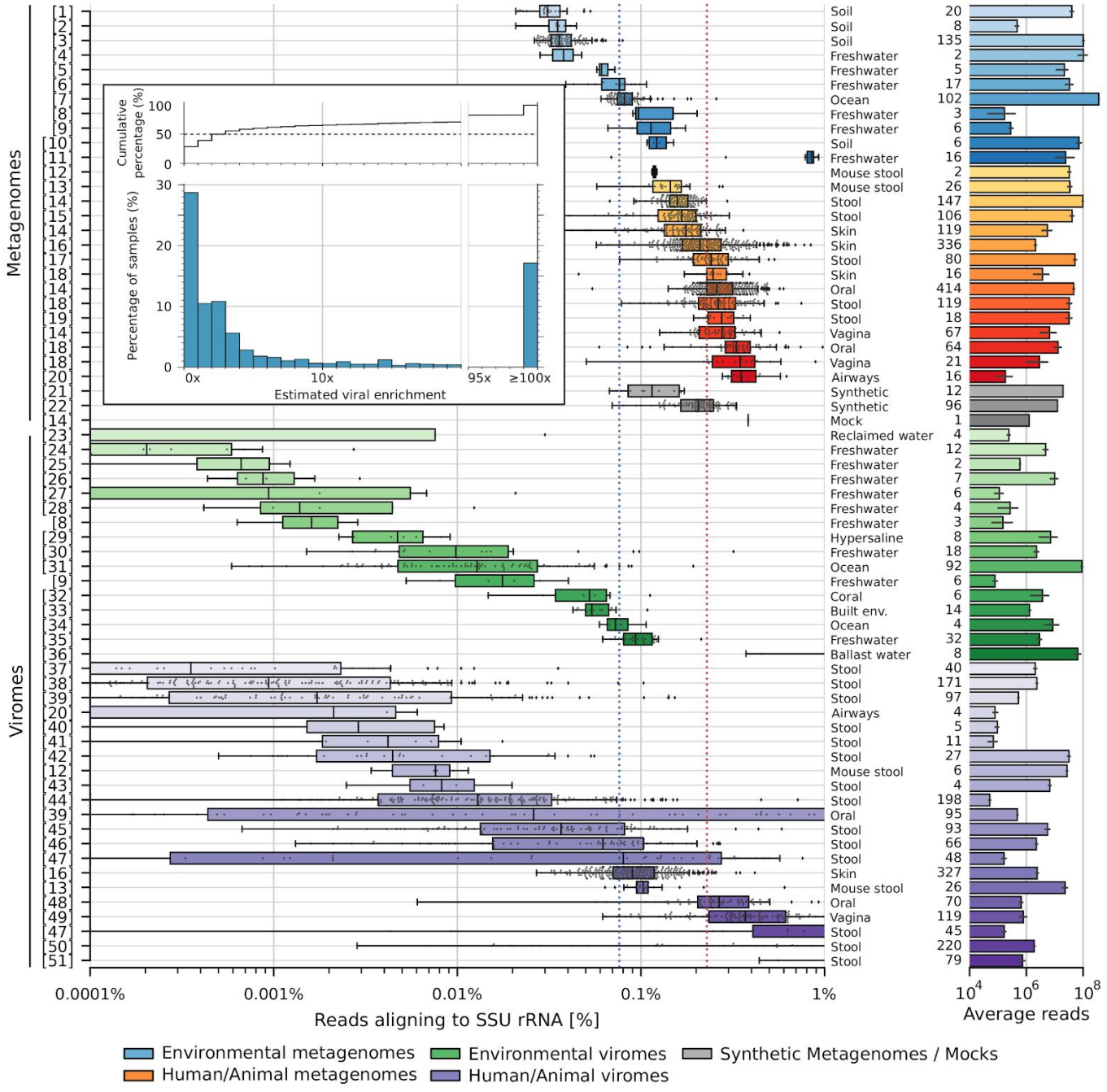


***Figure 3.1*. Survey of viral enrichment rates on 1,977 samples from 35 studies estimated as percentage of reads aligning to the small subunit rRNA gene.** *The vertical dotted lines indicate the median of median SSU rRNA abundances in human/animal (red dotted line) and environmental (blue dotted line) unenriched metagenomes, as a reference. The two medians are used to calculate the enrichment rate of each virome with respect to the reference metagenomes. The frequency of enrichment levels of all the 1,977 viromes that passed quality-control is represented in the inset histogram. The histogram on the right side shows the number of reads (bar height) and the number of samples (to the left of the bar) in each dataset. Datasets are grouped by type (environmental or*

*Human/animal). Datasets within the same group are ordered by median abundance. References to each dataset are provided in (**Supplementary Tables 3.1** and **3.2**). Error bars in the right barplot show the 95% confidence intervals. Boxes show the quartiles of each dataset, the central line in each box indicates the median, while whiskers extend to show data points within 1.5 IQR range. Data-points, including outliers, are overlaid to the boxes.*

Four VLP datasets (17, 26, 43, 44) were highly enriched in rRNA genes (i.e. median abundance > 10%), with peaks of 90% reads aligning to either the 16S/18S or 23S/28S rRNA gene subunits (datasets 36, 47, 50 and 51, see **Supplementary Table 3.1**). Conversely, the median rRNA gene abundance observed in unenriched real and synthetic metagenomes never exceeded 1% (**Supplementary Table 3.2**). The experimental design of these four studies pointed at the likely cause for this problem, as they involved DNA and RNA co-extraction, with DNA and retro-transcribed cDNA sequenced together. We hypothesize that higher rRNA abundance was observed due to incompletely-depleted structural rRNA in the samples (45). In a further 25 RNA viromes, we also found higher rRNA abundances than would be expected (4.18% median abundance when considering both rRNA subunits, maximum of 67.5%, **Supplementary Table 3.4**). As it was not possible to evaluate the VLP enrichment efficiency by estimating rRNA abundances for samples with atypically high levels of rRNA, we excluded from the downstream analysis datasets with more than 10% median abundance of rRNA genes, suggesting that viromes with such high levels of rRNA genes should be treated as unlikely to be useful in downstream genome assembly and analysis. In total, 307 samples were removed, all of which were from studies that sequenced DNA and RNA together. Thus, protocols of this type cannot be evaluated with the presented approach, though we note that for other applications, such as sequence-based detection of known pathogens, they may allow convenient analysis.

To corroborate viral enrichment estimates further, we also calculated the abundance of the large ribosomal subunit rRNA gene (LSU-rRNA), encoding for prokaryotic 23S and eukaryotic 28S rRNAs (**Fig. 3.2A**), and of 31 single-copy universal markers from bacterial and archaeal ribosomal proteins (46) (**Supplementary Fig. 3.2**). Some ribosomal proteins can be occasionally found in viral genomes (47), and this raises the possibility of designating viral genomes wrongly as contaminants. However, extensive mapping of the universal ribosomal markers against viral repositories (48, 49) provided evidence that these are rare events that are unlikely to affect the results (**Supplementary Note 3.1, Supplementary Fig. 3.3, Supplementary Table 3.5**), especially when considering as markers all the 31

single-copy universal markers. A few samples (11.8%) still harbored high levels of rRNA genes (i.e. >5% abundance) after the above filtration steps, mostly aligning to the large rRNA gene subunit (**Supplementary Fig. 3.4B, Supplementary Fig. 3.5**). However, the abundance quantification of rRNA genes (SSU and LSU) and genes coding for single-copy proteins were in agreement for most of the viromes. In 75.3% of the viromes, rRNA genes and single-copy marker abundances were either both below (67.1%) or above (8%) the reference unenriched-metagenomes medians (**Supplementary Fig. 3.4**). The abundances of the individual markers were highly correlated (**Fig. 3.2B**), as were the abundances of SSU rRNA and single-copy markers (Spearman's coefficient = 0.72 when considering the abundance of all 31 markers together). A weaker correlation was observed between LSU rRNA and single-copy markers (**Fig. 3.2B,** Spearman's coefficient = 0.47). While rRNA and single-copy marker abundances were generally in agreement, these differences suggest a multi-marker approach is necessary to estimate the level of viral enrichment accurately. For example, one of the studies considered (34) had high amounts of LSU rRNA genes, but was highly enriched if quantifying only SSU rRNA.

We hence defined a comprehensive enrichment score encompassing rRNA large and small subunit abundances and single-copy markers. The score expresses how much a sample is enriched with respect to the medians observed in unenriched metagenomes, and was computed by taking the minimum across the three single enrichment scores for both viromes and metagenomes (see **Methods**). Less than 50% of the analyzed viromes had an overall enrichment greater than 3x, less than 15% reached 30x enrichment, and only 10% of the viromes were more than 50x enriched. In fact, the majority of the analyzed viromes failed to meet even what we consider a low level of enrichment (i.e. 2-3x, **Fig. 3.2C**). Most of the studies had mixed levels of enrichment, confirming what observed previously on enrichments based on the SSU-rRNA gene only (**Fig. 3.2D**), with some datasets spanning between 1x and 100x. This further underscores how samples that underwent the same VLP-technique might have widely different levels of non-viral contamination.

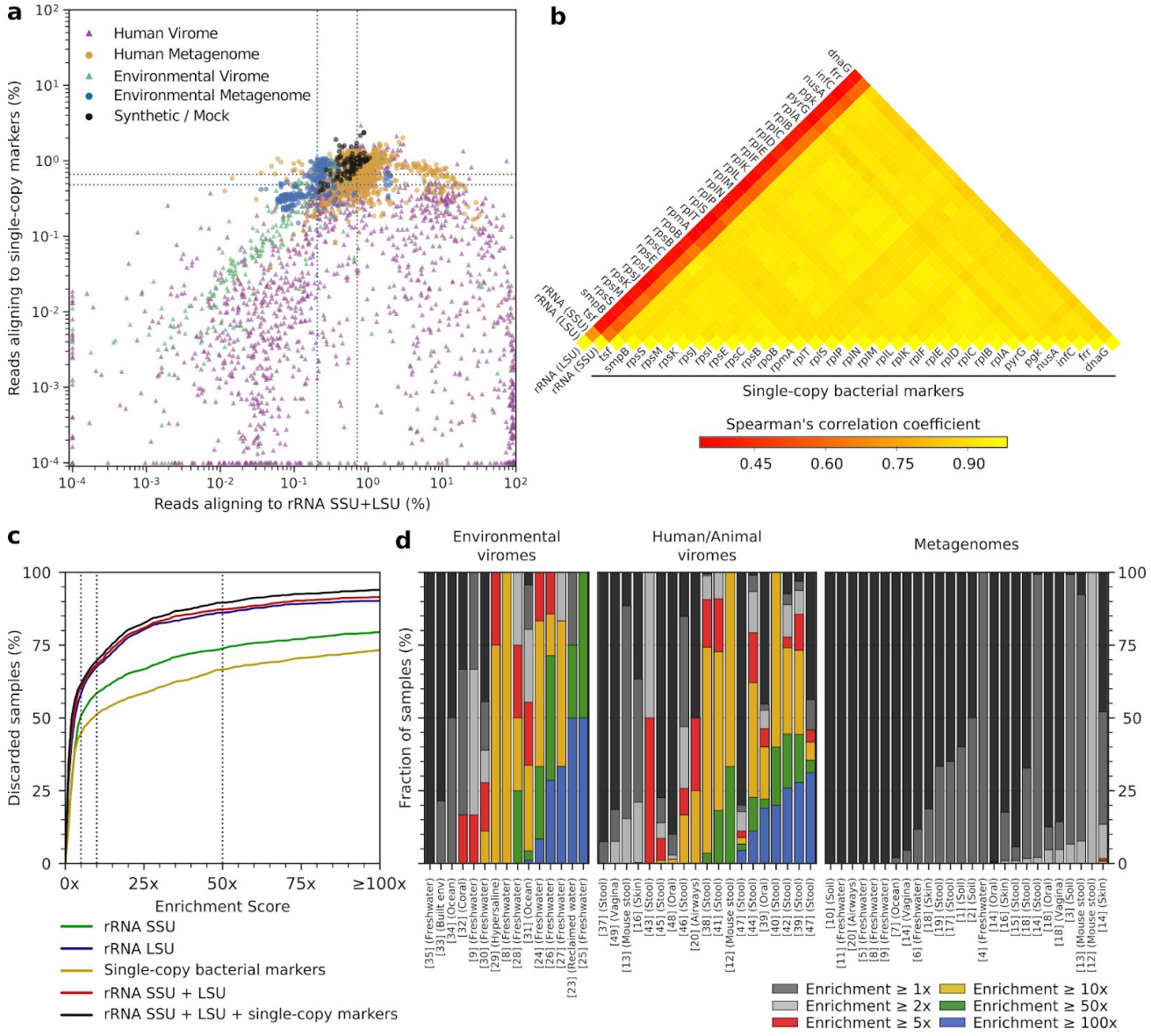


***Figure 3.2**. **Combined quantification of ribosomal genes and genes coding for universal proteins identifies the cross-study set of 101 samples with >100x VLP enrichment.** **(a)** The retrieved viromes were mapped against rRNA small and large subunits reference sequences (x-axis), and against 31 single-copy bacterial markers (y-axis). The scatter plot shows the percentage of aligned reads on 1,751 human and animal viromes (red) and 226 environmental viromes (blue). The dotted lines indicate the median abundances in the corresponding metagenomes. **(b)** Spearman's correlation coefficients between the 31 single-copy markers and the small and large subunits of the rRNA gene. **(c)** Fraction of the discarded viromes at different enrichment thresholds (dashed lines) and using different components to calculate the enrichment. The proposed threshold (rRNA SSU + LSU + single-copy markers) is drawn in black. **(d)** Enrichment scores of samples within each dataset grouped by dataset type together with metagenomes used as controls. The enrichment score is based on both SSU and LSU rRNAs and single-copy markers. References to each dataset are provided in **Supplementary Tables 3.1** and **3.2**.*

To further highlight the importance of quality control in untargeted virome metagenomics, we investigated the extent to which the viral enrichment score is connected with the ability to computationally identify viral genomes from virome samples subjected to metagenomic assembly. We assembled 1,445 untargeted virome samples and classified each of the resulting 2.09 x 10^7 contigs as viral or not-viral using VirSorter (8) (see **Methods**). The proportion of viral and potentially-viral contigs increased from an average of 7.9% to an average of 31% for samples with viral enrichment-scores of 1-2x and 5-9x, respectively. However, the proportion of predicted viral contigs did not significantly increase at higher enrichment values (**Supplementary Fig. 3.6**). Indeed, in the majority of samples enriched by a factor of 100x or more, for which there are, at best, just traces of ribosomal genes from prokaryotes and eukaryotes, less than 25% of the assembled nucleotides could be classified as "potentially viral" (i.e. VirSorter *category 1, 2 or 3*), and less than 4% was "surely viral" (i.e. *category 1)*. At such high enrichment rates, assembled contigs could all be considered of viral origin, and this calls for a substantial false negative rate, likely due to viral genomes not displaying enough similarity with known viruses and known characteristics of viral genomes, and the limitation of contig-based viral detection tools when analyzing contigs with relatively short length (8). Conversely, 55 of the 475 lowly enriched samples (i.e. <3x) had more than 20% of the assembled nucleotides labelled as potentially viral which is inconsistent with the high abundance of prokaryotic organisms that have much longer genomes and could suggest the presence of false positives in viral contig prediction. Overall, this analysis highlighted that caution is needed when interpreting the results of viral mining software, and that incorporating the concept of virome-enrichment in the existing untargeted virome investigations provides a useful tool for quality control and for better informed downstream analyses.

Conclusions

The level of infiltration of non-viral nucleic acids in untargeted virome metagenomics has been assessed in the past by several studies by PCR or by mapping-based approaches directed to either rRNA genes and other prokaryotic markers, but a comprehensive comparative analysis of thousands of samples was missing. The present analysis is thus trying to raise awareness of the issues of potential prokaryotic and eukaryotic contamination, as post-sequencing evaluation of non-viral contaminants in viromes prior to contig-based virus classification is still rarely performed in current investigations. Because the results of analyses conducted on VLP samples are interpreted differently from non-viral-enriched metagenomes, our read-based estimation of non-viral contamination can be used to guide the selection of tools and thresholds for downstream viral contig detection. This because, if metagenomic assembly is performed on poorly enriched samples, it raises the likelihood of increasing the number of contigs that are wrongfully considered viral by computational approaches based on sequence homology to viral genomes. Accordingly, we showed how a substantial part of the contigs assembled from highly purified viromes (i.e. enrichment >100x) cannot be labelled as viral by viral-detection approaches. Furthermore, there was no strong correlation between the fraction of assembled contigs that were labelled as viral and the enrichment score of the samples from which they were assembled. We hence emphasize the need for strict quality control of viromes prior to genome analysis. This is particularly important when datasets are retrieved and compared from public sources, and when metagenomic assembly is performed to characterize unknown viruses in samples (35).

The computational pipeline we introduced to analyze the enrichment efficacy of viromes can be a useful tool for better interpretation of untargeted virome sequencing experiments. Differently from previous methods which focused only 16S rRNA genes to address microbial contamination, the method we propose exploits the abundances of (a) 16S/18S rRNA genes, (b) 23S/28S rRNA genes, and (c) a panel of 31 universal bacterial genes. The rRNA detection method was validated on synthetic and 16S rRNA gene datasets and the abundance of the ribosomal and universal genes in unenriched metagenomes was used for a precise estimation of enrichment rate of viromes. The software is freely available at http://segatalab.cibio.unitn.it/tools/viromeqc. Analysis of 2,050 human/animal and environmental viromes highlighted that viromes can have widely different enrichments within and between studies, ranging from less than 1x to more than 100x. VLP-purity varied greatly even in samples processed with the same experimental methodology, and thus post-sequencing quality controlling VLP-enrichment should be performed regardless of the

technique used. Samples where DNA and RNA were co-extracted and pooled into the same library were highly enriched in rRNAs supporting the need to deplete structural rRNA prior to library preparation.

The identified enrichment biases may be due to different concentrations of non-viral material in the starting sample, different physical/chemical properties that affect only a subset of the samples, experimental biases or human errors. To at least partially control for these biases, we thus recommend considering non-viral contamination when performing untargeted virome analysis, as contaminants may lead to spurious conclusions when performing *de novo* viral discovery (i.e. false positives), or when analyzing differentially abundant viral genes and functions.

Methods

To estimate the viral enrichment in the 2,050 high-quality viromes, we combined two mapping approaches against the rRNA small and large subunits genes, and 31 single-copy bacterial markers. Viromes and control metagenomes were preprocessed to remove low-quality reads and were divided into four groups: a) environmental metagenomes, b) human/animal metagenomes, c) environmental viromes, d) human/animal viromes. For each sample, we calculated three enrichment scores based on: i) small subunit rRNA; ii) large subunit rRNA; iii) single-copy bacterial markers. The pipeline is released open-source with examples and documentation at http://segatalab.cibio.unitn.it/tools/viromeqc. We present below the approach in more detail.

Metagenomic and virome datasets considered

We recovered 2,050 viromes from 35 datasets (**Supplementary Table 3.1**) and 2,189 metagenomes from curatedMetagenomicData (38) and from other public sources (**Supplementary Table 3.2**). To calibrate the rRNA detection pipeline, we analyzed 956 16S rRNA gene amplicon sequencing samples (17, 19, 50–54), 108 synthetic microbial communities (40, 55) and 29 cDNA sequencing of retro-transcribed RNA samples from 5 of the datasets that also contained DNA viromes (20, 31, 56–58).

Metagenomes and viromes preprocessing

All samples were uniformly preprocessed to remove low quality (i.e. median Phred score < 20) and short (length < 75) sequences. On average, 91.9% of the reads passed the quality control: a total of 1.2×10^{11} reads were analyzed (1.4×10^{10} from viromes, 1×10^{11} from metagenomes, 5.4×10^{7} from 16S rRNA gene sequencing, 2.1×10^{7} from RNA-Seq and 1.4×10^{9} from synthetic metagenomes). Viromes and metagenomes with less than 20,000 high-quality reads, as well as 16S rRNA gene sequencing samples with less than 1,000 reads were excluded. Samples were screened for human DNA by mapping against the human genome using Bowtie2 (59) (mode --sensitive-local, index hg_19). In total, 1,977 viromes and 1,871 metagenomes passed quality control.

Reference database selection and curation

Ribosomal RNA genes sequences were downloaded from the SILVA-Arb database (60), release 132, files SSURef_Nr99 and LSURef. We manually curated both subunits sequences collections to remove low complexity entities. To avoid spurious mappings, we removed misclassified sequences (i.e. LSU rRNA sequences that were present in the SSU database and vice-versa), or that contained both subunits of the rRNA genes. To this end, we collected the sequences of both rRNA genes subunits of 14 annotated bacterial, archaeal and fungal reference genomes (accessions: NC_000913.3, NC_000964.3, NC_002696.2, NC_000908.2, NC_006840.2, NZ_ALVU00000000.2, NC_016830.1, NC_021149.1, NC_003888.3, NC_001144.5, NC_009515.1, NC_011852.1, NZ_LN846995.1 and NC_010513.1). Given the high conservation of rRNA genes, these few well-annotated reference sequences allowed to detect conserved regions among very unrelated organisms. We then used BLAST (61) to search the SILVA rRNA databases for misplaced rRNA sequences. SILVA-Arb LSU sequences that mapped against any of the 14 SSU rRNA genes with at least 70% identity were removed. The same procedure was followed for SSU sequences using LSU target references. Additionally, all entries labelled as chloroplasts and mitochondria were removed. After this procedure, we obtained 189,684 LSU and 686,807 SSU rRNA genes sequences and discarded 9,159 LSU and 8,364 SSU rRNA genes sequences.

We also considered and retrieved the 31 AMPHORA2 bacterial marker genes (46). These 31 marker genes are universally conserved and distributed among bacteria and the majority of them are single-copy housekeeping genes mainly involved in information processing (replication, transcription, and translation) or central metabolism. These markers are thus thought to be refractory to lateral gene transfer (62).

Abundance estimation for SSU and LSU rRNA genes

rRNA genes abundance was estimated by mapping the quality-filtered reads against the curated collection of rRNA large and small ribosomal subunit sequences. Reads were mapped using Bowtie2 (59) v. 2.3.4 (--very-sensitive-local mode) with two separate indexes for the large and small rRNA subunits.

To exclude spurious matches, we filtered the resulting alignments with the software tool CMSeq(63). Filtering parameters were calculated on the expected ratio of SSU-rRNA mapping in 108 synthetic microbial communities with known composition. Specifically, we

used 96 samples that were generated with StrainMetaSim as described by Quince and colleagues (41), and 12 samples generated with GemSIM (64) in the scope of a previous publication and that were retrieved from BioProject PRJNA339720. Detailed compositions of these samples are available in **Supplementary Table 3.4**. For each community, we then calculated the expected small subunit rRNA abundance. Briefly, for each genome in the community, we extracted the longest SSU-rRNA sequence of that species from SILVA-Arb. We then mapped each genome against its corresponding rRNA gene sequence with BLASTn. The expected abundance in each synthetic metagenomes was calculated as follows:

$$rRNA\ abundance = \sum_{i=0}^{n} \frac{\frac{\sum_{j=0}^{B_i}(a_{i,j})}{L_i} \cdot Q \cdot b_i}{Q} = \sum_{i=0}^{n} \frac{\sum_{j=0}^{B_i}(a_{i,j})}{L_i} \cdot b_i$$

where n is the number of genomes in the synthetic community, B_i is the number of SSU-rRNA alignments of genome i, $a_{i,j}$ is the length the j-th alignment in the i-th genome, L_i is the length of genome i, Q is the number of reads in the synthetic metagenome and b_i is the relative abundance of genome i in the synthetic metagenome. Alignments shorter than 250 nucleotides were not considered. The median expected SSU-rRNA abundance was 0.18%, similarly to what reported for real metagenomes in previous studies (34, 35).

Synthetic metagenomes were then aligned to SSU-rRNA sequences with Bowtie2 as described above. Alignments were filtered with different combinations of parameters, according to the alignment length (no filter vs 75 long alignments) and to the divergence with respect to the closest match (i.e alignment edit distance as of Bowtie2 XM:i value, divided by alignment length). The resulting rRNA gene abundances were then compared with the expected values (**Supplementary Table 3.4**). Additionally, 917 rRNA amplicon sequencing samples showed that even at moderate strictness, we could still recover more than 95% of the 16S rRNA gene reads (**Supplementary Table 3.4**). Higher stringency increased the accuracy of detection in synthetic communities, but lowered the detectability of true rRNAs. To cope with this trade-off, we selected a minimum alignment length of 75 nucleotides and a maximum of 7.5% divergence with respect to the closest rRNA reference sequence. All the abundances were computed by dividing the number of aligning reads by the total number of non-human high-quality reads.

Abundance estimation of universally conserved proteins

Samples were mapped against single-copy universal bacterial protein marker using Diamond (65) v. 0.9.9 (command blastx, parameters --id 50 --max-hsps 35 -k 0). The reported abundances are based on the reads aligning to any of the protein markers, with each read mapping at most once (i.e. only the best hit was selected). Read counts were divided by the total number of non-human high-quality reads. Single-copy markers were mapped against viral genomes using Diamond (command blastx, parameters --id 50 --evalue 0.0001 --very-sensitive) against the RefSeq viral database (release 91), and against IMG/VR (accessed in May 2019).

Definition of the overall enrichment score

The overall enrichment score was defined as the minimum of three enrichment scores calculated on LSU-rRNA, SSU-rRNA and single-copy markers abundances. Briefly, a baseline for each was calculated by considering the median of the median abundances in unenriched-metagenomes of each group (i.e. environmental and human/animal). Enrichments scores were calculated by dividing the metagenome-baseline by the abundance of either rRNAs or single-copy markers. Overall enrichment values are reported in **Fig. 3.2C-D** and **Supplementary Table 3.1** and **Supplementary Table 3.2**

Correlations and statistical tests were computed using Scipy (66) Stats module (version 1.2.1); to calculate the correlations in **Fig. 3.2B**, we excluded samples with fewer than 5 reads aligning either against rRNA or single-copy markers. Filtering was performed by custom scripts using PySam/htslib (67) and the internally developed CMSeq tool (63). We tested whether sequencing depth correlated with enrichment score and found that there was no correlation (**Supplementary Fig. 3.7**).

Metagenomic assembly and viral vs non-viral classification of the resulting contigs

Datasets with more than ten samples were preprocessed for metagenomic assembly with Trim Galore (68) (v. 0.4.4, parameters --stringency 5 --length 75 --quality 20 --max_n 2 --trim-n). Reads aligning to the human-genome were removed from the raw reads as described above, and samples were metagenomically assembled with metaSPAdes (69) (v. 3.10.1, default parameters) for paired-end samples and with MEGAHIT (70) (v. 1.1.1) for unpaired samples. A total of 1,445 samples were successfully assembled. Only contigs

longer than 500 nucleotides were retained. To classify the resulting viral contigs, VirSorter (8) (v. 1.0.5) was run in decontamination mode with the following parameters: --db 1 --virome --diamond). Only samples with a total assembly size of more than 1 million nucleotides were considered.

Data and Code Availability

The raw reads analyzed in this study are available from the respective studies accession numbers provided in **Supplementary Table 3.1** and **Supplementary Table 3.2**. Code and documentation are available at http://segatalab.cibio.unitn.it/tools/viromeqc.

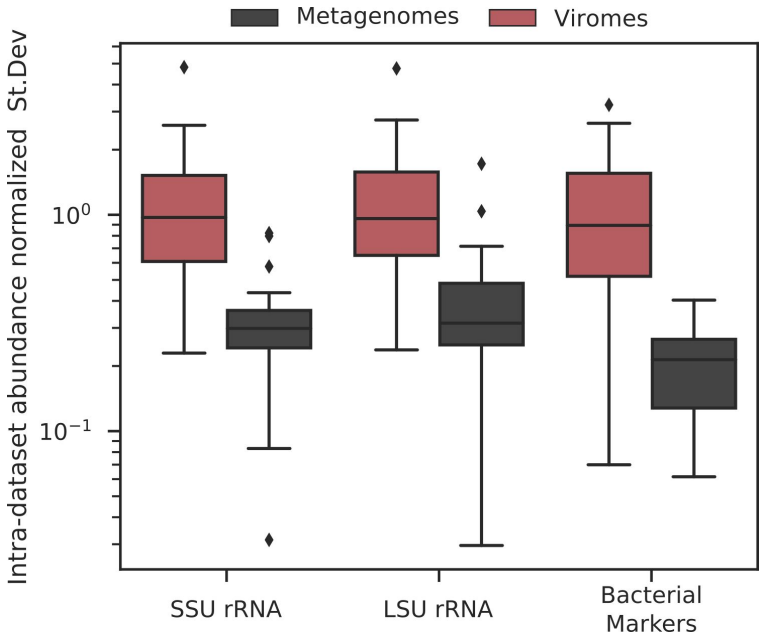
Supplementary Material

Supplementary Note 3.1 - Presence of ribosomal proteins in viral reference genomes.

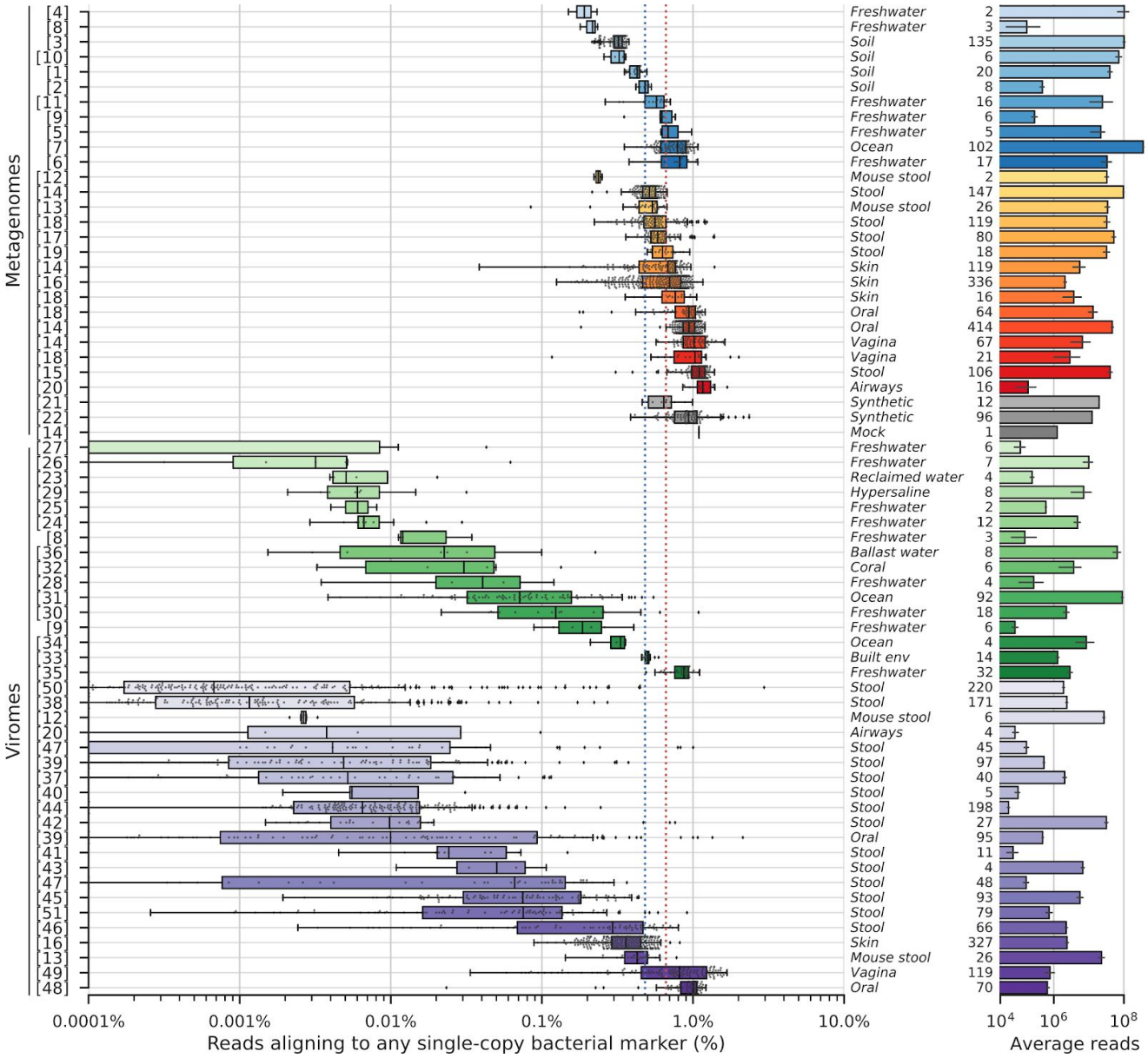
To evaluate the potential problem of the presence of prokaryotic genes encoding for ribosomal proteins in viral organisms, we mapped 31 the single-copy bacterial markers against the viral genomes in RefSeq(48) (see **Methods**). Only one marker (*dnaG*) could be detected in several viral genomes. Indeed, *dnaG* encodes for the bacterial DNA primase and was found in 45 Propionibacterium phage genomes. DNA primase is known to share homologies with bacteriophages primase-helicase proteins 55,56. Few viral reads could be aligned to the bacteriophage primase-helicase, and the bacterial *dnaG* marker had low identities with viral genomes (maximum identity: 53.7%, length 41 nucleotides). Moreover, *dnaG* only accounted for 6.5% of the overall read-abundance measure on the 31 markers, and its abundance across viromes and metagenomes was comparable with the abundance of other markers that were not found in viruses (**Fig. 3.3**). In addition to *dnaG* only four other ribosomal proteins were found in viral genomes, and in only one genome (**Supplementary Table 3.5**). Overall, only 16.1% of the ribosomal proteins could be found in a viral genome deposited in RefSeq.

Next, we mapped the markers also against IMG/VR (49), a broader database of viral genomes that also contains uncultivated (i.e. metagenomic assemblies) sequences. Indeed, the detectability of few ribosomal proteins in viral genomes increased. However, the greatest increase was observed among the “uncultivated viral genomes” (**Supplementary Table 3.5**). While these genomes may be affected by the same issues of performing metagenomic assembly on viromes with various degrees of purity, the presence of ribosomal proteins in viral genomes cannot be completely excluded. However, more than 97% of the viral genomes where a ribosomal protein could be found had only one of the 31 proteins, and in only one genome (IMG/VR uncultivated viral genome 3300007566, scaffold Ga0104970_1000046) we detected three proteins (*frr*, *rpsB*, and *tsf*). In none of the analyzed genomes we could detect more than 3 proteins (**Supplementary Table 3.5**). While it is not possible to completely exclude the possibility of ribosomal proteins to be present in viral genomes, the abundances of the 31 proteins were highly correlated (**Fig. 2B, Fig. 3**), indicating that it is unlikely that such abundances reflect sporadic hits to different proteins encoded by multiple viruses within the community. Therefore, the use of 31 markers together to detect microbial contamination makes the contribution of those few ribosomal proteins negligible.

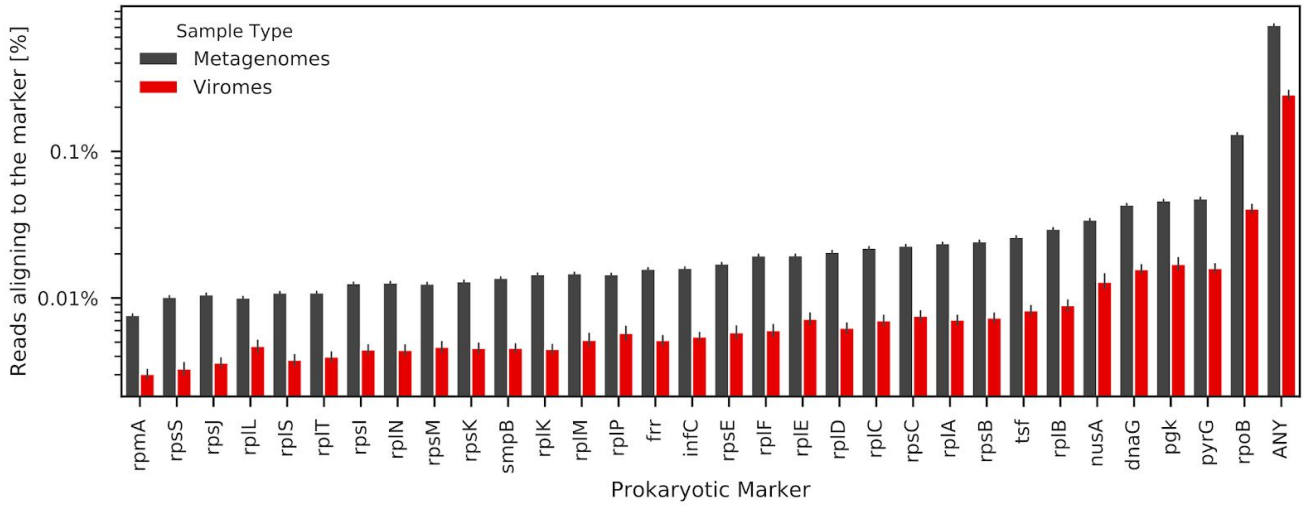
Supplementary Figures



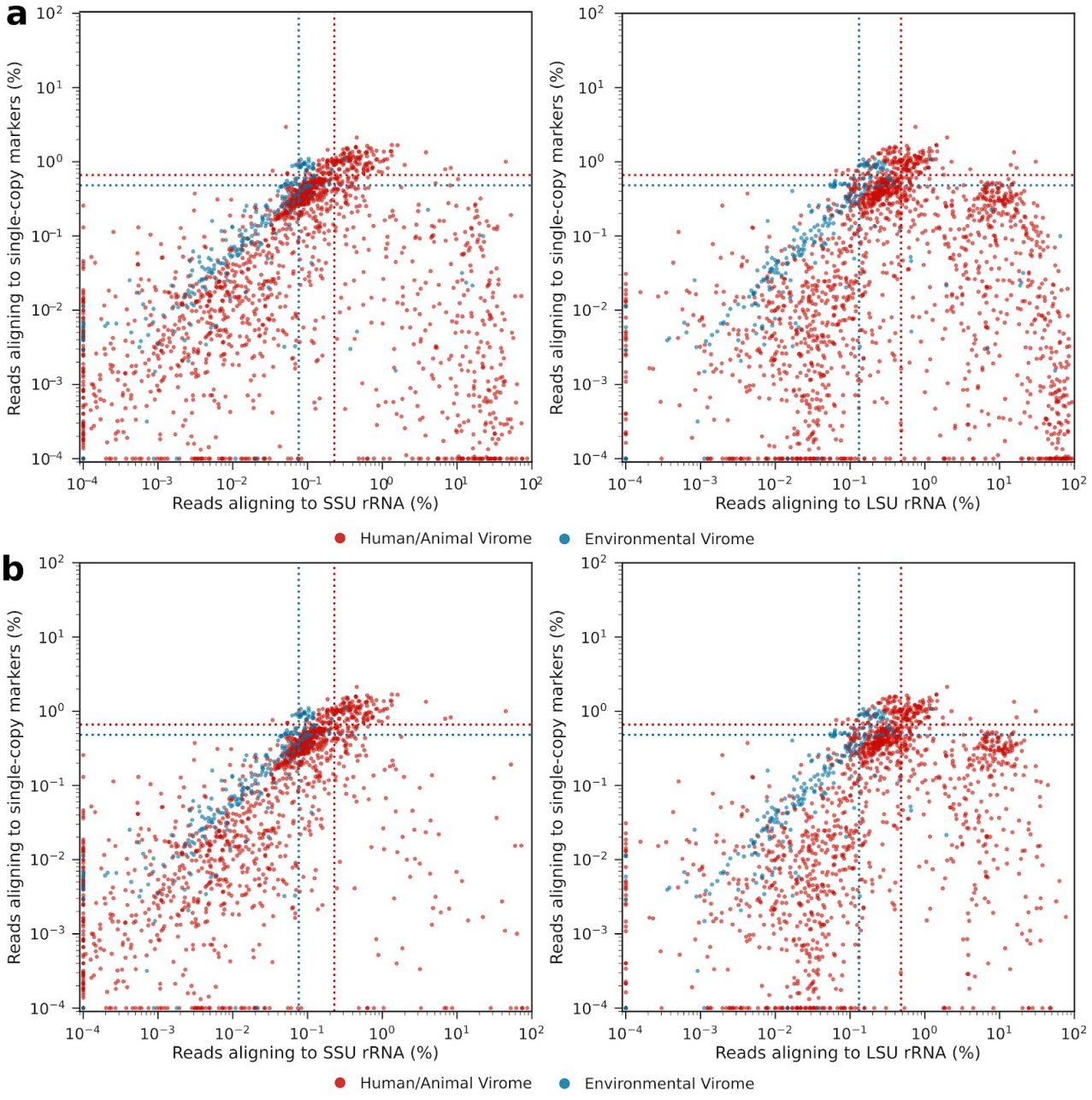
***Supplementary Figure 3.1*. Intra-dataset variability of 32 studies and 26 metagenome datasets.** *Boxplots represent the distribution of intra-dataset change coefficients calculated on the relative abundances of SSU rRNA, LSU rRNA, and single-copy bacterial markers. Human microbiome studies addressing more than one body-site were considered as separate datasets. Three virome datasets that co-sequenced DNA and RNA and that had more than 10% median abundance of either SSU or LSU rRNA were excluded. Viromes were always significantly more variable than Metagenomes (two-sided Mann–Whitney U test p-value of 7.5e-8, 2.2e-6, 2.8e-7 for SSU-rRNA, LSU-rRNA and bacterial-makers respectively). Boxes show the quartiles of each dataset, while whiskers extend to show the rest of the distribution. Outliers are drawn as black diamonds.*



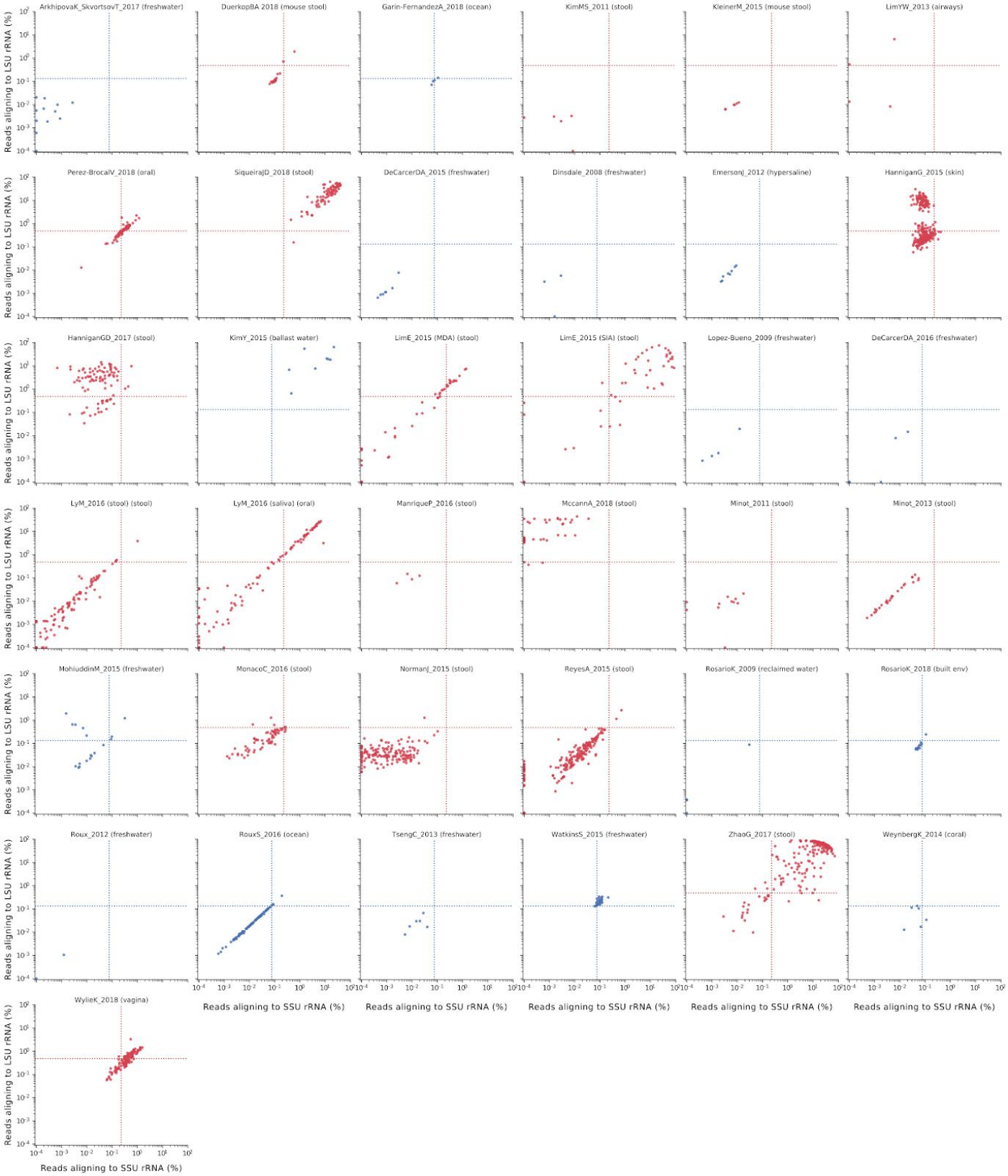
Supplementary Figure 3.2. Percentage of reads aligning to any of the 31 single-copy bacterial markers. *The vertical dotted lines indicate the median of median abundances in human/animal (red dotted line) and environmental (blue dotted line) unenriched metagenomes, as a reference. Datasets are grouped by type and body-site. Datasets within the same group are ordered by median abundance. References to each dataset are provided in (**Tables S1** and **S2**). Error bars in the right barplot reflect the 95% confidence intervals. Boxes show the quartiles of each dataset, while whiskers extend to show the rest of the distribution. Data-points, including outliers, are overlaid to the boxes.*



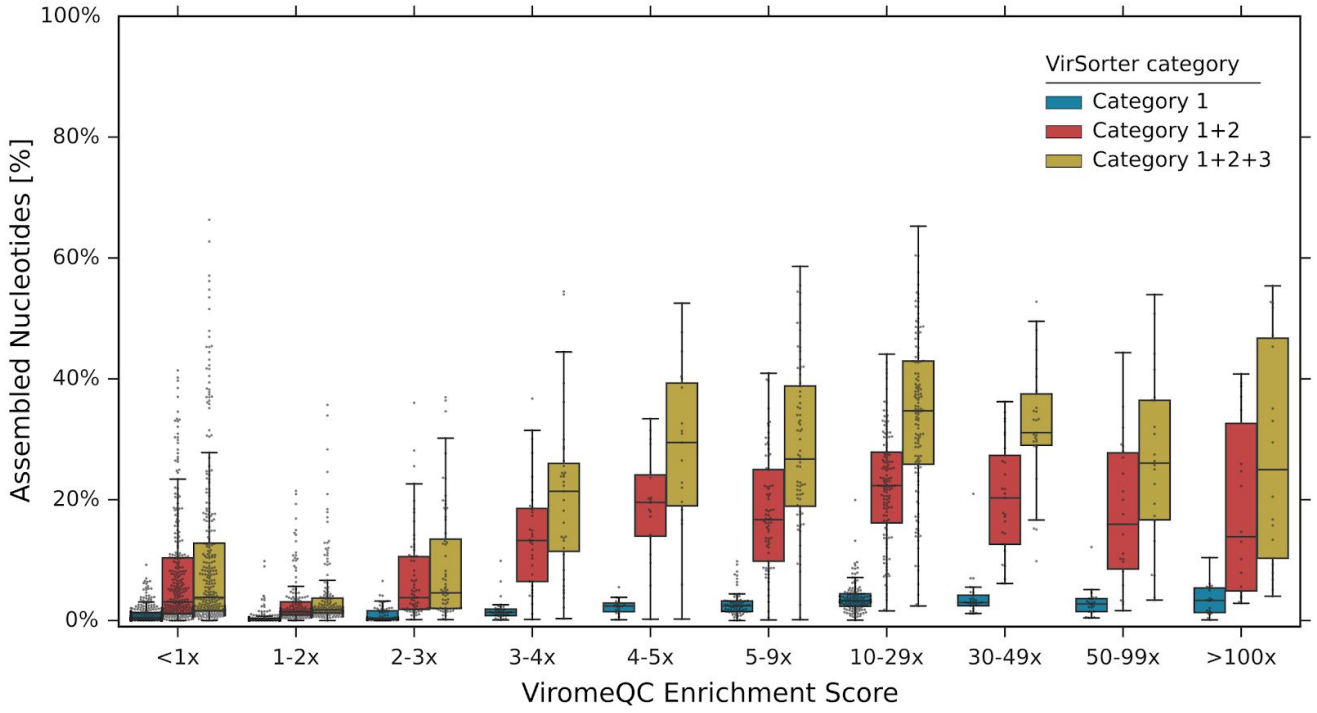
Supplementary Figure 3.3. Abundance single-copy bacterial markers. *The bars reflect the average percentage of reads aligning to each of the 31 single-copy bacterial markers and to any marker (rightmost bar). Metagenomes are represented in gray, while viromes are represented in red. The markers are ordered by the abundance in metagenomes. Error bars reflect the 95% confidence intervals.*



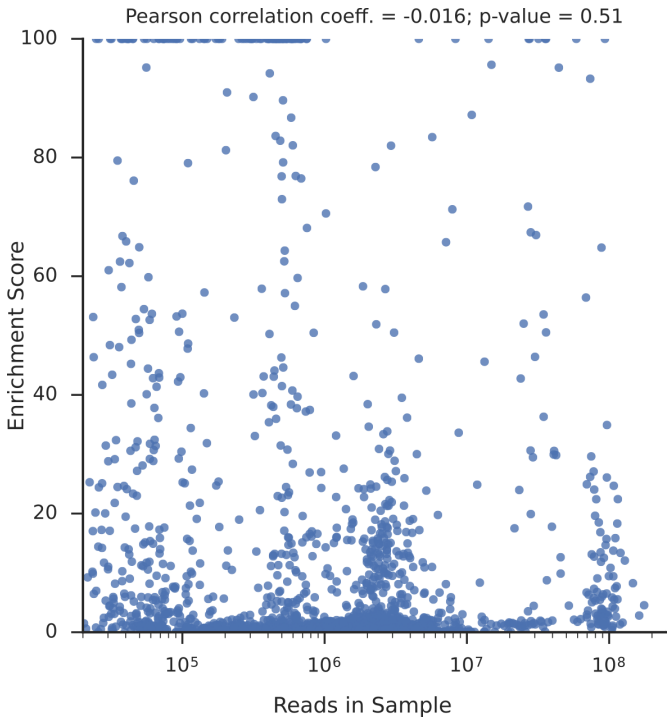
Supplementary Figure 3.4. rRNA and single-copy markers alignment rate. *The retrieved viromes were mapped against rRNA small and large subunits reference sequences (x-axis), and against 31 single-copy bacterial markers (y-axis). The scatter plots show the percentage of aligned reads on all the 1,977 viromes* ***(a)*** *and when removing datasets with a median rRNA (SSU or LSU) abundance greater than 10% and that co-sequenced DNA and RNA (1,640 viromes remaining)* ***(b)****. The dotted lines indicate the median abundances in the corresponding metagenomes and are colored accordingly.*


ArkhipovaK_SkvortsovT_2017 (freshwater)
DuerkopBA 2018 (mouse stool)
Garin-FernandezA_2018 (ocean)
KimMS_2011 (stool)
KleinerM_2015 (mouse stool)
LimYW_2013 (airways)
Perez-BrocalV_2018 (oral)
SiqueiraJD_2018 (stool)
DeCarcerDA_2015 (freshwater)
Dinsdale_2008 (freshwater)
EmersonJ_2012 (hypersaline)
HanniganG_2015 (skin)
HanniganGD_2017 (stool)
KimY_2015 (ballast water)
LimE_2015 (MDA) (stool)
LimE_2015 (SIA) (stool)
Lopez-Bueno_2009 (freshwater)
DeCarcerDA_2016 (freshwater)
LyM_2016 (stool) (stool)
LyM_2016 (saliva) (oral)
ManriqueP_2016 (stool)
MccannA_2018 (stool)
Minot_2011 (stool)
Minot_2013 (stool)
MohiuddinM_2015 (freshwater)
MonacoC_2016 (stool)
NormanJ_2015 (stool)
ReyesA_2015 (stool)
RosarioK_2009 (reclaimed water)
RosarioK_2018 (built env)
Roux_2012 (freshwater)
RouxS_2016 (ocean)
TsengC_2013 (freshwater)
WatkinsS_2015 (freshwater)
ZhaoG_2017 (stool)
WeynbergK_2014 (coral)
WylieK_2018 (vagina)
Reads aligning to LSU rRNA (%)
Reads aligning to SSU rRNA (%)

***Supplementary Figure 3.5 (previous page)**. **rRNA small and large subunits alignment rates by dataset.** The retrieved human/animal (red) and environmental (blue) viromes were mapped against rRNA small (x-axis) and large (y-axis) subunits reference sequences. The dotted lines indicate the median abundances in the corresponding category of unenriched metagenomes.*



***Supplementary Figure 3.6**. **Fraction of assembled contigs labelled as viral by VirSorter.** We assembled 1,445 samples and applied VirSorter to classify each assembled contig. The box plot reflects the distribution of the fraction of total assembled nucleotides in each sample, divided (x-axis) by class of enrichment. The three categories reflect the VirSorter classes of decreasing confidence for complete viruses. The boxes encompass the distribution from the 1st to the 3rd quartile, while the whiskers extend to show data points within 1.5 IQR range. Individual data-points, including outliers, are overlaid to the boxes. Only the 874 samples that generated more than 1 million nucleotides were included, and only contigs longer than 500 nucleotides were considered).*



Supplementary Figure 3.7. Sequencing depth and enrichment score. *The total enrichment score in viromes is plotted against the sequencing depth. The maximum value of the enrichment score is 100 (i.e. values exceeding 100 are forced to the maximum value).*

Supplementary Tables

Supplementary Tables are available at:
https://www.nature.com/articles/s41587-019-0334-5#Sec1. Captions are reported below.

Supplementary Table 3.1. Summary of the 2,050 Viromes datasets considered in the analysis. *Dataset sample sizes are related to the actual number of samples that could be classified as DNA VLP viromes according to the available metadata. The reference number refers to* ***Fig. 3.1****.* ***Fig. 3.2D*** *and* ***Supplementary Fig. 3.1****.*

Supplementary Table 3.2. Summary of the 2,189 Metagenomes and 109 synthetic metagenomes and mock communities considered in the analysis. *Dataset sample sizes are related to the actual number of samples that could be classified as DNA metagenomes according to the available metadata. The reference number refer to* ***Fig. 3.1****.* ***Fig. 3.2D*** *and* ***Supplementary Fig. 3.1****.*

Supplementary Table 3.3. Full dataset of metagenomes and viromes. *Contaminant abundances and enrichment data for all the 1,871 metagenomes, 1,670 viromes and 109 synthetic and mock communities that passed all the quality controls. Sample type and number of starting reads are provided, as well as the percentage of SSU and LSU rRNAs stratified by life domain.*

Supplementary Table 3.4. Validation of the rRNA mapping approach. *Expected abundances of 16S rRNA genes are reported for the 108 synthetic and mock communities (****Tab 1****) and 917 16S amplicon sequencing samples (****Tab 2****). Control metagenomes and 16S samples were mapped against the SSU-rRNA genes and filtered at different stringency thresholds (see* ***Methods****). For the amplicon 16S samples at the expected value was set to 100%. The selected threshold is highlighted in blue. The composition of each synthetic metagenome is reported in (****Tab 3****). The rRNA abundances in RNA viromes are reported in (****Tab 4****).*

Supplementary Table 3.5. Detection of single-copy bacterial markers in viral genomes. *Number of genomes in each database in which the 31 single-copy markers are detected. The IMG/VR database was split into Isolate Viruses and Uncultivated Viruses (****Tab 1****). Number of distinct single-copy markers detected in each database (****Tab 2****).*

Acknowledgements and Contributions

Funding. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 716575) to N.S. The work was also supported by MIUR 'Futuro in Ricerca'' RBFR13EWWI_001 and by the European Union (H2020-SFS-2018-1 project MASTER-818368 and H2020-SC1-BHC project ONCOBIOME-825410) to N.S.

Contributions. Study conception and design: M.Z. and N.S. Methodology and analysis: M.Z., F.P., F.A., A.T., F.B. and N.S. Public datasets collection and curation: M.Z. and P.M. All authors contributed to the writing of the final manuscript.

References

1. A. N. Shkoporov, C. Hill, Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell Host Microbe* **25**, 195–209 (2019).
2. C. A. Suttle, Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801 (2007).
3. D. Paez-Espino, *et al.*, Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
4. A. Reyes, *et al.*, Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).
5. C. Canchaya, G. Fournous, S. Chibani-Chennoufi, M. L. Dillmann, H. Brüssow, Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.* **6**, 417–424 (2003).
6. X. Wang, *et al.*, Cryptic prophages help bacteria cope with adverse environments. *Nat. Commun.* **1**, 147 (2010).
7. G. P. C. Salmond, P. C. Fineran, A century of the phage: past, present and future. *Nat. Rev. Microbiol.* **13**, 777–786 (2015).
8. S. Roux, F. Enault, B. L. Hurwitz, M. B. Sullivan, VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
9. J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, F. Sun, VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
10. S. Rampelli, *et al.*, ViromeScan: a new tool for metagenomic viral community profiling. *BMC Genomics* **17**, 165 (2016).
11. R. V. Thurber, M. Haynes, M. Breitbart, L. Wegley, F. Rohwer, Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).
12. M. Kleiner, L. V. Hooper, B. A. Duerkop, Evaluation of methods to purify virus-like particles for

metagenomic sequencing of intestinal viromes. *BMC Genomics* **16**, 1–15 (2015).

13. C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, N. Segata, Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

14. J. M. Norman, *et al.*, Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447–460 (2015).

15. K. M. Wylie, *et al.*, Metagenomic analysis of double-stranded DNA viruses in healthy adults. *BMC Biol.* **12**, 71 (2014).

16. C. L. Monaco, *et al.*, Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome. *Cell Host Microbe* **19**, 311–322 (2016).

17. E. S. Lim, *et al.*, Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* **21**, 1228–1234 (2015).

18. S. Roux, *et al.*, Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).

19. K. Arkhipova, *et al.*, Temporal dynamics of uncultured viruses: A new dimension in viral diversity. *ISME J.* **12**, 199–211 (2018).

20. K. Rosario, N. Fierer, S. Miller, J. Luongo, M. Breitbart, Diversity of DNA and RNA Viruses in Indoor Air As Assessed via Metagenomic Sequencing. *Environmental Science and Technology* **52**, 1014–1027 (2018).

21. S. Roux, M. Krupovic, D. Debroas, P. Forterre, F. Enault, Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* **3**, 130160 (2013).

22. F. Enault, *et al.*, Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J.* **11**, 237–247 (2017).

23. S. Minot, *et al.*, The human gut virome : Inter-individual variation and dynamic response to diet The human gut virome : Inter-individual variation and dynamic response to diet. *Genome Res.*, 1616–1625 (2011).

24. J. B. Emerson, *et al.*, Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Appl. Environ. Microbiol.* **78**, 6309–6320 (2012).

25. S. Minot, *et al.*, Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12450–12455 (2013).

26. Y. Kim, T. G. Aw, T. K. Teal, J. B. Rose, Metagenomic Investigation of Viral Communities in Ballast Water. *Environmental Science and Technology* **49**, 8396–8407 (2015).

27. M. Ly, *et al.*, Transmission of viruses via our microbiomes. *Microbiome* **4**, 64 (2016).

28. A. Reyes, *et al.*, Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11941–11946 (2015).

29. S. C. Watkins, *et al.*, Assessment of a metaviromic dataset generated from nearshore Lake Michigan. *Mar. Freshwater Res.* **67**, 1700–1708 (2016).

30. S. Roux, *et al.*, Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7** (2012).

31. K. D. Weynberg, E. M. Wood-Charlson, C. A. Suttle, M. J. H. van Oppen, Generating viral metagenomes from the coral holobiont. *Front. Microbiol.* **5**, 1–11 (2014).

32. G. D. Hannigan, *et al.*, The human skin double-stranded DNA virome: Topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. *MBio* **6** (2015).

33. D. A. de Cárcer, A. López-Bueno, J. M. Alonso-Lobo, A. Quesada, A. Alcamí, Metagenomic analysis of lacustrine viral diversity along a latitudinal transect of the Antarctic Peninsula. *FEMS Microbiol. Ecol.* **92**, 1–10 (2016).

34. A. McCann, *et al.*, Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* **6**, e4694 (2018).

35. A. N. Shkoporov, *et al.*, Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* **6**, 68 (2018).

36. B. L. Hurwitz, L. Deng, B. T. Poulos, M. B. Sullivan, Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environmental Microbiology* **15**, 1428–1440 (2013).

37. S. Roux, *et al.*, Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **4**, e2777 (2016).

38. E. Pasolli, *et al.*, Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).

39. R. Leinonen, H. Sugawara, M. Shumway, International Nucleotide Sequence Database Collaboration, The sequence read archive. *Nucleic Acids Res.* **39**, D19–21 (2011).

40. M. Zolfo, A. Tett, O. Jousson, C. Donati, N. Segata, MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. *Nucleic Acids Res.*, gkw837 (2016).

41. C. Quince, *et al.*, DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18**, 181 (2017).

42. Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

43. G. Zhao, *et al.*, Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proceedings of the National Academy of Sciences*, 201706359 (2017).

44. J. D. Siqueira, *et al.*, Complex virome in feces from Amerindian children in isolated Amazonian villages. *Nat. Commun.* **9**, 4270 (2018).

45. F. J. Stewart, E. A. Ottesen, E. F. DeLong, Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J.* **4**, 896–907 (2010).

46. M. Wu, A. J. Scott, Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033–1034 (2012).

47. C. M. Mizuno, *et al.*, Numerous cultivated and uncultivated viruses encode ribosomal proteins. *Nat. Commun.* **10**, 752 (2019).

48. J. R. Brister, D. Ako-Adjei, Y. Bao, O. Blinkova, NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–7 (2015).

49. D. Paez-Espino, *et al.*, IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Res.* **45**, D457–D465 (2017).

50. D. Bulgarelli, *et al.*, Structure and function of the bacterial root microbiota in wild and domesticated barley. *Cell Host Microbe* **17**, 392–403 (2015).

51. D. McDonald, *et al.*, American Gut: an Open Platform for Citizen Science Microbiome Research.

mSystems **3** (2018).

52. H. B. Nielsen, *et al.*, Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).

53. C.-H. Tseng, *et al.*, Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *ISME J.* **7**, 2374–2386 (2013).

54. All-Russia Research Institute for Agricultural Microbiology, Analysis of 16S rRNA soil metagenome in different parts of Russia.

55. C. Quince, *et al.*, DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18**, 181 (2017).

56. K. Rosario, C. Nilsson, Y. W. Lim, Y. Ruan, M. Breitbart, Metagenomic analysis of viruses in reclaimed water. *Environ. Microbiol.* **11**, 2806–2820 (2009).

57. Y. W. Lim, *et al.*, Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. *J. Cyst. Fibros.* **12**, 154–164 (2013).

58. A. Lopez-Bueno, *et al.*, High Diversity of the Viral Community from an Antarctic Lake. *Science* **326**, 858–861 (2009).

59. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

60. C. Quast, *et al.*, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).

61. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

62. R. Jain, M. C. Rivera, J. A. Lake, Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 3801–3806 (1999).

63. E. Pasolli, *et al.*, Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019).

64. K. E. McElroy, F. Luciani, T. Thomas, GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics* **13**, 74 (2012).

65. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

66. E. Jones, T. Oliphant, P. Peterson, {SciPy}: Open source scientific tools for {Python} (2001--).

67. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

68. F. Krueger, Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* (2015).

69. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

70. D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

Chapter 4

Discovering and exploring the hidden diversity of the human gut virome

4.1 | Contribution and Context

This paper presents a methodology to incorporate the notion of viral enrichment into the process of *de-novo* viral genome discovery. Viral reference genomes are essential for the optimal detection and characterization of viruses, but the number of available references is still extremely low. In this paper, we describe the approach we followed to reconstruct >150,000 potentially viral sequences, most of which are of unknown origin. To do so, we used ViromeQC (see **Chapter 3**) to select the viromes with the highest enrichment (i.e. >50x, which is 50 times the enrichment in a standard non-enriched metagenome). Because the contamination was extremely low, we considered the metagenomic contigs reconstructed from those samples to be high-confidence viral sequences. We extensively refined and analyzed the resulting set of sequences that we grouped in 3,944 viral sequence clusters (VSCs). Most of the retrieved sequences had no match to known viruses, in accordance with the available literature. In the paper we prove that the sequences reconstructed from highly-enriched viromes can be used to retrieve additional diversity in "regular" metagenomes and even in unenriched viromes. This proves the usefulness of our resource even further, as it underscores that there is a great amount of undiscovered diversity in viromes. To allow future studies on the VSCs, we released the collection of sequences for public use: http://segatalab.cibio.unitn.it/data/VDB_Zolfo_et_al.html.

In this paper, I performed the selection of the highly-enriched viromes and curated their preprocessing and metagenomic assembly. I wrote the scripts to perform the analysis and performed the evaluation of the prevalence of each viral cluster in metagenomes and viromes. I led the efforts for the phylogenetic analysis and characterized the initial set of sequences with the virome-analysis tools VirSorter, ViralVerify, and Seeker.

4.2 | Manuscript

Discovering and exploring the hidden diversity of the human gut virome

Moreno Zolfo [1], Andrea Silverj [1], Paolo Manghi [1], Omar Rota-Stabelli [2], Federica Pinto [1], Nicola Segata [1,*]

[1] Department CIBIO, University of Trento, Trento, Italy

[2] Fondazione Edmund Mach, S. Michele all'Adige, Italy

[*] Corresponding author: Nicola Segata (nicola.segata@unitn.it)

[In finalization for the submission to a scientific journal]

Abstract

Viruses are extremely abundant and relevant in the context of the human microbiome, but it is still surprisingly challenging to accurately profile them with cultivation-free metagenomics. The main limitations are the low representativeness of known viral genomes in microbiome samples and the issues related to expanding such set of references. As a result, many viruses are still transparent to metagenomics even when considering the power of de-novo metagenomic assembly and binning, which are effective for the discovery of bacterial species, but much less successful on viruses. In this work we describe a novel approach for the cataloging of new virome members of the human gut microbiome and present the resulting comprehensive human virome resource. We first retrieved >3,000 viromes and evaluated the efficiency of the Viral-Like Particle (VLP) enrichment, and then we used multiple steps involving thousands of metagenomes to expand and refine it. We report our findings on >162,000 highly-trusted viral sequences we recovered from thousands of metagenomes and viromes and started describing their characteristics. Most of the retrieved sequences were of unknown origin, and few sequences could be detected in >70% of the gut metagenomes we surveyed. We released the collection of sequences for further virome and metagenomic studies of the human microbiome.

Introduction

Viruses are the most abundant biological entity on Earth and are important players in many environments (1, 2). Among these environments, the human gut microbiome is of particular recent interest for its role in human health (3). Bacteriophages, in particular, compose the majority of the human gut viruses, the human "virome" (4, 5), and they have been associated with shifts in the resident microbial community in the gut, both as facilitators of lateral gene transfer (6) and as shapers of microbial communities (7, 8). Indeed, bacteriophages have been implied in many conditions and diseases including Inflammatory Bowel Disease (9, 10), malnutrition (11), cancer (12), and diabetes (13), but are likely to be even more crucial for promoting and maintaining gut and systemic health (14). Viromes are also intriguing for their potential as modulators of the gut ecosystem although their use as reproducible therapeutic tools is still far from being possible.

Metagenomics is the study of the overall genetic content of a sample (15–17) and enables the study of the human microbiome without the need for cultivation through shotgun next-generation sequencing (NGS). However, while NGS approaches can provide a census of the human microbiome at an unprecedented depth, the identification of viruses with metagenomics is still a daunting task for all the viral particles not already catalogued via experimental isolation. The most relevant limitations in sequence-based viral characterization are the huge diversity of virome members and the lack of universal genomic markers for viruses, which hinder the identification of viral entities from the raw sequencing reads. Numerous computational tools were developed to enable viral detection from metagenomes (18–22), some of which are also capable of *de-novo* viral prediction. However, all these approaches rely, to some extent, on the information available from previously characterized viruses. This is true both for tools that search for direct matches against viral genes, proteins, as well as for those that exploit viral genetic patterns or use trained machine learning models to predict viruses.

Developing genomic catalogs of viral genomes is therefore crucial as a fundamental step in phage description and as a reliable and comprehensive reference for NGS virome studies. However, to date, only 12,194 viral genomes are available in public repositories such as RefSeq (23), and most of them belong to pathogens of clinical relevance. While several recent studies exploited sequence assembly to reconstruct genomes (Metagenomic Assembly Genomes, or MAGs) from metagenomes and released atlases of unprecedentedly large bacterial diversity (24–27), these approaches rely on steps such as contig binning (28, 29) and marker-based quality control (30) that do not work properly on viruses. Such

computational limitations for the *de-novo* identification of divergent viral sequences (the so-called “viral dark matter”) are only partially overcome by experimental Viral Like Particle (VLP) enrichment protocols (4, 31, 32) that are a family of protocols used to selectively enrich for viruses prior to sequencing. These techniques include combinations of filtrations, concentrations, and selective gradient-purification in an aim to produce a sample composed mainly of viral sequences (i.e. a virome), but we showed that viral enrichment can have widely different efficiencies, even within the same dataset and when the same experimental procedure is used for all samples (33). While poorly enriched viromes can be useful for known viruses detection and profiling, if the goal is to perform *de-novo* viral discovery, the purity of the virome is a crucial limiting aspect. Overall, new tools and approaches are needed to expand the current knowledge of the virome fraction of human- and non-human-associated microbial systems.

In this study, we exploited highly enriched viromes to drive *de-novo* the discovery of potentially viral sequences from metagenomes and viromes. We considered more than 3,000 viromes and selected the ones with the highest viral enrichment for further analysis. Potential prevalent viruses reconstructed from these high-purity viromes were then mapped against >15,000 raw metagenomes and low-purity viromes. By combining these mapping with screening against known bacterial and archaeal genomes and collections of metagenome-assembled genomes, we built a catalog of 162,876 sequences that were clustered into 3,944 Viral Sequence Clusters (VSCs) representing high-confidence viral entities. While expected intestinal phages were successfully retrieved, most of such VSCs (85.1%) represented potentially uncharacterized viruses. We then surveyed the diversity and prevalence of such viral entities across many thousands of gut microbiome samples and showed that this resource will potentially serve as a computational base for viral detection. The resource is freely available at http://segatalab.cibio.unitn.it/data/VDB_Zolfo_et_al.html.

Results and Discussion

To identify and catalog truly novel viral genomes and expand current phage databases for the gut microbiome, we developed a 2-step strategy exploiting the increased availability of both metagenomically sequenced VLP-enriched samples (viromes) and unenriched metagenomic samples. First, we selected viromes with the highest enrichment efficiency among the thousands of publicly available samples and considered metagenomically assembled contigs as a highly trusted candidate set of viral sequences we call Highly Enriched Viral Contigs (HEVCs). We then applied a series of sequence-screening steps, not only involving public databases, but crucially exploiting the richness of more than 10,000 metagenomic samples and the associated set of >255M contigs and >150,000 *de-novo* reconstructed bacterial taxa.

For the first step of identifying high-confidence viral contigs from highly enriched viromes, we retrieved 3,044 publicly available viromes from 49 datasets available in NCBI-SRA (see **Methods**). Under the assumption that a higher viral enrichment correlates with a consequent low detection of bacterial, archaeal, and eukaryotic taxa, we used ViromeQC (33) to estimate the VLP-enrichment level of each sample. Of those, 255 samples had a ViromeQC enrichment score >= 50x and were then considered as "*highly enriched"* (**Table 4.1**, **Supplementary Table 4.1**) and a first source of novel high-confident viral diversity.

Selection of contigs from Highly Enriched Human Gut Viromes

We then proceeded to reconstruct genomes from the 255 high-purity virome samples. After uniform preprocessing to remove low quality and short reads (see **Methods**), we performed metagenomic assembly with SPAdes (34) or Megahit (35) (already validated on viromes (36), see **Methods**). In total, $1.5x10^9$ metagenomic reads were assembled, producing 1.3 million contigs longer than 500 nucleotides (**Table 4.1**) of which 120,041 contigs were assembled from human gut viromes and were retained for downstream analysis. Contigs were generally short (overall median contig length = 860, median N50 = 3,863) with 8,488 and 3,258 contigs (7.1%, 2.7%) longer than 5,000 and 10,000 nucleotides, respectively (**Fig. 4.1A**, **Supplementary Table 4.2**). The largest contig (i.e. 393,171 nucleotides), was reconstructed from a sample with an enrichment of 53x (SRR935337 from Minot *et al.* (37)).

To detect and remove contigs still showing evidence of non-viral origin, we annotated the contigs against available reference viral and prokaryotic genomes from isolate sequencing and Metagenome-Assembled Genomes (MAGs) (**Fig 4.1B, Supplementary Table 4.2**).

Overall, 3,932 out of 120,041 contigs had at least one BLAST hit against a viral reference genome in RefSeq (38) suggesting the presence of viruses closely related to known viral species. Conversely, 22,018 (18.34%) had at least one hit against a collection of representative reference bacterial and archaeal reference genomes (26, 39), and 34,424 (28.68%) aligned against at least one of the non-viral MAGs from Pasolli *et al*. (26). While truly viral sequences could match against bacterial genomes because of prophagic regions and MAGs could erroneously include non-bacterial contigs, it was intriguing that 81,212 (67.65%) contigs did not match against any sequence available as reference-based or reference-free catalogued microbial diversity (see **Methods**).

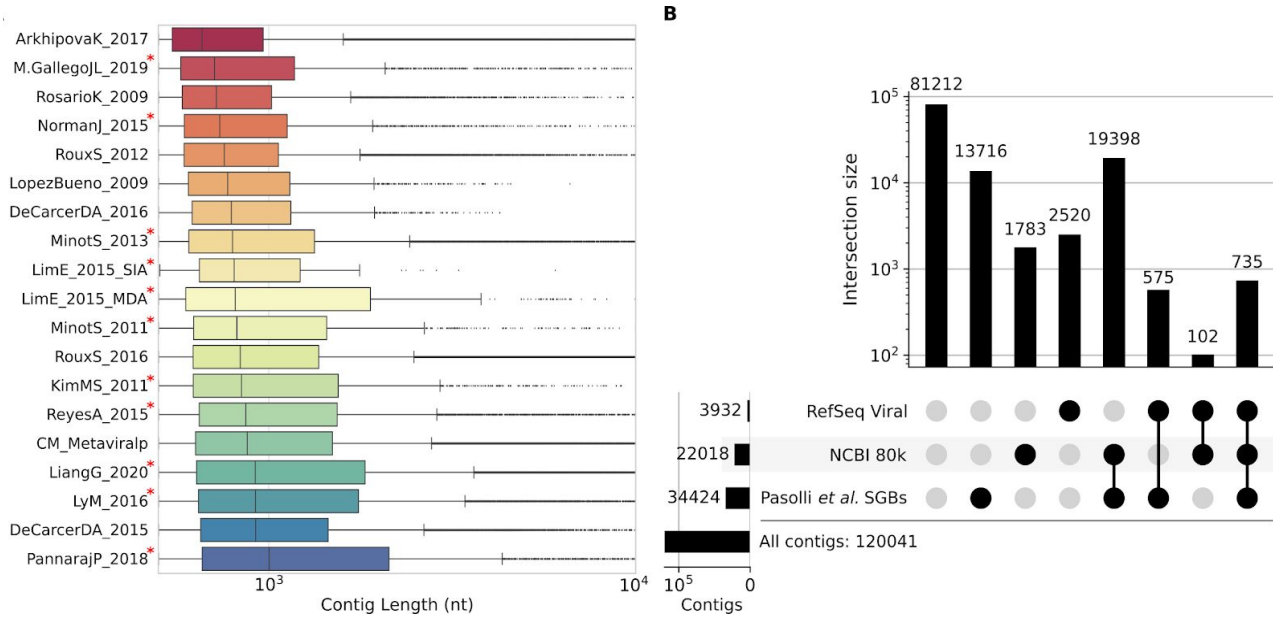


Figure 4.1. Length distribution and potential novelty of contigs assembled from highly enriched viromes. (A) *Length of the 1.3 million contigs assembled from viromes with a ViromeQC enrichment higher than 50x. Only contigs longer than 500 nucleotides were considered. The boxes show the quartiles of each distribution, whiskers extend to show the rest of the distribution, and points that exceed 1.5 IQR are represented as outliers. Red asterisks indicate human gut viromes. Additional data for each sample is available in* ***Supplementary Table 4.1****.* ***(B)*** *Annotation of the 120,041 contigs assembled from highly enriched gut viromes. The upset plot shows the cardinality of each set of contigs that are covered by hits against the RefSeq viral references, a set of bacterial reference genomes (NCBI 80k), and the MAGs from Pasolli et al., 2019. Contigs are annotated in each set if they have a cumulative breadth of coverage >50% with any reference sequence in the set. Intersection sets are highlighted by the continuous lines in the bottom plot.*

Dataset (Reference)	*Sample Type*	*Samples*	*HQ Reads*	*Contigs*	*Median VQC score*	*Median N50*	*Median Contig Length*	*Max Contig Length*
LiangG_2020 (40)	Human Stool	71	4.82E+08	59154	100	5266	890	175037
ReyesA_2015 (11)	Human Stool	45	2.34E+06	10019	87	2589	893	44781
LyM_2016 (41)	Human Stool	43	2.23E+07	12345	100	3009	875	98074
	Human Oral	21	9.84E+06	6201	100	3729	1.109	38713
LimE_2015_MDA (42)	Human Stool	17	2.21E+06	410	100	5440	969	84665
MinotS_2013 (37)	Human Stool	12	3.65E+08	27148	100	9221	789	393171
NormanJ_2015 (9)	Human Stool	6	1.51E+07	3195	63.5	2068	733	124399
DeCarcerDA_2015 (43)	Freshwater	5	5.30E+07	46462	95	1538	955	85920
ArkhipovaK_2017 (44)	Freshwater	4	2.05E+07	281233	70.5	1061	654	140124
PannarajP_2018 (45)	Human Stool	4	3.86E+07	2196	100	6722	948	71301
RouxS_2016 (46, 47)	Ocean	4	3.24E+08	706498	78.5	1864	835	200689
LimE_2015_SIA (42)	Human Stool	3	5.43E+05	120	100	1129	829	6039
Moreno-GallegoJL_2019 (48)	Human Stool	3	7.07E+06	3351	72	11143	702	187123
PannarajP_2018 (45)	Human Milk	3	2.34E+07	1785	96	4761	1.025	71301
RosarioK_2009 (49)	Reclaimed Water	3	7.30E+05	12407	100	967	720	19416
PintoF_2020 (unpublished)	Freshwater	2	8.97E+07	103294	61.5	2204	874	80726
DeCarcerDA_2016 (50)	Freshwater	2	2.57E+05	2579	100	1040	787	4264
KimMS_2011 (51)	Human Stool	2	2.20E+05	1235	92.5	2138	847	24667
MinotS_2011 (52)	Human Stool	2	8.47E+04	868	57	2501	821	31906
RouxS_2012 (53)	Freshwater	2	1.25E+06	46685	79.5	1062	761	27163
LopezBueno_2009 (54)	Freshwater	1	6.13E+05	1817	76	1075	773	6622
Total		**255**	**1.5E+09**	**1.3E+06**				

Table 4.1. Summary of the 255 highly-enriched viromes with a ViromeQC enrichment score >= 50x. *Samples are grouped by dataset and sample type. HQ reads refer to the number of reads that were retained after the quality-control step (see* ***Methods****). Additional data for each individual sample is available in* ***Supplementary Table 4.1****.*

To further minimize the chance of considering contigs from contaminant bacteria even in highly enriched virome samples, we identified matching contigs assembled in multiple highly enriched viromes. To this end, we performed an all-vs-all mapping-based search (55) and found that 3,765 contigs (6.34%) were present in more than one dataset, while 925 contigs (1.5%) were found in at least 4 distinct datasets (**Supplementary Fig. 4.1**). This highlights that a fraction of the retrieved viral contigs are also prevalent in the human gut microbiome, as already reported in other studies (14), and provides an additional confidence score for their viral origin, as it is increasingly unlikely that a bacterial contaminant is found multiple times in unrelated highly enriched virome samples.

Refinement of the viral contigs to remove further non-viral contamination

While a high ViromeQC enrichment score and detection of matching contigs across highly enriched samples provide evidence of low chances of non-viral origin, we further contextualize the 120,041 retrieved contigs using the collection of species-level genome bins (SGBs) retrieved from unenriched metagenomes by Pasolli *et al*. (26) to search for bacterial fragments. We accordingly excluded contigs that substantially overlapped (i.e. at least 1000 nucleotides at more than 80% identity) with the same MAG in any given sample or that were included in MAGs from more than 50 metagenomes. This procedure removed 29,237 (24.36%) contigs from the considered set of viral contigs.

Because virome enrichment has the potential to uncover viruses that could be at very low abundance or that are rare in the population, we then used the compendium of 9,428 available metagenomes to retain potential viruses of relevance for the human microbiome. We did this by focusing on the metagenomically-assembled contigs that were not grouped by contig binning procedures into MAGs, and should thus be enriched by viral contigs as viruses do not match the tetranucleotide frequency and coverage of enough other contigs containing bacterial markers. Only contigs that mapped against at least 20 samples in the unbinned fraction, but already found not to map against bacterial MAGs (see above), were selected. For example, the second-largest contig assembled (387,087 nucleotides from Minot et al, 2013) was detected in the unbinned fraction of >9,000 samples, but also in 2,038 samples as binned in kSGB-4285 (annotated as *Ruminococcus bromii*). The contig indeed overlapped to a *Ruminococcus bromii* genome (NCBI accession NNBY01000036) for 39,148 nucleotides at 98.7% identity and was hence discarded. This filtering removed a further 83,770 contigs. Additionally, 2,082 short contigs (i.e. less than 1,500 nucleotides) were removed. Overall, our approach selected 4,952 contigs that were potentially of viral origin, lowly contaminated by bacteria, and highly prevalent in metagenomes and viromes.

Finally, to better represent the viral diversity of the human gut, we added 699 full genomes of known bacteriophages from RefSeq to the set of 4,952 contigs. We selected 2,619 bacteriophages with a known bacterial host as per their description in NCBI and kept those that were found in at least 20 metagenomes from the unbinned cohort of Pasolli *et al*. (26). This brought the total number of sequences to 5,651.

Assessing the newly identified viral sequences with virome prediction tools

To assess how much our procedure captured uncharacterized viral diversity with the retrieved 5,651 viral contigs, we applied four available tools that have been developed to determine whether an assembled contig is a virus using direct or indirect concepts of similarity with known viruses (see **Methods**, (18–21)). We found that only a small fraction of the retrieved sequences could be labelled as viral by all tools (153 out of 5,651 if considering two operating modes of VirSorter; 186 out of 5,651 if considering Virsorter in *decontamination* mode) suggesting that our resource captures a large fraction of truly novel viral diversity (**Fig. 4.2**).

VirFinder labeled the highest number of sequences as viral (n=3175), while VirSorter in default mode only detected 1,000 contigs. Virsorter in decontamination mode labelled 692 genomes as viral and was in general agreement with the other tools (Viral Verify: 692, VirFinder: 632, Virsorter: 582, Seeker: 142, **Supplementary Table 4.3**). Importantly, a total of 1,052 contigs in the collection could not be labelled as viral by any tool.

When we applied the same tools to a random set of 200 bacterial MAGs and 2,619 bacteriophages genomes, viral-prediction tools were not all equally able to discern between viral and bacterial sequences, although bacterial MAGs contigs were less likely to be labelled as viral (**Supplementary Fig. 4.2**). Nonetheless, 7.9% of the bacterial contigs were labelled as viral by at least one tool, and only 31.1% of the 2,619 reference bacteriophages in RefSeq were labelled as viral by all tools (**Supplementary Fig. 4.2**). This suggests that, while virome-prediction tools can detect viruses from assembled contigs, some sequences may still be transparent to such approaches because of their divergence with respect to any known virus. Conversely, depending on the tool (or combination of tools) used to screen for viruses, bacterial sequences might be wrongfully considered as viral, leading to false positives and incorrect conclusions.

These results highlight that our newly generated resource, which is the only one available that does not rely on already known phages, is discovering truly novel viral diversity not identifiable by available methodologies that are also prone to false positive calls.

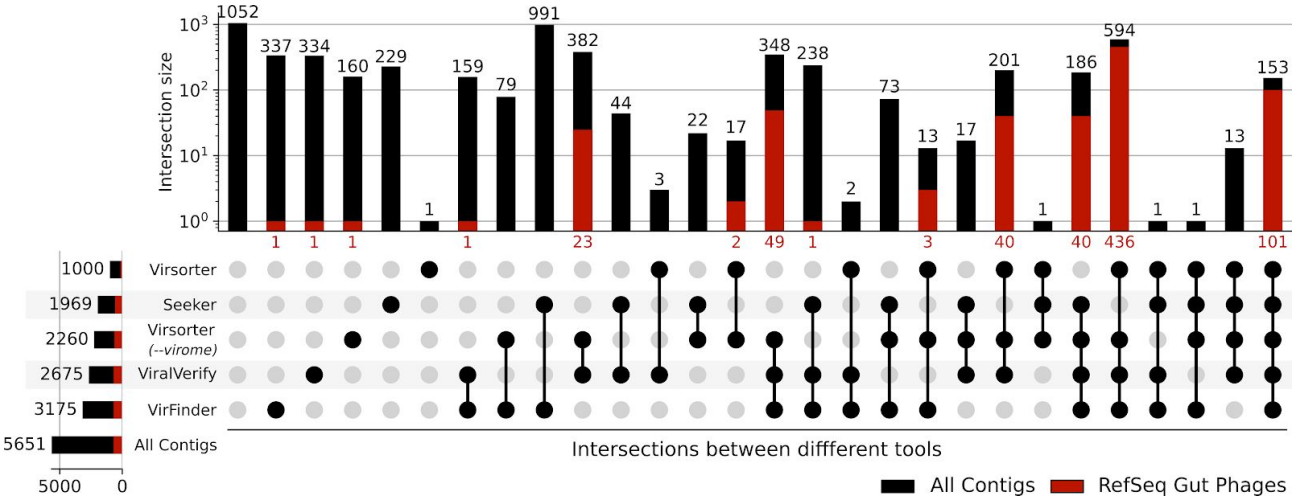


Figure 4.2. Comparison of different viral detection tools on the 5,651 contigs from RefSeq and Highly Enriched Gut Viromes. *The upset plot shows the cardinality of each intersection between sets of contigs that could be labelled as viral by Virsorter, Seeker, ViralVerify, and VirFinder. The set of analyzed contigs includes 4,952 contigs from highly enriched viromes (black bars) and 699 reference genomes of human gut bacteriophages from RefSeq (red bars). The size of each intersection is shown on the top (overall number of contigs) and bottom (RefSeq reference gut bacteriophages) of each bar. Virsorter was used both in normal and decontamination (i.e. --virome) mode.*

Grouping the 5,651 unique viral sequences into 3,944 VSCs

The concept of species is particularly loose for viruses, but because some of the retrieved viral contigs could have very high homology between themselves as they represent very closely related viruses retrieved from independent samples, we clustered the contigs into similarity bins.

The 5,651 contigs were thus clustered at 90% identity with VSEARCH (56), resulting in 3,944 clusters. Of these, 588 contained a viral reference genome and were thus labelled as “known Viral Species Cluster” (kVSC), and 3,357 were labelled as unknown or uncharacterized (uVSC). On average there were 1.38 sequences per cluster and the largest unknown viral cluster (vsearch_c84, see **Supplementary Table 4.3**) included 56 sequences. The largest known viral cluster (vsearch_c1586) grouped 19 sequences, and included the reference genome of the PhiX174 Coliphage (NC_001422.1), underscoring the presence of undepleted Illumina NGS Spike-in control in many metagenomes (57). The second and third largest clusters contained the sequences widely prevalent gut bacteriophage *crAssphage* (58) (NC_024711) and fragments of Bacteroides phage B12414 (NC_0167701), respectively.

Microbiome samples that are not highly enriched for viruses (i.e. standard unenriched metagenomes and viromes with a ViromeQC score < 50x) can contain near-complete viral sequences that could expand the genetic diversity of the identified VSCs. In light of this, we thus used Mash (59) to search for sequences similar to any of the 5,651 HEVCs into a) the contigs assembled from 3,044 viromes and b) contigs in the unbinned fraction of Pasolli *et al.* (minimum mash distance <= 10%, see **Methods**). This retrieved 425,309 homologous contigs that were similar to an HEVC and were included in a global re-clustering together with the original 5,651 HEVCs at 70% identity. Sequences that did not cluster with an HEVC were discarded to eliminate sequences that were too divergent. In total, 157,225 (36.97%) contigs were kept (126,894 from unbinned metagenomes, 30,331 from viromes) with an average of 44.37 (s.d. 125.78) contigs in each uVCS and 14.14 (s.d. 88.96) for kVCS (**Supplementary Table 4.3**). These sequences were added to the original 5,651 HEVCs and constitute our proposed resource of 162.876 novel high-confidence and potentially viral sequences.

To enable easy profiling of such sequences directly from raw metagenomes, we then organized the retrieved viral clusters into a collection of non-redundant sequences that could be easily used by other researchers. To reduce unnecessary sequence redundancy and at the same time to preserve the diversity of sequences within each VSC, we internally clustered all VSCs at 95% identity. We hence selected the longest sequence for each cluster and added the full set of 5,651 highly-enriched contigs, if they were not already the longest sequence within their sub-cluster (see **Methods**). This final step produced a total of 47,820 sequences that were then de-replicated at 99% identity over 90% of their length with CD-HIT (60). The final 45,872 sequences constituting the refined viral sequences collection were, for the most part, of metagenomic origin and assigned to unknown VSCs (**Table 4.2**). Sequences together with clustering information and additional annotations are available for phylogenetic and functional analyses and as resources for future studies (see **Code and Data Availability**).

Contig Source	Sequences	Avg Len.	St. Dev. Len.	Median Len.	Max Len.
Highly Enriched Viromes	3,467	5,774.23	9,950.35	2,991	167,064
of which in uVSCs	*(3,418)*	*(5,620.47)*	*(9,263.16)*	*(2,978)*	*(167,064)*
Viral RefSeq	684	58,780.93	44,793.82	41,718	240,413
Viromes	8,804	12,853.21	20,580.28	5,615	204,229
of which in uVSCs	*(8,437)*	*(12,288.42*	*(19,755.01)*	*(5,476)*	*(204,229)*
Metagenomes	32,917	20,923.09	30,367.33	7,592	198,476
of which in uVSCs	*(31,331)*	*(20,371.18)*	*(30,185.44)*	*(7,112)*	*(198,476)*
Any (Total)	**45,872**	**18793.83**	**28,758.97**	**6,485**	**240,413**

Table 4.2. Composition and origin of the 45,872 sequences in the VSCs representatives collection. *Sequences in each Viral Sequence Cluster (VSC) were further clustered at 95% identity and the longest sequence of each sub-cluster was selected for the final sequence collection. This produced 47,820 sequences, that were then de-replicated to avoid duplicates, and produced 45,872 sequences. Statistics of sequences in unknown VSCs (uVSCs) are reported in brackets below each row.*

Prevalence of the viral clusters in metagenomes and viromes

We then used the new catalog of human gut viruses to evaluate the prevalence of each viral cluster in the metagenomes and viromes originally used to retrieve the contigs. We defined the dataset-wide prevalence as the percentage of datasets in which a VSC could be retrieved in at least one sample-specific assembly. Out of the 3,944 clusters, 339 (8.59%) were retrieved at least once in half of the gut metagenomic datasets (**Supplementary Table 4.4**). Among the most prevalent known viral clusters there was the recently discovered *crAssPhage* (kVSC c72, detected in 45 datasets in at least one sample, average prevalence when detected 6.61% s.d. 5.26%) (58), as well as several Enterobacteria and Shigella phages (**Fig. 4.3A**). Of note, the fifth most prevalent viral cluster (kVSC c1586) was annotated as Coliphage phiX174, and was retrieved at least once in 15 out of 44 metagenomic datasets (34.1%) and in 10 out of 19 virome datasets. Moreover, across the viromes in which phiX174 could be retrieved, the cluster had an average prevalence of up to 60.4% (**Fig. 4.3C, Supplementary Table 4.4**), further pointing at this pervasive artifact of undepleted Illumina spike-in control.

The majority of highly prevalent VSCs were unknown (i.e. no RefSeq viral genomes were present in the cluster). Indeed, the three most prevalent clusters (c2353, c3062, c3861) attracted 8,838 sequences (8,333 of which from metagenomes). Further investigation on the contigs in these clusters indicated that sequences were extremely conserved, despite coming from different samples and datasets (median pairwise identity on multiple-sequence alignment > 98.7%, **Fig. 4.3B**). None of the sequences matched against any artificial vector, linker, adapter, and primer in UniVec, and only 4.25% of the sequences in cluster c3062 matched against UniVec entries that could be both of natural and artificial origin (see **Methods**). The seven representative sequences of the three clusters (i.e. the sequences extracted from highly-enriched viromes) were not identified as viral by VirSorter, ViralVerify, Seeker, and VirFinder, nor they had matches against sequences in viral-RefSeq. However, sequences matched against reference bacterial genomes of *Bacteroides fragilis* (cluster c2353), Clostridium sp. (cluster c3062), and *Dorea longicatena* (cluster c3861). Matches overlapped with plasmids or with bacterial genomic regions involved in plasmid replication. Interestingly, while contigs from metagenomes may include bacteria and plasmids, sequences extracted from highly enriched viromes originated from two studies that depleted free nucleic acids using DNase I (40, 41). While a viral origin of those sequences cannot be completely excluded, the presence of extremely prevalent and highly conserved sequences matching to plasmids demands further analysis on the effect of plasmid infiltration in Viral-Like Particle purification.

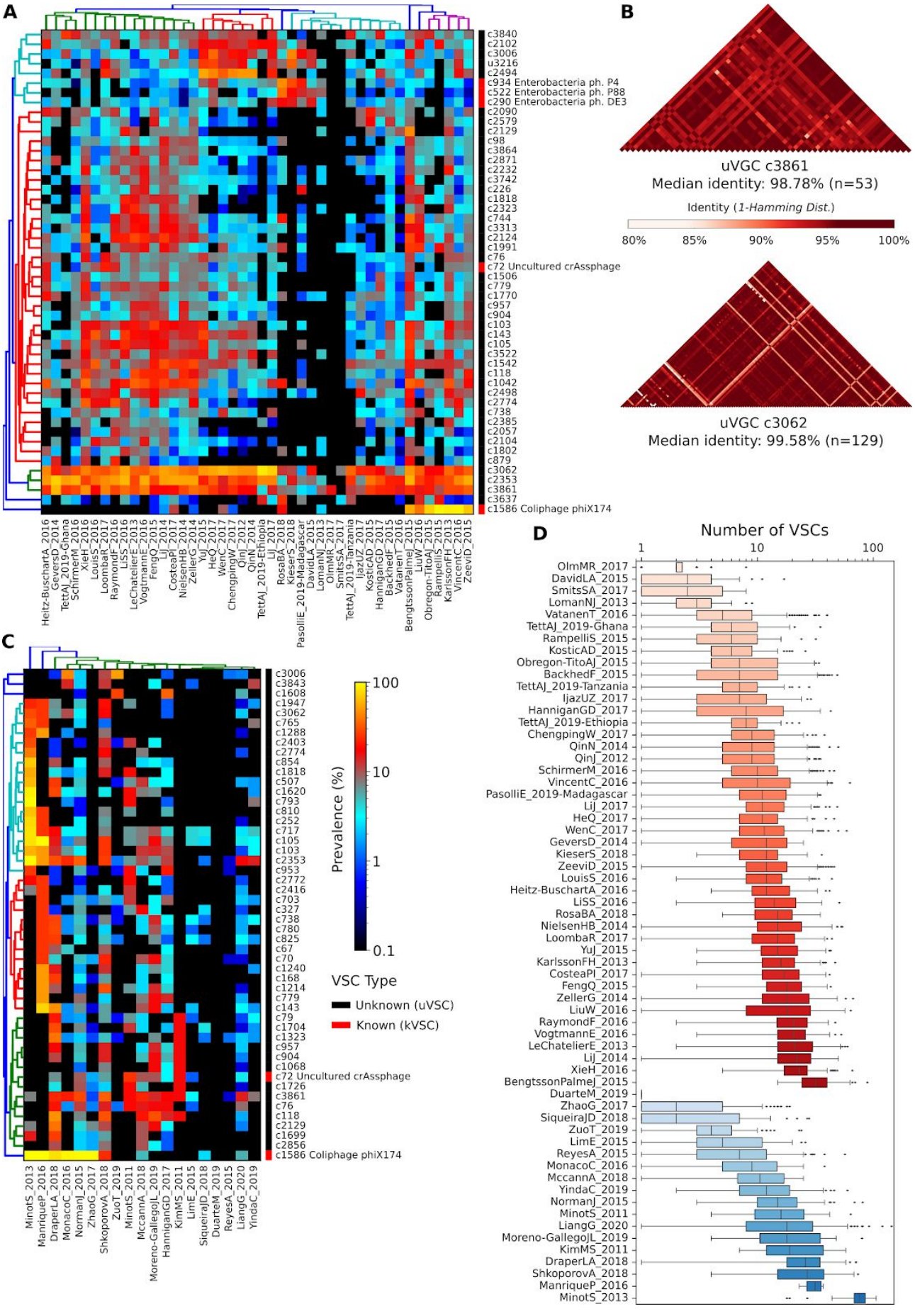

A
c3840
c2102
c3006
u3216
c2494
c934 Enterobacteria ph. P4
c522 Enterobacteria ph. P88
c290 Enterobacteria ph. DE3
c2090
c2579
c2129
c98
c3864
c2871
c2232
c3742
c226
c1818
c2323
c744
c3313
c2124
c1991
c76
c72 Uncultured crAssphage
c1506
c779
c1770
c957
c904
c103
c143
c105
c3522
c1542
c118
c1042
c2498
c2774
c738
c2385
c2057
c2104
c1802
c879
c3062
c2353
c3861
c3637
c1586 Coliphage phiX174
Heitz-BuschartA_2016
GeversD_2014
TettAJ_2019-Ghana
SchirmerM_2016
XieH_2016
LouisS_2016
LoombaR_2017
RaymondF_2016
LiSS_2016
LeChatelierE_2013
VogtmannE_2016
FengQ_2015
LiJ_2014
CosteaPI_2017
NielsenHB_2014
ZellerG_2014
YuJ_2015
HeQ_2017
WenC_2017
ChengpingW_2017
QinJ_2012
QinN_2014
TettAJ_2019-Ethiopia
LiJ_2017
RosaBA_2018
KieserS_2018
PasolliE_2019-Madagascar
DavidLA_2015
LomanNJ_2013
OlmMR_2017
SmitsSA_2017
TettAJ_2019-Tanzania
IjazUZ_2017
KosticAD_2015
HanniganGD_2017
BackhedF_2015
VatanenT_2016
BengtssonPalmeJ_2015
LiuW_2016
Obregon-TitoAJ_2015
RampelliS_2015
KarlssonFH_2013
VincentC_2016
ZeeviD_2015
B
uVGC c3861
Median identity: 98.78% (n=53)
Identity (1-Hamming Dist.)
80%
85%
90%
95%
100%
uVGC c3062
Median identity: 99.58% (n=129)
C
c3006
c3843
c1608
c1947
c3062
c765
c1288
c2403
c2774
c854
c1818
c507
c1620
c793
c810
c252
c717
c105
c103
c2353
c953
c2772
c2416
c703
c327
c738
c780
c825
c67
c70
c1240
c168
c1214
c779
c143
c79
c1704
c1323
c957
c904
c1068
c72 Uncultured crAssphage
c1726
c3861
c76
c118
c2129
c1699
c2856
c1586 Coliphage phiX174
Prevalence (%)
100
10
1
0.1
VSC Type
Unknown (uVSC)
Known (kVSC)
MinotS_2013
ManriqueP_2016
DraperLA_2018
MonacoC_2016
NormanJ_2015
ZhaoG_2017
ShkoporovA_2018
ZuoT_2019
MinotS_2011
MccannA_2018
Moreno-GallegoJL_2019
HanniganGD_2017
KimMS_2011
LimE_2015
SiqueiraJD_2018
DuarteM_2019
ReyesA_2015
LiangG_2020
YindaC_2019
D
Number of VSCs
1
10
100
OlmMR_2017
DavidLA_2015
SmitsSA_2017
LomanNJ_2013
VatanenT_2016
TettAJ_2019-Ghana
RampelliS_2015
KosticAD_2015
Obregon-TitoAJ_2015
BackhedF_2015
TettAJ_2019-Tanzania
IjazUZ_2017
HanniganGD_2017
TettAJ_2019-Ethiopia
ChengpingW_2017
QinN_2014
QinJ_2012
SchirmerM_2016
VincentC_2016
PasolliE_2019-Madagascar
LiJ_2017
HeQ_2017
WenC_2017
GeversD_2014
KieserS_2018
ZeeviD_2015
LouisS_2016
Heitz-BuschartA_2016
LiSS_2016
RosaBA_2018
NielsenHB_2014
LoombaR_2017
YuJ_2015
KarlssonFH_2013
CosteaPI_2017
FengQ_2015
ZellerG_2014
LiuW_2016
RaymondF_2016
VogtmannE_2016
LeChatelierE_2013
LiJ_2014
XieH_2016
BengtssonPalmeJ_2015
DuarteM_2019
ZhaoG_2017
SiqueiraJD_2018
ZuoT_2019
LimE_2015
ReyesA_2015
MonacoC_2016
MccannA_2018
YindaC_2019
NormanJ_2015
MinotS_2011
LiangG_2020
Moreno-GallegoJL_2019
KimMS_2011
DraperLA_2018
ShkoporovA_2018
ManriqueP_2016
MinotS_2013

Figure 4.3 (previous page). Prevalence of the Viral Sequence Clusters (VSCs) in assembled Metagenomes and Viromes. (A-C) *Number detected VSC across gut metagenomes* ***(A)*** *and viromes* ***(C)*** *normalized by the number of samples in each dataset. Side boxes indicate known (kVSC, red) and unknown (uVSCs, black) clusters.* ***(B)*** *Pairwise sequence identity of sequences within two of the most prevalent VSCs in metagenomes (c3861, prevalence = 97.7% and c3062, prevalence = 95.5%). Sequence identity was calculated from multiple-sequence alignment of contigs within 25% of the median length of the highly-enriched contigs in the cluster.* ***(D)*** *Number of distinct VSCs detected in each stool sample, grouped by dataset. Boxes encompass distribution quartiles, whiskers extend to 1.5 IQR. Red-shaded boxes refer to metagenomes, while blue-shaded boxes refer to viromes. Datasets with samples of multiple sources (e.g. stool + oral metagenomes) are not included in the figure.*

Phylogenetic analysis of a selection of the retrieved VSCs

Further analysis of selected Viral Sequence Clusters (VSCs) allowed us to reconstruct the phylogeny of several potentially viral sequence groups. Sequences were analyzed in a maximum likelihood framework supported by manual curation of the alignments (see **Methods**).

First, we analyzed known VSCs focusing on kVSC c718_c0 and kVSC c230_c0, two clusters containing the *Lactococcus phage bIL67* (accession NC_001629*,* **Fig. 4.4A**) and *Klebsiella phage KpnP* (accession NC_028670, **Fig. 4.4C**), respectively. Most of the aligned sequences originated from metagenomic contigs, with only one sequence retrieved from a virome from Liang *et al. (40)*. We were also able to reconstruct the phylogeny of 262 sequences of kVSC c72_c0 (crAssphage cluster, **Supplementary Fig. 4.3**). The crAssphage-like sequences retrieved from highly-enriched viromes were all within the same clade, while contigs from viromes and metagenomes extended the phylogeny. This demonstrates that even unenriched samples can be useful for viral discovery, as previously unseen viral diversity can be extracted from unenriched samples by exploiting highly enriched viral contigs.

We then incorporated metadata from the original studies into the phylogenetic trees, with the goal to identify individual-specific phage variants and to test whether they are retained when longitudinal sampling was available. In the Lactococcus phage phylogeny we identified at least two couples of samples of the same individual taken at different time-points (**Fig 4.4A**). This was also seen in several other VSCs in longitudinal datasets (37, 40, 61), where multiple samples per individual were analyzed (48, 62), or where individuals shared

households (11, 40, 41). As an example, in the unknown cluster uVSC c811_c0, one clade contained sequences from the same sample sequenced with different library preparations (Moreno-GallegoJL_2019), while the other clade was composed mainly by very similar sequences from the same individual sampled longitudinally (MinotS_2013, **Fig. 4.4B**). Also in uVSC c1292_c0, sequences from twins that shared the same household clustered in the same branch (ReyesA_2015, **Fig. 4.4E**). At least 38% of the Open Reading Frames of these three uVSCs matched a viral motif of V-FAM ((63), see **Methods**). All uVSCs had partial matches with MAGs of gut-associated bacterial taxa as *Holdemanella_biformis* (c811) *Bifidobacterium adolescentis* (c570) and *Prevotella copri* (c1292). Both VirFinder and ViralVerify labelled the three clusters as viral, while VirSorter labelled only one (c1292) as a sure virus, and the other two as potential viruses (**Supplementary Table 4.3**).

The detection of the same sequence in samples of the same individuals highlights that not only some VSCs are prevalent across datasets, but also that they can be retained in time by individuals and can be shared within households (**Fig. 4.4E, Supplementary Fig. 4.4**). This is similar to what has been found by other studies, both for bacteria and viruses in the human microbiome (48, 64–66). We were also able to reconstruct and analyze the phylogeny of the cluster that included the PhiX174 Coliphage. We noticed a complete absence of phylogenetic signal (median pairwise identity = 100%, average pairwise identity = 99.98%, n=1,133 sequences), further validating our approach with this genetically identical prevalent undepleted sequencing spike-in artifact.

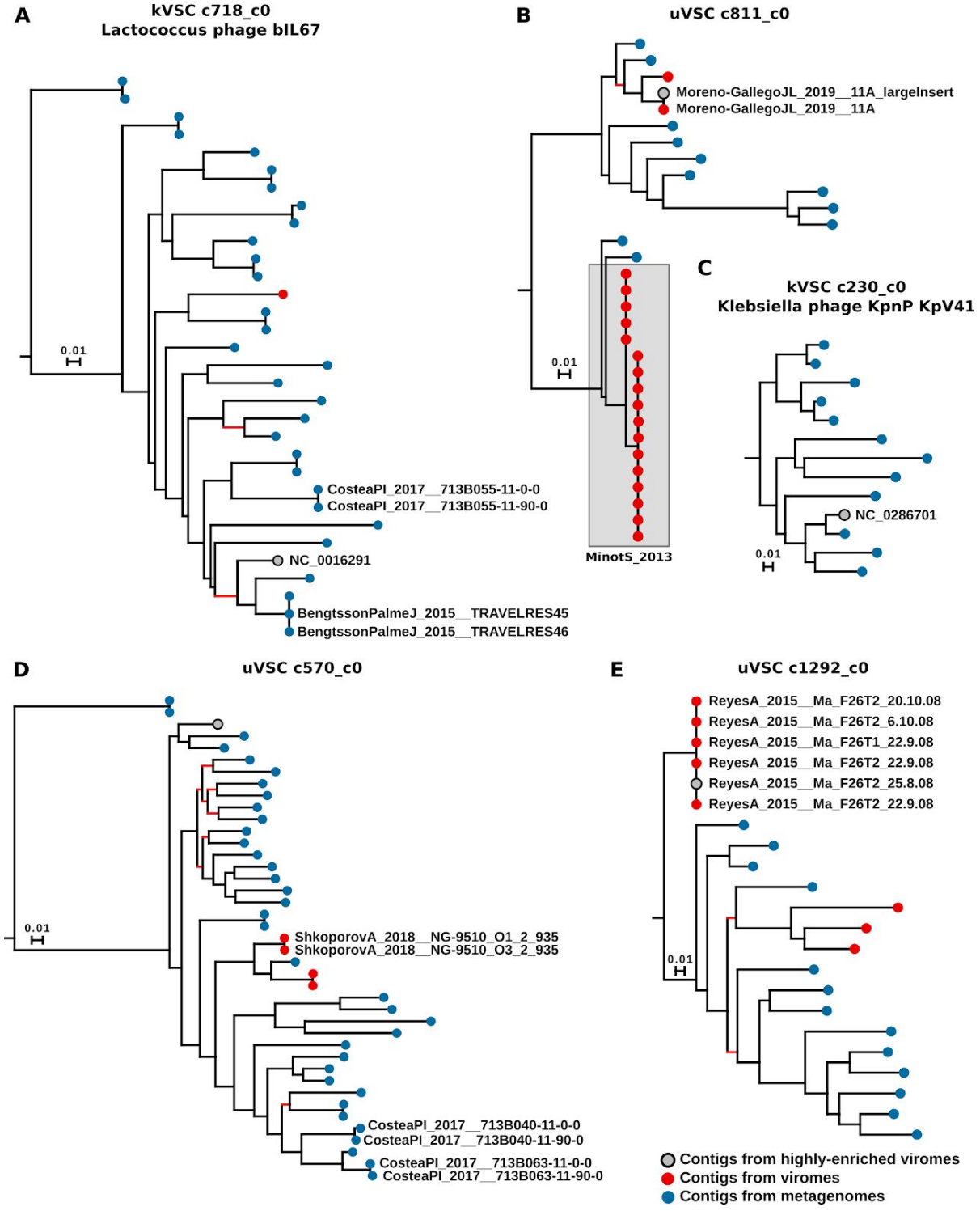

A
kVSC c718_c0
Lactococcus phage bIL67
0.01
CosteaPI_2017__713B055-11-0-0
CosteaPI_2017__713B055-11-90-0
NC_0016291
BengtssonPalmeJ_2015__TRAVELRES45
BengtssonPalmeJ_2015__TRAVELRES46
B
uVSC c811_c0
Moreno-GallegoJL_2019__11A_largeInsert
Moreno-GallegoJL_2019__11A
0.01
MinotS_2013
C
kVSC c230_c0
Klebsiella phage KpnP KpV41
NC_0286701
0.01
D
uVSC c570_c0
0.01
ShkoporovA_2018__NG-9510_O1_2_935
ShkoporovA_2018__NG-9510_O3_2_935
CosteaPI_2017__713B040-11-0-0
CosteaPI_2017__713B040-11-90-0
CosteaPI_2017__713B063-11-0-0
CosteaPI_2017__713B063-11-90-0
E
uVSC c1292_c0
ReyesA_2015__Ma_F26T2_20.10.08
ReyesA_2015__Ma_F26T2_6.10.08
ReyesA_2015__Ma_F26T1_22.9.08
ReyesA_2015__Ma_F26T2_22.9.08
ReyesA_2015__Ma_F26T2_25.8.08
ReyesA_2015__Ma_F26T2_22.9.08
0.01
Contigs from highly-enriched viromes
Contigs from viromes
Contigs from metagenomes

Figure 4.4 (previous page). Phylogenetic trees of five Viral Sequences Clusters. *A subset of the sequences in the same cluster underwent multiple-sequence alignment and a Maximum Likelihood phylogenetic tree was generated. Gray nodes indicate contigs from the original collection of 5,651 contigs from highly enriched samples (HEVCs). Colored nodes indicate sequences selected by similarity to HEVCs from unbinned metagenomes (red nodes) and from low-enrichment viromes (blue nodes). Leaves labels indicate contigs retrieved from the same individuals (**A,B,D**), or from subjects sharing households (**E**). Branches with bootstrap confidence support lower than 75% are highlighted in red. Only sequences within 25% of the median sequence length of the whole cluster (**A,B,C**) or within 15% of the median length of the highly-enriched-sequences in the cluster (**D-E**) were aligned. Alignment lengths are provided in **Supplementary Table 4.5**.*

Identification of prevalent VSCs in the human gut microbiome

To test the prevalence of our collection of VSCs in raw metagenomes, we mapped the 45,872 VSC representatives against 10,000 human gut metagenomes. As this mapping is done considering raw sequencing reads, the approach was expected to extend previous prevalence estimates because the detection of viral homologs is not dependent on successful assembly steps. Expanding on what observed when mapping VSCs against the unbinned fraction of assembled metagenomes (see above), we found extremely abundant clusters across multiple datasets, study cohorts, and countries (**Supplementary Table 4.3, Fig. 4.5**). The most prevalent VSCs in metagenomes and in unbinned contigs were in agreement (**Supplementary Table 4.6**), and seven unknown VSCs were found in more than half of the metagenomes. These seven uVSCs included the three VSCs that were already extremely prevalent in the assembled unbinned metagenomes described above: VSCs c2353, c3062 and c3861. The three clusters had a prevalence of 65.7%, 77.25%, and 78.23%, respectively. The most prevalent known VSCs were *Enterobacteria phage DE3* (prevalence = 27.56%), *crAssphage* (18.81%) and *Enterobacteria phage HK629* (17.24%). Generally, the prevalence of VSCs in raw metagenomes was higher if compared to assembled contigs, highlighting the superior sensitivity of mapping-based approach in detecting VSCs (**Fig. 4.5**, **Supplementary Fig. 4.5**). Of note, the prevalence of *crAssphage* was comparable with the findings of other studies (67).

We could also observe that groups of VSCs were often found together in the same samples. This happened both in the unbinned contigs, where different VSCs were found in the same datasets and with similar prevalences, and in the raw metagenomes, where different VSCs were detected with similar patterns in samples. This could suggest that

sequences of different VSCs belong to the same viral species. A further analysis on the similarities across the sequences in each VSC highlighted the presence of groups of sequences in different clusters that were still similar (>90% similarity, **Supplementary Fig. 4.6 and 4.7**). This could explain why some groups of clusters were always detected together in samples. While this could also be an effect of a group of co-abundant different viruses, this analysis calls for a cautionary approach if the VSCs are intended as individual single-species. This, however, affected only a few groups of clusters, indicating that this effect can be considered negligible.

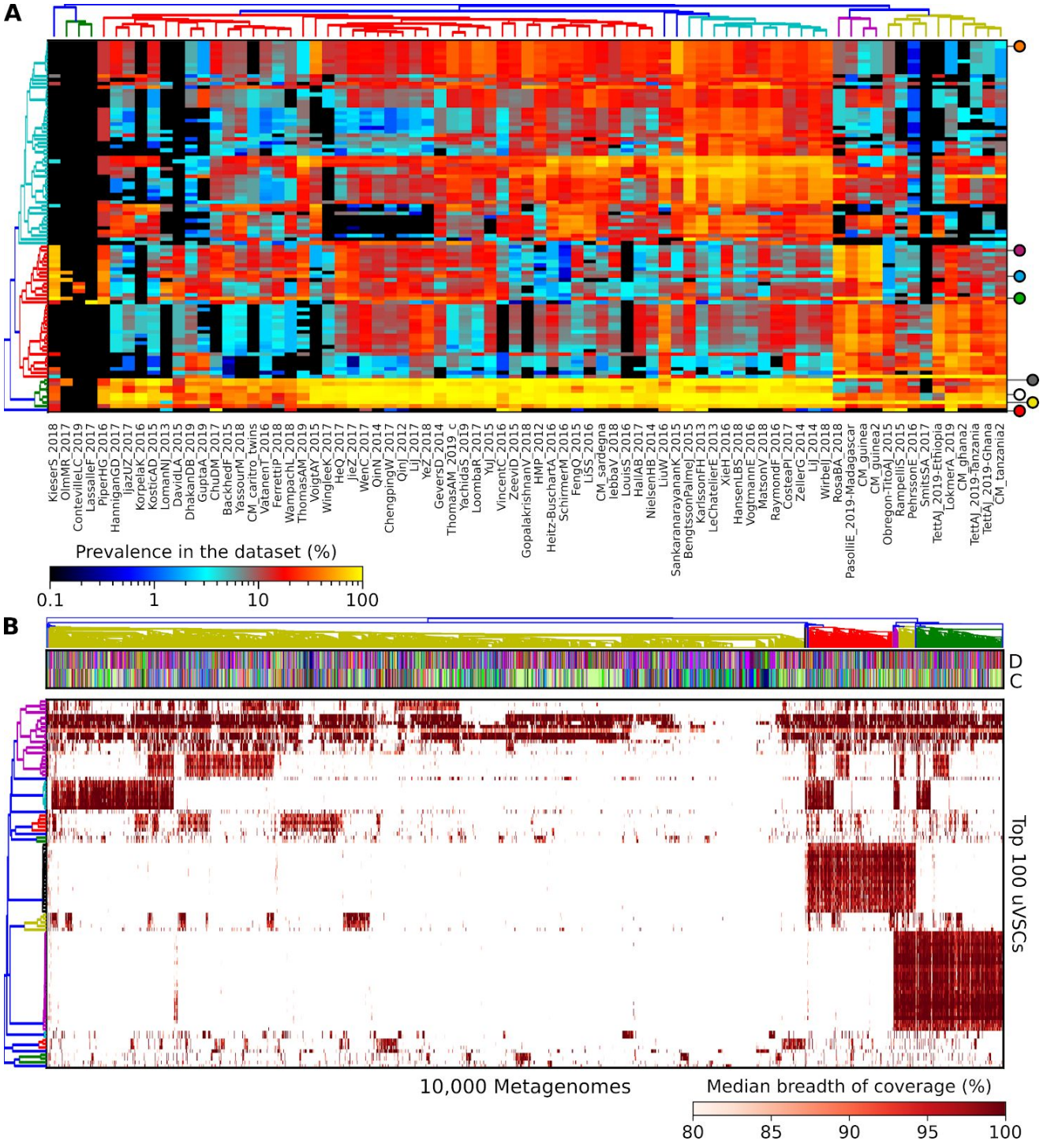

A
KieserS_2018
OlmMR_2017
ContevilleLC_2019
LassalleF_2017
PiperHG_2016
HanniganGD_2017
IjazUZ_2017
KorpelaK_2016
KosticAD_2015
LomanNJ_2013
DavidLA_2015
DhakanDB_2019
GuptaA_2019
ChuDM_2017
BackhedF_2015
YassourM_2018
CM_caritro_twins
VatanenT_2016
FerrettiP_2018
WampachL_2018
ThomasAM_2019
VoigtAY_2015
WingfieldK_2017
HeQ_2017
JieZ_2017
WenC_2017
QinN_2014
ChengpingW_2017
QinJ_2012
LiJ_2017
YeZ_2018
GeversD_2014
ThomasAM_2019_c
YachidaS_2019
LoombaR_2017
YuJ_2015
VincentC_2016
ZeeviD_2015
GopalakrishnanV_2018
HMP_2012
Heitz-BuschartA_2016
SchirmerM_2016
FengQ_2015
LiJS_2016
CM_sardegna
IebbaV_2018
LouisS_2016
HallAB_2017
NielsenHB_2014
LiuW_2016
SankaranarayananK_2015
BengtssonPalmeJ_2015
KarlssonFH_2013
LeChatelierE_2013
XieH_2016
HansenLBS_2018
VogtmannE_2016
MatsonV_2018
RaymondF_2016
CosteaPI_2017
ZellerG_2014
LiJ_2014
WirbelJ_2018
RosaBA_2018
PasolliE_2019-Madagascar
CM_guinea
CM_guinea2
Obregon-TitoAJ_2015
RampelliS_2015
PehrssonE_2016
SmitsSA_2017
TettAJ_2019-Ethiopia
LokmerA_2019
CM_ghana2
TettAJ_2019-Tanzania
TettAJ_2019-Ghana
CM_tanzania2
Prevalence in the dataset (%)
0.1
1
10
100
B
D
C
Top 100 uVSCs
10,000 Metagenomes
Median breadth of coverage (%)
80
85
90
95
100

Figure 4.5 (previous page). Prevalence of the 3,944 VSCs in 10,000 human gut metagenomes. (A) *Prevalence of the most 100 prevalent known and unknown Viral Sequence Clusters (VSCs, rows) in 10,000 samples of 77 datasets (columns). Prevalence is defined as the number of samples where a VSC is detected divided by the number of samples in the dataset. Colored circles on the right side of the heatmap indicate the most prevalent kVSCs. From top to bottom: crAssPhage (orange); Enterobacteria phages SfV (violet); HK629 (blue); DE3 (green); uVSC c3861 (gray); uVSC c3840 (white); uVSC c3062 (yellow); Coliphage phiX174 (red).* ***(B)*** *Breadth of coverage of the 100 most prevalent unknown VSCs in 10,000 metagenomes. For samples where more than one sequence was detected in the same VSC, the median breadth is reported. The minimum breadth of coverage for detection is set to 80% and values below this threshold were not reported. The upper metadata color bar indicates the dataset (D) and country of origin (C) of each sample.*

Conclusions

We presented here a novel strategy that exploits Viral Like Particle (VLP) enriched viromes as a first crucial step in a multiple-step pipeline for the identification and discovery of potentially new viral sequences. We focused on the metagenomic assembly of highly enriched viromes, under the assumption that they would yield more sequences of sure viral origin. We hence took into consideration >3,000 publicly available samples that underwent VLP enrichment, and applied ViromeQC, a tool to estimate the efficiency of VLP enrichment in viromes. We metagenomically assembled the 255 highly enriched viromes with a ViromeQC score greater than 50x. We used the Metagenome-Assembled Genomes (MAGs) from Pasolli *et. al.* and a set of bacterial and viral genomes in NCBI to exclude contigs that were similar to non-viral entities, thereby removing further sources of bacterial contamination. We also used the metagenomic assembled contigs from Pasolli *et al.* that were not binned into any MAG (i.e. the unbinned fraction) to select sequences that were found multiple times in different metagenomes. This led to a collection of 5,651 highly prevalent and potentially viral contigs, spanning 199 highly enriched viromes and 11 human gut datasets. We refer to these contigs as Highly-Enriched Viral Contigs (HEVCs).

Almost 20% of the 5,651 HEVCs were never identified as viral by any of the four viral-prediction tools we tested, and 39.1% were labelled as viral by one tool only (**Fig. 4.2**). This suggests that, even though these sequences are highly prevalent and originate from samples that are enriched in viruses, prediction tools that are based on homology to known viral genomes are unable to identify potentially viral sequences. This is important in the context of viral discovery, as approaches based solely on those viral prediction tools may miss a substantial part of the viral diversity still to be unveiled.

HEVCs were used to retrieve additional viral sequences in contigs from 9,721 metagenomes and the 3,044 viromes originally considered. After multiple rounds of sequence-clustering, we expanded the set of HEVCs with 157,225 homologous contigs, 80.7% of which came from the unbinned assembled fraction of unenriched metagenomes (26). Of those, more than 8,000 sequences belonged to clusters that contained one or more known viral genomes (known VSCs - kVSCs). We could identify several viral clusters that were retrieved in more than 50% of the datasets, with the most prevalent clusters reaching up to 37% and 34% overall prevalence in metagenomes and viromes, respectively.

We were able to reconstruct phylogenies of some of the most prevalent viral clusters, including Lactococcus and Klebsiella phages, as well as the recently characterized *crAssPhage* (58). Moreover, several of the most abundant viral clusters retrieved from

samples of the same individual were placed in the same phylogenetic clade. When the original 13 *crAssPhage* HEVC sequences were used to find homologous contigs in viromes and metagenomes, the 277 newly retrieved sequences formed a separate phylogenetic group that hence expanded the phylogeny.

Several prevalent unknown clusters were extremely conserved, with pairwise sequence identities as high as 99%. These sequences were shared across numerous different studies and individuals and matched against known plasmids. While a viral origin of such elements cannot be excluded, the fact that plasmid sequences may still be present in the highly-enriched viral contigs poses the need for extra caution in the interpretation of the results when such extremely-abundant sequences are found. To overcome this limitation, further research is needed to characterize the nature of the sequences contained in our collection.

These results show that even unenriched samples can be mined to characterize a rich virome in microbiome samples. Indeed, once a novel sequence is retrieved and selected from highly-enriched viromes, it can easily be searched in every metagenome or virome, regardless of its enrichment. Hence, while viral enrichment proves to be extremely useful to guide the discovery of viral genomes (i.e. viral dark matter discovery), studies performed with standard metagenomics can further expand the observed diversity of the viral realm. We released both the full dataset of 162,876 sequences and a collection of 45,872 de-replicated representative sequences for our VSCs at http://segatalab.cibio.unitn.it/data/VDB_Zolfo_et_al.html to enable future studies on the characterization and prevalence of such viral sequences.

Materials and Methods

The proposed pipeline for viral discovery exploits highly enriched viromes that undergo metagenomic assembly to produce potentially viral contigs. These contigs are then screened to remove unwanted residual contaminants (i.e. microbial sequences). Contigs are clustered by similarity and clusters are then extended by retrieving thousands of sequences from assembled metagenomes and viromes. Here we describe in detail the methodological aspects of the pipeline and its application to 3,044 viromes and 9,721 metagenomes to retrieve the 3,944 viral sequences clusters. The pipeline is represented in **Fig. 4.6** and **Supplementary Fig. 4.8.**

Retrieval of metagenomes and viromes used in the analysis

A total of 3,044 viromes were retrieved from NCBI-SRA with SRA-toolkit. This includes the viromes analyzed for viral-enrichment in the scope of a previous work (33) and viromes from 14 additional datasets (**Supplementary Table 4.1**). Samples that were replicated by more than one NCBI-SRA run were downloaded and processed independently. Prior to metagenomic assembly, viromes were preprocessed with Trim Galore, version 0.4.4 (68) to remove low quality (i.e. Phred quality < 20) and short (i.e. read length < 75) reads (parameters: --stringency 5 --length 75 --quality 20 --max_n 2 --trim-n). Reads aligning to the human genome *hg19* were removed by mapping with Bowtie2 version 2.4.1 (69) in "end-to-end" global mode. The preprocessed reads were kept ordered and split in forward, reverse, and singletons FASTQ files, when possible (i.e. for paired-end sequencing libraries). In total, 2.76×10^{10} high-quality reads were retrieved.

Sequences of the metagenome-assembled genomes (MAGs) from Pasolli *et al. (26)* were used to annotate the virome contigs. MAGs were retrieved from http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html, together with the assembled contigs that could not be binned by MetaBat2 (i.e. the unbinned fraction). Five datasets were binned as described in the original study but were not grouped into Species-Level Genome Bins (SGBs). These datasets were also included in the collection (*CM_ethiopia, Heitz-BuschartA_2016, KieserS_2018, RosaBA_2018, and ShiB_2015,* see Supplementary Table 1 of Pasolli *et al.*). Finally, MAGs and unbinned contigs from three additional metagenomic datasets from Tanzania, Ghana, and Madagascar were included (70). In this paper, binned contigs that were not assigned to any SGB are referred to as "lone genomes". Instead, contigs that are neither assigned nor binned into any SGB are referred to as "unbinned contigs" or "unbinned fraction".

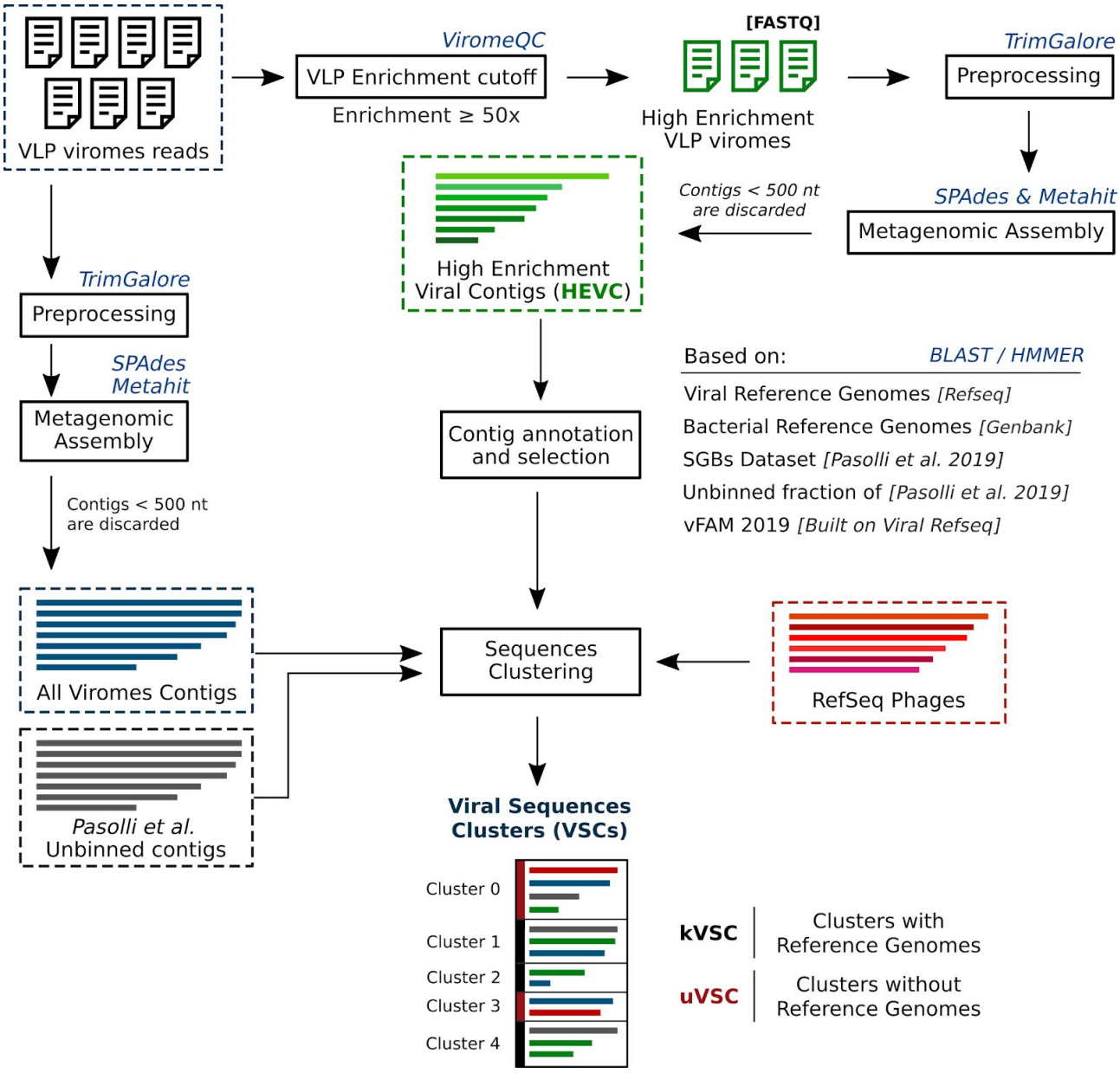


Figure 4.6. Overview of the pipeline for the assembly and identification of viral contigs. *The flowchart represents the steps followed to select for highly enriched viral contigs from viromes and to produce clusters of viral sequences integrating virome and metagenomic contigs. Viromes were preprocessed, assembled, and screened for enrichment with ViromeQC. Contigs from samples with an enrichment score higher than 50x were labelled as Highly Enriched Viral Contigs (HEVCs). HEVCs were screened to reduce further sources of non-viral contamination by mapping against reference bacterial genomes and Metagenomic Assembly Genomes (MAGs). The remaining HEVCs were clustered with a) reference viral genomes and b) sequences similar to HEVCs retrieved from unenriched metagenomes and viromes. This clustering step produced the set of Viral Sequence Clusters (VSCs), that were split into known (kVSCs, if they contained at least one reference viral genome) and unknown (uVSCs, if they did not contain any reference viral genome). A detailed representation of the Sequence Clustering procedure is shown in* ***Supplementary Fig. 4.8.***

Virome Enrichment Estimation

The viral enrichment of the VLP-viromes was assessed with ViromeQC (33), a computational tool that can be applied to raw metagenomic reads to estimate the efficacy of VLP enrichment. ViromeQC calculates the percentage of reads aligning against two sets of markers of sources of microbial contamination: a) the small and large subunits of the 16S/18S and 23S/28S rRNA genes and b) a set of single-copy bacterial markers. ViromeQC then outputs an enrichment score calculated on the abundances of the microbial-markers. A score lower than 1x indicates negative enrichment (i.e. the sample is less enriched than a metagenome). Samples with a ViromeQC score of 50x or higher were considered to be highly enriched (**Table 4.1**). RNA-viromes and samples where RNA and DNA were co-extracted were not considered for Viral-Enrichment estimation.

Metagenomic Assembly

After preprocessing, all the 3,044 viromes underwent metagenomic assembly. Paired-end samples were processed with custom python scripts to preserve the interleaved FASTQ format and were then assembled with metaSPAdes (71) version 3.10.1 (k-mer sizes: -k 21,33,55,77,99,127). Unpaired samples were assembled with Megahit version 1.1.1. (default parameters). Contigs shorter than 500 nucleotides were discarded, and 4.11 x 10^7 contigs were retained.

Contigs Annotation

In order to focus on human-gut associated viruses, only contigs originating from a gut virome (120,041 contigs, see **Supplementary Table 4.2**) were further analyzed. Open Reading Frames (ORFs) were predicted using Prokka (72), version 1.12 with --kingdom Viruses. To determine the similarity of the contigs to bacterial and viral reference genomes, contigs were mapped against a) all the reference viral genomes in RefSeq, release 91 (38); b) a set of 80,853 complete and draft bacterial genomes from NCBI GenBank (26, 73); c) the MAGs and from Pasolli *et al.* All mappings were performed with BLAST, version 2.6.0 (55) in nucleotides space and with the default parameters except for -max_target_seqs 1000. BLAST Databases were built with custom scripts in order to allow for better hits-tracking in the downstream analysis. Only hits with an alignment length >= 1000 nucleotides and a percentage of identity >= 80% were considered. We also computed the total

breadth-of-coverage of each contig against each database (i.e. the bacterial genomes, viral genomes, and SGBs), by calculating how many nucleotides of the contig were covered by at least one blast hit in the database. Values for each contig are reported in **Supplementary Table 4.2**.

Contigs were also mapped against the unbinned fraction of Pasolli *et al.* to identify potentially viral sequences in the residual unbinned fraction of binned MAGs. In this case, we set -max_target_seqs 500000 to maximize the number of hits to unbinned contigs. Only hits with length >= 1000 nucleotides and identity >= 80% were considered.

We calculated the proportion of ORFs that matched viral proteins in each contig, ORFs were mapped with hmmscan (74) against a database built with the V-Fam code provided by Skewes-Cox *et al.* (63). We used all the proteins from Viral Refseq (release 91). An ORF was considered to match a viral protein domain if it had at least one full-domain hit with an e-value lower than 10^{-5}.

To determine whether a contig was present in more than one sample or dataset, we ran an all-vs-all BLAST search on the 120,041 sequences. To be detected in another sample, one contig needed to have an overlap of at least 1000 nucleotides with at least 80% identity. The number of hits to other viromes is reported in **Supplementary Table 4.2** and is referred only to the 255 highly enriched viromes shown in **Table 4.1**.

Retrieval of 2,619 bacteriophages from RefSeq

We downloaded all the viral genomes from RefSeq (38) (Release 99; n=12,182) and selected the phages in two steps. First, we used the information present in the sequence ID to retrieve all the sequences containing the keyword “phage” (ignoring case distinctions) in the header (n=2,543); then, we relied on the NCBI taxonomy (taxid = 10239) to refine our selection, adding 76 more sequences to our set.

Selection of the 5,651 High-Enrichment Viral Contigs

To further remove contigs that were likely to be part of bacterial or archaeal genomes, we used the Pasolli *et al.* MAGs and metagenomes dataset to discard sequences that a) were found binned in the same Species-level Genome Bin (SGB) in more than 30 metagenomes; or b) were found binned in any SGB in more than 50 metagenomes. Finally, only contigs longer than 1,500 nucleotides and that were found in the unbinned fraction in more than 20 metagenomes were kept. Of the initial 120,041 contigs, 4,952 met the selection criteria. These contigs were added to the viral reference genomes of bacteriophages that could be

found in at least 20 metagenomes (n=699). In total, 5,651 sequences were then considered as “High-Enrichment Viral Contigs” (HEVC).

Analysis of the reconstructed contigs with viral-prediction software

Viral prediction tools were used to classify the retrieved contigs as viral or non-viral. Specifically, VirSorter (19) was run using diamond (75) as mapping tool (--diamond) with the RefSeq database (--db 1), both in standard and decontamination mode (virsorter --virome). VirFinder (18) was run in R 3.6.3 using the standard prediction model. ViralVerify was run with the provided database of virus-chromosome-specific HMMs (20). Seeker (21) was run with the default parameters.

Clustering of the 5,651 High-Enrichment Viral Contigs and clusters extension

The 5,651 sequences from highly enriched viral contigs and reference genomes were clustered into 3,944 clusters at 90% identity with VSearch (56), version 2.14.2 (parameters --cluster_fast --id 0.9 --strand both --maxseqlength 200000). Clusters that contained at least one RefSeq viral genome were labelled as known viral sequence clusters (kVSC); clusters that only contained novel sequences were labelled as unknown (uVSC).

Clusters were extended with contigs from viromes and metagenomes. We took contigs from the full dataset of 3,044 assembled viromes (i.e. highly enriched and non-highly enriched viromes, **Supplementary Table 4.1**) and from the unbinned fraction of metagenomes analyzed in Pasolli *et al.* (see above). First, these contigs were mapped against the 5,651 HEVCs with BLAST (55). Then, sequences with a percentage of identity of at least 80% over at least 1000 nucleotides to an HEVC were kept (69,484 contigs similar to an HEVC from viromes, 355,825 from unbinned metagenomes). The 425,309 contigs were used to build a sketch-database with Mash (59) version 2.0 (command: mash sketch -i -s 10000). Next, each VSC centroid (i.e. the longest sequence within each of the 3,944 initial clusters) was mapped against the initial set of contigs using Mash (command: mash dist -d 0.1 -v 0.05), and contigs with a distance lower than 10% (p-value <= 0.05) were assigned to the closest VSC cluster (i.e. the VSC with the minimum mash distance), hence extending clusters with new sequences and producing extended-HSVCs.

Extended-HSVCs were then further re-clustered at 70% identity with VSearch, and only clusters that contained at least one of the original 5,651 sequences were kept. Clusters that had more than one valid sub-cluster were kept separated. For example, cluster *vsearch_c1003* contained five sub clusters, two of which contained sequences from the

starting 5,651 HEVCs. Hence, two sub-clusters of *vsearch_c1003* (c1003_c0 and c1003_c3, see **Supplementary Table 4.3**) were kept. This step produced the final 4,077 extended-clusters.

Cluster Representatives selection and Prevalence analysis

We built a collection of representative sequences for each of our 3,944 Viral Sequence Clusters (VSCs) with the goal to select sequences that maximized the diversity within each cluster. Hence, we re-clustered all the sequences within each VSC at 95% identity with VSearch (parameters: --cluster_fast --id 0.95 --strand both). We then selected as final-representatives: a) the centroids of the 95% clustering; b) all the original 5,651 contigs from highly-enriched samples and reference genomes. In total, 47,820 clusters-representative sequences were selected (median length: 6,355 bp; maximum length: 240 kbp). Sequences were de-replicated at 99% identity and 90% overlap with CD-HIT (60) version 4.6.8 (parameters -n 10 -d 0 -c 0.99 -aL 0.9). The final collection of 45,872 sequences is publicly available together with clusters metadata (see **Data Availability**).

VSCs prevalence in raw metagenomes

To calculate the prevalence of each VSC in metagenomes, we mapped the raw reads of 10,000 publicly available metagenomes against the representative sequences of each VSC. Mapping was performed with Bowtie2 (69), version 2.4.1 in end-to-end global mode. Alignments were converted into BAM files with Samtools (76) version 1.3.1. Breadth and depth of coverage for each cluster were calculated with Bedtools (77) version 2.29.1. To compute the breadth of coverage of each sequence, we divided the number of nucleotides with a coverage >=3x by the length of the sequence. The overall breadth of coverage of each VSC was defined as the median breadth of coverage of all the sequences belonging to that cluster. The minimum breadth of coverage to consider a cluster as detected was set to 80%.

Phylogenetic analysis, Data Analysis and Visualization

To build phylogenetic trees, we selected the sequences with a length within 25% from the median length of all the sequences in each cluster. Sequences belonging to the original 5,651 HEVCs were kept regardless of their length, unless differently specified.

Multiple-sequence-alignments for elements of each Viral Sequences Cluster were performed with MAFFT (78), version 7.453 with automatic parameters selection (--auto).

Alignments were trimmed with Trimal (79), version 1.2rev59 with the -gappyout option. Phylogenetic trees were computed using RAxML (80) version 8.1.15 with the GTRGAMMA model (parameters: -p 48315 -x 48315 -# 100 -m GTRGAMMA -f a). We performed a manual curation and topology-checking of the resulting phylogenetic trees, branch lengths, and bootstrap values. In some cases we also took into consideration the corresponding alignment, to identify possible sources of non-phylogenetic signal. Trees with very low bootstrap values, very long and isolated branches, and/or ambiguous alignments were discarded.

Pairwise sequence identities of each viral sequence cluster were calculated on the multiple-sequence alignments. To ensure that only sequences of similar lengths were aligned, only sequences within 25% of the median sequence length of the highly-enriched contigs in the cluster were considered. Sequence identity of two sequences was defined as 1-Hamming_Distance (seq1,seq2), without considering gaps.

Hierarchical clustered heatmaps were generated with hclust2 (81) using Bray-Curtis distance to cluster features (VSCs) and correlation distance to cluster samples, unless otherwise specified. Average linkage was used in the hierarchical clustering. Plots and figures were drawn with matplotlib (82) and Seaborn. Upset plots (83) were drawn with the upsetplot python package, version 0.4. Bioinformatic analysis on sequences was performed in Python3 with packages BioPython (84), Pandas (85) version 1.0.1, and NumPy (86) version 1.18.1. Computation was performed on the High-Performance Computing infrastructure of the Segata Laboratory at the University of Trento, Italy.

Code and Data Availability

The complete code to reproduce the steps described in this section is available on GitHub at https://github.com/SegataLab/viromedb. The viromes used in this study are described in (33). Additionally, viromes from 14 other studies were used (**Supplementary Table 4.1**). The MAGs used to annotate viral contigs are available in Pasolli *et al.* and at http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html. The metagenomes used to extend viral sequences clusters are referred to in the same publication. The representative sequences of all the VSCs are available at http://segatalab.cibio.unitn.it/data/VDB_Zolfo_et_al.html.

Supplementary Tables

Supplementary Tables are available at:
http://segatalab.cibio.unitn.it/data/VDB_Zolfo_et_al.html. Captions are reported below.

Supplementary Table 4.1. Metadata and references for the 3,044 analyzed viromes. The table contains metadata and references to the viromes from 49 datasets that were analyzed in this paper. References and sample counts aggregated by dataset are presented in (**Tab 1**). Detailed statistics for each virome are shown in (**Tab 2**).

Supplementary Table 4.2. Annotation of the 120,041 contigs reconstructed from the metagenomic assembly of 255 highly enriched gut viromes. Each contig was annotated against 80,853 bacterial reference genomes, as well as the MAGs from Pasolli *et al*. 2019, and the viral reference genomes in RefSeq.

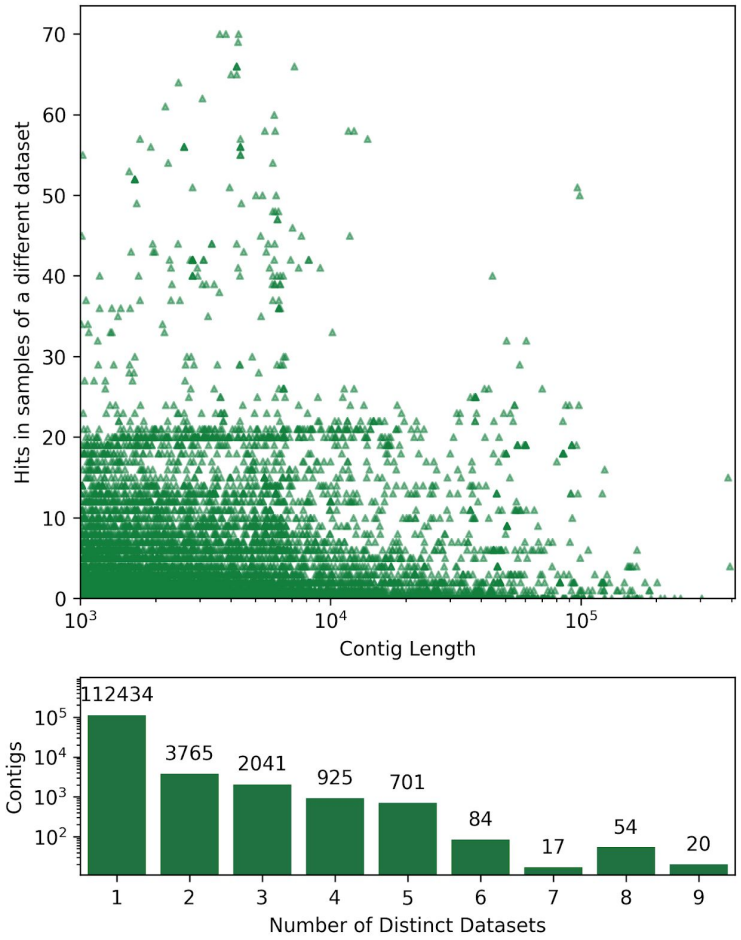
Supplementary Table 4.3. Selection of the 5,651 contigs from highly enriched viromes (HEVCs). Tab 1 contains the full set of 5,651 contigs that were selected from gut viromes according to their prevalence and low contamination. Detailed information on the clusters in which each contig is available is reported in the first six columns. The results of viral-classification tools applied to each contig are also included. **Tab 2** contains grouped statistics for each of the 3,944 clusters of sequences (see **Methods**).

Supplementary Table 4.4. Prevalence of each VSC in the surveyed metagenomes and viromes. Prevalence of each Viral Sequence Cluster in the analyzed metagenomes (**Tab 1**), viromes (**Tab 2**), and all-samples (**Tab 3**). Prevalence is defined as the percentage of samples in the dataset from which the cluster could be retrieved. Only samples where at least one VSC could be retrieved are included. The “percentage of datasets” column (4th and 10th columns of **Tabs 1-3**), contains the percentage of datasets in which the VSC could be retrieved from at least one sample. **Tab 4** contains the number of distinct samples in each dataset from which a VSC could be retrieved.

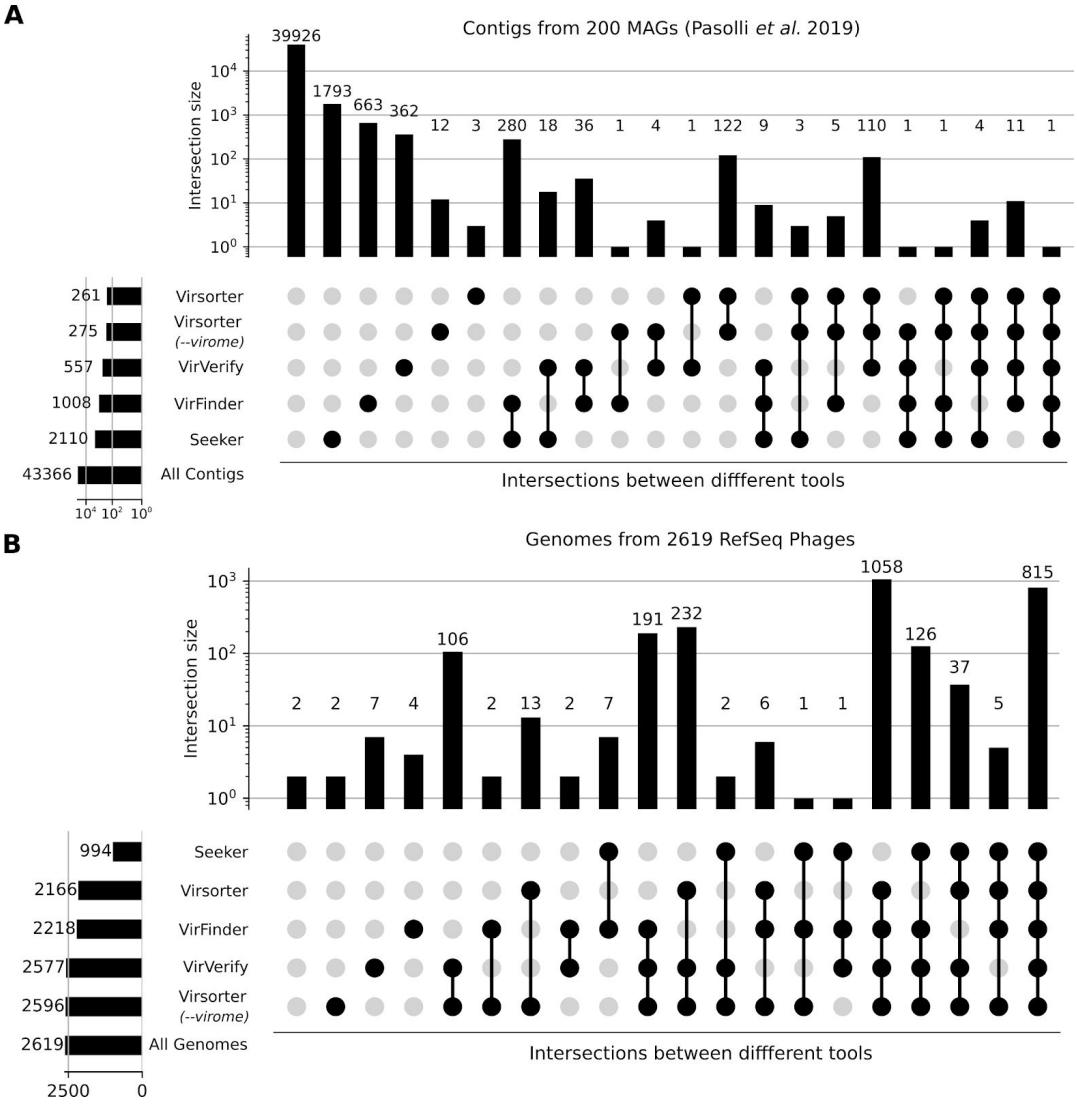
Supplementary Table 4.5. Multiple-Sequence Alignments Statistics. Phylogenetic trees were built from multiple-sequence alignments filtered to maximise the phylogenetic signal (see **Methods**). The table recapitulates the number of sequences in each tree and the lengths of sequences before and after the alignment. The selection strategy indicates how the sequences of each cluster were selected. Sequences were kept if they had a length within 25% of the median of the lengths in the cluster (strategy “A”), or within the 15% of the median of the sequences from highly-enriched contigs (strategy “B”).

Supplementary Table 4.6. Prevalence of the VSCs in 10,000 metagenomes. A selection of 10,000 human gut metagenomes from 77 datasets was mapped against the VSCs representatives to estimate the prevalence of each cluster in each dataset. Aggregated values are reported in **Tab 1**. Datasets names and references are reported in **Tab 2**. The raw data reporting the prevalence of each VSC in each dataset is reported in **Tab 3**.

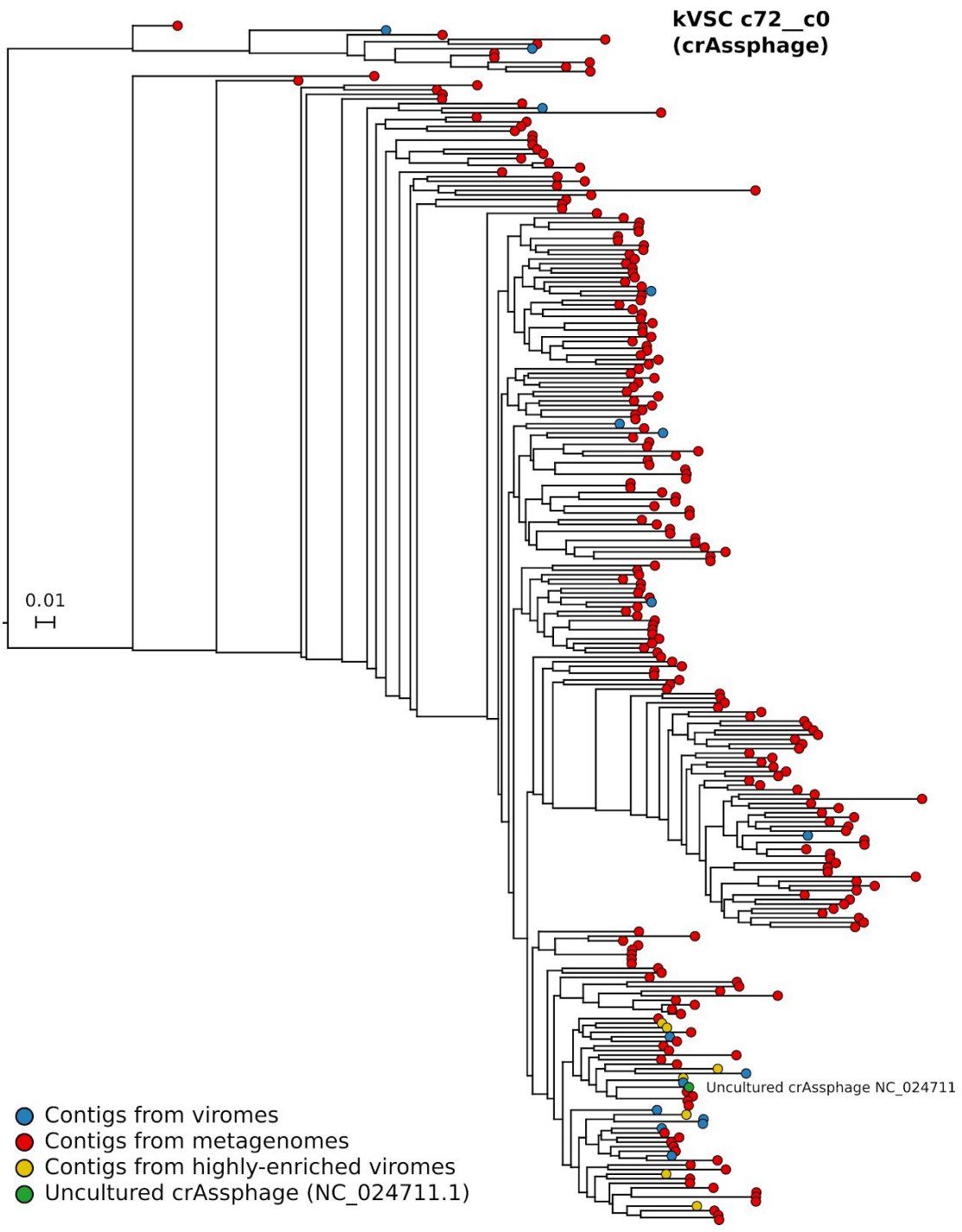
Supplementary Figures



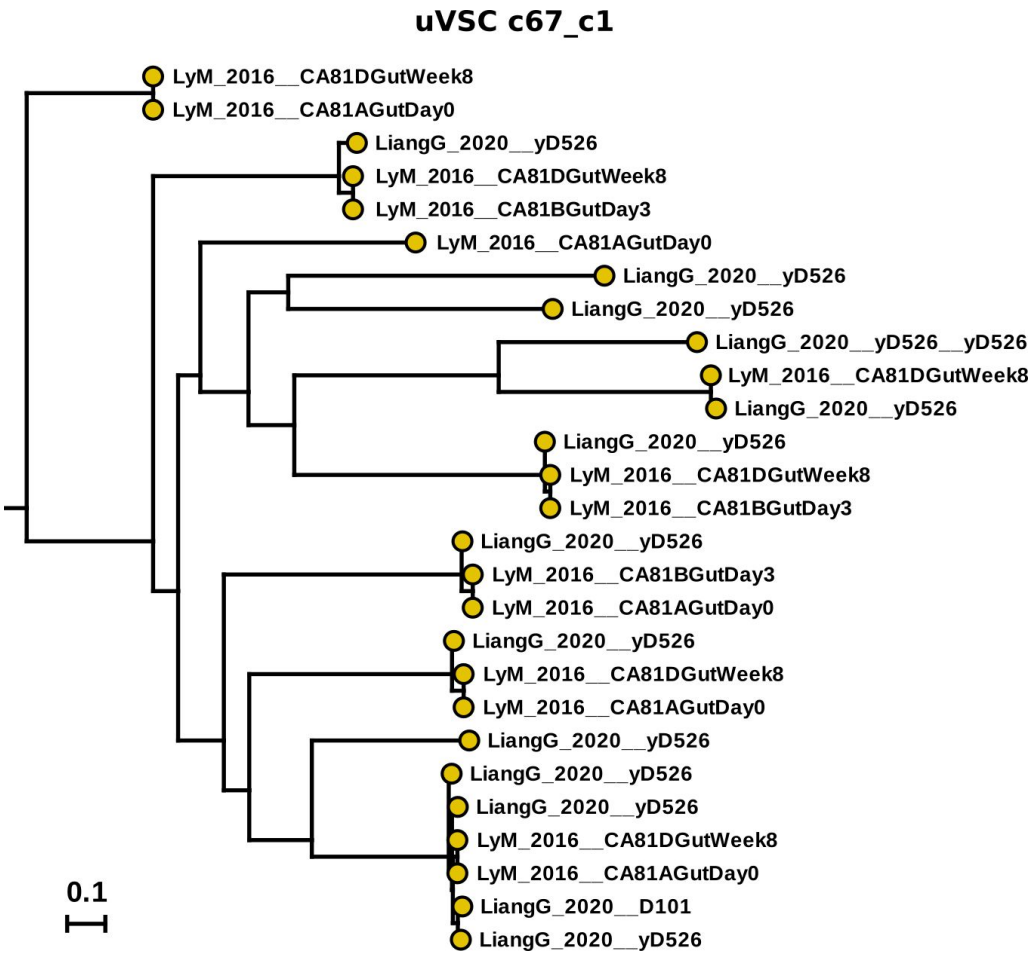
Supplementary Figure 4.1. High Enrichment Contigs prevalence across the 11 considered virome datasets. *The scatter shows each of the 120,041 contigs assembled from highly enriched gut viromes by length (x-axis) and the number of hits in samples of other datasets (y-axis). The bottom bar plot indicates the number of distinct datasets each contig is detected in. Detection of a contig in multiple samples was based on BLAST hits longer than 1000 nucleotides and with a percentage of identity of 80% or higher (see* ***Methods****).*



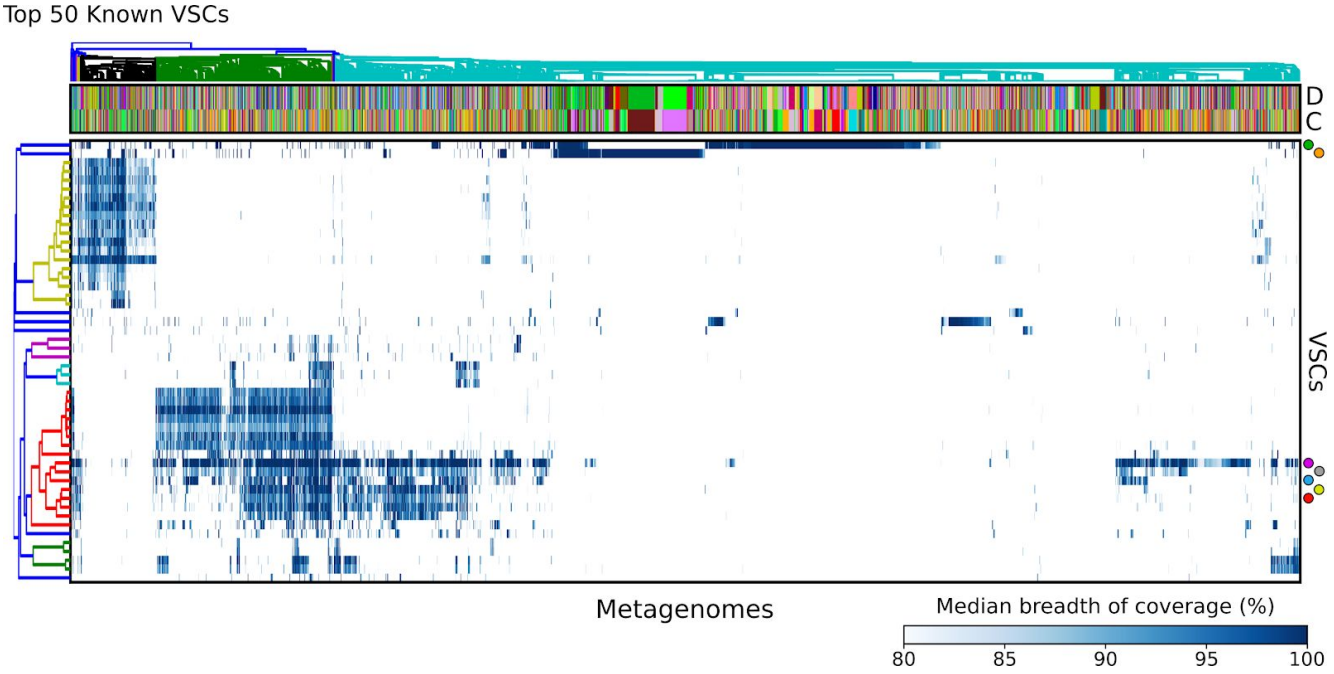
Supplementary Figure 4.2. Comparison of different Viral Detection tools on Viral and Bacterial Genomes. *Four viral-detection tools were applied to a set of contigs from medium and high-quality MAGs (**A**) and to 2,619 reference genomes of bacteriophages (**B**). The plot represents the cardinality of each intersection set. 'All genomes' and 'All contigs' refer to the full set of analyzed contigs and genomes.*



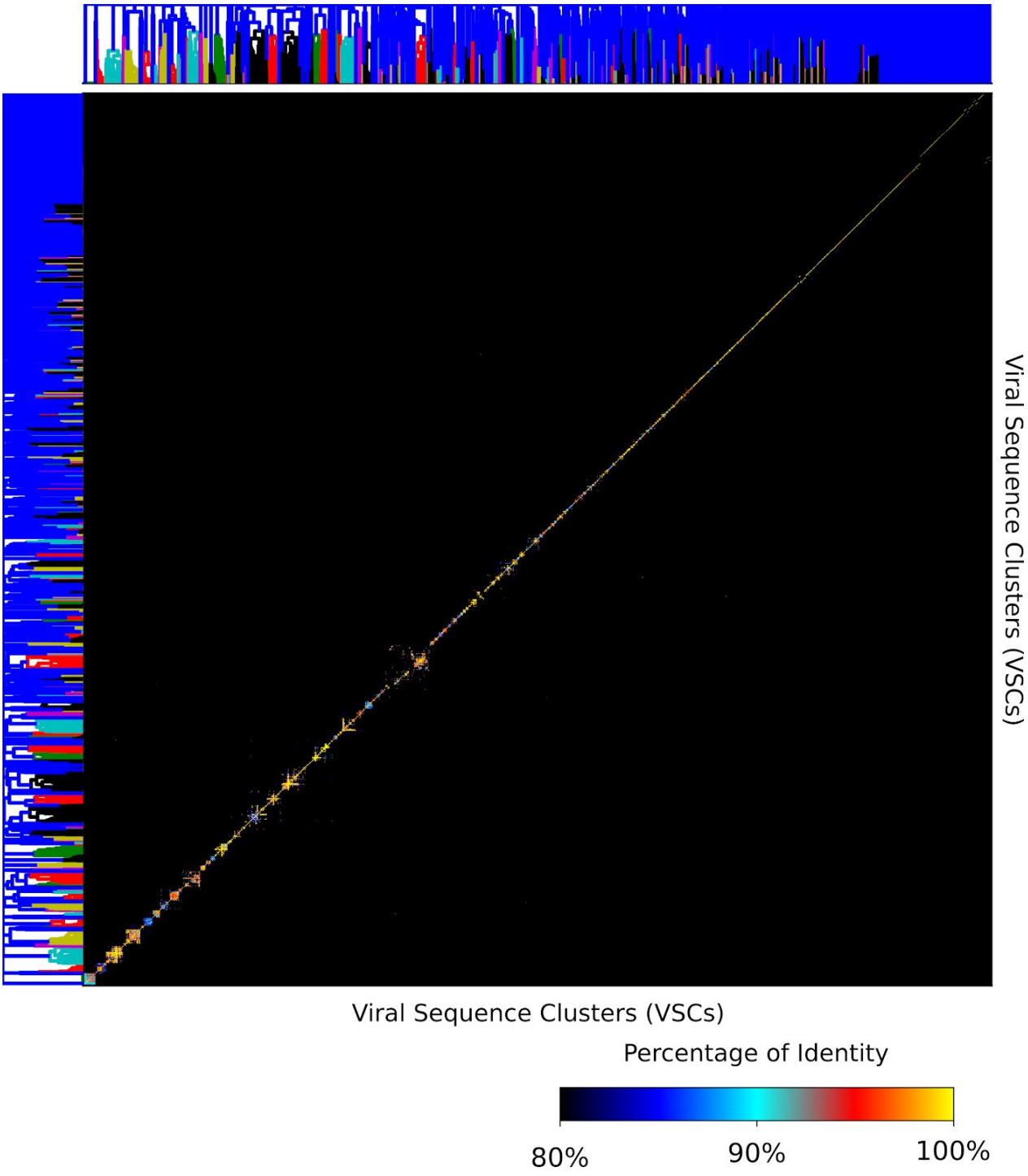
Supplementary Figure 4.3. Phylogenetic tree of crAssPhage-like sequences (cluster c72__c0). *The 577 sequences in cluster c72__c0 were selected to build a phylogenetic tree. All sequences with a length outside 25% of the median length of all sequences in the cluster were excluded from the multiple-sequences alignment. The remaining 262 contigs were reconstructed from highly enriched viral contigs. Leaves are colored according to the origin of each sequence. Alignment statistics are provided in* ***Supplementary Table 4.5.***



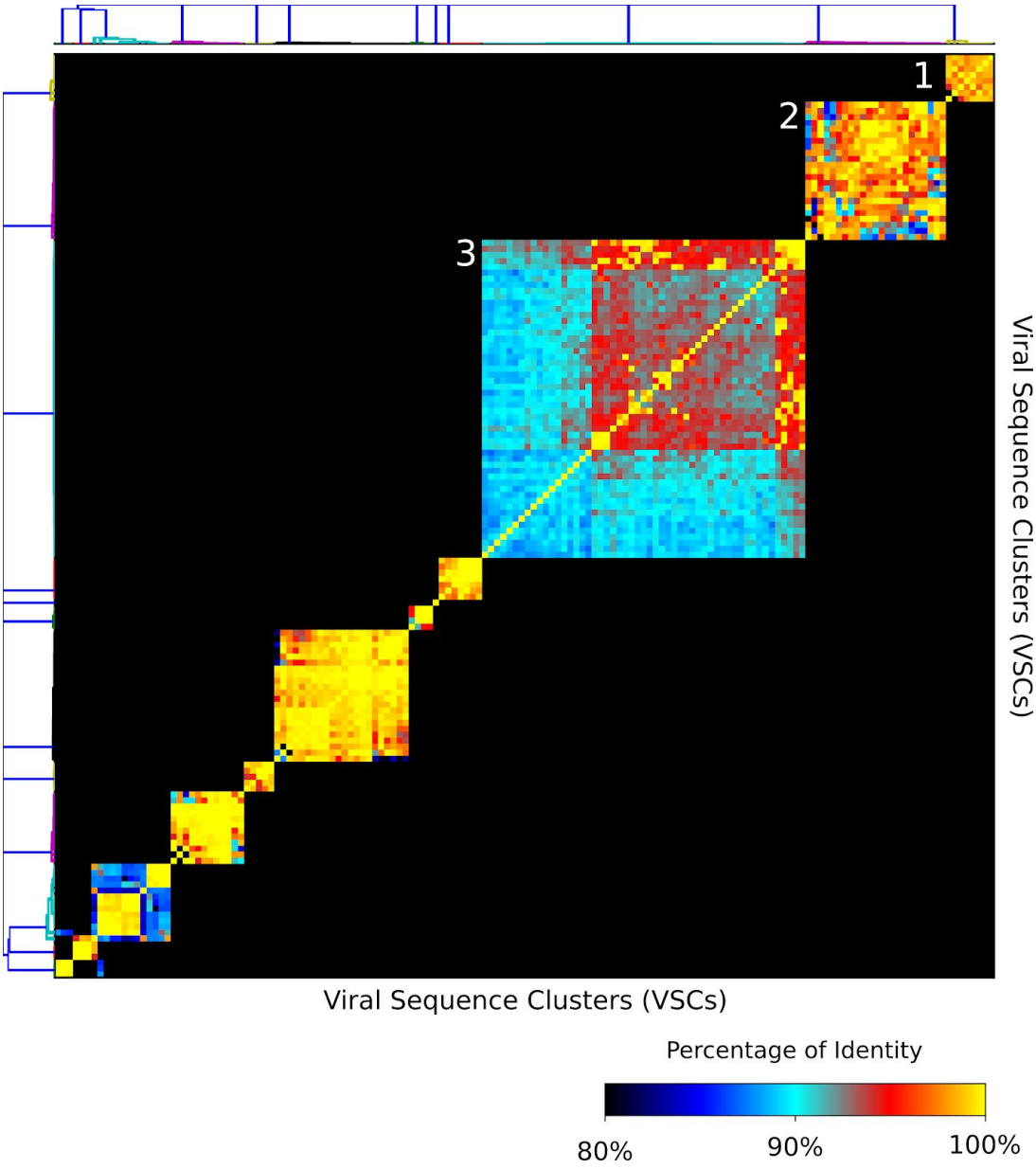
Supplementary Figure 4.4. Phylogenetic tree of unknown viral sequence cluster c67__c1. *The 123 sequences in cluster c67__c1 were selected to build a phylogenetic tree of cluster c67__c1. Only sequences with a length within 15% of the median length of highly-enriched contigs in the cluster were kept for multiple-sequences alignment. All the 27 remaining sequences belonged to contigs reconstructed from highly enriched viral contigs. Median sequence length = 3,882 bp ; Trimmed alignment length = 18,527 bp. Study ID and sample names are indicated in the label of each leaf. Alignment statistics are provided in* ***Supplementary Table 5.***



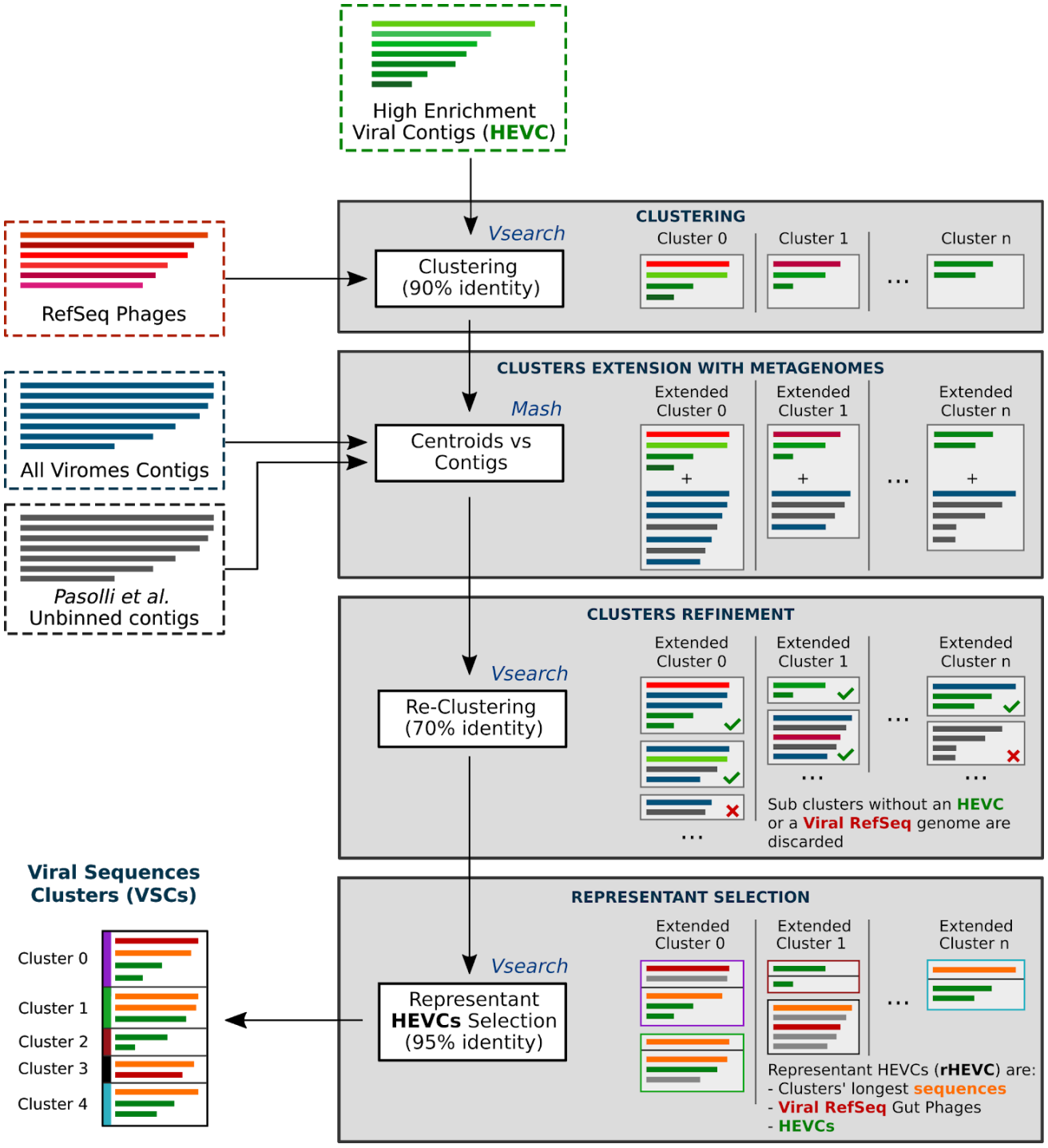
Supplementary Figure 4.5. Raw-reads mapping of kVSCs in 10,000 samples. *Breadth of coverage of the 50 most prevalent known VSCs in 10,000 metagenomes. For samples where more than one sequence was detected within the same VSC, the median breadth of coverage is reported. The minimum breadth of coverage for detection was set to 80%. Colored circles on the y-axis label the most prevalent known VSCs. From top to bottom: crAssphage (green), phiX174 (orange), Enterobacteria phages DE3 (purple), HK629 (grey), mEp460 (blue), and Sfv (yellow), Shigella phage SfIV (red).*



Supplementary Figure 4.6. VSCs cluster similarities. *Hierarchical Clustering of all-vs-all similarities between representative sequences of each Viral Sequence Cluster. Rows and columns of the heatmap are hierarchically clustered with hclust2 (average linkage, correlation distances). Values refer to the maximum BLAST percentage of identity between clusters (minimum percentage of identity: 80% over at least 1000 nucleotides).*



Supplementary Figure 4.7. VSCs cluster similarities (subset). *Hierarchical Clustering of all-vs-all similarities between representative sequences of each Viral Sequence Cluster. Rows and columns of the heatmap are hierarchically clustered with hclust2 (average linkage, correlation distances). Values refer to the maximum BLAST percentage of identity between clusters (minimum percentage of identity: 80% over at least 1000 nucleotides). The heatmap is a subset of the clusters with more than 50 hits against another cluster among those shown in* ***Supplementary Figure 5****. Group 1 contains known VSCs c184, c405, and c388 (Enterobacteria phages HK629 and HK544 and Bacteriophage HK022). Group 2 contains known VSCs c245, c273, and c278 (Staphylococcus aureus phages TEM123, 88 and phiNM2). Group 3 contains known VSCs c643, c620, and c645 (Propionibacterium phages P11, P100D, and Enoki). The other groups are unknown SVCs.*



Supplementary Figure 4.8. Detailed representation of the clustering procedure. *Contigs from highly enriched viromes (HEVCs, in green) were clustered together with gut bacteriophages reference genomes (in red). Clusters were extended by adding similar sequences from viromes (blue contigs) and unbinned metagenomes (gray contigs). After a second step of clustering at 70% identity, only clusters that contained at least one of the original sequences were kept. Finally, a third clustering step was performed to select the representative sequences of each cluster (Viral Sequences Clusters).*

References

1. W. B. Whitman, D. C. Coleman, W. J. Wiebe, Perspective Prokaryotes: The unseen majority 95, no (1998).

2. C. A. Suttle, Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801 (2007).

3. M. Brown-Jaque, M. Muniesa, F. Navarro, Bacteriophages in clinical samples can interfere with microbiological diagnostic tools. *Sci. Rep.* **6**, 33000 (2016).

4. M. Breitbart, *et al.*, Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**, 6220–6223 (2003).

5. A. Reyes, *et al.*, Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).

6. X. Wang, *et al.*, Cryptic prophages help bacteria cope with adverse environments. *Nat. Commun.* **1**, 147 (2010).

7. F. Rohwer, R. V. Thurber, Viruses manipulate the marine environment. *Nature* **459**, 207–212 (2009).

8. F. Rodriguez-Valera, *et al.*, Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* **7**, 828–836 (2009).

9. J. M. Norman, *et al.*, Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447–460 (2015).

10. G. Liang, *et al.*, The dynamics of the stool virome in very early onset inflammatory bowel disease. *J. Crohns. Colitis* (2020) https:/doi.org/10.1093/ecco-jcc/jjaa094.

11. A. Reyes, *et al.*, Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11941–11946 (2015).

12. G. Nakatsu, *et al.*, Alterations in Enteric Virome Are Associated With Colorectal Cancer and Survival Outcomes. *Gastroenterology* **155**, 529–541.e5 (2018).

13. G. Zhao, *et al.*, Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proceedings of the National Academy of Sciences*, 201706359 (2017).

14. P. Manrique, *et al.*, Healthy human gut phageome. *Proceedings of the National Academy of Sciences* **113**, 201601060 (2016).

15. J. C. Venter, *et al.*, Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).

16. Human Microbiome Project Consortium, A framework for human microbiome research. *Nature* **486**, 215–221 (2012).

17. C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, N. Segata, Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

18. J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, F. Sun, VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).

19. S. Roux, F. Enault, B. L. Hurwitz, M. B. Sullivan, VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).

20. D. Antipov, M. Raiko, A. Lapidus, P. A. Pevzner, metaviralSPAdes: assembly of viruses from

metagenomic data. *Bioinformatics* (2020) https:/doi.org/10.1093/bioinformatics/btaa490.

21. N. Auslander, A. B. Gussow, S. Benler, Y. I. Wolf, E. V. Koonin, Seeker: Alignment-free identification of bacteriophage genomes by deep learning. *bioRxiv*, 2020.04.04.025783 (2020).

22. M. J. Tisza, *et al.*, Discovery of several thousand highly diverse circular DNA viruses. *Elife* **9** (2020).

23. N. A. O'Leary, *et al.*, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).

24. D. H. Parks, *et al.*, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* (2017) https:/doi.org/10.1038/s41564-017-0012-7.

25. A. Almeida, *et al.*, A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).

26. E. Pasolli, *et al.*, Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019).

27. S. Nayfach, Z. J. Shi, R. Seshadri, K. S. Pollard, N. C. Kyrpides, New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).

28. D. D. Kang, J. Froula, R. Egan, Z. Wang, MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

29. J. Alneberg, *et al.*, Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).

30. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

31. R. V. Thurber, M. Haynes, M. Breitbart, L. Wegley, F. Rohwer, Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).

32. Y.-Z. Zhang, M. Shi, E. C. Holmes, Using Metagenomics to Characterize an Expanding Virosphere. *Cell* **172**, 1168–1172 (2018).

33. M. Zolfo, *et al.*, Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.* **37**, 1408–1412 (2019).

34. A. Bankevich, *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

35. D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

36. T. D. S. Sutton, A. G. Clooney, F. J. Ryan, R. P. Ross, C. Hill, Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **7**, 12 (2019).

37. S. Minot, *et al.*, Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12450–12455 (2013).

38. J. R. Brister, D. Ako-Adjei, Y. Bao, O. Blinkova, NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–7 (2015).

39. S. Manara, *et al.*, Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species. *Genome Biol.* **20**, 299

(2019).

40. G. Liang, *et al.*, The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* (2020) https:/doi.org/10.1038/s41586-020-2192-1.

41. M. Ly, *et al.*, Transmission of viruses via our microbiomes. *Microbiome* **4**, 64 (2016).

42. E. S. Lim, *et al.*, Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* **21**, 1228–1234 (2015).

43. D. Aguirre de Carcer, A. Lopez-Bueno, D. A. Pearce, A. Alcami, Biodiversity and distribution of polar freshwater DNA viruses. *Science Advances* **1**, e1400127–e1400127 (2015).

44. K. Arkhipova, *et al.*, Temporal dynamics of uncultured viruses: A new dimension in viral diversity. *ISME J.* **12**, 199–211 (2018).

45. P. S. Pannaraj, *et al.*, Shared and Distinct Features of Human Milk and Infant Stool Viromes. *Front. Microbiol.* **9**, 1162 (2018).

46. S. Roux, *et al.*, Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).

47. J. R. Brum, *et al.*, Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).

48. J. L. Moreno-Gallego, *et al.*, Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins. *Cell Host Microbe* **25**, 261–272.e5 (2019).

49. K. Rosario, C. Nilsson, Y. W. Lim, Y. Ruan, M. Breitbart, Metagenomic analysis of viruses in reclaimed water. *Environ. Microbiol.* **11**, 2806–2820 (2009).

50. D. A. de Cárcer, A. López-Bueno, J. M. Alonso-Lobo, A. Quesada, A. Alcamí, Metagenomic analysis of lacustrine viral diversity along a latitudinal transect of the Antarctic Peninsula. *FEMS Microbiol. Ecol.* **92**, 1–10 (2016).

51. M.-S. Kim, E.-J. Park, S. W. Roh, J.-W. Bae, Diversity and abundance of single-stranded DNA viruses in human feces. *Appl. Environ. Microbiol.* **77**, 8062–8070 (2011).

52. S. Minot, *et al.*, The human gut virome : Inter-individual variation and dynamic response to diet The human gut virome : Inter-individual variation and dynamic response to diet. *Genome Res.*, 1616–1625 (2011).

53. S. Roux, *et al.*, Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7** (2012).

54. A. Lopez-Bueno, *et al.*, High Diversity of the Viral Community from an Antarctic Lake. *Science* **326**, 858–861 (2009).

55. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

56. T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

57. S. Mukherjee, M. Huntemann, N. Ivanova, N. C. Kyrpides, A. Pati, Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand. Genomic Sci.* **10**, 18 (2015).

58. B. E. Dutilh, *et al.*, A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).

59. B. D. Ondov, *et al.*, Mash: fast genome and metagenome distance estimation using MinHash.

Genome Biol. **17**, 132 (2016).

60. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

61. P. I. Costea, *et al.*, Subspecies in the global human gut microbiome. *Mol. Syst. Biol.* **13**, 960 (2017).

62. A. N. Shkoporov, *et al.*, Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* **6**, 68 (2018).

63. P. Skewes-Cox, T. J. Sharpton, K. S. Pollard, J. L. DeRisi, Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS One* **9**, e105067 (2014).

64. P. Ferretti, *et al.*, Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* **24**, 133–145.e5 (2018).

65. T. S. Schmidt, *et al.*, Extensive transmission of microbes along the gastrointestinal tract. *Elife* **8** (2019).

66. I. L. Brito, *et al.*, Transmission of human-associated microbiota along family and social networks. *Nat Microbiol* **4**, 964–971 (2019).

67. T. P. Honap, *et al.*, Biogeographic study of human gut-associated crAssphage suggests impacts from industrialization and recent expansion. *PLOS ONE* **15**, e0226930 (2020).

68. F. Krueger, Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* (2015).

69. B. Langmead, S. L. Salzberg, Langmead. 2013. Bowtie2. *Nat. Methods* **9**, 357–359 (2013).

70. A. Tett, *et al.*, The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host & Microbe* **26**, 666–679.e7 (2019).

71. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

72. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

73. NCBI Resource Coordinators, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **41**, D8–D20 (2013).

74. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

75. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

76. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

77. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

78. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

79. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

80. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

81. D. T. Truong, *et al.*, MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).

82. J. D. Hunter, Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

83. A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, H. Pfister, UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).

84. P. J. A. Cock, *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

85. W. McKinney, Others, Data structures for statistical computing in python in *Proceedings of the 9th Python in Science Conference*, (Austin, TX, 2010), pp. 51–56.

86. S. van der Walt, S. C. Colbert, G. Varoquaux, The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).

Chapter 5

Discussion and Conclusions

5 | Discussion and Conclusions

Viral Like Particle enrichment (VLP) is a technique widely used to concentrate viral nucleic acids for NGS virome sequencing. Despite being a powerful tool that allows the study of viruses in challenging settings, it has been suggested multiple times that viral preparations can contain non-viral particles like bacterial cells. While this can be simply a matter of "lost-potential" with respect to the overall viral load that one is able to concentrate in the final sample, for some applications having a mixed community of viruses and bacteria can lead to wrong conclusions. This is the case, for example, of admixtures of bacterial and viral proteins that are all classified as viral, or the detection of bacterial genome fragments that are then labelled as viral, and propagated through public databases and meta-analyses. In this thesis, we analyzed the role of VLP enrichment in the context of the metagenomic analysis of viromes and the *de-novo* comprehensive reconstruction of viral genomes for the human gut microbiome.

In **Chapter 2** we presented an analysis of a step-by-step sequential VLP filtration on freshwater and lake sediments. Our goal was to understand the nature of the non-viral particles able to pass through the filters. Bacterial communities were profiled at each step of the filtration with 16S rRNA amplicon sequencing and shotgun sequencing. We found that each filter and sample type was characterized by a specific microbial community, with a general reduction of the diversity at each sequential filtration step. To quantify the final viral enrichment in samples, we used shotgun metagenomes collected at the end of each VLP enrichment and quantified the number of reads that mapped against the ribosomal rRNA genes. We designed this approach in order to use a common bacterial marker able to detect bacteria and archaea universally, without the need for a reference genome. Hence, we used the rRNA genes as a proxy for bacterial (i.e. contaminants, in the context of virome enrichment) abundance.

We estimated the amount of contaminants (i.e. non-viral sequences) in samples and found that the enrichment protocol worked better on water samples than on sediments. This was likely due to the intrinsic characteristics of the sediment samples (e.g. microbial diversity, porosity, minerals, and other particles interfering with the filtration protocol). However, even in water samples that had far less microbial rRNA, and were, therefore, less contaminated in principle, we still found microbial DNA. This indicates that VLP enrichment can have different outcomes depending on the nature of the sample analyzed and depending on the experimental procedure itself. Some of the bacterial contaminants we found exclusively in the final enriched sample (i.e. they were not retained by any of the filters) were

known to have a small cell size (e.g. candidate phyla radiation, CPR). This supports the hypothesis that small bacteria can pass through the filters, thus contaminating the final VLP preparation. However, the experimental procedure itself may push and squeeze other bacteria across the filters, allowing for some bigger cells to infiltrate in the final enriched sample. We conclude that bacteria are contaminants of VLP preparation, and we therefore advise caution when analysing viromes, as it cannot be assumed that metagenomes sequenced from VLP samples contain only viral sequences.

Few studies expressed the need for awareness with respect to the degree of contamination of viromes. However, testing the purity of the VLP preparation with PCR or qPCR against the 16S rRNA is a practice followed only by some, and overall there was no general assessment of the contamination in publicly available viromes. Nevertheless, awareness on the enrichment outcome of public data is essential when different datasets are combined and compared, as contamination may generate improper results and drive to spurious conclusions. We then refined the method to assess the contamination in viromes presented in Chapter 2 and designed ViromeQC, an open-source software to estimate the enrichment directly from the raw sequencing reads. We used ViromeQC to perform the largest meta-analysis on viral enrichment to date, which is described in detail in **Chapter 3**. Our analysis involved more than 2,000 environmental and human viromes that were screened to assess their level of non-viral contamination. Most viromes were indeed poorly enriched, with more than 50% viromes that were less than 3x enriched if compared to unenriched metagenomes of the same source.

We observed wide variability in the enrichment profiles of samples of the same dataset, and this suggests that even samples processed within the same experimental framework can have very different purities. While this may be related either to the intrinsic nature of each individual sample or to post-sampling contamination, it is nevertheless important to quantify the level of true enrichment in each sample to be aware of possible bias. ViromeQC does not allow to recover lowly enriched viromes, but it permits to rank them by enrichment efficacy so that differently enriched viromes can be considered as differently informative in any downstream analysis.

Thousands of metagenomes of different studies were used recently to build atlases of bacterial Metagenomic Assembly Genomes (MAGs), but the same had never been attempted for viruses. In **Chapter 4** we presented an approach to use highly enriched viromes in the discovery of thousands of novel potential bacteriophages in the human gut. We screened > 3000 viromes with ViromeQC and performed metagenomic assembly on the samples with the highest enrichment. Through careful annotation and identity-clustering of

the retrieved sequences, we defined 3,944 Viral Sequence Clusters (VSCs), and labelled them as known or unknown according to Viral RefSeq. In agreement with other studies, we found that the great majority (85.1%) of the clusters did not contain any known phages reference, and were then labelled as uVSCs (unknown VSCs). This highlights the great undiscovered potential in terms of viral diversity in viromes.

We extended our collection of sequences with >150,000 metagenomic contigs from thousands of metagenomes, demonstrating that even unenriched metagenomes and viromes can be exploited for *de-novo* viral discovery. We also showed that some of the sequences we could retrieve were extremely prevalent (up to 70% prevalence when mapping them back to 9,760 human gut metagenomes). Indeed, some of the most prevalent known viruses in the human gut were *Enterobacteria phage DE* and *crAssPhage*, found in 26.5% and 17.4% of the metagenomes, respectively. Finally, we are able to phylogenetically characterize several of the most prevalent viral entities and identified several cases of longitudinally sampled individuals that retained the same viral strain across time.

We are confident that this research may importantly contribute to future studies that address the human gut virome. To foster future research on the nature, prevalence, and abundance of the viral genomes we retrieved, we released the complete collection of sequences, as well as the bioinformatics pipeline to cluster the sequences into VSCs.

5.1 | Future Perspectives

The studies reported here highlight several interesting aspects of the role of VLP filtration in the characterization of the viral “dark matter”. Several intriguing scientific questions and potential future research directions emerge from the work I presented. Below are reported a few of the most interesting aspects I would like to follow-up in future studies.

- **Evaluation of different VLP enrichment techniques.** In **Chapter 2** we studied VLP enrichment protocols specific for freshwater and sediments, and several other protocols that exist to enrich for viruses. We also showed in **Chapter 3** that non-viral contamination affects multiple sample types regardless of the specific technique used. Nevertheless, it would be interesting to explore the nature of the contamination (i.e. the bacterial species able to pass through the filter) when using other protocols and other sample types. Examples are the cesium chloride ultracentrifugation or the tangential flow filtration.

- **Follow the enrichment along the filtration process via shotgun metagenomics.** In **Chapter 2** only 16S rRNA gene amplicon sequencing was performed on the filters of each filtration step, while shotgun metagenomics was used only for the starting and ending extremes of the process. To have a full picture of the enrichment dynamics of a VLP experiment it would be interesting to follow the process at each filtration step through shotgun metagenomics and ViromeQC analysis.
- ***De-novo* viral discovery from environmental samples.** We exploited VLP purification to select highly enriched viromes only from human gut studies. However, the same approach can be used to discover novel viruses in environmental samples as well. Metagenomic assembly of highly-enriched environmental viromes can be performed to support the discovery of previously uncharacterized viruses to be analyzed together with well-known ones to get a more complete picture of these still under-investigated organisms.
- **Decipher the role of VSCs in health and disease.** The VSC resource we curated and released can be used to further explore the human virome in multiple contexts. Many studied associated specific viral taxa, and especially bacteriophages, to human health and disease. Our collection of 3,944 viral sequence clusters encompasses more than 150k sequences, some of which are extremely prevalent. This resource can be exploited to identify viruses in the thousands of publicly available metagenomes and viromes, and to find associations with clinical metadata in the context of human health.
- **Extensive phylogenetics of the VSCs and microbial co-occurrence.** Temperate phages are extremely present in the human gut virome. Indeed, it is likely that many of the sequences composing our VSCs are of prophagic origin. This could also explain the reason why many contigs from highly enriched viromes are found in bacterial genomes. Phylogenetic comparison of the bacterial populations and the reconstructed viruses in each sample would also allow to study the co-evolution of bacteriophages and bacteria in the human microbiome. In order to perform such an analysis, sequences of each VSC could be reconstructed directly from mapping VSCs to metagenomes, and without the need to perform another round of metagenomic assembly. Moreover, by focusing on the datasets that follow individuals longitudinally in time, the temporal dimension could be added to phylogenetic (i.e. molecular clocking).

Chapter 6

Other main contributions

6.1 | Contribution and Context

Strain level profiling is a crucial aspect of computational metagenomics, but the most recent tools to achieve sub-species level identification have been developed, validated, and tested primarily on human metagenomics. In the following paper, I explored the use of three metagenomic strain-level profiling tools to analyze 1,614 urban-associated metagenomes collected in the public transportation systems of New York, Sacramento, and Boston. The main goal of the study was to demonstrate the feasibility of using three tools for strain-level profiling on environmental samples. We were able to track numerous strains of prevalent bacteria such as *Pseudomonas stutzeri* and *Stenotrophomonas maltophilia*, and we highlighted how the use of combinations of different tools can be beneficial to explore the microbial diversity of urban human-associated environments.

I was involved in this project through the CAMDA / MetaSUB challenge 2017. I retrieved, preprocessed, and analyzed the raw metagenomes, performed the computational analyses with the strain-tracking tools MetaPhlAn/StrainPhlAn, MetaMLST, and PanPhlAn, and wrote the manuscript.

6.2 | Manuscript

Profiling microbial strains in urban environments using metagenomic sequencing data

Moreno Zolfo [1], Francesco Asnicar [1], Paolo Manghi [1], Edoardo Pasolli [1], Adrian Tett [1], Nicola Segata [1,*]

[1] Department CIBIO, University of Trento, Trento, Italy

[*] Corresponding author: Nicola Segata (nicola.segata@unitn.it)

Published: Biology Direct 13, 9 - 2018

https://doi.org/10.1186/s13062-018-0211-z

Abstract

Background. The microbial communities populating human and natural environments have been extensively characterized with shotgun metagenomics, which provides an in-depth representation of the microbial diversity within a sample. Microbes thriving in urban environments may be crucially important for human health, but have received less attention than those of other environments. Ongoing efforts started to target urban microbiomes at a large scale, but the most recent computational methods to profile these metagenomes have never been applied in this context. It is thus currently unclear whether such methods, that have proven successful at distinguishing even closely related strains in human microbiomes, are also effective in urban settings for tasks such as cultivation-free pathogen detection and microbial surveillance. Here, we aimed at a) testing the currently available metagenomic profiling tools on urban metagenomics; b) characterizing the organisms in urban environments at the resolution of single strain and c) discussing the biological insights that can be inferred from such methods.

Results. We applied three complementary methods on the 1,614 metagenomes of the CAMDA 2017 challenge. With MetaMLST we identified 121 known sequence-types from 15 species of clinical relevance. For instance, we identified several *Acinetobacter* strains that were close to the nosocomial opportunistic pathogen *A. nosocomialis*. With StrainPhlAn, a generalized version of the MetaMLST approach, we inferred the phylogenetic structure of *Pseudomonas stutzeri* strains and suggested that the strain-level heterogeneity in environmental samples is higher than in the human microbiome. Finally, we also probed the functional potential of the different strains with PanPhlAn. We further showed that SNV-based and pangenome-based profiling provide complementary information that can be combined to investigate the evolutionary trajectories of microbes and to identify specific genetic determinants of virulence and antibiotic resistances within closely related strains.

Conclusion. We show that strain-level methods developed primarily for the analysis of human microbiomes can be effective for city-associated microbiomes. In fact, (opportunistic) pathogens can be tracked and monitored across many hundreds of urban metagenomes. However, while more effort is needed to profile strains of currently uncharacterized species, this work poses the basis for high-resolution analyses of microbiomes sampled in city and mass transportation environments.

Introduction

Complex communities of bacteria, fungi, viruses and micro-eukaryotes, called microbiomes, are an integral part of human and natural ecosystems (1, 2). Shotgun metagenomics (3) is a powerful tool to investigate such microbiomes. Indeed, metagenomics has enabled investigations such as those identifying associations between microbial communities and human diseases (1, 4–7) and it has even permitted the discovery of whole new bacterial phyla populating aquatic systems (8). However, while the microbiomes associated with the human body and with natural environments like soil and oceans have been extensively investigated (2, 9–11), there are instead only a few works characterizing the microbial communities associated with urban environments (12, 13).

The microbial communities populating the urban environment are in direct contact with the city's inhabitants and their associated microbiomes. Therefore, it is natural to assume there is interplay between the two, with the human inhabitants that have the ability to either acquire or deposit microbes as they travel through urban environments (13–15). Similarly to the ongoing efforts to characterize the role of microbiomes associated with the built environments (e.g. homes and offices) (16–19)microbial entities thriving within cities should also be considered for their potential interaction with the human microbiome. With the urban population projected to increase by 2.5 billion by 2050 (20–22), it is thus imperative to characterize the microbes that inhabit our cities and their genetic and functional diversity. Indeed, the study of urban microbiomes can be crucial for epidemiology and pathogen surveillance, but also for monitoring the spread of genetic microbial traits like genes responsible for resistance to antibiotics, similarly to what has recently been proposed in clinical settings (23, 24). Recently, endeavors like the MetaSUB Project have started to characterize the composition of the microbial inhabitants of urban environments (25), but the increasing effort in sampling and metagenomic sequencing from these environments has to be paralleled with either the development or adaptation of computational tools able to fully exploit this urban metagenomic data.

Computational metagenomic approaches for microbiome analysis are in part dependent on the source of the metagenome. The human gut microbiome, for example, can be successfully profiled by assembly-free methods (1) whereas environmental microbiomes characterized by a much larger diversity are typically more dependent on metagenomic assembly (26, 27) and binning (28, 29). The latest advances in computational metagenomics now permits profiling metagenomes at the sub-species resolution of single strains (30–35)

and these methods are particularly suited for the analysis of human microbiomes (36–39). However, little is known about the utility of existing profiling tools when applied to urban metagenomes, and strain-level analysis has never been applied in the urban setting.

In this work we tested, validated, post-processed and interpreted the application of three strain-level profiling tools originally developed for the human microbiome on a large set of urban metagenomic samples. We analyzed a total of 1,614 metagenomes of the MetaSUB dataset as distributed as a CAMDA challenge (from now on simply referred to as “MetaSUB dataset”).

Results and Discussion

We applied three strain-level computational profiling approaches for metagenomic data (MetaMLST (35), StrainPhlAn (34), PanPhlAn (33)) to a total of 1,614 environmental samples collected across the urban environment of three cities in the United States: New York (13), Boston (12), and Sacramento (unpublished data). The metagenomes were analyzed in the framework of the CAMDA 2017 Challenge conference and are herein referred to as the “MetaSUB data set” which includes the unpublished data of the Sacramento urban environment.

The methods adopted in this analysis have the capability to characterize microbial organisms from metagenomes at the resolution of single strains of known species and they exploit different genomic features, but they have never been applied to urban metagenomes (see Methods).

Strain typing by multi locus sequence typing using MetaMLST

The first strain typing approach we considered is based on Multi Locus Sequence Typing (MLST). MLST is an effective cultivation-based technique that is frequently used in clinical microbiology and epidemiology to identify and trace microbial pathogens (40, 41). The method exploits a reduced set of hypervariable loci (usually from 7 to 10) of the target species, which are subjected to Sanger amplicon sequencing and used to define an allelic profile for each strain called a sequence Type (ST) (42). MetaMLST (35) is a recent metagenomic cultivation-free extension of the approach that takes advantage of the hundreds of MLST typings available in public databases (43, 44) and performs an *in-silico* MLST analysis on the raw metagenomic reads. MetaMLST detects already observed STs but can also discover new ones that are diverged from the STs already publicly available (see **Methods**).

We applied MetaMLST to profile every species for which an established MLST schema is available. In the MetaSUB dataset a total of 551 samples were positive for at least one species and we recovered a total of 121 known and 510 novel STs of a total of 15 different species (**Table 6.1**). The most prevalent species found in the MetaSUB dataset by MetaMLST were *Acinetobacter baumannii*, *Enterobacter cloacae,* and *Stenotrophomonas maltophilia*, and the most prevalent STs were *A. baumannii* ST-71 (detected 20 times) and *Klebsiella oxytoca* ST-44 (detected 8 times).

Species	**Known**		**Novel**			
	samples	**STs**	**samples**	**STs**	**Prevalence in dataset**	**Most prevalent ST(#samples)**
Acinetobacter baumannii	69	22	123	117	11.90%	ST71 (20)
Enterobacter cloacae	63	39	89	85	9.42%	ST50 (7)
Stenotrophomonas maltophilia	15	12	98	90	7.00%	ST100009 (3)
Cronobacter spp.	2	2	66	56	4.21%	
Klebsiella pneumoniae	13	12	38	38	3.16%	
Bacillus cereus	15	11	17	17	1.98%	
Klebsiella oxytoca	12	5	18	18	1.86%	ST44 (8)
Achromobacter spp.	2	2	27	27	1.80%	
Enterococcus faecalis	6	5	18	18	1.49%	
Propionibacterium acnes	4	2	18	18	1.36%	
Escherichia coli	3	2	16	15	1.18%	
Pseudomonas fluorescens	1	1	7	6	0.50%	
Pseudomonas aeruginosa	5	1	1	4	0.37%	
Clostridium botulinum	1	5	5	1	0.37%	
Total	**211**	**121**	**541**	**510**		

Table 6.1. Results of MetaMLST applied to the 1614 samples of the MetaSUB dataset. *MetaMLST was applied on the full panel of 113 species, detecting in total 121 known and 510 previously unobserved profiles. The table reports the number of samples and STs found for both known and novel STs of the 15 species profiled in the MetaSUB dataset. The prevalence values are normalized over the total number of samples (1614).*

A. baumannii was originally described as an environmental bacterium and has been isolated from soil and water (45), but it can also be an opportunistic pathogen (46). It is one of the six members of the pathogenic group ESKAPE (47) and it is frequently responsible for nosocomial infections. *A. baumannii* and the closely related species *Acinetobacter*

calcoaceticus, *Acinetobacter pittii* and *Acinetobacter nosocomialis* are members of the ACB complex (48, 49), and due to the genetic similarity within this complex, a single MLST schema (50) is used for the whole group (51). Members of the ACB complex were detected in 192 New York urban metagenomes. When we modelled the detected STs and the reference isolates downloaded from public sources (43, 50) with the minimum spanning tree approach, we found that the majority of the strains from the MetaSUB samples belonged to *A. nosocomialis* and *A. calcoaceticus* STs (**Fig. 6.1A**). The majority of the detected STs fell outside the subtree with the known and labelled *A. baumannii* STs. Overall, this demonstrates the presence of *Acinetobacter* and therefore potentially opportunistic pathogens in the urban environment and highlights how a very well defined subtree of the group comprises strains that are found in the ecological niche of the urban environment.

Enterobacter cloacae and *Stenotrophomonas maltophilia* were the second and third most prevalent organisms in the MetaSub cohort, with a prevalence of 9.42% and 7% respectively. As *A. baumannii*, both *E. cloacae and S. maltophilia* pathogenic strains are responsible for nosocomial opportunistic infections, especially for immunosuppressed patients (52, 53). We applied MetaMLST to the MetaSub cohort and detected 124 Sequence Types of *E. cloacae* (85 of which were novel) and 102 STs of *S. maltophilia* (90 novel, **Table 6.1**). For both species, we did not detect any strong pattern associating specific STs to the MetaSub dataset. Both *E. cloacae* and *S. maltophilia* could be detected only in the New York samples, although this could be due to the overrepresentation of samples from New York in the overall cohort. MetaMLST allowed to highlight many prevalent STs of both species in the urban surfaces of New York's subway. Indeed, several of them were found multiple times (e.g. *E. cloacae* ST 50 was found 7 times and *S. maltophilia* novel STs 100006 and 100009 were found 3 times each, **Supplementary Fig. 6.2 and 6.3**). The detected sequence types were compared with STs retrieved from BigsDB (43) (for *S. maltophilia*) and with the STs inferred from 103 public reference genomes of *E. cloacae*, since no isolates metadata for *E. cloacae* were available in BigsDB (see **Methods**). We found no associations with the site of isolation (in *E. cloacae*, **Supplementary Fig. 6.2**) or with the bio-geographic location of the sample (*S. maltophilia*, **Supplementary F*ig. 6.3***).

We next focused on *Escherichia coli*, a common member of the human gut microbiome that is also found in the environment. *E. coli* has a large number of sequence types that can be classified in phylogroups, with the majority of commensal strains found within the phylogroups A and B1 (54, 55), and opportunistic pathogenic strains, such as ExPEC *E. coli*, falling in phylogroup B2 (56).

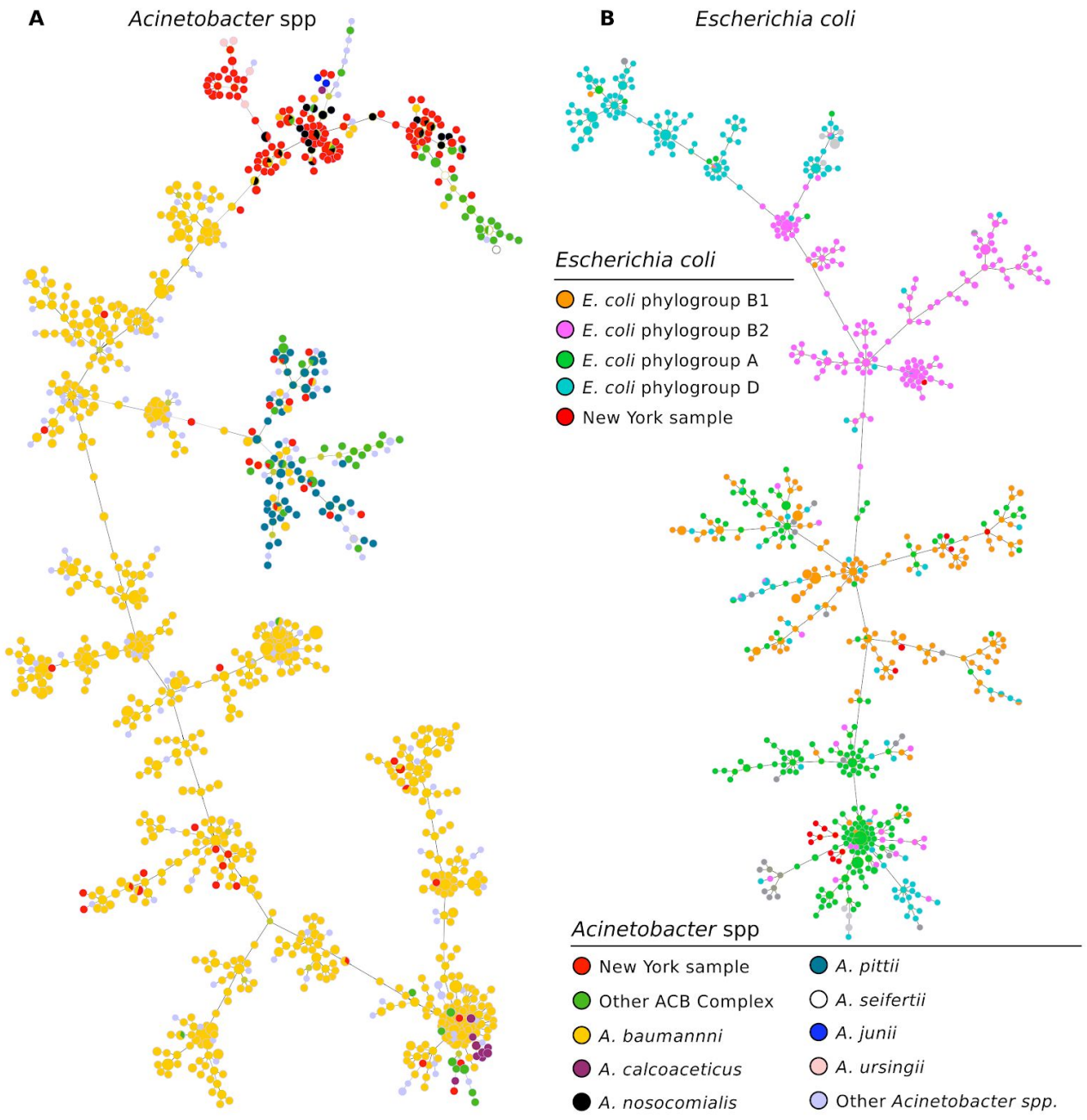


Figure 6.1. Application of MetaMLST to the 1614 urban metagenomes of the MetaSUB dataset. *Minimum spanning trees (MST) were generated on the basis of the allelic profile (57), where each node in the MST represents a Sequence Type (ST) and an edge connects similar STs (i.e. sharing at least one identical locus) with a length proportional to their allelic-profiles similarity. The two MSTs were built with PhyloViz (58). The 139 detected STs of A. baumannii* ***(A)*** *and the 17 STs of E. coli* ***(B)*** *are placed in the tree together with the available known STs for which the species is available. In both trees, the STs of the samples from the New York built environment are colored in red.*

MetaMLST detected *E. coli* in 19 New York subway samples and by comparing the recovered STs with the references available in BigsDB (43), we were able to assign the strains to the *E. coli* phylogroups **(Fig. 6.1B)**. The majority (53%) of the samples fell in the mainly non-harmful phylogroup A. One sample harboured a novel *E. coli* type (*adk 37; fumC 38; gyrB 19; icd 37; mdh NEW; purA 11; recA 26*) very closely related to ST-95 (3 SNVs over 3423 total nucleotides, profile) which is one of the most commonly found *E. coli* phylogroup B2 strains (59, 60). These results highlight that MetaMLST is capable of detecting microbes at the strain level in complex environmental communities thus enabling epidemiology modelling from urban samples.

Phylogenetic strain characterization using extended single nucleotide variant profiling

MetaMLST is a rapid method for the strain level profiling of a species for which a MLST schema exists and strains are identified by exploiting single nucleotide variants (SNVs) within a small set of genetic loci. With the goal of extending this approach, we recently developed StrainPhlAn (34) which characterizes strains in metagenomes by targeting the SNVs within clade-specific markers (>200 markers for each species). The increased number of loci enables a finer resolution for distinguishing closely related strains, and unlike MetaMLST is applicable to any species of interest for which at least one reference genome is available.

We applied StrainPhlAn to all the microbial species identified in the MetaSUB dataset by the species profiling tool MetaPhlAn2 (61). In total, we identified 539 microbial species with a relative abundance above 0.5%. Of these, 155 were present in more than 10 samples with only a minor correlation between the sequencing depth of each sample and the observed number of species (**Supplementary Fig. 6.1**). In samples from New York we found *Pseudomonas stutzeri and Stenotrophomonas maltophilia* to be the most abundant characterized species (**Supplementary Table 6.1**). Boston was instead dominated by *Propionibacterium acnes* as previously reported (12), while the city of Sacramento showed a high prevalence of species in the *Geodermatophilaceae* family and *Hymenobacter* genus, which are known environmental bacteria (62, 63). In addition, in the Sacramento samples we found other potential opportunistic pathogens such as *Halomonas* spp. (64) and *Kocuria* spp., which is a species commonly found both in soil and human skin (65–67).

The most prevalent species identified in New York, *P. stutzeri,* was identified in 967 samples across the New York dataset. Of those, 416 samples harboured *P. stutzeri* at a sufficient coverage to be profiled by StrainPhlAn. The StrainPhlAn inferred phylogeny highlighted the presence of three clusters of *P. stutzeri* strains that do not correlate with the

geographic area from which the sample was taken (**Fig. 6.2A**) nor are they correlated with other sample characteristics such as surface material (**Fig. 6.2B**). This may suggest that samples collected in high-density and high-transit urban environments may be extremely heterogeneous without evidence of sub-niche selection. Alternatively, this could be a reflection of these species being carried around between stations and other surfaces of the urban furniture by commuters, although this has never been previously observed; further research is needed to demonstrate such kind of events.

We next profiled *S. maltophilia*, which is the second most prevalent species in the New York dataset. *S. maltophilia* is not only a common environmental bacterium, but also a nosocomial opportunistic pathogen in immunocompromised patients (68). We found 654 samples in which *S. maltophilia* was present. Of those, 111 samples harboured *S. maltophilia* at a sufficient coverage to be profiled by StrainPhlAn and were considered in the phylogenetic analysis. From the ordination plot based on inter-strain genetic distances, we identified three main clusters (**Fig. 6.2C**) that, similarly to *P. stutzeri*, did not show any correlation with either the geography or the surface material from which the sample was taken, supporting the hypothesis that the genetic structures of microbial species and sample characteristics in urban environments tend to be uncoupled.

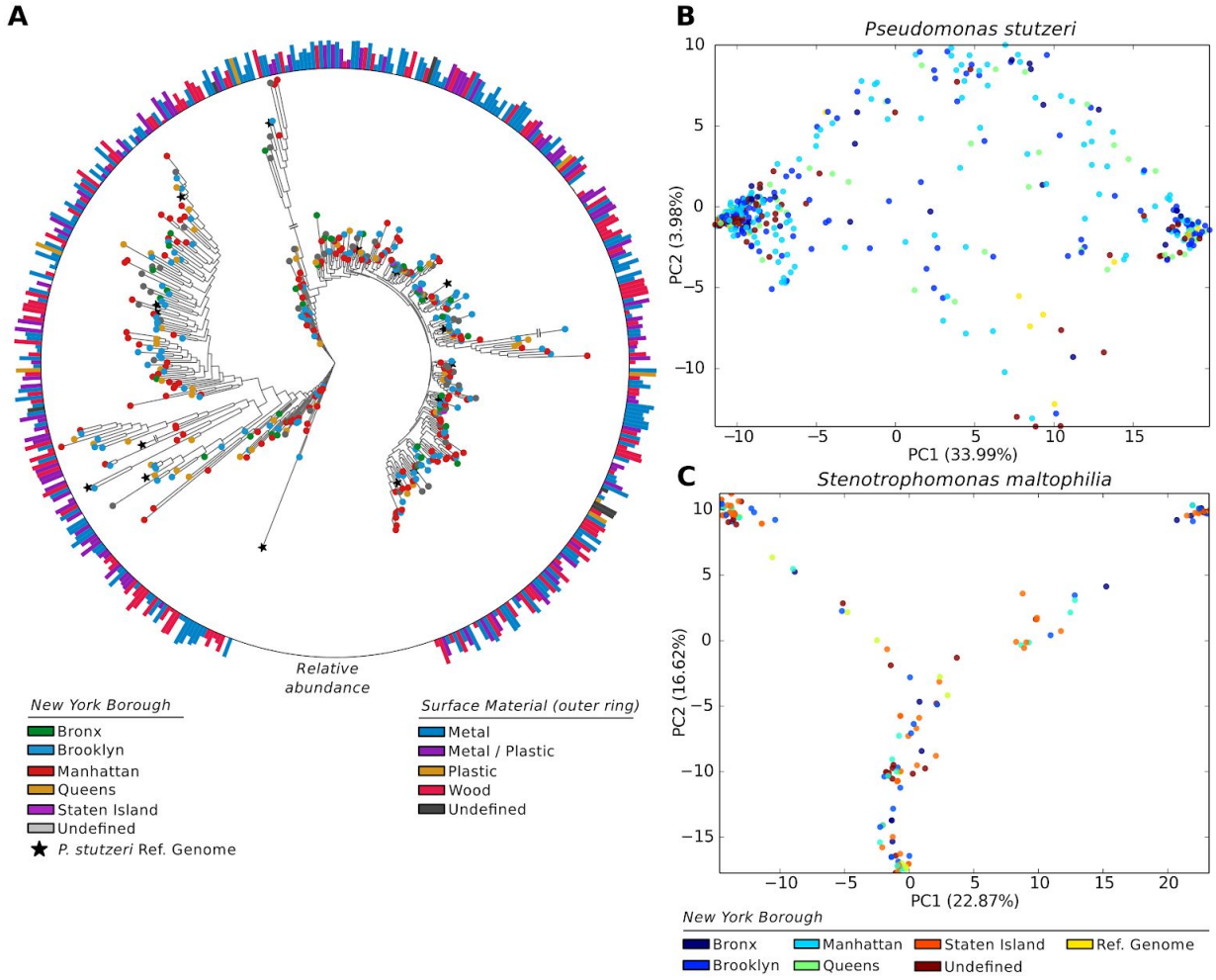


Figure 6.2. Strain-level phylogenetic analysis of the two most prevalent bacterial species identified in the metagenomic samples of the New York urban environment*. The phylogenetic trees are inferred by applying StrainPhlAn on the raw sequencing reads.* ***(A)*** *Maximum likelihood phylogeny of P. stutzeri (built with RAxML (69) internally in StrainPhlAn). The root of the phylogenetic tree is placed using P. putida as an outgroup. Nodes are colored by the New York Borough from which the sample was collected, with black stars representing reference genomes. The height of the bars of the bar-plot on the outer ring represents the relative abundance of P. stutzeri as computed by MetaPhlAn2, while the color represents the surface material of the sample. The lengths of branches marked with a double horizontal line are reduced by 50%* ***(B, C)*** *PCA plot based on the genetic distance computed on the species-specific markers sequences of 416 samples and 18 reference genomes of P. stutzeri* ***(B)*** *and 111 samples and 80 reference genomes of S. maltophilia* ***(C)****. The points are colored according to the New York Borough.*

Evidence for high intra-species strain heterogeneity in urban microbiome samples

Complex microbial communities can harbor multiple strains of the same species. This is a well-known characteristic for both human associated (34, 70) and environmental microbiomes, but profiling multiple related strains simultaneously within the same sample is currently very challenging (3). It is nonetheless important to quantify the strain level heterogeneity within a sample. Similarly to what we did previously for the human gut microbiome (34), we investigated the strain heterogeneity for the species in the urban microbiomes. This was performed by quantifying the rate of polymorphic nucleotides for each position along of the species' reads-to-markers alignments (see **Methods**). We computed the estimate of strain-heterogeneity for a number of the most prevalent species in each city (**Fig. 6.3**).

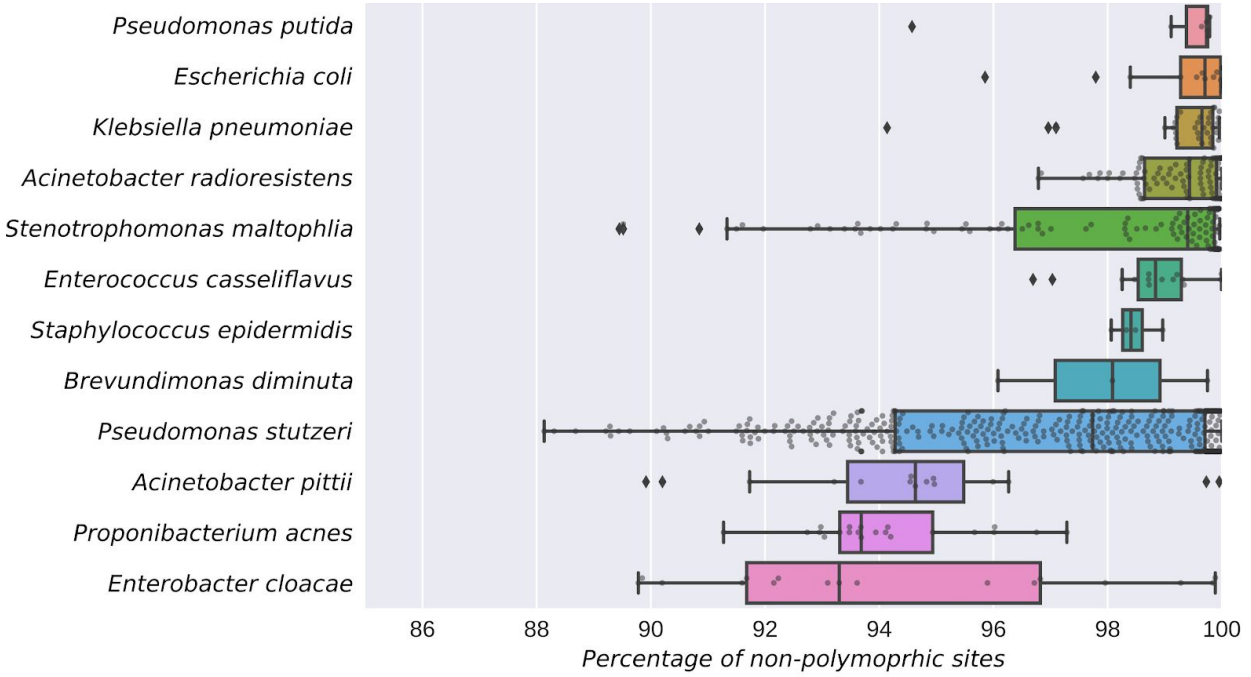


Figure 6.3. **Strain heterogeneity distribution for a set of highly prevalent species across the MetaSUB dataset.** For each species, we report the distribution of the average rate of non polymorphic sites in the sample (**see Methods**). The boxes show the first and third quartiles of the dataset, the bar inside the box represents the median (second quartile), while the whiskers extend to cover the 99.3% of the distribution. External points are represented as outliers.

We observed a higher intra-species variability in the MetaSUB dataset than what we previously found in the human gut microbiome (34), thus suggesting that the higher complexity and species richness of environmental microbiomes (71) is also reflected at the sub-species level. For instance, *E. cloacae* and *P. acnes* show high median polymorphic rates (**Fig. 6.3**) suggesting that more than one strain of the species is present within the sample. In contrast, for *P. putida* and *E. coli* a single strain dominates the community for most the samples. We also highlight the presence of species characterized by higher polymorphic-rates inter quantile ranges (IQR), like *P. stutzeri* and *S. maltophilia*, suggesting that these species are sometimes single-strain dominated and other times they are represented simultaneously by many distinct strains. We can speculate that the higher percentages of polymorphic rates can be due to the high number of distinct microbial sources (subway users) coming in contact with the sampled surfaces. Overall, these results highlight that the same species can harbour a substantial strain heterogeneity across samples, and that these strains can sometimes coexist in the same niche.

Functional profiling of strains based on species' pangenomes

MetaMLST and StrainPhlAn are based on the comparison of the SNVs within species-specific markers. Microbial species can also be profiled according to the presence or absence of their gene repertoire (72–74). In order to profile strains according to their genomic content (gene repertoires), we applied PanPhlAn, a software tool that outputs the gene presence-absence profile for a given species in a metagenome. In addition to the inference of the relatedness of strains, this approach can also be useful to identify specific strain-specific genomic traits. These include for instance antibiotic resistance and virulence determinants that can be present in only a subset of the strains in a species. In previous studies, PanPhlAn proved successful in detecting pathogenic species besides commensal strains of *E. coli* (33, 75), but again this was performed only in human-associated microbiomes.

To test whether differences in strains could be observed in the urban metagenomes, we applied PanPhlAn to target *E. coli* in the New York dataset. *E. coli* was detected at sufficient coverage for profiling in 19 samples, of which five were among those profiled with MetaMLST. Comparing the presence-absence profiles of this 19 *E. coli* with a selection of reference genomes (i.e. those contained in PanPhlAn), revealed that the New York samples had a genetic functional potential similar to the largely non-pathogenic phylogroups A and B1, similarly to what was shown with MetaMLST. Conversely, only two samples were close to phylogroup B2 (**Fig. 6.4A**).

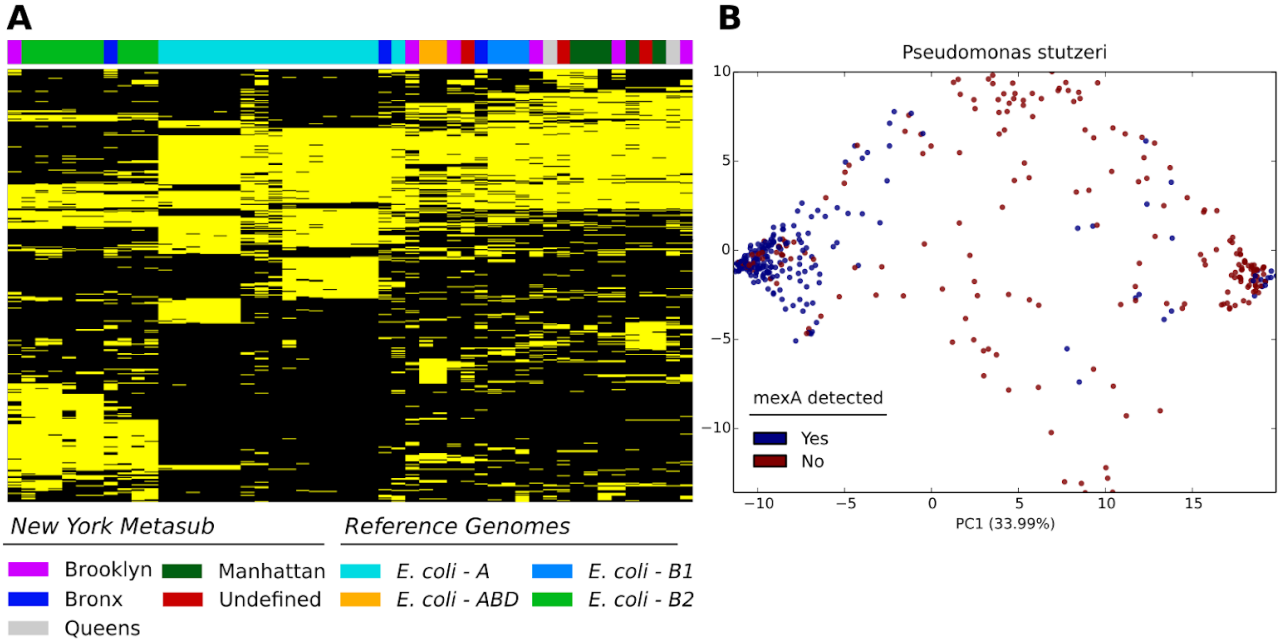


Figure 6.4. Functional profiling of the species of the MetaSUB dataset across the New York urban environment. (A) *PanPhlAn presence-absence matrix of Escherichia coli. The rows represent the gene families while columns represent the samples. The top colorbar highlights the New York Borough and the E. coli reference genomes' phylogroups. In the heatmap yellow corresponds to presence, black corresponds to absence. Only the gene-families present in less than 90% and more than 10% of the samples were included.* ***(B)*** *PCA plot based on the genetic distance computed on the species-specific markers sequences of 416 samples and 18 reference genomes of P. stutzeri as reported in* ***Fig. 6.2C****. Each point is a sample and is colored according to the presence-absence of the mexA component of the Pseudomonas MexAB-OprM efflux system.*

An analysis based on the genomic content of the species of interest can highlight the presence of specific traits of a species within a complex microbial community (76). For example, it would be useful for epidemiological and microbial surveillance to profile and trace directly specific antibiotic resistance genes or virulence factors. To test whether the identification of a specific genetic capability could be achieved in the urban environment, we applied PanPhlAn to profile a species commonly identified in the MetaSUB dataset, *P. stutzeri*, which is also known to encode for different antibiotic resistances (77, 78). As an example, we specifically targeted the presence of the *mexA* gene, a component of the MexAB-OprM efflux system, which can confer resistance to numerous antibiotics and other antimicrobial agents (79, 80). We found that *P. stutzeri mexA* strains were present in a subset of the New York samples (present in 372, absent in 56, **Fig. 6.4B**). In total, 372 New York samples encoded mexA, while 56 samples did not, and the PanPhlAn results were

generally in agreement with the three clusters model obtained with StrainPhlAn. Interestingly, while clusters of *P. stutzeri* grouped both according to the genetics and the presence/absence of *mexA*, few strains that contained mexA clustered genetically with strains that did not contain the gene and vice-versa. Indeed, the presence of the same protein encoded by two strains that are genetically very distant may imply that the presence of *mexA* in some of these strains is imputable to some degree of lateral gene-transfer.

Overall, these findings highlight that it is possible to type at the functional level populations in the urban metagenomes using strain-level approaches based on the overall genomic repertoire and that samples can be investigated at a deeper level to unravel the diversity of specific microbial genetic traits among complex communities.

Comparing strain profiling by SNVs and gene content.

The two approaches we presented so far can reflect the strain-level diversity within a species, either taking in consideration the genomic content of strains, or their phylogenies. However, the two methods can convey different information. For example, as highlighted above for the *mexA* gene in Pseudomonas stutzeri, two strains could be phylogenetically very similar while displaying different resistance capabilities, which is why these methods should be considered complementary. In order to further evaluate the consistency and complementarity of the two approaches to profile strains, we performed a comparison between the two distance measures of PanPhlAn and StrainPhlAn. We investigated a panel of the urban species already analyzed above, and computed the pairwise phylogenetic (StrainPhlAn) and phylogenomic (PanPhlAn) distances within the samples (see **Methods**).

We found that genetic and genomic variations within the same sample are generally correlated for all the six species considered, confirming that both measures are effective proxy for strain relatedness and identity across samples (**Fig. 6.5**). However, the correlation coefficient varied across species, spacing from 0.34 (p-value $5.2e^{-219}$) for *A. radioresistens* to 0.85 (p-value $6.9e^{-17}$) for *E. cloacae.* These values reflect a different consistency between the phylogenetic signal and the evolutionary modifications of the functional profiles.

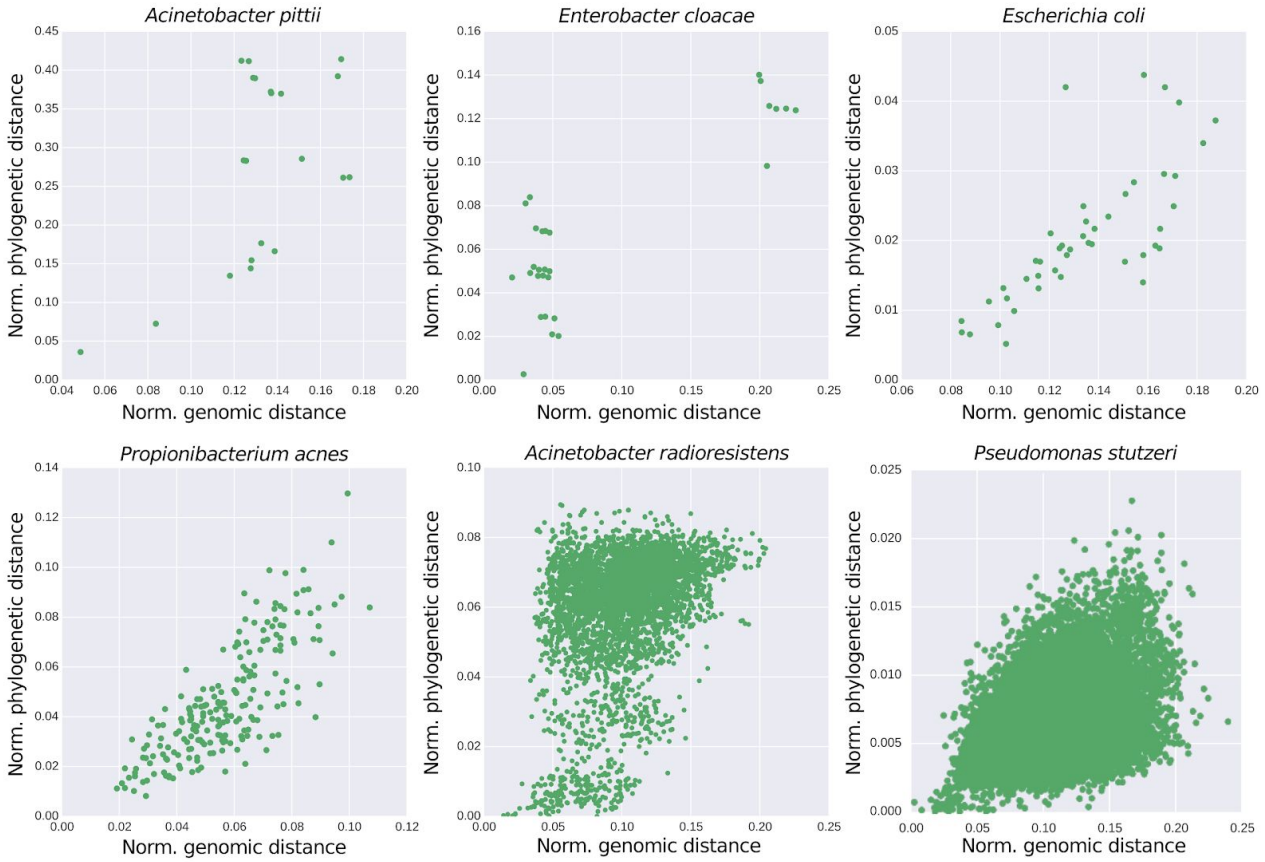


Figure 6.5. Normalised phylogenetic distance vs genomic-content distance within samples of six representative species of the MetaSub dataset. *Each data point refers to a pair of two strains of the same species in different samples. The genomic distance is defined as the normalised Hamming distance between binary vectors of presence-absence as reported by PanPhlAn. The phylogenetic distance is defined as the branch length distance of the two leaves in the StrainPhlAn phylogenetic tree, normalised over the total branch length of the tree. Pearson's correlation coefficients are A. pittii: 0.57, E. cloacae: 0.85, E. coli: 0.75, P. acnes: 0.79, A. radioresistens: 0.34 and P. stutzeri: 0.41. P-values are always lower than 1e-5.*

We also highlight the presence of samples that, regardless of the species, are much more functionally similar than the phylogenetic modeling would suggest, possibly reflecting convergent functional adaptation. Conversely, increased genomic content distances, suggests rapid functional divergence potentially due to plasmids, bacteriophages, or other lateral gene-transfer events. Such patterns, detected for example in *P. stutzeri* and *A. radioresistens*, are suggesting that strains can be very similar according to phylogeny and still be notably diverse in their functional potential.

Conclusions

We presented here the application of three strain level profiling tools to environmental urban metagenomics. While these tools were specifically developed for the context of the human microbiome, we highlighted that it is possible to apply them to efficiently perform strain profiling in the context of the urban environment. We provide evidence that potential pathogenic species can be recovered, typed, and traced across microbial communities that are wider and more complex than the ones we observe in the human microbiome. Moreover, the phylogenetic relation of strains in the same species and their functional repertoires can be simultaneously profiled, thus providing a more complete characterization of strains in the samples. These findings suggest that the tools presented above are effective for the purposes of pathogen surveillance and epidemiology in the context of environmental metagenomics.

The three methods presented in this work are capable of profiling microbes that are close to a reference sequences (MetaMLST), or for which a sequenced genome for the target species exists (StrainPhlAn and PanPhlAn). Because environmental microbiomes can contain a larger amount of unknown species (3) compared to human associated microbiomes, this dependency on already sequenced data can limit strain profiling to only a portion of the whole microbiomes. Additional profiling approaches can exploit metagenomically assembled contigs or genomes (3, 26–28, 30, 81) which are widely employed in environmental metagenomics and are necessary when targeting the fraction of not previously sequenced taxa. Our strain-profiling methods can be extended to use metagenomic assembled genomes as reference, and this would provide a combined assembly-based and assembly-free tool to explore the uncharacterized diversity in microbiome samples with strain-level resolution.

This work demonstrates that assembly-free strain-level profiling through SNVs and genomic content is a promising technique for comprehensive strain-resolved metagenomics in the urban environment.

Methods

We profiled a total of 1614 samples with three strain-level profiling tools described below. The dataset comprehended 1572 samples collected in the city of New York (NY, U.S.A., (13)), 24 samples collected in the city of Boston (MA, U.S.A., (12)) and 18 samples collected in the city of Sacramento (CA, U.S.A., unpublished). Samples from Boston and New York are publicly available at NCBI under codes PRJNA301589 and PRJNA271013, respectively.

MetaMLST

MetaMLST (35) is a tool for strain-level typing and identification from metagenomic data. It exploits the Multi Locus Sequence Typing (MLST) approach and performs an *in-silico* reconstruction of the MLST loci using a reference-guided majority rule consensus method. MetaMLST detects the sequence type (ST) of the most abundant strain the target species in the sample. Specifically, MetaMLST reconstructs the sequence of each locus from the raw metagenomic reads and compares it with a database of previously observed variants. Additionally, MetaMLST is capable of identifying new loci that diverge from the closest known sequence by up to 10 single nucleotide variants (SNVs). Hence, MetaMLST detects both known and novel (i.e. previously unobserved types) STs.

We applied MetaMLST version 1.1 to the entire MetaSUB dataset by mapping the raw reads against the MetaMLST database as of April 2017, consisting of 113 organisms, 798 loci, 46.2 Mbp and 12929 total profiles. The mapping was performed with bowtie2, version 2.2.6 (reference) as previously described (parameters: -D 20 -R 3 -N 0 -L 20 -i S,1,0.50 -a --no-unal) (35). Alignment files were sorted with Samtools version 1.3.1 (82). We reported only the species for which at least one known ST could be detected.

StrainPhlAn

StrainPhlAn (34) is a tool for identifying the specific strain of a given species within a metagenome. The tool is designed to track strains across large collections of samples and takes as input the raw metagenomic reads in FASTQ format. After mapping the reads against the set of species specific markers (>200 per species), StrainPhlAn reconstructs the sample specific marker loci using a variant calling approach and outputs the sequences of each sample-specific marker in FASTA format. Sequences are extracted from the raw reads using a reference-free majority rule that filters out noisy regions. The resulting sequences

were then concatenated and aligned by StrainPhlAn with Muscle version 3.8 (83). In this work, we applied StrainPhlAn to the whole MetaSUB dataset and investigated a panel of 12 species that were locally prevalent in the three cities of the MetaSUB dataset. The reconstructed markers were used to build the phylogenetic tree and the PCA plots of *P. stutzeri* and *S. maltophilia* (**Fig. 6.2**). The reads-to-markers alignments of the 12 species were used in the calculation of the polymorphic rate (**Fig. 6.3**). StrainPhlAn version 1.0 was used with default parameters, using the mpa_v20_m200 markers database of MetaPhlAn2 (61). The mapping against the markers was performed with Bowtie2, version 2.2.6, with the parameters implemented in the StrainPhlAn pipeline (34).

PanPhlAn

Pangenome-based Phylogenomic Analysis (PanPhlAn) (33) strain-level metagenomic profiling tool for identifying the gene composition of a strain of a given species within metagenomic samples. The approach of PanPhlAn is based on the identification of presence/absence patterns in the genomic content within the members of the same species, across complex metagenomic samples. As the pre-built PanPhlAn database did not include the pangenome of *Pseudomonas stutzeri*, we built a custom db from 19 high-quality reference genomes (NCBI accession numbers: ASM19510v1, ASM21960v1, ASM26754v1, ASM27916v1, ASM28055v1, ASM28295v1, PseStu2.0, ASM32706v1, PstNF13_1.0, PstB1SMN1_1.0, ASM59047v1, ASM66191v1, ASM95268v1, ASM98286v1, ASM103864v1, ASM106422v1, ASM127647v1, ASM157508v1) which were first annotated using Prokka (84) and then clustered into gene-families with Roary (85). We profiled the 1572 New York samples from the MetaSUB dataset with PanPhlAn version 1.2.1.3.

Visualization and statistical tools and phylogenetic distances

We defined the phylogenomic distance between two samples as the pairwise Hamming Distance on the PanPhlAn presence-absence profile for each sample, represented as binary vectors where 1 represents the presence of the gene, and 0 represents its absence. The phylogenetic distance was calculated as the minimal total branch-length distance between leaf nodes, normalized by the total branch length, using custom python scripts based on BioPython BaseTree (86, 87).

The phylogenetic trees were built with RAxML (69) version 8.1.15 (parameters: -p 1989 -m GTRCAT) and plotted with GraPhlAn (88). Minimum Spanning Trees were drawn with PHYLOViZ 2 (58) using the goeBURST Full MST algorithm (57). Isolates Sequence Types

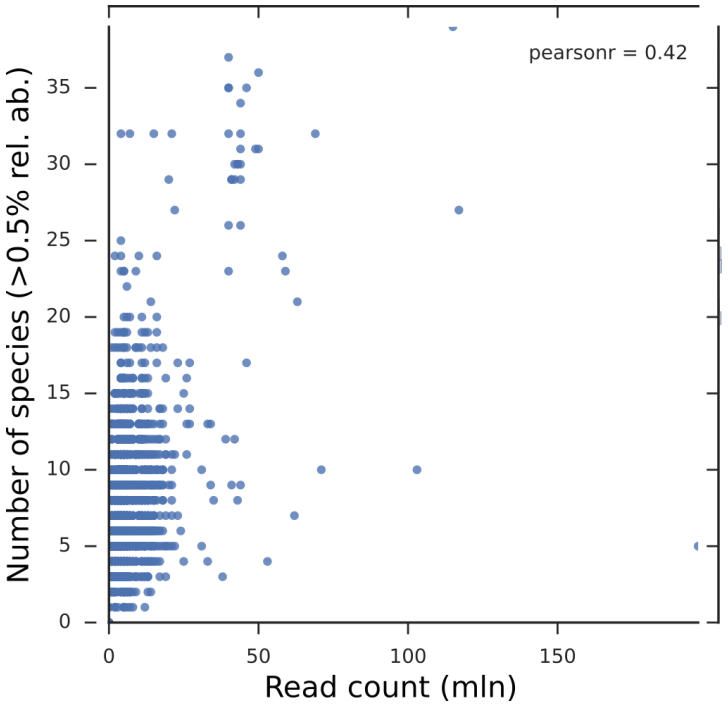
(STs) visualized with the MetaSub STs in **Fig. 6.1 and Supplementary Fig. 6.3** were retrieved from BIGSdb (43). STs of *E. cloacae* (**Supplementary Fig. 6.2)** were inferred from 103 assembled genomes retrieved from NCBI Assembly (89) (taxid = 550, Assembly Level equal to “complete genome”, “chromosome” or “scaffold”). The ST was inferred with BLAST by mapping the genomes against the MetaMLST database (35) for *E. cloacae* and by reconstructing the corresponding ST by the allelic profile (script mlst.py, default options). Only the genomes with all MLST loci and for which the ST was known were selected (71 out of 103 genomes). Metadata associated with each genome were extracted from the Sample’s GenBank Record.

The principal component analysis (PCA) plots were drawn with the scikit-learn package using the aligned concatenated marker sequences of StrainPhlAn as arrays of binary features. All the overlaid metadata used to colorize the trees and PCA plots came from the respective studies.

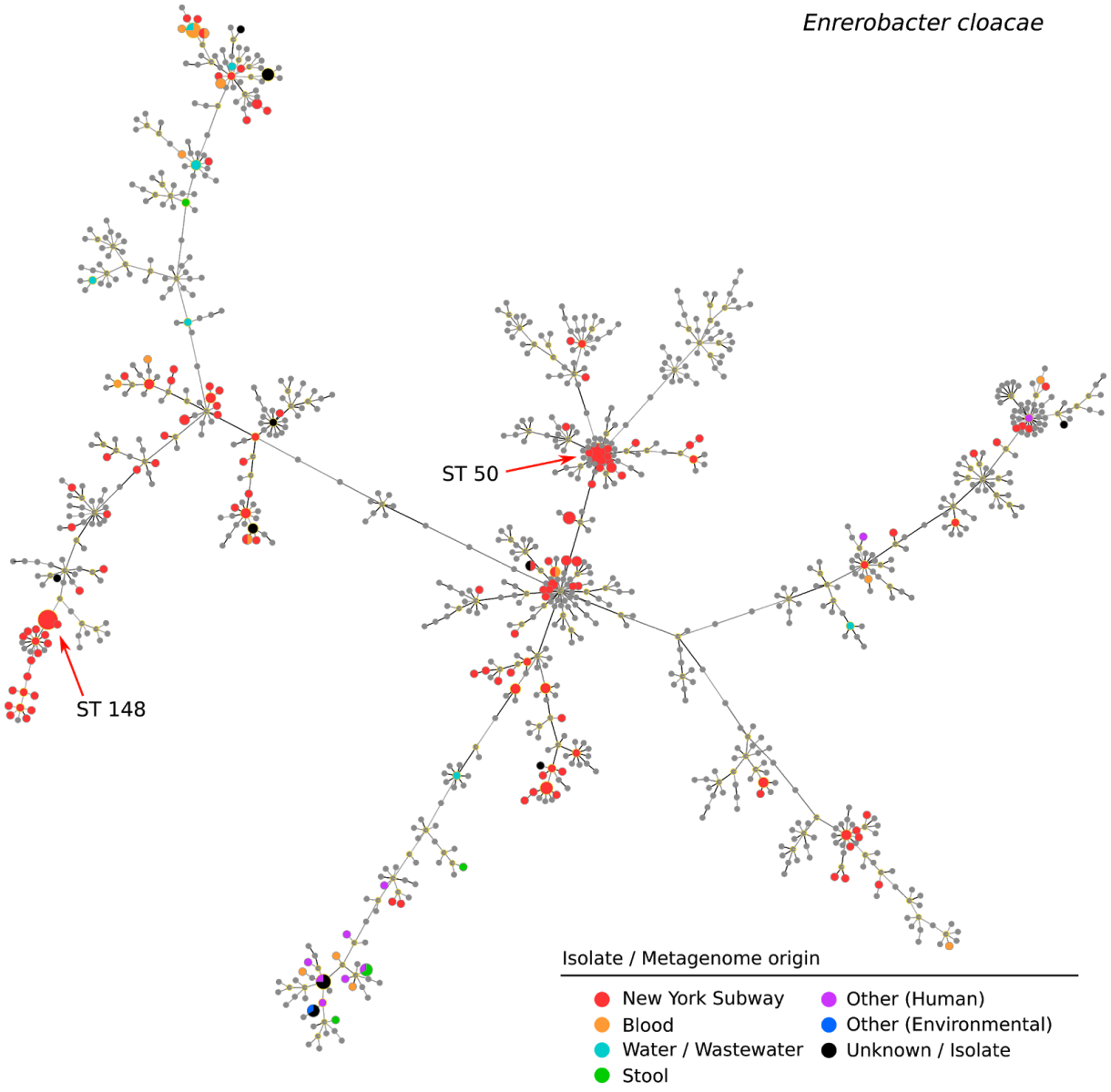
The presence of polymorphic sites within the reads-to-markers alignment was calculated and reported with StrainPhlAn (34), testing the non-polymorphic null hypothesis on a binomial test on the nucleotides distribution of each position in the alignment. The plots were drawn with python packages seaborn and matplotlib (90).

Supplementary Material

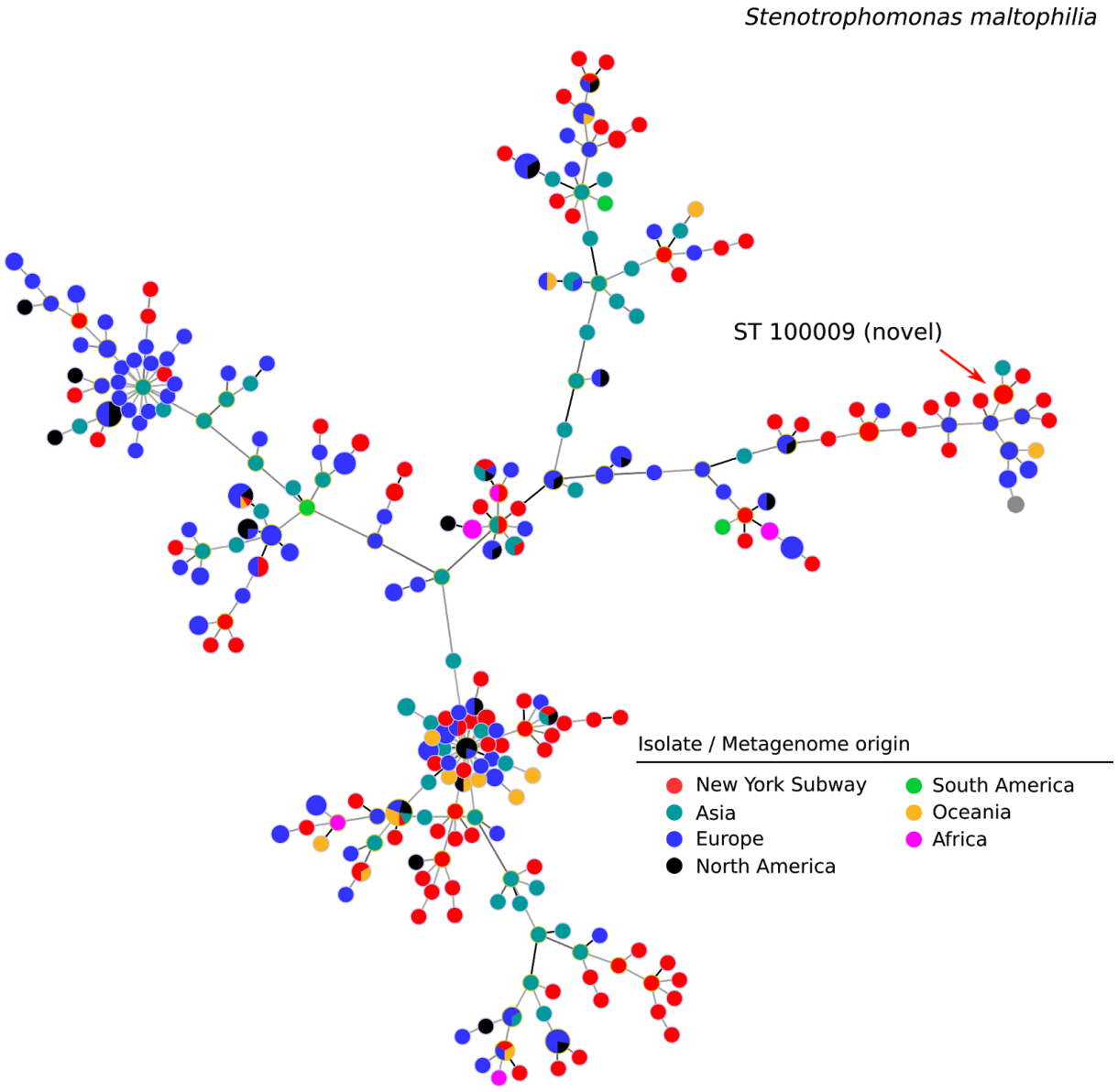
Supplementary Table 6.1: MetaPhlAn2 output table on the whole MetaSub dataset. Values represent the relative abundances detected for each sample (in the columns). Table available here: https://biologydirect.biomedcentral.com/articles/10.1186/s13062-018-0211-z#Sec18



Supplementary Figure 6.1: *Scatterplot contrasting for each sample the number of successfully profiled species against the metagenome size (in million reads). Each dot corresponds to a sample in the MetaSUB dataset. The number of detected species was calculated with MetaPhlAn2 by requiring a species to have a relative abundance higher than 0.5% within the sample.*



Supplementary Figure 6.2. Application of MetaMLST to detect E. cloacae in the 1614 urban metagenomes of the MetaSUB dataset. *The minimum spanning trees (MST) was generated on the basis of the allelic profile (57), where each node in the MST represents a Sequence Type (ST) and an edge connects similar STs (i.e. sharing at least one identical locus) with a length proportional to their allelic-profiles similarity. The two MSTs were built with PhyloViz (58). The 124 detected STs of E. cloacae (detected exclusively in the New York Subway, red dots) are placed in the tree together with 50 STs derived from 71 reference genomes of E. cloacae, isolated from different sources (other-colour nodes). The most abundant E. cloacae Sequence Types in the MetaSub cohort (STs 50 and ST 148) are indicated with a red arrow. The size of each node scales with the frequency of detection. When an ST is found in more than one continent, the samples' origins are represented through a coloured pie-chart.*



Supplementary Figure 6.3. Application of MetaMLST to detect S. maltophilia in the 1614 urban metagenomes of the MetaSUB dataset. *The minimum spanning trees (MST) was generated on the basis of the allelic profile (57), where each node in the MST represents a Sequence Type (ST) and an edge connects similar STs (i.e. sharing at least one identical locus) with a length proportional to their allelic-profiles similarity. The two MSTs were built with PhyloViz (58). The 102 detected STs of E. cloacae (detected exclusively in the New York Subway, red dots) are placed in the tree together with the available known STs from BIGSdb. Nodes are coloured according to the continent of origin, while New York Subway samples are represented in red. The most abundant Sequence Type in the MetaSub cohort (ST 100009, a novel ST) is indicated with a red arrow. The size of each node scales with the frequency of detection. When an ST is found in more than one continent, the samples' origins are represented through a coloured pie-chart.*

Acknowledgements and Declarations

Ethics approval and consent to participate / Consent for publication
Not applicable (the analyzed dataset was already publicly available)

List of abbreviations
Single nucleotide variant: SNV
MLST: Multi Locus Sequence Typing
Sequence Type: ST
Minimum Spanning tree: MST
Inter Quantile Range: IQR
Principal Component Analysis: PCA

Availability of data and material
The whole dataset is available at the NCBI BioProject accession numbers PRJNA301589 and PRJNA271013. The computational methods used are referenced in the manuscript.

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported by the European Research Council (ERC-STG project MetaPG), the European Union Framework Program 7 Marie-Curie grant (PCIG13-618833) and the MIUR (FIR RBFR13EWWI) to N.S.

Authors' contributions
M.Z and N.S. planned the experiment; M.Z., P.M, and E.P. conducted the analysis; M.Z., F.A., A.T., and N.S. interpreted the results and wrote the manuscript.

Acknowledgements
The authors would like to thank the members of the Computational Metagenomic Laboratory, Chris Mason and the MetaSUB and CAMDA initiative for their support and feedback.

Reviewers' comments
Reviewers’ reports are available in the published version of the article.

References

1. Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

2. S. Sunagawa, *et al.*, Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

3. C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, N. Segata, Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

4. J. Qin, *et al.*, A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).

5. S. Witkin, Faculty of 1000 evaluation for Vaginal microbiome of reproductive-age women. *F1000 - Post-publication peer review of the biomedical literature* (2010) https:/doi.org/10.3410/f.3689957.3413055.

6. K. Findley, D. R. Williams, E. A. Grice, V. L. Bonham, Health Disparities and the Microbiome. *Trends Microbiol.* **24**, 847–850 (2016).

7. J. L. Round, S. K. Mazmanian, The gut microbiota shapes intestinal immune responses during health and disease. *Nat. Rev. Immunol.* **9**, 313–323 (2009).

8. C. T. Brown, *et al.*, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

9. J. C. Venter, *et al.*, Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).

10. N. Fierer, *et al.*, Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J.* **6**, 1007–1017 (2012).

11. G. W. Tyson, *et al.*, Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).

12. T. Hsu, *et al.*, Urban Transit System Microbial Communities Differ by Surface Type and Interaction with Humans and the Environment. *mSystems* **1** (2016).

13. E. Afshinnekoo, *et al.*, Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Syst* **1**, 97–97.e3 (2015).

14. G. M. King, Urban microbiomes and urban ecology: how do microbes in the built environment affect human sustainability in cities? *J. Microbiol.* **52**, 721–728 (2014).

15. S. W. Kembel, *et al.*, Architectural design influences the diversity and structure of the built environment microbiome. *ISME J.* **6**, 1469–1479 (2012).

16. A. J. Prussin 2nd, L. C. Marr, Sources of airborne microorganisms in the built environment. *Microbiome* **3**, 78 (2015).

17. D. Hospodsky, *et al.*, Human occupancy as a source of indoor airborne bacteria. *PLoS One* **7**, e34867 (2012).

18. G. E. Flores, *et al.*, Diversity, distribution and sources of bacteria in residential kitchens. *Environ. Microbiol.* **15**, 588–596 (2013).

19. J. F. Meadow, *et al.*, Humans differ in their personal microbial cloud. *PeerJ* **3**, e1258 (2015).

20. J. E. Cohen, Human population: the next half century. *Science* **302**, 1172–1175 (2003).

21. ,*World Urbanization Prospects: The 2014 Revision* (UN, 2015).

22. J. E. Cohen, Population and climate change. *Proc. Am. Philos. Soc.* **154**, 158–182 (2010).

23. J. L. Gardy, N. J. Loman, Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* **19**, 9–20 (2018).

24. R. R. Miller, V. Montoya, J. L. Gardy, D. M. Patrick, P. Tang, Metagenomics for pathogen detection in public health. *Genome Med.* **5**, 81 (2013).

25. MetaSUB International Consortium, The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. *Microbiome* **4**, 24 (2016).

26. A. Bankevich, *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

27. D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

28. D. D. Kang, J. Froula, R. Egan, Z. Wang, MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

29. J. Alneberg, *et al.*, Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).

30. A. M. Eren, *et al.*, Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).

31. P. I. Costea, *et al.*, metaSNV: A tool for metagenomic strain level analysis. *PLoS One* **12**, e0182392 (2017).

32. C. Quince, *et al.*, DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18**, 181 (2017).

33. M. Scholz, *et al.*, Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* **13**, 435–438 (2016).

34. D. T. Truong, A. Tett, E. Pasolli, C. Huttenhower, N. Segata, Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).

35. M. Zolfo, A. Tett, O. Jousson, C. Donati, N. Segata, MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. *Nucleic Acids Res.* **45**, e7 (2017).

36. F. Asnicar, *et al.*, Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2** (2017).

37. J. Lloyd-Price, *et al.*, Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).

38. S. S. Li, *et al.*, Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* **352**, 586–589 (2016).

39. I. Sharon, *et al.*, Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* **23**, 111–120 (2013).

40. E. P. Price, *et al.*, Improved multilocus sequence typing of Burkholderia pseudomallei and closely related species. *J. Med. Microbiol.* **65**, 992–997 (2016).

41. L. Diancourt, V. Passet, J. Verhoef, P. A. D. Grimont, S. Brisse, Multilocus sequence typing of Klebsiella pneumoniae nosocomial isolates. *J. Clin. Microbiol.* **43**, 4178–4182 (2005).

42. M. C. Maiden, *et al.*, Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 3140–3145 (1998).

43. K. A. Jolley, M. C. J. Maiden, BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595 (2010).

44. D. M. Aanensen, B. G. Spratt, The multilocus sequence typing network: mlst.net. *Nucleic Acids Res.* **33**, W728–33 (2005).

45. P. Baumann, Isolation of Acinetobacter from soil and water. *J. Bacteriol.* **96**, 39–42 (1968).

46. L. Dijkshoorn, A. Nemec, H. Seifert, An increasing threat in hospitals: multidrug-resistant Acinetobacter baumannii. *Nat. Rev. Microbiol.* **5**, 939–951 (2007).

47. S. Santajit, N. Indrawattana, Mechanisms of Antimicrobial Resistance in ESKAPE Pathogens. *Biomed Res. Int.* **2016**, 2475067 (2016).

48. P. Gerner-Smidt, I. Tjernberg, J. Ursing, Reliability of phenotypic tests for identification of Acinetobacter species. *J. Clin. Microbiol.* **29**, 277–282 (1991).

49. H. Khayat, N. Sadeghifard, I. Pakzad, L. Azimi, S. Delfani, Determination of Different Fluoroquinolone Mechanisms Among Clinical Isolates of Acinetobacter baumannii in Tehran, Iran (2017).

50. L. Diancourt, V. Passet, A. Nemec, L. Dijkshoorn, S. Brisse, The population structure of Acinetobacter baumannii: expanding multiresistant clones from an ancestral susceptible genetic pool. *PLoS One* **5**, e10034 (2010).

51. J. W. Sahl, *et al.*, Evolution of a pathogen: a comparative genomics analysis identifies a genetic pathway to pathogenesis in Acinetobacter. *PLoS One* **8**, e54287 (2013).

52. D. Girlich, L. Poirel, P. Nordmann, Clonal distribution of multidrug-resistant Enterobacter cloacae. *Diagn. Microbiol. Infect. Dis.* **81**, 264–268 (2015).

53. J. S. Brooke, Stenotrophomonas maltophilia: an emerging global opportunistic pathogen. *Clin. Microbiol. Rev.* **25**, 2–41 (2012).

54. J. R. Johnson, A. L. Stell, Extended virulence genotypes of Escherichia coli strains from patients with urosepsis in relation to phylogeny and host compromise. *J. Infect. Dis.* **181**, 261–272 (2000).

55. B. Picard, *et al.*, The link between phylogeny and virulence in Escherichia coli extraintestinal infection. *Infect. Immun.* **67**, 546–553 (1999).

56. L. Micenková, J. Bosák, M. Vrba, A. Ševčíková, D. Šmajs, Human extraintestinal pathogenic Escherichia coli strains differ in prevalence of virulence factors, phylogroups, and bacteriocin determinants. *BMC Microbiol.* **16**, 218 (2016).

57. A. P. Francisco, M. Bugalho, M. Ramirez, J. A. Carriço, Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics* **10**, 152 (2009).

58. M. Nascimento, *et al.*, PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. *Bioinformatics* **33**, 128–129 (2017).

59. S. Adams-Sapper, B. A. Diep, F. Perdreau-Remington, L. W. Riley, Clonal composition and community clustering of drug-susceptible and -resistant Escherichia coli isolates from bloodstream infections. *Antimicrob. Agents Chemother.* **57**, 490–497 (2013).

60. S. Y. Tartof, O. D. Solberg, A. R. Manges, L. W. Riley, Analysis of a uropathogenic Escherichia coli clonal group by multilocus sequence typing. *J. Clin. Microbiol.* **43**, 5860–5864 (2005).

61. D. T. Truong, *et al.*, MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).

62. C. Urzì, *et al.*, Biodiversity of Geodermatophilaceae isolated from altered stones and monuments in the Mediterranean basin. *Environ. Microbiol.* **3**, 471–479 (2001).

63. K.-H. Kim, W.-T. Im, S.-T. Lee, Hymenobacter soli sp. nov., isolated from grass soil. *Int. J. Syst. Evol. Microbiol.* **58**, 941–945 (2008).

64. D. A. Stevens, J. R. Hamilton, N. Johnson, K. K. Kim, J.-S. Lee, Halomonas, a newly recognized human pathogen causing infections and contamination in a dialysis center: three new species. *Medicine* **88**, 244–249 (2009).

65. W.-J. Li, *et al.*, Kocuria aegyptia sp. nov., a novel actinobacterium isolated from a saline, alkaline desert soil in Egypt. *Int. J. Syst. Evol. Microbiol.* **56**, 733–737 (2006).

66. S. B. Kim, *et al.*, Kocuria marina sp. nov., a novel actinobacterium isolated from marine sediment. *Int. J. Syst. Evol. Microbiol.* **54**, 1617–1620 (2004).

67. F. Altuntas, *et al.*, Catheter-related bacteremia due to Kocuria rosea in a patient undergoing peripheral blood stem cell transplantation. *BMC Infect. Dis.* **4**, 62 (2004).

68. G. Berg, N. Roskot, K. Smalla, Genotypic and phenotypic relationships between clinical and environmental isolates of Stenotrophomonas maltophilia. *J. Clin. Microbiol.* **37**, 3594–3600 (1999).

69. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

70. S. Greenblum, R. Carr, E. Borenstein, Extensive strain-level copy-number variation across human gut microbiome species. *Cell* **160**, 583–594 (2015).

71. C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, N. Segata, Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

72. J. O. McInerney, A. McNally, M. J. O'Connell, Why prokaryotes have pangenomes. *Nat Microbiol* **2**, 17040 (2017).

73. H. Tettelin, *et al.*, Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13950–13955 (2005).

74. D. Medini, C. Donati, H. Tettelin, V. Masignani, R. Rappuoli, The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594 (2005).

75. D. V. Ward, *et al.*, Metagenomic Sequencing with Strain-Level Resolution Implicates Uropathogenic E. coli in Necrotizing Enterocolitis and Mortality in Preterm Infants. *Cell Rep.* **14**, 2912–2924 (2016).

76. N. J. Loman, *et al.*, A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. *JAMA* **309**, 1502–1510 (2013).

77. U. Tattawasart, J. Y. Maillard, J. R. Furr, A. D. Russell, Outer membrane changes in Pseudomonas stutzeri resistant to chlorhexidine diacetate and cetylpyridinium chloride. *Int. J. Antimicrob. Agents* **16**, 233–238 (2000).

78. M. Papapetropoulou, J. Iliopoulou, G. Rodopoulou, J. Detorakis, O. Paniara, Occurrence and

antibiotic-resistance of Pseudomonas species isolated from drinking water in southern Greece. *J. Chemother.* **6**, 111–116 (1994).

79. C. J. Papadopoulos, C. F. Carson, B. J. Chang, T. V. Riley, Role of the MexAB-OprM efflux pump of Pseudomonas aeruginosa in tolerance to tea tree (Melaleuca alternifolia) oil and its monoterpene components terpinen-4-ol, 1,8-cineole, and alpha-terpineol. *Appl. Environ. Microbiol.* **74**, 1932–1935 (2008).

80. H. Chalhoub, *et al.*, High-level resistance to meropenem in clinical isolates of Pseudomonas aeruginosa in the absence of carbapenemases: role of active efflux and porin alterations. *Int. J. Antimicrob. Agents* **48**, 740–743 (2016).

81. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

82. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

83. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

84. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

85. A. J. Page, *et al.*, Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

86. P. J. A. Cock, *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

87. E. Talevich, B. M. Invergo, P. J. A. Cock, B. A. Chapman, Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics* **13**, 209 (2012).

88. F. Asnicar, G. Weingart, T. L. Tickle, C. Huttenhower, N. Segata, Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029 (2015).

89. P. A. Kitts, *et al.*, Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* **44**, D73–80 (2016).

90. J. D. Hunter, Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

Chapter 7

Other contributions

7 | Other Contributions

In this chapter are reported my minor contributions to other research papers. A brief paragraph indicates, for each paper or group of papers, the context and my contribution to the works. Papers are grouped by topic and numbered sequentially.

7.1 | Profiling the increase in mappability of metagenomic reads with Metagenomic Assembly Genomes

The article from Pasolli *et al.* describes the retrieval of thousands of metagenomic assembled genomes (MAGs) from public datasets. Many of these genomes were previously unknown and represented entirely novel species. I mainly contributed to the assessment of the increase in the metagenomic "read mappability" granted by these new sequences. I designed and conducted the computational experiments to determine the fraction of reads that could be mapped with and without the MAGs in the 9,428 samples used in the study. This is reported in **Fig. 2** and **Supplementary Fig. 4** of the original paper (i.e. mappability from an average of 67.76% to 87.51% in the gut). I also contributed to other aspects of the sequence analysis reported in the main figures. Moreover, the MAGs of this paper have been pivotal in the experiments described in **Chapter 4** of this thesis.

I performed the same experimental setup to test the increment in read-mappability with MAGs from non-human primates in the article from Manara *et al.*

Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.

Pasolli, E., Asnicar, F.*, Manara, S.* , Zolfo, M.*, Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., Quince, C., Huttenhower, C., Segata, N.

* Equal Contribution | ***Cell*** 176, Issue 3, pp 649-662.E20 | https://doi.org/10.1016/j.cell.2019.01.001

Abstract. The body-wide human microbiome plays a role in health, but its full diversity remains uncharacterized, particularly outside of the gut and in international populations. We leveraged 9,428 metagenomes to reconstruct 154,723 microbial genomes (45% of high quality) spanning body sites, ages, countries, and lifestyles. We recapitulated 4,930 species-level genome bins (SGBs), 77% without genomes in public repositories (unknown SGBs [uSGBs]). uSGBs are prevalent (in 93% of well-assembled samples), expand underrepresented phyla, and are enriched in non-Westernized populations (40% of the total SGBs). We annotated 2.85 M genes in SGBs, many associated with conditions including infant development (94,000) or Westernization (106,000). SGBs and uSGBs permit deeper microbiome analyses and increase the average mappability of metagenomic reads from 67.76% to 87.51% in the gut (median 94.26%) and 65.14% to 82.34% in the mouth. We thus identify thousands of microbial genomes from yet-to-be-named species, expand the pangenomes of human-associated microbes, and allow better exploitation of metagenomic technologies.

Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species

Manara, S., Asnicar, F., Beghini, F., Bazzani, D., Cumbo, F., Zolfo, M., Nigro, E., Karcher, N., Manghi, P., Metzger, M. I., Pasolli, E.,Segata, N.

Genome Biology 20, 299 (2019) | https://doi.org/10.1186/s13059-019-1923-9

Abstract

Background.

Humans have coevolved with microbial communities to establish a mutually advantageous relationship that is still poorly characterized and can provide a better understanding of the human microbiome. Comparative metagenomic analysis of human and non-human primate (NHP) microbiomes offers a promising approach to study this symbiosis. Very few microbial species have been characterized in NHP microbiomes due to their poor representation in the available cataloged microbial diversity, thus limiting the potential of such comparative approaches.

Results.

We reconstruct over 1000 previously uncharacterized microbial species from 6 available NHP metagenomic cohorts, resulting in an increase of the mappable fraction of metagenomic reads by 600%. These novel species highlight that almost 90% of the microbial diversity associated with NHPs has been overlooked. Comparative analysis of this new catalog of taxa with the collection of over 150,000 genomes from human metagenomes points at a limited species-level overlap, with only 20% of microbial candidate species in NHPs also found in the human microbiome. This overlap occurs mainly between NHPs and non-Westernized human populations and NHPs living in captivity, suggesting that host lifestyle plays a role comparable to host speciation in shaping the primate intestinal microbiome. Several NHP-specific species are phylogenetically related to human-associated microbes, such as Elusimicrobia and *Treponema*, and could be the consequence of host-dependent evolutionary trajectories.

Conclusions.

The newly reconstructed species greatly expand the microbial diversity associated with NHPs, thus enabling better interrogation of the primate microbiome and empowering in-depth human and non-human comparative and co-diversification studies.

7.2 | Vertical Transmission of Bacteria and Phages from Mother to Infant

In the following papers, I contributed to analyze the bacteria and phages that were found both in mothers and their infants. The studies aimed to profile individual strains shared between mother and infant by analyzing the gut microbiome of newborns in the days and weeks following birth. All studies showed that part of the microbiome of newborns is acquired by their mother, and that transmitted species are more likely to persist into the newborn's gut. In Asnicar and Manara *et al.* I analyzed the viral communities in the metatranscriptomics of couples of mothers and infants and profiled the extremely abundant RNA plant virus Pepper Mild Mottle Virus in one couple. The virus was so abundant that its sequence could be completely reconstructed directly from the raw reads (see **Fig.2** of the original paper). I applied the same computational approach to reconstruct and analyze the phylogenies of abundant phages of bifidobacteria in Duranti *et al.*

In Ferretti *et al.* I contributed to the strain-level analysis of microbial species in mothers and infants and analyzed the metagenomes with PanPhlAn, a computational tool to profile microbes via their gene content.

Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome

Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D. T., Manara, S., Zolfo, M. Beghini, F., Bertorelli, R., De Sanctis, V., Bariletti, I., Canto, R., Clementi, R., Cologna, M., Crifò, T., Cusumano, G., Gottardi, S., Innamorati, C., Masè, C., Postai, D., Savoi, D., Duranti, S., Lugli, G. A., Mancabelli, L., Turroni, F., Ferrario, C., Milani, C., Mangifesta, M., Anzalone, R., Viappiani, A., Yassour, M., Vlamakis, H., Xavier, R., Collado, C. M., Koren, O., Tateo, S., Soffiati, M., Pedrotti, A., Ventura, M., Huttenhower, C., Bork, P., Segata, N.

Cell Host & Microbe Volume 24, Issue 1 p 133-145.e5 (2018) | https://doi.org/10.1016/j.chom.2018.06.005

Abstract

The acquisition and development of the infant microbiome are key to establishing a healthy host-microbiome symbiosis. The maternal microbial reservoir is thought to play a crucial role in this process. However, the source and transmission routes of the infant pioneering microbes are poorly understood. To address this, we longitudinally sampled the microbiome of 25 mother-infant pairs across multiple body sites from birth up to 4 months postpartum. Strain-level metagenomic profiling showed a rapid influx of microbes at birth followed by strong selection during the first few days of life. Maternal skin and vaginal strains colonize only transiently, and the infant continues to acquire microbes from distinct maternal sources after birth. Maternal gut strains proved more persistent in the infant gut and ecologically better adapted than those acquired from other sources. Together, these data describe the mother-to-infant microbiome transmission routes that are integral in the development of the infant microbiome.

Maternal inheritance of bifidobacterial communities and bifidophages in infants through vertical transmission.

Duranti, S., Lugli, G. A., Mancabelli, L., Armanini, F., Turroni, F., James, K., Ferretti, P., Gorfer, V., Ferrario, C., Milani, C., Mangifesta, M., Anzalone, R., Zolfo, M., Viappiani, A., Pasolli, E., Bariletti, I., Canto, R., Clementi, R., Cologna, M., Crifò, T., Cusumano, G., Fedi, S., Gottardi, S., Innamorati, C., Masè, C., Postai, D., Savoi, D., Soffiati, M., Tateo, S., Pedrotti, A., Segata, N., van Sinderen, D.,Ventura, M.

Microbiome 5, 66 (2017) | https://doi.org/10.1186/s40168-017-0282-6

Abstract

Background

The correct establishment of the human gut microbiota represents a crucial development that commences at birth. Different hypotheses propose that the infant gut microbiota is derived from, among other sources, the mother's fecal/vaginal microbiota and human milk.

Results

The composition of bifidobacterial communities of 25 mother-infant pairs was investigated based on an internal transcribed spacer (ITS) approach, combined with cultivation-mediated and genomic analyses. We identified bifidobacterial strains/communities that are shared between mothers and their corresponding newborns. Notably, genomic analyses together with growth profiling assays revealed that bifidobacterial strains that had been isolated from human milk are genetically adapted to utilize human milk glycans. In addition, we identified particular bacteriophages specific of bifidobacterial species that are common in the viromes of mother and corresponding child.

Conclusions

This study highlights the transmission of bifidobacterial communities from the mother to her child and implies human milk as a potential vehicle to facilitate this acquisition. Furthermore, these data represent the first example of maternal inheritance of bifidobacterial phages, also known as bifidophages in infants following a vertical transmission route.

Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling.

Asnicar, F. *, Manara, S. *, Zolfo, M., Truong, D. T., Scholz, M., Armanini, F., Ferretti, P., Gorfer, V., Pedrotti, A., Tett, A.,Segata, N. |

* Equal Contribution

mSystems 2:e00164-16 (2017) | https://doi.org/10.1128/mSystems.00164-16

Abstract

The gut microbiome becomes shaped in the first days of life and continues to increase its diversity during the first months. Links between the configuration of the infant gut microbiome and infant health are being shown, but a comprehensive strain-level assessment of microbes vertically transmitted from mother to infant is still missing. We collected fecal and breast milk samples from multiple mother-infant pairs during the first year of life and applied shotgun metagenomic sequencing followed by computational strain-level profiling. We observed that several specific strains, including those of Bifidobacterium bifidum, Coprococcus comes, and Ruminococcus bromii, were present in samples from the same mother-infant pair, while being clearly distinct from those carried by other pairs, which is indicative of vertical transmission. We further applied metatranscriptomics to study the in vivo gene expression of vertically transmitted microbes and found that transmitted strains of Bacteroides and Bifidobacterium species were transcriptionally active in the guts of both adult and infant. By combining longitudinal microbiome sampling and newly developed computational tools for strain-level microbiome analysis, we demonstrated that it is possible to track the vertical transmission of microbial strains from mother to infants and to characterize their transcriptional activity. Our work provides the foundation for larger-scale surveys to identify the routes of vertical microbial transmission and its influence on post infancy microbiome development.

Importance. Early infant exposure is important in the acquisition and ultimate development of a healthy infant microbiome. There is increasing support for the idea that the maternal microbial reservoir is a key route of microbial transmission, and yet much is inferred from the observation of shared species in mother and infant. The presence of common species, per se, does not necessarily equate to vertical transmission, as species exhibit considerable strain heterogeneity. It is therefore imperative to assess whether shared microbes belong to the same genetic variant (i.e., strain) to support the hypothesis of vertical transmission. Here we demonstrate the potential of shotgun metagenomics and strain-level profiling to identify vertical transmission events. Combining these data with metatranscriptomics, we show that it is possible not only to identify and track the fate of microbes in the early infant microbiome but also to investigate the actively transcribing members of the community. These approaches will ultimately provide important insights into the acquisition, development, and community dynamics of the infant microbiome.

7.3 | Strain-level analysis of known and unknown microbes

In the following papers, I contributed by performing strain-level profiling on different sample types. In particular, Tett *et al.* was the first community-wide metagenomics characterization of the microbiome associated with plaque psoriasis. In this study, I applied MetaMLST, a computational tool to perform an *in-silico* Multi-Locus Sequence Typing analysis from metagenomics I developed prior to the beginning of my Ph.D., to profile strains of *Cutibacterium acnes* (previously *Propionibacterium acnes)* and *Staphylococcus epidermidis*.

Similarly, in Asnicar *et al.* I applied MetaMLST to profile 208 *Escherichia coli* genomes and Metagenomic Assembled Genomes. The analysis allowed to confirm that a novel computational tools for phylogenetic analysis and placement was in agreement with the strain profiles reconstructed independently by other tools.

In Qin *et al.* I performed the strain-resolved phylogenetic analysis on several human-associated microbes retrieved in metagenomes from PM2.5 and PM10 air filters of Beijing. The analysis revealed the diversity of the human, potentially airborne, strains of, among others, *Acinetobacter lwoffii* and *Kocuria sp.*

Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis.

Tett, A., Pasolli *, E., Farina, S. *, Truong, D. T *., Asnicar, F., Zolfo, M., Beghini, F., Armanini, F., Jousson, O., De Sanctis, V., Bertorelli, R., Girolomoni, G., Cristofolini, M., Segata, N.

* Equal contribution

npj Biofilms Microbiomes 3, 14 (2017) | https://doi.org/10.1038/s41522-017-0022-5

Abstract
Psoriasis is an immune-mediated inflammatory skin disease that has been associated with cutaneous microbial dysbiosis by culture-dependent investigations and rRNA community profiling. We applied, for the first time, high-resolution shotgun metagenomics to characterise the microbiome of psoriatic and unaffected skin from 28 individuals. We demonstrate psoriatic ear sites have a decreased diversity and psoriasis is associated with an increase in Staphylococcus, but overall the microbiomes of psoriatic and unaffected sites display few discriminative features at the species level. Finer strain-level analysis reveals strain heterogeneity colonisation and functional variability providing the intriguing hypothesis of psoriatic niche-specific strain adaptation or selection. Furthermore, we accessed the poorly characterised, but abundant, clades with limited sequence information in public databases, including uncharacterised Malassezia spp. These results highlight the skins hidden diversity and suggests strain-level variations could be key determinants of the psoriatic microbiome. This illustrates the need for high-resolution analyses, particularly when identifying therapeutic targets. This work provides a baseline for microbiome studies in relation to the pathogenesis of psoriasis.

Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0.

Asnicar, F., Thomas, A., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan, M., Cumbo, F., Ms. M., Sanders, J., Zolfo, M., Kopylova, E., Pasolli, E., Knight, R., Mirarab, S., Huttenhower, C., Segata, N.

Nature Communications 11, 2500 (2020) | https://doi.org/10.1038/s41467-020-16366-7

Abstract
Microbial genomes are available at an ever-increasing pace, as cultivation and sequencing become cheaper and obtaining metagenome-assembled genomes (MAGs) becomes more effective. Phylogenetic placement methods to contextualize hundreds of thousands of genomes must thus be efficiently scalable and sensitive from closely related strains to divergent phyla. We present PhyloPhlAn 3.0, an accurate, rapid, and easy-to-use method for large-scale microbial genome characterization and phylogenetic analysis at multiple levels of resolution. PhyloPhlAn 3.0 can assign genomes from isolate sequencing or MAGs to species-level genome bins built from >230,000 publically available sequences. For individual clades of interest, it reconstructs strain-level phylogenies from among the closest species using clade-specific maximally informative markers. At the other extreme of resolution, it scales to large phylogenies comprising >17,000 microbial species. Examples including Staphylococcus aureus isolates, gut metagenomes, and meta-analyses demonstrate the ability of PhyloPhlAn 3.0 to support genomic and metagenomic analyses.

Longitudinal survey of microbiome associated with particulate matter in a megacity

Qin, N. *, Liang, P .*, Wu, C. *, Wang, G., Xu, Q., Xiong, X., Wang, T., Zolfo, M., Segata, N., Qin, H., Knight, R., Gilbert, J. A., Zhu, T. F.

* Equal contribution | ***Genome Biology*** 21, 55 (2020) | https://doi.org/10.1186/s13059-020-01964-x

Abstract
Background | While the physical and chemical properties of airborne particulate matter (PM) have been extensively studied, their associated microbiome remains largely unexplored. Here, we performed a longitudinal metagenomic survey of 106 samples of airborne PM2.5 and PM10 in Beijing over a period of 6 months in 2012 and 2013, including those from several historically severe smog events.
Results | We observed that the microbiome composition and functional potential were conserved between PM2.5 and PM10, although considerable temporal variations existed. Among the airborne microorganisms, *Propionibacterium acnes*, *Escherichia coli*, *Acinetobacter lwoffii*, *Lactobacillus amylovorus*, and *Lactobacillus reuteri* dominated, along with several viral species. We further identified an extensive repertoire of genes involved in antibiotic resistance and detoxification, including transporters, transpeptidases, and thioredoxins. Sample stratification based on Air Quality Index (AQI) demonstrated that many microbial species, including those associated with human, dog, and mouse feces, exhibit AQI-dependent incidence dynamics. The phylogenetic and functional diversity of air microbiome is comparable to those of soil and water environments, as its composition likely derives from a wide variety of sources.
Conclusions | Airborne particulate matter accommodates rich and dynamic microbial communities, including a range of microbial elements that are associated with potential health consequences.

7.4 | Sequence Analysis of unknown strains

In the following papers, I contributed with the development, maintenance, and execution of a software suite named CMSeq. CMSeq is a set of highly customizable APIs based on HTS-lib and PySam that allow to compute the breadth and depth of coverage of contigs with custom thresholds and quality filtering parameters.

I also adapted CMSeq to calculate the polymorphism rate of many novel Metagenomic Assembled Contigs (MAGs). In short, this is the measure of the likelihood that a MAG is composed of an admixture of sequences of more than one strain.

Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations

Karcher, N., Pasolli, E., Asnicar, F., Huang, K., Tett, A., Manara, S., Armanini, F., Bain, D., Duncan, S., Louis, P., Zolfo, M., Manghi, P., Valles-Colomer, M., Raffaeta, R., Rota-Stabelli, O., Carmen Collado, M., Zeller, G., Falush, D., Maixner, F., Walker, A., Huttenhower, C., Segata, N.

Genome Biology volume 21, Article: 138 (2020) | https://doi.org/10.1186/s13059-020-02042-y

Abstract

Background

Eubacterium rectale is one of the most prevalent human gut bacteria, but its diversity and population genetics are not well understood because large-scale whole-genome investigations of this microbe have not been carried out.

Results

Here, we leverage metagenomic assembly followed by a reference-based binning strategy to screen over 6500 gut metagenomes spanning geography and lifestyle and reconstruct over 1300 *E. rectale* high-quality genomes from metagenomes. We extend previous results of biogeographic stratification, identifying a new subspecies predominantly found in African individuals and showing that closely related non-human primates do not harbor *E. rectale*. Comparison of pairwise genetic and geographic distances between subspecies suggests that isolation by distance and co-dispersal with human populations might have contributed to shaping the contemporary population structure of *E. rectale*. We confirm that a relatively recently diverged *E. rectale* subspecies specific to Europe consistently lacks motility operons and that it is immotile in vitro, probably due to ancestral genetic loss. The same subspecies exhibits expansion of its carbohydrate metabolism gene repertoire including the acquisition of a genomic island strongly enriched in glycosyltransferase genes involved in exopolysaccharide synthesis.

Conclusions

Our study provides new insights into the population structure and ecology of *E. rectale* and shows that shotgun metagenomes can enable population genomics studies of microbiota members at a resolution and scale previously attainable only by extensive isolate sequencing.

The *Prevotella copri* Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations

Tett, A., Huang, K. D., Asnicar, F., Fehlner-Peach, H., Pasolli, E., Karcher, N., Armanini, F., Manghi, P., Bonham, K., Zolfo, M., De Filippis, F., Magnabosco, C., Bonneau, R., Lusingu, J., Amuasi, J., Reinhard, K., Rattei, T., Boulund, F., Engstrand, L., Zink, A., Collado, M. C., Littman, D. R., Eibach, D., Ercolini, D., Rota-Stabelli, O., Huttenhower, C., Maixner, F.,Segata, N.

Cell Host & Microbe Volume 26, Issue 5 p 666-679.e7 (2019) https://doi.org/10.1016/j.chom.2019.08.018

Abstract | *Prevotella copri* is a common human gut microbe that has been both positively and negatively associated with host health. In a cross-continent meta-analysis exploiting >6,500 metagenomes, we obtained >1,000 genomes and explored the genetic and population structure of *P. copri*. *P. copri* encompasses four distinct clades (>10% inter-clade genetic divergence) that we propose constitute the *P. copri* complex, and all clades were confirmed by isolate sequencing. These clades are nearly ubiquitous and co-present in non-Westernized populations. Genomic analysis showed substantial functional diversity in the complex with notable differences in carbohydrate metabolism, suggesting that multi-generational dietary modifications may be driving reduced prevalence in Westernized populations. Analysis of ancient metagenomes highlighted patterns of *P. copri* presence consistent with modern non-Westernized populations and a clade delineation time pre-dating human migratory waves out of Africa. These findings reveal that *P. copri* exhibits a high diversity that is underrepresented in Western-lifestyle populations.

Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation.

Thomas, A.M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., Gandini, S., Serrano, D., Tarallo, S., Francavilla, A., Gallo, G., Trompetto, M., Ferrero, G., Mizutani, S., Shiroma, H., Shiba, S., Shibata, T., Yachida, S., Yamada, T., Wirbel, J., Schrotz-King, P., Ulrich, C. M., Brenner, H., Arumugam, M., Bork, P., Zeller, G., Cordero, F., Dias-Neto, E., Setubal, J. C., Tett, A., Pardini, B., Rescigno, M., Waldron, L., Naccarati, A., Segata, N.

Nature Medicine 25, 667–678 (2019). | https://doi.org/10.1038/s41591-019-0405-7

Abstract | Several studies have investigated links between the gut microbiome and colorectal cancer (CRC), but questions remain about the replicability of biomarkers across cohorts and populations. We performed a meta-analysis of five publicly available datasets and two new cohorts and validated the findings on two additional cohorts, considering in total 969 fecal metagenomes. Unlike microbiome shifts associated with gastrointestinal syndromes, the gut microbiome in CRC showed reproducibly higher richness than controls ($P < 0.01$), partially due to expansions of species typically derived from the oral cavity. Meta-analysis of the microbiome functional potential identified gluconeogenesis and the putrefaction and fermentation pathways as being associated with CRC, whereas the stachyose and starch degradation pathways were associated with controls. Predictive microbiome signatures for CRC trained on multiple datasets showed consistently high accuracy in datasets not considered for model training and independent validation cohorts (average area under the curve, 0.84). Pooled analysis of raw metagenomes showed that the choline trimethylamine-lyase gene was overabundant in CRC ($P = 0.001$), identifying a relationship between microbiome choline metabolism and CRC. The combined analysis of heterogeneous CRC cohorts thus identified reproducible microbiome biomarkers and accurate disease-predictive models that can form the basis for clinical prognostic tests and hypothesis-driven mechanistic studies.

Acknowledgments

It would not be the end of a thesis without my personal and deep acknowledgment of all the people who contributed to making all of this possible. First and foremost, I wish to thank my supervisor, Prof. Nicola Segata, whose inspiring guidance and optimistic enthusiasm were essential to my development as a young researcher. In the last years, I had the opportunity to work in an extraordinarily stimulating research environment. It has been a long road, but It was worth every step.

Thanks to all my colleagues at the Laboratory of Computational Metagenomics. Life in the lab would not have been the same without you, and every one of you taught me something new on this journey. I promised years ago that I would return most of the pens I borrowed… I am still working on that, but I can start to catch up with all the scrounged coffees. Maybe.

My deepest gratitude goes to the greatest people of my life: my parents Luisa and Pasquale, my brother Beniamino, and my super-grandmother Rosa (who still asks me what job I am doing exactly). I could not be where I am now without your support, care, and love. Words can hardly express how thankful I am, and how important your support has been in these years.

A gigantic acknowledgment goes to my friends Serena, Francesco A & B, Mattia, Giulia, Marco, Jean, Marianna, Wainer, Davide, Roberta, Marta, Lia, and Giuseppe. Some of you are close, others are farther. However, distance makes absolutely no difference in regards to the love I feel for all of you.

A special mention of gratitude goes to my friends who patiently helped me to review the thesis, by withstanding computational jargon at best and meaningless gibberish at worst, or by taming my circadian rhythm. Marianna, Serena, Wainer: your help, together with Marco's coffee, was priceless.

Thanks to Serena, Veronica, Giulia, and Orsetta; my board-mates and friends at the non-profit charity "RagionevolMente". I am so proud of what we did together, and nothing would have happened without you. Sure, we have got piles and piles of human hair stocked everywhere [1] and this is starting to be a bit creepy, but the rest is awesome.

To all of you, within range or far away, thank you. I could not have dreamed of better adventure companions and friends. You all contributed to make me a better scientist and most importantly a better person. Thank you, sincerely!

[1] No criminal activity involved.