Programa de Doctorado en Biología Molecular y Celular

RECOVERY AND CHARACTERIZATION OF VIRAL DIVERSITY FROM AQUATIC SHORT- AND LONG-READ METAGENOMES

Asier Zaragoza Solas

Director/a de la tesis
Dr. D. Francisco Eduardo Rodríguez Valera

Codirector/a de la tesis
Dr. D. Mario López Pérez

Universidad Miguel Hernández de Elche

- 2022 -

La siguiente tesis doctoral, titulada "**RECOVERY AND CHARACTERIZATION OF VIRAL DIVERSITY FROM AQUATIC SHORT- AND LONG-READ METAGENOMES** ", se presenta bajo la modalidad de tesis por compendio de las siguientes publicaciones:

- Zaragoza-Solas A, Rodriguez-Valera F, López-Pérez M. Metagenome Mining Reveals Hidden Genomic Diversity of Pelagimyophages in Aquatic Environments. mSystems. 2020 Feb 18;5(1):e00905-19. doi: 10.1128/mSystems.00905-19. PMID: 32071164.
- Zaragoza-Solas A, Haro-Moreno JM, Rodriguez-Valera F, López-Pérez M. Long-Read Metagenomics Improves the Recovery of Viral Diversity from Complex Natural Marine Samples. mSystems. 2022 Jun 13:e0019222. doi: 10.1128/msystems.00192-22. PMID: 35695508.
- Coutinho FH, Zaragoza-Solas A, López-Pérez M, Barylski J, Zielezinski A, Dutilh BE, Edwards R, Rodriguez-Valera F. RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. Patterns (N Y). 2021 Jun 15;2(7):100274. doi: 10.1016/j.patter.2021.100274. PMID: 34286299.

INFORME DEL DIRECTOR Y CODIRECTOR

FRANCISCO EDUARDO RODRÍGUEZ VALERA, Catedrático de Microbiología de la Universidad Miguel Hernández de Elche, en calidad de director, y MARIO LÓPEZ PÉREZ, Profesor Ayudante Doctor de la Universidad Miguel Hernández de Elche, en calidad de codirector;

INFORMAN:

Que D. Asier Zaragoza Solas ha realizado bajo nuestra supervisión la tesis titulada "RECOVERY AND CHARACTERIZATION OF VIRAL DIVERSITY FROM AQUATIC SHORT- AND LONG-READ METAGENOMES" conforme a los términos y condiciones definidos en su Plan de Investigación y de acuerdo con el Código de Buenas Prácticas de la Universidad Miguel Hernández de Elche, cumpliendo los objetivos previstos de forma satisfactoria para su defensa pública como tesis doctoral.

Lo que firmamos para los efectos oportunos, en San Juan de Alicante a 15 de julio de 2022

Director de la tesis

Codirector de la tesis

(Firma y nombre)

INFORME DEL COORDINADOR DEL PROGRAMA DE DOCTORADO

ASIA FERNÁNDEZ CARVAJAL, Coordinadora del Programa de Doctorado en Biología Molecular y Celular del Instituto de Investigación, Desarrollo e Innovación en Biotecnología Sanitaria (IDiBE) de la Universidad Miguel Hernández de Elche;

INFORMA:

Que D. Asier Zaragoza Solas ha realizado bajo nuestra supervisión la tesis titulada "RECOVERY AND CHARACTERIZATION OF VIRAL DIVERSITY FROM AQUATIC SHORT- AND LONG-READ METAGENOMES" conforme a los términos y condiciones definidos en su Plan de Investigación y de acuerdo con el Código de Buenas Prácticas de la Universidad Miguel Hernández de Elche, cumpliendo los objetivos previstos de forma satisfactoria para su defensa pública como tesis doctoral.

Lo que firmo para los efectos oportunos, en Elche a 15 de julio de 2022

Coordinador/a del programa de doctorado
(firma y nombre)

FINANCIACIÓN

Durante la realización de esta Tesis, el doctorando fue financiado por las “Ayudas para contratos predoctorales para la formación de doctores 2017” (BES-2017-079993) dentro del Programa Estatal de Promoción del Talento y su empleabilidad en I+D+i, Ministerio de Economía y Competitividad. A su vez, las investigaciones recopiladas en este trabajo fueron financiadas por el Ministerio de Economía y Competitividad (“FLEX3GEN” - PID2020-118052GB-I00; “VIREVO” - CGL2016-76273-P) y por la Generalitat Valenciana (“HIDRAS3” - PROMETEO/2019/009).

SUMMARY

Viruses are the most abundant biological entities in marine ecosystems and play an essential role in global biogeochemical cycles. They have important ecological functions as drivers of bacterial populations through lytic infections and contribute to bacterial genetic diversification. Unfortunately, their study is severely limited by the difficulty to culture and isolate them in lab conditions. Culture-independent techniques such as metagenomics can complement culture-based approaches to capture more phage diversity. However, the vast majority of viral sequences recovered through these methods are uncharacterized and therefore do not provide any information about their interactions with the bacterial community, a phenomenon that has been named “viral dark matter”.

In this thesis, several bioinformatic techniques are applied to both short- and long-read metagenomic datasets to recover biological information from marine viral sequences contained therein. A pipeline for recovering viral sequences based on a reference genome was developed and applied to the study of myophages infecting the alphaproteobacterial SAR11 clade, one of the most abundant bacterioplankton groups in surface marine and freshwater ecosystems. We were able to recover 22 new genomes which include the first genomes of myophages infecting LD12, the SAR11 freshwater clade. These sequences are underrepresented in datasets derived from the viral fraction, suggesting a bias of either technical or biological nature. Surprisingly, this family of phages code for an operon which resembles the secretion system type VIII operon in *Escherichia coli*. The function of this phage operon is still unknown.

Next, a long-read dataset from the Mediterranean Sea was explored for viral contigs to contrast phage recovery between long- and short-read datasets. The analysis revealed that while long-read assemblies resulted in viral sequences of better quality, there was a sizable amount of intra-clade viral diversity that was not included in the assemblies. This viral diversity only found in long reads is even greater than previously thought. This untapped diversity could aid biotechnological efforts as evidenced by the discovery of new endolysins.

Finally, a tool (Random Forest Assignment of Hosts, or RaFAH) for assigning hosts to phage sequences obtained from metagenomic datasets was created. The tool is based on a machine learning tool trained with phage protein clusters generated *de novo*. Benchmarking shows that RaFAH is on par with other state-of-the-art classifiers and is able to classify phage contigs at the level of Kingdom, which makes it the first classifier to accurately detect Archaea viruses from metagenomic samples. A feature importance analysis reveals that the protein clusters with the most predictive power are those involved in host recognition.

RESUMEN

Los bacteriófagos (”fagos”) son los organismos más abundantes en los ecosistemas marinos y tienen un papel esencial en los ciclos biogeoquímicos globales. Asimismo, influencian la evolución de las poblaciones bacterianas que infectan y contribuyen a la diversificación del acervo genético bacteriano. Desgraciadamente, su estudio se ve limitado por la dificultad de cultivar y aislar estos organismos en el laboratorio. El uso de técnicas que no requieren cultivo, como la metagenómica, pueden complementar el cultivo en laboratorio para recuperar una mayor diversidad de fagos. Sin embargo, la inmensa mayoría de secuencias virales recuperadas mediante metagenómica no pueden ser caracterizadas, por lo que no proporcionan ninguna información sobre sus interacciones con la comunidad bacteriana, un fenómeno que se ha nombrado “materia oscura viral”.

En esta tesis se han utilizado múltiples procesos bioinformáticos en colecciones de metagenomas de lectura corta y larga para caracterizar las secuencias virales que contienen. Se ha desarrollado un procedimiento para recuperar secuencias virales a partir de un genoma de referencia y se ha aplicado al estudio de miofagos que infectan al clado SAR11 de las Alfaproteobacteria, uno de los grupos de bacterioplankton más abundantes en agua dulce y agua salada de superficie. Se consiguió recuperar 22 nuevos genomas que incluyen el primer genoma que infecta LD12, el subclado de SAR11 de agua dulce. Estos genomas están poco representados en colecciones obtenidas de la fracción viral, lo que sugiere que las afecta un sesgo técnico o biológico. Sorprendentemente, esta familia de fagos contiene un operón similar al sistema de secreción tipo VIII de *Escherichia coli*. La función de este operón es aún desconocida.

Asimismo, se contrastó la recuperación de secuencias víricas entre colecciones de lectura corta y larga utilizando colecciones obtenidas en el mar Mediterráneo. Los resultados muestran que aunque los ensamblajes derivados de las lecturas largas producen secuencias virales de mejor calidad, en el proceso se pierde una gran cantidad de diversidad intraclado. Esta diversidad es mucho mayor de la recuperada con lecturas cortas, y podría explotarse para aplicaciones biotecnológicas, como el descubrimiento de nuevas endolisinas.

Finalmente, se desarrolló un programa (Random Forest Assignment of Hosts, o RaFAH) para asignar hospedadores a secuencias virales obtenidas de colecciones metagenómicas. El programa se basa en el uso de algoritmos de *machine learning* entrenados con grupos de proteínas creados *de novo*. RaFAH muestra un rendimiento similar a otros clasificadores de secuencias y es capaz de clasificar secuencias víricas al nivel taxonómico de Reino, siendo así el primer clasificador capaz de detectar fagos que infectan arqueas con precisión. El análisis de importancia de rasgo revela que los grupos de proteínas con mayor poder predictivo son aquellos involucrados en el reconocimiento del hospedador.

LIST OF ABBREVIATIONS

AAI	Amino Acid Identity	**MAVG**	Metagenome-Assembled Viral Genome
AMG	Auxiliary Metabolic Gene	**mb / mbp**	Megabase pairs (10^6)
ANI	Average Nucleotide Identity	**MCL**	Markov Clustering Algorithm
ATP	Adenosine Triphosphate	**MDA**	Multiple Displacement Amplification
BATS	Bermuda Atlantic Time-series Study	**MGI**	Marine Group I
bp	Base pairs	**ML**	Machine-Learning
C-D	Constant-Diversity	**NCBI**	National Center for Biotechnology Information
Ca.	*Candidatus*	**NCDLV**	NucleoCytoplasmic DNA Large Viruses
CCS	Circular Consensus Sequencing	**NGS**	Next-Generation Sequencing
CDD	Conserved Domains Database	**OLC**	Overlap Layout Consensus
CMP	Cyanomyophage	**ORF**	Open Reading Frame
CMP	Cytosine Monophosphate	**PacBio**	Pacific Biosciences
Cryo-EM	Cryo-Electron Microscopy	**PC**	Protein Cluster
dbCAN2	Carbohydrate-active enzyme ANnotation database	**PDB**	Protein Data Bank
DOM	Dissolved Organic Matter	**PEG**	Polyethylene glycol
dsDNA	Double-stranded DNA	**PHROG**	Prokaryotic Virus Remote Homologous Groups
et al.	*et alia* (and others)	**pI**	Isoelectric Point
FAVS	Fluorescence-Activated Virus Sorting	**PMP**	Pelagimyophage
gb / gbp	Gigabase pairs (10^9)	**pVOG**	Prokaryotic Virus Orthologous Groups
GC (%)	Guanine-Citosine (%)	**RBP**	Receptor-Binding Protein
GLUVAB	Genomic Lineages of Uncultured Viruses of Archaea and Bacteria	**RPKG**	Reads recruited Per Kilobase of the genome per Gigabase of metagenome
GOV	Global Ocean Virome	**rRNA**	Ribosomal RNA
HGT	Horizontal Gene Transfer	**SAG**	Single-Amplified Genome
HMM	Hidden Markov Model	**SAM**	S-adenosyl-L-methionine
HRC	Host Recognition Cluster	**smRNA**	small mRNA
HVR	Hyper Variable Region	**SMRT**	Single Molecule Real Time
ICTV	International Committee on Nomenclature of Viruses	**SR**	Short Reads
kb / kbp	Kilobase pairs (10^3)	**SRa**	Short-Read Assembly
KEGG	Kyoto Encyclopedia of Genes and Genomes	**ssDNA**	Single-stranded DNA
KtW	Kill-the-Winner	**tb / tbp**	Tera base pairs (10^12)
LPS	Lipopolysaccharide	**TSS VIII**	Type VIII Secretion System

LR	Long Reads	**UPGMA**	Unweighted Pair Group Method with Arithmetic mean
LRa	Long-Read Assembly	**VP**	Viral Population
MAG	Metagenome-Assembled Genome	**vSAG**	Viral Single-Amplified Genome

Index

1. Introduction

1.1 Role of phages on ecology and biogeochemical cycles

1.1.1 Effects of phage predation in biogeochemistry

Considering that more than 70% of the surface of the Earth is covered in water, it is no wonder that marine environments are the most extensive habitats on the planet [1]. Despite being invisible to the naked eye, prokaryotic microorganisms constitute over 90% of the living biomass found in the ocean [2] and are the driving force behind the biogeochemical cycles that take place in this habitat, such as the carbon, nitrogen, sulphur and oxygen cycles [2,3]. As an example, up to 20% of the global oxygen production is done by *Prochlorococcus*, a small marine Cyanobacteria [4].

Although the study of bacteriophages (viruses that infect bacteria, hereinafter referred to as "phages") from aquatic environments is almost as old as the discovery of phages themselves [5,6], the importance of phage in marine ecosystems was underestimated for decades as the first studies on viral diversity, based on counts of plaque-forming units, detected few viruses able to infect bacteria [7,8]. It was not until the early '90s that a series of major discoveries forced a change of paradigm. The application of transmission electron microscopy to phage enumeration experiments revealed that these organisms were the most abundant members of the marine ecosystem, reaching abundances from 10^7 - 10^{10} virions per ml of seawater [9]. These results were corroborated by epifluorescence microscopy methods a few years later [10]. Meanwhile, a parallel study revealed that these phages were actively and successfully infecting a significant portion of the microbial community [11]. Intrigued by this discovery, scientists proceeded to calculate the rates of viral decay in natural conditions. They found that viruses are sensitive to a variety of environmental processes, and in order to maintain their population numbers, viruses would need to successfully infect and destroy around 20 - 40% of the bacterial population daily [12–14]. These discoveries marked the beginning of the "third age of phage", where their position as major drivers of the great planetary biogeochemical cycles and their importance as genetic reservoirs was realised [15].

Phage-mediated lysis is roughly at the level of grazing by protists and zooplankton as a source of prokaryote mortality [16,17]. However, phage-related mortality presents key differences that have profound effects on the ecosystem. Consider the marine carbon cycle, in which carbon enters the biological pool via photosynthesis, performed to a large extent by bacterial autotrophs [18]. Grazing moves nutrients up the trophic levels (bacteria are eaten by larger protozoa, which are in turn consumed by larger organisms). By contrast, lysis by phages releases dissolved organic matter (DOM), which is recycled mainly by heterotrophic bacteria. This process, in which phage lysis redirects the flow of nutrients back into the microbial loop, is called the "viral shunt", and allows for an increase in heterotrophic microbial secondary productivity in conditions where nutrient availability limits production [6,19,20] (**Figure 1)**. The viral shunt also plays a role in other limiting compounds such as nitrogen, phosphorus, sulphur and iron [6,21]. It is important to point out that phage activity also alters the efficiency of the biological pump, a process related to the capture of carbon from the surface to the deep ocean by the sinking of large particulate aggregates [2,6], but the role of phages in this

process is not as clear since their activity both detracts and contributes to this event, depending on the circumstances [22].

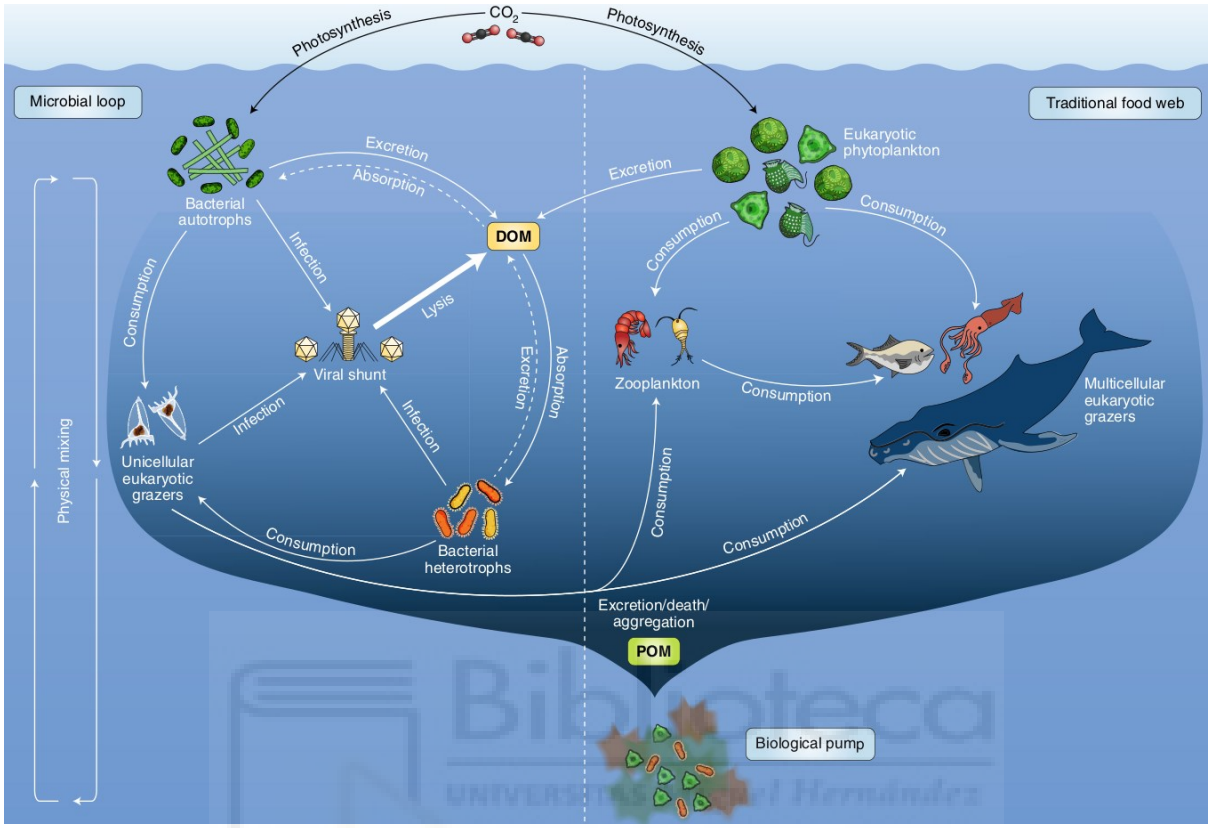


Figure 1. The marine food web, emphasising the role of the viral shunt in the recycling of dissolved organic matter (DOM) and particulate organic matter (POM). [20]

1.1.2 Effects of phage predation in bacterial diversity

Phage predation also plays a key role in what Hutchinson named "Paradox of the Plankton" [23]: Why are we able to find coexisting microbial species that exploit the same resources, when laboratory experiments suggest that one species should outcompete the rest?

Two characteristics of phage predation help untangle this apparent paradox. First, compared to grazer predation, which is relatively unspecific [24], phage predation is host specific. Phages find themselves in a predicament: they are dependent on their host's cellular machinery to complete their life cycle, but they can only infect a host once, since DNA injection into the host is an irreversible process. Therefore, phages are specialised predators that are calibrated to both the morphology and the cellular machinery of their host, in order to both recognize and exploit it [25]. In fact, the capability of phages to discern their host is so specific that it has been used for typing bacteria [26]. Second, phage predation is density-dependent [27]. Phages do not possess any means of independent locomotion and must rely instead on encountering their host by chance. Therefore, the frequency of infection will also depend on the abundance of the host in an environment. [28,29].

This dual nature of viral predation makes viruses powerful agents for controlling community composition, and results in a predator-prey dynamic named Kill-the-Winner (KtW) [30]. This hypothesis proposes that if an environmental change or a new mutation causes one specific bacterial species to thrive, it will be selected against, since it will have more encounters with its predating phage and successful infection cycles will mean that the number of phages that infect this fitter microbe will increase. In this way, predation by phages results in a fluctuating selection dynamic in which the fittest strain is selected against, protecting the general population against a sweep and maintaining population diversity.

The advent of genomics and the discovery of the great intra-species diversity in bacterial populations, the Kill-the-Winner model has since then expanded into the Constant Diversity (C-D) model [31], which both provides a link between predation and genotype and explains intra-species diversity. In the later, different clonal lineages in the same population are distinct from each other by both their capability to exploit resources and their different susceptibility to the population of phages that predate upon them. This predation is believed to provide an evolutionary advantage for the population. With phages acting as a control agent, the bacterial population avoids a clonal sweep, allowing it to maintain a gene pool that will allow the population to better exploit fluctuations in the environment **(Figure 2)**.

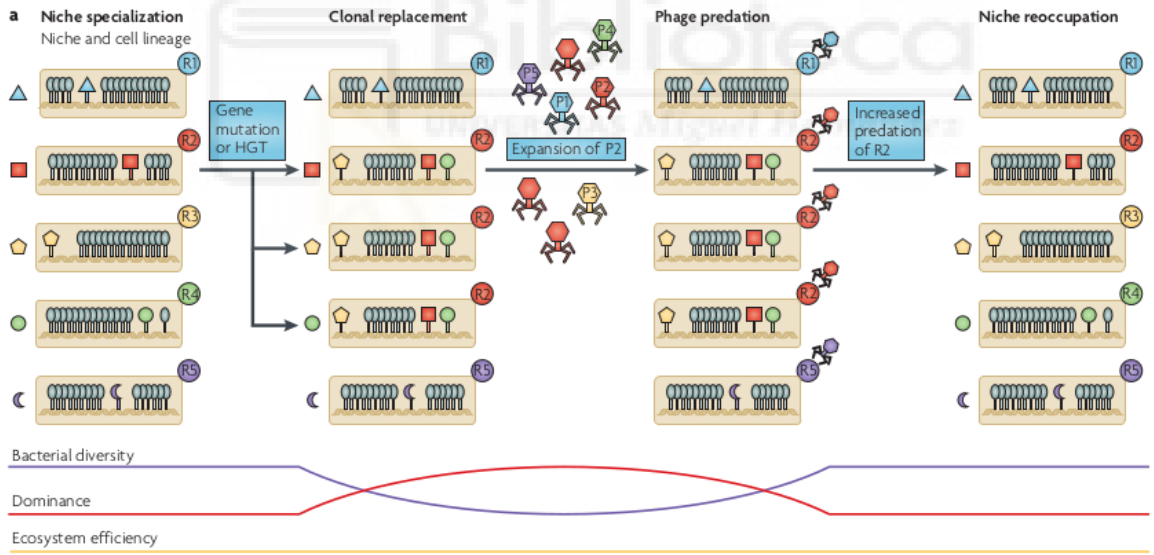


Figure 2. Population dynamics under the constant diversity model. A population of bacterial strains, with different gene content and phage receptors (R1-R5) is predated by a population of phages, each of them recognising a specific receptor. If a change in gene content causes a particular strain to be more fit and increase its population, it will be more predated upon, avoiding a clonal sweep. [31]

Last but not least, the inextricable antagonistic relationship between phages and their hosts results in a process of reciprocal adaptation and counter-adaptation to each other [32]. The first attempt to explain these coevolution dynamics was proposed in the '70s in the form of

the Red Queen dynamics model, which postulates that both members of an antagonistic evolutionary interaction (such as a prey-predator relationship) must continuously change for both lineages to survive [33]. Studies based on environmental data suggest that Red Queen dynamics explain evolutionary changes at the microdiversity scale, while C-D maintains gene pool diversity [34].

In addition to driving evolution directly maintaining microbial inter- and intra-species diversity, phages also indirectly contribute to the diversity of the bacterial community by serving as vectors for horizontal gene transfer (HGT) [35]. Considering their ubiquity and rate of infection, phages could arguably be the greatest reservoir of genetic diversity on Earth [36,37]. During infection, phages can randomly incorporate bacterial genomes fragments into their genomes or in phage capsids [38], which can then be transferred to subsequent hosts as phages infect other host cells. These transduction events can even cause the switch from a benign microorganism to pathogen: an example would be the filamentous phage CTX, which carries the toxin encoding genes responsible for the full virulence of *Vibrio cholerae* [39]. Moreover, various studies show phages from environmental samples containing other virulence factors, such as many toxins, antibiotic resistance genes or genes involved in host adhesion and invasion [40,41]

1.2 Phage genomics

1.2.1 Phage classification

Phage genomes are composed of either RNA or DNA, which can be single-stranded or double-stranded. This genetic material is packaged into a proteinaceous capsid that can be polyhedral (*Microviridae*, *Corticoviridae*, *Tectiviridae*, *Leviviridae* and *Cystoviridae*), filamentous (*Inoviridae*), pleomorphic (*Plasmaviridae*) or attached to a tail (*Caudovirales*) [42], the molecular machine involved in host recognition and DNA delivery [43,44]. We will focus on this last clade, as the vast majority of phage complete genomes (85%) belong to it. Traditionally, clades in Caudovirales have been based on tail morphology **(Figure 3)**. *Podoviridae* have short, non-contractile tails, *Siphoviridae* have long, non-contractile tails, and *Myoviridae* have long, contractile tails [45,46].

The host recognition module can be found at the distal end of the tail. In some phages this region is simple, but in others it has evolved in a complex structure called a baseplate [43,47]. In myophages, this structure is of utmost importance since it coordinates host recognition with sheath contraction [48]. The interaction between phage and host receptors is mediated by Receptor-Binding Proteins (RBPs). These proteins contain domains for recognizing bacterial structures found in the cell wall, such as lipopolysaccharides (LPS), transport proteins, porins, teichoic acids and peptidoglycan side chains [46,49,50]. Phages typically code for two types of RBPs: primary RBPs, which reversibly bind to a primary receptor and

allow the phage movement over the cell wall surface; and secondary RBPs, which bind irreversibly and trigger DNA injection [46,48].

The minimum genome size in tailed phages is about 19kbp, and the largest genome described to date reaches up to 735 kbp [51,52]. Compared to their bacterial hosts, phage genome size is influenced by its morphology. Capsid size plays a major role in genome size, both directly (the phage genome has to be able to fit inside the capsid) and indirectly (there is an optimal genome length for injection efficiency depending on the size of the capsid) [53]. Equally important is the tail structure, as different tail structures are capable of successfully delivering different amounts of DNA into the host cell. This results in mean genome sizes increasing with tail complexity: the mean genome size for *Podoviridae* is 40kbp, 60kbp for *Siphoviridae*, while *Myoviridae*, the family with the most elaborate injection mechanism, has larger genomes that start at 100kbp and can surpass 500kbp [51,53]. Deviations from this optimal size lead to a less efficient DNA injection step, which means phage structure applies an evolutionary pressure to either gain or lose DNA independent of gene function or replication fitness. In this way, DNA content in phage genomes is not aggressively removed for fitness and can thus provide a reservoir of genetic information for potential future use [53,54].

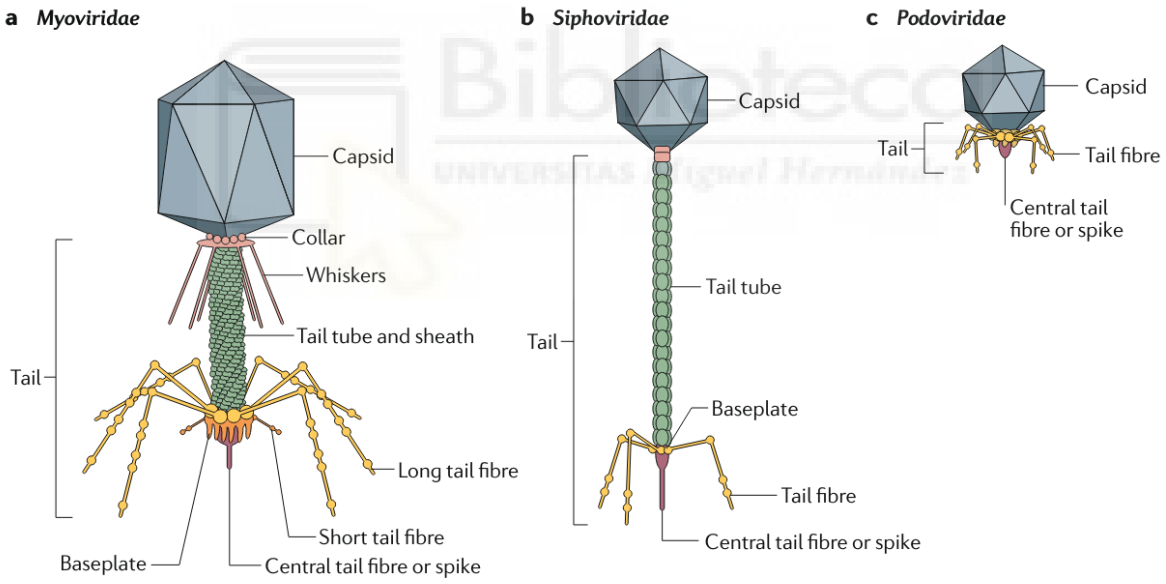


Figure 3. Representative structures of tailed phages: *Podoviridae* (Short tail, without tube), *Siphoviridae* (long, non-contractile tail) and *Myoviridae* (long, contractile tail with sheath). [43]

1.2.2 Phage sequence diversity

Phage proteins are extremely diverse, and two ortholog proteins can share little sequence similarity. In fact, most of the coding sequences from phage genomes (*ca.* 80%) [53] do not present any homology to other known proteins, and no gene with orthologous counterparts

in all phage genomes has been found that can be used as a reliable marker for phylogenetic studies [55,56]. This fact may appear puzzling when we consider the genome size constraints and the rigidity of the phage life cycle, which requires the phage to encode functions for host takeover, genome replication and virion biosynthesis. It would be expected that these basic functions would be extremely conserved and thus easily identifiable.

This diversity in phage proteins is derived from two main causes [57–59]. First, mutation and recombination rates are much higher in phages than in their hosts, causing a sequence divergence that might blur the homology signal. This divergence does not affect function, as protein function is more conserved than sequence [60], but as function annotation for newly identified genes relies on finding a homologous protein whose function is already known [61,62], it is possible that two sequences might maintain the same function and be divergent enough to not be identifiable.

Second, phage genomes are dense. Approximately 90% of the phage genome is occupied by protein-coding sequences, and the average gene length is lower than that of their hosts [63,64]. It is not uncommon for phages to present overlapping genes. Usually the overlaps are found between the initiation and termination codons of two different genes, but in some cases complete genes are encoded within each other [65,66]. Another form of gene compression is site-specific frameshifting to obtain different variants of the same protein [67,68]. Gene segments coding for functional domains are shuffled between different genes during evolution, resulting in mosaic proteins in a process known as “Domain shuffling”. This phenomenon is more prevalent in proteins related to host recognition, but also appears in other proteins [69]. These overlapping and frameshifting events complicate the prediction of open reading frames from the sequence, resulting in truncated or missing genes.

Perhaps the most striking feature of phage genomes is their extensive mosaicism. Genetic mosaicism refers to the phenomenon where different DNA regions of the genomes have distinct evolutionary stories [70–72]. This is easily identifiable when comparing two closely related phages, as regions that are almost identical will abruptly transition into others with no resemblance [70,73,74]. These regions are often the result of extensive HGT events, which have been extensively studied and reported in phages and include a wide variety of molecular mechanisms, including homologous (“relaxed”) and non-homologous (“illegitimate”) recombination [75–77]. The frequency of recombination events is not equal among phages, as evidence suggests they are heavily influenced by their life cycle and host [78].

Not all genes participate in mosaicism to the same degree. “Core” genes intimately interact with each other and are essential to the function of the phage (for example, the genes that form the head or the replication machinery), and usually conserve synteny and sequence similarity between closely-related phages. Conversely, “non-core” genes can be removed from the genome without affecting the function of the phage [53]. The degree of mosaicism depends on a multitude of factors, including their replication mode (virulent or temperate), their host and their morphological complexity. For example, phages from the *Tevenviridae* clade show large regions of core genes interspersed by hypervariable regions, whose content varies wildly from one genome to the next [53]. Localization of these flexible regions is usually found in the same place, and it is easily identifiable in recruitment assays for their under-

recruitment in metagenomes (metaviromic islands) [79]. This is a good way to detect where these hypervariable modules are located without the need to compare with other closely related genomes.

What kinds of genes can be found inside these flexible regions? The tail fiber module is the only structural region in *Caudovirales* that is not part of the phage core genome and is in fact its most plastic region. Tail fiber proteins are reported to show high variability [54,79] domain shuffling and gene fusion are pervasive. Some of these hypothetical proteins contain domains related to carbohydrate binding, which make them strong candidates to be RBPs [79]. Flexible regions also contain a wide variety of glycosyltransferases. These family of genes are expressed during virus replication and serve a variety of purposes. One example can be found in the T-even clade of phages, which express glycosyltransferases to modify the nucleotides that conform their genome, preventing its digestion by bacterial restriction enzymes [80]. Another example can be found among *Shigella* seroconverting phages, which express glycosyltransferases to modify the cell surface of its host and prevent co-infection by competing phages [80].

1.2.3 Auxiliary Metabolic Genes (AMGs)

The presence of virulence factors in phage genomes has been established for a long time, but phages may also include Auxiliary Metabolic Genes (AMGs). These genes allow the phage to modulate and expand the metabolic output of their host, increasing their fitness and therefore increasing viral propagation [81,82]. AMGs are acquired from the bacterial host, but continue to evolve separately from their bacterial orthologs, eventually repurposed to improve viral fitness. Evidence of this divergent evolution is that AMGs have reduced gene lengths compared to their bacterial counterparts [83], and some of them modify their functions to better serve the interests of the phage [83,84].

AMGs are classified into two distinct classes [85]. Class I AMGs include genes involved in central metabolic functions found in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [86], such as photosynthesis and amino acid metabolism. Class II AMGs cover genes with peripheral or undefined roles, such as antibiotic resistance and membrane transport. Genes involved in common viral processes, such as nucleotide synthesis and protein metabolism, are not considered AMGs since their functions directly affect the production of viral progeny, instead of acting indirectly by complement host function [85].

AMGs increase viral fitness by following two broad strategies. The straightforward approach is for AMGs to code for proteins that catalyse for rate-limiting reactions (transaldolase *talC*) [87], increase the uptake of limiting nutrients (phosphate-binding protein *pstS*, ammonium transporter *amt*) [81,88] or accelerate energy production (high light induced protein *hli,* plastoquinol terminal oxidase *ptox*) [88]. Another example would be proteins required to maintain the operation of the host cellular machinery, since host takeover stops host gene expression [81]. An example of this would be the proteins forming the photosynthesis reaction core (*psbA*, *psbD*), which are susceptible to photodamage and must be replaced periodically to ensure energy production [89,90]. Alternatively, phages may also code for

AMGs that redirect host metabolism towards pathways that promote viral productivity, either through manipulation of the expression of host metabolic genes (pyrophosphohydrolase *mazG*) [81] or manipulation of host pathways (Calvin cycle inhibitor *cp12*) [83].

Whilst all previous examples are extracted from cyanophages, AMGs are ubiquitous in phage genomes. AMGs have been found in a wide variety of phage families [91,92], including *Thaumarchaeota* viruses [93]. Since AMGs are directly acquired from the host genome, AMG distribution is highly correlated with host genera [88]. This correlation has been exploited to assign putative hosts to phage genomes that code for certain AMGs, as is the case for *Actinobacteria* phages and the transcription factor *whiB* [94].

Likewise, AMGs are involved in a wide array of metabolic pathways, either related to energy production or nutrient acquisition. To this date, AMGs have been found in pathways involving methane (methane monooxygenases *pmo*) [95], sulphur (sulfite reductase *dsr*, sulfane dehydrogenase *sox*) [96–98], nitrogen (ammonia transport protein *amt*, ammonia monooxygenase *amo*, nitrate reductase *nar*) [93,99,100] , phosphate (*ptsS*, phosphate uptake regulon *pho*) [101,102], carbohydrate (mannose-6-phosphate isomerase *manA*, ribose 5-phosphate isomerase *rpiB*, glycogen synthase *glgA*) [85,87], amino acids (cysteine synthase *cysK/M*, S-Adenosylmethionine synthetase *metK*, *N*-succinyldiaminopimelate aminotransferase *dapC*) [85,87], nucleotide (purine synthesis operon *pur*, cobalamin synthase *cobS*) [88] and photosynthetic energy production (*hli*, *psbA*) [84,89,90]. Metagenomic recruitment analyses suggest that the distribution pattern of AMGs varies according to environmental conditions. A correlation between the abundance of some AMGs and temperature has been observed [103], and AMGs involved in phosphate metabolism are overrepresented in the North Atlantic Subtropical Gyre compared to the richer North Pacific Subtropical Gyre [102]. However, not all AMGs are equally abundant in phage genomes, some are so prevalent that could be considered part of the core genome (*phoH*, *cobS*, *mazG*), while others are more sporadic (*talC*, *ptox*). It has been hypothesised that these variances reflect viral adaptation to changing environmental conditions [88].

1.2.4 Phage Life Cycle

One of the major differences between viruses is their strategy to replicate, with most phages classified into virulent or temperate lifestyles. Virulent phages follow what is called the lytic cycle. During this cycle, the phage DNA enters the host cell, the cellular machinery is hijacked to replicate its genome and assemble new virions, and finally ends with the lysis of the host cell and the release of the newly assembled virions [104]. In contrast, temperate phages can either carry out a lytic infection or enter a lysogenic cycle, in which the viral genome is integrated into the host chromosome. In this latent form, called “prophage”, the virus will replicate together with the host genome after each bacterial division. When certain conditions are met, the phage is able to excise from the host genome and follow a lytic cycle [104]. The conditions that cause temperate phages to switch from lysogenic to the lytic cycle are still not fully understood, but the switch has been reported to trigger after changes in the physiological state of the host, such as under conditions of stress like radiation, temperature or starvation [104]. Interestingly, meta-analyses of viral and bacterial densities in a variety of

environments suggest that temperate phages follow what is called the "Piggyback-the-Winner" strategy, in which phages will follow a lysogenic cycle while conditions for host growth are favourable, and switching to lytic infection otherwise [105,106]. Alternatively, there is also evidence that lysogenic phages coordinate their lytic cycles using a quorum-like system [107,108]. A less common, but widely studied phage life cycle would be the chronic infection found in some archaeal viruses and single-stranded DNA filamentous phages from the *Inoviridae* family. In this life cycle, virions are slowly shed from the host cell over a long time without obvious cell death [109,110].

However, there are no clear-cut divisions in nature. Some temperate phages, like coliphage P1, do not integrate into their host genome and exist in a plasmid form ("plasmid prophage") until they need to change into a lytic cycle [111]. Meanwhile, some lytic phages have been observed into what has been called a "carrier state", in which the phage is maintained within a host population without a measurable effect in cell growth [112]. We could also include in this section the defective bacteriophages, prophages that have lost via mutation the genes required to switch into the lytic phase and are therefore unable to complete their life cycle, replicating indefinitely in the bacterial genome [109].

1.2.5 Phage phylogeny

Historically, phage phylogeny has been based on their morphology, which was in turn derived from electron microscopy information. For example, the historical classification for tailed phages is based on work from Bradley, Ackermann and Eisenstark from the 70s, which classified the *Caudovirales* in morphotypes A, B and C; corresponding to the *Myoviridae*, *Siphoviridae* and *Podoviridae* families respectively [113,114]. Similar work was performed in other, well-studied viral groups (*Inoviridae*, *Microviridae*, *Tectiviridae*, etc), resulting in the well-known family-level classification system [115], with phylogenetic relationships based on the physical characteristics of the virion.

However, this clear-cut classification was not to last. Already in 2002, Rohwer and Edwards illustrated the difficulties of establishing a phage phylogeny based on sequence information in their phage proteomic tree paper [116] and highlighting inconsistencies within the morphology-based families. This seminal work already identified some key challenges in phage phylogeny studies, namely the lack of a universal marker gene and the issues derived from rampant HGT and mosaicism. The advent of viral genomics and viral metagenomic studies has done nothing but accelerate its collapse. Viral metagenome studies do not allow for the morphological characterisation of the recovered sequences, and determining virion morphology from sequence is a daunting task [57]. Furthermore, classification based on morphology is incompatible with a phylogeny determined from sequence features. For example, the genera *Lederbergvirus* and *Myxoctovirus* are both classified as *Podoviridae*, but their genomes share no orthologues [115]. Although various subfamilies were added into the phage tree of life, it quickly became obvious that these long-established families were not monophyletic and did not even fit into the same order [117–119].

This paraphyly has been corroborated at a greater scale by a plethora of studies such as the previously mentioned phage proteomic tree [116], a concatenated protein phylogeny of *Caudovirales* phages [120], a composite tool combining gene homologies and gene order (GRAViTy) [121], a virus domain orthologous groups approach (VDOG) [122] and a variety of network-based approaches [118,119,123,124], in which nodes represent phage genomes and edges the similarity between them. This latter approach has proven to be a more faithful representation of the relationships between the mosaic genomes. Since the first network representation of phage phylogeny in 2008 [118], a number of studies using different approaches have appeared, such as a bipartite network of shared genes [119] or networks based on shared predicted proteins [123,124]. In these network-based representations, closely-related phages still group into distinct clusters, which suggests that despite the mosaic nature of phages, they still evolve in somewhat isolated gene pools and are thus able to be phylogenetically classified. Interestingly, temperate phages are closely interconnected to one another, while lytic phages show less connectivity. This reinforces the role of temperate phages as the "brokers" of HGT between viruses [51]. Furthermore, the evolutionary relationships between phages is also related to their choice of host and the environment they inhabit [78].

Currently, the International Committee on Nomenclature of Viruses (ICTV) is in the process of a major phylogeny overhaul, transitioning into a sequence-based taxonomy, with new criteria for the demarcation of new clades. The order *Caudovirales* has been abolished and the subfamilies are being assessed on a case-by-case basis [115].

1.3 The study of marine viruses

1.3.1 Phage culture and isolation

Phage isolation from environmental samples is still the gold standard for viral identification. In this method, an environmental sample is added to a culture of host bacteria, then isolated [125]. The advantages of this method are considerable. Co-culture of a phage and its host is still the gold standard for establishing host-phage pairing, which is the fundamental piece of information on which all further analyses are based. Furthermore, cultures are still the only method to comprehensively characterise a virus infection cycle and molecular interactions with its host cell.

Unfortunately, culture-dependent phage isolation is reliant on host cultures, and the vast majority of microorganisms either cannot be cultivated in the lab or are remarkably fastidious to grow [126,127]. For example, the isolation procedure is dependent on the type of bacterial culture available: if the host can be grown on an agar plate, a plaque assay is preferred, but if the host can only be grown on liquid medium, it is required to perform a dilution-to-extinction assay. This last method is considerably slower, as an epifluorescence microscopy assay is required between dilution steps to confirm the lysis of the host [128]. Even when host cultures are available, some phages are not virulent enough to be isolated by these methods

[129,130] or they require growth conditions different from those of the host [131]. Consequently, it is not a surprise that the phages isolated through cultures are those that infect fast-growing, copiotrophic, well-studied hosts. For instance, the first marine phages to be sequenced infected the heterotroph *Pseudoalteromonas [132]* and the photoautotrophic *Synechococcus* [133] in the years 1999 and 2002, respectively. In recent years, methods for phage isolation have improved and phages have been isolated for fastidious clades such as SAR11 [134,135], SAR116 [136], OM43 [137,138] and *Roseobacter* RCA [139,140]. Still, the throughput of culture-based studies significantly lags behind other approaches.

1.3.2 Sequencing-based approaches: a brief primer on DNA extraction

Opposite to cultures stands sequencing-based approaches, in which DNA from a viral population is harvested from an environmental source and then sequenced. The main characteristic of this strategy is that virus discovery is uncoupled from virus isolation, as genetic material can be sequenced directly from a sample regardless of cultivability [141]. We will focus on the recovery of dsDNA phages from seawater samples, since it is the method relevant to the datasets used in this work. However, it is important to note that extraction and isolation of phages will vary wildly depending on the type of environment sample and the type of phages desired [127]. Protocols for these samples have been perfected and standardised thanks to decades of research, but protocols for other sample types (such as soil) or other virus clades (such as RNA phages) are much less developed [142,143].

In phage DNA isolation, the main problem is the extremely small size of phages. Although the abundance of phages exceeds that of prokaryotes by more than an order of magnitude, they represent only 5% of the total biomass [2]. This is reflected in their DNA content, there is only 10^{-17} grams of DNA per virion, compared to 10^{-15}g of DNA per bacteria [144]. This problem is compounded by the massive amount of free DNA in the environment, as even if more samples are processed to obtain more viral DNA, viral genomes are so small that any degree of contamination will overwhelm the viral signal [127].

Therefore, the first step is to concentrate the viral DNA present in the sample. To this end, seawater is sequentially filtered to remove the eukaryotic and prokaryotic fractions by the use of 20µm, 5µm, and 0.22µm pore size filters, with the viral fraction remaining in the filtered seawater [144,145]. As sample volumes are quite high, it is necessary to concentrate the viral DNA even further. The most widespread method is tangential flow filtration, based on the use of a filter parallel to the flow of the sample. With this setup, water and particles smaller than the pore filter pass through based on diffusion instead of being forced through the filter via pressure, resulting in a liquid concentrated sample [146,147]. Although this procedure ensures that virions are not damaged, it has a low throughput and results are highly variable [148,149]. Other techniques are available, such as iron chloride flocculation [148] and polyethylene glycol (PEG) precipitation [150]. These techniques are based on the binding of phages to a substrate, from which the virions are then recovered from. The former method is considered to better preserve the viral community [151]. If the sample is derived from biofilm, an extraction step might be required to remove the extracellular matrix from the sample. Contaminants that might muddle viral DNA can be eliminated by a purification step.

Extraneous DNA can be removed with a treatment with DNase, as phage DNA is protected by the capsid [127]. Cellular debris can be removed by a centrifugation step with a sucrose or CsCl gradient [152].

When the concentration and purification steps were not successful, it might be necessary to submit the extracted DNA sequences to an amplification process. The first amplification procedures were based on fosmid cloning, in which large fragments (up to 100 kbp) from an environmental source are cloned in fosmids (forming fosmid libraries), then transfected into *Escherichia coli* cells for natural amplification of the fosmid DNA [145,153]. This is a difficult and laborious process that found additional challenges in its application for phages, namely the presence of toxic genes (holins, lysozymes) in phage genomes and modified nucleotides that could not be synthesised by *E. coli* [154–156]. Nonetheless, the technique was eventually perfected and is still widely in use for the study of viruses, as the size of the insert is often large enough to contain entire phage genomes [157,158].

Fosmid cloning quickly gave way to simpler, more efficient methods such as Multiple Displacement Amplification (MDA). This protocol makes use of Phi29 DNA polymerase, an enzyme with strand-displacement activity that enables it to amplify genomic DNA using random primers and a single denaturation step [159,160]. However, there is ample evidence that this method is heavily biased towards certain types of sequences and thus distorting the natural viral community [160–162]. PCR-based methods such as linker amplification, based on ligating primers to sheared DNA to perform standard PCR amplification, perform better in this regard but still present issues of their own [163]. Improvements in second-generation sequencing (so less DNA is required) and the potential for larger reads of third-generation sequencing protocols (so shearing of DNA is not desirable) mean that this amplification step is ignored in favour of processing more sample.

1.3.3 First-Generation Sequencing (Sanger Sequencing)

The first sequencing method was introduced by Sanger et al in 1977 [164] and allowed for the sequencing of a 1kb sequence from a single template. By the end of the 2000s, state-of-the-art sanger sequencers could process 96 templates in a single run [165]. Considering that an environmental sample can contain thousands of different organisms, it is clear that these kinds of samples are beyond their scope [166].

Therefore, the first studies of environmental phage diversity were done in reduced DNA collections, usually based around the study of genome size or a universal marker gene to assess community diversity (“metataxomics”) [167]. As stated previously, phages lack such a universal marker, but some early studies were done based on conserved genes [168] or structural proteins [169,170] that could target broad subsets of phage communities. These studies revealed the massive diversity even within these subgroups. Sequences from cultured phages fell within a few defined clusters, but the vast majority of environmental fragments fell in new, undefined clades [170,171]. In some cases, the composition of the community was significantly different between samples taken only a few metres apart from each other [172,173], and even then, identical gene fragments were recovered from different oceans and even from freshwater [169]. These results were corroborated by other studies targeting different phage populations [174].

In 2002, Breitbart *et al*. published the first viral metagenomes ("viromes") from two marine samples (surface water and sediment), recovering more than 1,000 viral fragments [175]. At the time of publication, the majority of these sequences (*ca.* 65%) had no significant similarity to any known viral genomes. A subsequent analysis of a marine sediment sample by the same group reported an even larger uncharacterized diversity [176], with an estimated 1,000,000 phage genotypes per kg of processed sediment.

1.3.4 Second generation sequencing (Next-Generation Sequencing)

The early 2000s heralded the arrival of second-generation sequencing technologies, also known as Next-Generation Sequencing (NGS) [177]. A variety of platforms were available at the start of the decade, but the most widely used ones were Roche 454 pyrosequencing and Illumina dye sequencing. Differences between platforms are mostly based on the chemical process used in this last step, and are reflected in the sequenced fragments (reads): 454 pyrosequencing can produce reads up to 800bp, while Illumina dye sequencing produces reads of up to 150bp [177]. NGS platforms present two major advantages compared to Sanger sequencing, namely the significantly lower cost of per-base and the considerably higher throughput [165]. Although the former made previous fosmid and amplicon-based experiments cheaper and more efficient, the latter was the real breakthrough, as it was finally possible to sequence DNA environmental samples in a process named "shotgun sequencing", in which environmental DNA is randomly sheared into small fragments that are independently sequenced [178]. This finally allowed scientists to obtain the genomic content of a large part of the microbial community, providing the chance to analyse their composition and their metabolic potential.

It should not be surprising, then, that the study of environmental viruses was revolutionised by this technology. A milestone experiment using this technology was carried out in 2006, in which the group led by Forest Rohwer sequenced and studied the viral community of four distinct oceanic regions (the Sargasso Sea, the Gulf of Mexico, British Columbia coastal waters and the Arctic Ocean) using 454 pyrosequencing [179]. From this moment forward, virome studies explode in number. This trend can be followed by the number of viral contigs available in public databases, with only 84 contigs available in 2010, 35,000 in 2016, 775,000 in 2018, and 2,300,000 in 2020 [180]. Not all of these sequences are derived from viromes, however, as viral sequences can also be found in the cellular fraction in a variety of shapes: prophages, viral particles attached to cells and giant viruses (such as those from the Nucleocytoviricota clade) and cells infected with lytic phages (virocells) [181].

Without a doubt, the landmark studies of the decade are derived from large-scale sampling projects (both spatial and temporal), as they have provided comprehensive, large-scope datasets that have allowed researchers to explore viral diversity on a global scale. The first large-scale dataset of this kind was the Pacific Ocean Virome, which grouped 32 viromes from a variety of locations, depths and seasons in the Pacific Ocean [182]. In the spatial end of the axis, two consortia-driven sampling oceanographic expeditions (the French, surface-focused Tara Oceans and the Spanish, depth-focused Malaspina) have produced massive datasets (90 Tara Oceans viromes, 14 Malaspina viromes) [100,183]. In 2019, the second version of Global

Ocean Virome (GOV2) was published, adding 40 new viromes from the 2013 Tara Arctic expedition [184]. This collection of 184,009 viral populations is the most complete marine viral collection to date, spanning viromes from the whole globe (**Figure 4**). These expeditions have revealed that depth, salinity, temperature and oxygen concentration are the main drivers of viral community structure in the oceans [183]. On the temporal side of the axis, we find the Bermuda Atlantic Time-series Study (BATS), which now includes a dataset with more than 10 years of viral abundance in the Sargasso Sea [185]. Nevertheless, smaller-scale viromes have also had a large hand in shaping our understanding of marine viral diversity. In 2013, Mizuno et al. used large insert metagenomic fosmids from a Mediterranean deep chlorophyll maximum sample, resulting in a dataset of 208 complete phage genomes [153]. This increased the number of complete marine phages in marine databases by an order of magnitude [186] and paved the way for the discovery of metaviromic islands a year later [79].

As powerful of a tool as it is, shotgun sequencing analyses present several weaknesses. The crux of its issues is the small size of the sequenced reads, which require an assembly step to obtain sequences large enough to predict protein sequences. Even though assembler software has significantly improved, it still has issues assembling repeat regions [187] usually resulting in a collection of genome fragments of varying length. With time, the techniques, workflows and software employed to manipulate and study NGS datasets have improved considerably. An example would be binning, the process of grouping contigs from the same lineage based on sequence nucleotide composition and coverage data [189,190], which allows researchers to reconstruct Metagenome -Assembled Genomes (MAGs), composite genomes that provide insight on the capabilities of a clonal community [190–192].

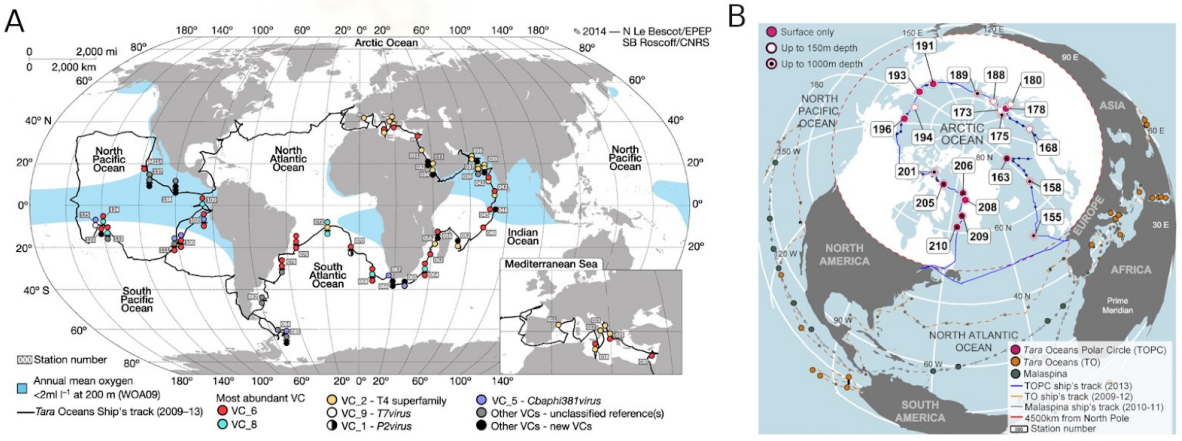


Figure 4. Sampling stations from Tara Oceans (A) and Tara Arctic (B). Each point in the map indicates a sampling site. [100,184]

A similar approach (based on overlapping contigs) has been employed for phage genomes in the past, resulting in Metagenome-Assembled Viral Genomes (MAVGs) (**Figure 5**) [193]. The microdiversity in the population is also lost, as assembly produces consensus contigs from the

dataset reads. This results in the loss of a large amount of phage diversity: most of the reads in a viral metagenomic dataset (>80%) fail to recruit against available phage reference genes, both cultivated and uncultivated [100].

A solution to this particular issue can be found in Single-Amplified Genomes (SAGs), which allow for the sequencing of one cell at a time. First, cells from an environmental sample are sorted, usually with a flow cytometer, and placed on individual plate wells. DNA from each cell is then extracted, amplified and finally sequenced [194]. This approach has been extensively used to study bacteria difficult to culture and assemble, with considerable success [194–196]. A similar approach has been employed for viruses. In 2017, Martinez-Hernandez et al reported the use of Fluorescence-Activated Virus Sorting (FAVS) to recover 2,234 distinct virus-like particles from environmental marine samples, out of which 44 were sequenced to recover viral SAGs (vSAGs) [197]. One of the recovered vSAGs was vSAG37-F6, a SAR11 phage that has been revealed to be the most abundant phage in the oceans [198].

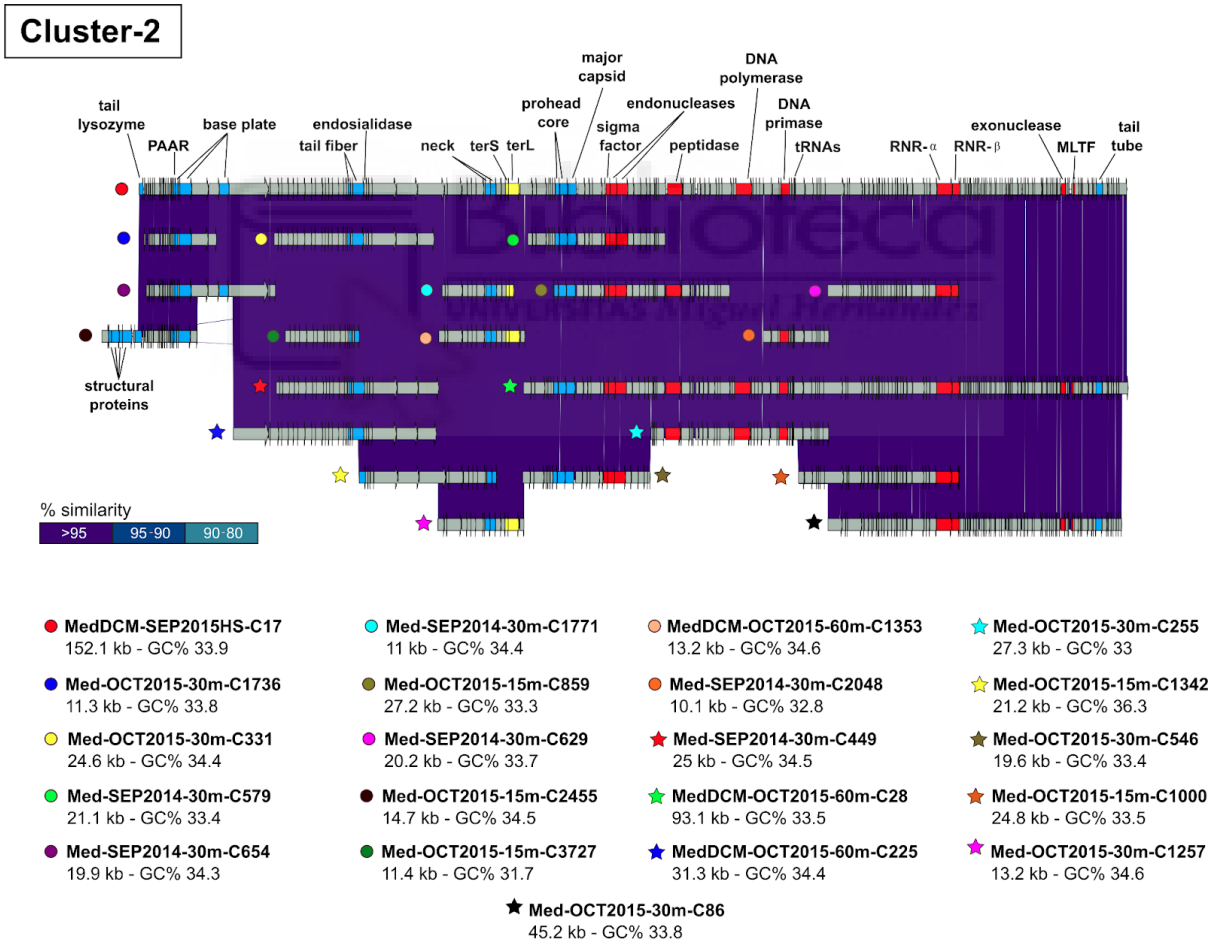


Figure 5. Reconstruction of a Metagenome-Assembled Viral Genome (MAVG), in this case the SAR11 myophage MAVG-2. Notice how every section of the genome is covered by at least two contigs [193].

1.3.5 Third generation sequencing

In 2011, Pacific Biosciences (PacBio) commercialised the "Single Molecule Real Time" (SMRT) sequencing technology [199], which was followed by the release of a competing long-read sequencing technology by Oxford Nanopore (MinION) three years later [200]. Although the chemistry behind each platform is different [201,202], both promised read lengths on the order of thousands to hundreds of thousands of base pairs, almost three orders of magnitude more than the previous generation of sequencers [203]. For metagenomics, these long reads would help overcome some of the limitations derived from the assembly step. Longer reads would allow for better resolution of genomic regions with repeats and for the recovery of abundant, highly diverse microbe populations.

These larger reads came with trade-offs. These technologies are more expensive, offer less throughput than the previous technology, and not least of all, are much more error-prone (on the order of 5-15%, compared to 0.1% for Illumina sequencing) [199,204]. A variety of methods have been developed to overcome this last hurdle, mainly correction of long reads by the use of a short-read dataset or, provided there is a high coverage, comparing the long reads against themselves [205,206]. However, the first approach requires to effectively sequence the same sample twice and the latter is not feasible for environmental samples, which never reach that level of high coverage. Changes in the technology can also help correct the error rate. For instance, the development of high-fidelity approaches such as PacBio circular consensus sequencing (CCS), which self-corrects by adding adapters at the ends in order to circularise and repeatedly sequence the DNA template [207].

Although the technology is still in its infancy, the application of long-read sequencing to viral metagenomes is particularly exciting, as reads can span several thousands and nucleotides and therefore it would be possible to recover complete or almost complete viral genomes from a single read. There are already some examples of long-read sequencing applied to the study of viromes using the Nanopore sequencing platform. Beaulaurier *et al.* recovered 1,864 new complete assembly-free virus genomes from three nanopore datasets [208]. On the other hand, Dugdale et al recovered around 2,500 viral contigs from the assembly Nanopore and Illumina datasets from the same seawater sample of the Western English Channel [209], showing that a hybrid or long read-only assembly improved the recovery of viral contigs and their metaviromic islands compared to short read assemblies.

1.3.6 More is less? Viral dark matter

The widespread adoption of high-throughput DNA sequencing has made the number of phage sequences available to increase dramatically. However, with this deluge of new data came the realisation that the vast majority (63-93%) of environmental phage sequences cannot be assigned a host or functionally annotated [182], and the immense majority of viral sequences found in nature cannot be taxonomically classified [183]. Moreover, each virome dataset adds more and more of these unknown sequences to the databases, providing no insight more than the huge diversity of the virosphere that seems to be bottomless. These uncharacterized sequences have been termed the 'viral dark matter' [7,153,210], and our inability to characterise them have severely limited our understanding of marine viruses.

Here again we encounter the crucial limitation of metagenomic studies: Identifying the genome sequence of a novel phage is only the first step towards understanding its role in the microbial ecosystem. A phage genome extracted from an environmental sample carries no direct information that can link it to its host, especially if it is a novel virus with no significantly detectable homologs in genomic databases [211]. Combined with the fact that these are usually recovered in a fragmented state, and that the vast majority of viral ORFs also have no homologues in protein databases, it is easy to see how viral metagenomic datasets can quickly become mere strings of nucleotides, with no information about their function and reducing the function of the scientist to that of a mere genome collector [212].

However, isolation of phages from abundant, difficult-to-culture hosts [134,136,138] and the refinement of analysis tools for viral sequences (see below) have allowed the advancement in our understanding of viral phage communities. Although this community and its diversity is still largely unexplored, a few common patterns have emerged. Viral species are globally distributed (everything is everywhere), but the relative abundance of each species is restricted by local selection [183,213]. There are also a few general trends: phage community composition is linked to bacterial composition, which means it is affected by marine currents and water column stratification [214,215]. There is evidence of niche specialisation between photic and aphotic zones, but there is a significant downward flux of viruses, presumably driven by particles [183,216], which explains the presence of cyanophages in aphotic marine sediments [217,218]. Interestingly, cyanophages that date over 50 years old recovered from these sediment samples are still infective, suggesting these sediments could be reservoirs of genetic diversity [6]. Furthermore, even with as diverse as marine viruses are, constant phage communities can be found. For instance, out of the hundreds of cyanophages sequenced, some of them can be grouped into discrete populations, which suggests there are constraints to gene transfer between phages [219–221]. These populations are also temporally stable, with the same phages found several years apart with the only differences being in the non-core genome and RBPs [219,220].

The abundance of single-stranded DNA and RNA viruses is currently unclear. As TEM microscopy is expensive and laborious, current viral quantification assays are based on epifluorescence with SYBR green as a staining agent, which insufficiently stains RNA and ssDNA molecules [222–224]. Studies employing alternative quantification protocols indicate that marine samples are dominated by non-tailed virus particles, but epifluorescence-based counts result in higher viral abundances, suggesting that the contribution of ssDNA and RNA viruses might not be much [225,226]. Further studies are required to determine the contributions of RNA and ssDNA viruses to the marine phage population.

1.3.7 Illuminating viral dark matter: predicting hosts from phage genomes

Cultivation-independent in vitro assays for host assignment exist, usually based on co-detection of host and viral marker in a host cell, either with fluorescence markers (PhageFISH) [227], droplet digital PCR [228] or physical linkage (epicPCR) [229]. Another method would be viral tagging [219,230], based on the fluorescent labelling of phages followed by adsorption

to host cells. However, we will not focus on these methods of host identification as they do not help identify phages from an already sequenced sample.

We will focus our attention instead on the analysis of computational signals. As stated beforehand, phages and their host have co-evolved through thousands of years. This shared history is reflected in the genome sequence of both the phage and its host. Therefore, careful analysis of the molecular signals derived from this coevolution can help discern the host of these orphan phage sequences [211]. Multiple methods to assign a host have been developed over the years, based on a multitude of different genomic signatures.

Alignment-based approaches are based on finding regions of homology between the phage genome and a database of sequences, based on the fact that genetic exchange between viral and host genomes is indicative of virus-host associations. When comparing phage genomes to host genomes, these regions of sequence similarity reveal different co-evolution processes. Large similarity regions usually indicate the presence of a prophage closely related to the query phage in the host genome. Small matches require robust search strategies and stringent criteria, since they can be attributed to random chance. For example, CRISPR spacers are sequences 26-72bp long [231], found in specific regions of the genome and their presence implies a successful defence of the bacteria against a closely related phage. Likewise, Phages integrating into tRNAs carry a phage attachment site (attP) that is an exact match of a host tRNA gene (bacterial attachment site, attB). Therefore, a phage carrying an integrase and a putative attP site identical to a host tRNA gene fragment is strong evidence towards that phage-host relationship [153]. The gene content of the viral sequences is also of importance, as certain genes, specially AMGs or other elements of the flexible genome, can be indicative of the host. Examples of these marker genes are photosynthesis genes for cyanophages (*psbA* and *hli*) [34] or the *whi*B transcription factor for Actinobacteria [232].

Alignment-based methods present high accuracy, but low recall [211,233]. This is due to two factors. First, not all phage-host pairs exhibit these signals (for example, not all bacteria encode CRISPRs) [234]. Second, predicting power depends on the database used. Unfortunately, the number of isolate viral genomes in databases remains limited and is heavily biased towards certain clades. For comparison, a paper published in 2015 noted that more than 25,000 bacterial and archaeal host genomes are available in NCBI RefSeq, whereas only 1,531 of their viruses were entirely sequenced and most (86%) of these derive from only 3 of 61 known host phyla [235]. MAGs or SAGs derived from the same sample as the phage are regularly used to complement the cultured genome databases, but this approach risks prediction errors by contamination [236].

Alternatively, alignment-free approaches are based on detecting shared sequence composition features between phage and host, based on the theory that over time, phages adapt their sequence composition to that of their host to make better use of their replication, transcription, and translation machinery [237,238]. A variety of similarity measures such as GC content, k-mer profiles and codon usage have been suggested over the years [239]. Although these methods have the benefit of not needing the presence of a host or a closely-related phage in the databases, they tend to display a lower accuracy than the former approach [211,233]. Co-occurrence of phage and host in a sample is another type of

alignment-free signal, as we only expect to find phages in an environment if their host is also present. However, this method requires the availability of various metagenomes and viromes spaced in time, as the correlation between host and phage abundance is lagged. This reduces their applicability to combined studies such as *Tara* Oceans, where correlation between host and phage were used to assign hosts to eight phage sequences [240]. Moreover, these correlations can be difficult to interpret, especially in complex communities for which phage–host correlation patterns vary depending on sampling date or in phage-host relationships with unusual infection dynamics. Experimental procedures can also affect the analysis, as phage and microbial metagenomes are often isolated and sequenced separately following different protocols, which might confound the signal [241].

Recently, the more sophisticated host prediction pipelines follow an ensemble approach, in which multiple signals (both alignment-free and alignment-based) are evaluated at the same time. As evaluating multiple signatures with distinct properties is a significant challenge, this task is usually delegated to a machine-learning (ML) algorithm. Although these algorithms can process hundreds of different signals at a time, choosing a good set of features is still a crucial step to avoid noise. Therefore, ensemble pipelines are distinguished by their chosen set of predicting features and their algorithm. Examples of these pipelines include BacteriophageHostPrediction, which uses more than 200 features derived from receptor-binding proteins with a variety of ML algorithms [242], VirHostMatcher-Net, which analyses various alignment-free signals with a Markov random field [243].

1.4 The SAR11 clade and its phage

1.4.1 The host: SAR11 clade

The SAR11 clade of Alphaproteobacteria are a group of aerobic, free-living, chemoheterotrophic bacteria [244]. As evidenced by the drab name, they were one of the bacterial groups discovered in a famous metataxomic study in the Sargasso Sea. The prefix SAR refers to the Sargasso Sea, and SAR11 was the 11th clone in the 16S rRNA genome library [244]. Even though the clade passed unnoticed until the appearance of DNA-based surveying, direct cell counts by FISH assays reveal that the clade is the most abundant microorganism in the sea, with SAR11 cells constituting 20 - 40% of all cells in the epipelagic zone, and around 20% in mesopelagic and bathypelagic regions [245].

The phylogeny of the clade is still a hotly contested debate and could easily be the subject for an entire thesis [246–248], but for our purposes, it will suffice to indicate that phylogenomic studies reveal the clade to be exceptionally diverse, including various freshwater clades [249], and that even in the ocean environment there are multiple ecotypes showing preferences for temperature and depth [250–252]. These studies also reveal that surprisingly, the collective core genome of the SAR11 clade is extremely conserved across all members, with only subtle differences in gene content explaining the adaptation of each ecotype to its ecological niche [250,253].

Unlike other microorganisms discovered in the aforementioned study, a subgroup of marine strains (the clade *Pelagibacterales*) were eventually cultured in lab conditions once its unusual growth requirements were elucidated [254]. Once cultured, isolated and sequenced, it was revealed that the main culprit for these extravagant nutrient requirements was its genome. SAR11 bacteria possess dense, very small genomes (*ca*. 1.34 Mbp) that contain only the bare minimum to ensure the reproductive success of the cell [253], with the only exception being the hypervariable regions that contain horizontally transferred genes [253,255]. This was one of the first examples of what is now defined as a streamlined genome, a genome under a selection pressure that favours minimization of cell size and complexity. Examples of this streamlining process can be found in various marine taxa, such as the cyanobacteria *Prochlorococcus* [256], *Ca*. Actinomarinales [257] and the methylotrophic Betaproteobacteria clade OM43 [258]. It is believed that this streamlining confers an advantage in nutrient-poor environments, where either gathering a larger share of resources or using them more efficiently can increase reproductive fitness [259]. As an example, a smaller cell size requires fewer building blocks to replicate and results in a larger surface-to-volume ratio, which is advantageous for scavenging nutrients from the environment [259].

This genome reduction has important consequences in its lifestyle. Depending on the SAR11 strain, various metabolic pathways present gaps (e.g. lack of *thiC* gene for 4-amino-5-Hydroxymethyl-2-methylpyrimidine biosynthesis) [260] or are outright missing in some strains (e.g. glycolysis pathway) [261]. This is the reason for their unusual growth requirements, as the cell must compensate for these incomplete pathways by scavenging these specific metabolites from the environment [244]. However, not all metabolic pathways are equally affected: SAR11 genomes contain a large metabolic repertoire for the capture and oxidation of dissolved organic compounds, particularly the C1 compounds that it uses for energy production [262,263]. SAR11 was the first cultured microorganism shown to code for a light-dependent proton pump proteorhodopsin [255], which uses light to support ATP synthesis in the cell. Surprisingly, the presence of proteorhodopsin has little impact on growing cells but instead allows carbon-starved cells to conserve their biomass [264]. The reason for this behaviour is still unknown, but it is theorised that organic respiration might be incompatible with the proton pump [244].

Genes pertaining to regulation pathways are also sacrificed. This clade has lost the vast majority of its regulation pathways. For example, the clade only codes for two sigma factors and four two-component regulatory systems, compared to seven and 29 in *E. coli*, respectively [244]. Furthermore, the regulation paths that are preserved are simplified variations with fewer genes, or have switched to other regulation mechanisms such as riboswitches [265,266]. The main consequence of the loss of these regulation capabilities is a reduced growth rate, as it has lost the capability to adapt to fluctuations in nutrient concentrations. For example, there is a strong correlation between maximal growth rate and rRNA operon number [267], and SAR11 cells only code for one rRNA operon [255].

Considering the capabilities described above, the ecological niche suggested by Giovanonni *et al.* is the oxidation of one-carbon compounds present at low ambient concentrations,

where they are highly successful thanks to their high surface-to-area ratio and low cost of replication [244].

1.4.2 The parasites: SAR11 bacteriophages

Being the most abundant bacteria in the oceans, it is to be expected that the SAR11 bacteria are under a lot of pressure by predation of phages. However, their defences against phages are also lacking. The only two active defence mechanisms found in the entire clade are a predicted CRISPR region in a SAG from the bathypelagic clade Ic [250] and a DNA phosphorylation system that might protect against phage lysozymes by modifying the DNA backbone [268]. Although their growth rate is not particularly high, the contribution of SAR11 to bacterial heterotrophic production is greater than its relative abundance, which suggests that they would not be able to avoid phage predation.

However, this does not mean that SAR11 is completely defenceless: All pelagibacter strains contain genome region HVR2, a hypervariable region flanked by conserved rRNA genes [269]. This region is rich in genes involved in synthesis of cell surface proteins, such as glycosyltransferases, sugar isomerases and pilins [253]. Considering the large population size and the fact that recombination rates in SAR11 are among the highest ever recorded [270], suggests that the SAR11 clade is able to protect against phage-mediated predation by sheer abundance. High population densities increase predation by phages but they also increase the possibilities of finding other SAR11 cells or their DNA, providing the opportunity for genetic recombination by conjugation/transformation [134]. Recombination is a faster way to propagate genetic elements in populations than clonal replication, and thus potentially offers an advantage with regards to immunity spread [134].

As explained previously, SAR11 are notoriously fastidious to cultivate, which significantly complicates the task of isolating their phages. Nonetheless, the first four SAR11 phages were first isolated from seawater samples in 2013, roughly 10 years after the first axenic SAR11 cultures were reported (**Figure 6**) [134]. Over time, the number of isolated phages has slowly increased bringing the total number to 44 [135,271–274]. These sequences have been supported by several sequences retrieved from a myriad of culture-independent single-cell genomic or metagenomic studies, which have allowed to cement the phylogeny of the isolates [198,273].

The vast majority of isolated phages (41 out of 44) belong to the *Podoviridae* family and can be divided in three distinct clades (all names are tentative): HTVC019P-like, which belong to the *Autographivirinae* subfamily, forming a clade close to the P60-like cyanophages [275]; HTVC010P-like, divergent podoviruses [135]; and HTVC023P-like, which include vSAG37-F6, the most abundant phage in the marine environment [198,272]. Both HTVC019 and HTVC010P-like clades include freshwater members, found by metagenome mining [273]. Lysogeny seems to be widespread among all three clades, as evidenced by PCR assays and mining of sequences that include hybrid *attL* and *attR* sites [271,276]. The *attB* integration site is usually found near a tRNA gene in the host. The remaining three phages belong to the *Siphoviridae* (Kolga phage) [274] and the *Myoviridae* (HTVC008M, Mosig phage) [134,274],

respectively. Both HTVC008M and Mosig belong to the *Tevenviridae* subfamily and are phylogenetically related to other isolated cyanobacteria myoviruses (S-SSM7, P-SSM2, Syn1) [134,274], while Kolga is not related to any other known viruses [274].

Recruitment analyses show an interesting pattern: while pelagiphages seem to be widely distributed in the epipelagic zone across the longitudinal gradient, a few groups are highly abundant (HTVC010P-like and HTVC023P-like lytic subgroups), while the rest of phages present much lower recruitment values [134,277,278]. An exception to this rule is found in some niche-specific sequences, as seen with Pelagipodophage *Greip* being exclusively found in arctic datasets [274].

Significant differences were found between the podophages and the myophage in regards to burst size (40+ for podophages, less than 10 for the myophage) and latent times (an average of 20 hours for podophages, 17 hours for the myophage) [134]. Values for both groups, however, are lower than those of their closest neighbours, the cyanophages [279]. Nonetheless, the most puzzling difference between isolated phages has to be in their infection processes. The first SAR11 phages presented infection dynamics similar to those of cyanophages, in which host density in the infected culture falls as a consequence of viral lysis [134]. In contrast, other pelagiphage infections result in the host culture growing into a steady state, but at a lower cell density than uninfected cultures, with no association to phage phylogeny or gene content related to lysogeny [274]. This pattern of infection has also been reported in crAssphages cultures isolated from the human gut, with their authors proposing that these viruses may only complete the viral cycle in a subset of host cells, while the rest present alternate interactions such as pseudolysogeny [274]. The switch from this dormancy state to lytic cycle would be controlled by a yet unknown set of genes. This theory would also explain the low lytic activity and the decoupled abundances related to their hosts both observed in pelagiphages *in situ* and in production experiments [277]. In fact, SAR11 prophages recently detected in SAR11 strains NP1 and NP2 have been found to increase virion production under carbon-replete conditions in a process that is not accompanied by an increase in host cell lysis [276].

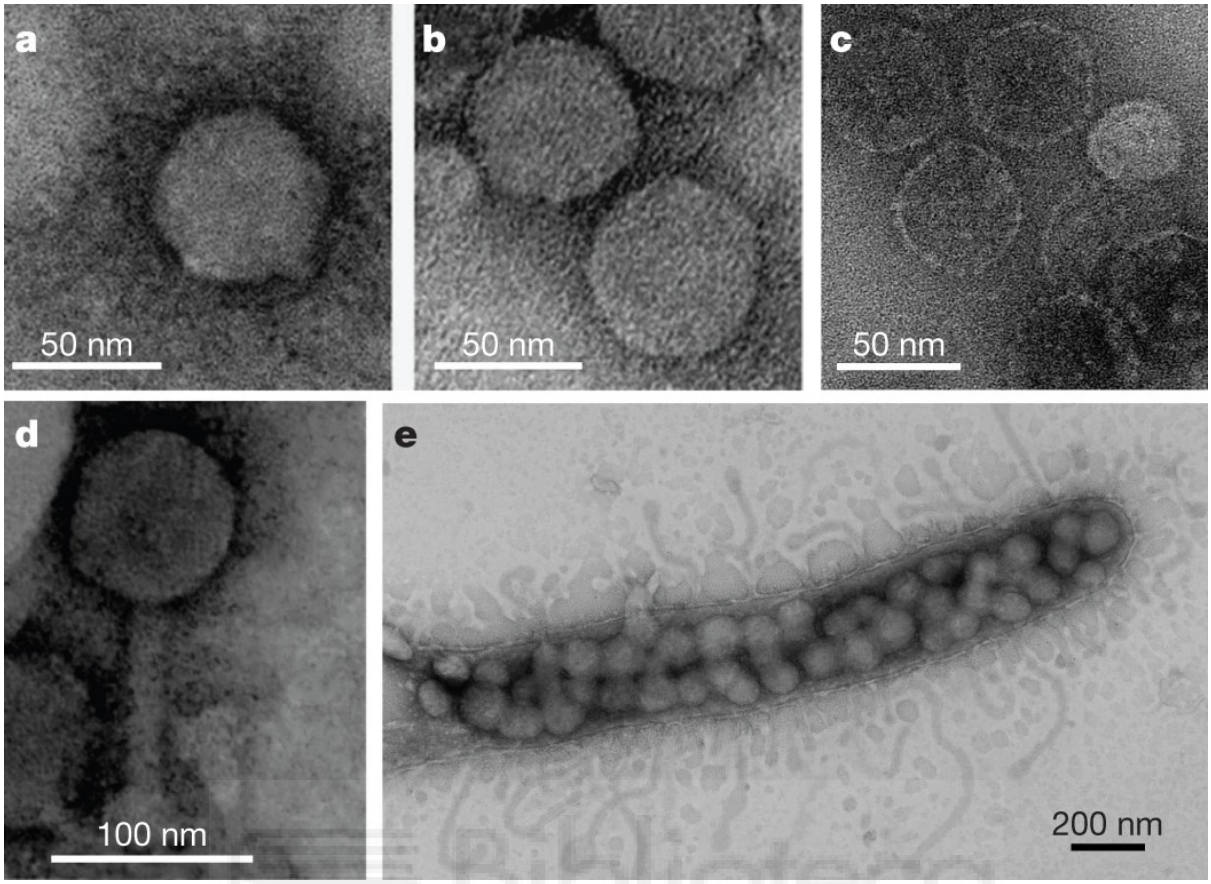


Figure 6. Electronic microscopy images of SAR11 phages. **a**, Pelagipodovirus HTVC011P. **b**, Pelagipodovirus HTVC019P. **c**, Pelagipodovirus HTVC010P. **d**, Pelagimyovirus HTVC008M. **e**, Host cell of '*Candidatus* P. ubique' HTCC1062 infected with HTVC011P immediately before lysis. [134]

Like other marine viruses [280,281], pelagiphages show diel cycling. Metatranscriptomic datasets from the North Pacific Ocean and the Osaka Bay show that pelagiphage abundance and transcriptomic activity closely follows that of its host, with a peak abundance/activity at night and a nadir around midday [282]. This pattern has been observed in other heterotrophic bacteria, probably as an adaptation to diel cycling of autotrophs in the community [281,283,284].

2. Objectives

Viruses are the most abundant biological entities in marine ecosystems and play an essential role in global geochemical cycles. They have important ecological functions as drivers of bacterial populations through lytic infections and contribute to bacterial genetic diversification. Unfortunately, their study is severely limited by the difficulty to culture and isolate them in lab conditions. Culture-independent techniques such as metagenomics can complement culture-based approaches to capture more phage diversity. However, the vast majority of viral sequences recovered through these methods are uncharacterized and therefore do not provide any information about their interactions with the bacterial community ("viral dark matter"). For this reason, it is essential to develop methods that can provide context to viral metagenomic data to describe and quantify all biodiversity as well as to improve our understanding of the processes of diversification and viral evolution.

The objectives of this thesis are thus:

- Develop an approach to detect and extract all viral diversity from metagenomic samples, obtain more complete genomes by reassembly and provide context to the recovered sequences.

- Application of the previous approach to phages of the order SAR11 *Pelagibacterales*, one of the most ubiquitous lineages of free-living heterotrophic bacteria in the world's oceans.

- Evaluate the resolving power of third generation sequencing (long reads, PacBio CCS) and compare it with short reads (Illumina sequencing) to study the metagenome of a known marine sample from the mixed epipelagic water column of the winter Mediterranean with regards to viral recovery.

- Develop bioinformatic tools to explore and discover virus-host interactions.

3. Materials and methods

3.1 Viral genome mining

Two different approaches were followed to obtain viral contigs, depending on the degree of specificity desired. To recover Pelagimyophage (PMP) contigs from metagenomic samples, we followed the workflow shown in **Annex 1**:Figure S1A. The reference cultivated PMP genome (HTVC008M) [134] and metagenomic PMP sequences MAVG-2, MAVG-4, MAVG-5, and Io7-C40 [193], were used as bait to comb through a vast quantity of contigs derived from several metagenome and virome samples (**Annex 1**:Table S1) [153,193,285–289]. First, a hidden Markov model (HMM) made from an alignment of *terL* gene sequences was used to identify viral contigs larger than 5 kb. The *terL* gene from the extracted contigs was then used to construct a phylogenetic tree (**Annex 1**:Figure S7A). The position of the *terL* gene of the reference PMP in this tree was then used to recover a set of candidate contigs (**Annex 1**:Figure S1B and **Annex 1**:Figure S5). At the time this work was made, the closest taxa to PMPs were Cyanomyophages (CMPs), which are expected to be present in significant quantities in the surveyed metagenomes. To remove all CMP-related contigs from the candidates, two collections of gene clusters were built, (i) one of them derived from 28 CMP genomes downloaded from the NCBI Refseq database [290] and (ii) another derived from the reference PMP genomes. Gene clusters shared between both collections were removed. HMMs built from both cluster collections were used to classify the contigs, keeping only those that had at least a match to a PMP gene cluster and no matches to any CMP gene cluster (**Annex 1**: Figure S1).

Conversely, recovery of viral diversity from long-read datasets was performed in two steps. Bacteria and archaea viral contigs larger than 1kb were recovered using VIBRANT [291] with default parameters. Eukaryotic viruses were recovered via manual curation. Each dataset was dereplicated using CD-HIT [292] at 95% identity to remove redundant sequences. Contigs were considered unique based on the definition of 'Viral population' as described in Gregory *et al* [184], that is, contigs were considered part of the same population if they had hits with at least 95% identity and the sum of distinct alignment lengths resulted in a coverage of at least 70% across the smallest contig using BLASTN [293].

3.2 MAVG cross-assembly

To obtain PMP genomes as complete as possible, the contigs recovered from the sequence mining step were subjected to a cross-assembly step. Identical sequences were removed from the analysis, always keeping the longer contig if they did not have the same length. Contigs were then separated into bins of overlapping contigs based on an all-versus-all comparison (**Annex 1**:Figure S1). Next, bins were assembled manually into MAVGs as described previously [193] provided that (i) overlaps between contigs had a nucleotide sequence identity of >99%, an alignment length of >1,000 nt, and gaps of <10 nt, (ii) all overlaps were corroborated by more than two contigs, and (iii) sample metadata were ecologically coherent for all involved contigs (for example, not assembling contigs from freshwater and marine samples together).

3.3 Recruitment analysis

To assess the distribution and abundance patterns of recovered phage sequences, filtered raw reads were aligned against a phage genome. By quantifying the number of reads aligned to each genome, it is possible to obtain a measurement of abundance of that genome across several metagenomic datasets. Genomes were recruited using pblat [294] with a threshold of 95% nucleotide identity over at least 50 nucleotides. Each read was mapped only to the viral contig with the best match. Normalisation was performed by calculating RPKG (reads recruited per kilobase of the genome per gigabase of the metagenome) so recruitment values could be compared across samples. Marine phages were recruited against a collection of datasets that include the *Tara* Oceans metagenomes [286], GEOTRACES cellular metagenomes [285], and a collection of Mediterranean metagenomes [105,193]. For the freshwater PMP group PMP-D, genomes were also recruited against the virome data sets they were recovered from and against samples from other freshwater environments (Lake Biwa, Lake Simoncouche, Lake Kivu, Baltic Sea) [189,295–297].

Linear metagenomic recruitments (see **Annex 1**:Figure S3) were performed by alignment of reads using BLASTN [293], with a cut-off of 70% nucleotide identity over a minimum alignment length of 50 nucleotides. The resulting alignments were plotted using the ggplot2 package in R.

3.4 Genome Functional Annotation

Genes and tRNA sequences were identified using Prodigal [298] in metagenomic mode and tRNAscan-SE [299], respectively. As phage protein sequences are extremely divergent, functional annotation of predicted features followed a consensus-based approach. Phage protein sequences clustered at 30% identity, 50% query coverage using MMSeqs2 [300]. Clusters with less than 10 sequences were expanded with MMSeqs2 and the uniclust30 database [301].

Individual proteins in each cluster were annotated against functional annotation databases. Proteine were first annotated against the uniref90 database [302], using DIAMOND [303]. A second round of annotation was done with the Conserved Domains Database (CDD) [304] serving as a general function database, while pVOGs [305] PHROGs [306] served as phage-specific databases. Protein alignments were downloaded from their respective databases, then converted to Hidden Markov Models (HMMs) using hmmbuild from the HMMER suite [307]. For each database, each gene was assigned the best hit with an E-value of at least 10^{-5} and a query coverage of at least 50%. Annotations for each cluster were manually curated to ensure the annotations were coherent for all proteins in the cluster. In the cases that presented discrepancies, the second and third best hits were used to verify the annotation. Finally, the remaining clusters without annotation were annotated using HMM vs HMM annotation, a much more sensitive procedure. Clusters were finally converted into HMMs then compared to the PDB HMM database [308] using the HH-suite3 software suite [309].

3.5 Co-occurrence matrix

In order to analyse the organisation of phage genomes, we constructed a co-occurrence matrix, which links genes if they are part of the same operon. Operon determination was done as follows: Terminator sequences were predicted using Transterm_HP [310], while early promoter sequences were predicted using BPROM [311]. Prediction of middle and late promoter sequences was attempted following the steps described previously [312] but was unsuccessful in PMP genomes. As prediction of all promoter regions proved unfeasible, genes were grouped into operons based on terminator positions and strand changes. These operons were then used as the basis for a co-occurrence matrix. Two protein clusters (nodes) were linked to each other if they were present in two genomes and were part of the same operon, with edge strength representing the number of genome pairs where this was the case. Edges with edge strength representing 0.05% of the total were removed from the matrix. The matrix was then used to build a network in Cytoscape [313]. The add-on ClusterMaker2 [314] was used to separate the co-occurrence network into clusters (MCL algorithm, 2.5 granularity).

3.6 Phylogenetic reconstruction

Different methods to construct phylogenetic trees were used depending on the evolutionary distance between the genomes and the type of sequence information available. If the genomes available are reasonably complete and phylogenetically closely related (e.g. pertaining to the same family), as was the case for the PMP tree (**Annex 1**:Figure 1) and the nucleocytoplasmic large DNA viruses (NCDLV) tree (**Annex 2**:Figure S1) a phylogenetic tree based on a concatenate of conserved proteins was constructed. Marker proteins common to all genomes are found with software such as GET_HOMOLOGUES [315] or ncldv_markersearch [316]. For the *Tevenviridae* tree, the chosen proteins were the large and small subunits of terminase, VrlC protein, tail tube monomer gp18, and baseplate wedge protein gp8, while the NCDLV tree includes the major capsid protein mcp, DNA Polymerase beta subunit PolB, DEAD/SNF2-like helicase SFII, Poxvirus Late Transcription Factor VLTF3 and the Packaging ATPase A32. The recovered marker proteins are aligned using a protein aligner such as MUSCLE [317] or ClustalOmega [318], then concatenated. The resulting sequences are then used as input for a phylogenomic inference software, such as IQ-TREE2 [319] or FastTree [320], which will build a phylogenetic tree. Confidence on the tree branches is assessed by producing bootstrap replicates by building the tree again after reshuffling nucleotides between sequences.

This approach is not feasible in cases in which genomes are distantly related to each other, as the divergence in phage protein sequences results in poor detection of homologues. To recover as many homologues as possible, Benler *et al.* describe the following method [321]. Marker viral proteins in the target genomes are detected via hmmsearch [307] against the PHROGs database [306] and merged into a single dataset. This dataset is then grouped with mmseqs2 [300] into clusters with 50% amino acid identity and a coverage of 70%, which are then aligned using ClustalOmega [318] and compared to each other using hhsearch [309]. A distance matrix is calculated by calculating distances following the formula $-\ln(S_{A,B} / \min(S_{A,A}, S_{B,B}))$, where $S_{A,B}$ is the raw score per alignment length. This matrix is then used to build a

dendrogram (UPGMA method), which acts as a guide to merge clusters using ClustalOmega [318], resulting in larger protein alignments. The resulting protein alignments were filtered to remove sites with more than 50% gaps, then used to build trees using phylogenomic inference software.

3.7 Protein analysis

3.7.1 Putative endolysin discovery and analysis

Putative endolysins in the Mediterranean datasets were extracted following the method described in Fernandez-Ruiz *et al* [322]. The predicted proteins from each contig were compared against a curated database of endolysins using DIAMOND [303]. Matches were classified as putative endolysins if the match had >50% identity, covered at least 30% of the query sequence, the alignment was at least 50 aa long and the e-value was at least 10^{-3}. A phylogenetic tree including both the reference dataset and new putative sequences was built following the method described above (See *Phylogenetic reconstruction of viral marker proteins*). Protein domains were detected by using hmmsearch [307] against the CDD [304] and dbCAN2 [323] databases, considering a match as valid if the match had 70% of HMM coverage and E-value of at least 10^{-5}. Proteins of the C4 clade were tested for the presence of a Signal-Arrest-Release domain following the method explained in Oliveira *et al* [324].

3.7.2 Protein isoelectric point determination

To determine the isoelectric point distribution patterns of the phage genomes, calculations of all predicted proteins for both genomes were calculated with the pepstats software from the EMBOSS package [325]. The resulting isoelectric point values were plotted using the ggplot2 package in R.

3.8 Genomic pairwise comparison

Average nucleotide identity (ANI) and coverage between a pair of genomes were calculated using the Jspecies software with default parameters [326].

3.9 Statistical testing

Wilcoxon rank sumStatistical tests were performed using the coin package in R [327]. The Effect size for Kruskal-Wallis test was calculated using the rstatix package (https://CRAN.R-project.org/package=rstatix).

3.10 Host Assignment

Different host assignment protocols were used depending on the objectives of the study. For phage sequences obtained through mining with phage-specific HMMs (See **Annex 1**:Figure S1A), we were only interested in recovering phages that infected SAR11. Therefore, the

recovered phages filtered by size (>100 kb), GC content (30 to 35%, which is the GC% range of SAR11), the number of proteins matching to SAR11 (>70% of identity), and tRNA gene matches to SAR11 (>95% of identity).

Larger collections of viral contigs were taxonomically annotated following the method described in Beaulaurier *et al* [208]. The predicted proteins from each contig were annotated against the NCBI Viral Genomes database [290] using LAST [328]. Viral contigs were annotated at the order level if they contained one, three or five or more proteins with top hits to phages that infect the same host genus. The choice of threshold seems to only affect the number of phages classified, not the community composition (**Annex 3**:Figure 1).

Lastly, the GLUVAB database [329] used for training and validation of RaFAH were assigned putative hosts using classical approaches. We used three lines of evidence for virus-host associations: CRISPR spacers, homology matches, and shared tRNAs. CRISPR spacers were identified in the RefSeq genomes as previously described [330]. The obtained spacers were queried against the sequences of bona fide viral sequences using BLASTN v2.6.0+ (task blastn-short, 100% identity, 100% query coverage, no mismatches) [293]. Homology matches were performed by querying viral sequences against the databases of prokaryote genomes using BLASTN (alignment length ≥ 500bp, identity ≥ 95%, evalue ≤ 0.001) [293]. tRNAs were identified in viral scaffolds using tRNAScan-SE [299] using the bacterial models. The obtained viral tRNAs were queried against the RefSeq database of prokaryote genomes using BLASTN (alignment length ≥ 60bp, identity ≥ 97%, mismatches ≤ 10) [293]. These steps for host assignment did not include the prophages in the Genomic Lineages of Uncultured Viruses of Archaea and Bacteria (GLUVAB) database, as we were already confident of their host assignments.

All GLUVAB genomes were clustered into viral populations (VPs) on the basis of 95% average nucleotide identity and 80% shared genes [183]. For each virus-taxon association signal detected (i.e., homology, tRNA, or CRISPR), 3 points were added to the taxon if it was a CRISPR match, 2 points if it was a homology match, and 1 point if it was a shared tRNA. The taxon that displayed the highest score was defined as the host of the viral population. With this approach we ensured that all the genomes in the same VP were assigned to the same host and that no sequences had to be excluded due to ambiguous predictions.

3.11 Random Forest input dataset

To build the input table used to train the random forest model, protein sequences were identified in viral genomes using Prodigal [298] in metagenomic mode. Proteins were then clustered using the MMseqs2 software suite into clusters with at least 35% sequence identity and an alignment coverage of at least 70% of all proteins [300]. The resulting following Protein Clusters (PCs) were aligned with QuickProbs [331] using default parameters, then converted into HMMs using the hmmake program from the HMMER suite [307]. The HMMs obtained in this way were annotated against the pVOG database [305] using the HH-suite3 software suite [309], keeping all annotations with target coverage ≥50% and e-value ≤1^{-10}.

Finally, individual viral proteins were mapped to the HMM profiles using the hmmsearch program from the HMMER suite [307], keeping all hits with e-value ≤10^{-5}, alignment length ≥70% for both the viral protein and the HMM, and minimum score of 50. These results were parsed into a matrix of viral genomes × PCs in which the values of each cell corresponded to the bit score of the best hit of each protein to a given PC, or zero if the protein and the PC did not match or if the score of the match was below the aforementioned 50 cutoff. Once the matrix of genomes × PC was defined, we calculated Pearson correlation coefficients (r) between all possible pairwise combinations of PCs. To remove redundancies, we grouped PCs into superclusters if they presented r ≥ 0.9, and only a single PC from each supercluster was kept for subsequent analysis. This reduced table of genomes versus PC scores (25,879 genomes × 43,644 PCs) was used as input to train, validate, and test the random forest models. The taxonomic classification of each genus up to the domain level was obtained by parsing the NCBI Taxonomy database with a custom script.

3.12 Random Forest model training

Random forest models were built using the Ranger [332] package in R. The response variable was the genus-level host assignment of the viral sequences while the input parameters were the scores of viral genomes to each PC. To ensure that the resulting random forests could be used for all virus genomes, the multi-class random forests were built with 1,000 trees, 5,000 variables to possibly split at in each node, and using probabilistic mode. When training the models and reporting predictions, we assumed that a virus can only infect a single genus. Due to the probabilistic nature of the random forests, all genera are associated with a score (which ranges from 0 to 1). The putative host of a viral genome was selected as the taxon with the highest probability score yielded by the random forest. Variable importance was estimated using the impurity method.

Three models were built and validated on independent datasets. Models 1 and 2 were used as proof-of-principle models, and Model 3 was the definitive model used for testing and which is provided to the users and used for all subsequent analyses. Model 1 was trained on Training Set 1, which comprised 80% randomly selected non-redundant viral genomes from NCBI RefSeq. The performance of this model was evaluated on Training Set 1 and Validation Set 1, which comprised the remaining 20% of non-redundant RefSeq genomes. Model 2 was trained on Training Set 2, which comprised 100% of the RefSeq genomes, and validated on Validation Set 2, which was comprised of genomes from the GLUVAB database that could be assigned to a host at the level of genus by the pipeline described above. Finally, Model 3 was built based on Training Set 3, which comprised all of the RefSeq viral genomes and the GLUVAB genomes that could be assigned to a host at the level of genus (i.e., a combination of Training Set 2 and Validation Set 2). In this dataset each genus was represented by a median of three genomes, and for 187 out of 617 (30.3%) genera the model was trained with a single genome (**Annex 3**:Table S5).

3.13 Random Forest model testing

Viral genome completeness is likely to influence the performance of the models. A tool trained solely on complete or nearly complete genomes might not be capable of producing accurate predictions for the genome fragments that are often obtained with metagenomic datasets. Completeness of the 25,879 sequences used to train RaFAH was estimated with CheckV [333] which indicated that this dataset encompassed both complete viral genomes as well as partial viral contigs. Partial viral genomes were the majority of sequences used to train RaFAH. Altogether, the genomes used for training displayed an average completeness of 53.6% ± 32.3%. According to CheckV, these sequences were classified as complete genomes (709 sequences), high-quality genome fragments (5,823), medium-quality genome fragments (5,493), low-quality genome fragments (13,707) and not determined (147).

We used three independent test sets to evaluate the performance of RaFAH Model 3. Test Set 1 comprised a non-redundant dataset (95% nucleotide identity and 50% alignment length of the shorter sequence) viral genomes retrieved from NCBI Genomes database [290]. We took several steps to make sure that Test Set 1 represented a challenging dataset for the random forest model by excluding any genomes made public before November 2019 or that shared more than 70% of proteins (>70% average AAI) with genomes present in RaFAH Model 3. These steps resulted in an independent Test Set 1 consisting of 561 (out of the initial 3,427) genomes. Host assignment was derived from NCBI Annotation.

Test Set 2 comprised viral genomes identified in SAGs from marine samples [196]. A total of 418 viral sequences were extracted from 4,751 SAGs (with completeness ≥50% and contamination ≤5% as estimated by CheckM [334]) using VIBRANT [291]. Host assignment was derived from the SAGs, as we assumed that the viral sequences in the SAGs infected the organisms from which these SAGs were derived, either because they were derived from integrated prophages or from viral particles attached or inside host cells. Viruses from SAGs that could not be classified were excluded from the precision and recall analyses.

Test Set 3 comprised a collection of 61,647 viral genomic sequences from studies that spanned multiple samples from permafrost [335], marine [336], human gut [321], freshwater[94] , soil [142], hypersaline lakes [337], hydrothermal springs (Fredrickson *et al.*, unpublished data obtained from IMG/VR [288]), and sludge bioreactor [338] habitats. These sequences were assigned to putative hosts through a classical host-prediction pipeline (See *"Host assignment"*, above). Bootstrap analysis was applied to evaluate the precision of RaFAH in this last dataset. We assumed that the hosts predicted by the classical approaches were the true hosts of the viral genomes on Test Set 3. Random subsamples representing 20% of the full data were generated in 1,000 replicates. Precision was estimated for each replicate.

4. Results

4.1 Results derived from the work "Metagenome Mining Reveals Hidden Genomic Diversity of Pelagimyophages in Aquatic Environments"

4.1.1 Summary

In this study, a collection of metagenome and virome datasets (Ca. 45 Gbp) were surveyed to recover myophages similar to the SAR11 phage HTVC008M (Pelagimyophages, PMPs). Despite being is one of the most abundant bacterioplankton groups in surface waters, only 15 SAR11 phages have been isolated thus far, and only one of them belongs to the Myoviridae family. We recovered 26 new myophages that putatively infect the SAR11 clade, including the first that putatively infect *Candidatus* Fonsibacter (freshwater SAR11) and another group putatively infecting bathypelagic SAR11 phylogroup Ic. The recovered genomes have similar sizes and maintain overall synteny with HTVC008M despite low average nucleotide identity values, revealing high similarity to marine cyanomyophages (CMPs). Contrary to CMPs, PMP genomes include a large hypervariable region in the tail structural region that contains genes related to the host cell wall. Proteins from freshwater representatives display an isoelectric point shift, suggesting an adaptation to freshwater. Interestingly, 25 of the 26 new sequences have been recovered from datasets derived from the cellular fraction and not from the viral fraction, which could explain their poor representation in databases. Supporting this theory, recruitment analysis of the phages reveal they are widely distributed across marine environments but always much more from datasets derived from the cellular fraction, a pattern distinct to that of other related phages. A co-occurrence gene network was built to analyse the gene content of PMPs, finding some AMGs unique to this group as the 30S ribosomal protein S21 and a cluster of curli-related proteins. The function of this last cluster is not fully understood.

4.1.2 Genomic features of PMPs

MAVG completeness was verified either by the presence of identical repeated sequences (>10 nucleotides) at the 5'- and 3'-terminal regions or by showing a similar synteny and gene content to the cultivated PMP HTVC008M [134]. The genome size of the 13 complete genomes ranges from 132 to 164 kb (**Annex 1**:Table 1). To study the relationships of the recovered phages, the 31 PMP genomes were compared in a phylogenomic tree using four CMP genomes as an outgroup. The five proteins common to all 35 genomes (large and small subunits of terminase, VrlC protein, tail tube monomer gp18, and baseplate wedge protein gp8) were merged into a concatemer. The phylogenomic tree clustered PMPs into five different groups (PMP-A to PMP-E), with group PMP-A containing the reference phage HTVC008M (**Annex 1**:Figure 1). Host assignment within different SAR11 subclades was not possible (except for group D [see below]) due to (i) lack of tRNA genes (only 18 genomes had them, and the ones present were all under 95% identity to SAR11 known tRNAs), which suggests that either we do not have genome representatives for the hosts they infect, or they have a broad host range, (ii) similarity of shared proteins provided inconclusive results (same identity to distantly related host-groups) and (iii) there is only one report of a CRISPR-cas

system in SAR11, which is found only in the bathypelagic ecotype Ic [250]. The enormous diversity of the SAR11 clade probably complicates the process of host assignment.

Annex 1:Figure 2A shows the alignment of two genomes of group PMP-A (one of them the pure culture HTVC008M), while alignments of one representative genome from each cluster are shown in **Annex 1**:Figure 2B. Overall, synteny was well preserved in all sequences once they were rearranged to start from the major capsid gene (gp23), and all of the sequences displayed the characteristic patchwork architecture of the Tevenvirinae subfamily, with remarkably conserved core modules (DNA replication and virion structure) separated by variable regions, designated as hypervariable [54,339] (**Annex 1**:Figure 2A and B). The most remarkable feature is the presence of a large non syntenic island located in the middle of the structural region, always between the VrlC gene and the neck protein gene gp14 (**Annex 1**: Figure 2C). On the basis of its variable character and the presence of tail fibers, we have designated this variable region the host recognition cluster (HRC) (**Annex 1**: Figure 2C). In other T4-like phages, this region contains only the tail fiber module [54,340]. This large hypervariable region has been already described in CMPs, usually containing several structural genes and AMGs [340]. In PMPs, this region is larger (mean HRC size of 44.6 kb versus 34.2 kb in CMPs), and contains, along with the expected tail fiber genes, a large number of genes seemingly unrelated to the tail fiber module, the most conspicuous of which are several glycosyltransferases, typically involved in the synthesis of the O-chain of the lipopolysaccharide that is located in the outer layer of the Gram-negative cell envelope [80,312] (**Annex 1**: Figure 2C). In PMPs, 63 out of the 162 lipopolysaccharide (LPS)-related proteins found are inside the HRC, while CMP HRCs have more identifiable tail fiber-related proteins. However, the latter could be attributed to the fact that CMPs are better represented in the sequence databases and are thus easier to annotate. The comparison of the CMP and PMP genomes showed strong conservation of all modules, including the HRC (**Annex 1**:Figure 3A). However, unlike the latter, in some CMP genomes, the baseplate module is divided by another plastic region (**Annex 1**:Figure 3A).

The two most similar complete genomes were MAGV3 and MAGV16, found in cluster B (average nucleotide identity [ANI] of 72.0% and coverage of 38.6%), although they were assembled from the Western Arctic ocean and the Mediterranean Sea, respectively (Fig. 3B). In the case of these two, the HRC was much more similar and differed only by the addition of some gene cassettes related to radical SAM (S-adenosyl-L-methionine) proteins (**Annex 1**:Figure 3B). Their comparison seems to indicate that the divergence of this region is a gradual process rather than a complete replacement, as described for replacement flexible genomic islands in prokaryotic cells [341]. The genes located downstream from VrlC, which are the tail fibers in most genomes, show high similarity, indicating a possible host overlap of these two phages.

4.1.3 Recruitment from cellular metagenomes and viromes

To evaluate the abundance and elucidate possible patterns of distribution of these phages, we performed recruitment analysis by comparing each sequence to 395 metagenomes from Mediterranean depth profile [105,342], Tara Oceans [343] and Geotraces [285] data sets as

well as several freshwater metagenomes (see *2.3 Recruitment analysis*). We considered only those samples where at least one PMP recruited more than five reads per kilobase of genome and gigabase of metagenome (RPKG) with an identity of >95%. PMP genomes showed a wide, if uneven, oceanic distribution along the Tara Oceans transect [343] (**Annex 1**:Table S2). All genomes except the freshwater PMP-D group (see below) recruited significantly in several marine samples from different geographic regions, with maximum recruitment typically found in the 5-to-45-m-depth range. Figure 4A shows the recruitment of both families of SAR11 phages (Podoviridae and Myoviridae) and their host in both the cellular and viral fractions from Tara Oceans. In addition, we have also included the other most relevant and widespread marine group, Cyanobacteria, and their myophages. While the presence of podophages was mainly restricted to viromes, both groups of myophages were present in both fractions (cellular and viral) (**Annex 1**:Figure 4A), although pelagimyophage genomes recruited significantly more from cellular metagenomes than from viromes. The abundance of viral DNA in the cellular fraction indicates that a high number of microbial cells are undergoing the lytic cycle, which acts as a natural amplification of viral DNA [153,193]. Another interesting observation was that a significant amount of SAR11 DNA was present in viromes, probably because some SAR11 cells might be small enough to pass through the 0.2-µm filter used frequently to retain bacteria (**Annex 1**:Figure 4A) [244,344]. A latitude transect from 50°N to 50°S in the West Atlantic Ocean was analyzed using the Geotraces database [285]. However, latitude did not seem to be a significant factor in their distribution (**Annex 1**:Table S3).

The recruitment results as a whole suggest that PMP amplification is biased, as this group of genomes always recruited much more from cellular metagenomes than from viromes. The nature of this bias (either biological or technical) is still unclear. We also observed significant differences in recruitment values between the Mediterranean viromes treated with multiple displacement amplification (MDA) and those that had not been amplified (**Annex 1**:Fig. 4B). Although there is no direct evidence of their effect over myoviruses, MDA amplification might have played a part in these differential recruitment. MDA has been reported to be biased toward certain nucleic acid structures and sequences [161,345].

However, we were able to distinguish some groups with different patterns of recruitment. One genome of group PMP-A (PMP-MAVG-4) predominantly recruits below 200m in both the Geotraces and *Tara* Oceans data sets, supporting its association to bathypelagic Pelagibacterales clade Ic [250] (**Annex 1**:Figure S2; **Annex 1**:Tables S2 and S3), although the assignment is tentative, since it could not be proven by sequence analysis. Due to the scarcity of samples from the deep ocean, we can confirm its presence only in temperate zones of the Pacific and Atlantic Oceans (**Annex 1**:Tables S2 and S3). In Mediterranean samples, it appears only in areas below the deep chlorophyll maximum (75 to 90 m) but not at bathypelagic depths, probably due to the Mediterranean relatively warm water column, although Ic representatives have been detected there (Fig. 4B) [346]. Unique genes to this putatively "deep ecotype" include a GMP reductase and various genes involved in heme biosynthesis (coprophyrinogen oxidase, porphobilinogen deaminase) as well as a formate dehydrogenase, an enzyme that transforms formate into CO_2 and $2H^+$ [347]. This could be an adaptation to generate a proton gradient in the absence of light, as SAR11 cells can generate it via

rhodopsins. Two other PMP-A representatives, MAGV05 and Io7-C40, showed tolerance for brackish waters, as demonstrated by their recruitment from Baltic Sea cellular metagenomes (**Annex 1**:Figure 4B). Group D recruits only from freshwater samples, making them the first described freshwater myophages of the SAR11 clade (see below) (Fig. 4B). Linear recruitments (**Annex 1**:Figure S3A) showed that although genomes recruit along their entire lengths, most of the reads were recruited at more than 99% identity. The genome regions that recruit vertically down to 80% identity correspond to the structural and DNA replication-related genome regions described previously, which are very well conserved among all the members of the subfamily [54,312]. The HRC usually underrecruited, indicating the highly variable nature of this region (**Annex 1**:Figure S3A). The same pattern was observed in cellular metagenomes and viromes with and without MDA (**Annex 1**:Figure S3A).

4.1.4 First genomes of PMPs infecting Ca. Fonsibacter

Genomic analysis of the two genomes in group PMP-D showed that both contained tRNA genes with the best match to tRNAs from the recently isolated Candidatus Fonsibacter ubiquis LSUCC0530, a member of the LD12 subclade [348]. Metagenomic recruitment showed clear evidence that group PMP-D was associated with freshwater samples (**Annex 1**:Figure 4B). To our knowledge, these are the first genomes of myophages that putatively infect Ca. Fonsibacter (fonsimyophages). Both are remarkably similar to each other but present different degrees of completeness. PMP-MAVG-15 is considered complete, while PMP-MAVG-20 is lacking the DNA replication module. Recently, a shift toward basic values was described in the relative frequency of predicted isoelectric points when comparing freshwater and marine microbes [349]. Along these lines, we found a significant difference in PMPs infecting *Ca*. Fonsibacter compared to the reference genome HTVC008M (**Annex 1**:Figure S3B). However, synteny was well preserved between marine and freshwater groups (**Annex 1**:Figure S3C).

Recruitments show the recovered fonsimyophages to be present in various lakes from Canada (Erie, Ontario, Simoncouche) in both the cellular and viral fraction (**Annex 1**:Figure 4B). We also found recruitment matches at lower identity (<80%) in other freshwater samples (Lake Biwa, Lake Kivu). Linear recruitments for group D phages against freshwater viromes are different from those originating from their marine counterparts (**Annex 1**:Figure S3), showing that diversity in fonsimyophages is lower than that of the marine PMPs. This fact might reflect the reduced intrapopulation diversity of their host compared to other SAR11 subclades [348].

Gene content comparisons between marine or freshwater SAR11 PMPs shed little light on possible adaptations to the latter. However, the freshwater genomes do not contain genes related to LPS, substrate transport, radical SAM proteins, or the curli operon (see below). Nevertheless, it has some unique genes, such as speH (involved in polyamine salvaging), various genes involved in lipid biosynthesis (fabF, stearoyl-coenzyme A [CoA] desaturase) and a 2OGFeDO superfamily protein, which catalyzes nucleic acid modifications [350,351]. Strikingly, some proteins core to all PMPs (peptide deformylase, ribosomal protein S21, and aspartyl/asparaginyl beta-hydroxylase) are present in group PMP-D but are different enough to be separated in independent protein clusters.

4.1.5 Comparative genomics

To maximize our ability to annotate phage proteins, we clustered orthologous genes into protein clusters (PCs) and annotated their function following a consensus-based approach (see *2.4 Genome Functional Annotation*). The PCs with the most differences in abundance between PMPs and CMPs have been collected in **Annex 1**:Table S4. Furthermore, to examine the organization of the PCs into operons in both groups of phages, we built a co-occurrence matrix (**Annex 1**:Figure S4A), which links genes if they are in the same operon. Previously described methods to detect middle and late promoters in CMPs [312] did not provide satisfactory results when applied to PMPs, so we delimited operons by terminators and strand changes (see Materials and Methods). The co-occurrence matrix reveals differences in the structural organization of the operons containing conserved PCs. While structural operons contain only structural or hypothetical proteins, operons containing DNA metabolism genes are more diverse, containing AMGs of various types. Furthermore, genes involved in the same function are not in the same operon unless they are subunits of the same protein or the presence of one is meaningless without the other. An example of this phenomenon would be the photosynthesis-related AMGs in CMPs. Photosystem II D1 and D2 subunits are always in the same cluster, but the reaction center protein PsbN is not.

Structural genes. Structural modules are well conserved among both groups of phages, as we identified homologs for the majority of typically conserved structural capsid and tail proteins. Despite the structural conservation of core components in all *Tevenvirinae* phages, we were unable to identify some conserved but highly divergent proteins, like the tape measure or tail fiber proteins. The structural region with the most differences compared to the T4 phage was the baseplate. Both groups contain homologs for a large number of the genes involved in the internal structure of the baseplate of T4-type phages [48], which is involved in baseplate assembly, initiation, and sheath contraction [352]. A remarkable difference is the absence of T4 Gp7, which appears to be substituted in both groups of phages by the VrlC protein. VrlC is particularly meaningful, as it is considered an integral component of the two-layered baseplate structure [353,354], so we can predict that both groups possess this type of baseplate. The other regions of the baseplate appear to be less conserved. Within this large structural operon, we also found various unidentified structural proteins that contain domains linked to carbohydrate-binding and host recognition (specifically, YHYH domains, concanavalin A domains, triple collagen repeats, major tropism determinant domains, and YadA domains) [79,355–358]. These putative receptor-binding proteins could be part of the tail fiber complex or the baseplate, as double-layered baseplates have been reported to contain these kind of proteins [353]. Last, the gp5 gene shows a much larger divergence than the VrlC protein, with both groups of phages coding for various gp5 PCs. As gp5 is involved in cell puncturing and local cell wall degradation [359], we can assume that the differences in gp5 PCs are an adaptation to the specific cell wall of the host.

DNA transcription and translation. Transcription regulation in PMPs seems to be quite similar to that of CMPs, with both groups lacking homologs to the T4 genes involved in regulating early and middle transcription (*alt*, *modA*, *modB*, *asi* and *motA*) [360,361]. Some genomes of

group PMP-A code for an homolog of the L12 ribosomal protein, which is the binding site for several factors involved in protein synthesis [362], and a tRNA(Ile)-lysidine synthetase, which is an uncommon nucleoside usually seen only in tRNA and involved in solving differences between the elongation methionine tRNA and isoleucine tRNA [363]. The most significant difference between both groups of phages related to the translation process is that the latter group codes for a homolog of the 30S ribosomal protein S21. This protein is responsible for the recognition of complex mRNA templates during translation and has been described only as an AMG in HTVC008M [364,365]. S21 is not part of any specific gene cluster, which, assuming the protein follows the same rules as the other AMGs, suggests that no other viral factors are required for its functionality.

Auxiliary metabolic genes. CMPs frequently contain AMGs, homologs of host genes, to modify host metabolism during infection [366]. We have analysed the occurrence of this type of genes in the PMP genomes and compared it with the occurrence in CMPs (**Annex 1**:Table S5), which have been widely studied [367]. Both groups of phages had the three classic AMGs involved in nucleotide biosynthesis (cobS, cobT, both subunits of ribonucleotide reductase) [366,368] (**Annex 1**:Table S5). However, Both PMP-A and PMP-B groups code for the adenylate kinase *adk*, which is involved in the interconversion between adenine nucleotides [369], while group C has two different thymidylate synthases and a deoxycytidylate CMP deaminase, which provides the substrate for both [370,371] (**Annex 1**:Table S4). A peptide deformylase involved in protein maturation was present in all PMPs in the core genome, inside a DNA metabolism operon, while in their cyanobacterial counterparts, it was found only in a few and inside the flexible genome, together with the photosystem AMGs [372].

We found fewer genes dedicated to regulation in PMPs than in CMPs. Typical CMP regulation AMGs such as mazG are absent in PMPs, and regulation genes shared by both groups such as the Pho regulon PhoH or Sm/Lsm RNA-binding proteins are more abundant in CMPs than in PMPs (**Annex 1**:Table S5). However, genes related to the *sprT* family (a gene involved in the regulation of the stress factor BolA) are much more prevalent in PMPs than in CMPs. bolA has many effects on cell morphology, cell growth, cell division, and biofilm development in the stationary phase and under starvation conditions [373]. These differences in regulatory proteins are not surprising, since it has been proposed that SAR11 cells are not as tightly regulated as cyanobacteria [244]; hence, their regulatory systems would be significantly different (as mentioned above, the starvation system *mazE*/*mazG* does not exist in SAR11 but it is present in picocyanobacteria) [244]. Regulation in SAR11 seems to be less dependent on proteins, being directed by riboswitches and other small mRNA (smRNA) molecules instead [244]. However, a search of these regulatory mRNAs with the tool Riboswitch Scanner [374] found no evidence of their presence in either group of phages.

Another type of AMG found in PMP genomes are genes related to the production of the O-chain of bacterial lipopolysaccharides, usually found as part of the HRC, but also distributed along the genome in clusters of two or three genes. This category of genes is also found in CMPs but is much less abundant. The LPS-related genes are either enzymes involved in the synthesis of deoxy-sugars to use as building blocks (rfaE, UDP-glucose 6-dehydrogenase) [375,376] or are glycosyltransferases, involved in adding specific sugar residues to a molecule

[377]. Glycosyltransferases in bacteriophages are involved in the glycosylation of viral DNA to protect against the host restriction-modification systems or in the modification of the O-antigen chain of the host to protect against coinfection by other phages [80]. Considering that the glycosyltransferase family most represented in PMPs is GT8, which is mainly involved in LPS biosynthesis [377], and that only one SAR11 genome out of more than 100 sequenced thus far codes for a restriction-modification system [248], it seems likely that glycosyltransferases in this group are involved in the modification of the O-chain of their host.

Curli operon. Between the DNA replication and structural modules, there is a hypervariable region containing a variable number of genes with little synteny among the different PMP representatives (**Annex 1**:Figure 2A and Figure S2A). Within this variable region, we found two homologs of the type VIII secretion system (TSS VIII) present in all PMP groups but the fonsimyophages (**Annex 1**:Figure 2). To our knowledge, this is the first report of phages that code for proteins of this secretion system. The co-occurrence network shows that these proteins are part of a well-defined operon that includes the proteins CsgF, CsgG, two hypothetical proteins and a curli-associated protein. The phylogenetic tree of the PMP and bacterial curli proteins clustered closer to the Alphaproteobacteria representatives (**Annex 1**:Figure S4B).

TSS VIII has not been detected in SAR11, but it has been described in other bacterial groups [378] as the transporter of curli, surface-associated amyloid fibers mainly involved in adhesion to surfaces, biofilm formation, and interaction with host factors and the host immune system [379,380]. The two proteins identified as part of the TSS VIII in PMPs are CsgF, an extracellular chaperone involved in anchoring curli fibers to the outer membrane [381], and CsgG, which form the outer membrane diffusion channel [382]. Both hypothetical proteins in the operon are of the same size, similarly to csgA and csgB genes [383], while the curli-associated protein is of the same size as CsgE, although no similarity could be detected at the sequence level or predicted structural level. Several experiments have shown that the only proteins required for curli phenotype expression are CsgA, CsgB, CsgF, and CsgG (CsgE increases almost 20-fold the amount of curli released, but it is not essential) [381,384]. Therefore, CsgA and CsgB are the only proteins missing in PMPs for the infected cells to express a curli phenotype.

4.2 Results derived from the work "Long-read metagenomic improves the recovery of viral diversity from complex natural marine samples"

4.2.1 Summary

In this study, we analysed a single marine sample from offshore Mediterranean waters that was sequenced with Illumina and PacBio Sequel II, then assembled twice: first using only the Illumina short reads, resulting in the Short Read Assembly dataset (SRa); then a hybrid assembly using both the Illumina short reads and the PacBio long reads, resulting in the Long-Read Assembly dataset (LRa). We decided on the hybrid assembly over a long-read only assembly based on previous results (12). In order to evaluate the possible biases introduced by the assembly process, we also analysed the PacBio CCS15 reads (PacBio consensus reads created by comparing at least 15 subreads, LR) before assembly.

Although the sample is derived from the cellular fraction, the presence of replicating viruses inside cells during the lytic cycle produces a natural amplification that makes it possible to find abundant sequences of viral origin. We have found a major wedge of the expected marine viral diversity directly recovered by the raw PacBio circular consensus sequencing (CCS) reads. More than 30,000 sequences were detected only in this dataset with no homologous in the long- and short-read assembly and ca. 26,000 had no homologues to the large dataset of the Global Ocean Virome 2 (GOV2), highlighting the information gap created by the assembly bias. No novel major clades of viruses were found, but there was an increase of the intra-clade genomic diversity recovered by long-reads that produced an enriched assessment of the real diversity and allowed the discovery of novel genes with biotechnological potential (e. g. endolysins).

4.2.2 Viral sequence recovery and statistics

First of all, we wanted to compare the efficiency of viral sequence recovery between the three datasets (**Annex 2**:Table 1). The first step in the pre-processing pipeline was to run VIBRANT [291] for all those sequences >1kb to determine those in each dataset that were of viral origin. Viral sequences turned out to be quite numerous in both datasets, with 5% of the total sequences from the SRa and LRa and 2.5% of the LR dataset classified as viral contigs. After a step of clustering at 95% sequence identity to remove redundant reads from the LR dataset, we recovered a total of 54,082 putative viral sequences (10,979 in SRa, 947 in LRa and 42,156 in LR) (**Annex 2**:Table 1). In order to assess if the different assembly methods recovered the same viral community, we identified unique sequences in each dataset by comparing the three datasets against each other (see Material and Methods). Most sequences from the LRa were also found in the SRa, with only 36 unique LRa contigs. Remarkably, while the SRa dataset contained a fair number of unique sequences (5,886), most of the unique sequences were found in the LR dataset (30,203; 71% of total viral LR sequences), revealing a large genomic diversity not recovered by the assemblies. This diversity gap was also present when comparing a marker gene such as the terminase large subunit (*ter*L), with the LR dataset containing 393 unique terminase genes (clustering at 95% amino acid identity), compared to

30 and 2 in the SR and LRa datasets respectively. The GC content showed a light (effect size = 0.022) but significant (Kruskal-Wallis test, p-value < 10-15) skew towards high GC values when PacBio CCS reads were added to the datasets (**Annex 2**:Table 1). The SRa dataset presented an average GC content of 35.45% compared to 36.9% for the LRa and 38.13% for the LR (**Annex 2**:Table 1). This bias could arise from the fact that assemblies usually only recover the core genome. In this sample (marine surface water), SAR11 clade is the most abundant organism [204], with an average GC content of 34%. LRs recover more of the flexible genome, which can present GC fluctuations compared to the core and would thus explain this variation from 34 to 38%. Regarding sequences shared between the three datasets, **Annex 2**:Table S1 shows the relationship between contigs that were considered part of the same phage (identity over 95%, 70% overlap of the smallest contig). When comparing the ratio of recovered sequences between SRa and the combined LRa and LR datasets for shared ones, we found that in 2,463 out of 3,316 shared instances (*ca*. 75%), the LR datasets contained longer contigs than their SRa counterpart (**Annex 2**:Table S1). These results show that the use of long reads in assembly result in larger contigs compared to assembly with only SR.

Next, we were interested in assessing if this novel diversity had been captured by previous studies, so we compared the three datasets against the Global Ocean Virome 2 (GOV2) [184], the largest database of seawater phages to date (195,728 marine populations, containing 6,685,706 proteins). This dataset was created from viromes obtained from 145 samples from the Malaspina [385], *Tara* Oceans [183] and *Tara* Arctic [184] expeditions, therefore representing marine phage communities from different environments from all around the world. We found 30,997 viral sequences in our whole dataset (SRa, LR and LRa) not found in GOV2, with the vast majority (26,766) of these unique sequences belonging to the LR dataset.

Regarding size and completeness, the hybrid PacBio assembly LRa resulted in the largest viral contigs, with a maximum size of 428,169 bp and an average contig size of 32,260 bp (**Annex 2**:Table 1). We recovered 24 complete phage genomes (based on circular redundancy at the ends) from both assembled datasets (15 in LRa, 9 in SRa). As expected, due to their small estimated average size (*ca*. 5Kb), we were unable to recover any complete genomes directly from the LRs. However, we can make an estimated guess of the quality of the remaining contigs using VIBRANT's quality statistics, which classify contigs based on the estimated completeness of the genome. If we consider only contigs marked as high quality (70% of the estimated phage genome), we found that only 53 (0.4%) of the SRa contigs belonged to this category, while in the LRa dataset there were 114 (12.5%) (**Annex 2**:Table 1). Some complete phage genomes were shared by the LRa and SRa datasets. The SRa contigs resulted in a maximum contig size of approximately half of that found in LRa (188,349 bp), with an average contig size on par with the LR dataset, more than six times smaller than the average in LRa (*ca*. 32Kb) (**Annex 2**:Table 1). These results, together with the fact that the average protein size in all three datasets is similar and the number of proteins recovered from the LR dataset are an order of magnitude larger than in the assembled datasets, suggests that PacBio CCS15 reads could be used for viral protein calling without the need for assembly, as previously stated [208].

4.2.3 Putative host prediction

An important part of the biological significance of viruses depends on knowledge of the host they infect. We attempted to assign a host to contigs in all three datasets (SRa/ LRa/ LR). To this end, phage contigs were classified against the RefSeq database. We assigned hosts to each sequence following the method described in Beaulaurier *et al.* [208], which was applied to phages obtained by Nanopore sequencing. The method is based on protein homology against a reference database, assigning a host to a sequence based on the number of best hits (See Materials and methods). **Annex 2**:Figure 1A shows the results at a 3-protein threshold, including both all contigs from a dataset and those unique to their specific dataset (de-replicated) and before de-replication. Considering the SRa and LR datasets, both unique and de-replicated variants presented a similar host assignment rate (*ca*. 30%) with no differences at the taxonomic level suggesting that the differences between the two datasets could be beyond the order level. As might be expected, Alphaproteobacteria and Cyanobacteria were the most abundant hosts in all three datasets, as they were the most abundant groups in the sample [204] and were also the most represented in the reference databases (Fig. 1A). The recent addition of various *Methylophilales [274]* and *Flavobacteria [386]* phage genomes to the reference databases has resulted in a highly increased Gammaproteobacteria and Flavobacteria phage count compared to previous analysis of the Mediterranean virome [105]. The LRa dataset provided the highest rate of host assignment. In the non-de-replicated sample, almost 75% of the contigs had a host assigned, compared to a 30% rate for the LR and SRa datasets. This is probably due to the fact that they were, on average, larger contigs, and as such contain more information to reliably assign a host. However, we were unable to assign host to any of the 36 unique sequences in the LRa dataset (3.2% of the total). Host taxonomy was similar to that seen in the previous datasets, the main difference being an increase in eukaryotic and archaeal viruses (20% of total contigs), mainly Marine Group I Thaumarchaeota (*Marthavirus*) [387].

Comparison between the sequences obtained by assembly (LRa and SRa) also revealed differences between the viral groups. As a general rule, LRa contigs were on average larger than their SRa counterparts, even if the latter can result in similar maximum size. For example, in Alphaproteobacterial phages (**Annex 2**:Figure 1B), we recovered 52 sequences over 30Kb in the SRa dataset, compared to 126 in the LRa dataset. We found a similar case for the cyanophages (**Annex 2**:Figure 1C), where 14 sequences were over 50kb in the SRa dataset compared to 68 sequences in the LRa dataset. Special attention deserves the Nucleocytoplasmic Large DNA Viruses, (NCLDV, proposed order *Megavirales*) (**Annex 2**:Figure 1D), whose assemblies in the LRa dataset were larger and more numerous (24 sequences over 20Kb, including the largest contig of 428Kb) than those in the SRa dataset (21 sequences, 2 over 50Kb, max size 61Kb).

To analyse the phylogenomic diversity of the NCLDV sequences found, we used only sequences that contained five key markers highly conserved in this type of virus: the major capsid protein mcp, DNA Polymerase beta subunit PolB, DEAD/SNF2-like helicase SFII, Poxvirus Late Transcription Factor VLTF3 and the Packaging ATPase A32 [316]. **Annex 2**:Figure S1 shows a phylogenetic tree based on a concatenate of these five proteins, including reference genomes from RefSeq and the collection of 444 marine NCDLV MAGs from

Moniruzzaman *et al* [316]. The tree shows these new eukaryotic sequences fell into the *Mimiviridae* family (16 sequences) and the *Phycodnaviridae* family (8 sequences).

4.2.4 Relative abundance in marine samples

Next, we wanted to analyse whether all the diversity found only in the LR dataset was abundant and representative in nature. For that reason, we performed a recruitment analysis of SRa/LRa/de-replicated LR viral sequences against the entire *Tara* Oceans metagenome dataset [286]. We considered the presence of a sequence in a metagenomic sample if they recruited at least five reads per kilobase of genome and gigabase of metagenome (RPKG), with an identity of 95% and a contig coverage of 50%. The results are shown in **Annex 2**:Figure 2. Although pelagiphages and cyanophages (viruses that infect *Ca. Pelagibacter* and Cyanobacteria, respectively) show a similar abundance, they present different patterns of recruitment. The most cosmopolitan phages are cyanophages, particularly those who infect the genus *Prochlorococcus*. On the other hand, pelagiphages show a more endemic distribution, specially pelagimyophages, which tend to recruit only in a few stations at a time (in this case, as could be expected, in the *Tara* stations in the Mediterranean), while pelagipodophages tend to appear in more stations (**Annex 2**:Figure 2). In each of the plots, the recruitment means for each dataset were represented as a line, showing that in all three cases (*Alphaproteobacteria*, *Cyanobacteria* and Other phages) the sequences recovered by LR prior to assembly are significantly more abundant than their assembled counterparts (Wilcoxon rank sum test, p-value < 10^{-5}). Furthermore, this difference in RPKG was accentuated when comparing phages that infect taxa which are typically difficult to assemble, such as those infecting *Alphaproteobacteria* [248]. These results suggest that the de-replicated (non-redundant) LR sequences represent an untapped and abundant reservoir of genomic diversity.

Since the phage sequences were obtained from the cell fraction, we were interested to know if they were abundant and could also be recovered in the viral fraction. To that end, we recruited all phage datasets in metagenomes and viromes at different depths obtained at the same location from which the sample was collected [105,342]. When comparing the recruitment values in both types of samples (**Annex 2**:Figure 2B), we observed that the vast majority of sequences recruited significantly more in the viral fraction at the three depths surveyed (Wilcoxon rank sum test, p-value <10^{-16}, for all three depths). Therefore, we can confirm that the phage genomes recovered from the cellular fraction are representative of the community found in the virion fraction as well, and as such represent a valid method to recover the viral diversity of a sample.

4.2.5 New diversity recovered from LR

Once we discovered that there is a larger amount of viral sequences in LR not contained in the other datasets, probably lost in the assembly process, and that is abundant in nature, we decided to analyse this diversity. Given that there is no universal marker for analysing viral diversity, we use a number of different phage-specific markers (Large terminase subunit *terL*,

replicative DNA helicase *dnaB*, tail tube protein, major capsid protein, spanin) as well as several well-characterised auxiliary metabolic genes (AMGs) (thymidilate synthase *thyX*, Phosphoheptose isomerase *gmhA*, Ribonucleoside-diphosphate reductase *nrdA*, Ribonucleotide Reductase large subunit, Phosphate starvation-inducible protein *phoH*).

We analysed the diversity of these markers in the same sample for the three datasets by building phylogenetic trees (**Annex 2**:Figure 3A, S2, S3) and also by comparing with GOV2 the dereplicated sequence distribution (**Annex 2**:Figure 3B, **Annex 2**:Table S2). The phylogenetic trees showed that none of the clades were composed only of LR-unique proteins, so we can conclude that the unique sequences recovered from the LR dataset do not belong to novel phage taxa, but to known clades. Comparing the distribution of unique proteins between our three datasets, the LR dataset usually contained more unique sequences by an order of magnitude compared to the assembled datasets (**Annex 2**:Table S2). Moreover, the percentage of unique variants was always higher in the LR.

After including the GOV2 dataset in the comparison, it quickly becomes apparent that this dataset contained most of the unique sequences (*ca*. 90% of all unique proteins). This was expected considering the vast size and breadth of sampling of the GOV dataset (144 samples), it was therefore surprising that a dataset derived from a single sample contains a tenth of the diversity, especially considering that the ten proteins selected are conserved proteins in phage genomes. Out of this slice of diversity, the vast majority of the unique contigs derive from the unassembled LR dataset, as seen in the case of DnaB (149 different proteins versus 26 in the assembled datasets) and the RrdA (150 versus 19 in assembled datasets) (**Annex 2**:Table S2).

It is important to emphasise that the fact that LR do not reveal novel phage clades does not mean that their novelty is not relevant. An example of this would be the endolysins, a remarkably diverse group of catalytic enzymes that degrade the cell wall of the host so the phage progeny can escape [388]. In recent years, these proteins have awakened increased interest for their potential to be used as antimicrobial agents [389,390]. Culture-free approaches have been applied to great effect in order to broaden the diversity of endolysins. In a previous study [322], 2,628 putative endolysins were retrieved from a collection of 183,298 assembled viral genomes, pooled from a variety of metagenomic datasets. We applied the same pipeline to our samples to evaluate if this novel diversity found by LR would also apply to proteins with more diversity than the usual protein markers.

We recovered 335, 106 and 841 putative endolysins from the SRa, LRa and LR datasets respectively, yielding a total of 1,216 new sequences. A phylogenetic tree of the sequences (**Annex 2**:Figure S4) reveals that although most of the sequences are distributed along previously described endolysin groups, there were four clades not found in the previous endolysin environmental collection, which we will name C1 to C4. An analysis of their domains revealed them to be glycoside hydrolases from families 24, 104, 23 and 24, respectively. These are lytic transglycosylases that have the well-known α+β lysozyme [391] fold, with differences in activity and specificity supposed to be determined by the environment surrounding the active site. Each family includes several well-characterised phage lysozymes. No domains related to cell wall binding were found. Interestingly, the C4 clade contains a signal-arrest-

release motif, a mechanism not reported in the original dataset [322]. This motif first directs the endolysin to the periplasm by first attaching it to the membrane, where it remains inactive until it is released as a soluble active enzyme in the periplasm [392]. No other domains related to protein export or cell wall binding were found.

4.2.6 Functional characterization

Finally, our last question was if there was any functional category more enriched in the LR dataset compared to the assemblies. To this end, we analysed the protein content at the level of functionality, annotating the proteins against the KEGG [86] and Conserved Domain Database (CDD) [304]. Then we compared the number of proteins with each annotation in the LR dataset against the proteins found in the assembled datasets. The LR dataset was particularly enriched in repeat-containing proteins, such as MORN repeats (37 times higher in LR than in the assembled datasets), pentapeptide repeats (26 times higher), Ankyrin repeats (10 times higher), and Kelch repeats (9 times higher). Pentapeptide and Kelch repeats are widespread through bacterial and viral proteins [393,394], ankyrin repeats have been found in a novel AMG which protects the infected bacteria from eukaryotes [395], and MORN repeats have been found in bacteriophage endolysins [396]. The appearance of these proteins was not surprising, as repeats are the main cause for fragmented assemblies [397]. A similar argument could be made for the prevalence of integrases (18 times higher), reverse transcriptases (not found in the assembled datasets) and transposases (9 times higher). Although these proteins are widespread in phage genomes [398–400], they present a large amount of microdiversity, which is also difficult for assemblers to solve [204]. No groups of proteins were noticeably less abundant in LR compared to its assembled counterparts. These results suggest that long reads can help recover parts of the viral genome difficult to retrieve due to assembly bias.

4.3 Results derived from the work "RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content"

4.3.1 Summary

In this study, we apply a machine learning algorithm (Random Forest) to the assignment of hosts to complete or fragmented genomes of viruses of Bacteria and Archaea. Previous studies have shown that random forest algorithms are suitable for classifying viruses according to their hosts and that protein domains can be used to achieve accurate host predictions. Based on these findings, we postulate that random forest classifiers could be applied to protein content to build a classifier based on identifying combinations of genes that are indicative of virus-host associations.

Through this approach, we were able to design RaFAH (Random Forest Assignment of Hosts), a classifier that displays a comparable performance with that of other state-of-the-art methods in a wide variety of metagenomic datasets, including various viromes of medical, biotechnological and environmental relevance. Importance analysis reveals that the best predictors for host assignment are proteins related to virus-host interaction (lysins, tail fiber proteins, Rz-like proteins). Furthermore, our analyses led to the identification of 537 sequences of archaeal viruses representing unknown lineages, whose genomes encode novel auxiliary metabolic genes.

4.3.2 Performance of RaFAH against other host prediction software

We tested the performance of RaFAH and other host-prediction approaches on an independent dataset of isolated viral genomes that did not overlap with those used for training the models (Test Set 1, composed of RefSeq viral genomes with less than 70% average amino acid identity when compared with those in Training Set 3, see **3.10 Host Assignment**). When using RaFAH and the other tested methods without score or prediction probability cutoff (considering as valid all host predictions with no thresholds for their probability value or bit score), RaFAH outperformed alignment-independent, hybrid, and alignment-dependent approaches for host prediction at every taxonomic level based on the F1 score (**Annex 3**:Figure 2A). This difference in performance became gradually more evident from domain to genus level. Next, we evaluated how the performance of these tools responded to thresholding (i.e., applying a cutoff on their probability value or bit score) and only considering predictions that were above the cutoffs. These analyses revealed that homology matches, CRISPR and tRNA (the classical approaches) displayed the lowest recall (**Annex 3**:Figure 2B) but the highest precision (**Annex 3**:Figure 2C). HostPhinder and CRISPR displayed high precision only at the strictest score cutoffs. As a consequence, these two methods displayed very low recall when the highest cutoffs for predictions were established. RaFAH, WIsH, and VirHostMatcher-Net displayed higher recall than the other approaches, especially at the range of more permissive score cutoffs (0). Yet this higher recall came at the expense of lower precision for WIsH and VirHostMatcher-Net. Meanwhile the precision of RaFAH outperformed these tools even when no cutoffs were applied. Together, precision, recall, and

F1 score suggest that RaFAH can predict more virus-host interactions than the other tested approaches while maintaining high precision, particularly for divergent viral genomes that escape detection by the classical approaches (**Annex 3**:Figure S1).

We evaluated how the similarity among the genomes in Test Set 1 with those used to train the model (Training Set 3) affected the performance of RaFAH. For this purpose, we assessed how the precision of RaFAH changed by setting a threshold on the maximum allowed average amino acid identity (AAI) between the genomes on Test Set 1 and those on Training Set 3. As expected, a positive association was observed between these variables (**Annex 3**:Figure S2), meaning that the more similar the testing genomes are to the ones used for training, the more likely RaFAH is to correctly predict their hosts at all taxonomic levels. Based on this analysis, 75% of the class-level host predictions will be correct (precision: ~0.75) for viruses that possess <60% AAI to the ones in the database, when no cutoffs on prediction scores are applied.

Host-prediction tools were further validated on a dataset of viral genomic sequences derived from marine SAGs, Test Set 2 [196]. These sequences represent an ideal test dataset because they are uncultured viruses, not represented in the NCBI database used for training, and can confidently be assigned hosts because these viruses were inside or attached to the host cells during sample processing. Based on the F1 score, HostPhinder displayed the best performance at the levels of domain and class, followed by RaFAH slightly behind (**Annex 3**:Figure S3A). Yet at the level of phylum WIsH displayed the best performance, again followed closely by RaFAH. At the levels of order, family, and genus, WIsH displayed the highest F1 scores followed by the combined classical approaches. The recall (**Annex 3**:Figure S3B) and precision (**Annex 3**:Figure S3C) of RaFAH on Test Set 2 was lower than that obtained for Test Set 1. Nevertheless, a negative association between precision and recall as a function of the score cutoff was also observed for RaFAH and the other tested tools on Test Set 2 (**Annex 3**:Figure S3D). Taken together, these results are evidence that RaFAH also performed well when predicting hosts of uncultured viruses from the marine ecosystem.

4.3.3 Performance of RaFAH based on environment

To test the performance of RaFAH on samples from other habitats, we applied it to predict hosts of a dataset of viral genomes obtained from metagenomes of eight different ecosystems (Test Set 3, see 3.12). For comparison, we also applied the other tested methods of host prediction (HostPhinder did not scale to the more than 60,000 genomes in this dataset, and analyses did not complete after running for several days). According to the F1 score, RaFAH outperformed WIsH and VirHostMatcher-Net for this dataset as well (**Annex 3**:Figure S4A), due to slightly higher recall (**Annex 3**:Figure S4B) and precision (**Annex 3**:Figure S4C). RaFAH was also superior when the strictest cutoffs were applied, whereby both precision and recall were markedly superior to VirHostMatcher-Net (**Annex 3**:Figure S4D). On this dataset, RaFAH achieved 43.13% precision at the level of genus when no score threshold was applied. Bootstrap analysis revealed that this level of precision was consistent across 1,000 replicates (mean 43.02% ± 2.1%). This result indicates that the precision of RaFAH on Test Set 3 was not biased by uneven viral genome diversity among the samples that made up

this dataset. When using classical approaches for host prediction, the majority of viruses remained unassigned regardless of ecosystem, and the best performance of these approaches was among the human gut dataset, in which only about 25% of sequences (lengthwise) could be assigned to a host at the level of phylum (**Annex 3**:Figure 3). Meanwhile, when set to the 0.14 cutoff, which yielded 92% phylum level precision on Test Set 1 (**Annex 3**:Figure S1) and 90% on Test Set 3 (**Annex 3**:Figure S4D), RaFAH was capable of predicting putative hosts to the majority of viral sequences across all ecosystems except for the permafrost dataset, likely because viruses derived from this ecosystem are poorly represented in reference databases.

Interestingly, the host predictions yielded by RaFAH were markedly different across ecosystems. Viruses of Proteobacteria were the dominant group in all ecosystems except the human gut. As expected, the most abundant targeted hosts of the viruses from each ecosystem were the most abundant taxa that reside in those habitats. Viruses of Cyanobacteria were the second most abundant group among the marine dataset, a position that was occupied by viruses of Actinobacteria and Bacteroidetes among the freshwater dataset. Viruses of Firmicutes and Bacteroidetes were the dominant group among the dataset of human gut viruses while viruses of Firmicutes, Bacteroidetes, and Actinobacteria were among the most abundant among the soil and permafrost datasets. Viruses of Euryarchaeota were the second most abundant group among the hypersaline dataset, a position that was occupied by viruses of Crenarcheaota in the thermal springs dataset. These results are in accordance with the known prokaryote diversity that dwells in each of these ecosystems [94,321,335,338,343,401,402].

4.3.4 Effect of genome completeness on host prediction

We assessed how genome completeness affected the performance of RaFAH. For this purpose, we used Test Set 3 as it displayed the necessary range of genome completeness values necessary for this purpose, while Test Set 1 was mostly made up of complete genomes and Test Set 2 was mostly made up of low-completeness genomes. We assumed that the predictions yielded by the combined classical approaches represented the true hosts of Test Set 3, although this assumption is likely to lead to an underestimation of the true precision of RaFAH. We found weak positive associations (Pearson $R^2 > 0.6$, $p < 10$ 13 for all taxonomic levels) between the precision of RaFAH and genome completeness at all taxonomic levels (**Annex 3**:Figure S5A). These curves tended to reach a plateau around ~25%–50% genome completeness and increased further for the lower taxonomic ranks (genus, family, and order) for genomes that were >85% complete. Coupled with the observations of the performance of RaFAH on Test Set 2, we suggest that RaFAH is better suited for viral genomes with 50% or more completeness.

We used Test Set 3 to analyse the relationship between genome completeness, sequence length, and RaFAH prediction score across the eight different ecosystems (**Annex 3**:Figure S5B). This revealed a positive correlation between those variables (Pearson $R^2 = 0.65$, $p < 2.2e^{-16}$ for the combined set of all ecosystems). Likewise, significant albeit weaker positive correlations were also detected between prediction score and sequence length (Pearson $R^2 =$

0.14, $p < 2.2e^{-16}$), and prediction score and genome completeness (Pearson $R^2 = 0.11$, $p < 2.2e^{-16}$). We found that regardless of taxonomic level, precision did not consistently increase through thresholding for genome length, providing further evidence that shorter sequences do not necessarily yield worse predictions (and vice versa) (**Annex 3**:Figure S5C). These results suggest that the precision of RaFAH cannot be explained by genome length/completeness alone, likely because RaFAH was trained on a dataset with a majority of genome fragments.

4.3.5 Diversity and AMGs of Archaea viruses

Based on the finding that RaFaH achieved nearly perfect precision for domain-level host predictions, and the fact that viruses of Archaea are under-represented in databases, we subsequently focused on the description of these viruses. Few large-scale studies have addressed the diversity of uncultured viruses of Archaea, and they focused mostly on marine samples. Here, we describe viruses from seven other ecosystems: soil, permafrost, freshwater, sludge, hypersaline lakes, thermal springs, and the human gut. Applying RaFAH to only eight metagenomic datasets led to the prediction that 537 genomic sequences represent viruses of Archaea (prediction score R=0.14). To put this figure in context, there are only 96 genomes of viruses of Archaea deposited in the NCBI RefSeq database. We took several steps to ensure that these genomes were truly derived from viruses of Archaea and consistently found compelling evidence to support our claim. First, these genomes could be linked to archaeal genomes either through homology matches or alignment-independent approaches, which provided further evidence that 423 out of the 537 genomes (79%) were indeed derived from archaeal viruses (**Annex 3**:Table S2). Second, much like the RefSeq genomes of archaeal viruses, these sequences were enriched in Pfam domains annotated as exclusive of Archaea, eukaryotes, and their viruses (**Annex 3**:Figure S6). Third, these genomes were enriched in ribosomal binding site motifs that are also enriched among RefSeq viruses of Archaea (**Annex 3**:Figure S7).

Next, we manually inspected the gene content of the viruses predicted to infect Archaea in search of novel auxiliary metabolic genes (AMGs) and new mechanisms of interaction with the host molecular machinery. The small number of reference genomes of Archaea and their viruses makes it difficult to describe the gene content of the archaeal viruses that we discovered because most of their genes have no taxonomic or functional annotation. However, we found several sequences containing genes coding for thermosomes, group II chaperonins involved in the correct folding of proteins, homologous to their bacterial counterparts, GroEL/GroES [403]. Other AMGs found among archaeal viruses were those involved in the synthesis of cobalamin cobS, recently associated with Marine Group I (MGI) Thaumarchaeota virus infection [387] as well as genes that encoded 7-cyano-7-deazaguanine synthase QueC involved in archaeosine tRNA modification [404]. One of the AMGs most prevalent among archaeal viral genomes coded for a molybdopterin biosynthesis MoeB protein (ThiF family). This family of proteins is involved in the first of the three steps that make up the ubiquitination process [405]. This system regulates several cellular processes through post-translational modification of proteins such as their function, location, and degradation, making it an ideal target from the point of view of viruses to facilitate their replication [406].

5. Discussion

5.1 The role of the cellular fraction in phage metagenomics

It has been demonstrated that despite the pore size, the cellular fraction (> 0.2 µm) contains a large amount of phage DNA (*ca.* 15%) [145,407], which gets roughly translated into 5% of assembled contigs in both PacBio and Illumina datasets. It should not be a surprise that the origin of this phage DNA differs between the cellular and viral fractions. Phage DNA in the viral fraction originates from the virions present in the water, while in the cellular fraction the recovered phage DNA likely belongs to cells undergoing the lytic cycle, which produces a natural amplification of the sequences. However, other sources are also a possibility, including lysogenic viruses (either integrated or as a plasmid) or phages encapsulated in virions larger than the filter pore (> 0.2 µm) or attached to the cell.

These latter possibilities open the door to the proposition that cellular metagenomes could contain complementary information missing in viromes. A number of phage genomes could then be better represented in the cellular fraction compared to the viral fraction, revealing a new slice of viral diversity. Our results in this area are inconclusive. Our comparison of recruitment values in metagenomes and viromes of the PacBio metagenomic viruses revealed that the vast majority recruit more in the viral fraction. However, we noticed a bias with regards to podovirus and myoviruses that infect SAR11 and Cyanobacteria, two of the most abundant bacteria in marine environments: CMPs and PMPs are more abundant in the cellular fraction, while podophages that infect the same hosts display the inverse recruitment pattern. As viral fraction DNA is derived from virions present in the sample, it is reasonable to expect that phages with features that allow for their virions to prevail in the environment regardless of host abundance (larger burst size, more stable virions, strictly lytic cycle) will be overrepresented on viral fraction datasets. The study of phage content in the cellular fraction could then help expand our knowledge of phage diversity.

There is also the matter of DNA extraction amount and its relation to sequencing, especially with regards to third-generation platforms. DNA extraction from the viral fraction is an arduous process, requiring a large amount of sample (around 200 seawater litres for a single Illumina sequencing run) and specialised equipment. The use of DNA amplification techniques such as MDA is not a solution as it introduces a noticeable bias on the phage community, as it has been reported in this work and many others [160–162]. With PacBio DNA requirements being at least an order of magnitude larger than that required for Illumina sequencing [204], the study of viruses within the cellular fraction might be a good alternative.

5.2 Metagenome mining and PMP recovery

The bioinformatic approach developed in this work to recover PMP contigs can be applied to other microbes difficult to cultivate but with some isolates already sequenced. Likewise, cross-assembly (joining contigs obtained from metagenomes or metaviromes) can help reconstruct more complete genomes. In fact, the crAssphage-like family of phages was discovered by combining contigs from a plethora of metagenomic gut datasets [408]. Results might differ depending on how abundant the host is in the environment. In the case of SAR11, its prevalence in surface waters of the ocean and other aquatic habitats played in our favour,

and we have been able to uncover a remarkable diversity of viral entities different from the cultured reference. Surprisingly, the vast majority of PMPs were recovered from cellular fraction (>0.2 μm) datasets, which points to a bias against these genomes in the viral fraction datasets. Similar methods could be applied to other relevant, hard to grow microbes with recently isolated phages such as the marine methylotrophs clade OM43 of the Gammaproteobacteria. It could also be applied to samples from other environments to recover similar phages that might have undergone an environmental transition. Such is the case of the two recovered PMPs that putatively infect *Candidatus Fonsibacter*, the SAR11 freshwater clade. The study of phylogenetically related phages across multiple environments can provide an insight into the adaptations needed for successful infection under different conditions.

The pelagibacter phages recovered in this thesis have features that contrast with their host. While the SAR11 clade cells are characterised by their small size and the marked streamlining of their genomes, myoviruses are large phages with big and complex genomes. In fact, the ones described here are even more complex than the classic *E. coli* phage T4, with a larger repertoire of genes in their flexible genome and novel sets of AMGs. A clear example among these is the large hypervariable island found near the tail genes, which we have named Host Recognition Cluster (HRC) due to the fact that it contains genes annotated as tail fiber proteins or genes containing carbohydrate-binding domains. This region is characterised by its under-recruitment in metaviromic datasets, similar to the metagenomic islands described in prokaryotic genomes and the metaviromic islands first described by Mizuno *et al. [79]*. Contrary to other T4-like phages [312], the HRC region in PMPs remarkably includes glycosyltransferases, which in phages are usually involved in protection of viral DNA against restriction enzymes or in host serotype conversion [80]. Considering that the SAR11 clade genomes are not known to code for restriction-modification systems [244], it is more likely that these genes are involved in the latter function, using a similar mechanism to those already described for other marine and non-marine podoviruses [409–411]. In this, the glycosyltransferases coded by the phage modify the polysaccharides forming the O-chain of the host cell wall. As these structures are usually the target of phage recognition proteins, a change in the host envelope will cause other competing phages to miss the already infected host. That these large phages of SAR11 require protection against superinfection events is not surprising, given the potentially sharp competition with, for example, SAR11 podophages that have much larger burst sizes (42 ± 7 versus 9 ± 2 for the cultured representatives) [134,271].

In recent years, the team led by Ben Temperton has made significant progress in the study of T4-like marine phages. First, they published the draft genome of phage Mosig, the second SAR11 myophage isolated from culture [274]. Their second, and perhaps more interesting discovery is the isolation of a new group of T4-like marine phages, named Melnitz [138]. This group of phages infects the OM43 clade, a streamlined type I methylotroph of the class *Gammaproteobacteria*. Like the SAR11 clade, it is abundant in coastal ecosystems, with abundance peaks coinciding with phytoplankton blooms [138]. Melnitz phages present a remarkable similarity to PMPs, both in gene content and synteny, with phylogenetic analysis suggesting they are both part of the same clade [138]. Furthermore, the tRNA and two-piece tmRNA *ssrA* genes coded by Melnitz are more similar to those found in SAR11 cells than to

their host. Considering that host ranges appear to not overlap between Melnitz and Mosig phages, it is suggested that Melnitz recently underwent a host transition from SAR11 to OM43 [138].

The analysis of the genomic content of these phages also reveals other adaptations to their streamlined hosts. An example would be the presence of a glutamine riboswitch controlling the aforementioned *ssrA* gene. Riboswitches are a common regulatory mechanism in streamlined marine bacteria due to their low metabolic maintenance cost compared to protein-encoded operators and repressors [244], and would therefore be expected to find them in phage genomes, as they are reliant on controlling the host's metabolic machinery to complete their life cycle. We did not find any instances of riboswitches in our collection of PMPs, probably due to the difficulty to detect them in genomes. We expect that further research on this field will help elucidate these and other methods phages use to control streamlined hosts.

5.3 The phage-encoded curli operon (Type VIII secretion system)

Perhaps the best example of new AMGs found in PMPs is the phage TSS VIII operon, which was first described in this work. In bacteria, this operon is involved in the production and secretion of the functional amyloid curli, an integral part of biofilm [379,380]. Compared to the *E. coli* TSS VIII operon, the phage version only includes two out of the four proteins involved in curli secretion (CsgG and CsgF). In this way, the phage TSS VIII operon resembles those found in Alphaproteobacteria, which usually code for curli subunits CsgF, CsgG and CsgH, a CsgC-like protein implicated in amyloid inhibition [378,412]. No homolog of CsgH/CsgC has been found in phage operons to this date, but it is not needed for the curli phenotype in *E. coli [413,414]*.

The origin of the operon in viruses is still unclear, but the recent report of the phage TSS VIII operon also being found in OM43-infecting myophages can provide a few clues as to its origin. The similarity between the phage CsgF and CsgG proteins and their bacterial homologues suggests it is the product of a lateral transfer event, but the curli operon has not been described in either of the hosts [244,258]. Considering both phages coding for the operon infect hosts in different classes and the comparable synteny between phage and Alphaproteobacteria TSS VIII operons, we suggest that the lateral transfer event of the TSS VIII operon occurred in an Alphaproteobacteria-infecting phage ancestor, followed by viral speciation.

As stated previously, phage operons only code for CsgG and CsgF, which code for a pore-forming complex [382] and an extracellular chaperone [381], respectively. CsgG and CsgF form an 18-mer heterodimer with 1:1 stoichiometry, in which the CsgG pore spans the entire outer membrane and CsgF forms a secondary channel outside the cell that acts as a chaperone in curli nucleation [384]. In *E. coli*, the transport system includes a periplasmic accessory protein (CsgE), which increases translocation efficiency of curli but it is not required for the curli phenotype [381,384]. The CsgG pore is too narrow to allow for virion exit for the cell (the CsgG pore has 40-Å inner diameter, while the HTVC008M capsid diameter is 550 Å)

[134,384], which rules out the possibility of the operon acting as a virion release mechanism. There are reports of amyloid proteins in eukaryotic viruses, where they play the role of inhibiting programmed cell death of their eukaryotic host by sequestering effector proteins [415], which both requires the still missing amyloid-forming proteins and does not require the presence of the curli transporter.

Recently, an extensive structural analysis of the operon was performed by Buchholtz *et al,* in which they analyse the 3D structure of the CsgFG complex via *de novo* modelling using AlphaFold2 [416], which sheds more light in the possible function of the operon. 3D structure comparison between phage CsgG / CsgF reveal low structural similarity to their bacterial homologues. Surprisingly, while phage CsgG maintains a similar structure seen in its bacterial counterpart, phage-encoded CsgF presents little structural similarity, including the appearance of an extra α-helix and ending in a β-sheet [138]. Based on these results, Buchholz *et al*. conclude the proteins may no longer form a heterodimer, with CsgG retaining its function as a pore and CsgF evolving independently to provide an alternative function [138]. Another clue towards identifying the operon's function can be found analysing the genomic location of the operon. Genes *csgGF* are located immediately downstream of thymidylate synthase *thyX* and *ssrA* and transcription coactivator genes, which are involved in lysis regulation in other phages. Therefore, the curli operon could play a putative role in regulation of the timing of cell lysis. The unusual lysis curves in SAR11 myophages support this hypothesis [274,277].

Even with the information available right now, the function of this operon in bacteriophages is also a mystery. Based on their results, Buchholz *et al.* propose two putative functions for the operon. First, they argue CsgG is functionally analogous as pinholins, proteins used by phage λ to regulate cell lysis [417]. However, the pinholin channel only spans the inner membrane, as its function is to activate membrane-bound lysins and allow them to access the peptidoglycan layer by membrane depolarization [418]. In contrast, the CsgG pore is situated in the outer membrane, as evidenced by the presence of a Sec/SPII signal peptide in phage CsgG (SignalP 6.0, probability 0.9966). Pinholins are also a more efficient system for cell lysis, as they are arranged in a heptamer configuration and include a regulation domain to trigger cell lysis [418]. Considering CsgG maintains the structure of its bacterial homolog, we would expect another protein (probably CsgF, given their interaction) to act as the regulatory element. In this is case, it would imply that a group of phages that infect cells in nutrient-starved environments employ a larger (nonamer vs heptamer), two-protein structure to perform a function already done by other proteins widespread in bacteriophages, such as holins, spanins and Rz-lysis proteins [419].

The second function proposed by Buchholz *et al.* is based on the divergent structure of CsgF, where they suggest that CsgF's position in the CsgGF complex is inverted so that the extended α-helices of CsgF point into the periplasm. In this conformation, CsgGF is structurally similar to a secretin, a large protein superfamily involved in macromolecule transport across the membrane that includes the pIV system in filamentous phages [420,421]. It is important to note that the 3D structure of *E. coli* CsgF produced by Alphafold does not match Cryo-EM structures of the same protein [422], which does include the missing β-sheet. Although the

possibility of a chronic infection in oligotrophic phages is an exciting proposition, the CsgG pore (40-Å inner diameter) is too narrow to allow for myophage virion exit from the cell. While filamentous phages have a 70Å diameter [423], the diameter of the HTVC008M capsid is an order of magnitude larger.

We speculated that, in diluted environments such as oligotrophic waters, curli production might induce aggregation of potential successive hosts, increasing the chances of the newly produced virions of finding a host. However, Buchholz *et al.* did not observe evidence of clumping in their Melnitz phage cultures, neither in cytograms nor TEMs. Nevertheless, the case for curli production in bacteriophages is still open. Although the phage TSS VIII operon does not include homologs of CsgA or CsgB, it does code for various hypothetical proteins that could be their functional equivalents. Such proteins might not be found by sequence homology, as the proteins annotated as CsgA and CsgB in databases correspond to those in *E. coli*, and curlin protein homologs are highly variable even among Bacteria, varying in the number, position and type of repeat motifs [378]. In fact, we have observed that the TSS VIII operon in the cultured SAR11 phage HTVC008M includes 2 hypothetical proteins with the aforementioned motifs that conform into the β-sheet-rich secondary structure typical of amyloid-forming proteins [424]. Likewise, structural similarity between bacterial and phage CsgG is greatest in the periplasmic-facing α-helices of CsgG, which interact with CsgA in the bacterial complex [425].

5.4. PacBio long reads

Another of the goals of this thesis was to evaluate the suitability of third-generation sequencing technologies for viral metagenomics. This study aimed to understand whether the third-generation sequencing technology (PacBio) has addressed its characteristic high error rate and therefore was suitable for metagenomics, by comparing the phage communities recovered with those obtained from an Illumina dataset. With this technical focus in mind, we analysed a Mediterranean water column sample, as it is a well-known environment that has been extensively studied by multiple approaches, including fosmid cloning [153] and Illumina sequencing, both the viral fraction [105] and the viruses from the cellular fraction [193].

The results obtained here demonstrate once again that it is possible to recover a representative sample of the viral community from a cellular fraction metagenomic dataset. The benefits of this approach only increase with LR sequencing, not only in terms of DNA amount requirements (as stated before, PacBio DNA requirements are an order of magnitude larger than those of Illumina), but also with regard to the quality of the viral genomes recovered. Most of the viral species (*ca.* 75%) recovered in this study are either only recovered from LR datasets or the quality of the assembled genome is improved by the addition of the LR dataset to the assembly. In the case of the Nucleocytoplasmic Large DNA Viruses, the improvement is dramatic, with LR assembly resulting in contigs more than six times larger than their SRa counterparts. We believe this might be due to the fact that eukaryotic genomes have many repeats and other features that make their assembly from short-read metagenomes less efficient [426].

This is to our knowledge the first study of viruses extracted from a metagenomic sample that was sequenced using PacBio sequencing. Previous studies on the viral community with long-read datasets used Oxford Nanopore sequencing and applied either a complementary short-read dataset [209] or a correction based on coverage [208] to compensate for the high error rate. As evidenced by the results, PacBio CCS reads achieve an error rate similar to that of Illumina without the need for a correction step. However, this technology is more expensive and requires both more environmental DNA and of better quality (DNA template length can now be a limiting factor on read size, so DNA extraction protocols will also need to adapt to recover less fragmented DNA). It is important to note that third-generation sequencing is still in its infancy and improvements to the sequencing processes are still possible. As an example, Oxford Nanopore recently announced Kit 14 chemistry for Q20+ sequencing, which promises error rates similar to those of PacBio CCS. Nonetheless, even if error rate and read length issues can be ameliorated through improvements in the sequencing and DNA extraction processes respectively, the sizable increase in read length and dataset size means that long reads require the development of new pipelines and specialised software suites for their analysis.

The most damning evidence of this claim is our discovery of a large quantity of inter-clade diversity present in the long reads that is missing in both short-read and hybrid assemblies, laying bare the inadequacies of current assembly methods. There has been an effort in developing new assemblers for long-read prokaryote whole genome sequencing, which mainly differ in the approach employed to deal with noisy reads [427]. Most of them have abandoned the *de Brujin* graph approach popular in short-read assemblers in favour of Overlap Layout Consensus (OLC) algorithms, which can take advantage of the extra read length [427]. A recent benchmark paper reveals that while current software can result in circularised assemblies, continued development is needed to obtain better, more efficient assemblers [427]. A number of hybrid assemblers (assemblers that leverage the value of both short- and long-reads to perform the assembly) are also available, but in these assemblers long-read data is mostly employed to bridge gaps and resolve repeat regions in the short-read assembly [428]. Unfortunately, development of assemblers that can deal with the idiosyncrasies of metagenomic datasets is even further behind. For this particular sample, the long-read assemblers available (HiCanu and metaFlye) resulted in a lower quality assembly (smaller assembly size, smaller mean contig size) that the hybrid assembly produced by hybridSPAdes [204]. However, as the long-read sequences only complement the short read contigs, there is a large amount of information only found in the long read dataset that does not get included into the final assembly. Therefore, we end up with a contig collection that is technically better than the short-read assembly but is still subjected to its biases. It is clear that more development in the computational tools is needed in order to exploit the information contained in long read datasets.

Even if development of long-read assembly software continues, it could be argued that the best way forward for metagenomics involves forgoing the assembly step altogether, as long reads are large enough to perform gene calling. This argument is stronger for the analysis of phages for two reasons. First, phage genomes are significantly smaller than those of prokaryotes, with most bacteriophages ranging from 30 to 250kb [51]. Considering the mean

length of long reads is 15kb, a phage genome could be collected in a few overlapping reads or even a single read. Second, phage genomes are incredibly diverse. There have been multiple reports of short-read assemblies being biased against genomes from microdiverse populations [429], and single-cell sequencing confirmed the existence of this unexplored phage diversity [197]. Long-read sequencing has proved capable of recovering this lost microdiversity, as seen first in the analysis of genome-level nucleotide diversity performed by Dugdale *et al*, which revealed that phage contigs extracted from long reads were three times more diverse than those from short-read assemblies [209], and further corroborated by our results. With sequencing allowing a more thorough surveying of the viral community than single-cell experiments, we expect that as more long-read datasets become available we will obtain a more realistic picture of viral diversity. Although our results do not suggest that completely new clades of phages will be revealed by long-read sequencing, a more comprehensive catalogue of phages that infect the same host will deepen our understanding of host-phage interactions. The enhanced recovery of their flexible genome will also provide researchers with a larger gene pool from which to extract new genes with biotechnological applications. An example of this is the endolysin search performed in this work, in which four new clades of endolysins were found, with clade C4 including a motif not found in the previous dataset that affects its function.

5.5 RaFAH & Host-phage prediction

The results of previous studies meant facing viral dark matter once again, it is possible to recover a large amount of viral diversity that we know almost nothing about?. Arguably, the most critical information to know about a phage is the host(s) it can infect, both from an ecological (virus-host interactions are essential to understand the evolution of phages and their effect on the microbial community) and a biotechnological perspective (for example, phage therapy is based on exploiting these interactions to leverage the bactericidal effects of phages in a target host).

With this motivation, we have developed RaFAH, a new tool that uses a random forest that combines protein content features with the speed and flexibility of machine learning. Compared to other host prediction software, RaFAH can accurately predict more-host virus interactions from large environmental datasets, even for divergent viral genomes that escape detection by classical approaches. Furthermore, RaFAH reports accurate host predictions in samples from a wide variety of ecosystems and for phages that infect different kingdoms. In fact, RaFAH host predictions match the prokaryotic host community in each ecosystem sample. Although this agreement between virus and host community composition is to be expected, it is seldom observed in studies of viral ecology based on metagenomics because classical methods leave the majority of viruses without host predictions.

A common misconception is that machine learning approaches are fundamentally different from alignment-based or alignment-free methods. In fact, machine-learning-based approaches are still based on the same biological signals that previous methods employ, just transformed into measurable attributes the computer can understand. Therefore, it is not surprising that this choice of feature set, not the learning method, is the main factor that

determines the resolutive power of the model [430]. For RaFAH, this feature set consisted of similarity scores to protein clusters extracted *de novo* from the training set. The use of marker genes and RBPs for host assignment is an approach that has been successfully used to identify cyanobacteria-infecting phages and more recently, in a host classifier for phages infecting the ESKAPE group (an acronym standing for *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* species) [242]. The use of HMMs instead of direct protein-to-protein comparisons allowed us to alleviate the sequence divergence found between phage genomes. Perhaps more important is the capability of RaFAH to accurately detect phages that infect Archaea, allowing for a significant expansion of the archaeal virosphere and shed light on their yet poorly understood content of AMGs. Our archaeal AMG analysis in this work is severely constrained by protein function predictions, but our results show that viruses in Archaea include AMGs also found in bacteriophages, such as those related to nucleotide biosynthesis (*cobS*), maintenance of homeostasis (thermosomes) and post-translational regulation (*moeB*).

An advantage of machine learning approaches is that once the prediction model is trained, it can be probed for biological insights of the features used. As RaFAH uses protein clusters as its predicting features, a feature importance analysis will reveal which proteins are crucial for host prediction, which will include these proteins essential to the specific virus-host interaction or novel AMGs. When applied to RaFAH, the most important predictor was annotated as an Rz-like phage lysis protein (**Annex 3**:Table S1). Among the protein clusters that ranked among the 50 most important were multiple lysins, tail, and tail fiber proteins. These proteins are known to determine virus-host range, as they play fundamental roles in virus entry and exit and host recognition. In all, these results can provide insight on which proteins are relevant to the interactome between host and phage.

However, even with a carefully-chosen feature set the performance of the model is still subjected to the availability of a suitable training set. We are usually therefore constrained to making predictions about the small fraction of cellular life with known host-phage links. The use of HMMs alleviates but does not solve this problem, as tests show predictions made with RaFAH are better the more similar the query genomes are to the training set. The quality of the training set will also affect the performance, as the more diverse and complete the genomes comprising the training set are, the easier it is to find protein clusters that provide discriminatory power. For example, a protein cluster composed of conserved genes such as a terminase subunit will not be as useful as a cluster of tail fiber protein, which are specific for the cell wall of the target host. An example of such shortcomings can be found in the performance of RaFAH in Test Set 2, a dataset of viral genomic sequences derived from marine single amplified genomes (SAGs) [196]. Here, RaFAH scored below other classifiers. We consider these results to be caused by some features of the training set. First, most of the viruses identified in Test Set 2 were derived from single-cell genomes classified as either *Pelagibacter*, *Puniceispirillum*, *Prochlorococcus* or *Synechococcus*. This is expected considering these are the most abundant organisms in the ecosystem from which this dataset is derived. Nevertheless, this relatively low diversity of taxa has implications for the assessment of host-prediction tools. For instance, the genera *Prochlorococcus* and *Synechococcus* have no determined taxonomy at the level of class in the GTDB. Therefore,

predictions at this level do not count toward precision for these particular taxa and the precision of all host-prediction tools displayed a steep decrease at this taxonomic level. Second, the majority of bacteriophage genomes in Test Set 2 have very low completeness (median 6.85%, estimated by CheckV). The low diversity of hosts and the very low genome completeness likely impacted the performance of RaFAH on this dataset, even though RaFAH was trained on a dataset with a majority of genome fragments. We speculate that this is due to the fact that marker proteins or RBPs are usually located in metaviromic islands that are usually missing from short-read assemblies.

We also performed analysis of the combined effects of the relevant variables and how those, together, affected precision, recall, and the F1 score of RaFAH using Test Set 3. Taken together, these results demonstrated that the performance of RaFAH on a given genome is dependent on ecosystem source, genome completeness, similarity of the genome to those in the training dataset, and the taxonomic level being considered (see **Annex 3**:Table S6). For this reason, there is not a single score threshold that is ideal for all use cases. Nevertheless, we make the following recommendations. For differentiating between viruses of Bacteria and Archaea, RaFAH has nearly 100% precision even at the most permissive cutoff (0), thus for this particular purpose it can be applied without threshold. For a broad characterization of multiple viral genomes from an ecosystem, permissive thresholds are acceptable. For example, to compare viral host prevalences across different metagenomes at the level of phylum, we recommend a threshold of 0.14. This yields a precision of approximately 90% without sacrificing recall (**Annex 3**:Figures S1 and S4D), regardless of ecosystem source, genome length, completeness, or similarity to the training dataset. At lower taxonomic levels, stricter cutoffs are necessary. Users can select cutoffs according to the desired precision based on the curves depicted in **Annex 3**:Figures S1 and S4D. As a rule, longer, more complete genomes (Based on our tests, RaFAH is better suited for viral genomes with 50% or more completeness) with higher maximum AAI values to genomes in the training set should allow more permissive cutoffs.

6. Conclusions

CONCLUSIONS

1. Metagenome mining is a viable approach to increase the diversity of phage genomes from highly abundant, difficult-to-culture prokaryotes, as long as there is a reference genome available. The use of techniques such as cross-assembly can help mitigate biases against some kinds of phage genomes. The application of this method recovered 22 new myoviruses infecting SAR11, including the first freshwater representative.

2. Myoviruses infecting SAR11 present a T4-like genomic organization, with the inclusion of a large hypervariable region situated between structural operons. This region codes for several genes related to the phage-host interaction, such as tail fiber proteins and glycosyltransferases. We have designated this variable region the Host Recognition Cluster (HRC).

3. Pelagimyophages code for an operon that includes components of the type VIII secretion system. This operon is part of the viral flexible genome and in bacteria is related to the secretion of functional curli. No system of this kind has been described in his host. The function of this operon is still unclear, but it is plausible that it is involved in a response to external stimuli.

4. It is possible to recover a representative sample of the phage community from a long-read sequencing dataset derived from the cellular fraction, and it analysis is a good complementary alternative to the study of viral fraction datasets. No differences in phage community composition were detected and the recovered genomes in the long-read dataset are more complete.

5. Direct analysis of long reads revealed a vast phage diversity lost at the level of reads that did not lead to the discovery of new phage taxa, but did lay bare a wealth of inter-clade diversity. This loss of diversity is lost due to current assembler software not being able to fully exploit long-read datasets.

6. Long reads can help recover fragments of the viral genome difficult to retrieve due to assembly bias, as demonstrated by the fact that protein clusters predominantly found in long-read datasets are enriched in repeat regions and eukaryotic viruses showing a dramatic improvement in assembly size.

7. Protein cluster similarity is a suitable feature set to predict hosts in phage genomes obtained from metagenomic datasets, alleviating issues derived from phage protein divergence and allowing classification of all contigs at the taxonomic level of kingdom. However, it is still dependent on a suitable training set to improve over other feature sets such as tetranucleotide composition.

8. RaFAH is the first tool that can reliably classify Archaea phages, allowing for a significant expansion of the archaeal virosphere. Archaea AMGs follow the same categories as

bacterial AMGs, including genes related to nucleotide synthesis (*cobS*), chaperones (thermosomes) and regulation of post-translational modification (*moeB*).

CONCLUSIONES

1. El minado de metagenomas es un método viable para aumentar la diversidad de genomas de fagos que infectan a huéspedes procariotas que son abundantes en el entorno y difíciles de aislar, siempre y cuando haya disponible un genoma de referencia. El uso de técnicas como el *cross-assembly* pueden ayudar a mitigar los sesgos existentes hacia ciertos grupos de fagos. La aplicación de estas técnicas permitió la recuperación de 22 nuevos miovirus que infectan a SAR11, incluyendo el primer representante de agua dulce.

2. Los miovirus de SAR11 presentaun una organización genómica tipo T4, con la inclusión de una gran región hipervariable situada entre dos operones estructurales. Dicha región codifica diversos genes relacionados con la interacción fago-hospedador, como las proteínas de la fibra de la cola y glicosiltranferasas. Hemos denominado a esta región la Región de Reconocimiento del Hospedador (HRC, por sus siglas en inglés).

3. Los miovirus de SAR11 contienen un operón que incluye componentes del sistema de secreción VIII. Este operón es parte del genoma flexible viral y en bacterias está relacionado con la secreción de curlina. Este sistema de secreción no ha sido descrito en el hospedador. La función de este operón aún no se conoce, pero es posible que esté relacionado con la respuesta a estímulos externos.

4. Es posible recuperar una muestra representativa de la comunidad de fagos a partir de una set de datos de lectura larga derivada de la fracción celular. Asimismo, su análisis es una alternativa complementaria y razonable al estudio de sets de datos de la fracción vírica. No se detectaron diferencias en la composición de comunidad vírica y los genomas recuperados son más completos.

5. El análisis de las lecturas largas sin ensamblar reveló una gran diversidad de secuencias víricas que si bien no escondía nuevos clados virales, sí mostró una gran variación intra-clado no encontrada hasta la fecha. Esta pérdida de diversidad se debe a que los ensambladores modernos no es capaz de aprovechar la información extra contenida en las lecturas largas.

6. Las lecturas largas pueden recuperar regiones del genoma viral difíciles de conseguir por culpa de los sesgos de ensamblaje, como demuestra el hecho de que los sets de datos de lectura larga están enriquecidos con secuencias repetitivas y virus eucariotas mostrando una mejora dramática en tamaño de ensamblaje.

7. La similaridad entre clústeres de proteínas en una propiedad viable para predecir hospedadores de genomas víricos obtenidos de muestras metagenómicas, aliviando problemas derivados de la divergencia entre proteínas virales y permitiendo la

clasificación de todas las secuencias a nivel taxonómico de Reino. Sin embargo, se requiere un set de entrenamiento adecuado para mejorar los resultados obtenidos con otras propiedades, como la composición de tetranucleótidos.

8. RaFAH es la primera herramienta capaz de clasificar de manera fiable los fagos que infectan a arqueas, permitiendo una expansión de las colecciones de fagos. Los AMGs de fagos de arqueas contienen las mismas categorías que los AMGs bacterianos, incluyendo genes relacionados con la síntesis de nucleótidos (*cobS*), chaperonas (termosomas) y la regulación de la modificación post-traduccional (*moeB*).

7. References

1. Karl DM. Microbial oceanography: paradigms, processes and promise. Nat Rev Microbiol. 2007;5: 759–769.

2. Suttle CA. Marine viruses — major players in the global ecosystem. Nature Reviews Microbiology. 2007. pp. 801–812. doi:10.1038/nrmicro1750

3. Pinet PR. Invitation to Oceanography. Jones & Bartlett Publishers; 2011.

4. Goericke R, Welschmeyer NA. The marine prochlorophyte Prochlorococcus contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. Deep Sea Res Part I. 1993;40: 2283–2294.

5. Børsheim KY. Native marine bacteriophages. FEMS Microbiol Ecol. 1993;11: 141–159.

6. Wilhelm SW, Suttle CA. Viruses and Nutrient Cycles in the Sea. BioScience. 1999. pp. 781–788. doi:10.2307/1313569

7. Brum JR, Sullivan MB. Rising to the challenge: accelerated pace of discovery transforms marine virology. Nat Rev Microbiol. 2015;13: 147–159.

8. Torrella F, Morita RY. Evidence by electron micrographs for a high incidence of bacteriophage particles in the waters of Yaquina Bay, oregon: ecological and taxonomical implications. Appl Environ Microbiol. 1979;37: 774–778.

9. Bergh O, Børsheim KY, Bratbak G, Heldal M. High abundance of viruses found in aquatic environments. Nature. 1989;340: 467–468.

10. Noble RT, Fuhrman JA. Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aquat Microb Ecol. 1998;14: 113–118.

11. Hennes KP, Suttle CA. Direct counts of viruses in natural waters and laboratory cultures by epifluorescence microscopy. Limnology and Oceanography. 1995. pp. 1050–1055. doi:10.4319/lo.1995.40.6.1050

12. Suttle CA, Chan AM, Cottrell MT. Infection of phytoplankton by viruses and reduction of primary productivity. Nature. 1990;347: 467–469.

13. Proctor LM, Fuhrman JA. Viral mortality of marine bacteria and cyanobacteria. Nature. 1990;343: 60–62.

14. Suttle CA, Chen F. Mechanisms and rates of decay of marine viruses in seawater. Appl Environ Microbiol. 1992;58: 3721–3729.

15. Mann NH. The third age of phage. PLoS Biol. 2005;3: e182.

16. Tuomi P. Viral lysis and grazing loss of bacteria in nutrient- and carbon-manipulated brackish water enclosures. Journal of Plankton Research. 1999. pp. 923–937. doi:10.1093/plankt/21.5.923

17. Steward GF, Smith DC, Azam F. Abundance and production of bacteria and viruses in the Bering and Chukchi Seas. Mar Ecol Prog Ser. 1996;131: 287–300.

18. Li WKW. Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting. Limnol Oceanogr. 1994;39: 169–175.

19. Weitz JS, Wilhelm SW. Ocean viruses and their effects on microbial communities and biogeochemical cycles. F1000 Biol Rep. 2012;4: 17.

20. Breitbart M, Bonnain C, Malki K, Sawaya NA. Phage puppet masters of the marine microbial realm. Nat Microbiol. 2018;3: 754–766.

21. Pourtois J, Tarnita CE, Bonachela JA. Impact of Lytic Phages on Phosphorus- vs. Nitrogen-Limited Marine

Microbes. Frontiers in Microbiology. 2020. doi:10.3389/fmicb.2020.00221

22. Poulton AJ. Shunt or shuttle. Nat Geosci. 2021;14: 181–183.

23. Hutchinson GE. The paradox of the plankton. Am Nat. 1961;95: 137–145.

24. Sintes E, Del Giorgio PA. Feedbacks between protistan single-cell activity and bacterial physiological structure reinforce the predator/prey link in microbial foodwebs. Front Microbiol. 2014;5: 453.

25. Leggett HC, Buckling A, Long GH, Boots M. Generalism and the evolution of parasite virulence. Trends Ecol Evol. 2013;28: 592–596.

26. Korn-Wendisch, Schneider. Phage typing—a useful tool in actinomycete systematics. Gene. Available: https://www.sciencedirect.com/science/article/pii/0378111992905657

27. Anderson RM, May RM. The population dynamics of microparasites and their invertebrate hosts. Philos Trans R Soc Lond B Biol Sci. 1981;291: 451–524.

28. Brockhurst MA, Fenton A, Roulston B, Rainey PB. The impact of phages on interspecific competition in experimental populations of bacteria. BMC Ecol. 2006;6: 19.

29. Williams HTP. Phage-induced diversification improves host evolvability. BMC Evol Biol. 2013;13: 17.

30. Thingstad TF. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnol Oceanogr. 2000;45: 1320–1328.

31. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7: 828–836.

32. Koskella B, Brockhurst MA. Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. FEMS Microbiol Rev. 2014;38: 916–931.

33. Van Valen L. A new evolutionary law. Evol Theory. 1973;1: 1–30.

34. Ignacio-Espinoza JC, Ahlgren NA, Fuhrman JA. Long-term stability and Red Queen-like strain dynamics in marine viruses. Nat Microbiol. 2020;5: 265–271.

35. Jiang SC, Paul JH. Gene transfer by transduction in the marine environment. Appl Environ Microbiol. 1998;64: 2780–2787.

36. Modi SR, Lee HH, Spina CS, Collins JJ. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. Nature. 2013;499: 219–222.

37. Grose JH, Casjens SR. Understanding the enormous diversity of bacteriophages: the tailed phages that infect the bacterial family Enterobacteriaceae. Virology. 2014;468-470: 421–443.

38. Touchon M, Moura de Sousa JA, Rocha EP. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. Curr Opin Microbiol. 2017;38: 66–73.

39. Pant A, Das B, Bhadra RK. CTX phage of Vibrio cholerae: Genomics and applications. Vaccine. 2020. pp. A7–A12. doi:10.1016/j.vaccine.2019.06.034

40. Rohwer F, Thurber RV. Viruses manipulate the marine environment. Nature. 2009;459: 207–212.

41. Moon K, Jeon JH, Kang I, Park KS, Lee K, Cha C-J, et al. Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. Microbiome. 2020;8: 75.

42. Kutter E, Sulakvelidze A. Bacteriophages: biology and applications. [cited 26 May 2022]. doi:10.1201/9780203491751/bacteriophages-elizabeth-kutter-alexander-sulakvelidze

43. Nobrega FL, Vlot M, de Jonge PA, Dreesens LL, Beaumont HJE, Lavigne R, et al. Targeting mechanisms of

tailed bacteriophages. Nat Rev Microbiol. 2018;16: 760–773.

44. Zinke M, Schröder GF, Lange A. Major tail proteins of bacteriophages of the order Caudovirales. J Biol Chem. 2022;298: 101472.

45. Ackermann H-W. Tailed Bacteriophages: The Order Caudovirales. In: Maramorosch K, Murphy FA, Shatkin AJ, editors. Advances in Virus Research. Academic Press; 1998. pp. 135–201.

46. Garcia-Doval C, van Raaij MJ. Bacteriophage receptor recognition and nucleic acid transfer. Subcell Biochem. 2013;68: 489–518.

47. Leiman PG, Arisaka F, van Raaij MJ, Kostyuchenko VA, Aksyuk AA, Kanamaru S, et al. Morphogenesis of the T4 tail and tail fibers. Virol J. 2010;7: 355.

48. Taylor NMI, Prokhorov NS, Guerrero-Ferreira RC, Shneider MM, Browning C, Goldie KN, et al. Structure of the T4 baseplate and its function in triggering sheath contraction. Nature. 2016;533: 346–352.

49. Rakhuba DV, Kolomiets EI, Dey ES, Novik GI. Bacteriophage receptors, mechanisms of phage adsorption and penetration into host cell. Pol J Microbiol. 2010;59: 145–155.

50. Letarov AV, Kulikov EE. Adsorption of Bacteriophages on Bacterial Cells. Biochemistry . 2017;82: 1632–1658.

51. Dion MB, Oechslin F, Moineau S. Phage diversity, genomics and phylogeny. Nat Rev Microbiol. 2020;18: 125–138.

52. Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, Devoto A, et al. Clades of huge phages from across Earth's ecosystems. Nature. 2020;578: 425–431.

53. Hatfull GF, Hendrix RW. Bacteriophages and their genomes. Curr Opin Virol. 2011;1: 298–303.

54. Comeau AM, Bertrand C, Letarov A, Tétart F, Krisch HM. Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. Virology. 2007;362: 384–396.

55. Casjens SR. Comparative genomics and evolution of the tailed-bacteriophages. Curr Opin Microbiol. 2005;8: 451–458.

56. Koonin EV, Dolja VV, Krupovic M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology. 2015;479-480: 2–25.

57. Lopes A, Tavares P, Petit M-A, Guérois R, Zinn-Justin S. Automated classification of tailed bacteriophages according to their neck organization. BMC Genomics. 2014;15: 1027.

58. Khor BY, Tye GJ, Lim TS, Choong YS. General overview on structure prediction of twilight-zone proteins. Theor Biol Med Model. 2015;12: 15.

59. Lopes A, Amarir-Bouhram J, Faure G, Petit M-A, Guerois R. Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. Nucleic Acids Research. 2010. pp. 3952–3962. doi:10.1093/nar/gkq096

60. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins. 2009;77: 499–508.

61. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, et al. Protein function annotation by homology-based inference. Genome Biol. 2009;10: 207.

62. Sinha S, Lynn AM, Desai DK. Implementation of homology based and non-homology based computational methods for the identification and annotation of orphan enzymes: using Mycobacterium tuberculosis H37Rv as a case study. BMC Bioinformatics. 2020;21: 466.

63. Russell PW, Müller UR. Construction of bacteriophage luminal diameterX174 mutants with maximum genome sizes. J Virol. 1984;52: 822–827.

64. Hatfull GF, Jacobs-Sera D, Lawrence JG, Pope WH, Russell DA, Ko C-C, et al. Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. J Mol Biol. 2010;397: 119–143.

65. Cahill J, Rajaure M, O'Leary C, Sloan J, Marrufo A, Holt A, et al. Genetic Analysis of the Lambda Spanins Rz and Rz1: Identification of Functional Domains. G3 . 2017;7: 741–753.

66. Chirico N, Vianelli A, Belshaw R. Why genes overlap in viruses. Proc Biol Sci. 2010;277: 3809–3817.

67. Christie GE, Temple LM, Bartlett BA, Goodwin TS. Programmed translational frameshift in the bacteriophage P2 FETUD tail gene operon. J Bacteriol. 2002;184: 6522–6531.

68. Xu J, Hendrix RW, Duda RL. Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. Mol Cell. 2004;16: 11–21.

69. Steczkiewicz K, Prestel E, Bidnenko E, Szczepankowska AK. Expanding Diversity of Firmicutes Single-Strand Annealing Proteins: A Putative Role of Bacteriophage-Host Arms Race. Front Microbiol. 2021;12: 644622.

70. Hatfull GF. Bacteriophage genomics. Current Opinion in Microbiology. 2008. pp. 447–453. doi:10.1016/j.mib.2008.09.004

71. Casjens SR, Thuman-Commike PA. Evolution of mosaically related tailed bacteriophage genomes seen through the lens of phage P22 virion assembly. Virology. 2011;411: 393–415.

72. Lawrence JG, Hatfull GF, Hendrix RW. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J Bacteriol. 2002;184: 4891–4905.

73. Casjens SR, Grose JH. Contributions of P2- and P22-like prophages to understanding the enormous diversity and abundance of tailed bacteriophages. Virology. 2016;496: 255–276.

74. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci U S A. 1999;96: 2192–2197.

75. Martinsohn JT, Radman M, Petit M-A. The lambda red proteins promote efficient recombination between diverged sequences: implications for bacteriophage genome mosaicism. PLoS Genet. 2008;4: e1000065.

76. De Paepe M, Hutinet G, Son O, Amarir-Bouhram J, Schbath S, Petit M-A. Temperate phages acquire DNA from defective prophages by relaxed homologous recombination: the role of Rad52-like recombinases. PLoS Genet. 2014;10: e1004181.

77. Morris P, Marinelli LJ, Jacobs-Sera D, Hendrix RW, Hatfull GF. Genomic characterization of mycobacteriophage Giles: evidence for phage acquisition of host DNA by illegitimate recombination. J Bacteriol. 2008;190: 2172–2182.

78. Mavrich TN, Hatfull GF. Bacteriophage evolution differs by host, lifestyle and genome. Nature Microbiology. 2017. doi:10.1038/nmicrobiol.2017.112

79. Mizuno CM, Ghai R, Rodriguez-Valera F. Evidence for metaviromic islands in marine phages. Front Microbiol. 2014;5: 27.

80. Markine-Goriaynoff N, Gillet L, Van Etten JL, Korres H, Verma N, Vanderplasschen A. Glycosyltransferases encoded by viruses. J Gen Virol. 2004;85: 2741–2754.

81. Warwick-Dugdale J, Buchholz HH, Allen MJ, Temperton B. Host-hijacking and planktonic piracy: how phages command the microbial high seas. Virol J. 2019;16: 15.

82. Jacobson TB, Callaghan MM, Amador-Noguez D. Hostile Takeover: How Viruses Reprogram Prokaryotic

Metabolism. Annu Rev Microbiol. 2021;75: 515–539.

83. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, et al. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci U S A. 2011;108: E757–64.

84. Dammeyer T, Bagby SC, Sullivan MB, Chisholm SW, Frankenberg-Dinkel N. Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. Curr Biol. 2008;18: 442–448.

85. Hurwitz BL, Brum JR, Sullivan MB. Depth-stratified functional and taxonomic niche specialization in the "core" and "flexible" Pacific Ocean Virome. ISME J. 2015;9: 472–484.

86. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28: 27–30.

87. Hurwitz BL, Hallam SJ, Sullivan MB. Metabolic reprogramming by viruses in the sunlit and dark ocean. Genome Biol. 2013;14: R123.

88. Crummett LT, Puxty RJ, Weihe C, Marston MF, Martiny JBH. The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. Virology. 2016. pp. 219–229. doi:10.1016/j.virol.2016.09.016

89. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. Photosynthesis genes in marine viruses yield proteins during host infection. Nature. 2005;438: 86–89.

90. Lindell D, Jaffe JD, Coleman ML, Futschik ME, Axmann IM, Rector T, et al. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature. 2007;449: 83–86.

91. Huang X, Jiao N, Zhang R. The genomic content and context of auxiliary metabolic genes in roseophages. Environ Microbiol. 2021;23: 3743–3757.

92. Miller ES, Heidelberg JF, Eisen JA, Nelson WC, Durkin AS, Ciecko A, et al. Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. J Bacteriol. 2003;185: 5220–5233.

93. Ahlgren NA, Fuchsman CA, Rocap G, Fuhrman JA. Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. ISME J. 2019;13: 618–631.

94. Kavagutti VS, Andrei A-Ş, Mehrshad M, Salcher MM, Ghai R. Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. Microbiome. 2019. doi:10.1186/s40168-019-0752-0

95. Chen L-X, Méheust R, Crits-Christoph A, McMahon KD, Nelson TC, Slater GF, et al. Large freshwater phages with the potential to augment aerobic methane oxidation. Nat Microbiol. 2020;5: 1504–1515.

96. Roux S, Hawley AK, Beltran MT, Scofield M, Schwientek P, Stepanauskas R, et al. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. eLife. 2014. doi:10.7554/elife.03125

97. Anantharaman K, Duhaime MB, Breier JA, Wendt KA, Toner BM, Dick GJ. Sulfur oxidation genes in diverse deep-sea viruses. Science. 2014;344: 757–760.

98. Kieft K, Zhou Z, Anderson RE, Buchan A, Campbell BJ, Hallam SJ, et al. Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. Nat Commun. 2021;12: 3503.

99. Cassman N, Prieto-Davó A, Walsh K, Silva GGZ, Angly F, Akhter S, et al. Oxygen minimum zones harbour novel viral communities with low diversity. Environ Microbiol. 2012;14: 3043–3065.

100. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature. 2016;537: 689–693.

101. Zeng Q, Chisholm SW. Marine Viruses Exploit Their Host's Two-Component Regulatory System in

Response to Resource Limitation. Current Biology. 2012. pp. 124–128. doi:10.1016/j.cub.2011.11.055

102. Kelly L, Ding H, Huang KH, Osburne MS, Chisholm SW. Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. ISME J. 2013;7: 1827–1841.

103. Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI, et al. The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS One. 2008;3: e1456.

104. Morgan GJ, Pitts WB. Evolution without species: The case of mosaic bacteriophages. Br J Philos Sci. 2008;59: 745–765.

105. Coutinho FH, Rosselli R, Rodríguez-Valera F. Trends of Microdiversity Reveal Depth-Dependent Evolutionary Strategies of Viruses in the Mediterranean. mSystems. 2019. doi:10.1128/msystems.00554-19

106. Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, Cobián-Güemes AG, et al. Lytic to temperate switching of viral communities. Nature. 2016;531: 466–470.

107. Erez Z, Steinberger-Levy I, Shamir M, Doron S, Stokar-Avihail A, Peleg Y, et al. Communication between viruses guides lysis–lysogeny decisions. Nature. 2017. pp. 488–493. doi:10.1038/nature21049

108. Trinh JT, Székely T, Shao Q, Balázsi G, Zeng L. Cell fate decisions emerge as phages cooperate or compete inside their host. Nat Commun. 2017;8: 14341.

109. Abedon ST. Chapter 1 Phage Evolution and Ecology. Advances in Applied Microbiology. 2009. pp. 1–45. doi:10.1016/s0065-2164(08)01001-0

110. Rakonjac J, Bennett NJ, Spagnuolo J, Gagic D, Russel M. Filamentous bacteriophage: biology, phage display and nanotechnology applications. Curr Issues Mol Biol. 2011;13: 51–76.

111. Sternberg N, Austin S. The maintenance of the P1 plasmid prophage. Plasmid. 1981;5: 20–31.

112. Bastías R, Higuera G, Sierralta W, Espejo RT. A new group of cosmopolitan bacteriophages induce a carrier state in the pandemic strain of Vibrio parahaemolyticus. Environ Microbiol. 2010;12: 990–1000.

113. Bradley DE. Ultrastructure of bacteriophage and bacteriocins. Bacteriological Reviews. 1967. pp. 230–314. doi:10.1128/br.31.4.230-314.1967

114. Ackermann H-W, Eisenstark A. The Present State of Phage Taxonomy. Intervirology. 1974. pp. 201–219. doi:10.1159/000149758

115. Turner D, Kropinski AM, Adriaenssens EM. A Roadmap for Genome-Based Phage Taxonomy. Viruses. 2021. p. 506. doi:10.3390/v13030506

116. Rohwer F, Edwards R. The Phage Proteomic Tree: a Genome-Based Taxonomy for Phage. Journal of Bacteriology. 2002. pp. 4529–4535. doi:10.1128/jb.184.16.4529-4535.2002

117. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. ViPTree: the viral proteomic tree server. Bioinformatics. 2017. pp. 2379–2380. doi:10.1093/bioinformatics/btx157

118. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Reticulate Representation of Evolutionary and Functional Relationships between Phage Genomes. Molecular Biology and Evolution. 2008. pp. 762–777. doi:10.1093/molbev/msn023

119. Iranzo J, Krupovic M, Koonin EV. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. mBio. 2016. doi:10.1128/mbio.00978-16

120. Low SJ, Džunková M, Chaumeil P-A, Parks DH, Hugenholtz P. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales.

Nature Microbiology. 2019. pp. 1306–1315. doi:10.1038/s41564-019-0448-z

121. Aiewsakun P, Adriaenssens EM, Lavigne R, Kropinski AM, Simmonds P. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy. J Gen Virol. 2018;99: 1331–1343.

122. Andrade-Martínez JS, Moreno-Gallego JL, Reyes A. Defining a Core Genome for the Herpesvirales and Exploring their Evolutionary Relationship with the Caudovirales. Sci Rep. 2019;9: 11342.

123. Bolduc B, Jang HB, Doulcier G, You Z-Q, Roux S, Sullivan MB. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect*Archaea*and*Bacteria*. PeerJ. 2017. p. e3243. doi:10.7717/peerj.3243

124. Jang HB, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nature Biotechnology. 2019. pp. 632–639. doi:10.1038/s41587-019-0100-8

125. Lederberg EM, Lederberg J. GENETIC STUDIES OF LYSOGENICITY IN ESCHERICHIA COLI. Genetics. 1953. pp. 51–64. doi:10.1093/genetics/38.1.51

126. Grimes DJ, Atwell RW, Brayton PR, Palmer LM, Rollins DM, Roszak DB, et al. The fate of enteric pathogenic bacteria in estuarine and marine environments. Microbiol Sci. 1986;3: 324–329.

127. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. Nat Protoc. 2009;4: 470–483.

128. Nagasaki K, Bratbak G. Isolation of viruses infecting photosynthetic and nonphotosynthetic protists. Manual of Aquatic Viral Ecology. 2010. pp. 92–101. doi:10.4319/mave.2010.978-0-9845591-0-7.92

129. Fischer CR, Yoichi M, Unno H, Tanji Y. The coexistence of*Escherichia coli*serotype O157:H7 and its specific bacteriophage in continuous culture. FEMS Microbiology Letters. 2004. pp. 171–177. doi:10.1016/j.femsle.2004.10.017

130. Carey-Smith GV, Billington C, Cornelius AJ, Andrew Hudson J, Heinemann JA. Isolation and characterization of bacteriophages infecting*Salmonella*spp. FEMS Microbiology Letters. 2006. pp. 182–186. doi:10.1111/j.1574-6968.2006.00217.x

131. Clokie MRJ, Kropinski A. Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions. Humana Press; 2008.

132. Männistö RH, Kivelä HM, Paulin L, Bamford DH, Bamford JKH. The Complete Genome Sequence of PM2, the First Lipid-Containing Bacterial Virus To Be Isolated. Virology. 1999. pp. 355–363. doi:10.1006/viro.1999.9837

133. Chen F, Lu J. Genomic Sequence and Evolution of Marine Cyanophage P60: a New Insight on Lytic and Lysogenic Phages. Applied and Environmental Microbiology. 2002. pp. 2589–2594. doi:10.1128/aem.68.5.2589-2594.2002

134. Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC, et al. Abundant SAR11 viruses in the ocean. Nature. 2013;494: 357–360.

135. Du S, Qin F, Zhang Z, Tian Z, Yang M, Liu X, et al. Genomic diversity, life strategies and ecology of marine HTVC010P-type pelagiphages. Microbial Genomics. 2021. doi:10.1099/mgen.0.000596

136. Kang I, Oh H-M, Kang D, Cho J-C. Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. Proceedings of the National Academy of Sciences. 2013. pp. 12343–12348. doi:10.1073/pnas.1219930110

137. Yang M, Xia Q, Du S, Zhang Z, Qin F, Zhao Y. Genomic Characterization and Distribution Pattern of a Novel Marine OM43 Phage. Frontiers in Microbiology. 2021. doi:10.3389/fmicb.2021.651326

138. Buchholz HH, Bolaños LM, Bell AG, Michelsen ML, Allen MJ, Temperton B. A Novel and Ubiquitous Marine Methylophage Provides Insights into Viral-Host Coevolution and Possible Host-Range Expansion in Streamlined Marine Heterotrophic Bacteria. Applied and Environmental Microbiology. 2022. doi:10.1128/aem.00255-22

139. Cai L, Ma R, Chen H, Yang Y, Jiao N, Zhang R. A newly isolated roseophage represents a distinct member of Siphoviridae family. Virology Journal. 2019. doi:10.1186/s12985-019-1241-6

140. Zhang Z, Chen F, Chu X, Zhang H, Luo H, Qin F, et al. Diverse, Abundant, and Novel Viruses Infecting the Marine *Roseobacter* RCA Lineage. mSystems. 2019. doi:10.1128/msystems.00494-19

141. Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. Curr Opin Virol. 2012;2: 63–77.

142. Göller PC, Haro-Moreno JM, Rodriguez-Valera F, Loessner MJ, Gómez-Sanz E. Uncovering a hidden diversity: optimized protocols for the extraction of dsDNA bacteriophages from soil. Microbiome. 2020;8: 17.

143. Greninger AL. A decade of RNA virus metagenomics is (not) enough. Virus Res. 2018;244: 218–229.

144. Edwards RA, Rohwer F. Viral metagenomics. Nature Reviews Microbiology. 2005. pp. 504–510. doi:10.1038/nrmicro1163

145. Ghai R, Martin-Cuadrado A-B, Molto AG, Heredia IG, Cabrera R, Martin J, et al. Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. ISME J. 2010;4: 1154–1166.

146. Cai L, Yang Y, Jiao N, Zhang R. Evaluation of Tangential Flow Filtration for the Concentration and Separation of Bacteria and Viruses in Contrasting Marine Environments. PLOS ONE. 2015. p. e0136741. doi:10.1371/journal.pone.0136741

147. Sun G, Xiao J, Wang H, Gong C, Pan Y, Yan S, et al. Efficient purification and concentration of viruses from a large body of high turbidity seawater. MethodsX. 2014. pp. 197–206. doi:10.1016/j.mex.2014.09.001

148. John SG, Mendez CB, Deng L, Poulos B, Kauffman AKM, Kern S, et al. A simple and efficient method for concentration of ocean viruses by chemical flocculation. Environmental Microbiology Reports. 2011. pp. 195–202. doi:10.1111/j.1758-2229.2010.00208.x

149. Corinaldesi C, Tangherlini M, Dell'Anno A. From virus isolation to metagenome generation for investigating viral diversity in deep-sea sediments. Scientific Reports. 2017. doi:10.1038/s41598-017-08783-4

150. Colombet J, Robin A, Lavie L, Bettarel Y, Cauchie HM, Sime-Ngando T. Virioplankton "pegylation": Use of PEG (polyethylene glycol) to concentrate and purify viruses in pelagic ecosystems. Journal of Microbiological Methods. 2007. pp. 212–219. doi:10.1016/j.mimet.2007.08.012

151. Langenfeld K, Chin K, Roy A, Wigginton K, Duhaime MB. Comparison of ultrafiltration and iron chloride flocculation in the preparation of aquatic viromes from contrasting sample types. PeerJ. 2021;9: e11111.

152. Bekliz M, Brandani J, Bourquin M, Battin T, Peter H. Benchmarking protocols for the metagenomic analysis of stream biofilm viromes. doi:10.7287/peerj.preprints.27914v1

153. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. Expanding the Marine Virosphere Using Metagenomics. PLoS Genetics. 2013. p. e1003987. doi:10.1371/journal.pgen.1003987

154. Wang I-N, Smith DL, Young R. Holins: The Protein Clocks of Bacteriophage Infections. Annual Review of Microbiology. 2000. pp. 799–825. doi:10.1146/annurev.micro.54.1.799

155. Rohwer F, Seguritan V, Choi DH, Segall AM, Azam F. Production of Shotgun Libraries Using Random Amplification. BioTechniques. 2001. pp. 108–118. doi:10.2144/01311rr02

156. Warren RA. Modified bases in bacteriophage DNAs. Annu Rev Microbiol. 1980;34: 137–158.

157. Mtimka S, Pillay P, Rashamuse K, Gildenhuys S, Tsekoa TL. Functional screening of a soil metagenome for DNA endonucleases by acquired resistance to bacteriophage infection. Molecular Biology Reports. 2020. pp. 353–361. doi:10.1007/s11033-019-05137-3

158. Zhou M, Abid M, Yin H, Wu H, Teklue T, Qiu H-J, et al. Establishment of an Efficient and Flexible Genetic Manipulation Platform Based on a Fosmid Library for Rapid Generation of Recombinant Pseudorabies Virus. Front Microbiol. 2018;9: 2132.

159. Spits C, Le Caignec C, De Rycke M, Van Haute L, Van Steirteghem A, Liebaers I, et al. Whole-genome multiple displacement amplification from single cells. Nature Protocols. 2006. pp. 1965–1970. doi:10.1038/nprot.2006.326

160. Binga EK, Lasken RS, Neufeld JD. Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. The ISME Journal. 2008. pp. 233–241. doi:10.1038/ismej.2008.10

161. Marine R, McCarren C, Vorrasane V, Nasko D, Crowgey E, Polson SW, et al. Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. Microbiome. 2014. doi:10.1186/2049-2618-2-3

162. Yilmaz S, Allgaier M, Hugenholtz P. Multiple displacement amplification compromises quantitative analysis of metagenomes. Nature Methods. 2010. pp. 943–944. doi:10.1038/nmeth1210-943

163. Sabina J, Leamon JH. Bias in Whole Genome Amplification: Causes and Considerations. Methods Mol Biol. 2015;1347: 15–41.

164. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74: 5463–5467.

165. Shokralla S, Spall JL, Gibson JF, Hajibabaei M. Next-generation sequencing technologies for environmental DNA research. Mol Ecol. 2012;21: 1794–1805.

166. Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS One. 2011;6: e17497.

167. Hambly E, Suttle CA. The viriosphere, diversity, and genetic exchange within phage communities. Curr Opin Microbiol. 2005;8: 444–450.

168. Short SM, Suttle CA. Use of the polymerase chain reaction and denaturing gradient gel electrophoresis to study diversity in natural virus communities. In: Zehr JP, Voytek MA, editors. Molecular Ecology of Aquatic Communities. Dordrecht: Springer Netherlands; 1999. pp. 19–32.

169. Short CM, Suttle CA. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. Appl Environ Microbiol. 2005;71: 480–486.

170. Zhong Y, Chen F, Wilhelm SW, Poorvin L, Hodson RE. Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. Appl Environ Microbiol. 2002;68: 1576–1584.

171. Marston MF, Sallee JL. Genetic diversity and temporal variation in the cyanophage community infecting marine Synechococcus species in Rhode Island's coastal waters. Appl Environ Microbiol. 2003;69: 4639–4647.

172. Wilson WH, Fuller NJ, Joint IR, Mann NH. Analysis of cyanophage diversity and population structure in a south-north transect of the Atlantic Ocean. BULLETIN-INSTITUT OCEANOGRAPHIQUE MONACO-NUMERO SPECIAL-. 1999; 209–216.

173. Frederickson CM, Short SM, Suttle CA. The physical environment affects cyanophage communities in British Columbia inlets. Microb Ecol. 2003;46: 348–357.

174. Breitbart M, Miyake JH, Rohwer F. Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiol Lett. 2004;236: 249–256.

175. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, et al. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A. 2002;99: 14250–14255.

176. Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P, et al. Diversity and population structure of a near-shore marine-sediment viral community. Proc Biol Sci. 2004;271: 565–574.

177. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17: 333–351.

178. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Front Plant Sci. 2014;5: 209.

179. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, et al. The marine viromes of four oceanic regions. PLoS Biol. 2006;4: e368.

180. Roux, Borodovsky. Current Techniques and Approaches for Metagenomic Exploration of Phage Diversity. Virus Bioinformatics. Available: https://books.google.com/books?hl=en&lr=&id=50k2EAAAQBAJ&oi=fnd&pg=PA17&dq=Current+Techniques+and+approaches+for+Metagenomic+Exploration+of+Phage+Diversity&ots=iUifV0p6Zj&sig=dHkYy2JhbuQumtqxq--TXpKe0do

181. Rodriguez-Valera F, Mizuno CM, Ghai R. Tales from a thousand and one phages. Bacteriophage. 2014;4: e28265.

182. Hurwitz BL, Sullivan MB. The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. PLoS One. 2013;8: e57355.

183. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science. 2015;348: 1261498.

184. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. Cell. 2019;177: 1109–1123.e14.

185. Parsons RJ, Breitbart M, Lomas MW, Carlson CA. Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. ISME J. 2012;6: 273–284.

186. Perez Sepulveda B, Redgwell T, Rihtman B, Pitt F, Scanlan DJ, Millard A. Marine phage genomics: the tip of the iceberg. FEMS Microbiol Lett. 2016;363. doi:10.1093/femsle/fnw158

187. Lapidus AL, Korobeynikov AI. Metagenomic data assembly - the way of decoding unknown microorganisms. Front Microbiol. 2021;12: 613791.

188. Nelson WC, Maezato Y, Wu Y-W, Romine MF, Lindemann SR. Identification and Resolution of Microdiversity through Metagenomic Sequencing of Parallel Consortia. Appl Environ Microbiol. 2016;82: 255–267.

189. Hugerth LW, Larsson J, Alneberg J, Lindh MV, Legrand C, Pinhassi J, et al. Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biol. 2015;16: 279.

190. Sharon I, Banfield JF. Microbiology. Genomes from metagenomics. Science. 2013;342: 1057–1058.

191. Bendall ML, Stevens SL, Chan L-K, Malfatti S, Schwientek P, Tremblay J, et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 2016;10: 1589–1601.

192. Sangwan N, Xia F, Gilbert JA. Recovering complete and draft population genomes from metagenome datasets. Microbiome. 2016;4: 8.

193. López-Pérez M, Haro-Moreno JM, Gonzalez-Serrano R, Parras-Moltó M, Rodriguez-Valera F. Genome

diversity of marine phages recovered from Mediterranean metagenomes: Size matters. PLoS Genet. 2017;13: e1007018.

194. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016;17: 175–188.

195. Berube PM, Biller SJ, Hackl T, Hogle SL, Satinsky BM, Becker JW, et al. Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. Sci Data. 2018;5: 180154.

196. Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, et al. Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. Cell. 2019;179: 1623–1635.e11.

197. Martinez-Hernandez F, Fornas O, Lluesma Gomez M, Bolduc B, de la Cruz Peña MJ, Martínez JM, et al. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nat Commun. 2017;8: 15892.

198. Martinez-Hernandez F, Fornas Ò, Lluesma Gomez M, Garcia-Heredia I, Maestre-Carballa L, López-Pérez M, et al. Single-cell genomics uncover Pelagibacter as the putative host of the extremely abundant uncultured 37-F6 viral population in the ocean. ISME J. 2019;13: 232–236.

199. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13: 278–289.

200. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17: 239.

201. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009;4: 265–270.

202. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323: 133–138.

203. Athanasopoulou K, Boti MA, Adamopoulos PG, Skourou PC, Scorilas A. Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. Life. 2021;12. doi:10.3390/life12010030

204. Haro-Moreno JM, López-Pérez M, Rodriguez-Valera F. Enhanced Recovery of Microbial Genes and Genomes From a Marine Water Column Using Long-Read Metagenomics. Front Microbiol. 2021;12: 708782.

205. Zhang H, Jain C, Aluru S. A comprehensive evaluation of long read error correction methods. BMC Genomics. 2020;21: 889.

206. Fu S, Wang A, Au KF. A comparative evaluation of hybrid error correction methods for error-prone long reads. Genome Biol. 2019;20: 26.

207. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37: 1155–1162.

208. Beaulaurier J, Luo E, Eppley JM, Uyl PD, Dai X, Burger A, et al. Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. Genome Res. 2020;30: 437–446.

209. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, et al. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. PeerJ. 2019;7: e6800.

210. Youle M, Haynes M, Rohwer F. Scratching the Surface of Biology's Dark Matter. In: Witzany G, editor. Viruses: Essential Agents of Life. Dordrecht: Springer Netherlands; 2012. pp. 61–81.

211. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage-host relationships. FEMS Microbiol Rev. 2016;40: 258–272.

212. Canuti M, van der Hoek L. Virus discovery: are we scientists or genome collectors? Trends Microbiol. 2014;22: 229–231.

213. Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 2005;13: 278–284.

214. Mojica KDA, Brussaard CPD. Factors affecting virus dynamics and microbial host-virus interactions in marine environments. FEMS Microbiol Ecol. 2014;89: 495–515.

215. Johannessen TV, Larsen A, Bratbak G, Pagarete A, Edvardsen B, Egge ED, et al. Seasonal Dynamics of Haptophytes and dsDNA Algal Viruses Suggest Complex Virus-Host Relationship. Viruses. 2017;9. doi:10.3390/v9040084

216. Goldsmith DB, Brum JR, Hopkins M, Carlson CA, Breitbart M. Water column stratification structures viral community composition in the Sargasso Sea. Aquat Microb Ecol. 2015;76: 85–94.

217. Hargreaves KR, Anderson NJ, Clokie MRJ. Recovery of viable cyanophages from the sediments of a eutrophic lake at decadal timescales. FEMS Microbiol Ecol. 2013;83: 450–456.

218. Broman E, Holmfeldt K, Bonaglia S, Hall POJ, Nascimento FJA. Cyanophage Diversity and Community Structure in Dead Zone Sediments. mSphere. 2021;6. doi:10.1128/mSphere.00208-21

219. Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, et al. Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. Nature. 2014;513: 242–245.

220. Marston MF, Martiny JBH. Genomic diversification of marine cyanophages into stable ecotypes. Environ Microbiol. 2016;18: 4240–4253.

221. Ignacio-Espinoza JC, Solonenko SA, Sullivan MB. The global virome: not as big as we thought? Curr Opin Virol. 2013;3: 566–571.

222. Brussaard CP, Marie D, Bratbak G. Flow cytometric detection of viruses. J Virol Methods. 2000;85: 175–182.

223. Tomaru Y, Nagasaki K. Flow cytometric detection and enumeration of DNA and RNA viruses infecting marine eukaryotic microalgae. J Oceanogr. 2007;63: 215–221.

224. Holmfeldt K, Odić D, Sullivan MB, Middelboe M, Riemann L. Cultivated single-stranded DNA phages that infect marine Bacteroidetes prove difficult to detect with DNA-binding stains. Appl Environ Microbiol. 2012;78: 892–894.

225. Bettarel Y, Sime-Ngando T, Amblard C, Laveran H. A comparison of methods for counting viruses in aquatic systems. Appl Environ Microbiol. 2000;66: 2283–2289.

226. Weinbauer MG, Suttle CA. Comparison of epifluorescence and transmission electron microscopy for counting viruses in natural marine waters. Aquat Microb Ecol. 1997;13: 225–232.

227. Allers E, Moraru C, Duhaime MB, Beneze E, Solonenko N, Barrero-Canosa J, et al. Single-cell and population level viral infection dynamics revealed by phage FISH , a method to visualize intracellular and free viruses. Environmental Microbiology. 2013. pp. 2306–2318. doi:10.1111/1462-2920.12100

228. Tadmor AD, Ottesen EA, Leadbetter JR, Phillips R. Probing Individual Environmental Bacteria for Viruses by Using Microfluidic Digital PCR. Science. 2011. pp. 58–62. doi:10.1126/science.1200758

229. Sakowski EG, Arora-Williams K, Tian F, Zayed AA, Zablocki O, Sullivan MB, et al. Interaction dynamics and virus–host range for estuarine actinophages captured by epicPCR. Nature Microbiology. 2021. pp. 630–642. doi:10.1038/s41564-021-00873-4

230. Deng L, Gregory A, Yilmaz S, Poulos BT, Hugenholtz P, Sullivan MB. Contrasting Life Strategies of Viruses That Infect Photo- and Heterotrophic Bacteria, as Revealed by Viral Tagging. mBio. 2013. doi:10.1128/mbio.00516-12

231. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics. 2007;8: 172.

232. Ghai R, Mehrshad M, Mizuno CM, Rodriguez-Valera F. Metagenomic recovery of phage genomes of uncultured freshwater actinobacteria. ISME J. 2017;11: 304–308.

233. Versoza CJ, Pfeifer SP. Computational Prediction of Bacteriophage Host Ranges. Microorganisms. 2022;10. doi:10.3390/microorganisms10010149

234. Horvath P, Barrangou R. CRISPR/Cas, the Immune System of Bacteria and Archaea. Science. 2010. pp. 167–170. doi:10.1126/science.1179555

235. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015;3: e985.

236. Coclet C, Roux S. Global overview and major challenges of host prediction methods for uncultivated phages. Curr Opin Virol. 2021;49: 117–126.

237. Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol. 1997;44: 383–397.

238. Pride DT, Wassenaar TM, Ghose C, Blaser MJ. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics. 2006;7: 8.

239. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. eLife. 2015. doi:10.7554/elife.08490

240. Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, et al. Ocean plankton. Determinants of community structure in the global plankton interactome. Science. 2015;348: 1262073.

241. Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, et al. Viral and microbial community dynamics in four aquatic environments. ISME J. 2010;4: 739–751.

242. Boeckaerts D, Stock M, Criel B, Gerstmans H, De Baets B, Briers Y. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. Sci Rep. 2021;11: 1467.

243. Wang W, Ren J, Tang K, Dart E, Ignacio-Espinoza JC, Fuhrman JA, et al. A network-based integrated framework for predicting virus-prokaryote interactions. NAR Genom Bioinform. 2020;2: lqaa044.

244. Giovannoni SJ. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. Ann Rev Mar Sci. 2017;9: 231–255.

245. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, et al. SAR11 clade dominates ocean surface bacterioplankton communities. Nature. 2002;420: 806–810.

246. Ferla MP, Thrash JC, Giovannoni SJ, Patrick WM. New rRNA gene-based phylogenies of the Alphaproteobacteria provide perspective on major groups, mitochondrial ancestry and phylogenetic instability. PLoS One. 2013;8: e83383.

247. Luo H. Evolutionary origin of a streamlined marine bacterioplankton lineage. ISME J. 2015;9: 1423–1433.

248. Haro-Moreno JM, Rodriguez-Valera F, Rosselli R, Martinez-Hernandez F, Roda-Garcia JJ, Gomez ML, et al. Ecogenomics of the SAR11 clade. Environ Microbiol. 2020;22: 1748–1763.

249. Salcher MM, Pernthaler J, Posch T. Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria “that rule the waves” (LD12). ISME J. 2011;5: 1242–1252.

250. Thrash JC, Temperton B, Swan BK, Landry ZC, Woyke T, DeLong EF, et al. Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. ISME J. 2014;8: 1440–1451.

251. Carlson CA, Morris R, Parsons R, Treusch AH, Giovannoni SJ, Vergin K. Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. ISME J. 2009;3: 283–295.

252. Brown MV, Lauro FM, DeMaere MZ, Muir L, Wilkins D, Thomas T, et al. Global biogeography of SAR11 marine bacteria. Mol Syst Biol. 2012;8: 595.

253. Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, et al. Streamlining and core genome conservation among highly divergent members of the SAR11 clade. MBio. 2012;3. doi:10.1128/mBio.00252-12

254. Rappé MS, Connon SA, Vergin KL, Giovannoni SJ. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature. 2002;418: 630–633.

255. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, et al. Genome streamlining in a cosmopolitan oceanic bacterium. Science. 2005;309: 1242–1245.

256. García-Fernández Jose Manuel, de Marsac Nicole Tandeau, Diez Jesús. Streamlined Regulation and Gene Loss as Adaptive Mechanisms in Prochlorococcus for Optimized Nitrogen Utilization in Oligotrophic Environments. Microbiol Mol Biol Rev. 2004;68: 630–638.

257. López-Pérez Mario, Haro-Moreno Jose M., Iranzo Jaime, Rodriguez-Valera Francisco, Bowman Jeff. Genomes of the "Candidatus Actinomarinales" Order: Highly Streamlined Marine Epipelagic Actinobacteria. mSystems. 5: e01041–20.

258. Jimenez-Infante F, Ngugi DK, Vinu M, Alam I, Kamau AA, Blom J, et al. Comprehensive Genomic Analyses of the OM43 Clade, Including a Novel Species from the Red Sea, Indicate Ecotype Differentiation among Marine Methylotrophs. Appl Environ Microbiol. 2016;82: 1215–1226.

259. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. ISME J. 2014;8: 1553–1565.

260. Carini P, Campbell EO, Morré J, Sañudo-Wilhelmy SA, Thrash JC, Bennett SE, et al. Discovery of a SAR11 growth requirement for thiamin's pyrimidine precursor and its distribution in the Sargasso Sea. ISME J. 2014;8: 1727–1738.

261. Schwalbach MS, Tripp HJ, Steindler L, Smith DP, Giovannoni SJ. The presence of the glycolysis operon in SAR11 genomes is positively correlated with ocean productivity. Environ Microbiol. 2010;12: 490–500.

262. Sun J, Steindler L, Thrash JC, Halsey KH, Smith DP, Carter AE, et al. One carbon metabolism in SAR11 pelagic marine bacteria. PLoS One. 2011;6: e23973.

263. Sun J, Todd JD, Thrash JC, Qian Y, Qian MC, Temperton B, et al. The abundant marine bacterium Pelagibacter simultaneously catabolizes dimethylsulfoniopropionate to the gases dimethyl sulfide and methanethiol. Nat Microbiol. 2016;1: 16065.

264. Steindler L, Schwalbach MS, Smith DP, Chan F, Giovannoni SJ. Energy starved Candidatus Pelagibacter ubique substitutes light-mediated ATP production for endogenous carbon respiration. PLoS One. 2011;6: e19725.

265. Smith DP, Kitner JB, Norbeck AD, Clauss TR, Lipton MS, Schwalbach MS, et al. Transcriptional and translational regulatory responses to iron limitation in the globally distributed marine bacterium Candidatus pelagibacter ubique. PLoS One. 2010;5: e10487.

266. Smith Daniel P., Thrash J. Cameron, Nicora Carrie D., Lipton Mary S., Burnum-Johnson Kristin E., Carini Paul, et al. Proteomic and Transcriptomic Analyses of "Candidatus Pelagibacter ubique" Describe the First PII-Independent Response to Nitrogen Limitation in a Free-Living Alphaproteobacterium. MBio. 4:

e00133–12.

267. Vieira-Silva S, Rocha EPC. The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet. 2010;6: e1000808.

268. Wang L, Chen S, Vergin KL, Giovannoni SJ, Chan SW, DeMott MS, et al. DNA phosphorothioation is widespread and quantized in bacterial genomes. Proc Natl Acad Sci U S A. 2011;108: 2963–2968.

269. Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ. Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. Biol Direct. 2007;2: 27.

270. López-Pérez M, Haro-Moreno JM, Coutinho FH, Martinez-Garcia M, Rodriguez-Valera F. The Evolutionary Success of the Marine Bacterium SAR11 Analyzed through a Metagenomic Perspective. mSystems. 2020;5. doi:10.1128/mSystems.00605-20

271. Zhao Y, Qin F, Zhang R, Giovannoni SJ, Zhang Z, Sun J, et al. Pelagiphages in the Podoviridae family integrate into host genomes. Environ Microbiol. 2019;21: 1989–2001.

272. Zhang Z, Qin F, Chen F, Chu X, Luo H, Zhang R, et al. Culturing novel and abundant pelagiphages in the ocean. Environ Microbiol. 2021;23: 1145–1161.

273. Chen L-X, Zhao Y, McMahon KD, Mori JF, Jessen GL, Nelson TC, et al. Wide Distribution of Phage That Infect Freshwater SAR11 Bacteria. mSystems. 2019;4. doi:10.1128/mSystems.00410-19

274. Buchholz HH, Michelsen ML, Bolaños LM, Browne E, Allen MJ, Temperton B. Efficient dilution-to-extinction isolation of novel virus–host model systems for fastidious heterotrophic bacteria. The ISME Journal. 2021. pp. 1585–1598. doi:10.1038/s41396-020-00872-z

275. Zhao Y, Qin F, Zhang R, Giovannoni SJ, Zhang Z, Sun J, et al. Pelagiphages in the *Podoviridiae* family integrate into host genomes. doi:10.1101/410191

276. Morris RM, Cain KR, Hvorecny KL, Kollman JM. Lysogenic host–virus interactions in SAR11 marine bacteria. Nature Microbiology. 2020. pp. 1011–1015. doi:10.1038/s41564-020-0725-x

277. Eggleston EM, Hewson I. Abundance of Two Pelagibacter ubique Bacteriophage Genotypes along a Latitudinal Transect in the North and South Atlantic Oceans. Frontiers in Microbiology. 2016. doi:10.3389/fmicb.2016.01534

278. Alonso-Sáez L, Morán XAG, Clokie MRJ. Low activity of lytic pelagiphages in coastal marine waters. The ISME Journal. 2018. pp. 2100–2102. doi:10.1038/s41396-018-0185-y

279. Kirzner S, Barak E, Lindell D. Variability in progeny production and virulence of cyanophages determined at the single-cell level. Environ Microbiol Rep. 2016;8: 605–613.

280. Aylward FO, Boeuf D, Mende DR, Wood-Charlson EM, Vislova A, Eppley JM, et al. Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. Proceedings of the National Academy of Sciences. 2017. pp. 11446–11451. doi:10.1073/pnas.1714821114

281. Yoshida T, Nishimura Y, Watai H, Haruki N, Morimoto D, Kaneko H, et al. Locality and diel cycling of viral production revealed by a 24 h time course cross-omics analysis in a coastal region of Japan. The ISME Journal. 2018. pp. 1287–1295. doi:10.1038/s41396-018-0052-x

282. Martinez-Hernandez F, Luo E, Tominaga K, Ogata H, Yoshida T, DeLong EF, et al. Diel cycling of the cosmopolitan abundant Pelagibacter virus 37-F6: one of the most abundant viruses on earth. Environ Microbiol Rep. 2020;12: 214–219.

283. Ottesen EA, Young CR, Gifford SM, Eppley JM, Marin R 3rd, Schuster SC, et al. Ocean microbes. Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. Science. 2014;345: 207–212.

284. Gasol JM, Doval MD, Pinhassi J, Calderón-Paz JI, Guixa-Boixareu N, Vaqué D, et al. Diel variations in bacterial heterotrophic activity and growth in the northwestern Mediterranean Sea. Mar Ecol Prog Ser. 1998;164: 107–124.

285. Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, et al. Marine microbial metagenomes sampled across space and time. Sci Data. 2018;5: 180176.

286. Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, et al. Open science resources for the discovery and analysis of Tara Oceans data. Sci Data. 2015;2: 150023.

287. Haro-Moreno JM, Rodriguez-Valera F, López-Pérez M. Prokaryotic Population Dynamics and Viral Predation in a Marine Succession Experiment Using Metagenomics. Front Microbiol. 2019;10: 2926.

288. Paez-Espino D, Roux S, Chen I-MA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. Nucleic Acids Res. 2019;47: D678–D686.

289. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res. 2012;40: D115–22.

290. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. Nucleic Acids Res. 2015;43: D571–7.

291. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome. 2020;8: 90.

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

293. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–3402.

294. Wang M, Kong L. pblat: a multithread blat algorithm speeding up aligning sequences to genomes. BMC Bioinformatics. 2019;20: 28.

295. Okazaki Y, Nishimura Y, Yoshida T, Ogata H, Nakano S-I. Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. Environ Microbiol. 2019;21: 4740–4754.

296. Tran P, Ramachandran A, Khawasik O, Beisner BE, Rautio M, Huot Y, et al. Microbial life under ice: Metagenome diversity and in situ activity of Verrucomicrobia in seasonally ice-covered Lakes. Environ Microbiol. 2018;20: 2568–2584.

297. Mohiuddin M, Schellhorn HE. Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. Front Microbiol. 2015;6: 960.

298. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119.

299. Chan PP, Lowe TM. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. Methods Mol Biol. 2019;1962: 1–14.

300. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35: 1026–1028.

301. Mirdita M, den Driesch L von, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Research. 2017. pp. D170–D176. doi:10.1093/nar/gkw1081

302. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant

UniProt reference clusters. Bioinformatics. 2007;23: 1282–1288.

303. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature Methods. 2015. pp. 59–60. doi:10.1038/nmeth.3176

304. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD: NCBI's conserved domain database. Nucleic Acids Res. 2015;43: D222–6.

305. Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. 2017;45: D491–D498.

306. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinform. 2021;3: lqab067.

307. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Informatics 2009. PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO.; 2009. doi:10.1142/9781848165632_0019

308. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242.

309. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019;20: 473.

310. Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol. 2007;8: R22.

311. Solovyev VV, Shahmuradov IA, Salamov AA. Identification of Promoter Regions and Regulatory Sites. Methods in Molecular Biology. 2010. pp. 57–83. doi:10.1007/978-1-60761-854-6_5

312. Sullivan MB, Huang KH, Ignacio-Espinoza JC, Berlin AM, Kelly L, Weigele PR, et al. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environ Microbiol. 2010;12: 3035–3056.

313. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13: 2498–2504.

314. Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, et al. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics. 2011;12: 436.

315. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 2013;79: 7696–7701.

316. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nature Communications. 2020. doi:10.1038/s41467-020-15507-2

317. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5: 113.

318. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 2018;27: 135–145.

319. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. Corrigendum to: IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37: 2461.

320. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26: 1641–1650.

321. Benler S, Yutin N, Antipov D, Rayko M, Shmakov S, Gussow AB, et al. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. Microbiome. 2021;9: 78.

322. Fernández-Ruiz I, Coutinho FH, Rodriguez-Valera F. Thousands of Novel Endolysins Discovered in Uncultured Phage Genomes. Front Microbiol. 2018;9: 1033.

323. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46: W95–W101.

324. Oliveira H, Melo LDR, Santos SB, Nóbrega FL, Ferreira EC, Cerca N, et al. Molecular aspects and comparative genomics of bacteriophage endolysins. J Virol. 2013;87: 4558–4570.

325. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16: 276–277.

326. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106: 19126–19131.

327. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. Implementing a Class of Permutation Tests: The**coin**Package. Journal of Statistical Software. 2008. doi:10.18637/jss.v028.i08

328. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21: 487–493.

329. Coutinho FH, Edwards RA, Rodríguez-Valera F. Charting the diversity of uncultured viruses of Archaea and Bacteria. BMC Biol. 2019;17: 109.

330. Díez-Villaseñor C, Rodriguez-Valera F. CRISPR analysis suggests that small circular single-stranded DNA smacoviruses infect Archaea instead of humans. Nat Commun. 2019;10: 294.

331. Gudyś A, Deorowicz S. QuickProbs 2: Towards rapid construction of high-quality alignments of large protein families. Sci Rep. 2017;7: 41553.

332. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017;77: 1–17.

333. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39: 578–585.

334. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25: 1043–1055.

335. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, et al. Host-linked soil viral ecology along a permafrost thaw gradient. Nat Microbiol. 2018;3: 870–880.

336. Luo E, Eppley JM, Romano AE, Mende DR, DeLong EF. Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. ISME J. 2020;14: 1304–1315.

337. Roux S, Enault F, Ravet V, Colombet J, Bettarel Y, Auguet J-C, et al. Analysis of metagenomic data reveals common features of halophilic viral communities across continents. Environ Microbiol. 2016;18: 889–903.

338. Liu R, Qi R, Wang J, Zhang Y, Liu X, Rossetti S, et al. Phage-host associations in a full-scale activated sludge plant during sludge bulking. Appl Microbiol Biotechnol. 2017;101: 6495–6504.

339. Petrov VM, Ratnayaka S, Nolan JM, Miller ES, Karam JD. Genomes of the T4-related bacteriophages as windows on microbial genome evolution. Virol J. 2010;7: 292.

340. Millard AD, Zwirglmaier K, Downey MJ, Mann NH, Scanlan DJ. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of Synechococcus host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution. Environ Microbiol. 2009;11:

2370–2387.

341. López-Pérez M, Rodriguez-Valera F. Pangenome Evolution in the Marine Bacterium Alteromonas. Genome Biol Evol. 2016;8: 1556–1570.

342. Haro-Moreno JM, López-Pérez M, de la Torre JR, Picazo A, Camacho A, Rodriguez-Valera F. Fine metagenomic profile of the Mediterranean stratified and mixed water columns revealed by assembly and recruitment. Microbiome. 2018;6: 128.

343. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science. 2015;348: 1261359.

344. Luef B, Frischkorn KR, Wrighton KC, Holman H-YN, Birarda G, Thomas BC, et al. Diverse uncultivated ultra-small bacterial cells in groundwater. Nat Commun. 2015;6: 6372.

345. Haible D, Kober S, Jeske H. Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. J Virol Methods. 2006;135: 9–16.

346. Martín-Cuadrado A-B, López-García P, Alba J-C, Moreira D, Monticelli L, Strittmatter A, et al. Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. PLoS One. 2007;2: e914.

347. Boyington JC, Gladyshev VN, Khangulov SV, Stadtman TC, Sun PD. Crystal structure of formate dehydrogenase H: catalysis involving Mo, molybdopterin, selenocysteine, and an Fe4S4 cluster. Science. 1997;275: 1305–1308.

348. Henson MW, Lanclos VC, Faircloth BC, Thrash JC. Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. ISME J. 2018;12: 1846–1860.

349. Cabello-Yeves PJ, Rodriguez-Valera F. Marine-freshwater prokaryotic transitions require extensive changes in the predicted proteome. Microbiome. 2019;7: 117.

350. Cliffe LJ, Siegel TN, Marshall M, Cross GAM, Sabatini R. Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. Nucleic Acids Res. 2010;38: 3923–3935.

351. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 2009;324: 930–935.

352. Yap ML, Klose T, Arisaka F, Speir JA, Veesler D, Fokine A, et al. Role of bacteriophage T4 baseplate in regulating assembly and infection. Proc Natl Acad Sci U S A. 2016;113: 2654–2659.

353. Nováček J, Šiborová M, Benešík M, Pantůček R, Doškař J, Plevka P. Structure and genome release of Twort-like Myoviridae phage with a double-layered baseplate. Proc Natl Acad Sci U S A. 2016;113: 9351–9356.

354. Habann M, Leiman PG, Vandersteegen K, Van den Bossche A, Lavigne R, Shneider MM, et al. Listeria phage A511, a model for the contractile tail machineries of SPO1-related bacteriophages. Mol Microbiol. 2014;92: 84–99.

355. Kadirvelraj R, Foley BL, Dyekjaer JD, Woods RJ. Involvement of water in carbohydrate-protein binding: concanavalin A revisited. J Am Chem Soc. 2008;130: 16933–16942.

356. Mühlenkamp M, Oberhettinger P, Leo JC, Linke D, Schütz MS. Yersinia adhesin A (YadA)--beauty & beast. Int J Med Microbiol. 2015;305: 252–258.

357. Smith NL, Taylor EJ, Lindsay A-M, Charnock SJ, Turkenburg JP, Dodson EJ, et al. Structure of a group A streptococcal phage-encoded virulence factor reveals a catalytically active triple-stranded β-helix. Proceedings of the National Academy of Sciences. 2005;102: 17652–17657.

358. Yu Z, An B, Ramshaw JAM, Brodsky B. Bacterial collagen-like proteins that form triple-helical structures. J

Struct Biol. 2014;186: 451–461.

359. Nakagawa H, Arisaka F, Ishii S. Isolation and characterization of the bacteriophage T4 tail-associated lysozyme. J Virol. 1985;54: 460–466.

360. Clokie MRJ, Millard AD, Mann NH. T4 genes in the marine ecosystem: studies of the T4-like cyanophages and their role in marine ecology. Virol J. 2010;7: 291.

361. Hinton DM. Transcriptional control in the prereplicative phase of T4 development. Virol J. 2010;7: 289.

362. Diaconu M, Kothe U, Schlünzen F, Fischer N, Harms JM, Tonevitsky AG, et al. Structural basis for the function of the ribosomal L7/12 stalk in factor binding and GTPase activation. Cell. 2005;121: 991–1004.

363. Soma A, Ikeuchi Y, Kanemasa S, Kobayashi K, Ogasawara N, Ote T, et al. An RNA-modifying enzyme that governs both the codon and amino acid specificities of isoleucine tRNA. Mol Cell. 2003;12: 689–698.

364. Mizuno CM, Guyomar C, Roux S, Lavigne R, Rodriguez-Valera F, Sullivan MB, et al. Numerous cultivated and uncultivated viruses encode ribosomal proteins. Nat Commun. 2019;10: 752.

365. Van Duin J, Wijnands R. The function of ribosomal protein S21 in protein synthesis. Eur J Biochem. 1981;118: 615–619.

366. Breitbart M, Thompson L, Suttle C, Sullivan M. Exploring the vast diversity of marine viruses. Oceanography . 2007;20: 135–139.

367. Gao E-B, Huang Y, Ning D. Metabolic Genes within Cyanophage Genomes: Implications for Diversity and Evolution. Genes . 2016;7. doi:10.3390/genes7100080

368. Hurwitz BL, U'Ren JM. Viral metabolic reprogramming in marine ecosystems. Curr Opin Microbiol. 2016;31: 161–168.

369. Esmon BE, Kensil CR, Cheng CH, Glaser M. Genetic analysis of Escherichia coli mutants defective in adenylate kinase and sn-glycerol 3-phosphate acyltransferase. J Bacteriol. 1980;141: 405–408.

370. Ross P, O'Gara F, Condon S. Cloning and characterization of the thymidylate synthase gene from Lactococcus lactis subsp. lactis. Appl Environ Microbiol. 1990;56: 2156–2163.

371. Moore JT, Silversmith RE, Maley GF, Maley F. T4-phage deoxycytidylate deaminase is a metalloprotein containing two zinc atoms per subunit. J Biol Chem. 1993;268: 2288–2291.

372. Frank JA, Lorimer D, Youle M, Witte P, Craig T, Abendroth J, et al. Structure and function of a cyanophage-encoded peptide deformylase. ISME J. 2013;7: 1150–1160.

373. Santos JM, Freire P, Vicente M, Arraiano CM. The stationary-phase morphogene bolA from Escherichia coli is induced by stress during early stages of growth. Mol Microbiol. 1999;32: 789–798.

374. Mukherjee S, Sengupta S. Riboswitch Scanner: an efficient pHMM-based web-server to detect riboswitches in genomic sequences. Bioinformatics. 2016;32: 776–778.

375. Petit C, Rigg GP, Pazzani C, Smith A, Sieberth V, Stevens M, et al. Region 2 of the Escherichia coli K5 capsule gene cluster encoding proteins for the biosynthesis of the K5 polysaccharide. Mol Microbiol. 1995;17: 611–620.

376. Valvano MA, Marolda CL, Bittner M, Glaskin-Clay M, Simon TL, Klena JD. The rfaE gene from Escherichia coli encodes a bifunctional protein involved in biosynthesis of the lipopolysaccharide core precursor ADP-L-glycero-D-manno-heptose. J Bacteriol. 2000;182: 488–497.

377. Lairson LL, Henrissat B, Davies GJ, Withers SG. Glycosyltransferases: structures, functions, and mechanisms. Annu Rev Biochem. 2008;77: 521–555.

378. Dueholm MS, Albertsen M, Otzen D, Nielsen PH. Curli functional amyloid systems are phylogenetically widespread and display large diversity in operon and protein structure. PLoS One. 2012;7: e51274.

379. Barnhart MM, Chapman MR. Curli biogenesis and function. Annu Rev Microbiol. 2006;60: 131–147.

380. Evans ML, Chapman MR. Curli biogenesis: order out of disorder. Biochim Biophys Acta. 2014;1843: 1551–1558.

381. Nenninger AA, Robinson LS, Hultgren SJ. Localized and efficient curli nucleation requires the chaperone-like amyloid assembly protein CsgF. Proc Natl Acad Sci U S A. 2009;106: 900–905.

382. Robinson LS, Ashman EM, Hultgren SJ, Chapman MR. Secretion of curli fibre subunits is mediated by the outer membrane-localized CsgG protein. Mol Microbiol. 2006;59: 870–881.

383. Hammer ND, Schmidt JC, Chapman MR. The curli nucleator protein, CsgB, contains an amyloidogenic domain that directs CsgA polymerization. Proc Natl Acad Sci U S A. 2007;104: 12494–12499.

384. Van Gerven N, Klein RD, Hultgren SJ, Remaut H. Bacterial amyloid formation: structural insights into curli biogensis. Trends Microbiol. 2015;23: 693–706.

385. Acinas SG, Sánchez P, Salazar G, Cornejo-Castillo FM, Sebastián M, Logares R, et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Commun Biol. 2021;4: 604.

386. Bartlau N, Wichels A, Krohne G, Adriaenssens EM, Heins A, Fuchs BM, et al. Highly diverse flavobacterial phages isolated from North Sea spring blooms. ISME J. 2022;16: 555–568.

387. López-Pérez M, Haro-Moreno JM, de la Torre JR, Rodriguez-Valera F. Novel *Caudovirales* associated with Marine Group I Thaumarchaeota assembled from metagenomes. Environmental Microbiology. 2019. pp. 1980–1988. doi:10.1111/1462-2920.14462

388. Pimentel M. Genetics of Phage Lysis. Molecular Genetics of Mycobacteria. 2015. pp. 121–133. doi:10.1128/9781555818845.ch6

389. Haddad Kashani H, Schmelcher M, Sabzalipoor H, Seyed Hosseini E, Moniri R. Recombinant endolysins as potential therapeutics against antibiotic-resistant staphylococcus aureus: Current status of research and novel delivery strategies. Clin Microbiol Rev. 2018;31. doi:10.1128/CMR.00071-17

390. Ramos-Vivas J, Elexpuru-Zabaleta M, Samano ML, Barrera AP, Forbes-Hernández TY, Giampieri F, et al. Phages and Enzybiotics in Food Biopreservation. Molecules. 2021;26. doi:10.3390/molecules26175138

391. Grütter MG, Weaver LH, Matthews BW. Goose lysozyme structure: an evolutionary link between hen and bacteriophage lysozymes? Nature. 1983;303: 828–831.

392. Xu M, Struck DK, Deaton J, Wang I-N, Young R. A signal-arrest-release sequence mediates export and control of the phage P1 endolysin. Proc Natl Acad Sci U S A. 2004;101: 6415–6420.

393. Bateman A, Murzin AG, Teichmann SA. Structure and distribution of pentapeptide repeats in bacteria. Protein Sci. 1998;7: 1477–1480.

394. Prag S, Adams JC. Molecular phylogeny of the kelch-repeat superfamily reveals an expansion of BTB/kelch proteins in animals. BMC Bioinformatics. 2003;4: 42.

395. Jahn MT, Arkhipova K, Markert SM, Stigloher C, Lachnit T, Pita L, et al. A Phage Protein Aids Bacterial Symbionts in Eukaryote Immune Evasion. Cell Host & Microbe. 2019. pp. 542–550.e5. doi:10.1016/j.chom.2019.08.019

396. Buck M, Gerken T. Two Hands Grip Better Than One for Tight Binding and Specificity: How a Phage Endolysin Fits into the Cell Wall of Its Host. Structure . 2019. pp. 1350–1352.

397. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019;47: 10994–11006.

398. Toussaint A, Rice PA. Transposable phages, DNA reorganization and transfer. Curr Opin Microbiol. 2017;38: 88–94.

399. Sharifi F, Ye Y. MyDGR: a server for identification and characterization of diversity-generating retroelements. Nucleic Acids Res. 2019;47: W289–W294.

400. Groth AC, Calos MP. Phage integrases: biology and applications. J Mol Biol. 2004;335: 667–678.

401. Johnston ER, Hatt JK, He Z, Wu L, Guo X, Luo Y, et al. Responses of tundra soil microbial communities to half a decade of experimental warming at two critical depths. Proc Natl Acad Sci U S A. 2019;116: 15096–15105.

402. Ghai R, Pašić L, Fernández AB, Martin-Cuadrado A-B, Mizuno CM, McMahon KD, et al. New Abundant Microbial Groups in Aquatic Hypersaline Environments. Scientific Reports. 2011. doi:10.1038/srep00135

403. Marine RL, Nasko DJ, Wray J, Polson SW, Wommack KE. Novel chaperonins are prevalent in the virioplankton and demonstrate links to viral biology and ecology. ISME J. 2017;11: 2479–2491.

404. Turner B, Burkhart BW, Weidenbach K, Ross R, Limbach PA, Schmitz RA, et al. Archaeosine Modification of Archaeal tRNA: Role in Structural Stabilization. J Bacteriol. 2020;202. doi:10.1128/JB.00748-19

405. Makarova KS, Koonin EV. Archaeal ubiquitin-like proteins: functional versatility and putative ancestral involvement in tRNA modification revealed by comparative genomic analysis. Archaea. 2010;2010. doi:10.1155/2010/710303

406. Randow F, Lehner PJ. Viral avoidance and exploitation of the ubiquitin system. Nat Cell Biol. 2009;11: 527–534.

407. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, et al. Community genomics among stratified microbial assemblages in the ocean's interior. Science. 2006;311: 496–503.

408. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nature Communications. 2014. doi:10.1038/ncomms5498

409. Duhaime MB, Solonenko N, Roux S, Verberkmoes NC, Wichels A, Sullivan MB. Comparative Omics and Trait Analyses of Marine Pseudoalteromonas Phages Advance the Phage OTU Concept. Front Microbiol. 2017;8: 1241.

410. Pickard D, Toribio AL, Petty NK, van Tonder A, Yu L, Goulding D, et al. A conserved acetyl esterase domain targets diverse bacteriophages to the Vi capsular receptor of Salmonella enterica serovar Typhi. J Bacteriol. 2010;192: 5746–5754.

411. Hardies SC, Hwang YJ, Hwang CY, Jang GI, Cho BC. Morphology, physiological characteristics, and complete sequence of marine bacteriophage φRIO-1 infecting Pseudoalteromonas marina. J Virol. 2013;87: 9189–9198.

412. Taylor JD, Hawthorne WJ, Lo J, Dear A, Jain N, Meisl G, et al. Electrostatically-guided inhibition of Curli amyloid nucleation by the CsgC-like family of chaperones. Sci Rep. 2016;6: 24656.

413. Hammar M, Arnqvist A, Bian Z, Olsén A, Normark S. Expression of two csg operons is required for production of fibronectin- and congo red-binding curli polymers in Escherichia coli K-12. Mol Microbiol. 1995;18: 661–670.

414. Gibson DL, White AP, Rajotte CM, Kay WW. AgfC and AgfE facilitate extracellular thin aggregative fimbriae synthesis in Salmonella enteritidis. Microbiology. 2007;153: 1131–1140.

415. Tetz G, Tetz V. Prion-like Domains in Eukaryotic Viruses. Sci Rep. 2018;8: 8931.

416. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589.

417. Pang T, Savva CG, Fleming KG, Struck DK, Young R. Structure of the lethal phage pinhole. Proc Natl Acad Sci U S A. 2009;106: 18966–18971.

418. Steger LME, Kohlmeyer A, Wadhwani P, Bürck J, Strandberg E, Reichert J, et al. Structural and functional characterization of the pore-forming domain of pinholin S2168. Proc Natl Acad Sci U S A. 2020;117: 29637–29646.

419. Young R. Phage lysis: three steps, three choices, one outcome. J Microbiol. 2014;52: 243–258.

420. Narulita E, Addy HS, Kawasaki T, Fujie M, Yamada T. The involvement of the PilQ secretin of type IV pili in phage infection in Ralstonia solanacearum. Biochem Biophys Res Commun. 2016;469: 868–872.

421. Korotkov KV, Pardon E, Steyaert J, Hol WGJ. Crystal structure of the N-terminal domain of the secretin GspD from ETEC determined with the assistance of a nanobody. Structure. 2009;17: 255–265.

422. Schubeis T, Spehr J, Viereck J, Köpping L, Nagaraj M, Ahmed M, et al. Structural and functional characterization of the Curli adaptor protein CsgF. FEBS Lett. 2018;592: 1020–1029.

423. Conners R, McLaren M, Łapińska U, Sanders K, Stone MRL, Blaskovich MAT, et al. CryoEM structure of the outer membrane secretin channel pIV from the f1 filamentous bacteriophage. Nat Commun. 2021;12: 6316.

424. Morimoto A, Irie K, Murakami K, Masuda Y, Ohigashi H, Nagao M, et al. Analysis of the Secondary Structure of β-Amyloid (Aβ42) Fibrils by Systematic Proline Replacement. Journal of Biological Chemistry. 2004. pp. 52781–52788. doi:10.1074/jbc.m406262200

425. Zhang M, Shi H, Zhang X, Zhang X, Huang Y. Cryo-EM structure of the nonameric CsgG-CsgF complex and its implications for controlling curli biogenesis in Enterobacteriaceae. PLoS Biol. 2020;18: e3000748.

426. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2011;13: 36–46.

427. Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Research. 2021. p. 2138. doi:10.12688/f1000research.21782.4

428. Brown CL, Keenum IM, Dai D, Zhang L, Vikesland PJ, Pruden A. Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes. Sci Rep. 2021;11: 3753.

429. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ. 2017;5: e3817.

430. Young F, Rogers S, Robertson DL. Predicting host taxonomic information from viral genomes: A comparison of feature representations. PLoS Comput Biol. 2020;16: e1007894.

8. Annex 1

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE
Applied and Environmental Science

Metagenome Mining Reveals Hidden Genomic Diversity of Pelagimyophages in Aquatic Environments

Asier Zaragoza-Solas,[a] Francisco Rodriguez-Valera,[a,b] Mario López-Pérez[a]

[a]Evolutionary Genomics Group, División de Microbiología, Universidad Miguel Hernández, Alicante, Spain
[b]Laboratory for Theoretical and Computer Research on Biological Macromolecules and Genomes, Moscow Institute of Physics and Technology, Dolgoprudny, Russia

ABSTRACT The SAR11 clade is one of the most abundant bacterioplankton groups in surface waters of most of the oceans and lakes. However, only 15 SAR11 phages have been isolated thus far, and only one of them belongs to the *Myoviridae* family (pelagimyophages). Here, we have analyzed 26 sequences of myophages that putatively infect the SAR11 clade. They have been retrieved by mining ca. 45 Gbp aquatic assembled cellular metagenomes and viromes. Most of the myophages were obtained from the cellular fraction (0.2 μm), indicating a bias against this type of virus in viromes. We have found the first myophages that putatively infect *Candidatus* Fonsibacter (freshwater SAR11) and another group putatively infecting bathypelagic SAR11 phylogroup Ic. The genomes have similar sizes and maintain overall synteny in spite of low average nucleotide identity values, revealing high similarity to marine cyanomyophages. Pelagimyophages recruited metagenomic reads widely from several locations but always much more from cellular metagenomes than from viromes, opposite to what happens with pelagipodophages. Comparing the genomes resulted in the identification of a hypervariable island that is related to host recognition. Interestingly, some genes in these islands could be related to host cell wall synthesis and coinfection avoidance. A cluster of curli-related proteins was widespread among the genomes, although its function is unclear.

IMPORTANCE SAR11 clade members are among the most abundant bacteria on Earth. Their study is complicated by their great diversity and difficulties in being grown and manipulated in the laboratory. On the other hand, and due to their extraordinary abundance, metagenomic data sets provide enormous richness of information about these microbes. Given the major role played by phages in the lifestyle and evolution of prokaryotic cells, the contribution of several new bacteriophage genomes preying on this clade opens windows into the infection strategies and life cycle of its viruses. Such strategies could provide models of attack of large-genome phages preying on streamlined aquatic microbes.

KEYWORDS Fonsibacter, pelagiphages, SAR11, genome-resolved metagenomics, myophages

In marine ecosystems, bacteriophages (viruses that infect bacterial cells) are extremely abundant, with an estimated $>10^{10}$ viral particles per liter of seawater (1, 2). Their lytic lifestyle is responsible for the mortality of nearly 10% to 50% of the microbial population per day (3). Therefore, it should not come as a surprise that bacteriophages are important players in the functioning of the marine microbial ecosystem. For example, they affect nutrient cycling through the "viral shunt" (4), influence microbial community composition and diversity (5), and drive host evolution, both by favoring genetic exchange and by predation pressure. The latter is of special importance as it favors high diversity at the population level, especially at loci that code for phage resistance traits (6, 7).

Citation Zaragoza-Solas A, Rodriguez-Valera F, López-Pérez M. 2020. Metagenome mining reveals hidden genomic diversity of pelagimyophages in aquatic environments. mSystems 5:e00905-19. https://doi.org/10.1128/mSystems.00905-19.

Editor Jillian Petersen, University of Vienna

Address correspondence to Francisco Rodriguez-Valera, frvalera@umh.es, or Mario López-Pérez, mario.lopezp@umh.es.

Received 22 December 2019
Accepted 27 January 2020
Published 18 February 2020

The SAR11 clade (including the order *Pelagibacterales*) is one of the most abundant bacteria in marine ecosystems, constituting approximately 20% to 40% of all planktonic cells in the oceanic photic zone (8). A particular subclade within SAR11 (LD12) is also important in freshwaters, lakes, and rivers, although less prevalent (9). Recently, a representative of this freshwater subgroup was isolated in pure culture and named *Candidatus* Fonsibacter ubiquis (9). Considering the facts described above, we would expect that members of this clade are prime targets for phage predation. To date, only 15 SAR11 phages have been isolated, all belonging to the order *Caudovirales* (10, 11). This order of viruses is the most prevalent in aquatic environments and can be divided into the families *Myoviridae*, *Siphoviridae*, and *Podoviridae* on the basis of their morphological characteristics (12). SAR11 phages belonging to the *Podoviridae* family are found more often both in pure culture (10, 11) and metagenomic collections (13–15) compared to the other two families. Most of these phages belong to the subfamily *Autographivirinae*, and it has been suggested that many are temperate phages that use tRNA genes as integration sites (11). Only one of the isolated SAR11 phages belongs to the *Myoviridae* family, and despite the abundance of cultivation-independent metagenomic sequencing techniques, only four more myophage genomes have been found in the form of metagenome assembled viral genomes (MAVGs) (14). This scarcity of pelagimyophage (PMP) genomes is surprising, since several metagenomic studies from aquatic environments have shown that T4-like phages constitute the dominant fraction of the viral community (16–19).

The PMP genomes discovered thus far are all part of the *Tevenvirinae* subfamily. This subfamily of double-stranded DNA, contractile-tailed phages owe their name to their remarkable gene homology and genomic synteny to the well-studied *Escherichia coli*-infecting T-even phages, which are represented by T4 (20). Members of this subfamily have been isolated from a variety of hosts (21–24) and can be clustered into three phylogenetic groups based on the genetic divergence of the major capsid protein: Far T4, Near T4, and Cyano T4 (25). PMP HTVC008M is included within the Cyano T4 group (10), together with viral isolates of *Sinorhizobium meliloti* (23), *Stenotrophomonas maltophilia* (26), and the marine cyanobacteria *Synechococcus* and *Prochlorococcus* spp. (24). The latter group is known as the cyanomyophages (CMPs) and is the clade most closely related to HTVC008M. CMPs are generalist phages, successfully infecting hosts from different cyanobacterial species (27), and even genera (28). All CMPs share a set of core genes related to virion structure, DNA replication, and auxiliary metabolic genes (AMGs) (24, 29, 30), which are involved in supplementing host metabolism during infection (31).

Given their large genomes and complex morphology, myoviruses can provide rich information about their hosts and life cycle. In this study, we analyzed 26 new sequences of myophages that putatively infect the SAR11 clade retrieved by mining aquatic metagenomes. This alternative approach to culture-dependent methods has succeeded in discovering new viruses from uncultured microbes earlier (32, 33). Together, these findings increased sixfold the SAR11 myophage repertoire and allowed us to discover different PMP clades, including the first myophage specific of the freshwater genus *Ca*. Fonsibacter and the bathypelagic SAR11 phylogroup Ic (9, 34). This recovery effort has increased their genome diversity enough to be able to perform genomic comparisons with the closest well-studied CMPs to elucidate peculiarities of the PMP infection model.

RESULTS

Figure S1A in the supplemental material shows the workflow that we used to recover sequences of myophages that putatively infect the SAR11 clade from several cellular metagenomic and viromic samples (see Table S1 in the supplemental material). In the end, we were able to recover 26 new PMP MAVGs that, together with the reference sequences, add up to 31 genomes (Table 1). Interestingly, 25 of the 26 new sequences have been recovered from the cellular fraction and not from the viral fraction, which could explain their poor representation in databases.

TABLE 1 Genomic features for the pelagimyophages analyzed in this study

PMP	Group	Mean igs (bp)[a]	Length (bp)	GC content (%)	No. of tRNAs	No. of genes	Completeness[b]	No. of matches[c] to:		Habitat[d]	Sample type[e]	Reference(s)
								SAR11	PMP core			
HTVC008M	A	23.87	147,284	33.45	0	199	Yes (Cu)	9	23	M	C	10
Io7-C40	A	21.35	103,430	33.11	2	117	No	11	17	M	MG	13
MAVG02	A	25.5	157,661	33.98	0	216	Yes (Al)	10	20	M	MG	14
MAVG05	A	21.49	164,624	32.74	2	228	Yes (Al)	15	37	M	MG	14
PMP-MAVG-4	A	21.59	179,730	32.04	0	242	Yes (Al)	21	24	M	MG	93
PMP-MAVG-12	A	15.54	104,791	33.36	0	131	No	5	20	M	MG	92
PMP-MAVG-18	A	23.35	153,977	32.58	1	197	No	17	25	M	MG	93
PMP-MAVG-21	A	24.53	135,163	31.59	0	195	No	11	24	M	MG	93
PMP-MAVG-25	A	25.56	142,712	31.7	0	204	Yes (Al)	19	24	M	MG	93
PMP-MAVG-8	A	14.28	118,694	31.91	0	159	No	13	14	M	MG	91
PMP-MAVG-2	B	15.66	139,426	32.4	0	189	No	7	28	M	MG	92
PMP-MAVG-3	B	16.98	147,773	32.66	0	200	Yes (Al)	8	23	M	MG	14, 92
PMP-MAVG-14	B	18.64	136,460	32.92	2	186	No	11	27	M	V	91
PMP-MAVG-16	B	28.57	132,453	32.99	3	179	Yes (TR)	5	25	M	MG	93
PMP-MAVG-19	B	24.69	149,077	34.83	2	199	Yes (TR)	9	18	M	MG	93
PMP-MAVG-26	B	25.6	142,788	32.48	0	193	No	7	29	M	MG	91
PMP-MAVG-1	C	26.18	118,124	33.71	1	154	No	4	11	M	MG	41
MAVG04	C	26.64	159,588	34.12	2	211	Yes (Al)	5	12	M	MG	14
PMP-MAVG-9	C	21.81	124,621	33.95	1	165	No	6	10	M	MG	41
PMP-MAVG-10	C	13	127,706	32.6	0	177	No	8	15	M	V	91
PMP-MAVG-17	C	21.52	149,073	34.51	3	200	No	5	13	M	MG	93
PMP-MAVG-22	C	15.6	103,989	34.17	0	129	No	2	10	M	MG	93
PMP-MAVG-24	C	21.72	116,502	34.74	1	162	No	1	11	M	MG	93
PMP-MAVG-15	D	21.52	144,833	31.3	3	193	Yes (TR)	6	6	F	V	93
PMP-MAVG-20	D	21.3	122,912	31.08	3	174	No	8	6	F	V	93
PMP-MAVG-5	E	26.22	149,934	33.6	3	190	Yes (TR)	4	10	M	MG	41
PMP-MAVG-6	E	27.22	135,833	33.58	1	176	No	4	17	M	MG	41
PMP-MAVG-7	E	32.87	135,598	33.82	2	171	No	2	14	M	MG	41
PMP-MAVG-11	E	27.05	141,312	34.54	1	177	Yes (Al)	5	16	M	MG	41
PMP-MAVG-13	E	24.74	155,847	34.2	0	208	Yes (Al)	3	16	M	V	91
PMP-MAVG-23	E	19.87	110,977	34.96	2	146	No	4	10	M	MG	93

[a]Igs, intergenic spacer.
[b]How completeness was found is shown is parentheses: Cu, cultivated; Al, alignment; TR, terminal repeats.
[c]Protein matches, based on BLASTN hits with at least 70% similarity and an alignment length between 70% and 130% of the length of the smaller protein.
[d]M, marine; F, freshwater.
[e]C, culture; MG, metagenome; V, virome.

Genomic features. MAVG completeness was verified either by the presence of identical repeated sequences (>10 nucleotides [nt]) at the 5′- and 3′-terminal regions or by showing a similar synteny and gene content to the cultivated PMP HTVC008M (10). The genome size of the 13 complete genomes ranges from 132 to 164 kb (Table 1). To study the relationships of the recovered phages, the 31 PMP genomes were compared in a phylogenomic tree using four CMP genomes as an outgroup. The five proteins common to all 35 genomes (large and small subunits of terminase, VrlC protein, tail tube monomer gp18, and baseplate wedge protein gp8) were merged into a concatemer. The phylogenomic tree clustered PMPs into five different groups (PMP-A to PMP-E), with group PMP-A containing the reference phage HTVC008M (Fig. 1). Host assignment within different SAR11 subclades was not possible (except for group D [see below]) due to (i) lack of tRNA genes (only 18 genomes had them, and the ones present were all under 95% identity to SAR11 known tRNAs), which suggests that either we do not have genome representatives for the hosts they infect, or they have a broad host range, (ii) similarity of shared proteins provided inconclusive results (same identity to distantly related host-groups) and (iii) there is only one report of a CRISPR-cas system in SAR11, which is found only in the bathypelagic ecotype Ic (34). The enormous diversity of the SAR11 clade probably complicates the process of host assignment.

Figure 2A shows the alignment of two genomes of group PMP-A (one of them the pure culture HTVC008M), while alignments of one representative genome from each

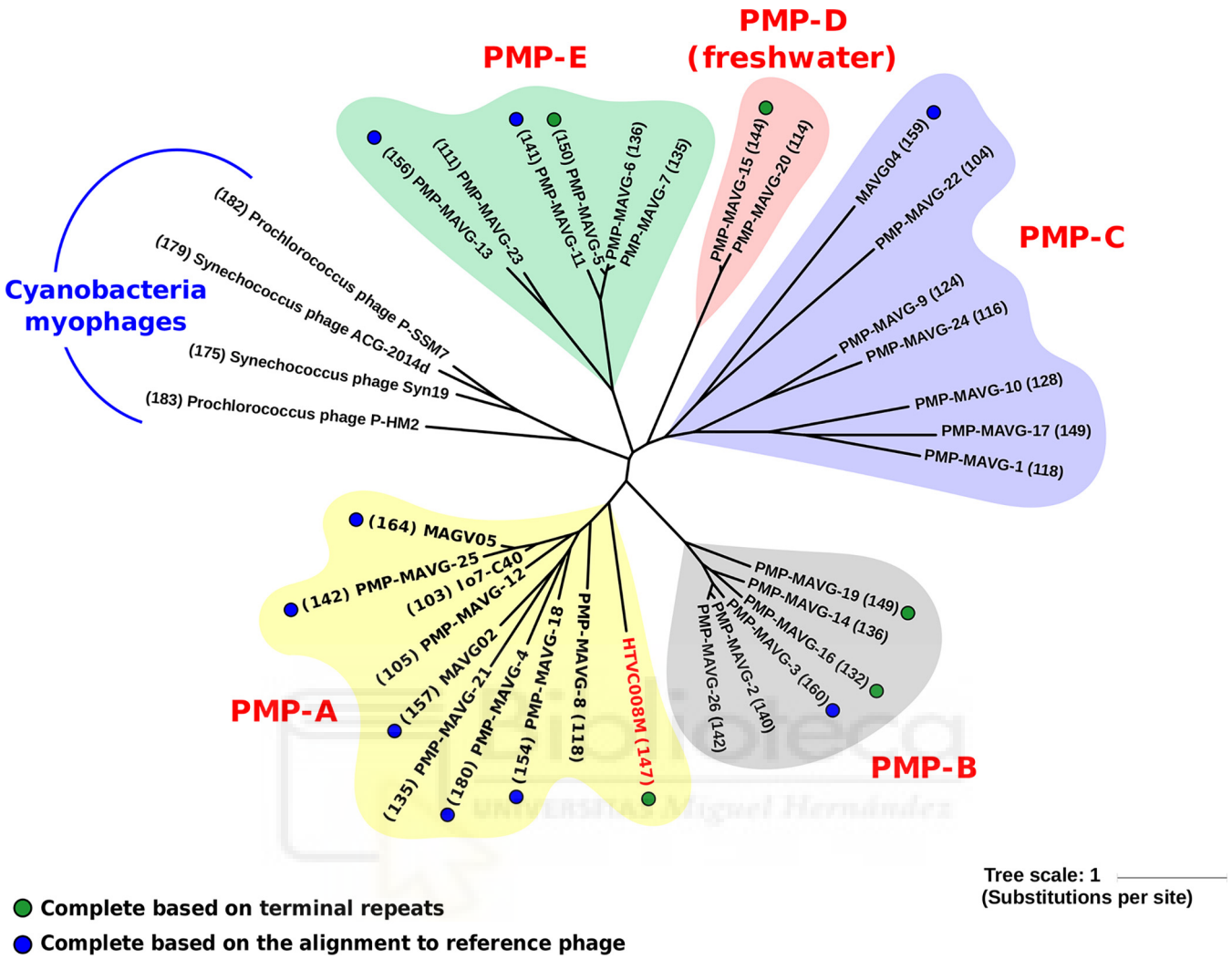


FIG 1 Unrooted phylogenomic tree of concatenated conserved proteins (terminase small subunit, terminase large subunit, tail tube monomer, tail tube monomer, baseplate wedge protein gp8, and VrlC protein) found in pelagimyophages (PMPs) and in the cyanomyophage outgroup. The reference cultured PMP is highlighted in red. The size (in kilobases) of each MAVG is shown in parentheses next to each branch, with complete PMP MAVGs marked with solid circles.

cluster are shown in Fig. 2B. Overall, synteny was well preserved in all sequences once they were rearranged to start from the major capsid gene (*gp23*), and all of the sequences displayed the characteristic patchwork architecture of the *Tevenvirinae* subfamily, with remarkably conserved core modules (DNA replication and virion structure) separated by variable regions, designated as hypervariable (21, 35) (Fig. 2A and B). The most remarkable feature is the presence of a large nonsyntenic island located in the middle of the structural region, always between the VrlC gene and the neck protein gene *gp14* (Fig. 2C). On the basis of its variable character and the presence of tail fibers, we have designated this variable region the host recognition cluster (HRC) (Fig. 2C). In other T4-like phages, this region contains only the tail fiber module (30, 35). This large hypervariable region has been already described in CMPs, usually containing several structural genes and AMGs (30). In PMPs, this region is larger (mean HRC size of 44.6 kb versus 34.2 kb in CMPs), and contains, along with the expected tail fiber genes, a large number of genes seemingly unrelated to the tail fiber module, the most conspicuous of which are several glycosyltransferases, typically involved in the synthesis of the O chain of the lipopolysaccharide that is located in the outer layer of the Gram-negative cell envelope (24, 36) (Fig. 2C). In PMPs, 63 out of the 162 lipopolysaccharide (LPS)-related proteins found are inside the HRC, while CMP HRCs have more identifiable tail

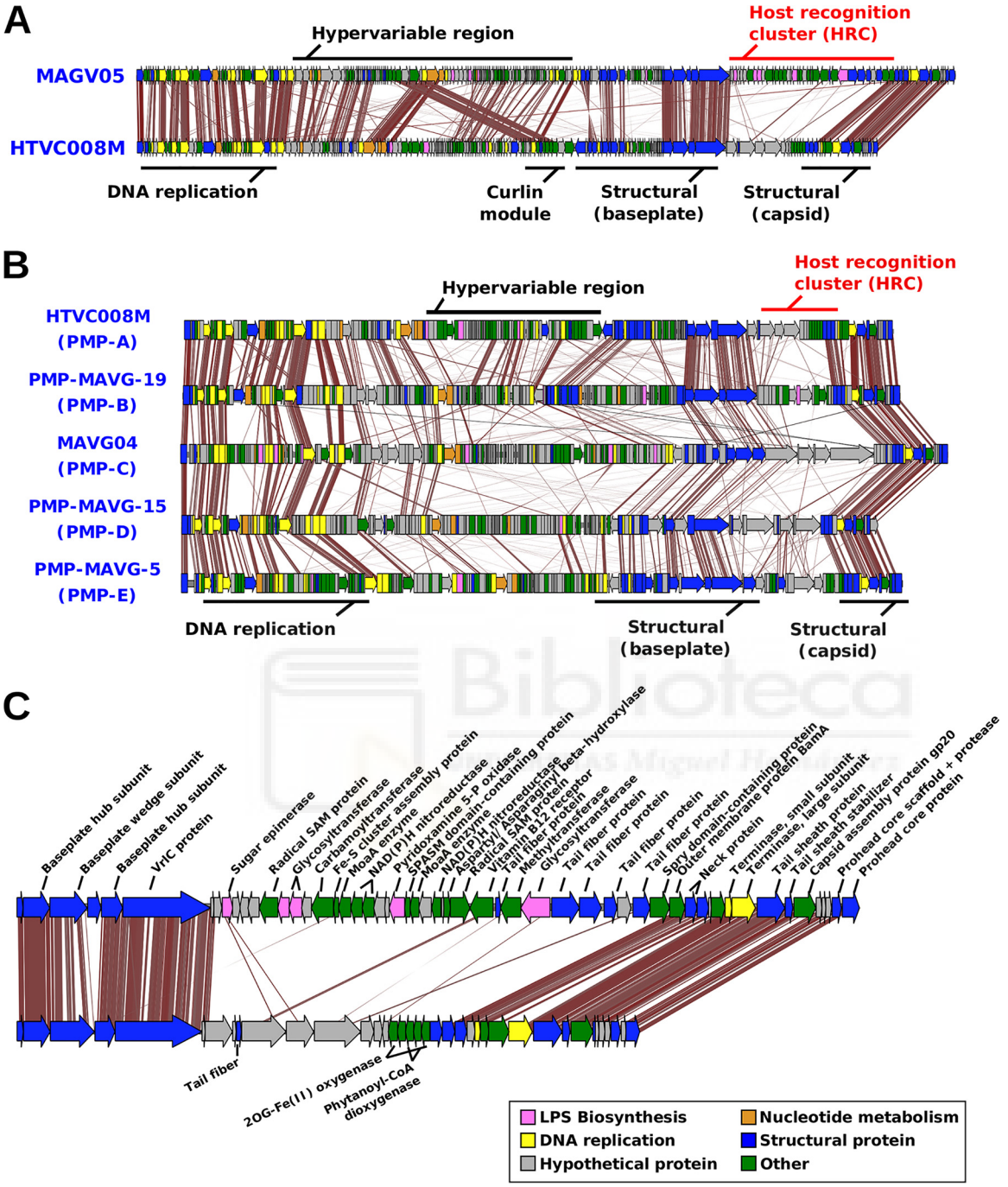


FIG 2 Alignment of pelagimyophage genomes (tblastx, 30% identity). (A) Whole-genome alignment of two PMP-A group genomes. The different modules and hypervariable regions are labeled with black lines over the genomes, while the host recognition module (HRC) is highlighted in red. (B) Whole-genome alignment of a complete representative of each PMP group. (C) Close-up view of the HRC. Genes are colored according to their predicted function.

fiber-related proteins. However, the latter could be attributed to the fact that CMPs are better represented in the sequence databases and are thus easier to annotate. The comparison of the CMP and PMP genomes showed strong conservation of all modules, including the HRC (Fig. 3A). However, unlike the latter, in some CMP genomes, the baseplate module is divided by another plastic region (Fig. 3A).

The two most similar complete genomes were MAGV3 and MAGV16, found in cluster B (average nucleotide identity [ANI] of 72.0% and coverage of 38.6%), although

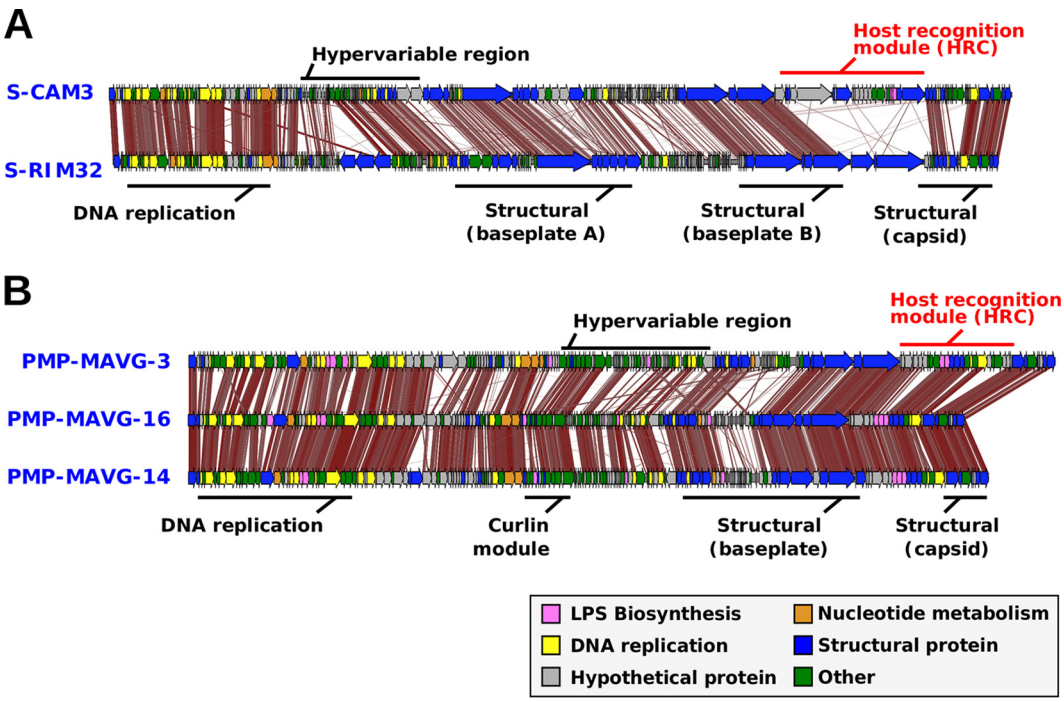


FIG 3 Alignment of pelagimyophage and cyanomyophage (CMP) phages (tblastx, 30% identity). Gene modules are labeled with black lines over the genomes, with the host recognition cluster highlighted in red. (A) Alignment of two CMPs. (C) Alignment of three PMPs from group PMP-B with a similar HRC.

they were assembled from the Western Arctic ocean and the Mediterranean Sea, respectively (Fig. 3B). In the case of these two, the HRC was much more similar and differed only by the addition of some gene cassettes related to radical SAM (*S*-adenosyl-L-methionine) proteins (Fig. 3B). Their comparison seems to indicate that the divergence of this region is a gradual process rather than a complete replacement, as described for replacement flexible genomic islands in prokaryotic cells (37). The genes located downstream from VrlC, which are the tail fibers in most genomes, show high similarity, indicating a possible host overlap of these two phages.

Recruitment from cellular metagenomes and viromes. To evaluate the abundance and elucidate possible patterns of distribution of these phages, we performed recruitment analysis by comparing each sequence to 395 metagenomes from Mediterranean depth profile (38, 39), *Tara* Oceans (40) and Geotraces (41) data sets as well as several freshwater metagenomes (see Materials and Methods). We considered only those samples where at least one PMP recruited more than five reads per kilobase of genome and gigabase of metagenome (RPKG) with an identity of >95%. PMP genomes showed a wide, if uneven, oceanic distribution along the *Tara* Oceans transect (40) (Table S2). All genomes except the freshwater PMP-D group (see below) recruited significantly in several marine samples from different geographic regions, with maximum recruitment typically found in the 5-to-45-m-depth range. Figure 4A shows the recruitment of both families of SAR11 phages (*Podoviridae* and *Myoviridae*) and their host in both the cellular and viral fractions from *Tara* Oceans. In addition, we have also included the other most relevant and widespread marine group, *Cyanobacteria*, and their myophages. While the presence of podophages was mainly restricted to viromes, both groups of myophages were present in both fractions (cellular and viral) (Fig. 4A), although pelagimyophage genomes recruited significantly more from cellular metagenomes than from viromes. The abundance of viral DNA in the cellular fraction indicates that a high number of microbial cells are undergoing the lytic cycle, which acts as a natural amplification of viral DNA (13, 14). Another interesting observation was that a significant amount of SAR11 DNA was present in viromes, probably because

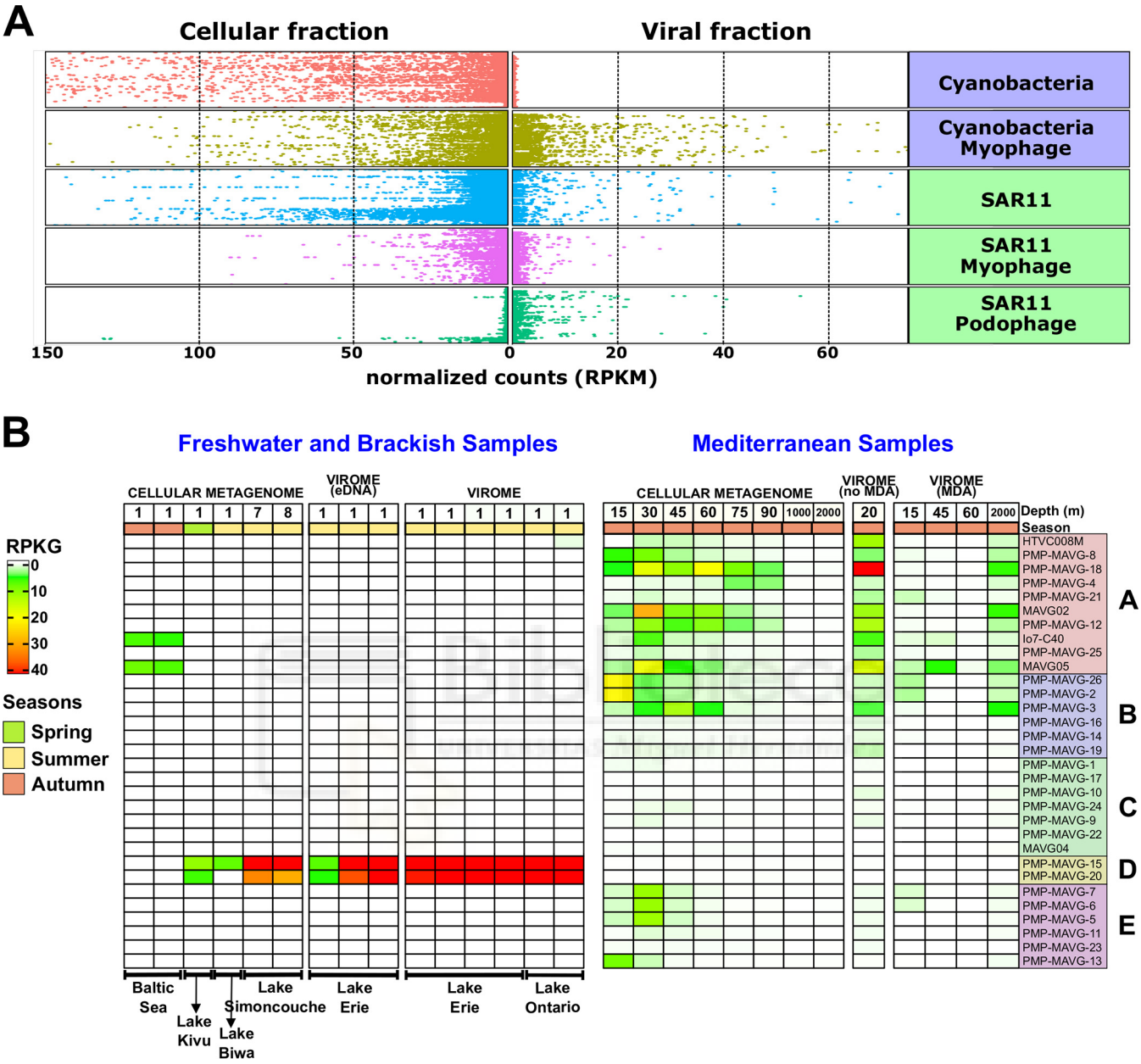


FIG 4 Recruitment of pelagimyophages. (A) Relative abundance of PMPs reads in Mediterranean, Geotraces, and *Tara* Oceans metagenomes and viromes is shown along with the abundances of SAR11 bacteria, SAR11 podoviruses, and *Cyanobacteria* and their myophages. The horizontal axis shows the normalized count of reads per kilobase pair of genome and megabase pair of metagenome (RPKM), while the vertical axis shows the sampling stations (like in reference 33). (B) Heatmap of abundance of PMPs in freshwater and Mediterranean cellular metagenomes and viromes. Normalization of the abundance was performed by calculating RPKG (reads recruited per kilobase of the genome per gigabase of the metagenome).

some SAR11 cells might be small enough to pass through the 0.2-μm filter used frequently to retain bacteria (Fig. 4A) (8, 42). A latitude transect from 50°N to 50°S in the West Atlantic Ocean was analyzed using the Geotraces database (41). However, latitude did not seem to be a significant factor in their distribution (Table S3).

The recruitment results as a whole suggest that PMP amplification is biased, as this group of genomes always recruited much more from cellular metagenomes than from viromes. The nature of this bias (either biological or technical) is still unclear. We also observed significant differences in recruitment values between the Mediterranean viromes treated with multiple displacement amplification (MDA) and those that had not been amplified (Fig. 4B). Although there is no direct evidence of their effect over

myoviruses, MDA amplification might have played a part in these differential recruitment. MDA has been reported to be biased toward certain nucleic acid structures and sequences (43, 44).

However, we were able to distinguish some groups with different patterns of recruitment. One genome of group PMP-A (PMP-MAVG-4) predominantly recruits below 200 m in both the Geotraces and TARA data sets, supporting its association to bathypelagic *Pelagibacterales* clade Ic (34) (Fig. S2 and Tables S2 and S3), although the assignment is tentative, since it could not be proven by sequence analysis. Due to the scarcity of samples from the deep ocean, we can confirm its presence only in temperate zones of the Pacific and Atlantic Oceans (Tables S2 and S3). In Mediterranean samples, it appears only in areas below the deep chlorophyll maximum (75 to 90 m) but not at bathypelagic depths, probably due to the Mediterranean relatively warm water column, although Ic representatives have been detected there (Fig. 4B) (45). Unique genes to this putatively "deep ecotype" include a GMP reductase and various genes involved in heme biosynthesis (coprophyrinogen oxidase, porphobilinogen deaminase) as well as a formate dehydrogenase, an enzyme that transforms formate into CO_2 and $2H^+$ (46). This could be an adaptation to generate a proton gradient in the absence of light, as SAR11 cells can generate it via rhodopsins. Two other PMP-A representatives, MAGV05 and Io7-C40, showed tolerance for brackish waters, as demonstrated by their recruitment from Baltic Sea cellular metagenomes (Fig. 4B). Group D recruits only from freshwater samples, making them the first described freshwater myophages of the SAR11 clade (see below) (Fig. 4B). Linear recruitments (Fig. S3A) showed that although genomes recruit along their entire lengths, most of the reads were recruited at more than 99% identity. The genome regions that recruit vertically down to 80% identity correspond to the structural and DNA replication-related genome regions described previously, which are very well conserved among all the members of the subfamily (24, 35). The HRC usually underrecruited, indicating the highly variable nature of this region (Fig. S3A). The same pattern was observed in cellular metagenomes and viromes with and without MDA (Fig. S3A).

First genomes of PMPs infecting *Ca.* Fonsibacter. Genomic analysis of the two genomes in group PMP-D showed that both contained tRNA genes with the best match to tRNAs from the recently isolated *Candidatus* Fonsibacter ubiquis LSUCC0530, a member of the LD12 subclade (9). Metagenomic recruitment showed clear evidence that group PMP-D was associated with freshwater samples (Fig. 4B). To our knowledge, these are the first genomes of myophages that putatively infect *Ca*. Fonsibacter (fonsimyophages). Both are remarkably similar to each other but present different degrees of completeness. PMP-MAVG-15 is considered complete, while PMP-MAVG-20 is lacking the DNA replication module. Recently, a shift toward basic values was described in the relative frequency of predicted isoelectric points when comparing freshwater and marine microbes (47). Along these lines, we found a significant difference in PMPs infecting *Ca*. Fonsibacter compared to the reference genome HTVC008M (Fig. S3B). However, synteny was well preserved between marine and freshwater groups (Fig. S3C).

Recruitments show the recovered fonsimyophages to be present in various lakes from Canada (Erie, Ontario, Simoncouche) in both the cellular and viral fraction (Fig. 4B). We also found recruitment matches at lower identity (<80%) in other freshwater samples (Lake Biwa, Lake Kivu). Linear recruitments for group D phages against freshwater viromes are different from those originating from their marine counterparts (Fig. S3), showing that diversity in fonsimyophages is lower than that of the marine PMPs. This fact might reflect the reduced intrapopulation diversity of their host compared to other SAR11 subclades (9).

Gene content comparisons between marine or freshwater SAR11 PMPs shed little light on possible adaptations to the latter. However, the freshwater genomes do not contain genes related to LPS, substrate transport, radical SAM proteins, or the curli operon (see below). Nevertheless, it has some unique genes, such as *speH* (involved in

polyamine salvaging), various genes involved in lipid biosynthesis (*fabF*, stearoyl-coenzyme A [CoA] desaturase) and a 2OGFeDO superfamily protein, which catalyzes nucleic acid modifications (48, 49). Strikingly, some proteins core to all PMPs (peptide deformylase, ribosomal protein S21, and aspartyl/asparaginyl beta-hydroxylase) are present in group PMP-D but are different enough to be separated in independent protein clusters.

Comparative genomics. To maximize our ability to annotate phage proteins, we clustered orthologous genes into protein clusters (PCs) and annotated their function following a consensus-based approach (see Materials and methods). The PCs with the most differences in abundance between PMPs and CMPs have been collected in Table S4. Furthermore, to examine the organization of the PCs into operons in both groups of phages, we built a cooccurrence matrix (Fig. S4A), which links genes if they are in the same operon. Previously described methods to detect middle and late promoters in CMPs (24) did not provide satisfactory results when applied to PMPs, so we delimited operons by terminators and strand changes (see Materials and Methods). The cooccurrence matrix reveals differences in the structural organization of the operons containing conserved PCs. While structural operons contain only structural or hypothetical proteins, operons containing DNA metabolism genes are more diverse, containing AMGs of various types. Furthermore, genes involved in the same function are not in the same operon unless they are subunits of the same protein or the presence of one is meaningless without the other. An example of this phenomenon would be the photosynthesis AMGs in CMPs. Photosystem II D1 and D2 subunits are always in the same cluster, but the reaction center protein PsbN is not.

(i) Structural genes. Structural modules are well conserved among both groups of phages, as we identified homologs for the majority of typically conserved structural capsid and tail proteins. Despite the structural conservation of core components in all *Tevenvirinae* phages, we were unable to identify some conserved but highly divergent proteins, like the tape measure or tail fiber proteins. The structural region with the most differences compared to the T4 phage was the baseplate. Both groups contain homologs for a large number of the genes involved in the internal structure of the baseplate of T4-type phages (50), which is involved in baseplate assembly, initiation, and sheath contraction (51). A remarkable difference is the absence of T4 Gp7, which appears to be substituted in both groups of phages by the VrlC protein. VrlC is particularly meaningful, as it is considered an integral component of the two-layered baseplate structure (52, 53), so we can predict that both groups possess this type of baseplate. The other regions of the baseplate appear to be less conserved. Within this large structural operon, we also found various unidentified structural proteins that contain domains linked to carbohydrate-binding and host recognition (specifically, YHYH domains, concanavalin A domains, triple collagen repeats, major tropism determinant domains, and YadA domains) (54–58). These putative receptor-binding proteins could be part of the tail fiber complex or the baseplate, as double-layered baseplates have been reported to contain these kind of proteins (52). Last, the *gp5* gene shows a much larger divergence than the VrlC protein, with both groups of phages coding for various gp5 PCs. As gp5 is involved in cell puncturing and local cell wall degradation (59), we can assume that the differences in gp5 PCs are an adaptation to the specific cell wall of the host.

(ii) DNA transcription and translation. Transcription regulation in PMPs seems to be quite similar to that of CMPs, with both groups lacking homologs to the T4 genes involved in regulating early and middle transcription (*alt*, *modA*, *modB*, *asi*, and *motA*) (60, 61). Some genomes of group PMP-A code for an homolog of the L12 ribosomal protein, which is the binding site for several factors involved in protein synthesis (62), and a tRNA(Ile)-lysidine synthetase, which is an uncommon nucleoside usually seen only in tRNA and involved in solving differences between the elongation methionine tRNA and isoleucine tRNA (63). The most significant difference between both groups of phages related to the translation process is that the latter group codes for a homolog

of the 30S ribosomal protein S21. This protein is responsible for the recognition of complex mRNA templates during translation and has been described only as an AMG in HTVC008M (64, 65). S21 is not part of any specific gene cluster, which, assuming the protein follows the same rules as the other AMGs, suggests that no other viral factors are required for its functionality.

Auxiliary metabolic genes. CMPs frequently contain AMGs, homologs of host genes, to modify host metabolism during infection (66). We have analyzed the occurrence of this type of genes in the PMP genomes and compared it with the occurrence in CMPs (Table S5), which have been widely studied (67).

Both groups of phages had the three classic AMGs involved in nucleotide biosynthesis (*cobS*, *cobT*, both subunits of ribonucleotide reductase) (66, 68) (Table S5). However, Both PMP-A and PMP-B groups code for the adenylate kinase *adk*, which is involved in the interconversion between adenine nucleotides (69), while group C has two different thymidylate synthases and a deoxycytidylate CMP deaminase, which provides the substrate for both (70, 71) (Table S4). A peptide deformylase involved in protein maturation was present in all PMPs in the core genome, inside a DNA metabolism operon, while in their cyanobacterial counterparts, it was found only in a few and inside the flexible genome, together with the photosystem AMGs (72).

We found fewer genes dedicated to regulation in PMPs than in CMPs. Typical CMP regulation AMGs such as *mazG* are absent in PMPs, and regulation genes shared by both groups such as the Pho regulon *PhoH* or Sm/Lsm RNA-binding proteins are more abundant in CMPs than in PMPs (Table S5). However, genes related to the *sprT* family (a gene involved in the regulation of the stress factor BolA) are much more prevalent in PMPs than in CMPs. *bolA* has many effects on cell morphology, cell growth, cell division, and biofilm development in the stationary phase and under starvation conditions (73). These differences in regulatory proteins are not surprising, since it has been proposed that SAR11 cells are not as tightly regulated as cyanobacteria (8); hence, their regulatory systems would be significantly different (as mentioned above, the starvation system *mazE*/*mazG* does not exist in SAR11 but it is present in picocyanobacteria) (8). Regulation in SAR11 seems to be less dependent on proteins, being directed by riboswitches and other small mRNA (smRNA) molecules instead (8). However, a search of these regulatory mRNAs with the tool Riboswitch Scanner (74) found no evidence of their presence in neither group of phages.

Another type of AMG found in PMP genomes are genes related to the production of the O-chain of bacterial lipopolysaccharides, usually found as part of the HRC, but also distributed along the genome in clusters of two or three genes. This category of genes is also found in CMPs but is much less abundant. The LPS-related genes are either enzymes involved in the synthesis of deoxy-sugars to use as building blocks (*rfaE*, UDP-glucose 6-dehydrogenase) (75, 76) or are glycosyltransferases, involved in adding specific sugar residues to a molecule (77). Glycosyltransferases in bacteriophages are involved in the glycosylation of viral DNA to protect against the host restriction-modification systems or in the modification of the O-antigen chain of the host to protect against coinfection by other phages (36). Considering that the glycosyltransferase family most represented in PMPs is GT8, which is mainly involved in LPS biosynthesis (78), and that only one SAR11 genome out of more than 100 sequenced thus far codes for a restriction-modification system (79), it seems likely that glycosyltransferases in this group are involved in the modification of the O-chain of their host.

Curli operon. Between the DNA replication and structural modules, there is a hypervariable region containing a variable number of genes with little synteny among the different PMP representatives (Fig. 2A and Fig. S2A). Within this variable region, we found two homologs of the type VIII secretion system (TSS VIII) present in all PMP groups but the fonsimyophages (Fig. 2). To our knowledge, this is the first report of phages that code for proteins of this secretion system. The cooccurrence network shows that these proteins are part of a well-defined operon that includes the proteins CsgF, CsgG, two hypothetical proteins and a curli-associated protein. The phylogenetic

tree of the PMP and bacterial curli proteins clustered them closer to the *Alphaproteobacteria* representatives (Fig. S4B).

TSS VIII has not been detected in SAR11, but it has been described in other bacterial groups (80) as the transporter of curli, surface-associated amyloid fibers mainly involved in adhesion to surfaces, biofilm formation, and interaction with host factors and the host immune system (81, 82). The two proteins identified as part of the TSS VIII in PMPs are CsgF, an extracellular chaperone involved in anchoring curli fibers to the outer membrane (83), and CsgG, which form the outer membrane diffusion channel (84). Both hypothetical proteins in the operon are of the same size, similarly to *csgA* and *csgB* genes (85), while the curli-associated protein is of the same size as CsgE, although no similarity could be detected at the sequence level or predicted structural level. Several experiments have shown that the only proteins required for curli phenotype expression are CsgA, CsgB, CsgF, and CsgG (CsgE increases almost 20-fold the amount of curli released, but it is not essential) (83, 86). Therefore, CsgA and CsgB are the only proteins missing in PMPs for the infected cells to express a curli phenotype.

DISCUSSION

The kind of bioinformatic approach utilized here can be applied to other microbes difficult to cultivate but with some isolates already sequenced. The diversity of sequences retrieved indicate that similar methods could provide much more complete pictures of the biodiversity of viruses infecting relevant but hard to grow microbes such as SAR11. In this case, its prevalence in superficial waters of the ocean and other aquatic habitats played in our favor, and we have been able to uncover a remarkable diversity of viral entities different from the cultivated reference. It seems clear that the amplification of PMPs in viromes is negatively affected by one or more biases, with MDA amplification being a prime suspect, and the same might be true for other myoviruses. This application of metagenomics complements culture to capture more phage diversity in natural environments (14).

The host cells belonging to the SAR11 clade are characterized by marked streamlining of the genomes (8). Myophages, on the other hand, are very large phages with big and complex genomes. In fact, the ones described here are even more complex than *E. coli* phage T4, with a large host recognition hypervariable island and novel sets of AMGs. They are actually closest to CMPs, a group of myophages whose host range also includes streamlined microbes (e.g., *Prochlorococcus*) inhabiting a similar habitat, an interesting convergence considering the phylogenetic distance between the hosts. Among the special features of the PMP genomes, it is remarkable that the large hypervariable region involved in host recognition in addition to several tail fibers, often contained glycosyltransferases, which might be involved in surface alterations that could lead to changes in phage recognition, preventing coinfection by other phages preying on the same host. That these large phages of SAR11 require a change in the host surface is not surprising, given the potentially sharp competition with, for example, SAR11 podophages that have much larger burst sizes (42 ± 7 versus 9 ± 2 for the cultured representatives) (10, 11). The genes provided by the phage might induce a change in the structures responsible for phage recognition and act as a serotype conversion mechanism to avoid superinfection by other phages (87). Similar mechanisms have been described for other marine and nonmarine podoviruses (88–90).

PMPs are, to our knowledge, the first phages that code for a partial curli-secreting system. The origin of this operon is unclear, since so far, the TSS VIII secretion system has not been described in the SAR11 clade. However, its remarkable similarity to the TSS VIII operon described in *Alpha-* and *Gammaproteobacteria* suggests that it is a product of a lateral transfer event. The function of such a system in viruses is also a mystery. The only two proteins identified as part of the TSS VIII in PMPs are CsgF and CsgG, which implies that if no other proteins in the operon are functional, it would code for only an extracellular chaperone and a pore-forming complex, respectively. The CsgG pore is too small to allow for virion exit (the CsgG pore has 40-Å inner diameter, while the HTVC008M capsid diameter is 550 Å) (10, 86), and the only report of functional

amyloids in viruses was in eukaryotic viruses, where they have the role of inhibiting programmed cell death of their eukaryotic host by sequestering effector proteins (89), which does not require the presence of the curli transporter. The simplest explanation would be that the pore structure might enhance the uptake of larger molecules. However, this does not explain the presence of CsgF, as it is not needed for the assembly of CsgG (82, 86) or the other genes present within the operon. Another, bolder hypothesis would be the involvement of these genes in the production of myeloid-like fibers. Some of the hypothetical proteins in the curlin cluster could be functional equivalents of CsgA and CsgB (86). If this were the case, they might induce aggregation, facilitating the acquisition of new host cells to the released virions. Thus, the curli gene cluster would act as a capture mechanism by retaining in close proximity the recently divided cells, that would be successive hosts, leading to a much larger phage offspring. This strategy could be called "sibling capture," and would be highly desirable in diluted environments such as the pelagic habitat in oligotrophic waters.

MATERIALS AND METHODS

Genome mining strategy and output. Following the workflow shown in Fig. S1A in the supplemental material, the reference cultivated PMP genome (HTVC008M) (10) and metagenomic PMP sequences MAVG-2, MAVG-4, MAVG-5, and Io7-C40 (14), were used as bait to comb through a vast quantity of contigs derived from several metagenomic and viromic samples (Table S1) (13, 14, 41, 91–94). First, a hidden Markov model (HMM) made from an alignment of *terL* gene sequences was used to identify viral contigs larger than 5 kb. The *terL* gene from the extracted contigs was then used to construct a phylogenetic tree (Fig. S7A). The position of the *terL* gene of the reference PMP in this tree was then used to recover a set of candidate contigs (Fig. S1B and S5). As mentioned previously, the closest group to PMPs are CMPs, which are expected to be present in significant quantities in the surveyed metagenomes. To remove all CMP-related contigs from the candidates, two collections of gene clusters were built, (i) one of them derived from 28 CMP genomes downloaded from the NCBI Refseq database (95) and (ii) another derived from the reference PMP genomes. Gene clusters shared between both collections were removed. HMMs built from both cluster collections were used to classify the contigs, keeping only those that had at least a match to a PMP gene cluster and no matches to any CMP gene cluster (Fig. S1).

MAVG cross-assembly. The contigs obtained from the genome-mining step were subjected to a cross-assembly step. Identical sequences were removed from the analysis, always keeping the longer contig if they did not have the same length. Contigs were then separated into bins of overlapping contigs based on an all-versus-all comparison (Fig. S1). Next, the bins were assembled manually into MAVGs as described previously (14) provided that (i) overlaps between contigs had a nucleotide sequence identity of >99%, an alignment length of >1,000 nt, and gaps of <10 nt, (ii) all overlaps were corroborated by more than two contigs, and (iii) sample metadata were ecologically coherent for all involved contigs (for example, not assembling contigs from freshwater and marine samples together). After this cross-assembly step, we obtained 14,748 sequences with an average length of 28 kb (Fig. S1B). Finally, contigs recovered were filtered by size (>100 kb), GC content (30 to 35%, which is the GC% range of the host), the number of proteins matching to SAR11 (>70% of identity), and tRNA gene matches (>95% of identity).

Recruitment analysis. To assess the distribution and abundance patterns of the recovered PMP MAVGs, the genomes were recruited using BLASTN (96) against the *Tara* Oceans metagenomes (40, 91), Geotraces cellular metagenomes (41), and the Mediterranean metagenomes described previously (14, 39). PMP group PMP-D were also recruited against the virome data sets they were recovered from (97) and against samples from other freshwater environments (Lake Biwa [98], Lake Simoncouche [99], Lake Kivu [GOLD Study identifier {ID} Gs0127566], Baltic Sea [100]). Normalization was performed by calculating RPKG (reads recruited per kilobase of the genome per gigabase of the metagenome) so recruitment values could be compared across samples. For linear metagenomic recruitments, metagenomic reads were aligned using BLASTN, with a cutoff of 70% nucleotide identity over a minimum alignment length of 50 nucleotides. The resulting alignments were plotted using the ggplot2 package in R. Figure 3A (cellular fraction versus viral fraction plot) was plotted following the scripts included in reference 33.

Phylogenetic tree of the recovered genomes. Common proteins to all 35 genomes were calculated using the GET_HOMOLOGUES (101) software package. The five common proteins identified were concatenated and aligned using MUSCLE (102) and a maximum-likelihood tree was then constructed using RAxML (103) with the following parameters: "-f a" algorithm, 100 bootstrap replicates, PROTGAMMAJTT model.

Protein isoelectric point determination. To determine the isoelectric point distribution patterns of the phage genomes, calculations of all predicted proteins for both genomes were calculated with the Pepstats software from the EMBOSS package (104). The resulting isoelectric point values were plotted using the ggplot2 package in R.

Genomic pairwise comparison. Average nucleotide identity (ANI) and coverage between a pair of genomes were calculated using the Jspecies software with default parameters (105).

Genome annotation. Genes and tRNAs were predicted using Prodigal (106) and tRNAscan-SE (107), respectively. Functional annotation of predicted genes followed a consensus-based approach. First, the genes from all PMPs and the reference CMPs were annotated against the uniref90 protein database (108) (using DIAMOND [109]) and the CDD (110) and pVOG (111) HMM databases (using hmmscan [112]). For each database, we assigned to each gene sequence the best hit with an E value of at least $<10^{-5}$ and an alignment length of between 70% and 130% of the query length. Genes were then clustered using GET_HOMOLOGUES (101) and the annotations for each cluster were manually curated to ensure that the annotations were coherent for all genes in the cluster. In the cases where we found discrepancies, the second and third best hits were used to verify the annotation. Finally, the remaining clusters without annotation were compared against the PDB HMM database (113) using hhblits (114). Clusters with less than 10 sequences were first inflated by using the uniclust30 (115) database.

Cooccurrence matrix. Terminator sequences were predicted for both CMP and PMP genomes using Transterm_HP (116), while early promoter sequences were predicted using BPROM (117). Prediction of middle and late promoter sequences was attempted following the steps described previously (24) but was unsuccessful in PMP genomes. Genes that pertain to a protein cluster (obtained in the genome annotation step) in each genome were then grouped into operons based on terminator positions and strand changes. These operons were then used as the basis for a cooccurrence matrix. Two protein clusters (nodes) were linked to each other if they were present in two genomes and were part of the same operon, with edge strength representing the number of genome pairs where this was the case. Edges with edge strength representing 0.05% of the total were removed from the matrix. The matrix was then used to build a network in Cytoscape (118). The add-on ClusterMaker2 (119) was used to separate the cooccurrence network into clusters (MCL algorithm, 2.5 granularity).

Data availability. Viral sequences presented in this article have been submitted to NCBI and are available under BioProject accession number PRJNA588231.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.2 MB.
FIG S2, PDF file, 0.2 MB.
FIG S3, PDF file, 2.4 MB.
FIG S4, PDF file, 0.2 MB.
FIG S5, PDF file, 0.1 MB.
TABLE S1, PDF file, 0.4 MB.
TABLE S2, XLSX file, 0.04 MB.
TABLE S3, XLSX file, 0.1 MB.
TABLE S4, PDF file, 0.2 MB.
TABLE S5, PDF file, 0.2 MB.

ACKNOWLEDGMENTS

This work was supported by grant "VIREVO" CGL2016-76273-P (AEI/FEDER, EU) (cofounded with FEDER funds) from the Spanish Ministerio de Economía, Industria y Competitividad and by grant PROMETEO/2019/009 "HIDRAS3" from Generalitat Valenciana, both granted to F.R.-V. F.R.-V. was also a beneficiary of the 5top100-program of the Ministry for Science and Education of Russia. A.Z.-S. was supported by a Ph.D. fellowship from the Spanish Ministerio de Economía y Competitividad (BES-2017-079993). M.L.-P. was supported by a postdoctoral fellowship from the Spanish Ministerio de Economía, Industria y Competitividad (IJCI-2017-34002).

M.L.-P. conceived the study. A.Z.-S. and F.R.-V. analyzed the data. M.L.-P., F.R.-V., and A.Z.-S. contributed to writing the manuscript.

We declare that we have no competing interests.

REFERENCES

1. Wommack KE, Colwell RR. 2000. Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev 64:69–114. https://doi.org/10.1128/mmbr.64.1.69-114.2000.
2. Fuhrman JA. 1999. Marine viruses and their biogeochemical and ecological effects. Nature 399:541–548. https://doi.org/10.1038/21119.
3. Weinbauer MG. 2004. Ecology of prokaryotic viruses. FEMS Microbiol Rev 28:127–181. https://doi.org/10.1016/j.femsre.2003.08.001.
4. Wilhelm SW, Suttle CA. 1999. Viruses and nutrient cycles in the sea: viruses play critical roles in the structure and function of aquatic food webs. Bioscience 49:781–788. https://doi.org/10.2307/1313569.
5. Suttle CA. 2007. Marine viruses — major players in the global ecosystem. Nat Rev Microbiol 5:801–812. https://doi.org/10.1038/nrmicro1750.
6. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F, Mira A. 2009. Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7:828–836. https://doi.org/10.1038/nrmicro2235.
7. Rohwer F, Thurber RV. 2009. Viruses manipulate the marine environment. Nature 459:207–212. https://doi.org/10.1038/nature08060.
8. Giovannoni SJ. 2017. SAR11 bacteria: the most abundant plankton in the oceans. Annu Rev Mar Sci 9:231–255. https://doi.org/10.1146/annurev-marine-010814-015934.

9. Henson MW, Lanclos VC, Faircloth BC, Thrash JC. 2018. Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. ISME J 12:1846–1860. https://doi.org/10.1038/s41396-018-0092-2.
10. Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC, Ellisman M, Deerinck T, Sullivan MB, Giovannoni SJ. 2013. Abundant SAR11 viruses in the ocean. Nature 494:357–360. https://doi.org/10.1038/nature11921.
11. Zhao Y, Qin F, Zhang R, Giovannoni SJ, Zhang Z, Sun J, Du S, Rensing C. 2019. Pelagiphages in the Podoviridae family integrate into host genomes. Environ Microbiol 21:1989–2001. https://doi.org/10.1111/1462-2920.14487.
12. Ackermann H-W. 2003. Bacteriophage observations and evolution. Res Microbiol 154:245–251. https://doi.org/10.1016/S0923-2508(03)00067-6.
13. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. 2013. Expanding the marine virosphere using metagenomics. PLoS Genet 9:e1003987. https://doi.org/10.1371/journal.pgen.1003987.
14. López-Pérez M, Haro-Moreno JM, Gonzalez-Serrano R, Parras-Moltó M, Rodriguez-Valera F. 2017. Genome diversity of marine phages recovered from Mediterranean metagenomes: size matters. PLoS Genet 13:e1007018. https://doi.org/10.1371/journal.pgen.1007018.
15. Eggleston EM, Hewson I. 2016. Abundance of two Pelagibacter ubique bacteriophage genotypes along a latitudinal transect in the North and South Atlantic Oceans. Front Microbiol 7:1534. https://doi.org/10.3389/fmicb.2016.01534.
16. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99:14250–14255. https://doi.org/10.1073/pnas.202488399.
17. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science 311:496–503. https://doi.org/10.1126/science.1120250.
18. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia J-M, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC. 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5:e16. https://doi.org/10.1371/journal.pbio.0050016.
19. Filée J, Tétart F, Suttle CA, Krisch HM. 2005. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. Proc Natl Acad Sci U S A 102:12471–12476. https://doi.org/10.1073/pnas.0503404102.
20. Ackermann H-W, Krisch HM. 1997. A catalogue of T4-type bacteriophages. Arch Virol 142:2329–2345. https://doi.org/10.1007/s007050050246.
21. Petrov VM, Ratnayaka S, Nolan JM, Miller ES, Karam JD. 2010. Genomes of the T4-related bacteriophages as windows on microbial genome evolution. Virol J 7:292. https://doi.org/10.1186/1743-422X-7-292.
22. Marti R, Zurfluh K, Hagens S, Pianezzi J, Klumpp J, Loessner MJ. 2013. Long tail fibres of the novel broad-host-range T-even bacteriophage S16 specifically recognize Salmonella OmpC. Mol Microbiol 87:818–834. https://doi.org/10.1111/mmi.12134.
23. Brewer TE, Stroupe ME, Jones KM. 2014. The genome, proteome and phylogenetic analysis of Sinorhizobium meliloti phage ΦM12, the founder of a new group of T4-superfamily phages. Virology 450-451:84–97. https://doi.org/10.1016/j.virol.2013.11.027.
24. Sullivan MB, Huang KH, Ignacio-Espinoza JC, Berlin AM, Kelly L, Weigele PR, DeFrancesco AS, Kern SE, Thompson LR, Young S, Yandava C, Fu R, Krastins B, Chase M, Sarracino D, Osburne MS, Henn MR, Chisholm SW. 2010. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environ Microbiol 12:3035–3056. https://doi.org/10.1111/j.1462-2920.2010.02280.x.
25. Comeau AM, Krisch HM. 2008. The capsid of the T4 phage superfamily: the evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. Mol Biol Evol 25:1321–1332. https://doi.org/10.1093/molbev/msn080.
26. Chen C-R, Lin C-H, Lin J-W, Chang C-I, Tseng Y-H, Weng S-F. 2007. Characterization of a novel T4-type Stenotrophomonas maltophilia virulent phage Smp14. Arch Microbiol 188:191–197. https://doi.org/10.1007/s00203-007-0238-5.
27. Waterbury JB, Valois FW. 1993. Resistance to co-occurring phages enables marine Synechococcus communities to coexist with cyanophages abundant in seawater. Appl Environ Microbiol 59:3393–3399. https://doi.org/10.1128/AEM.59.10.3393-3399.1993.
28. Sullivan MB, Waterbury JB, Chisholm SW. 2003. Cyanophages infecting the oceanic cyanobacterium Prochlorococcus. Nature 424:1047–1051. https://doi.org/10.1038/nature01929.
29. Mann NH, Clokie MRJ, Millard A, Cook A, Wilson WH, Wheatley PJ, Letarov A, Krisch HM. 2005. The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine Synechococcus strains. J Bacteriol 187:3188–3200. https://doi.org/10.1128/JB.187.9.3188-3200.2005.
30. Millard AD, Zwirglmaier K, Downey MJ, Mann NH, Scanlan DJ. 2009. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of Synechococcus host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution. Environ Microbiol 11:2370–2387. https://doi.org/10.1111/j.1462-2920.2009.01966.x.
31. Crummett LT, Puxty RJ, Weihe C, Marston MF, Martiny J. 2016. The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. Virology 499:219–229. https://doi.org/10.1016/j.virol.2016.09.016.
32. López-Pérez M, Haro-Moreno JM, de la Torre JR, Rodriguez-Valera F. 2019. Novel Caudovirales associated with marine group I Thaumarchaeota assembled from metagenomes. Environ Microbiol 21:1980–1988. https://doi.org/10.1111/1462-2920.14462.
33. Philosof A, Yutin N, Flores-Uribe J, Sharon I, Koonin EV, Béjà O. 2017. Novel abundant oceanic viruses of uncultured marine group II Euryarchaeota. Curr Biol 27:1362–1368. https://doi.org/10.1016/j.cub.2017.03.052.
34. Thrash JC, Temperton B, Swan BK, Landry ZC, Woyke T, DeLong EF, Stepanauskas R, Giovannoni SJ. 2014. Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. ISME J 8:1440–1451. https://doi.org/10.1038/ismej.2013.243.
35. Comeau AM, Bertrand C, Letarov A, Tétart F, Krisch HM. 2007. Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. Virology 362:384–396. https://doi.org/10.1016/j.virol.2006.12.031.
36. Markine-Goriaynoff N, Gillet L, Van Etten JL, Korres H, Verma N, Vanderplasschen A. 2004. Glycosyltransferases encoded by viruses. J Gen Virol 85:2741–2754. https://doi.org/10.1099/vir.0.80320-0.
37. López-Pérez M, Rodriguez-Valera F. 2016. Pangenome evolution in the marine bacterium Alteromonas. Genome Biol Evol 8:1556–1570. https://doi.org/10.1093/gbe/evw098.
38. Haro-Moreno JM, López-Pérez M, de la Torre JR, Picazo A, Camacho A, Rodriguez-Valera F. 2018. Fine metagenomic profile of the Mediterranean stratified and mixed water columns revealed by assembly and recruitment. Microbiome 6:128. https://doi.org/10.1186/s40168-018-0513-5.
39. Coutinho FH, Rosselli R, Rodríguez-Valera F. 2019. Trends of microdiversity reveal depth-dependent evolutionary strategies of viruses in the Mediterranean. mSystems 4:e00554-19. https://doi.org/10.1128/mSystems.00554-19.
40. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Tara Oceans coordinators, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P. 2015. Structure and function of the global ocean microbiome. Science 348:1261359. https://doi.org/10.1126/science.1261359.
41. Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, Hogle SL, Coe A, Bergauer K, Bouman HA, Browning TJ, De Corte D, Hassler C, Hulston D, Jacquot JE, Maas EW, Reinthaler T, Sintes E, Yokokawa T, Chisholm SW. 2018. Marine microbial metagenomes sampled across space and time. Sci Data 5:180176. https://doi.org/10.1038/sdata.2018.176.
42. Luef B, Frischkorn KR, Wrighton KC, Holman HYN, Birarda G, Thomas BC, Singh A, Williams KH, Siegerist CE, Tringe SG, Downing KH, Comolli LR, Banfield JF. 2015. Diverse uncultivated ultra-small bacterial cells in groundwater. Nat Commun 6:6372. https://doi.org/10.1038/ncomms7372.
43. Haible D, Kober S, Jeske H. 2006. Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. J Virol Methods 135:9–16. https://doi.org/10.1016/j.jviromet.2006.01.017.
44. Marine R, McCarren C, Vorrasane V, Nasko D, Crowgey E, Polson SW,

Wommack KE. 2014. Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. Microbiome 2:3. https://doi.org/10.1186/2049-2618-2-3.
45. Martín-Cuadrado AB, López-García P, Alba JC, Moreira D, Monticelli L, Strittmatter A, Gottschalk G, Rodríguez-Valera F. 2007. Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. PLoS One 2:e914. https://doi.org/10.1371/journal.pone.0000914.
46. Boyington JC, Gladyshev VN, Khangulov SV, Stadtman TC, Sun PD. 1997. Crystal structure of formate dehydrogenase H: catalysis involving Mo, molybdopterin, selenocysteine, and an Fe4S4 cluster. Science 275:1305–1308. https://doi.org/10.1126/science.275.5304.1305.
47. Cabello-Yeves PJ, Rodriguez-Valera F. 2019. Marine-freshwater prokaryotic transitions require extensive changes in the predicted proteome. Microbiome 7:117. [CrossRef] https://doi.org/10.1186/s40168-019-0731-5.
48. Cliffe LJ, Siegel TN, Marshall M, Cross GAM, Sabatini R. 2010. Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. Nucleic Acids Res 38:3923–3935. https://doi.org/10.1093/nar/gkq146.
49. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324:930–935. https://doi.org/10.1126/science.1170116.
50. Taylor NMI, Prokhorov NS, Guerrero-Ferreira RC, Shneider MM, Browning C, Goldie KN, Stahlberg H, Leiman PG. 2016. Structure of the T4 baseplate and its function in triggering sheath contraction. Nature 533:346–352. https://doi.org/10.1038/nature17971.
51. Yap ML, Klose T, Arisaka F, Speir JA, Veesler D, Fokine A, Rossmann MG. 2016. Role of bacteriophage T4 baseplate in regulating assembly and infection. Proc Natl Acad Sci U S A 113:2654–2659. https://doi.org/10.1073/pnas.1601654113.
52. Nováček J, Šiborová M, Benešík M, Pantůček R, Doškař J, Plevka P. 2016. Structure and genome release of Twort-like Myoviridae phage with a double-layered baseplate. Proc Natl Acad Sci U S A 113:9351–9356. https://doi.org/10.1073/pnas.1605883113.
53. Habann M, Leiman PG, Vandersteegen K, Van den Bossche A, Lavigne R, Shneider MM, Bielmann R, Eugster MR, Loessner MJ, Klumpp J. 2014. Listeria phage A511, a model for the contractile tail machineries of SPO1-related bacteriophages. Mol Microbiol 92:84–99. https://doi.org/10.1111/mmi.12539.
54. Kadirvelraj R, Foley BL, Dyekjær JD, Woods RJ. 2008. Involvement of water in carbohydrate-protein binding: concanavalin A revisited. J Am Chem Soc 130:16933–16942. https://doi.org/10.1021/ja8039663.
55. Mühlenkamp M, Oberhettinger P, Leo JC, Linke D, Schütz MS. 2015. Yersinia adhesin A (YadA) – beauty & beast. Int J Med Microbiol 305:252–258. https://doi.org/10.1016/j.ijmm.2014.12.008.
56. Mizuno CM, Ghai R, Rodriguez-Valera F. 2014. Evidence for metaviromic islands in marine phages. Front Microbiol 5:27. https://doi.org/10.3389/fmicb.2014.00027.
57. Smith NL, Taylor EJ, Lindsay AM, Charnock SJ, Turkenburg JP, Dodson EJ, Davies GJ, Black GW. 2005. Structure of a group A streptococcal phage-encoded virulence factor reveals a catalytically active triple-stranded β-helix. Proc Natl Acad Sci U S A 102:17652–17657. https://doi.org/10.1073/pnas.0504782102.
58. Yu Z, An B, Ramshaw JAM, Brodsky B. 2014. Bacterial collagen-like proteins that form triple-helical structures. J Struct Biol 186:451–461. https://doi.org/10.1016/j.jsb.2014.01.003.
59. Nakagawa H, Arisaka F, Ishii SI. 1985. Isolation and characterization of the bacteriophage T4 tail-associated lysozyme. J Virol 54:460–466. https://doi.org/10.1128/JVI.54.2.460-466.1985.
60. Clokie MRJ, Millard AD, Mann NH. 2010. T4 genes in the marine ecosystem: studies of the T4-like cyanophages and their role in marine ecology. Virol J 7:291. https://doi.org/10.1186/1743-422X-7-291.
61. Hinton DM. 2010. Transcriptional control in the prereplicative phase of T4 development. Virol J 7:289. https://doi.org/10.1186/1743-422X-7-289.
62. Diaconu M, Kothe U, Schlünzen F, Fischer N, Harms JM, Tonevitsky AG, Stark H, Rodnina MV, Wahl MC. 2005. Structural basis for the function of the ribosomal L7/12 stalk in factor binding and GTPase activation. Cell 121:991–1004. https://doi.org/10.1016/j.cell.2005.04.015.
63. Soma A, Ikeuchi Y, Kanemasa S, Kobayashi K, Ogasawara N, Ote T, Kato JI, Watanabe K, Sekine Y, Suzuki T. 2003. An RNA-modifying enzyme that governs both the codon and amino acid specificities of isoleucine tRNA. Mol Cell 12:689–698. https://doi.org/10.1016/s1097-2765(03)00346-0.
64. Mizuno CM, Guyomar C, Roux S, Lavigne R, Rodriguez-Valera F, Sullivan MB, Gillet R, Forterre P, Krupovic M. 2019. Numerous cultivated and uncultivated viruses encode ribosomal proteins. Nat Commun 10:752. [CrossRef] https://doi.org/10.1038/s41467-019-08672-6.
65. Van Duin J, Wijnands R. 1981. The function of ribosomal protein S21 in protein synthesis. Eur J Biochem 118:615–619. https://doi.org/10.1111/j.1432-1033.1981.tb05563.x.
66. Breitbart M, Thompson LR, Suttle CA, Sullivan MB. 2007. Exploring the vast diversity of marine viruses. Oceanography 20:135–139. https://doi.org/10.5670/oceanog.2007.58.
67. Gao EB, Huang Y, Ning D. 2016. Metabolic genes within cyanophage genomes: implications for diversity and evolution. Genes (Basel) 7:E80. https://doi.org/10.3390/genes7100080.
68. Hurwitz BL, U'Ren JM. 2016. Viral metabolic reprogramming in marine ecosystems. Curr Opin Microbiol 31:161–168. https://doi.org/10.1016/j.mib.2016.04.002.
69. Esmon BE, Kensil CR, Cheng CHC, Glaser M. 1980. Genetic analysis of Escherichia coli mutants defective in adenylate kinase and sn-glycerol 3-phosphate acyltransferase. J Bacteriol 141:405–408. https://doi.org/10.1128/JB.141.1.405-408.1980.
70. Ross P, O'Gara F, Condon S. 1990. Cloning and characterization of the thymidylate synthase gene from Lactococcus lactis subsp. lactis. Appl EnvironMicrobiol56:2156–2163.https://doi.org/10.1128/AEM.56.7.2156-2163.1990.
71. Moore JT, Silversmith RE, Maley GF, Maley F. 1993. T4-phage deoxycytidylate deaminase is a metalloprotein containing two zinc atoms per subunit. J Biol Chem 268:2288–2291.
72. Frank JA, Lorimer D, Youle M, Witte P, Craig T, Abendroth J, Rohwer F, Edwards RA, Segall AM, Burgin AB. 2013. Structure and function of a cyanophage-encoded peptide deformylase. ISME J 7:1150–1160.
73. Santos JM, Freire P, Vicente M, Arraiano CM. 1999. The stationary-phase morphogene bolA from Escherichia coli is induced by stress during early stages of growth. Mol Microbiol 32:789–798. https://doi.org/10.1046/j.1365-2958.1999.01397.x.
74. Mukherjee S, Sengupta S. 2016. Riboswitch Scanner: an efficient pHMM-based web-server to detect riboswitches in genomic sequences. Bioinformatics 32:776–778. https://doi.org/10.1093/bioinformatics/btv640.
75. Petit C, Rigg GP, Pazzani C, Smith A, Sieberth V, Stevens M, Boulnois G, Jann K, Roberts IS. 1995. Region 2 of the Escherichia coli K5 capsule gene cluster encoding proteins for the biosynthesis of the K5 polysaccharide. Mol Microbiol 17:611–620. https://doi.org/10.1111/j.1365-2958.1995.mmi_17040611.x.
76. Valvano MA, Marolda CL, Bittner M, Glaskin-Clay M, Simon TL, Klena JD. 2000. The rfaE gene from Escherichia coli encodes a bifunctional protein involved in biosynthesis of the lipopolysaccharide core precursor ADP-L-glycero-D-manno-heptose. J Bacteriol 182:488–497. https://doi.org/10.1128/jb.182.2.488-497.2000.
77. Lairson LL, Henrissat B, Davies GJ, Withers SG. 2008. Glycosyltransferases: structures, functions, and mechanisms. Annu Rev Biochem 77:521–555. https://doi.org/10.1146/annurev.biochem.76.061005.092322.
78. Campbell JA, Davies GJ, Bulone V, Henrissat B. 1997. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. Biochem J 326:929–939. https://doi.org/10.1042/bj3260929u.
79. Haro-Moreno JM, Rodriguez-Valera F, Rosselli R, Martinez-Hernandez F, Roda-Garcia JJ, Lluesma Gomez M, Fornas O, Martinez-Garcia M, López-Pérez M. 15 December 2019. Ecogenomics of the SAR11 clade. Environ Microbiol https://doi.org/10.1111/1462-2920.14896.
80. Dueholm MS, Albertsen M, Otzen D, Nielsen PH. 2012. Curli functional amyloid systems are phylogenetically widespread and display large diversity in operon and protein structure. PLoS One 7:e51274. https://doi.org/10.1371/journal.pone.0051274.
81. Barnhart MM, Chapman MR. 2006. Curli biogenesis and function. Annu Rev Microbiol 60:131–147. https://doi.org/10.1146/annurev.micro.60.080805.142106.
82. Evans ML, Chapman MR. 2014. Curli biogenesis: order out of disorder. Biochim Biophys Acta 1843:1551–1558. https://doi.org/10.1016/j.bbamcr.2013.09.010.
83. Nenninger AA, Robinson LS, Hultgren SJ. 2009. Localized and efficient curli nucleation requires the chaperone-like amyloid assembly protein CsgF. Proc Natl Acad Sci U S A 106:900–905. https://doi.org/10.1073/pnas.0812143106.

84. Robinson LS, Ashman EM, Hultgren SJ, Chapman MR. 2006. Secretion of curli fibre subunits is mediated by the outer membrane-localized CsgG protein. Mol Microbiol 59:870–881. https://doi.org/10.1111/j.1365-2958.2005.04997.x.
85. Hammer ND, Schmidt JC, Chapman MR. 2007. The curli nucleator protein, CsgB, contains an amyloidogenic domain that directs CsgA polymerization. Proc Natl Acad Sci U S A 104:12494–12499. https://doi.org/10.1073/pnas.0703310104.
86. Van Gerven N, Klein RD, Hultgren SJ, Remaut H. 2015. Bacterial amyloid formation: structural insights into curli biogensis. Trends Microbiol 23:693–706. https://doi.org/10.1016/j.tim.2015.07.010.
87. Iyer LM, Koonin EV, Aravind L. 2002. Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. Genome Biol 3:research0012. https://doi.org/10.1186/gb-2002-3-3-research0012.
88. Duhaime MB, Solonenko N, Roux S, Verberkmoes NC, Wichels A, Sullivan MB. 2017. Comparative omics and trait analyses of marine Pseudoalteromonas phages advance the phage OTU concept. Front Microbiol 8:1241. https://doi.org/10.3389/fmicb.2017.01241.
89. Hardies SC, Hwang YJ, Hwang CY, Jang GI, Cho BC. 2013. Morphology, physiological characteristics, and complete sequence of marine bacteriophage ϕRIO-1 infecting Pseudoalteromonas marina. J Virol 87:9189–9198. https://doi.org/10.1128/JVI.01521-13.
90. Pickard D, Toribio AL, Petty NK, Van Tonder A, Yu L, Goulding D, Barrell B, Rance R, Harris D, Wetter M, Wain J, Choudhary J, Thomson N, Dougan G. 2010. A conserved acetyl esterase domain targets diverse bacteriophages to the Vi capsular receptor of Salmonella enterica serovar Typhi. J Bacteriol 192:5746–5754. https://doi.org/10.1128/JB.00659-10.
91. Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, Iudicone D, Karsenti E, Speich S, Troublé R, Dimier C, Searson S, Tara Oceans Consortium Coordinators. 2015. Open science resources for the discovery and analysis of Tara Oceans data. Sci Data 2:150023. [CrossRef] https://doi.org/10.1038/sdata.2015.23.
92. Haro-Moreno JM, Rodriguez-Valera F, López-Pérez M. 2019. Prokaryotic population dynamics and viral predation in a marine succession experiment using metagenomics. Front Microbiol 10:2926. https://doi.org/10.3389/fmicb.2019.02926.
93. Paez-Espino D, Roux S, Chen IMA, Palaniappan K, Ratner A, Chu K, Huntemann M, Reddy TBK, Pons JC, Llabrés M, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC. 2019. IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. Nucleic Acids Res 47:D678–D686. https://doi.org/10.1093/nar/gky1127.
94. Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res 40:D115–D122. https://doi.org/10.1093/nar/gkr1044.
95. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. 2015. NCBI viral genomes resource. Nucleic Acids Res 43:D571–D577. https://doi.org/10.1093/nar/gku1207.
96. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389.
97. Mohiuddin M, Schellhorn HE. 2015. Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. Front Microbiol 6:960. https://doi.org/10.3389/fmicb.2015.00960.
98. Okazaki Y, Nishimura Y, Yoshida T, Ogata H, Nakano S. 2019. Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. Environ Microbiol 21:4740–4754. https://doi.org/10.1111/1462-2920.14816.
99. Tran P, Ramachandran A, Khawasik O, Beisner BE, Rautio M, Huot Y, Walsh DA. 2018. Microbial life under ice: metagenome diversity and in situ activity of Verrucomicrobia in seasonally ice-covered lakes. Environ Microbiol 20:2568–2584. https://doi.org/10.1111/1462-2920.14283.
100. Hugerth LW, Larsson J, Alneberg J, Lindh MV, Legrand C, Pinhassi J, Andersson AF. 2015. Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biol 16:279. [CrossRef] https://doi.org/10.1186/s13059-015-0834-7.
101. Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol 79:7696–7701. https://doi.org/10.1128/AEM.02411-13.
102. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.
103. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
104. Rice P, Longden L, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.
105. Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 106:19126–19131. https://doi.org/10.1073/pnas.0906412106.
106. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
107. Eddy SR, Lowe TM. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964.
108. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282–1288. https://doi.org/10.1093/bioinformatics/btm098.
109. Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176.
110. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222–D226. https://doi.org/10.1093/nar/gku1221.
111. Grazziotin AL, Koonin EV, Kristensen DM. 2017. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res 45:D491–D498. https://doi.org/10.1093/nar/gkw975.
112. Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform 23:205–211. https://doi.org/10.1142/9781848165632_0019.
113. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN. 2000. The Protein Data Bank. Nucleic Acids Res 28:235–242. https://doi.org/10.1093/nar/28.1.235.
114. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. 2019. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20:473. https://doi.org/10.1186/s12859-019-3019-7.
115. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res 45:D170–D176. https://doi.org/10.1093/nar/gkw1081.
116. Kingsford CL, Ayanbule K, Salzberg SL. 2007. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol 8:R22. https://doi.org/10.1186/gb-2007-8-2-r22.
117. Solovyev V, Salamov A. 2011. Automatic annotation of microbial genomes and metagenomic sequences, p 61–78. *In* Li RW (ed), Metagenomics and its applications in agriculture, biomedicine and environmental studies. Nova Science Publishers, Hauppauge, NY.
118. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303.
119. Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, Bader GD, Ferrin TE. 2011. ClusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics 12:436. https://doi.org/10.1186/1471-2105-12-436.

Metagenome Mining Reveals Hidden Genomic Diversity of Pelagimyophages in Aquatic Environments

Asier Zaragoza-Solas[1], Francisco Rodriguez-Valera[1,2,*] , Mario López-Pérez[1,*]

[1]Evolutionary Genomics Group, División de Microbiología, Universidad Miguel Hernández, 03550 San Juan de Alicante, Spain
[2]Laboratory for Theoretical and Computer Research on Biological Macromolecules and Genomes, Moscow Institute of Physics and Technology, Dolgoprudny, Russia

*Corresponding Authors: Francisco Rodriguez-Valera, frvalera@umh.es, or Mario López-Pérez, mario.lopezp@umh.es.

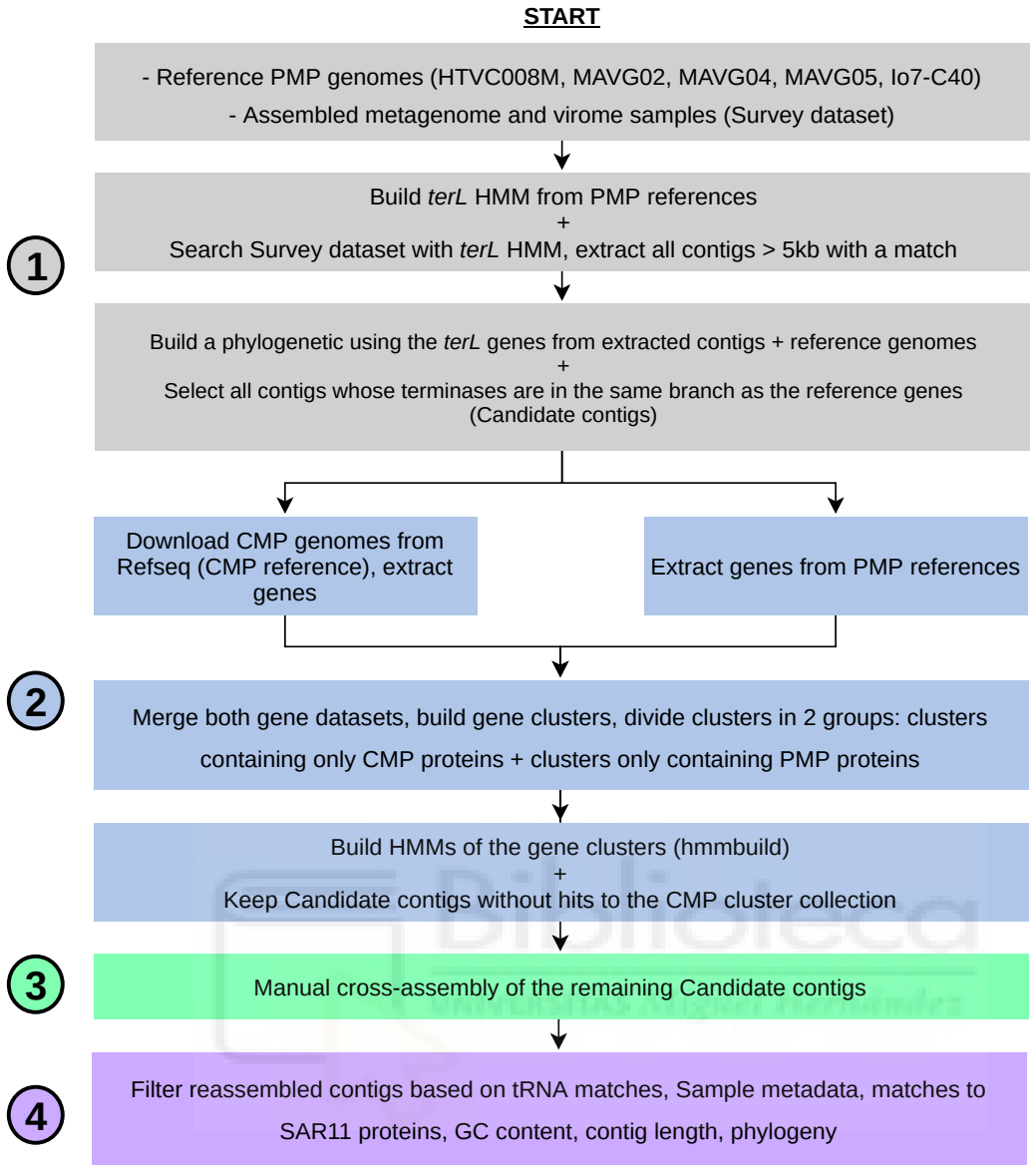
Evolutionary Genomics Group, División de Microbiología, Universidad Miguel Hernández, Apto 18, San Juan de Alicante, 03550 Alicante, Spain.
Phone +34 965919313, Fax +34 965919457

SUPPLEMENTARY FILES

Table S2 and Table S3 can be found at the publisher's website:
https://doi.org/10.1128/mSystems.00905-19

Genome mining flowchart

A



B

	Step	#Contigs	Average contig size (kbp)
1 →	*terL* hmm filtering	126,467	12.56
2 →	Gene cluster hmm filtering	30,331	25.89
3 →	Manual cross-assembly	28,431	26.42
3 →	Duplicate removal	14,748	28.30
4 →	Contigs > 100kbp	688	152.05
4 →	GC% between 30 – 35	197	143.03
4 →	Phylogeny + Gene filtering	31	137.19

Fig. S1. Genome mining pipeline. (A) Workflow describing the steps used in the genome mining process, color-keyed based on the step. (B) Contigs remaining in the analysis after each step, along with the average contig size. Each step includes a key to the workflow in (A).

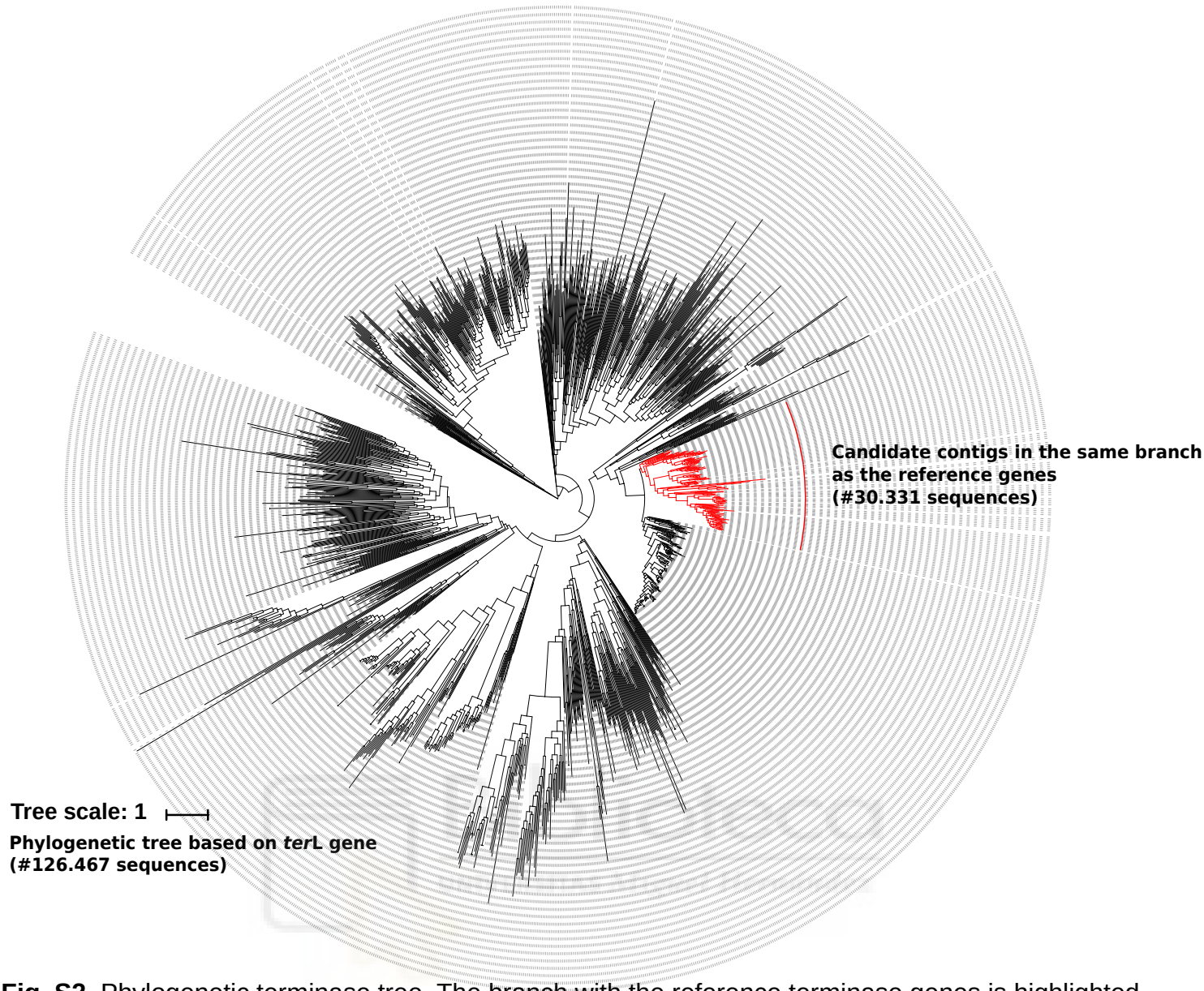


Fig. S2. Phylogenetic terminase tree. The branch with the reference terminase genes is highlighted in red.

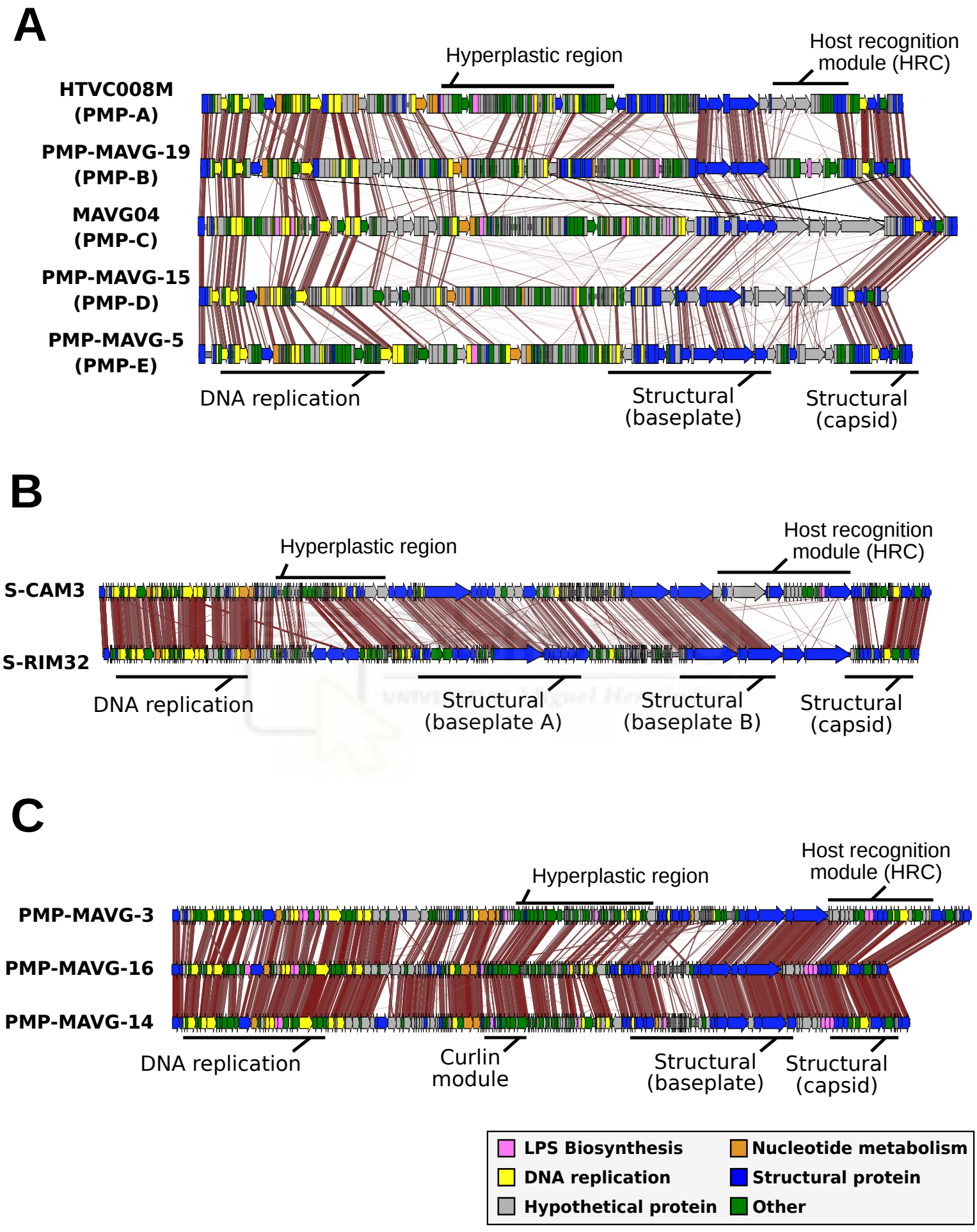


Fig. S3. Alignment of pelagimyophage (PMP) and cyanomyophage (CMP) phages (tblastx, 30% identity). Gene modules are labelled with black lines over the genomes. (A) Whole genome alignment of representatives of the five PMP groups. (B) Alignment of two CMPs. (C) Alignment of three PMPs from PMP group B with a similar Host Recognition Cluster (HRC).

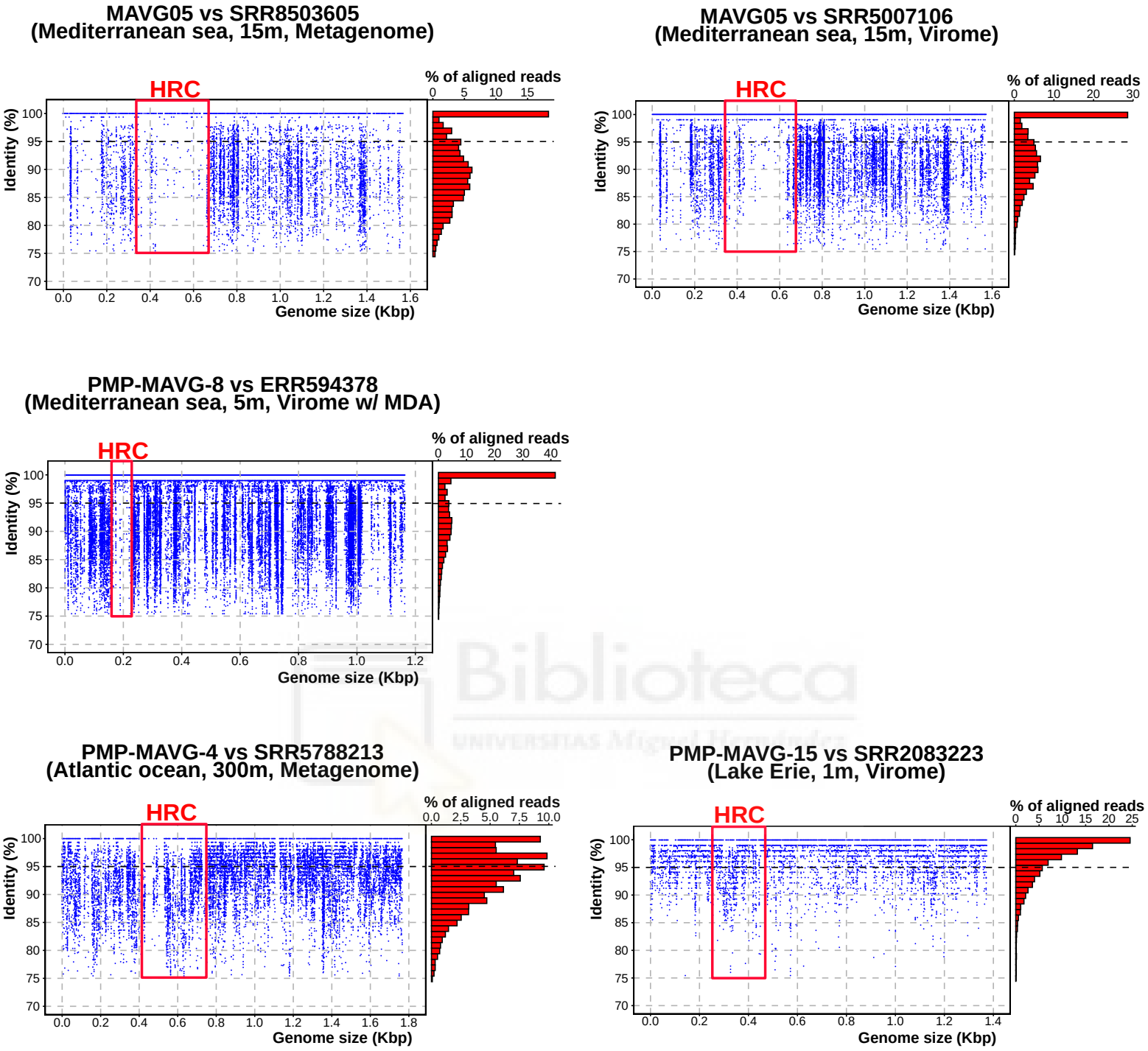


Fig. S4. Linear recruitments of various PMPs against various metagenomes and viromes. The Host Recognition Cluster (HRC) is highlighted with a red rectangle.

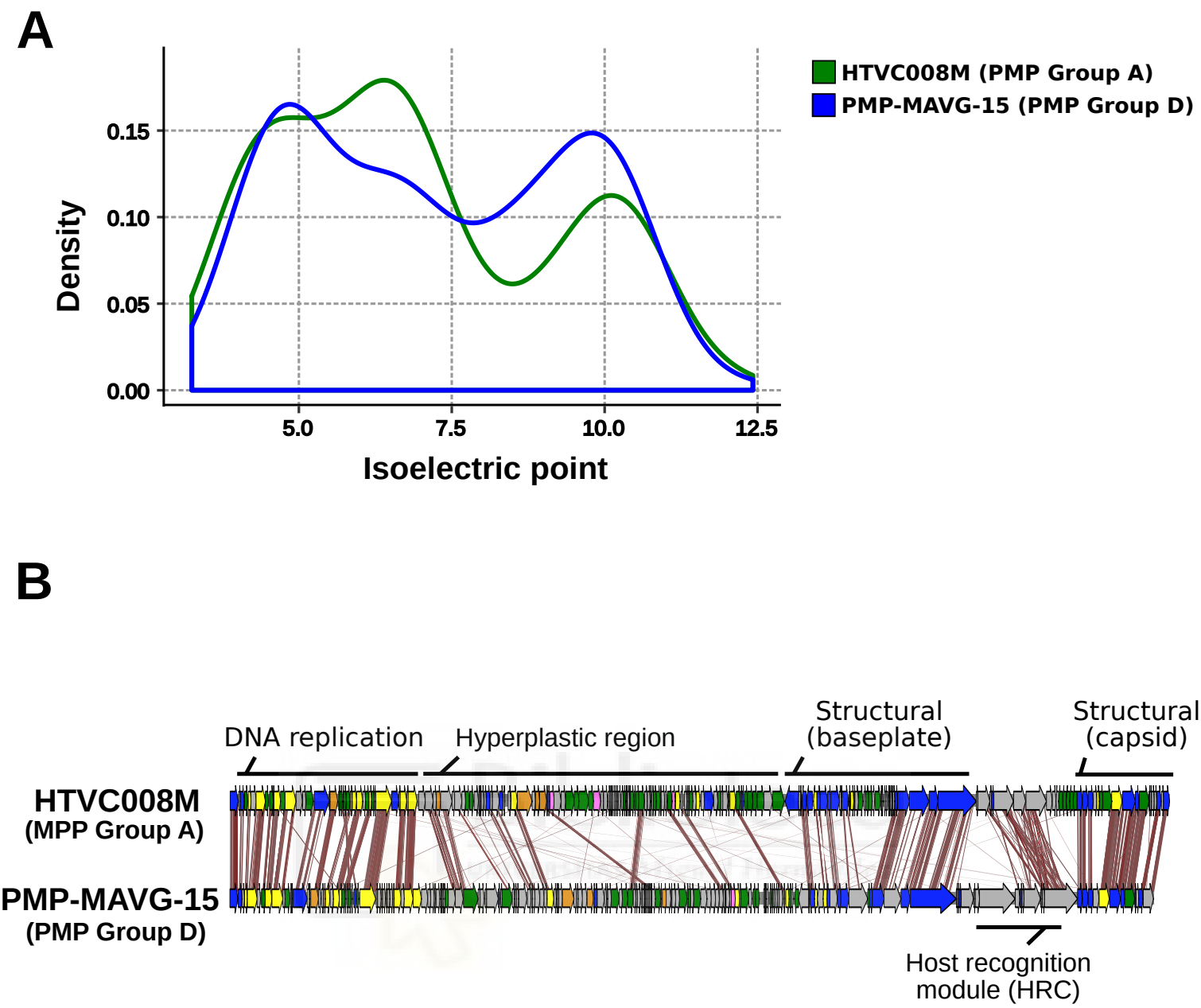


Fig. S5. Freshwater PMP group. (A) Isoelectric point versus density plot of marine (HTVC008M) and freshwater (PMP-MAVG-15) PMPs. (B) Genome alignment of HTVC008M and PMP-MAVG-15. Genomic modules are labelled with black bars over the genomes.

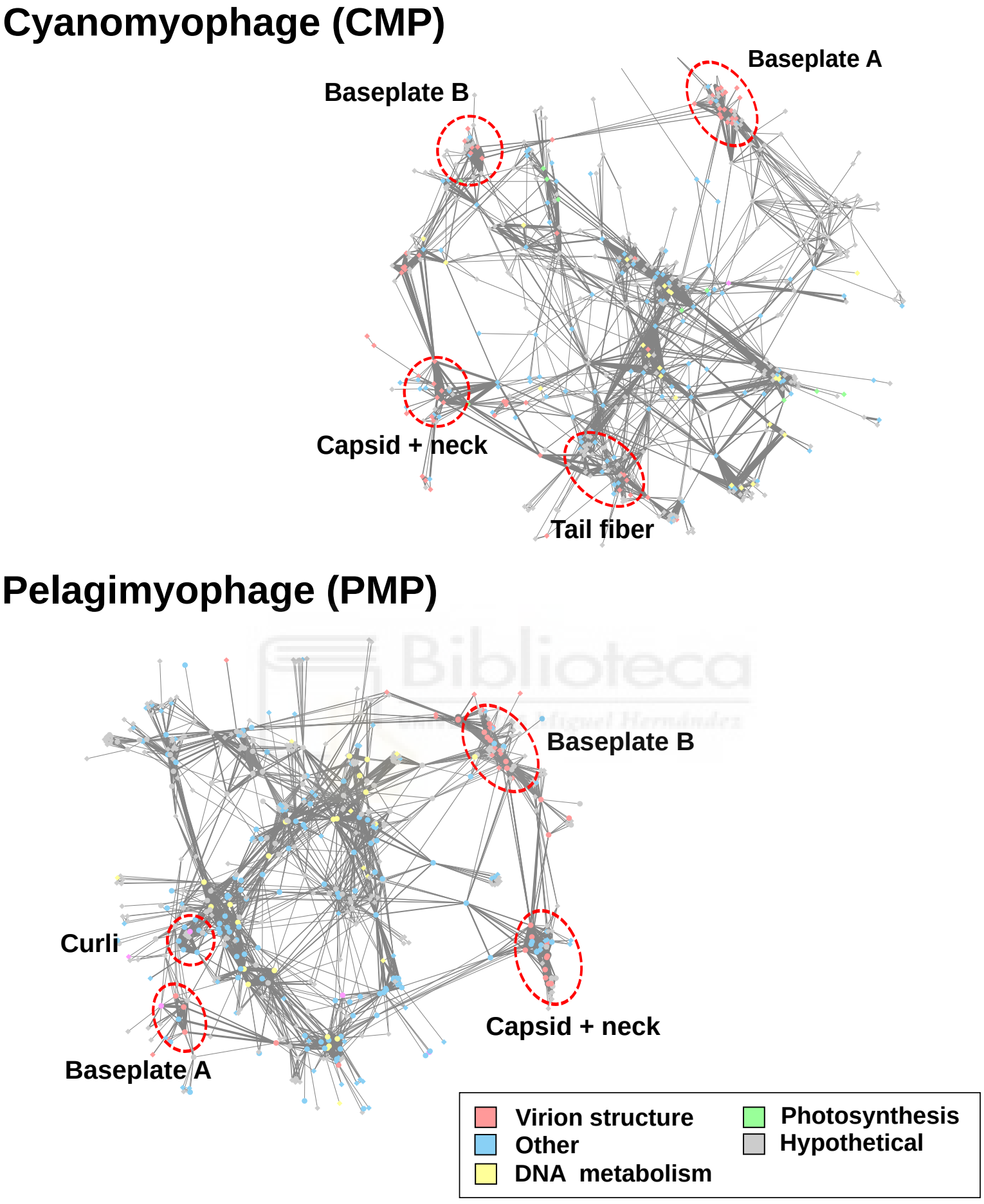
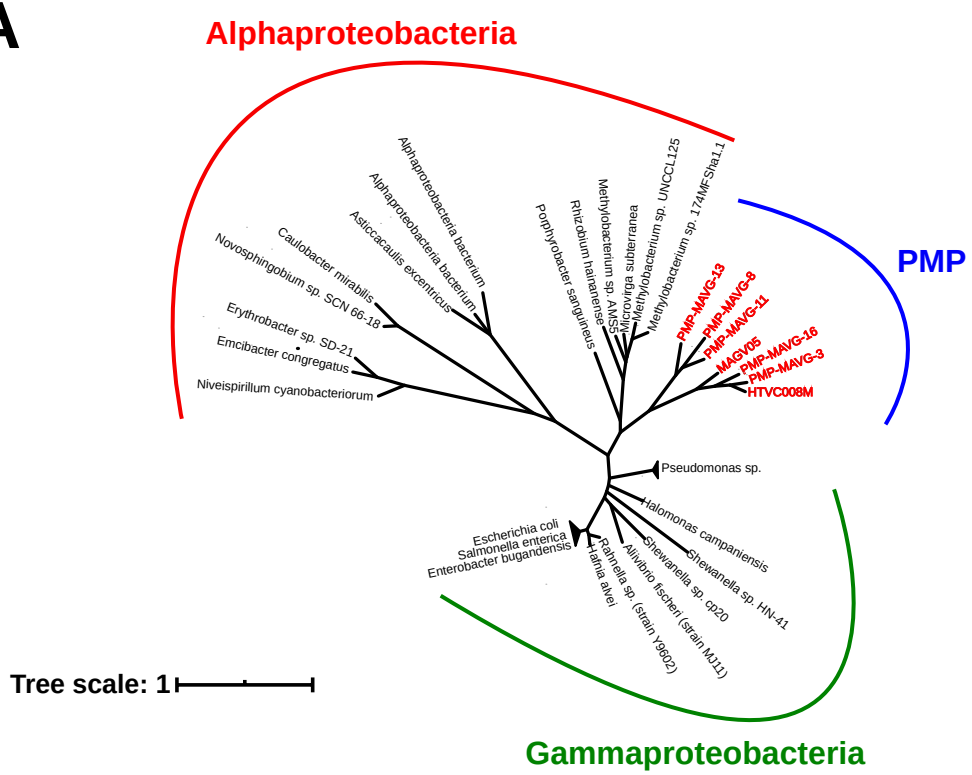


Fig. S6. Co-ocurrence networks of Cyanomyophage (CMP) and Pelagimyophage (PMP) genomes. Each node represents a Gene cluster (GCs), while each edge indicates that the two nodes it connects are present in the same operon in at least a pair of genomes. Edge thickness indicates the number of genomes the two GCs are present in the same operon. Each node is colored according to its predicted function.

A

Curli production assembly/ transport component (CsgG)



B

Curli production assembly/ transport component (CsgF)

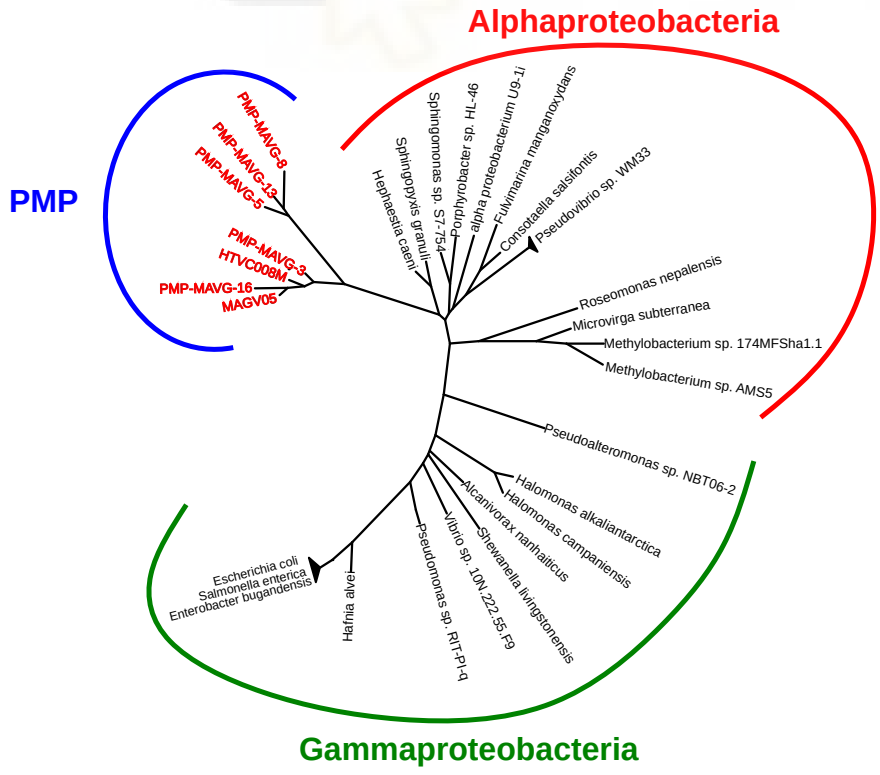


Fig. S7. Phylogenetic trees of two curli operon proteins (csgF, csgG) from Pelagimyophages (PMP), Alphaproteobacteria and Gammaproteobacteria. PMP genes are highlighted in red.

Table S1. Metagenomic databases samples utilized in this study.

Name	Type	Habitat	Reference	Raw size (Gbp)1	Number of contigs >5kb
GEOTRACES	Metagenome	Marine	[1]	7,152	703,045
HOT / BATS Time Series	Metagenome	Marine	[1]	2,206	225,141
Malaspina	Metagenome & Virome	Marine	[2]	1,302	102,569
IMG/M (Aquatic metagenome subset)[2]	Metagenome & Virome	Marine & Freshwater	[3]	13,811	1,196,310
TARA	Metagenome & Virome	Marine	[4], [5]	8,611	851,642
Mediterranean contig collection	Metagenome & Virome	Marine	[6], [7]	1,701	108,453
Global Oceanic Virome (GOV)[2]	Virome	Marine	[8]	1,046	109,862
IMG/VR[2]	Metagenome & Virome	Marine & Freshwater	[9]	11,837	715,672

(1) Gbp = 1,000,000 bp.
(2) Contigs from these datasets were not assembled in house but are instead the ones provided in their respective repositories.

REFERENCES

[1] Biller SJ et al. "Marine microbial metagenomes sampled across space and time." Scientific Data. 2018 Sep 4;5:180176.
[2] Silva G. Acinas et al. "Metabolic Architecture of the Deep Ocean Microbiome." bioRxiv 635680; doi: https://doi.org/10.1101/635680.
[3] Chen, I-Min A et al. "IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes." Nuc. acids res. 2019 Jan 8;47(D1):D666-D677.
[4] Pesant S et al. "Open science resources for the discovery and analysis of Tara Oceans data." Scientific Data. 2015 May 26;2:150023.
[5] Alberti A et al. "Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition." Scientific Data. 2017 Aug 1;4:170093.
[6] Mizuno CM et al. "Expanding the marine virosphere using metagenomics." PLoS Genetics. 2013;9(12):e1003987.
[7] López-Pérez M et al. "Genome diversity of marine phages recovered from Mediterranean metagenomes: Size matters." PLoS Genetics. 2017 Sep 25;13(9):e1007018.
[8] Roux S et al. "Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses." Nature. 2016 Sep 29;537(7622):689-693.
[9] Paez-Espino D et al. "IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes." Nuc. acids res. 2019 Jan 8;47(D1):D678-D686

Table S4. Protein clusters (PC) with size >10 genes and a log fold difference of at least 1. Normalized gene counts were calculated as (Total number of genes in all genomes of the group / total genomes in the group).

Protein ID	Normalized protein count (Norm. PMP)	Normalized protein count (Norm. CMP)	Log(Norm. PMP/Norm. CMP)	Category
Putative lipase	0.37	0.00	5.00	Lipid biosynthesis
30S Ribosomal protein S21	0.83	0.00	5.00	DNA translation
5'(3')-deoxyribonuclease	0.60	0.00	5.00	Nucleotide biosynthesis
Adenine-specific RNA methyltransferase	0.37	0.00	5.00	Methyltransferase
Adenylate kinase	0.37	0.00	5.00	Nucleotide biosynthesis
Adenylate kinase adk	0.37	0.00	5.00	Regulation
Adenylate/Guanylate cyclase catalytic domain w/ CHASE2 domain	0.60	0.00	5.00	Regulation
Baseplate hub subunit gp5	0.60	0.00	5.00	Structural/Baseplate
Cell wall hydrolase	0.43	0.00	5.00	Hydrolase activity
Curlin-associated protein	0.77	0.00	5.00	Cell envelope
DNA end protector protein	0.87	0.00	5.00	DNA packaging
FeCR domain-containing protein	0.60	0.00	5.00	Substrate transport
Glycosyltransferase, family 8	1.03	0.00	5.00	LPS biosynthesis
Glycosyltransferase, family 8	0.60	0.00	5.00	LPS biosynthesis
Iron-Sulfur cluster assembly accesory protein IscA	0.53	0.00	5.00	Fe-S cluster
Iron-Sulfur cluster assembly accesory protein IscU	0.40	0.00	5.00	Fe-S cluster
Molybdenum cofactor biosynthesis protein MoaA	0.60	0.00	5.00	Radical SAM
Peptide-modifying protein SuiA/SuiB	0.57	0.00	5.00	Protein modification
Radical SAM protein + SPASM domain	0.50	0.00	5.00	Radical SAM
T4-like lysozyme	0.43	0.00	5.00	Hydrolase activity
tRNA(Ile)-lysidine synthetase	0.40	0.00	5.00	DNA translation
Type VIII secretion system (T8SS), CsgF protein	0.70	0.00	5.00	Cell envelope
Type VIII secretion system (T8SS), CsgG protein	0.70	0.00	5.00	Cell envelope
VrlC Protein	0.80	0.03	4.68	Structural/Baseplate
Cytidylyltransferase	0.63	0.03	4.34	Other
C-5 cytosine-specific DNA methylase	0.33	0.03	3.42	Methyltransferase
Peptide deformylase	0.60	0.06	3.26	Protein modification
Baseplate wedge subunit gp53	0.87	0.13	2.79	Structural/Baseplate
Ribonuclease H RnaseH	0.77	0.13	2.62	DNA replication
collagen triple repeat domain-containing protein	0.30	0.06	2.26	Cell envelope
Peroxiredoxin-like protein	0.50	0.13	2.00	Cell homeostasis
Tail protein w/ Immoglobulin fold	0.73	0.28	1.38	Structural/Fiber
SprT-like protein	0.43	0.19	1.21	Regulation
2OG-Fe(II) oxygenase superfamily protein	0.27	0.13	1.09	2OG-Fe(II) superfamily
Endonuclease YncB-like	0.40	0.81	-1.02	DNA replication
2OG-Fe(II) oxygenase superfamily protein	0.13	0.28	-1.08	2OG-Fe(II) superfamily
Glutaredoxin	0.63	1.38	-1.12	Cell homeostasis
Ferredoxin-dependent bilin reductase	0.10	0.22	-1.13	Iron metabolism
Protein structurally similar to Sm/Lsm-like RNA-binding proteins	0.77	1.75	-1.19	Regulation
Sm-like domain-containing protein	0.53	1.25	-1.23	Regulation
DNA methylase	0.10	0.25	-1.32	Methyltransferase
Tail tube protein	0.53	1.34	-1.33	Structural/Tail
2OG-Fe(II) oxygenase superfamily protein	0.13	0.34	-1.37	2OG-Fe(II) superfamily
Rnf-Nqr	0.13	0.34	-1.37	Nitrogen metabolism
2OG-Fe(II) oxygenase superfamily protein	0.70	1.88	-1.42	2OG-Fe(II) superfamily
2OG-Fe(II) oxygenase superfamily protein	0.10	0.28	-1.49	2OG-Fe(II) superfamily
Phosphate starvation-inducible protein PhoH	0.30	0.97	-1.69	Regulation
Prolyl-4 hydroxylase	0.13	0.44	-1.71	2OG-Fe(II) superfamily
Prolyl-4 hydroxylase	0.17	0.59	-1.83	2OG-Fe(II) superfamily
FAD-dependent thymidylate synthase ThyX	0.20	0.94	-2.23	Nucleotide biosynthesis
PKHD-type hydroxylase	0.07	0.31	-2.23	2OG-Fe(II) superfamily
Baseplate tail tube cap	0.13	0.66	-2.30	Structural/Tail
2OG-Fe(II) oxygenase superfamily protein	0.20	1.09	-2.45	2OG-Fe(II) superfamily
Ferrochelatase	0.07	0.41	-2.61	Iron metabolism
Baseplate wedge tail fiber connector gp9/gp10	0.07	0.50	-2.91	Structural/Baseplate
Cytitidyltransferase	0.07	0.78	-3.55	Other
6-phosphogluconate dehydrogenase	0.00	0.44	-5.00	Energy metabolism
ABC-type phosphate transport system, periplasmic component PstS	0.00	0.44	-5.00	Substrate transport
Antenna protein CpeT-like	0.00	0.75	-5.00	Photosynthesis
Baseplate wedge initiator w/ Concavalin A-like domain + YHYH domain	0.00	0.47	-5.00	Structural/Baseplate
Baseplate wedge initiator w/ YHYH domain	0.00	0.44	-5.00	Structural/Baseplate
Baseplate wedge subunit gp53	0.00	0.78	-5.00	Structural/Baseplate
cAMP phosphodiesterase	0.00	0.44	-5.00	Regulation
CP12 domain-containing protein	0.00	0.84	-5.00	Energy metabolism
DNA adenine methylase dam	0.00	0.91	-5.00	Methyltransferase
fructose-6-phosphate aldolase TalC	0.00	0.91	-5.00	Energy metabolism
Glucose 6-phosphate dehydrogenase	0.00	0.34	-5.00	Energy metabolism
High light inducible protein	0.00	0.44	-5.00	Photosynthesis
High light inducible protein	0.00	1.53	-5.00	Photosynthesis
LlaGI endonuclease	0.00	0.69	-5.00	DNA replication
Major outer membrane protein OMP1	0.00	0.97	-5.00	Cell envelope
MazE protein	0.00	0.47	-5.00	Regulation
MazG protein	0.00	0.88	-5.00	Other
PA14 domain-containing protein	0.00	0.31	-5.00	Other
Phage tail lysozyme	0.00	0.47	-5.00	Structural/Tail
Photosystem II protein, D1/D2 subunit	0.00	1.59	-5.00	Photosynthesis
Plastocyanin	0.00	0.63	-5.00	Photosynthesis
Plastoquinol terminal oxidase	0.00	0.53	-5.00	Electron transfer
S-adenosylmethionine decarboxylase speD	0.00	0.38	-5.00	Photosynthesis
Tail fiber protein	0.00	0.47	-5.00	Structural/Fiber
Tail fiber protein	0.00	0.50	-5.00	Structural/Fiber
Tail spike protein gp5	0.00	0.50	-5.00	Structural/Baseplate
VrlC protein	0.00	0.75	-5.00	Structural/Baseplate

Table S5. AMGs detected in the analysed genomes

AMG Name	Function	PMP	CMP
Carbamoyltransferase	Nucleotide metabolism	X	X
CobS	Cobalamin biosynthesis	X	X
CobT	Cobalamin biosynthesis	X	X
Ferrochelatase	Heme biosynthetic pathway	X	X
Heme oxygenase	Heme degradation	X	X
Heat shock protein hsp20	Helps refold proteins in stressful conditions	X	X
Pyrophosphatase MazG	Limits the effects of mazEF in response to aminoacid starvation	X	X
PhoH	Phosphate starvation regulon Pho	X	X
Tryptophan halogenase prnA	Degradation of aromatic compounds, antibiotic biosynthesis	X	X
ABC-type phosphate transport system PstS	Part of PstABC, involved in phosphate import	X	X
PurM	Purine biosynthesis de novo pathway	X	X
Thioredoxin	Redox protein, plays a role in many biological processes	X	X
Ribonucleotide reductase RNA	Provides precursors for DNA synthesis (NTP → dNTP)	X	X
Prolyl-4 hydroxylase	DNA repair	X	X
Peroxiredoxin	Oxidative stress control	X	X
Acyl carrier protein acpP	Lipid biosynthesis	X	X
Rnf-Nqr	Nitrogen fixation	X	X
Peptide deformylase	Protein maturation	X	X
Ferredoxin, ISC System	Small electron carrier	X	
Iron-Sulfur cluster assembly protein IscA	Scaffold protein for Fe-S cluster biosynthesis	X	
Iron-Sulfur cluster assembly protein IscU	Scaffold protein for Fe-S cluster biosynthesis	X	
Iron-Sulfur cluster assembly protein SufE	Fe-S cluster biosynthesis	X	
asparagine synthase asnB	Synthesis of asparagine from aspartate	X	
30S ribosomal protein S21	Protein translation, recognition of Shine-dalgarno sequence	X	
50S ribosomal protein L7/L12	Protein translation, binding site for several factors	X	
Cytochrome C	Electron transport	X	
Alternative oxidase AOX	Electron transfer from reduced ubiquinol to oxygen, forming water	X	
Stearoyl-CoA desaturase (Delta-9 desaturase)	Lipid biosynthesis	X	
beta-ketoacyl-acyl-carrier-protein synthase II FabF	Lipid biosynthesis	X	
Adenylate kinase adk	Nucleotide conversion (NTP + NMP → 2 NDP)	X	
Cold shock protein CpsA	Regulation, involved in RNA folding	X	
L-lactate permease	Substrate transport	X	
TonB-dependent vitamin B12 receptor	Substrate transport	X	
Vitamin B3 transporter PnuC	Substrate transport	X	
Tripartite tricarboxylate transporter, TctA family	Substrate transport	X	
L-Aspartate-alpha-decarboxylase	CoA biosynthesis	X	
6-phosphogluconate dehydrogenase gnd	Pentose phosphate pathway		X
cAMP phosphodiesterase	Alters gene expression of genes controlled by cAMP		X
CP12	Pentose phosphate pathway		X
cpeT chromophore lyase	Site-selective attachment of cromophores		X
Glucose-6-phosphate-1 dehydrogenase zwf	Pentose phosphate pathway		X
high light inducible protein	Collect energy from protons and transfer it to photosystems		X
Plastocyanin petE	Electron transport between photosystems		X
Ferredoxin petF	Small electron carrier, part of the photsynthetic system		X
Taurine catabolism dioxygenase TauD	Part of the photosystem II reaction center		X
Photosystem II D2 protein PsbD	Part of the photosystem II reaction center		X
Plastoquinol terminal oxidase PtoX	May have a role in mantaining the reduction state of electron transport chain components		X
PurC	Purine biosynthesis de novo pathway		X
PurH	Purine biosynthesis de novo pathway		X
PurL	Purine biosynthesis de novo pathway		X
PurN	Purine biosynthesis de novo pathway		X
PurS	Purine biosynthesis de novo pathway		X
PyrE	Purine biosynthesis de novo pathway		X
S-adenosylmethionine decarboxylase proenzyme SpeD	Polyamine biosynthesis		X
Transaldolase TalC	Pentose phosphate pathway		X
Taurine catabolism dioxygenase TauD	Sulfur salvage from Taurine		X

9. Annex 2

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

Long-Read Metagenomics Improves the Recovery of Viral Diversity from Complex Natural Marine Samples

Asier Zaragoza-Solas,[a] Jose M. Haro-Moreno,[a] Francisco Rodriguez-Valera,[a] Mario López-Pérez[a]

[a]Evolutionary Genomics Group, División de Microbiología, Universidad Miguel Hernández, San Juan, Alicante, Spain

ABSTRACT The recovery of DNA from viromes is a major obstacle in the use of long-read sequencing to study their genomes. For this reason, the use of cellular metagenomes (>0.2-μm size range) emerges as an interesting complementary tool, since they contain large amounts of naturally amplified viral genomes from prelytic replication. We have applied second-generation (Illumina NextSeq; short reads) and third-generation (PacBio Sequel II; long reads) sequencing to compare the diversity and features of the viral community in a marine sample obtained from offshore waters of the western Mediterranean. We found that a major wedge of the expected marine viral diversity was directly recovered by the raw PacBio circular consensus sequencing (CCS) reads. More than 30,000 sequences were detected only in this data set, with no homologues in the long- and short-read assembly, and ca. 26,000 had no homologues in the large data set of the Global Ocean Virome 2 (GOV2), highlighting the information gap created by the assembly bias. At the level of complete viral genomes, the performance was similar in both approaches. However, the hybrid long- and short-read assembly provided the longest average length of the sequences and improved the host assignment. Although no novel major clades of viruses were found, there was an increase in the intraclade genomic diversity recovered by long reads that produced an enriched assessment of the real diversity and allowed the discovery of novel genes with biotechnological potential (e.g., endolysin genes).

IMPORTANCE We explored the vast genetic diversity of environmental viruses by using a combination of cellular metagenome (as opposed to virome) sequencing using high-fidelity long-read sequences (in this case, PacBio CCS). This approach resulted in the recovery of a representative sample of the viral population, and it performed better (more phage contigs, larger average contig size) than Illumina sequencing applied to the same sample. By this approach, the many biases of assembly are avoided, as the CCS reads recovers (typically around 5 kb) complete genes and even operons, resulting in a better discovery of the viral gene diversity based on viral marker proteins. Thus, biotechnologically promising genes, such as endolysin genes, can be very efficiently searched with this approach. In addition, hybrid assembly produces more complete and longer contigs, which is particularly important for studying little-known viral groups such as the nucleocytoplasmic large DNA viruses (NCLDV).

KEYWORDS PacBio CCS long reads, bacteriophage, long-read sequencing, metagenome, viral diversity, virome

Marine viruses are the most abundant biological entities in oceanic marine environments, with an estimated population density of 10^7 per mL of seawater (1). It is therefore no wonder that they are critical drivers of ocean biogeochemistry, both via the release of organic matter as a by-product of their predation upon phytoplankton and heterotrophic bacteria (2, 3) (viral shunt) and via the manipulation of host metabolism during infection (4, 5).

Editor Julie A. Huber, Woods Hole Oceanographic Institution

Address correspondence to Mario López-Pérez, mario.lopezp@umh.es, or Francisco Rodriguez-Valera, frvalera@umh.es.

The authors declare no conflict of interest.

Received 24 February 2022
Accepted 11 May 2022

During the last 2 decades, the study of the viral community in marine environments has been driven by metagenomics, thanks to the advances in short-read (SR) sequencing (6, 7). However, the advent of long-read (LR) sequencing technology, spearheaded by Oxford Nanopore and Pacific Biosciences (PacBio), has the potential to solve major issues that have plagued SR sequencing-based studies for years, mainly the low recovery of both high-diversity microbes (8) and the flexible genome (9). Unfortunately, the high error rate derived from these technologies has delayed their application in metagenomics and requires the use of either a complementary short-read data set (10) or very high coverage (11) to correct these sequencing errors. The development of high-fidelity approaches such as PacBio circular consensus sequencing (CCS) with an error rate similar to that of the Illumina system opens a new avenue for the study of prokaryotic communities in their natural environments. An example of the advantages of this new technology can be found in a recent work, in which a well-known marine sample from the Mediterranean water column was analyzed using both SR and LR sequencing (12). Results suggested that PacBio Sequel II CCS is particularly suitable for cellular metagenomics due to its large read size and its low error rate. Reads in LR metagenomes are large enough to perform gene prediction directly, bypassing the biases inherent in the assembly process. The assembly step is also improved with this kind of sample by using hybrid assembly of LR and SR, allowing reconstruction of genomes, including the flexible genome and even streamlined genomes, such as those from *Pelagibacterales* (12).

The benefits of LR sequencing can be even more pronounced for the study of viruses, as the size of individual reads may be sufficient to recover complete genomes. There are already some examples of LR sequencing applied to the study of viromes using the Nanopore sequencing platform. Beaulaurier et al. recovered 1,864 new complete assembly-free virus genomes from three Nanopore data sets (11). On the other hand, Warwick-Dugdale et al. recovered around 2,500 viral contigs from the assembly of Nanopore and Illumina data sets from the same seawater sample of the western English Channel, showing that a hybrid or long-read-only assembly improved the recovery of viral contigs and their metaviromic islands compared to short-read assemblies (10). These results have been corroborated in other virome studies using the same technology (13, 14).

However, the study of viromes by LR sequencing is limited by the large amount of DNA required for this type of technology and the scarcity of viral DNA that can be collected from environmental samples. Therefore, as an alternative to the study of the virome, we used the viral DNA present in a cellular metagenome (>0.22-μm size range). A high presence of viral DNA (around 10% to 15%) in marine metagenomes has been reported (15). The vast majority of this viral DNA likely belongs to cells undergoing the lytic cycle, although other sources might be possible, including lysogenized viruses (either integrated or as a plasmid) or virions larger than the filter pore (>0.2 μm) (15). The aim of this study was to compare the efficiency of LR sequencing for the study of the viral community (with and without an assembly step) with the classical approach using Illumina (short reads).

RESULTS AND DISCUSSION

To evaluate the viral genomic diversity resolution power of LR metagenomics and compare it to that of SR sequencing, we analyzed a single marine sample from offshore Mediterranean waters during winter, when the epipelagic water column was mixed. The presence of replicating viruses inside cells during the lytic cycle produces a natural amplification that makes it possible to find abundant sequences of viral origin in the cell fraction of metagenomic samples. This sample was sequenced with Illumina and PacBio Sequel II systems and then assembled twice, first using only the Illumina short reads, resulting in the short-read assembly data set (SRa), and then in a hybrid assembly using both the Illumina short reads and the PacBio long reads, resulting in the long-read assembly data set (LRa). We decided on the hybrid assembly rather than a long-read-only assembly based on previous results (12). In order to evaluate the possible biases introduced by the assembly process, we also analyzed the PacBio CCS15

TABLE 1 Summary statistics of viral sequence recovery for the short-read assembly (SRa), long-read assembly (LRa), and raw-read (LR) data sets

Statistic	Illumina assembly (SRa)	PacBio assembly (LRa)	PacBio CCS15 reads (LR)
Starting sequences	149,018	19,982	1,535,891
Putative phages (VIBRANT)	10,979	947	50,296
95% identity clustering	10,979	947	42,156
Unique sequences[a]	5,886	36	30,203
Nucleotides sequenced (Gb)	23.4	31.0	7.6
Unique sequences/Gbp sequenced	251.53	1.16	3,974
Unique sequences (versus GOV2)[b]	4,196	35	26,766
No. complete (high quality)[c]	9 (53)	15 (114)	0 (27)
Min–max sequence length (bp)	1,000–188,349	1,353–428,169	1,011–17,836
Avg sequence length (bp)	4,906	32,260	5,261
Min–max GC content (%)	19.40–65.25	19.56–69.93	14.25–86.03
Avg GC content (%)	35.45	36.9	38.13
Total proteins[d]	80,487	41,599	330,157
Unique terminase (*terL*) proteins	30	2	393
Avg proteins/sequence	7.33	43.92	7.83
Avg protein length (aa)	190.29	223.42	177.9

[a]Sequences not present in the other data sets (BLASTN, 95%; coverage of at least 70% of the smallest sequence).
[b]Sequences not present in the other data sets or the Global Ocean Virome 2.0 (BLASTN, 95%; coverage of at least 70% of the smallest sequence).
[c]VIBRANT defines a high-quality sequence as one that likely contains the majority of a virus's complete genome (~70% completeness).
[d]Values shown here represent protein numbers after dereplication (CD-HIT, 95% identity).

reads (PacBio consensus reads created by comparing at least 15 subreads [LR]) before assembly.

Viral sequence recovery and statistics. First, we wanted to compare the efficiency of viral sequence recovery between the three data sets (Table 1). The first step in the preprocessing pipeline was to run VIBRANT (16) for all sequences >1 kb to identify those in each data set that were of viral origin. Viral sequences turned out to be quite numerous in both data sets, with 5% of the total sequences from the SRa and LRa and 2.5% of the LR data set classified as viral contigs. After a step of clustering at 95% sequence identity to remove redundant reads from the LR data set, we recovered a total of 54,082 putative viral sequences (10,979 in the SRa, 947 in the LRa, and 42,156 in the LR) (Table 1). In order to assess if the different assembly methods recovered the same viral community, we identified unique sequences in each data set by comparing the three data sets against each other (see Materials and Methods). Most sequences from the LRa were also found in the SRa, with only 36 unique LRa contigs. Remarkably, while the SRa data set contained a fair number of unique sequences (5,886), most of the unique sequences were found in the LR data set (30,203; 71% of total viral LR sequences), revealing a large genomic diversity not recovered by the assemblies. This diversity gap was also present when results for a marker gene, such as that encoding the terminase large subunit (*terL*), were compared, with the LR data set containing 393 unique terminase genes (clustering at 95% amino acid identity), compared to 30 and 2 in the SRa and LRa data sets, respectively. The GC content showed a slight (effect size = 0.022) but significant (Kruskal-Wallis test, P value $< 10^{-15}$) skew toward high GC values when PacBio CCS reads were added to the data sets (Table 1). The SRa data set presented an average GC content of 35.45% compared to 36.9% for the LRa and 38.13% for the LR (Table 1). This bias could arise from the fact that assemblies usually recover only the core genome. In this sample (marine surface water), clade SAR11 is the most abundant organism (12), with an average GC content of 34%. LRs recover more of the flexible genome, which can present GC fluctuations compared to the core and would thus explain this variation from 34% to 38%. Regarding sequences shared between the three data sets, Table S1 shows the relationship between contigs that were considered part of the same phage (identity over 95%, 70% overlap of the

smallest contig). When comparing the ratio of recovered sequences between SRa and the combined LRa and LR data sets for shared ones, we found that in 2,463 of 3,316 shared instances (ca. 75%), the LR data sets contained longer contigs than their SRa counterpart (Table S1). These results show that the use of long reads in assembly result in larger contigs than assembly with only SR.

Next, we were interested in assessing if this novel diversity had been captured by previous studies, so we compared the three data sets against the Global Ocean Virome 2 (GOV2) (17), the largest database of seawater phages to date (195,728 marine populations, containing 6,685,706 proteins). This data set was created from viromes obtained from 145 samples from the Malaspina (18), *Tara* Oceans (6) and *Tara* Arctic (17) expeditions, therefore representing marine phage communities from different environments all around the world. We found 30,997 viral sequences in our whole data set (SRa, LR, and LRa) not found in GOV2, with the vast majority (26,766) of these unique sequences belonging to the LR data set.

Regarding size and completeness, the hybrid PacBio LRa resulted in the largest viral contigs, with a maximum size of 428,169 bp and an average contig size of 32,260 bp (Table 1). We recovered 24 complete phage genomes (based on circular redundancy at the ends) from both assembled data sets (15 in the LRa and 9 in the SRa). As expected, due to their small estimated average size (ca. 5 kb), we were unable to recover any complete genomes directly from the LRs. However, we can make an estimated guess of the quality of the remaining contigs using VIBRANT's quality statistics, which classify contigs based on the estimated completeness of the genome. When we considered only contigs marked as high quality (70% of the estimated phage genome), we found that only 53 (0.4%) of the SRa contigs belonged to this category, while in the LRa data set there were 114 (12.5%) (Table 1). Some complete phage genomes were shared by the LRa and SRa data sets. The SRa contigs resulted in a maximum contig size approximately half of that found in the LRa (188,349 bp), with an average contig size on a par with the LR data set, more than six times smaller than the average in the LRa (ca. 32 kb) (Table 1). These results, together with the facts that the average protein size in all three data sets is similar and the number of proteins recovered from the LR data set is an order of magnitude larger than those from the assembled data sets, suggest that PacBio CCS15 reads could be used for viral protein calling without the need for assembly, as previously stated (12).

Putative host prediction. An important part of the biological significance of viruses depends on knowledge of the host they infect. We attempted to assign a host to contigs in all three data sets (SRa, LRa, and LR). To this end, phage contigs were classified against the RefSeq database. We assigned hosts to each sequence following the method described by Beaulaurier et al. (11), which was applied to phages obtained by Nanopore sequencing. The method is based on protein homology against a reference database, assigning a host to a sequence based on the number of best hits (see Materials and Methods). Figure 1A shows the results at a 3-protein threshold, including all contigs from a data set and those unique to their specific data set (dereplicated) and before dereplication. Considering the SRa and LR data sets, both unique and dereplicated variants presented a similar host assignment rate (ca. 30%) with no differences at the taxonomic level, suggesting that the differences between the two data sets could be beyond the order level. As might be expected, *Alphaproteobacteria* and *Cyanobacteria* were the most abundant hosts in all three data sets, as they were the most abundant groups in the sample (12) and were also the most represented in the reference databases (Fig. 1A). The recent addition of various *Methylophilales* (19) and *Flavobacteria* (20) phage genomes to the reference databases has resulted in a highly increased *Gammaproteobacteria* and *Flavobacteria* phage count compared to previous analysis of the Mediterranean virome (7).

The LRa data set provided the highest rate of host assignment. In the nondereplicated sample, almost 75% of the contigs had a host assigned, compared to a 30% rate for the LR and SRa data sets. This is probably due to the fact that they were, on

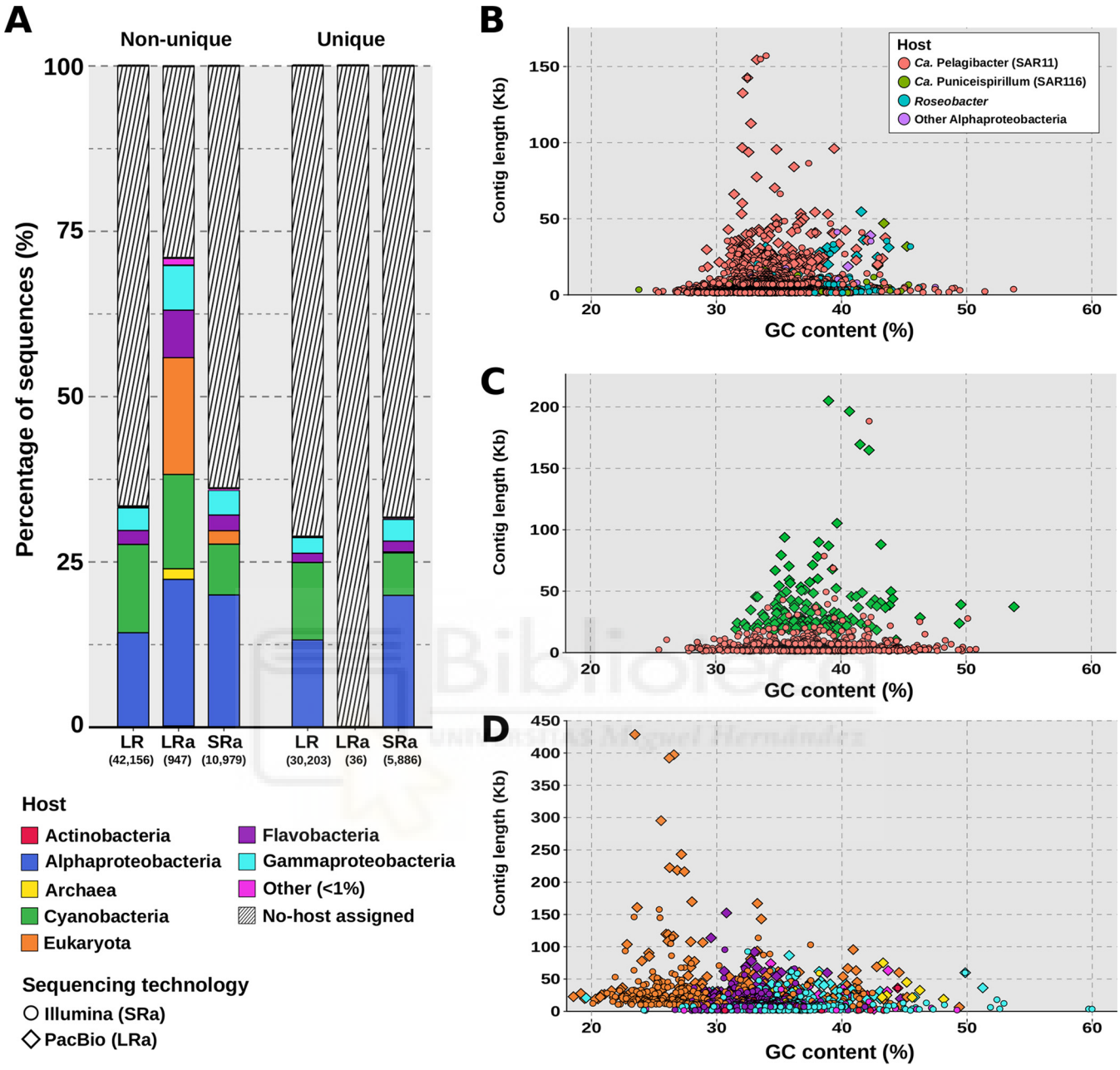


FIG 1 (A) Taxonomic affiliations of viral contigs expressed in percentages, separated into those found in that data set (non-unique) and those unique to that data set (unique). The number in parentheses below each bar is the number of contigs in that category. (B) Distribution of assembled viral contigs that infect *Alphaproteobacteria* by contig length and GC content. Circles represent short-read assemblies (Illumina), while diamonds represent hybrid assemblies (PacBio + Illumina). Shapes are colored according to their host. (C) Distribution of assembled viral contigs that infect *Cyanobacteria* by contig length and GC content. Orange circles represent short-read assemblies (Illumina), while green diamonds represent hybrid assemblies (PacBio + Illumina). (D) Distribution of viral contigs by contig length and GC content. Circles represent short-read assemblies (Illumina), while green diamonds represent hybrid assemblies (PacBio + Illumina). Shapes are colored according to their host.

average, larger contigs and as such contain more information to reliably assign a host. However, we were unable to assign a host to any of the 36 unique sequences in the LRa data set (3.2% of the total). Host taxonomy was similar to that seen in the previous data sets, the main difference being an increase in eukaryotic and archaeal viruses (20% of total contigs), mainly marine group I *Thaumarchaeota* (marthavirus) (21).

Comparison between the sequences obtained by assembly (LRa and SRa) also revealed differences between the viral groups. As a general rule, LRa contigs were on

average larger than their SRa counterparts, even if the latter can result in similar maximum sizes. For example, in alphaproteobacterial phages (Fig. 1B), we recovered 52 sequences over 30 kb in the SRa data set, compared to 126 in the LRa data set. We found a similar case for the cyanophages (Fig. 1C), where 14 sequences were over 50 kb in the SRa data set compared to 68 sequences in the LRa data set. The nucleocytoplasmic large DNA viruses (NCLDV, proposed order *Megavirales*) (Fig. 1D) deserve special attention, as their assemblies in the LRa data set were larger and more numerous (24 sequences over 20 kb, including the largest contig of 428 kb) than those in the SRa data set (21 sequences, 2 over 50 kb; maximum size, 61 kb). We believe this might be due to the fact that eukaryotic genomes have many repeats and other features that make their assembly from short-read metagenomes less efficient (22).

To analyze the phylogenomic diversity of the NCLDV sequences found, we used only sequences that contained five key markers highly conserved in this type of virus: the major capsid protein (MCP), the DNA polymerase beta subunit (PolB), the DEAD/SNF2-like helicase SFII, the poxvirus late transcription factor VLTF3, and the packaging ATPase A32 (23). Figure S1 shows a phylogenetic tree based on a concatenation of these five proteins, including reference genomes from RefSeq and the collection of 444 marine NCLDV Metagenome-Assembled Genomes from the work of Moniruzzaman et al. (23). The tree shows that these new eukaryotic sequences fall in the family *Mimiviridae* (16 sequences) and the family *Phycodnaviridae* (8 sequences).

Relative abundance in marine samples. Next, we wanted to analyze whether all the diversity found only in the LR data set was abundant and representative in nature. For that reason, we performed a recruitment analysis of SRa, LRa, and dereplicated LR viral sequences against the entire *Tara* Oceans metagenome data set (24). We considered a sequence present in a metagenomic sample if the sequence recruited at least five reads per kilobase of sequence and gigabase of metagenome (RPKG), with an identity of 95% and a contig coverage of 50%. The results are shown in Fig. 2. Although pelagiphages and cyanophages (viruses that infect "*Candidatus* Pelagibacter" and *Cyanobacteria*, respectively) show a similar abundance, they present different patterns of recruitment. The most cosmopolitan phages are cyanophages, particularly those that infect the genus *Prochlorococcus*. On the other hand, pelagiphages show a more endemic distribution, especially pelagimyophages, which tend to appear in only a few stations at a time (in this case, as could be expected, in the *Tara* stations in the Mediterranean), while pelagipodophages tend to appear in more stations (Fig. 2). In each of the plots, the recruitment means for each data set were represented as a line, showing that in all three cases (*Alphaproteobacteria*, *Cyanobacteria*, and other phages), the sequences recovered by LR prior to assembly are significantly more abundant than their assembled counterparts (Wilcoxon rank sum test, P value $< 10^{-5}$). Furthermore, this difference in RPKG was accentuated when comparison was made with phages that infect taxa which are typically difficult to assemble, such as those infecting *Alphaproteobacteria* (8). These results suggest that the dereplicated (nonredundant) LR sequences represent an untapped and abundant reservoir of genomic diversity.

Since the phage sequences were obtained from the cell fraction, we were interested to know if they were abundant and could also be recovered in the viral fraction. To that end, we recruited all phage data sets in metagenomes and viromes at different depths obtained at the same location from which the sample was collected (7, 25). When comparing the recruitment values in both types of samples (Fig. 2B), we observed that the vast majority of sequences accumulated significantly more in the viral fraction at the three depths surveyed (Wilcoxon rank sum test, P value $<10^{-16}$, for all three depths). Therefore, we can confirm that the phage genomes recovered from the cellular fraction are representative of the community found in the virion fraction as well and therefore represent a valid method for recovering the viral diversity of a sample.

New diversity recovered from LR. Once we discovered that there is a large amount of viral sequences in LR that is not contained in the other data sets (probably lost in the assembly process) and that is abundant in nature, we decided to analyze this diversity. Given that there is no universal marker for analyzing viral diversity, we used a number of

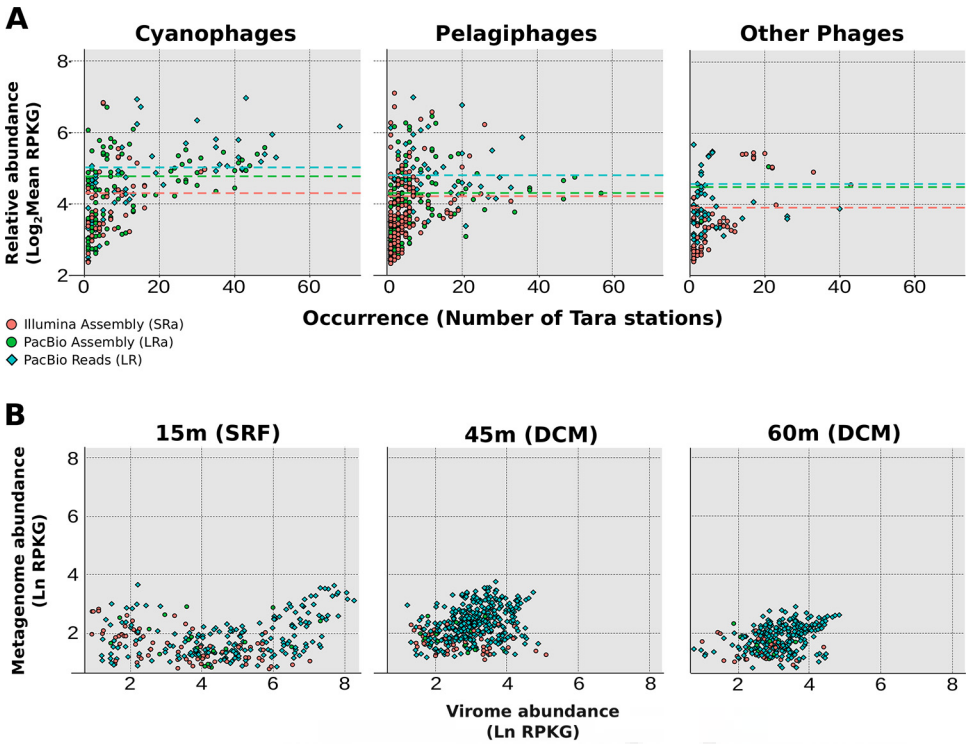
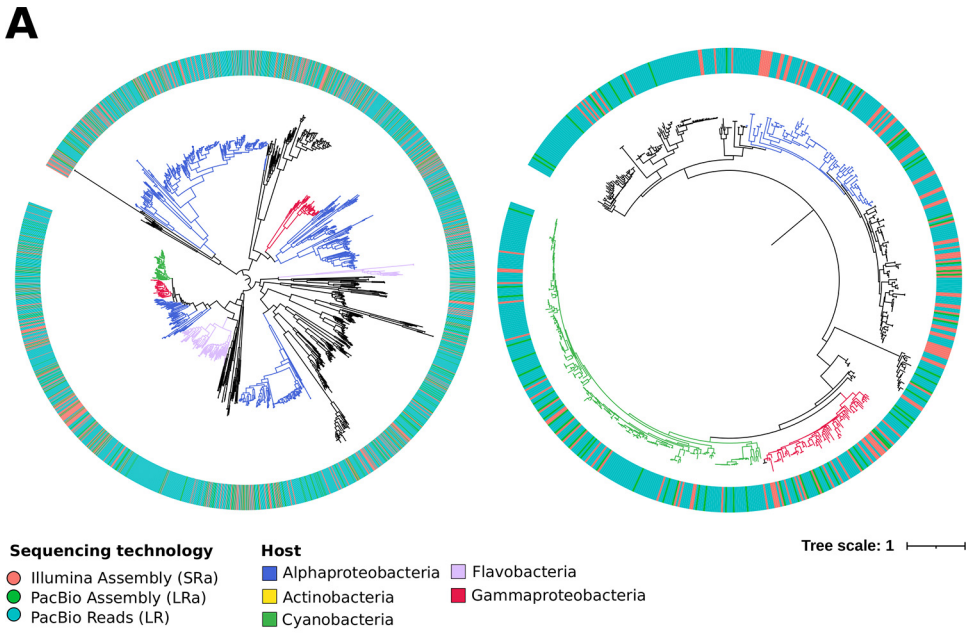


FIG 2 (A) Relative abundance of viral sequences measured by their recruitment values in metagenomes from *Tara* Oceans expeditions for cyanophages, pelagiphages, and other phages. The *x* axis shows the number of *Tara* stations where the contig accumulated over the coverage thresholds, while the *y* axis shows the combined recruitment value (in RPKG). Circles represent contigs derived from assembly (green for hybrid assembly, orange for Illumina assembly), while blue diamonds represent raw PacBio reads. (B) Relative abundance of viral sequences in viromes (*x* axis) and metagenomes (*y* axis) obtained from the same sample at 15, 45, and 60 m, measured in ln RPKG. Circles represent contigs derived from assembly (green for hybrid assembly, orange for Illumina assembly), while blue diamonds represent raw PacBio CCS15 reads. SRF, surface; DCM, deep chlorophyll maximum.

different phage-specific markers (large terminase subunit [*terL*], replicative DNA helicase [*dnaB*], tail tube protein, major capsid protein, and spanin) as well as several well-characterized auxiliary metabolic genes (AMGs) (thymidylate synthase [*thyX*], phosphoheptose isomerase [*gmhA*], ribonucleoside-diphosphate reductase [*nrdA*], ribonucleotide reductase large subunit, and phosphate starvation-inducible protein [*phoH*]).

We analyzed the diversity of these markers in the same sample for the three data sets by building phylogenetic trees (Fig. 3A; Fig. S2 and S3) and also by comparing the dereplicated sequence distribution with GOV2 (Fig. 3B; Table S2). The phylogenetic trees showed that none of the clades were composed only of LR-unique proteins, so we can conclude that the unique sequences recovered from the LR data set belong not to novel phage taxa but to known clades. Comparing the distribution of unique proteins between our three data sets, the LR data set usually contained more unique sequences by an order of magnitude compared to the assembled data sets (Table S2). Moreover, the percentage of unique variants was always higher in the LR.

After including the GOV2 data set in the comparison, it quickly became apparent that this data set contained most of the unique sequences (ca. 90% of all unique proteins). This was expected, considering the vast size and breadth of sampling of the GOV data set (144 samples); it was therefore surprising that a data set derived from a single sample contains a tenth of the diversity, especially considering that the 10 proteins selected are conserved proteins in phage genomes. Out of this slice of diversity, the vast majority of the unique contigs derive from the unassembled LR data set, as seen in the case of DnaB (149 different proteins versus 26 in the assembled data sets) and RrdA (150 versus 19 in assembled data sets) (Table S2).



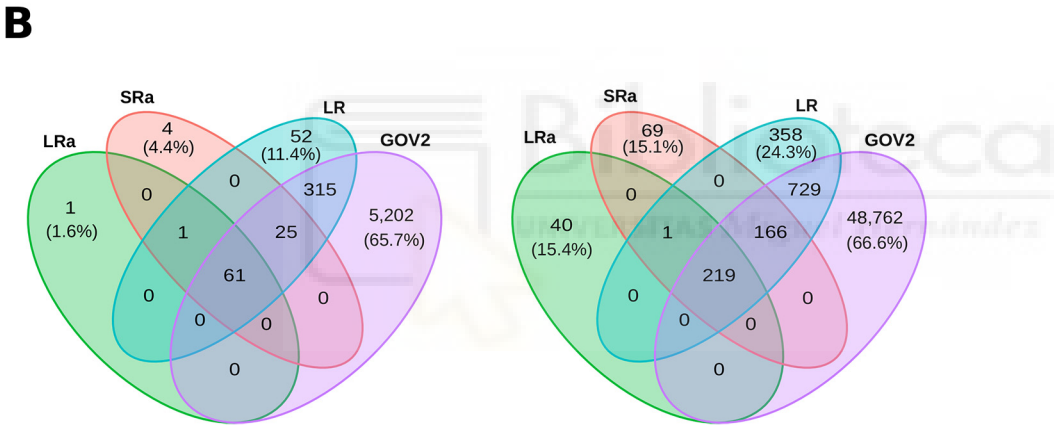


FIG 3 (A) Phylogenetic trees based on the terminase large subunit (TerL) and thymidylate synthase (PhyX). Branches are colored according to the assigned host, while the color of the outer circle indicates the data set the contig was obtained from (orange for Illumina assembly, green for PacBio assembly, blue for PacBio CCS15 reads). (B) Venn diagrams showing shared and unique sequences among the three data sets and GOV2 for the terminase large subunit (TerL) and thymidylate synthase (PhyX). The number inside each intersection leaf indicates the number of proteins shared by those data sets. In the unique section for each data set, the number in parentheses is the percentage of unique proteins in that data set compared to the total.

It is important to emphasize that the fact that LRs do not reveal novel phage clades does not mean that their novelty is not relevant. An example of this would be the endolysins, a remarkably diverse group of catalytic enzymes that degrade the cell wall of the host so that the phage progeny can escape (26). In recent years, these proteins have awakened increased interest for their potential to be used as antimicrobial agents (27, 28). Culture-free approaches have been applied to great effect in order to broaden the diversity of endolysins. In a previous study (29), 2,628 putative endolysins were retrieved from a collection of 183,298 assembled viral genomes, pooled from a variety of metagenomic data sets. We applied the same pipeline to our samples to evaluate if this novel diversity found by LR would also apply to proteins with more diversity than the usual protein markers.

We recovered 335, 106, and 841 putative endolysins from the SRa, LRa, and LR data sets, respectively, yielding a total of 1,216 new sequences. A phylogenetic tree of the sequences (Fig. S4) reveals that although most of the sequences are distributed among previously described endolysin groups, there were four clades not found in the

previous endolysin environmental collection, which we name C1 to C4. An analysis of their domains revealed them to be glycoside hydrolases from families 24, 104, 23, and again 24, respectively. These are lytic transglycosylases that have the well-known $\alpha+\beta$ lysozyme (30) fold, with differences in activity and specificity thought to be determined by the environment surrounding the active site. Each family includes several well-characterized phage lysozymes. No domains related to cell wall binding were found. Interestingly, the C4 clade contains a signal-arrest-release motif, a mechanism not reported in the original data set (29). This motif first directs the endolysin to the periplasm by first attaching it to the membrane, where it remains inactive until it is released as a soluble active enzyme in the periplasm (31). No other domains related to protein export or cell wall binding were found.

Functional characterization. Finally, our last question was if there was any functional category more enriched in the LR data set than in the assemblies. To answer this, we analyzed the protein content at the level of functionality, annotating the proteins against the KEGG (32) and Conserved Domain Database (CDD) (33). Then we compared the number of proteins with each annotation in the LR data set against the proteins found in the assembled data sets. The LR data set was particularly enriched in repeat-containing proteins, such as MORN repeats (37 times higher in LR than in the assembled data sets), pentapeptide repeats (26 times higher), ankyrin repeats (10 times higher), and Kelch repeats (9 times higher). Pentapeptide and Kelch repeats are widespread through bacterial and viral proteins (34, 35), ankyrin repeats have been found in a novel AMG which protects the infected bacteria from eukaryotes (36), and MORN repeats have been found in bacteriophage endolysins (37). The appearance of these proteins was not surprising, as repeats are the main cause of fragmented assemblies (38). A similar argument could be made for the prevalence of integrases (18 times higher), reverse transcriptases (not found in the assembled data sets) and transposases (9 times higher). Although these proteins are widespread in phage genomes (39–41), they present a large amount of microdiversity, which is also difficult for assemblers to solve (12). No groups of proteins were noticeably less abundant in LR than its assembled counterparts. These results suggest that long reads can help recover parts of the viral genome that are difficult to retrieve due to assembly bias.

Conclusions. The results obtained here demonstrate that it is possible to recover a representative sample of the viral community fraction of the viral community from the cellular fraction using LR sequencing approaches (e.g., PacBio Sequel II with CCS). This has already been observed with Illumina data sets (15), but the benefits of this approach improve with LR sequencing. The amount of DNA required for a PacBio run is at least an order of magnitude larger than that required for Illumina sequencing, and considering that DNA extraction from the viral fraction is an arduous process, requiring a large amount of sample as well as specialized equipment, studying these recovered viral genomes within the cell size fraction (e.g., $>$0.2 μm) may be a good alternative. The benefits of LR sequencing for the study of viral sequences are important even compared with the already-proven advantages for cellular metagenome analysis (12). CCS15 long reads are the equivalent of the average Illumina contig both in terms of length and reliability and therefore allow similarly reliable gene calling and protein identification.

We have also revealed that viral genomic diversity is even greater than previously thought. As the discovery of new endolysins demonstrates, this untapped diversity could aid biotechnological efforts, such as the search for biological agents for medicine and the application of bio-industry to agriculture or food production.

MATERIALS AND METHODS

Viral contig recovery and dereplication. Contigs larger than 1 kb from the three metagenome data sets were described by Haro-Moreno et al. (12). For the LR data set, we decided to analyze the CCS15 data set, consisting of PacBio long reads that have been resequenced at least 15 times. This process results in reads with 99.95% base calling accuracy, which is similar to Illumina error rates and results in accurate gene calling (12). Phage sequence recovery was performed in two steps. Bacteria and archaea viral contigs were recovered using VIBRANT (16) with default parameters. Eukaryotic viruses were recovered via manual curation. Each data set was dereplicated using CD-HIT (42) at 95% identity to remove redundant sequences. Contigs were considered unique based on the definition of "viral population" as

described by Gregory et al. (17); that is, contigs were considered part of the same population if they had hits with at least 95% identity and the sum of distinct alignment lengths resulted in a coverage of at least 70% across the smallest contig using BLASTN (43).

Genome annotation. Predicted viral contigs were taxonomically annotated following the method described by Beaulaurier et al. (11). The predicted proteins from each contig were annotated against the NCBI Viral Genomes database (44) (downloaded in September 2021) using LAST (45). Viral contigs were annotated at the order level if they contained one, three, or five or more proteins with top hits to phages that infect the same host genus. The choice of threshold seems to affect only the number of phages classified, not the community composition. The contigs were also functionally annotated following a variation of the method described by Zaragoza-Solas et al. (46). Protein alignments were downloaded from the PHROG (47) and CDD (33) databases and then converted to hidden Markov models (HMMs) using hmmbuild (48). Protein sequences from the three data sets were annotated against the previously built HMMs using hmmscan (48). For each database, we assigned to each gene the best hit with an E value of at least 10^{-5} and a query coverage of at least 50%. Proteins were then clustered at 30% identity and 50% query coverage using MMSeqs2 (49), and the annotations for each cluster were manually curated to ensure that the annotations were coherent for all proteins in the cluster. All contigs were searched for the presence of tRNAs using tRNA-scan-SE (50).

Read recruitment. Viral contigs from the SRa and LRa data sets and the unique contigs from the SR data set were mapped against the *Tara* Oceans metagenomes using pblat (51), using a cutoff of 95% nucleotide identity over at least 50 nucleotides. Each read was mapped only to the viral contig with the best match. Normalization was performed by calculating RPKG (reads recruited per kilobase of the genome per gigabase of the metagenome) so that recruitment values could be compared across samples.

Phylogenetic reconstruction of viral marker proteins. Phylogenetic trees of marker viral proteins were constructed adapting the method described by Benler et al. (52). Marker viral proteins in the SRa, LRa, and LR data sets were detected via hmmsearch (48) against the PHROG (47) database (see "Genome annotation") and merged into a single data set. This data set was then grouped with mmseqs2 (49) into clusters with 50% amino acid identity and a coverage of 70%, which were then aligned using ClustalOmega (53) and compared to each other using hhsearch (48). A distance matrix was calculated by calculating distances following the formula $-\ln[S_{A,B}/\min(S_{A,A}, S_{B,B})]$, where $S_{A,B}$ is the raw score per alignment length. *A* and *B* are the different clusters being compared. $S_{A,A}$ and $S_{B,B}$ are the raw alignment score of those clusters aligned to themselves. This matrix was used to build a dendrogram (unweighted pair group method using average linkages [UPGMA]), which acted as a guide to merge clusters using ClustalOmega, resulting in larger protein alignments. The resulting protein alignments were filtered to remove sites with more than 50% gaps and then used to build trees using FastTree (54) (substitution matrix, BLOSUM45; James-Taylor-Thornton model).

Phylogenetic reconstruction of eukaryotic viruses. To assess the phylogeny of the contigs categorized as eukaryotic viruses, a concatenation of 5 marker proteins (PolB, SFII, A32, VLTF3, and MCP) was built. ncldv_markersearch (23) was used to identify and align individual marker proteins, and then a Python script was used to build the final concatenation. Contigs were included in the concatenation if they had at least 3 of the 5 marker proteins. A database of 622 NCLDV Metagenome-Assembled Genomes from the previous study (23) and reference NCLDV from RefSeq (55) were added to the concatenation. Finally, a phylogenetic tree was built using IQ-TREE2 (56), using the model VT+F+I+G4 and producing 1,000 ultrafast bootstraps to assess confidence.

Putative endolysin discovery and analysis. Putative endolysins in the Mediterranean data sets were extracted following the method described by Fernández-Ruiz et al. (29). The predicted proteins from each contig were compared against a curated database of endolysins using DIAMOND (57). Matches were classified as putative endolysins if the match had >50% identity, covered at least 30% of the query sequence, the alignment was at least 50 aa long and the E value was at least 10^{-3}. A phylogenetic tree including both the reference data set and new putative sequences was built following the method described above (see "Phylogenetic reconstruction of viral marker proteins"). Protein domains were detected by using hmmsearch (48) against the CDD (33) and dbCAN2 (58) databases, with a match being considered valid if it had 70% HMM coverage and an E value of at least 10^{-5}. Proteins of the C4 clade were tested for the presence of a signal-arrest-release domain following the method of Oliveira et al. (59).

Statistical testing. Wilcoxon rank sum tests were performed using the coin package in R (60). The effect size for Kruskal-Wallis test was calculated using the rstatix package (https://cran.r-project.org/web/packages/rstatix/index.html).

Data availability. Metagenomic data sets (Illumina reads, MedWinter-FEB2019-I; PacBio CCS reads, MedWinter-FEB2019-PBCCS15; and PacBio raw reads, MedWinter-FEB2019-PB) are available in the NCBI BioProject database under accession number PRJNA674982.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, TIF file, 2.6 MB.

FIG S2, TIF file, 2.5 MB.

FIG S3, TIF file, 2.2 MB.

FIG S4, TIF file, 2 MB.

TABLE S1, XLSX file, 0.4 MB.

TABLE S2, XLSX file, 0.01 MB.

ACKNOWLEDGMENTS

This work was supported by the grants VIREVO CGL2016-76273-P [AEI/FEDER, EU] and FLEX3GEN PID2020-118052GB-I00 (cofunded with FEDER funds) from the Spanish Ministerio de Economía, Industria y Competitividad and HIDRAS3 PROMETEU/2019/009 from Generalitat Valenciana. A.Z.-S. was supported by a Ph.D. fellowship from the Spanish Ministerio de Economía y Competitividad (BES-2017-079993).

M.L.-P. and F.R.-V. conceived the study. A.Z.-S. analyzed the data. J.M.H.-M. and F.R.-V. contributed to writing the manuscript. All authors revised the manuscript and approved the final version.

We declare that we have no competing interests.

REFERENCES

1. Bergh Ø, Børsheim KY, Bratbak G, Heldal M. 1989. High abundance of viruses found in aquatic environments. Nature 340:467–468. https://doi.org/10.1038/340467a0.
2. Suttle CA. 2007. Marine viruses—major players in the global ecosystem. Nat Rev Microbiol 5:801–812. https://doi.org/10.1038/nrmicro1750.
3. Fuhrman JA, Noble RT. 1995. Viruses and protists cause similar bacterial mortality in seawater. Limnol Oceanogr 40:1236–1242. https://doi.org/10.4319/lo.1995.40.7.1236.
4. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. 2005. Photosynthesis genes in marine viruses yield proteins during host infection. Nature 438:86–89. https://doi.org/10.1038/nature04111.
5. Fedida A, Lindell D. 2017. Two Synechococcus genes, two different effects on cyanophage infection. Viruses 9:136. https://doi.org/10.3390/v9060136.
6. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, de Vargas C, Gasol JM, Gorsky G, Gregory AC, Guidi L, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Poulos BT, Schwenck SM, Speich S, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Bork P, Bowler C, Sunagawa S, Wincker P, Karsenti E, Sullivan MB, Tara Oceans Coordinators. 2015. Patterns and ecological drivers of ocean viral communities. Science 348:1261498. https://doi.org/10.1126/science.1261498.
7. Coutinho FH, Rosselli R, Rodríguez-Valera F. 2019. Trends of microdiversity reveal depth-dependent evolutionary strategies of viruses in the Mediterranean. mSystems 4:e00554-19. https://doi.org/10.1128/mSystems.00554-19.
8. Haro-Moreno JM, Rodriguez-Valera F, Rosselli R, Martinez-Hernandez F, Roda-Garcia JJ, Lluesma Gomez M, Fornas O, Martinez-Garcia M, López-Pérez M. 2020. Ecogenomics of the SAR11 clade. Environ Microbiol 22:1748–1763. https://doi.org/10.1111/1462-2920.14896.
9. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F, Mira A. 2009. Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7:828–836. https://doi.org/10.1038/nrmicro2235.
10. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, Sullivan MB, Temperton B. 2019. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. PeerJ 7:e6800. https://doi.org/10.7717/peerj.6800.
11. Beaulaurier J, Luo E, Eppley JM, Den Uyl P, Dai X, Burger A, Turner DJ, Pendelton M, Juul S, Harrington E, DeLong EF. 2020. Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. Genome Res 30:437–446. https://doi.org/10.1101/gr.251686.119.
12. Haro-Moreno JM, López-Pérez M, Rodriguez-Valera F. 2021. Enhanced recovery of microbial genes and genomes from a marine water column using long-read metagenomics. Front Microbiol 12:708782. https://doi.org/10.3389/fmicb.2021.708782.
13. Yahara K, Suzuki M, Hirabayashi A, Suda W, Hattori M, Suzuki Y, Okazaki Y. 2021. Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria. Nat Commun 12:27. https://doi.org/10.1038/s41467-020-20199-9.
14. Cao J, Zhang Y, Dai M, Xu J, Chen L, Zhang F, Zhao N, Wang J. 2020. Profiling of human gut virome with Oxford Nanopore technology. Med Microecol 4:100012. https://doi.org/10.1016/j.medmic.2020.100012.
15. López-Pérez M, Haro-Moreno JM, Gonzalez-Serrano R, Parras-Moltó M, Rodriguez-Valera F. 2017. Genome diversity of marine phages recovered from Mediterranean metagenomes: size matters. PLoS Genet 13:e1007018. https://doi.org/10.1371/journal.pgen.1007018.
16. Kieft K, Zhou Z, Anantharaman K. 2020. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome 8:90. https://doi.org/10.1186/s40168-020-00867-0.
17. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova K, Carmichael M, Cruaud C, Dimier C, Domínguez-Huerta G, Ferland J, Kandels S, Liu Y, Marec C, Pesant S, Picheral M, Pisarev S, Poulain J, Tremblay J-É, Vik D, Babin M, Bowler C, Culley AI, de Vargas C, Dutilh BE, Iudicone D, Karp-Boss L, Roux S, Sunagawa S, Wincker P, Sullivan MB, Acinas SG, Babin M, Bork P, Boss E, Bowler C, Cochrane G, de Vargas C, Follows M, Gorsky G, Grimsley N, Guidi L, Hingamp P, Iudicone D, Jaillon O, Kandels LS, Karp-Boss L, Karsenti E, Tara Oceans Coordinators, et al. 2019. Marine DNA viral macro- and microdiversity from pole to pole. Cell 177:1109–1123.E14. https://doi.org/10.1016/j.cell.2019.03.040.
18. Acinas SG, Sánchez P, Salazar G, Cornejo-Castillo FM, Sebastián M, Logares R, Royo-Llonch M, Paoli L, Sunagawa S, Hingamp P, Ogata H, Lima-Mendez G, Roux S, González JM, Arrieta JM, Alam IS, Kamau A, Bowler C, Raes J, Pesant S, Bork P, Agustí S, Gojobori T, Vaqué D, Sullivan MB, Pedrós-Alió C, Massana R, Duarte CM, Gasol JM. 2021. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Commun Biol 4:604. https://doi.org/10.1038/s42003-021-02112-2.
19. Buchholz HH, Michelsen ML, Bolaños LM, Browne E, Allen MJ, Temperton B. 2021. Efficient dilution-to-extinction isolation of novel virus–host model systems for fastidious heterotrophic bacteria. ISME J 15:1585–1598. https://doi.org/10.1038/s41396-020-00872-z.
20. Bartlau N, Wichels A, Krohne G, Adriaenssens EM, Heins A, Fuchs BM, Amann R, Moraru C. 2022. Highly diverse flavobacterial phages isolated from North Sea spring blooms. ISME J 16:555–568. https://doi.org/10.1038/s41396-021-01097-4.
21. López-Pérez M, Haro-Moreno JM, de la Torre JR, Rodriguez-Valera F. 2019. Novel Caudovirales associated with Marine Group I Thaumarchaeota assembled from metagenomes. Environ Microbiol 21:1980–1988. https://doi.org/10.1111/1462-2920.14462.
22. Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13:36–46. https://doi.org/10.1038/nrg3117.
23. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. 2020. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nat Commun 11:1710. https://doi.org/10.1038/s41467-020-15507-2.
24. Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, Iudicone D, Karsenti E, Speich S, Trouble R, Dimier C, Searson S, Tara Oceans Consortium Coordinators. 2015. Open science resources for the discovery and analysis of Tara Oceans data. Sci Data 2:150023. https://doi.org/10.1038/sdata.2015.23.
25. Haro-Moreno JM, López-Pérez M, de la Torre JR, Picazo A, Camacho A, Rodriguez-Valera F. 2018. Fine metagenomic profile of the Mediterranean stratified and mixed water columns revealed by assembly and recruitment. Microbiome 6:128. https://doi.org/10.1186/s40168-018-0513-5.
26. Pimentel M. 2014. Genetics of phage lysis. Microbiol Spectr 2:MGM2. https://doi.org/10.1128/microbiolspec.MGM2-0017-2013.
27. Haddad Kashani H, Schmelcher M, Sabzalipoor H, Seyed Hosseini E, Moniri R. 2018. Recombinant endolysins as potential therapeutics against antibiotic-resistant Staphylococcus aureus: current status of research and novel delivery strategies. Clin Microbiol Rev 31:e00071-17. https://doi.org/10.1128/CMR.00071-17.

28. Ramos-Vivas J, Elexpuru-Zabaleta M, Samano ML, Barrera AP, Forbes-Hernández TY, Giampieri F, Battino M. 2021. Phages and enzybiotics in food biopreservation. Molecules 26:5138. https://doi.org/10.3390/molecules26175138.
29. Fernández-Ruiz I, Coutinho FH, Rodriguez-Valera F. 2018. Thousands of novel endolysins discovered in uncultured phage genomes. Front Microbiol 9:1033. https://doi.org/10.3389/fmicb.2018.01033.
30. Grütter MG, Weaver LH, Matthews BW. 1983. Goose lysozyme structure: an evolutionary link between hen and bacteriophage lysozymes? Nature 303:828–831. https://doi.org/10.1038/303828a0.
31. Xu M, Struck DK, Deaton J, Wang I-N, Young R. 2004. A signal-arrest-release sequence mediates export and control of the phage P1 endolysin. Proc Natl Acad Sci U S A 101:6415–6420. https://doi.org/10.1073/pnas.0400957101.
32. Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27–30. https://doi.org/10.1093/nar/28.1.27.
33. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222–D226. https://doi.org/10.1093/nar/gku1221.
34. Bateman A, Murzin AG, Teichmann SA. 1998. Structure and distribution of pentapeptide repeats in bacteria. Protein Sci 7:1477–1480. https://doi.org/10.1002/pro.5560070625.
35. Prag S, Adams JC. 2003. Molecular phylogeny of the kelch-repeat superfamily reveals an expansion of BTB/kelch proteins in animals. BMC Bioinformatics 4:42. https://doi.org/10.1186/1471-2105-4-42.
36. Jahn MT, Arkhipova K, Markert SM, Stigloher C, Lachnit T, Pita L, Kupczok A, Ribes M, Stengel ST, Rosenstiel P, Dutilh BE, Hentschel U. 2019. A phage protein aids bacterial symbionts in eukaryote immune evasion. Cell Host Microbe 26:542–550.E5. https://doi.org/10.1016/j.chom.2019.08.019.
37. Buck M, Gerken T. 2019. Two hands grip better than one for tight binding and specificity: how a phage endolysin fits into the cell wall of its host. Structure 27:1350–1352. https://doi.org/10.1016/j.str.2019.08.006.
38. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ, Anisimova M, Jakobsen KS, Linke D. 2019. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res 47:10994–11006. https://doi.org/10.1093/nar/gkz841.
39. Toussaint A, Rice PA. 2017. Transposable phages, DNA reorganization and transfer. Curr Opin Microbiol 38:88–94. https://doi.org/10.1016/j.mib.2017.04.009.
40. Sharifi F, Ye Y. 2019. MyDGR: a server for identification and characterization of diversity-generating retroelements. Nucleic Acids Res 47:W289–W294. https://doi.org/10.1093/nar/gkz329.
41. Groth AC, Calos MP. 2004. Phage integrases: biology and applications. J Mol Biol 335:667–678. https://doi.org/10.1016/j.jmb.2003.09.082.
42. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. https://doi.org/10.1093/bioinformatics/btl158.
43. McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 32:W20–W25. https://doi.org/10.1093/nar/gkh435.
44. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. 2015. NCBI Viral Genomes Resource. Nucleic Acids Res 43:D571–D577. https://doi.org/10.1093/nar/gku1207.
45. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. Genome Res 21:487–493. https://doi.org/10.1101/gr.113985.110.
46. Zaragoza-Solas A, Rodriguez-Valera F, López-Pérez M. 2020. Metagenome mining reveals hidden genomic diversity of pelagimyophages in aquatic environments. mSystems 5:e00905-19. https://doi.org/10.1128/mSystems.00905-19.
47. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, Toussaint A, Petit M-A, Enault F. 2021. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genomics Bioinforma 3:lqab067. https://doi.org/10.1093/nargab/lqab067.
48. Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform 23:205–211. https://doi.org/10.1142/9781848165632_0019.
49. Mirdita M, Von Den Driesch L, Galiez C, Martin MJ, Soding J, Steinegger M. 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res 45:D170–D176. https://doi.org/10.1093/nar/gkw1081.
50. Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964. https://doi.org/10.1093/nar/25.5.955.
51. Wang M, Kong L. 2019. pblat: a multithread blat algorithm speeding up aligning sequences to genomes. BMC Bioinformatics 20:28. https://doi.org/10.1186/s12859-019-2597-8.
52. Benler S, Yutin N, Antipov D, Rayko M, Shmakov S, Gussow AB, Pevzner P, Koonin EV. 2021. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. Microbiome 9:78. https://doi.org/10.1186/s40168-021-01017-w.
53. Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci 27:135–145. https://doi.org/10.1002/pro.3290.
54. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
55. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189.
56. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015.
57. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176.
58. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 46:W95–W101. https://doi.org/10.1093/nar/gky418.
59. Oliveira H, Melo LDR, Santos SB, Nobrega FL, Ferreira EC, Cerca N, Azeredo J, Kluskens LD. 2013. Molecular aspects and comparative genomics of bacteriophage endolysins. J Virol 87:4558–4570. https://doi.org/10.1128/JVI.03277-12.
60. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. 2008. Implementing a class of permutation tests: the coin package. J Stat Soft 28:1–23.

Long-Read Metagenomics Improves the Recovery of Viral Diversity from Complex Natural Marine Samples

Asier Zaragoza-Solas, Jose M. Haro-Moreno, Francisco Rodriguez-Valera* and Mario López-Pérez*

Evolutionary Genomics Group, División de Microbiología, Universidad Miguel Hernández, 03550 San Juan de Alicante, Spain

*Corresponding Authors: Francisco Rodriguez-Valera, frvalera@umh.es, or Mario López-Pérez, mario.lopezp@umh.es.

Evolutionary Genomics Group, División de Microbiología, Universidad Miguel Hernández, Apto 18, San Juan de Alicante, 03550 Alicante, Spain.
Phone +34 965919313, Fax +34 965919457

SUPPLEMENTARY FILES

Table S2 can be found at the publisher's website:
https://doi.org/10.1128/msystems.00192-22

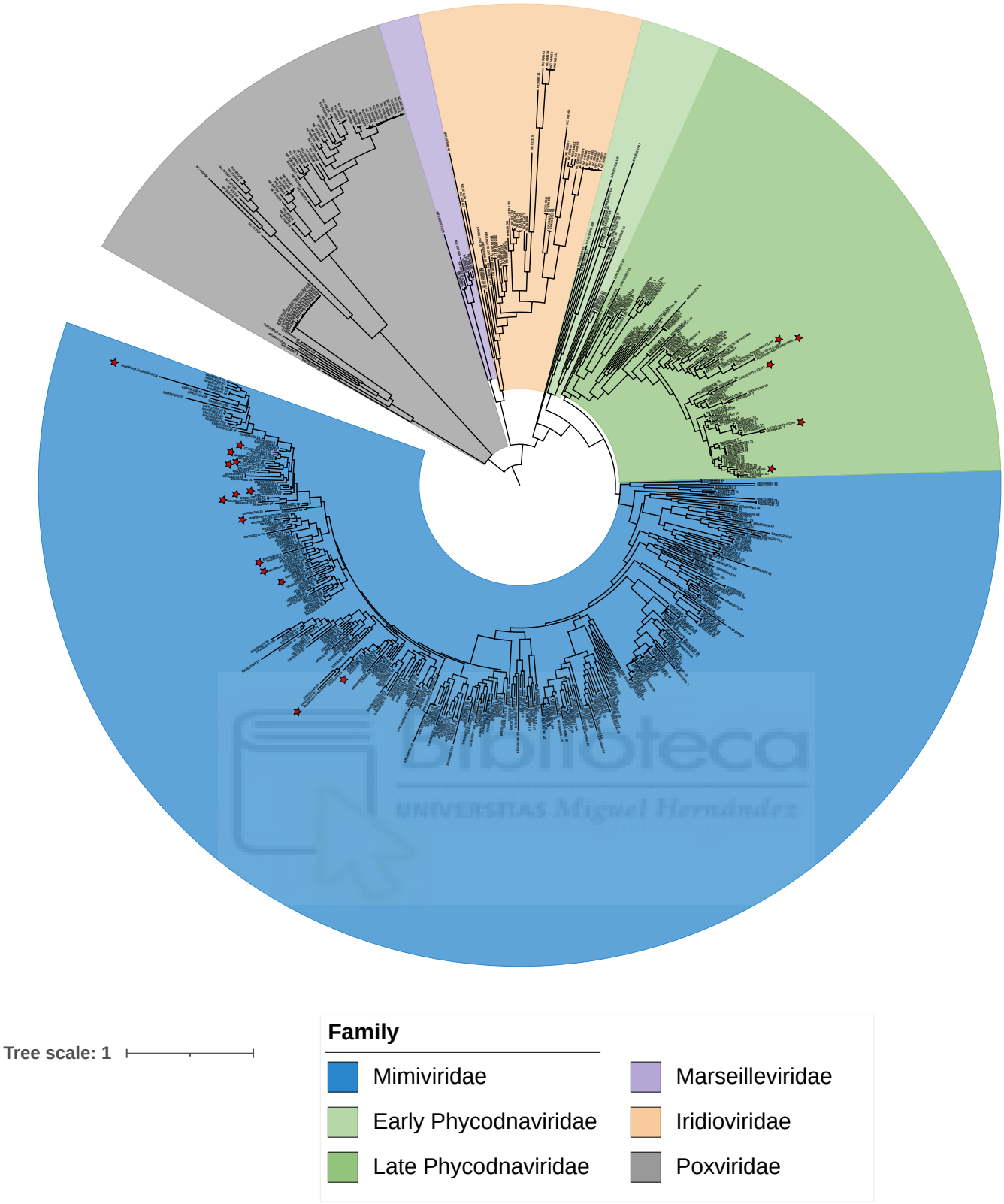


Fig. S1. Phylogeny of the NCLDV obtained in this study together with 444 reference genomes from Moniruzzaman et al and the NCDLV genomes present in the RefSeq database, using a concatenated of 5 highly conserved marker genes (mcp, VLTF3, A32, SFII and PolB). The sections of the tree are colored based on the taxonomic family. A red star in a tree branch indicates which sequences have been recovered in this study.

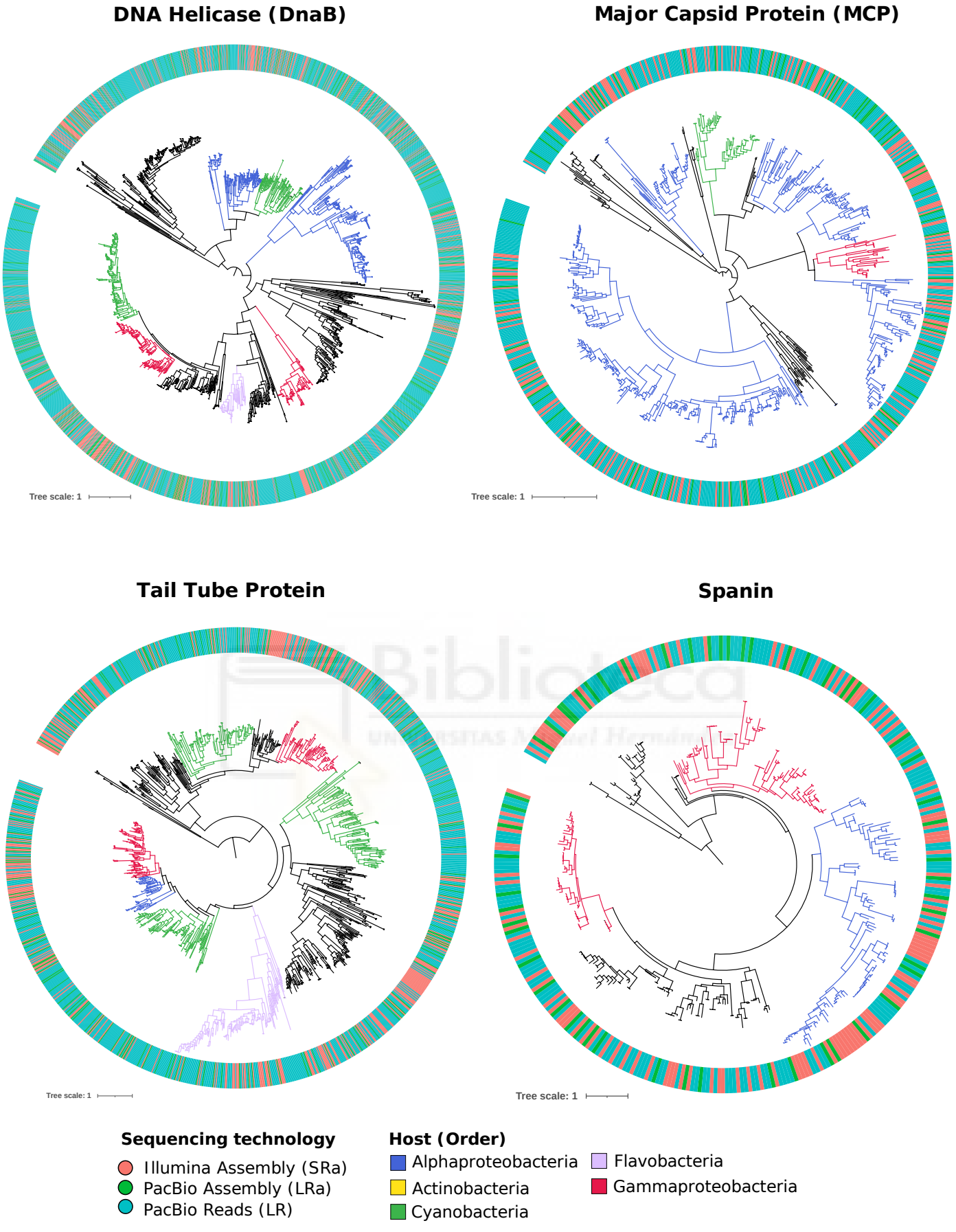


Fig. S2. Phylogenetic trees for the DNA helicase (DnaB), major capsid protein (mcp), Tail tube protein and Spanin. Branches are coloured according to the assigned host, while the colour of the outer circle indicates the dataset the contig was obtained from (orange for Illumina assembly, green for PacBio assembly, blue for raw PacBio reads).

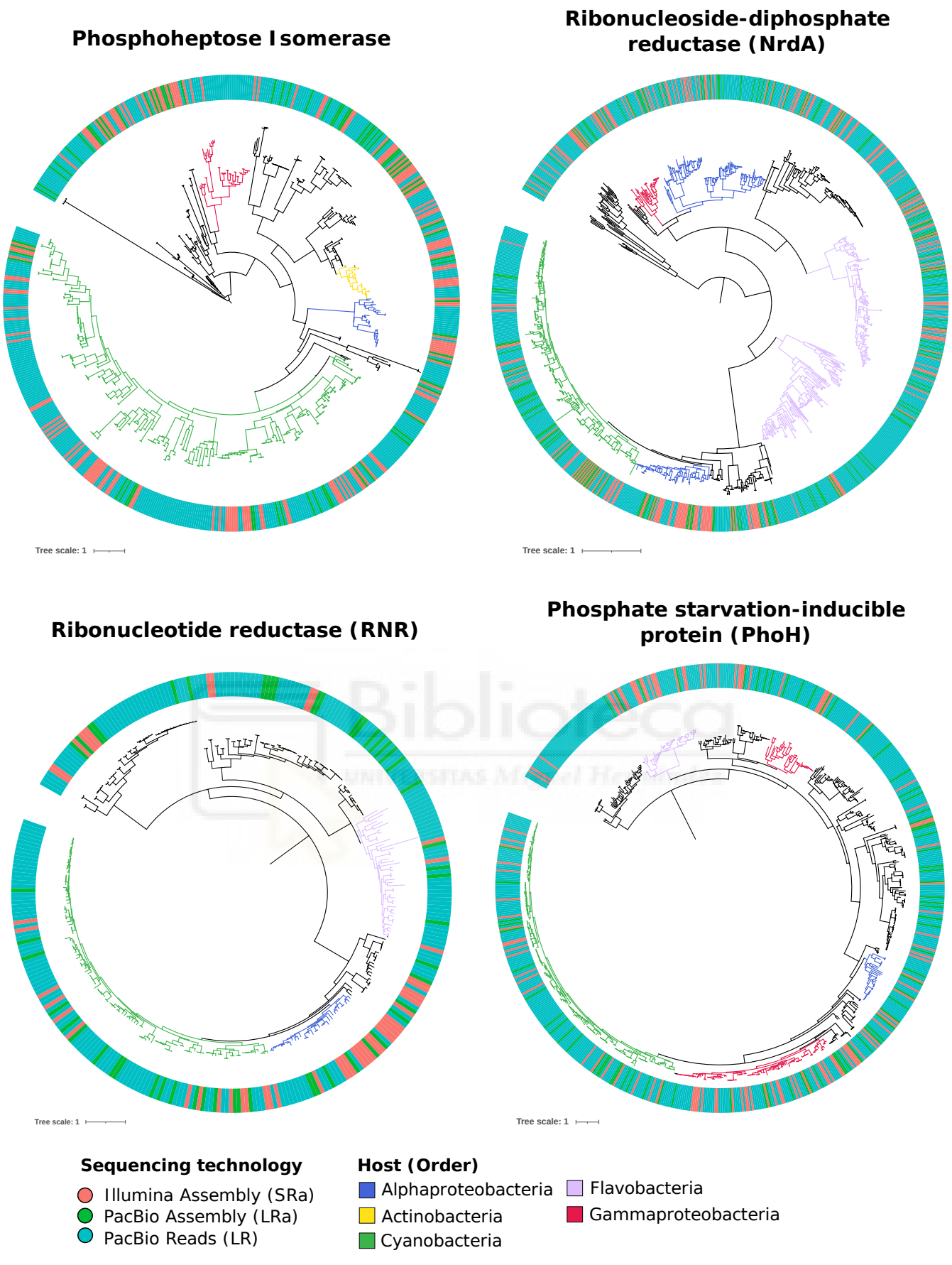


Fig. S3: Phylogenetic Trees for the AMGs Phosphoheptose Isomerase, Ribonucleoside-diphosphate reductase (NrdA), Ribonucleotide reductase (RNR) and Phosphate starvation-inducible protein (PhoH). Branches are coloured according to the assigned host, while the colour of the outer circle indicates the dataset the contig was obtained from (orange for Illumina assembly, green for PacBio assembly, blue for raw PacBio reads).

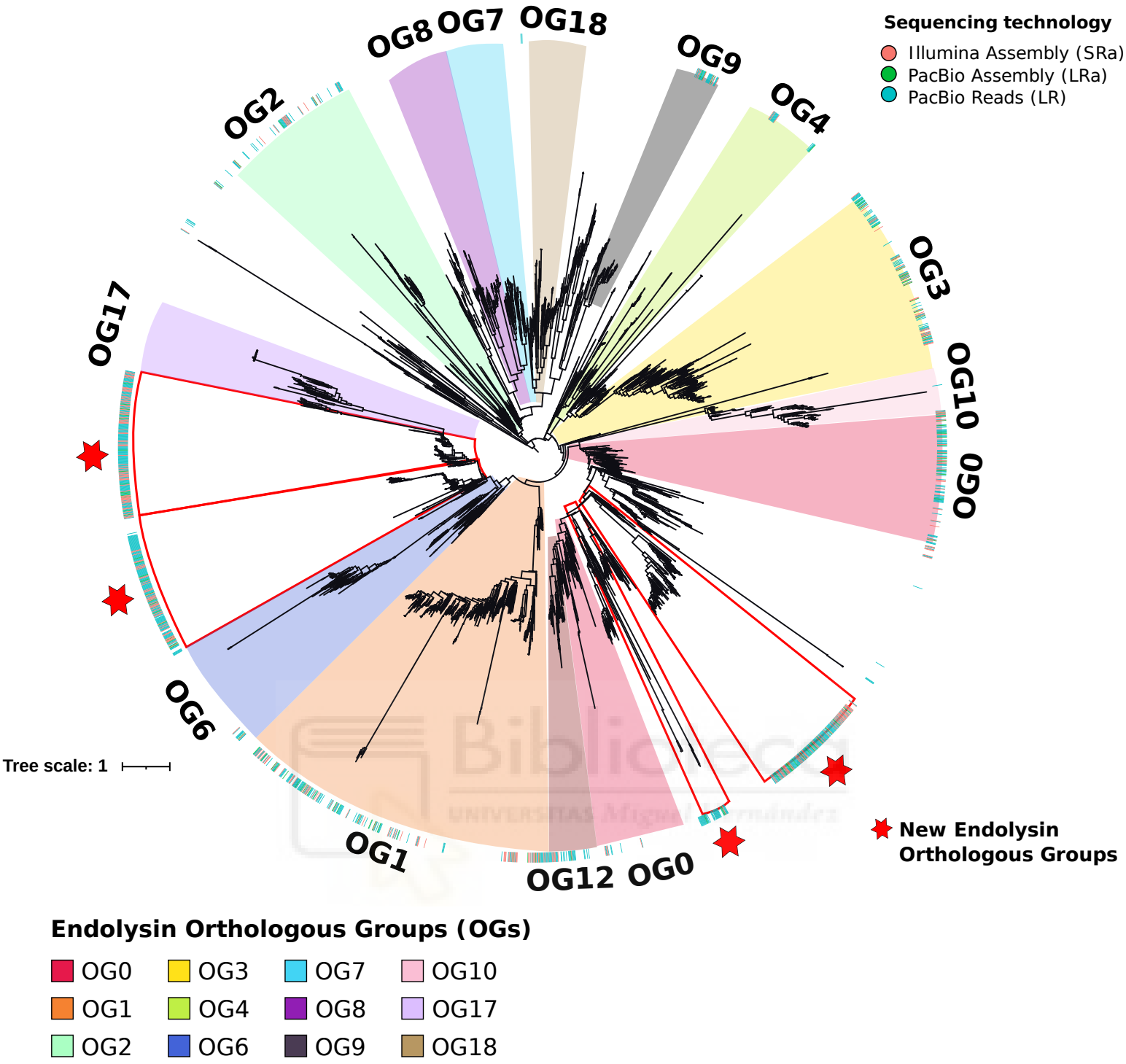


Fig. S4. Phylogenetic tree of endolysins recovered in this study, along with the endolysin dataset described in Fernandez-Ruiz et al. Sections of the tree are colored based on the orthologous groups described in the aforementioned paper, while sections bordered in red indicate new orthologous groups. An outer circle indicates which proteins come from this study and from which dataset (orange for Illumina assembly, green for PacBio assembly, blue for raw PacBio reads).

Table S2. Shared and unique proteins among the three datasets and GOV2 for the phage Proteins and AMGs shown in Figures S2 and S3.

PHROGS ID	Name	Matches (GOV2)	Unique Matches (GOV2)	Unique% (GOV2)	Matches (SRa)	Unique Matches (SRa)	Unique% (SRa)	Matches (LRa)	Unique Matches (LRa)	Unique% (LRa)	Matches (LR)	Unique Matches (LR)	Unique% (LR)
phrog_19	DNA Helicase DnaB	31,620	21,583	68.26%	358	19	5.31%	213	7	3.29%	1,090	149	13.67%
phrog_45	Tail Tube Protein	13,170	9,501	72.14%	377	205	54.38%	170	71	41.76%	917	499	54.42%
phrog_170	Phosphate starvation-inducible protein PhoH	7,337	4,018	54.76%	110	8	7.27%	75	0	0.00%	512	97	18.95%
phrog_267	Major Capsid Protein	12,161	5,967	49.07%	216	7	3.24%	84	0	0.00%	533	34	6.38%
phrog_739	Rz-like Spanin	3,924	2,250	57.34%	140	22	15.71%	59	7	11.86%	183	29	15.85%
phrog_1238	Phosphoheptose isomerase	6,566	5,171	78.75%	92	11	11.96%	51	5	9.80%	272	78	28.68%
phrog_3844	Aerobic NDP reductase NrdA, large subunit	7,968	3,099	38.89%	147	12	8.16%	99	7	7.07%	546	150	27.47%
phrog_3987	Ribonucleotide reductase, large subunit	2,212	994	44.94%	45	2	4.44%	44	3	6.82%	204	69	33.82%

10. Annex 3

Descriptor

Patterns

RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content

Highlights

- RaFAH was developed to predict the hosts of viruses of Bacteria and Archaea
- RaFAH displayed comparable or superior performance to other host-prediction tools
- RaFAH performed well across viromes from eight different ecosystems
- RaFAH identified hundreds of genomic sequences as derived from viruses of Archaea

Authors

Felipe Hernandes Coutinho, Asier Zaragoza-Solas, Mario López-Pérez, ..., Bas E. Dutilh, Robert Edwards, Francisco Rodriguez-Valera

Correspondence

fhernandes@icm.csic.es

In brief

We developed a machine-learning tool called RaFAH that uses genomic data to predict the hosts of viruses of Bacteria and Archaea. This tool led to an expansion of the catalog of known archaeal viruses and has potential to help better characterize the global virosphere by linking viruses to their hosts with high precision and recall.

Coutinho et al., 2021, Patterns 2, 100274
July 9, 2021
https://doi.org/10.1016/j.patter.2021.100274

CellPress

Patterns

CellPress
OPEN ACCESS

Descriptor

RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content

Felipe Hernandes Coutinho,[1,8,*] Asier Zaragoza-Solas,[1] Mario López-Pérez,[1] Jakub Barylski,[2] Andrzej Zielezinski,[3] Bas E. Dutilh,[4,5] Robert Edwards,[6] and Francisco Rodriguez-Valera[1,7]

[1]Evolutionary Genomics Group, Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández, Aptdo. 18., Ctra. Alicante-Valencia N-332, s/n, San Juan de Alicante, 03550 Alicante, Spain

[2]Molecular Virology Research Unit, Faculty of Biology, Adam Mickiewicz University Poznan, 61-614 Poznan, Poland

[3]Department of Computational Biology, Faculty of Biology, Adam Mickiewicz University Poznan, 61-614 Poznan, Poland

[4]Centre for Molecular and Biomolecular Informatics (CMBI), Radboud University Medical Centre/Radboud Institute for Molecular Life Sciences, 6525 GA Nijmegen, the Netherlands

[5]Theoretical Biology and Bioinformatics, Science for Life, Utrecht University (UU), 3584 CH Utrecht, the Netherlands

[6]College of Science and Engineering, Flinders University, Bedford Park, SA 5042, Australia

[7]Moscow Institute of Physics and Technology, Dolgoprudny 141701, Russia

[8]Lead contact

*Correspondence: fhernandes@icm.csic.es

https://doi.org/10.1016/j.patter.2021.100274

THE BIGGER PICTURE Viruses that infect Bacteria and Archaea are ubiquitous and extremely abundant. Recent advances have led to the discovery of many thousands of complete and partial genomes of these biological entities. Understanding the biology of these viruses and how they influence their ecosystems depends on knowing which hosts they infect. We developed a tool that uses data from complete or fragmented genomes to predict the hosts of viruses using a machine-learning approach. Our tool, RaFAH, displayed performance comparable with or superior to that of other host-prediction tools. In addition, it identified hundreds of sequences as derived from the genomes of viruses of Archaea, which are one of the least characterized fractions of the global virosphere.

1 **2** 3 4 5 **Proof-of-Concept:** Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Culture-independent approaches have recently shed light on the genomic diversity of viruses of prokaryotes. One fundamental question when trying to understand their ecological roles is: which host do they infect? To tackle this issue we developed a machine-learning approach named Random Forest Assignment of Hosts (RaFAH), that uses scores to 43,644 protein clusters to assign hosts to complete or fragmented genomes of viruses of Archaea and Bacteria. RaFAH displayed performance comparable with that of other methods for virus-host prediction in three different benchmarks encompassing viruses from RefSeq, single amplified genomes, and metagenomes. RaFAH was applied to assembled metagenomic datasets of uncultured viruses from eight different biomes of medical, biotechnological, and environmental relevance. Our analyses led to the identification of 537 sequences of archaeal viruses representing unknown lineages, whose genomes encode novel auxiliary metabolic genes, shedding light on how these viruses interfere with the host molecular machinery. RaFAH is available at https://sourceforge.net/projects/rafah/.

INTRODUCTION

Viruses that infect Bacteria and Archaea are the most abundant and diverse biological entities on Earth. Because of their sheer abundance, genomic diversity, and the fact that most viruses are only found in specific ecological niches, they remain elusive. Culture-independent techniques such as metagenomics[1] have been pivotal in the effort to describe viral biodiversity. Computational approaches have been developed to link these novel viruses to putative hosts[2] by identifying genomic signals that are indicative of a virus-host association.

First, alignment-free methods such as *k*-mer profiles use nucleotide composition to predict the host of a viral genome. Some viruses adapt their oligonucleotide composition to that of the host they infect, a process that may be driven by the adaptation of the codon usage to the translational machinery and tRNA pool available in the host cell, exchange of the genetic material, co-evolution of regulatory sequences, and/or an evasion of the host defense systems. Hence, by identifying the prokaryote genome with the highest significant similarity to a viral genome, tools that exploit *k*-mer profiles assume that prokaryote genome to be the host of the virus in question. Alignment-free methods (e.g., WIsH) show very high recall (i.e., percentage of viral genomes linked to a host) but usually have low precision (i.e., percentage of correct virus-host associations among the predicted virus-host associations), with reported host-prediction accuracy for genus-level predictions between 33% and 64% depending on the dataset.[2–4] Similarities in *k*-mer profiles between viruses can also be used for host prediction following the same rationale (e.g., HostPhinder).[5]

Second, there are alignment-dependent approaches to assess similarity between viral and prokaryote genomes. These methods assume that genetic information exchange between viral and prokaryote genomes is indicative of virus-host associations. Specific genetic fragments, although short, might be informative for this purpose, such as CRISPR spacers and tRNA genes, while longer matches such as whole genes or integrated prophages can also provide an indication of virus-host linkage.[2] Both aforementioned approaches are limited by the fact that they require the genome of the host to be present in the reference database. That host should contain an active CRISPR system whose array should contain a spacer targeting (a close relative of) the phage, allowing identification of a protospacer without too many mismatches. Alignment-dependent approaches also require that detectable genetic exchange has taken place between virus and host. Hybrid approaches leverage on information from both alignment-free and alignment-dependent approaches for host prediction (e.g., VirHostMatcher-Net).[6]

Third, the gene content of viral sequences can be investigated in search of specific marker genes that are indicative of the host, such as photosynthesis genes for cyanophages.[7] This low-throughput approach may have high precision, but usually the recall of such predictions is low and the procedure is extremely time-consuming.

All of these approaches have been used extensively in viral metagenomic studies to predict hosts to uncultured viruses.[1,7–9] An ideal tool for virus-host prediction should combine the precision of alignment-dependent methods and the recall of alignment-free approaches. Furthermore, it should be independent of host genomes so as not to be limited by database completeness biases. Previous studies have shown that random forest algorithms are suitable for classifying viruses according to their hosts[10] and that protein domains can be used to achieve accurate host predictions.[11,12] Based on these findings, we postulated that random forest classifiers could be applied to protein content to build a classifier based on identifying combinations of genes that are indicative of virus-host associations. Through this approach, we were able to design RaFAH (Random Forest Assignment of Hosts), a classifier that combined the precision of manual curation, the recall of alignment-free approaches, and the speed and flexibility of machine learning (Figure 1).

RESULTS AND DISCUSSION

We tested the performance of RaFAH and other host-prediction approaches on an independent dataset of isolated viral genomes that did not overlap with those used for training the models (Test Set 1, composed of RefSeq viral genomes with less than 70% average amino acid identity when compared with those in Training Set 3, see experimental procedures). When using RaFAH and the other tested methods without score or prediction probability cutoff (i.e., considering as valid all host predictions with no thresholds for their probability value or bit score), RaFAH outperformed alignment-independent, hybrid, and alignment-dependent approaches for host prediction at every taxonomic level based on the F1 score (Figure 2A). This difference in performance became gradually more evident from domain to genus level. Next, we evaluated how the performance of these tools responded to thresholding (i.e., applying a cutoff on their probability value or bit score) and only considering predictions that were above the cutoffs. These analyses revealed that homology matches, CRISPR, tRNA, and combined classical approaches (i.e., homology matches, CRISPR, and tRNA, see experimental procedures) displayed the lowest recall (Figure 2B) but the highest precision (Figure 2C). HostPhinder and CRISPR displayed high precision only at the strictest score cutoffs. As a consequence, these two methods displayed very low recall when the highest cutoffs for predictions were established. RaFAH, WIsH, and VirHostMatcher-Net displayed higher recall than the other approaches, especially at the range of more permissive score cutoffs (0). Yet this higher recall came at the expense of lower precision for WIsH and VirHostMatcher-Net. Meanwhile the precision of RaFAH outperformed these tools even when no cutoffs were applied. Together, precision, recall, and F1 score suggest that RaFAH can predict more virus-host interactions than the other tested approaches while maintaining high precision, particularly for divergent viral genomes that escape detection by the classical approaches (Figure S1).

We evaluated how the similarity among the genomes in Test Set 1 with those used to train the model (Training Set 3) affected the performance of RaFAH. For this purpose, we assessed how the precision of RaFAH changed by setting a threshold on the maximum allowed average amino acid identity (AAI) between the genomes on Test Set 1 and those on Training Set 3. As expected, a positive association was observed between these variables (Figure S2), meaning that the more similar the testing genomes are to the ones used for training, the more likely RaFAH is to correctly predict their hosts at all taxonomic levels. Based on this analysis, 75% of the class-level host predictions will be correct (precision: ~0.75) for viruses that possess <60% AAI to the ones in the database, when no cutoffs on prediction score are applied.

We applied importance analysis to determine which protein clusters were most relevant for predicting viral hosts using RaFAH. The most important predictor was annotated as an Rz-like phage lysis protein (Table S1). Among the protein clusters that ranked among the 50 most important were multiple lysins, tail, and tail fiber proteins. These proteins are known to determine virus-host range, as they play fundamental roles in virus

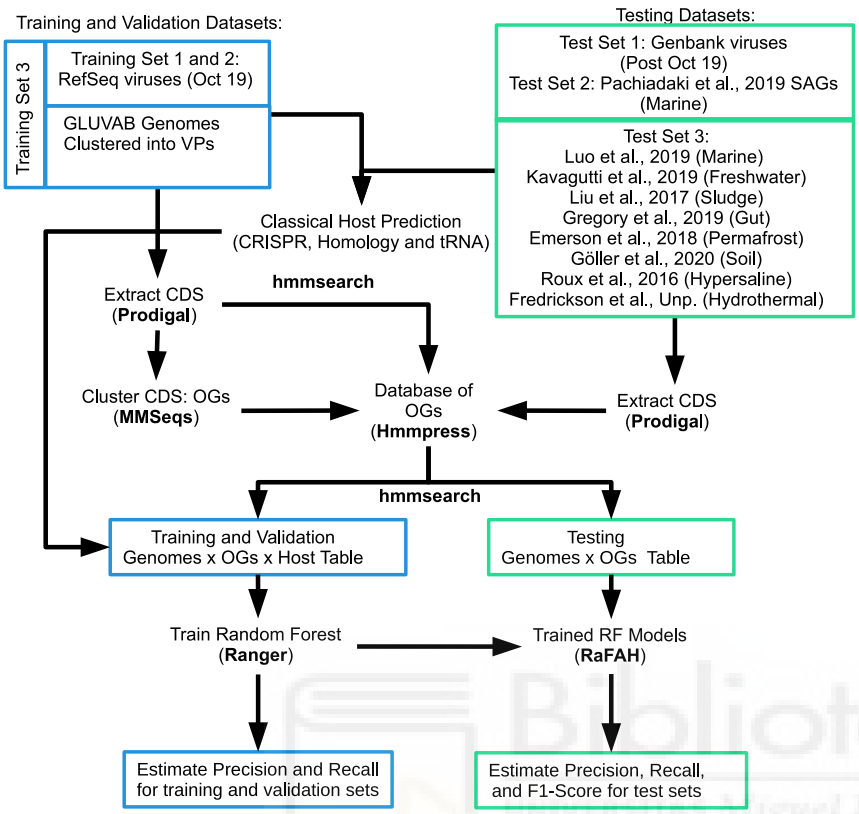


Figure 1. Overview of the strategy used to train, validate, and test random forest models

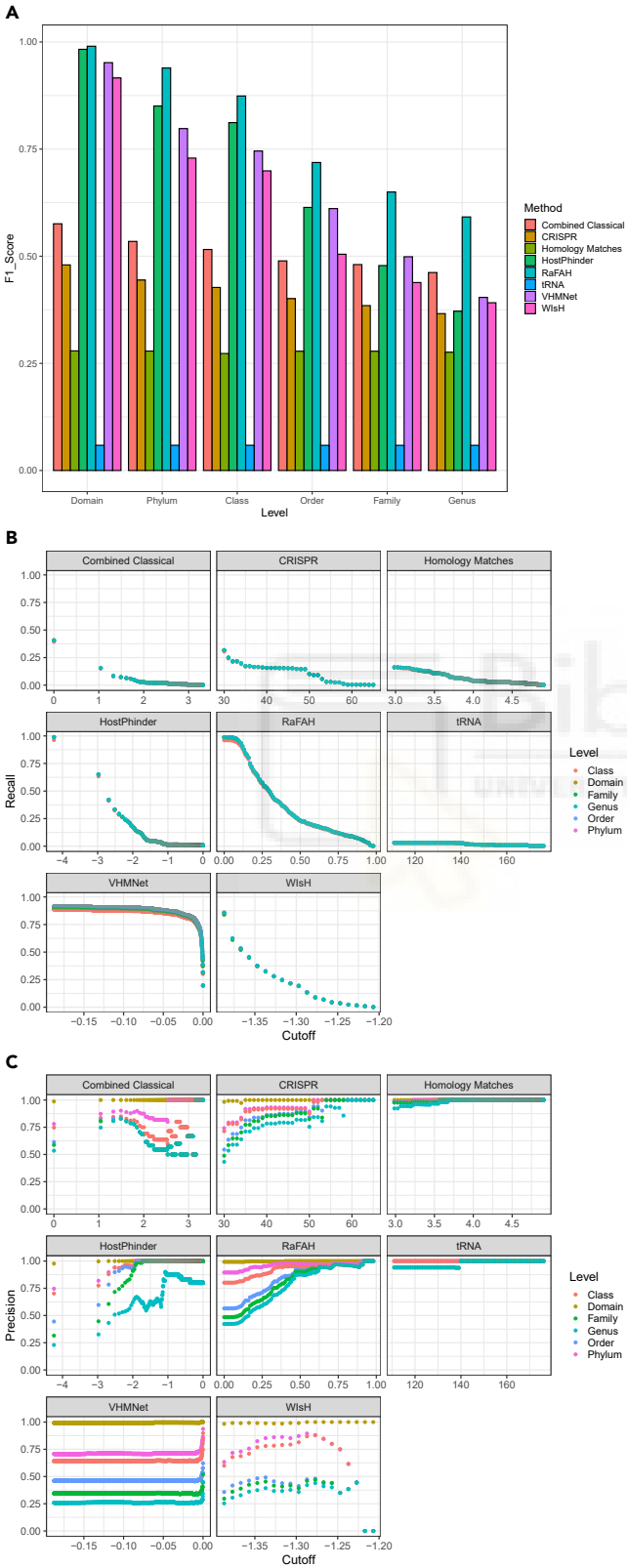
The training and validation sets were composed of viral RefSeq genome sequences published until October 2019 and viral genomic sequences derived from GLUVAB. GLUVAB sequences were clustered into viral populations (VPs) and assigned putative hosts through classical approaches (tRNA, homology matches, and CRISPR) on a per-population basis. Coding DNA sequences (CDS) were extracted from these sets and clustered into orthologous groups (OGs), aligned, and pressed into a database of hidden Markov model (HMM) profiles. Next, CDS were queried against this database to compute the bit scores of each CDS against each HMM, from which a matrix of Genomes × OG scores was derived. This matrix was used to train the random forest model. The performance of the model was evaluated on the training and validation sets according to precision and recall. The test sets comprised viral RefSeq genomes published after October 2019 (Test Set 1), viral genome fragments retrieved from marine SAGs (Test Set 2), and metagenomes/viromes from eight distinct ecosystems (Test Set 3). Similarly, CDS were extracted from these sets and queried against the HMM database derived from the training set to compute the bit scores of each CDS against each HMM, from which the testing matrix of Genomes × OG scores was derived and analyzed through RaFAH. From these results, the precision, recall, and F1 score of RaFAH were evaluated on the Test Sets.

entry and exit and host recognition.[13] The fact that these proteins ranked among the most important for RaFAH predictions is evidence that it learned to predict virus-host associations based on proteins that are directly involved in virus-host molecular interactions.

Host-prediction tools were further validated on a dataset of viral genomic sequences derived from marine single amplified genomes (SAGs), Test Set 2.[14] These sequences represent an ideal test dataset because they are uncultured viruses, not represented in the National Center for Biotechnology Information (NCBI) database used for training, and can confidently be assigned hosts because these viruses were inside or attached to the host cells during sample processing. Based on the F1 score, HostPhinder displayed the best performance at the levels of domain and class, followed by RaFAH slightly behind (Figure S3A). Yet at the level of phylum WIsH displayed the best performance, again followed closely by RaFAH. At the levels of order, family, and genus, WIsH displayed the highest F1 scores followed by the combined classical approaches. The recall (Figure S3B) and precision (Figure S3C) of RaFAH on Test Set 2 was lower than that obtained for Test Set 1. Nevertheless, a negative association between precision and recall as a function of the score cutoff was also observed for RaFAH and the other tested tools on Test Set 2 (Figure S3D). Taken together, these results are evidence that RaFAH also performed well when predicting hosts of uncultured viruses from the marine ecosystem.

Some features of Test Set 2 must be considered when interpreting these results. First, most of the viruses identified in Test Set 2 were derived from single-cell genomes classified as either *Pelagibacter*, *Puniceispirillum*, *Prochlorococcus*, and *Synechococcus*. This is expected considering these are the most abundant organisms at the ecosystem from which this dataset is derived. Nevertheless, this relatively low diversity of taxa has implications for the assessment of host-prediction tools. For instance, the genera *Prochlorococcus* and *Synechococcus* have no determined taxonomy at the level of class. Therefore, predictions at this level do not count toward precision for these particular taxa. As a consequence, the precision of all host-prediction tools displayed a steep decrease at this taxonomic level. This was particularly noticeable for VHM-Net for which all correct predictions were restricted to the two aforementioned taxa, which led to 0% precision at the level of class. Second, the majority of bacteriophage genomes in Test Set 2 have very low completeness (median 6.85%, estimated by CheckV,[15] see experimental procedures). The low diversity of hosts and the very low genome completeness (as evaluated below) likely impacted the performance of RaFAH on this dataset. Third, because most viral genomes in Test Set 2 belong to four genera, RaFAH is likely to have its performance hindered due to the number of phage genomes that infect these genera available in Training Set 3. Meanwhile, approaches that rely on host genomes are likely to be hindered by the number of these genomes available in the reference database.

To test the performance of RaFAH on samples from other habitats, we applied it to predict hosts of a dataset of viral genomes obtained from metagenomes of eight different ecosystems (Test Set 3). For comparison, we also applied the other tested methods of host prediction (HostPhinder did not scale to the more than



60,000 genomes in this dataset, and analyses did not complete after running for several days). According to the F1 score, RaFAH outperformed WIsH and VirHostMatcher-Net for this dataset as well (Figure S4A), due to slightly higher recall (Figure S4B) and precision (Figure S4C). RaFAH was also superior when the strictest cutoffs were applied, whereby both precision and recall were markedly superior to VirHostMatcher-Net (Figure S4D). On this dataset, RaFAH achieved 43.13% precision at the level of genus when no score threshold was applied. Bootstrap analysis revealed that this level of precision was consistent across 1,000 replicates (mean 43.02% ± 2.1%). This result indicates that the precision of RaFAH on Test Set 3 was not biased by uneven viral genome diversity among the samples that made up this dataset.

When using classical approaches for host prediction, the majority of viruses remained unassigned regardless of ecosystem, and the best performance of these approaches was among the human gut dataset, in which only about 25% of sequences (lengthwise) could be assigned to a host at the level of phylum (Figure 3). Meanwhile, when set to the 0.14 cutoff, which yielded 92% phylum level precision on Test Set 1 (Figure S1) and 90% on Test Set 3 (Figure S4D), RaFAH was capable of predicting putative hosts to the majority of viral sequences across all ecosystems except for the permafrost dataset, likely because viruses derived from this ecosystem are poorly represented in reference databases.

Interestingly, the host predictions yielded by RaFAH were markedly different across ecosystems. Viruses of Proteobacteria were the dominant group in all ecosystems except the human gut. As expected, the most abundant targeted hosts of the viruses from each ecosystem were the most abundant taxa that reside in those habitats. Viruses of Cyanobacteria were the second most abundant group among the marine dataset, a position that was occupied by viruses of Actinobacteria and Bacteroidetes among the freshwater dataset. Viruses of Firmicutes and Bacteroidetes were the dominant group among the dataset of human gut viruses while viruses of Firmicutes, Bacteroidetes, and Actinobacteria were among the most abundant among the soil and permafrost datasets. Viruses of Euryarchaeota were the second most abundant group among the hypersaline dataset, a position that was occupied by viruses of Crenarcheaota in the thermal springs dataset. These results are in accordance with the known prokaryote diversity that dwells in each of these ecosystems.[8,16–22] Although this agreement between virus and host community composition is to be expected, it is seldom observed in studies of viral ecology based on metagenomics because classical methods leave the majority of viruses without host predictions. RaFAH circumvents these issues by providing an accurate and complete description of viral communities regarding their targeted hosts.

Figure 2. Performance of RaFAH compared with alignment-free and classical host-prediction approaches on Test Set 1

(A) F1 score of methods when considering all predictions regardless of score at multiple taxonomic levels.

(B) Association between score cutoff and recall of predictions for each method.

(C) Association between score cutoff and precision of predictions for each method. The score cutoffs for HostPhinder, Homology matches, VirHostMatcher-Net, and combined classical are shown on the $\log_{10}$ scale. Figure S8 depicts the association between precision and score cutoff of VirHostMatcher-Net for score values above the 75th percentile.

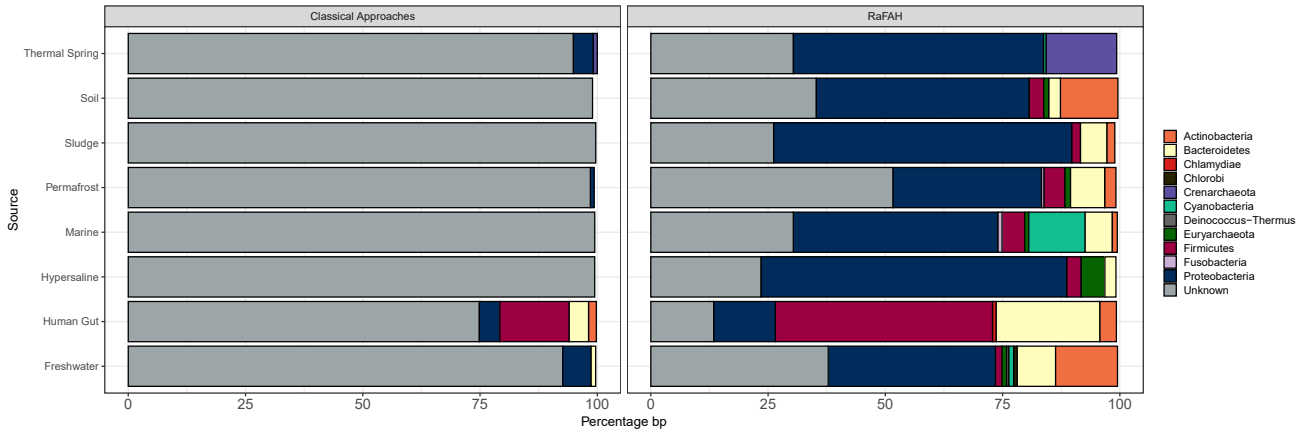


Figure 3. Description of the viromes of eight ecosystems using combined classical host-prediction approaches and RaFaH

For each dataset we calculated the fraction of the assembly predicted to each putative host phylum by each method. Phyla that represent less than 0.5% of the total assembly are not shown.

We assessed how genome completeness affected the performance of RaFAH. For this purpose, we used Test Set 3 as it displayed the necessary range of genome completeness values necessary for this purpose, while Test Set 1 was mostly made up of complete genomes and Test Set 2 was mostly made up of low-completeness genomes. We assumed that the predictions yielded by the combined classical approaches represented the true hosts of Test Set 3, although this assumption is likely to lead to an underestimation of the true precision of RaFAH. We found weak positive associations (Pearson $R^2 > 0.6$, $p < 10^{-13}$ for all taxonomic levels) between the precision of RaFAH and genome completeness at all taxonomic levels (Figure S5A). These curves tended to reach a plateau around ~25%–50% genome completeness and increased further for the lower taxonomic ranks (genus, family, and order) for genomes that were >85% complete. Coupled with the observations of the performance of RaFAH on Test Set 2, we suggest that RaFAH is better suited for viral genomes with 50% or more completeness. We used Test Set 3 to analyze the relationship between genome completeness, sequence length, and RaFAH prediction score across the eight different ecosystems (Figure S5B). This revealed a positive correlation between those variables (Pearson R^2 = 0.65, p < 2.2e−16 for the combined set of all ecosystems). Likewise, significant albeit weaker positive correlations were also detected between prediction score and sequence length (Pearson R^2 = 0.14, p <2.2e−16), and prediction score and genome completeness (Pearson R^2 = 0.11, p <2.2e−16). We found that regardless of taxonomic level, precision did not consistently increase through thresholding for genome length, providing further evidence that shorter sequences do not necessarily yield worst predictions (and vice versa) (Figure S5C). These results suggest that the precision of RaFAH cannot be explained by genome length/completeness alone, likely because RaFAH was trained on a dataset with a majority of genome fragments.

We also performed analysis of the combined effects of the relevant variables and how those, together, affected precision, recall, and the F1 score of RaFAH using Test Set 3. Taken together, these results demonstrated that the performance of RaFAH on a given genome is dependent on each of ecosystem source, genome completeness, similarity of the genome to those in the training dataset, and the taxonomic level being considered (see Table S6 at https://doi.org/10.6084/m9.figshare.14365562). For this reason, there is not a single score threshold that is ideal for all use cases. Nevertheless, we make the following recommendations. For differentiating between viruses of Bacteria and Archaea, RaFAH has nearly 100% precision even at the most permissive cutoff (0), thus for this particular purpose it can be applied without threshold. For a broad characterization of multiple viral genomes from an ecosystem, permissive thresholds are acceptable. For example, to compare viral host prevalences across different metagenomes at the level of phylum, we recommend a threshold of 0.14. This yields a precision of approximately 90% without sacrificing recall (Figures S1 and S4D), regardless of ecosystem source, genome length, completeness, or similarity to the training dataset. At lower taxonomic levels, stricter cutoffs are necessary. Users can select cutoffs according to the desired precision based on the curves depicted in Figures S1 and S4D. As a rule, longer, more complete genomes with higher maximum AAI values to genomes in the test set should allow more permissive cutoffs.

Based on the finding that RaFaH achieved nearly perfect precision for domain-level host predictions, and the fact that viruses of Archaea are under-represented in databases, we subsequently focused on the description of these viruses. Few large-scale studies have addressed the diversity of uncultured viruses of Archaea, and they focused mostly on marine samples.[23–26] Here, we describe viruses from seven other ecosystems: soil, permafrost, freshwater, sludge, hypersaline lakes, thermal springs, and the human gut. Applying RaFAH to only eight metagenomic datasets led to the prediction that 537 genomic sequences represent viruses of Archaea (prediction score $\geq$0.14). To put this figure in context, there are only 96 genomes of viruses of Archaea deposited in the NCBI RefSeq database.

We took several steps to ensure that these genomes were truly derived from viruses of Archaea and consistently found compelling evidence to support our claim. First, these genomes could be linked to archaeal genomes either through homology matches or alignment-independent approaches, which provided further evidence that 423 out of the 537 genomes (79%) were indeed

derived from archaeal viruses (Table S2). Second, much like the RefSeq genomes of archaeal viruses, these sequences were enriched in Pfam domains annotated as exclusive of Archaea, eukaryotes, and their viruses (Figure S6). Third, these genomes were enriched in ribosomal binding site motifs that are also enriched among RefSeq viruses of Archaea (Figure S7).

Next, we manually inspected the gene content of the viruses predicted to infect Archaea in search of novel auxiliary metabolic genes (AMGs) and new mechanisms of interaction with the host molecular machinery. The small number of reference genomes of Archaea and their viruses makes it difficult to describe the gene content of the archaeal viruses that we discovered because most of their genes have no taxonomic or functional annotation. However, we found several sequences containing genes coding for thermosomes, group II chaperonins involved in the correct folding of proteins, homologous to their bacterial counterparts, GroEL/GroES.[27] Other AMGs found among archaeal viruses were those involved in the synthesis of cobalamin *cob*S, recently associated with Marine Group I (MGI) Thaumarchaeota virus infection[26] as well as genes that encoded 7-cyano-7-deazaguanine synthase QueC involved in archaeosine tRNA modification.[28] One of the AMGs most prevalent among archaeal viral genomes encoded for a molybdopterin biosynthesis MoeB protein (ThiF family). This family of proteins is involved in the first of the three steps that make up the ubiquitination process.[29] This system regulates several cellular processes through post-translational modification of proteins such as their function, location, and degradation, making it an ideal target from the point of view of viruses to facilitate their replication.[30]

In conclusion, we developed a new tool that uses a random forest classifier based on protein content for virus-host prediction with great potential for studies of viral biodiversity and ecology. RaFAH frequently outperformed other methods that we tested and displayed high accuracy and recall in a dataset of cultured viruses, which extended to uncultured viruses from a diverse set of ecosystems. By analyzing metagenomic datasets from eight different ecosystems, RaFAH allowed for a significant expansion of the archaeal virosphere and shed light on their yet poorly understood content of AMGs. Future studies will describe even more uncultured viral sequences, and RaFAH will likely play a role on describing their hosts and allowing us to decipher their ecological roles. The addition of new viruses with predicted or experimentally verified hosts will allow RaFAH to evolve to identify viruses for an even larger diversity of hosts, and possibly at deeper taxonomic levels such as species. Likewise, these advancements will likely contribute to increasing the accuracy of RaFAH at all taxonomic levels.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Felipe Hernandes Coutinho, fhernandes@icm.csic.es.

Materials availability

This study did not generate new unique materials or reagents.

Data and code availability

All the data (viral and prokaryote genomes) analyzed in this study are freely available from public repositories. The data were also made available as part of the supplemental information. RaFAH and the associated files necessary to run it are freely available online at https://sourceforge.net/projects/rafah/. In addition, we created a Docker container with all the necessary dependencies, scripts, and files available at https://hub.docker.com/r/fhcoutinho/rafah.

Viral genomes database for model training and validation

Two datasets of viral genomes were used for both training and validating the random forest models. The first dataset contained the genomes of viruses of Bacteria and Archaea from NCBI RefSeq available on October 2019, which comprised 2,668 genomes along with their associated host data (Table S3). To avoid overestimating precision due to identical and nearly identical genomes in the database, this dataset was made non-redundant using CD-HIT[31] at a clustering cutoff of 95% identity over 50% alignment of the shorter sequence. The second dataset comprised the 195,698 GLUVAB genomes.[32] GLUVAB is a database of uncultured viral genomes compiled from multiple studies that covered several ecosystems. Only those sequences classified as bona fide viruses of prokaryotes in the original publication were used in subsequent analysis (Table S4).

Classical host prediction for GLUVAB genomes

To use GLUVAB genomes for training and validation of the random forest models, we first had to assign them to putative hosts using classical approaches. To minimize errors during this step we opted for using only alignment-dependent methods due to their higher precision.[2] The RefSeq genomes of Bacteria and Archaea were used as the reference database. We used three lines of evidence for virus-host associations: CRISPR spacers, homology matches, and shared tRNAs. CRISPR spacers were identified in the RefSeq genomes as previously described.[33] The obtained spacers were queried against the sequences of bona fide viral sequences using BLASTn v2.6.0+ (task blastn-short). The cutoffs defined for these searches were minimum identity of 100%, minimum query coverage of 100%, with no mismatches and maximum e-value of 1. Homology matches were performed by querying viral sequences against the databases of prokaryote genomes using BLASTn.[34] The cutoffs defined for these searches were minimum alignment length of 500 bp, minimum identity of 95%, and maximum e-value of 0.001. tRNAs were identified in viral scaffolds using tRNAScan-SE v1.2 [35] using the bacterial models. The obtained viral tRNAs were queried against the RefSeq database of prokaryote genomes using BLASTn. The cutoffs defined for these searches were minimum alignment length of 60 bp, minimum identity of 97%, minimum query coverage of 95%, maximum of 10 mismatches, and maximum e-value of 0.001. These steps for host assignment did not include the prophages in the GLUVAB database, as we were already confident of their host assignments.

We developed a per-viral population scoring method. First, all GLUVAB genomes were clustered into viral populations (VPs) on the basis of 95% average nucleotide identity and 80% shared genes.[36] For each virus-taxon association signal detected (i.e., homology, tRNA, or CRISPR), 3 points were added to the taxon if it was a CRISPR match, 2 points if it was a homology match, and 1 point if it was a shared tRNA. The taxon that displayed the highest score was defined as the host of the viral population. With this approach we ensured that all the genomes in the same VP were assigned to the same host and that no sequences had to be excluded due to ambiguous predictions.

Protein cluster inference and annotation

Protein sequences were identified in viral genomes using Prodigal[37] in metagenomic mode. Hidden Markov models (HMMs) for the phage proteins were built as follows. The 4,701,074 identified proteins were clustered by the cluster workflow of the MMseqs2 software suite,[38] with parameters: 35% sequence identity and alignment coverage had to cover at least 70% of both proteins. Protein clusters (PCs) were aligned into multiple sequence alignments (MSAs) using QuickProbs[39] with default parameters, then converted into HMMs using the hmmake program from the HMMER suite,[40] which resulted in 144,613 HMMs. The HMM profiles were annotated by performing HMM-to-HMM annotation against the pVOG database[41] using the HH-suite3 software suite.[42] First, the MSAs provided on the pVOGs website and the ones built in the previous step were converted into the hhsuite proprietary HMM format using hhmake. The pVOG HHMs were built into an HH-suite3 database, which was then used to find matches to the phage protein HMMs using hhsearch. All HMMs could be annotated through this approach, but only 4,578 matches displayed target coverage $\geq$50% and e-value $\leq 1^{-10}$.

Finally, individual viral proteins were mapped to the HMM profiles using the hmmsearch program limiting hits to those with e-value $\leq 10^{-5}$, alignment length $\geq$70% for both proteins, and minimum score of 50. These results were parsed into a matrix of viral genomes × PCs in which the values of each cell corresponded to the bit score of the best hit of each protein to a given PC, or zero if the protein and the PC did not match or if the score of the match was below the aforementioned 50 cutoff. Once the matrix of genomes × PC was defined, we calculated Pearson correlation coefficients (r) between all possible pairwise combinations of PCs. To remove redundancies, we grouped PCs into superclusters if they presented $r \geq 0.9$, and only a single PC from each supercluster was kept for subsequent analysis. This reduced table of genomes versus PC scores (25,879 genomes × 43,644 PCs) was used as input to train, validate, and test the random forest models.

Random forest training, validation, and testing

Our rationale was that the machine could learn the associations between genes and hosts much more efficiently than a human while also using the information contained in the hypothetical proteins. Hence, random forest models were built using the Ranger[43] package in R.[44] The response variable was the genus-level host assignment of the viral sequences while the input parameters were the scores of viral genomes to each PC. Multi-class random forests were built with 1,000 trees, 5,000 variables to possibly split at in each node, and using probabilistic mode. This classification approach ensured that a single model could be used for all virus genomes. The putative host of a viral genome was selected as the taxon with the highest probability score yielded by the random forest. The taxonomic classification of each genus up to the domain level was obtained by parsing the NCBI Taxonomy database with a custom script. Next, variable importance was estimated using the impurity method. When training the models and reporting predictions, we assumed that a virus can only infect a single genus. Due to the probabilistic nature of the random forests, all genera are associated with a score (which ranges from 0 to 1). Users interested in multi-genera viruses can search for those genomes that have close or equal scores as preliminary evidence that the viral genome in question might infect across multiple genera.

Three models were built and validated on independent datasets. Model 1 was trained on Training Set 1, which comprised 80% randomly selected non-redundant viral genomes from NCBI RefSeq. The performance of this model was evaluated on Training Set 1 and Validation Set 1, which comprised the remaining 20% of non-redundant RefSeq genomes. This process was repeated for a 10-fold cross-validation. Even without thresholding, these models exhibited high precision for both the training (mean 99.96% ± 0.026%) and validation sets (mean 76.47% ± 1.523%) at the genus level. Model 2 was trained on Training Set 2, which comprised 100% of the RefSeq genomes, and validated on Validation Set 2, which was comprised of GLUVAB genomes that could be assigned to a host at the level of genus by the pipeline described above. Finally, Model 3 was built based on Training Set 3, which comprised all of the RefSeq viral genomes and the GLUVAB genomes that could be assigned to a host at the level of genus (i.e., a combination of Training Set 2 and Validation Set 2). In this dataset each genus was represented by a median of three genomes, and for 187 out of 617 (30.3%) genera the model was trained with a single genome (Table S5). Models 1 and 2 were used as proof-of-principle models, and Model 3 was the definitive model used for testing and which is provided to the users and used for all subsequent analyses.

Viral genome completeness is likely to influence the performance of the models. A tool trained solely on complete or nearly complete genomes might not be capable of producing accurate predictions for the genome fragments that are often obtained with metagenomic datasets. Completeness of the 25,879 sequences used to train RaFAH was estimated with CheckV,[15] which indicated that this dataset encompassed both complete viral genomes as well as partial viral contigs. Partial viral genomes were the majority of sequences used to train RaFAH. Altogether, the genomes used for training displayed an average completeness of 53.6% ± 32.3%. According to CheckV, these sequences were classified as complete genomes (709 sequences), high-quality genome fragments (5,823), medium-quality genome fragments (5,493), low-quality genome fragments (13,707) and not determined (147).

We used three independent test sets to evaluate the performance of RaFAH Model 3. Test Set 1 comprised viral genomes retrieved from NCBI Genomes database in January 2021. We took several steps to make sure that Test Set 1 represented a challenging dataset for the random forest model so as to assess its ability to extrapolate. First, we excluded from Test Set 1 any genomes made public before November 2019. Second, Test Set 1 was made non-redundant at 95% nucleotide identity and 50% alignment length of the shorter sequence. Third, protein sequences derived from Test Set 1 were compared with the protein sequences of Training Set 3 using DIAMOND.[45] Any genomes that shared more than 70% of proteins or more than 70% average AAI with any genome from Training Set 3 were removed from Test Set 1. These steps resulted in an independent Test Set 1 consisting of 561 (out of the initial 3,427) genomes with no overlap to the genomes used to train the models.

Test Set 2 comprised viral genomes identified in SAGs from marine samples.[14] A total of 4,751 SAGs (with completeness $\geq$50% and contamination $\leq$5% as estimated by CheckM)[46] were classified at the level of genus using BAT.[47] This algorithm provides taxonomic affiliations to microbial genomes based on consensus taxa of proteins matches to the NCBI-nr database. Next, viral sequences were extracted from the SAGs using VIBRANT,[48] which identified 418 viral sequences. We assumed that the viral sequences in the SAGs infected the organisms from which these SAGs were derived, either because they were derived from integrated prophages or from viral particles attached or inside host cells. Viral sequences for which the host taxon predicted by RaFAH was the same taxon of the SAG as determined by BAT were considered as correct host predictions. Viruses from SAGs that could not be classified were excluded from the precision and recall analyses.

Test Set 3 comprised a collection of 61,647 viral genomic sequences from studies that spanned multiple samples from permafrost,[8] marine,[49] human gut,[50] freshwater,[19] soil,[51] hypersaline lakes,[52] hydrothermal springs (Fredrickson et al., unpublished data obtained from IMG/VR),[53] and sludge bioreactor[18] habitats. These sequences were assigned to putative hosts through the classical host-prediction pipeline described above for the GLUVAB genomes and also using RaFAH. Bootstrap analysis was applied to evaluate the precision of RaFAH in this dataset. For this, we assumed that the hosts predicted by the classical approaches were the true hosts of the viral genomes on Test Set 3. Random subsamples representing 20% of the full data were generated in 1,000 replicates. Precision was estimated for each replicate. Also, we estimated the completeness of viral genomes on Test Set 3 with CheckV[15] and analyzed the association between genome completeness and the precision of RaFAH.

RaFAH was tested on an Intel Xeon Gold 6140 CPU @ 2.30-GHz machine. Timing calculations were performed using randomly selected genomes of Test Set 3 using 24 threads in both the training and prediction modes (Figure S9). These results showed that the time to perform computations varied exponentially as a function of input genomes. Using 10,000 input genomes, RaFAH took 184 min to fit models and 495 min to predict hosts.

Comparison with other methods for host prediction

To assess the performance of RaFAH compared with other host-prediction tools, we assessed the performance of the alignment-free methods HostPhinder[5] and WIsH,[3] the alignment-dependent approaches based on homology matches, shared tRNAs and CRISPR spacers (and the three combined as described above for assigning hosts to GLUVAB genomes), and a hybrid approach, VirHostMatcher-Net.[6] We compared these tools on Test Sets 1, 2, and 3. HostPhinder, VirHostMatcher-Net, and WIsH were run with default parameters. The classical host predictions (CRISPR, tRNA, and homology matches) for Test Set 1 were performed using the same parameters described above for the GLUVAB genomes and for Test Set 3. Three performance metrics were evaluated at different taxonomic levels (domain to genus): Recall is the percentage of viral sequences for which a host was predicted by a given tool. Each viral sequence that was associated to a host was counted toward recall, regardless of the host association being correct or not. Recall was calculated as the number of sequences associated with a host divided by the total number of sequences in the dataset. For approaches that provided multiple host predictions for the same viral sequence (i.e., homology matches, tRNA, and CRISPR), each individual viral sequence counted toward recall only once. Precision is the percentage of host predictions that were correct. Each viral sequence that was associated with a host by a given tool was counted toward precision if the host association matched the true host of the sequence. Precision was calculated as the total of matching host predictions divided by the total number of predictions. Approaches that provided multiple host predictions for the same viral sequence counted toward precision if at least one

of the predictions was correct, but each sequence was counted toward precision only once. Finally, the F1 score was calculated as the harmonic mean between precision and recall.

For the approaches that required reference host genomes (i.e., WIsH, CRISPR, tRNA, and homology matches), the database of host genomes was the NCBI RefSeq genomes of Bacteria and Archaea and the genomes of Uncultured Bacteria and Archaea from the Genome Taxonomy Database.[54] To minimize false positives due to homology between viruses and mobile genetic elements, we removed all sequences that matched the keyword "plasmid" in their description field from the database of reference host genomes.

Assessment of archaeal virus-host predictions

To confirm the prediction of 537 genomes predicted by RaFAH as archaeal viruses, we used Mash v.2.1.[55] Mash calculates Jaccard distance between two genomes based on the number of shared *k*-mers with a certain length. We used *k*-mer sizes from 13 to 20 nucleotides. For each *k*-mer size we calculated distances of every phage genomic sequence against all potential host genomes. This database included 17,134 bacterial genomes and 4,716 archaeal genomes retrieved from RefSeq and GenBank. For each phage genome, we selected the potential host with the smallest Mash distance. In addition to Mash distance, we also calculated Manhattan distances and correlation scores between phage and host *k*-mer frequencies using $k = 6$ as described in Edwards et al.[2] and Ahlgren et al.[4] Finally, all 537 phages were used as BLASTn queries against the whole NR database. For each phage we determined a potential host by selecting the top-scoring non-viral hit as described in Edwards et al.[2] In addition, we compared the prevalence of ribosomal binding site motifs (defined by Prodigal[37] gene predictions) between viral sequences predicted to infect Bacteria and Archaea, from both the eight metagenomic datasets and RefSeq viruses. A similar analysis was performed to compare the prevalence of Pfam domains among these groups. For this analysis, protein sequences were queried against the Pfam database using hmmsearch with maximum e-value set to 10^{-3}.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100274.

ACKNOWLEDGMENTS

This work was supported by grants "VIREVO" CGL2016-76273-P [MCI/AEI/FEDER, EU] (cofounded with FEDER funds) from the Spanish Ministerio de Ciencia e Innovación and "HIDRAS3" PROMETEU/2019/009 from Generalitat Valenciana. F.R.-V. was also a beneficiary of the 5top100-program of the Ministry for Science and Education of Russia. F.H.C. was supported by APOSTD/2018/186 post-doctoral fellowships from Generalitat Valenciana. A.Z. was funded by the Polish National Science Centre (2018/31/D/NZ2/00108). J.B.'s research was supported by the National Center for Research and Development (NCBR, Poland), grant number LIDER/5/0023/L-10/18/NCBR/2019. B.E.D. was supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004 and by the European Research Council Consolidator grant 865694: DiversiPHI. R.E. was supported by National Institutes of Health grant RC2 DK116713-01A1.

AUTHOR CONTRIBUTIONS

F.H.C. conceived and designed the experiments. F.H.C., A.Z.-S., M.L.-P., A.Z., J.B., B.E.D., and R.E. analyzed the data. All authors contributed to writing the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 28, 2020
Revised: November 23, 2020
Accepted: May 7, 2021
Published: June 15, 2021

REFERENCES

1. Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., Poulos, B.T., Solonenko, N., Lara, E., Poulain, J., et al. (2016). Ecogenomics and biogeochemical impacts of uncultivated globally abundant ocean viruses. Nature *537*, 589–693.
2. Edwards, R.A., McNair, K., Faust, K., Raes, J., and Dutilh, B.E. (2016). Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol. Rev. *40*, 258–272.
3. Galiez, C., Siebert, M., Enault, F., Vincent, J., and Söding, J. (2017). WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics *33*, 3113–3114.
4. Ahlgren, N.A., Ren, J., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017). Alignment-free d2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res. *45*, 39–53.
5. Villarroel, J., Kleinheinz, K.A., Jurtz, V.I., Zschach, H., Lund, O., Nielsen, M., and Larsen, M.V. (2016). HostPhinder: a phage host prediction tool. Viruses *8*, 116.
6. Wang, W., Ren, J., Tang, K., Dart, E., Ignacio-Espinoza, J.C., Fuhrman, J.A., Braun, J., Sun, F., and Ahlgren, N.A. (2020). A network-based integrated framework for predicting virus–prokaryote interactions. NAR Genomics Bioinform. *2*, 505768.
7. Ignacio-Espinoza, J.C., Ahlgren, N.A., and Fuhrman, J.A. (2020). Long-term stability and Red Queen-like strain dynamics in marine viruses. Nat. Microbiol. *5*, 265–271.
8. Emerson, J.B., Roux, S., Brum, J.R., Bolduc, B., Woodcroft, B.J., Jang, H.B., Singleton, C.M., Solden, L.M., Naas, A.E., Boyd, J.A., et al. (2018). Host-linked soil viral ecology along a permafrost thaw gradient. Nat. Microbiol. *3*, 870–880.
9. López-Pérez, M., Haro-Moreno, J.M., Gonzalez-Serrano, R., Parras-Moltó, M., and Rodriguez-Valera, F. (2017). Genome diversity of marine phages recovered from Mediterranean metagenomes: size matters. PLoS Genet. *13*, e1007018.
10. Zhang, M., Yang, L., Ren, J., Ahlgren, N.A., Fuhrman, J.A., and Sun, F. (2017). Prediction of virus-host infectious association by supervised learning methods. BMC Bioinformatics *18*, 60.
11. Young, F., Rogers, S., and Robertson, D.L. (2020). Predicting host taxonomic information from viral genomes: a comparison of feature representations. PLoS Comput. Biol. *16*, e1007894.
12. Leite, D.M.C., Brochet, X., Resch, G., Que, Y.-A., Neves, A., and Peña-Reyes, C. (2018). Computational prediction of inter-species relationships through omics data analysis and machine learning. BMC Bioinformatics *19*, 420.
13. de Jonge, P.A., Nobrega, F.L., Brouns, S.J.J., and Dutilh, B.E. (2018). Molecular and evolutionary determinants of bacteriophage host-range. Trends Microbiol. *27*, 51–63.
14. Pachiadaki, M.G., Brown, J.M., Brown, J., Bezuidt, O., Berube, P.M., Biller, S.J., Poulton, N.J., Burkart, M.D., La Clair, J.J., Chisholm, S.W., et al. (2019). Charting the complexity of the marine microbiome through single-cell genomics. Cell *179*, 1623–1635.e11.
15. Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2020). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. https://doi.org/10.1038/s41587-020-00774-7.
16. Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Structure and function of the global ocean microbiome. Science *348*, 1261359.
17. Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011). Enterotypes of the human gut microbiome. Nature *473*, 174–180.

18. Liu, R., Qi, R., Wang, J., Zhang, Y., Liu, X., Rossetti, S., Tandoi, V., and Yang, M. (2017). Phage-host associations in a full-scale activated sludge plant during sludge bulking. Appl. Microbiol. Biotechnol. *101*, 6495–6504.

19. Kavagutti, V.S., Andrei, A.-Ş., Mehrshad, M., Salcher, M.M., and Ghai, R. (2019). Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. Microbiome *7*, 135.

20. Johnston, E.R., Hatt, J.K., He, Z., Wu, L., Guo, X., Luo, Y., Schuur, E.A.G., Tiedje, J.M., Zhou, J., and Konstantinidis, K.T. (2019). Responses of tundra soil microbial communities to half a decade of experimental warming at two critical depths. Proc. Natl. Acad. Sci. *116*, 201901307.

21. Ghai, R., Pašić, L., Fernández, A.B., Martin-Cuadrado, A.B., Mizuno, C.M., McMahon, K.D., Papke, R.T., Stepanauskas, R., Rodriguez-Brito, B., Rohwer, F., et al. (2011). New abundant microbial groups in aquatic hypersaline environments. Sci. Rep. *1*, 135.

22. Menzel, P., Gudbergsdóttir, S.R., Rike, A.G., Lin, L., Zhang, Q., Contursi, P., Moracci, M., Kristjansson, J.K., Bolduc, B., Gavrilov, S., et al. (2015). Comparative metagenomics of eight geographically remote terrestrial hot springs. Microb. Ecol. *70*, 411–424.

23. Philosof, A., Yutin, N., Flores-Uribe, J., Sharon, I., Koonin, E.V., and Béjà, O. (2017). Novel abundant oceanic viruses of uncultured marine group II Euryarchaeota identified by genome-centric metagenomics. Curr. Biol. *27*, 1362–1368.

24. Vik, D.R., Roux, S., Brum, J.R., Bolduc, B., Emerson, J.B., Padilla, C.C., Stewart, F.J., and Sullivan, M.B. (2017). Putative archaeal viruses from the mesopelagic ocean. PeerJ *5*, e3428.

25. Ahlgren, N.A., Fuchsman, C.A., Rocap, G., and Fuhrman, J.A. (2019). Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. ISME J. *13*, 618–631.

26. López-Pérez, M., Haro-Moreno, J.M., de la Torre, J.R., and Rodriguez-Valera, F. (2019). Novel caudovirales associated with marine group I Thaumarchaeota assembled from metagenomes. Environ. Microbiol. *21*, 1980–1988.

27. Marine, R.L., Nasko, D.J., Wray, J., Polson, S.W., and Wommack, K.E. (2017). Novel chaperonins are prevalent in the virioplankton and demonstrate links to viral biology and ecology. ISME J. *11*, 2479–2491.

28. Turner, B., Burkhart, B.W., Weidenbach, K., Ross, R., Limbach, P.A., Schmitz, R.A., de Crécy-Lagard, V., Stedman, K.M., Santangelo, T.J., and Iwata-Reuyl, D. (2020). Archaeosine modification of archaeal tRNA: role in structural stabilization. J. Bacteriol. *202*. https://doi.org/10.1128/JB.00748-19.

29. Makarova, K.S., and Koonin, E.V. (2010). Archaeal ubiquitin-like proteins: functional versatility and putative ancestral involvement in tRNA modification revealed by comparative genomic analysis. Archaea *2010*, 9–13.

30. Randow, F., and Lehner, P.J. (2009). Viral avoidance and exploitation of the ubiquitin system. Nat. Cell Biol. *11*, 527–534.

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

32. Coutinho, F.H., Edwards, R.A., and Rodríguez-Valera, F. (2019). Charting the diversity of uncultured viruses of Archaea and Bacteria. BMC Biol. *17*, 109.

33. Díez-Villaseñor, C., and Rodriguez-Valera, F. (2019). CRISPR analysis suggests that small circular single-stranded DNA smacoviruses infect Archaea instead of humans. Nat. Commun. *10*, 294.

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

35. Lowe, T.M., and Chan, P.P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. *44*, W54–W57.

36. Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J.M., et al. (2015). Patterns and ecological drivers of ocean viral communities. Science *348*, 1261498.

37. Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.

38. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028.

39. Gudyś, A., and Deorowicz, S. (2017). QuickProbs 2: towards rapid construction of high-quality alignments of large protein families. Sci. Rep. *7*, 41553.

40. Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A., and Eddy, S.R. (2015). HMMER web server: 2015 update. Nucleic Acids Res. *43*, W30–W38.

41. Grazziotin, A.L., Koonin, E.V., and Kristensen, D.M. (2017). Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. *45*, D491–D498.

42. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics *20*, 473.

43. Wright, M.N., and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. *77*. https://doi.org/10.18637/jss.v077.i01.

44. R Core Team (2016). R : A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

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

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

47. von Meijenfeldt, F.A.B., Arkhipova, K., Cambuy, D.D., Coutinho, F.H., and Dutilh, B.E. (2019). Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. Genome Biol. *20*, 530188.

48. Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome *8*, 90.

49. Luo, E., Eppley, J.M., Romano, A.E., Mende, D.R., and DeLong, E.F. (2020). Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. ISME J. *14*, 1304–1315.

50. Gregory, A.C., Zablocki, O., Howell, A., Bolduc, B., and Sullivan, M.B. (2019). The human gut virome database. bioRxiv. https://doi.org/10.1101/655910.

51. Göller, P.C., Haro-Moreno, J.M., Rodriguez-Valera, F., Loessner, M.J., and Gómez-Sanz, E. (2020). Uncovering a hidden diversity: optimized protocols for the extraction of dsDNA bacteriophages from soil. Microbiome *8*, 17.

52. Roux, S., Enault, F., Ravet, V., Colombet, J., Bettarel, Y., Auguet, J.C., Bouvier, T., Lucas-Staat, S., Vellet, A., Prangishvili, D., et al. (2016). Analysis of metagenomic data reveals common features of halophilic viral communities across continents. Environ. Microbiol. *18*, 889–903.

53. Paez-Espino, D., Roux, S., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K., Huntemann, M., Reddy, T.B.K., Pons, J.C., Llabrés, M., et al. (2018). IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. Nucleic Acids Res. *47*, 678–686.

54. Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. *36*, 996–1004.

55. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132.

RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content

Felipe Hernandes Coutinho,[1,*] Asier Zaragoza-Solas,[1] Mario López-Pérez,[1] Jakub Barylski,[2] Andrzej Zielezinski,[3] Bas E. Dutilh,[4,5] Robert Edwards,[6] and Francisco Rodriguez-Valera[1,7]

[1]Evolutionary Genomics Group, Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández, 03550 San Juan de Alicante, Spain
[2]Molecular Virology Research Unit, Faculty of Biology, Adam Mickiewicz University Poznan, 61-614 Poznan, Poland
[3]Department of Computational Biology, Faculty of Biology, Adam Mickiewicz University Poznan, 61-614 Poznan, Poland
[4]Centre for Molecular and Biomolecular Informatics (CMBI), Radboud University Medical Centre/Radboud Institute for Molecular Life Sciences, 6525 GA Nijmegen, the Netherlands
[5]Theoretical Biology and Bioinformatics, Science for Life, Utrecht University (UU), 3584 CH Utrecht, the Netherlands
[6]College of Science and Engineering, Flinders University, Bedford Park, SA 5042, Australia
[7]Moscow Institute of Physics and Technology, Dolgoprudny 141701, Russia

*Corresponding Author: Felipe Hernandes Coutinho, fhernandes@icm.csic.es

SUPPLEMENTARY FILES

Tables S1 - S5 can be found at the publisher's website:
https://doi.org/10.1016/j.patter.2021.100274

Supplementary material:

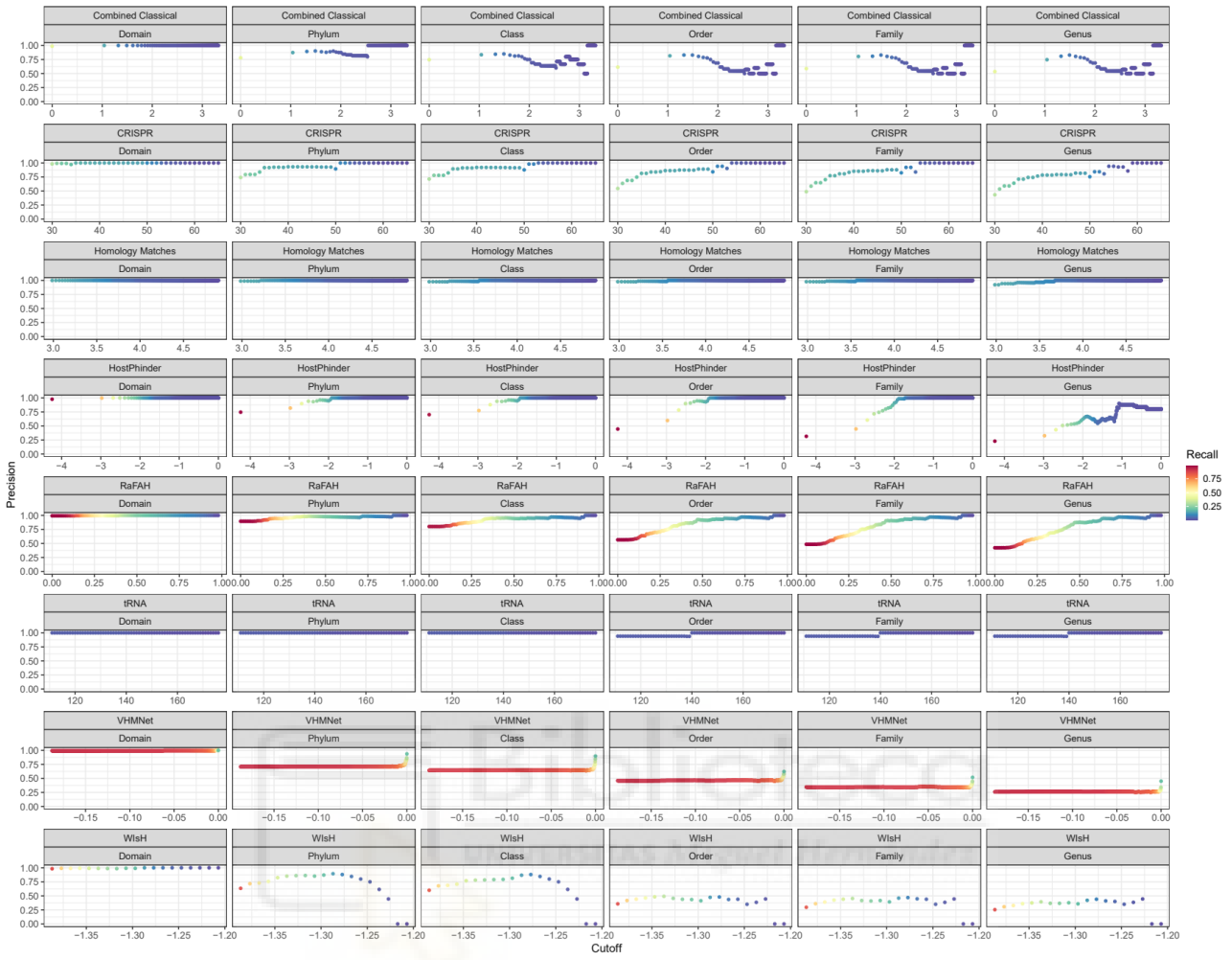


Figure S1: Associations between score cutoff, precision, and recall for RaFAH, the alignment-free (WIsH and HostPhinder), hybrid (VirHostMatcher-Net), and classical (CRISPR, tRNA and homology matches) host prediction approaches on Test Set 1. The score cutoffs for HostPhinder, Homology matches, VirHostMatcher-Net and Combined Classical are shown in the Log10 scale.

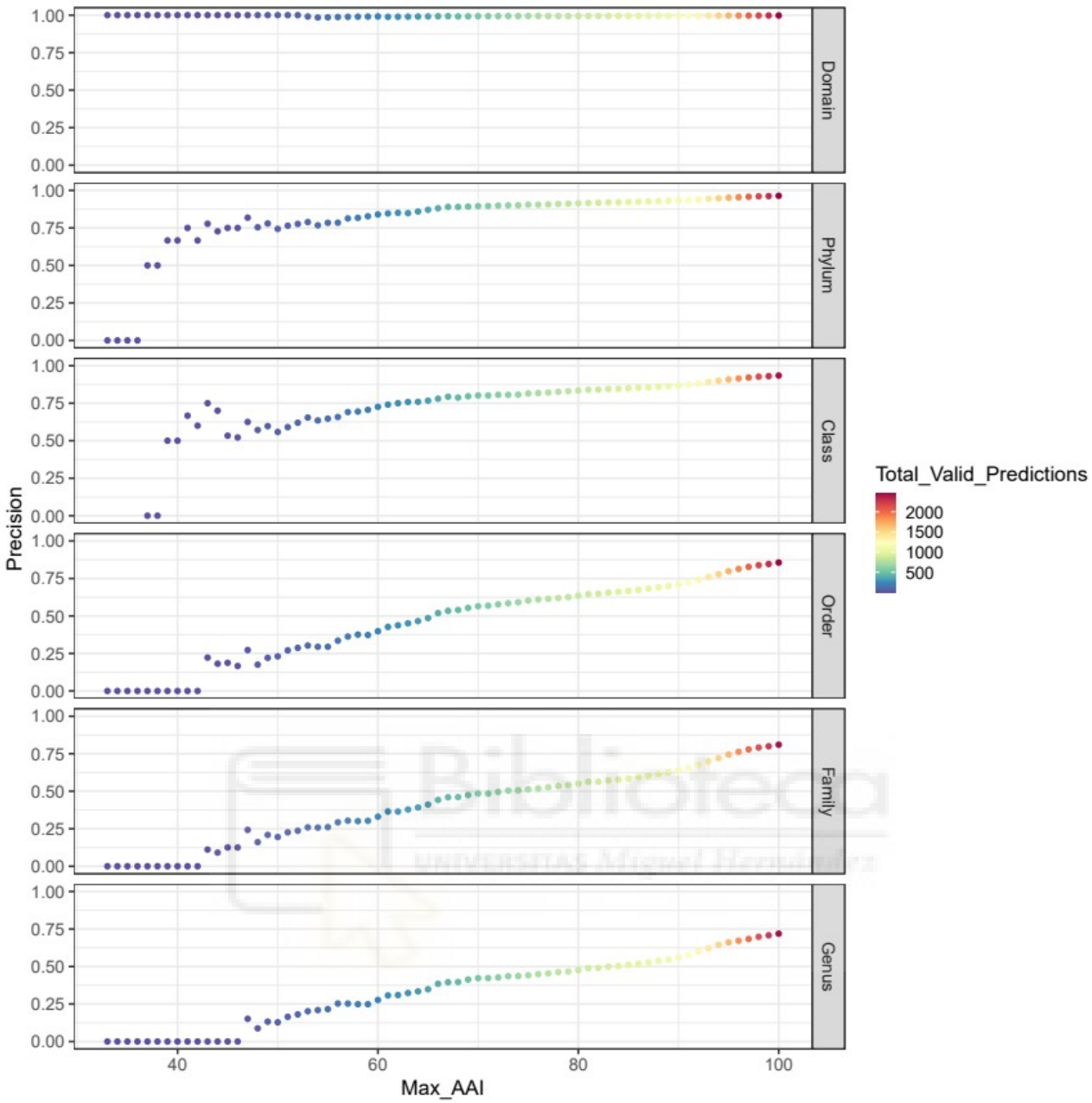


Figure S2: Associations between precision and similarity among Test Set 1 and Training Set 3 genomes. Each panel represents a different taxonomic level. X axis displays the maximum Average Amino acid Identity (AAI) among genomes of the two sets. Y axis displays the precision of RaFAH. Points are coloured according to the number of valid predictions (host taxon predicted by RaFAH for a non "NA/undef/Unknown" host genome in Test Set 1) yielded at each taxonomic level and AAI cutoff. For this particular analysis all non-redundant genomes in Test Set 1 were used while in all other instances this dataset was filtered for maximum 70% AAI and 70% matched proteins.

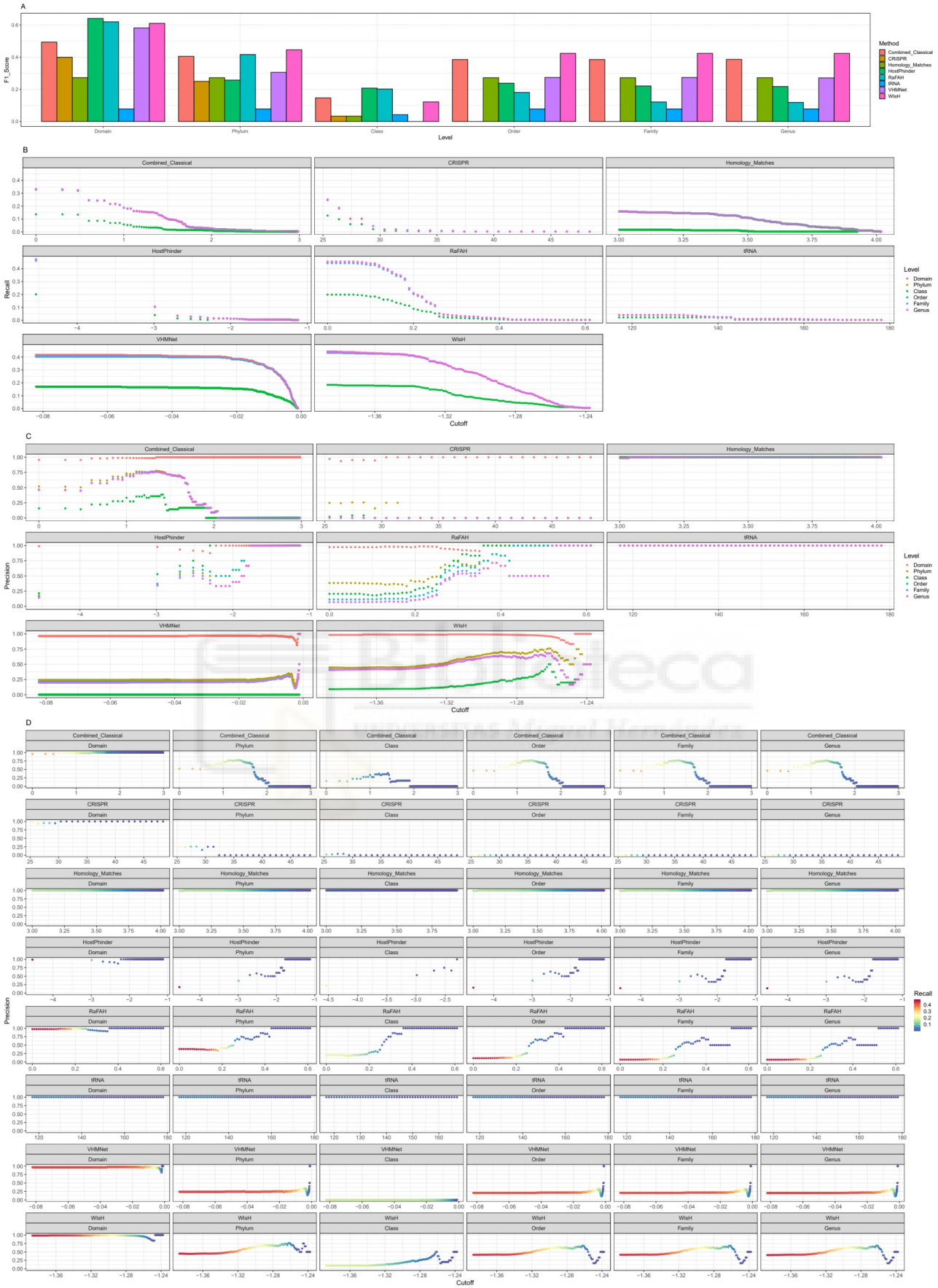


Figure S3: Performance of host prediction tools on Test Set 2: Associations between score cutoff, precision, and recall for RaFAH, the alignment-free (WIsH and HostPhinder), hybrid

(VirHostMatcher-Net), and classical (CRISPR, tRNA and homology matches) host prediction approaches. A) F1-score of methods when considering all predictions regardless of score at multiple taxonomic levels. B) Association between score cutoff and recall for each taxonomic level. C) Association between score cutoff and precision for each taxonomic level D) Associations between precision and recall in function of score cutoff. Figure S8 depicts the association between precision and score cutoff of VirHostMatcher-Net for score values above the 75th percentile.

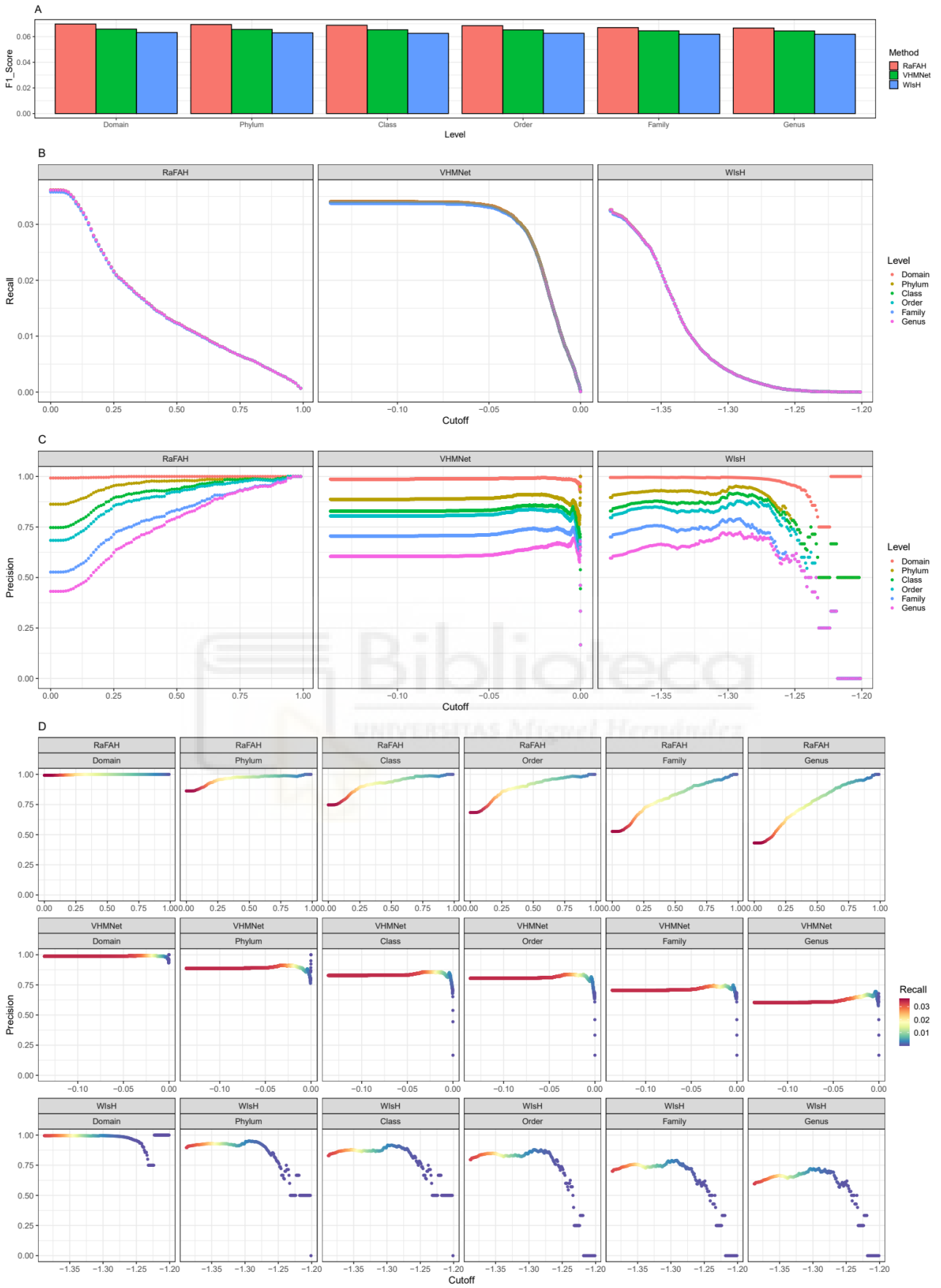


Figure S4: Performance of host prediction tools on Test Set 3: Associations between score cutoff, precision, and recall for RaFAH, WIsH and VirHostMatcher-Net. The hosts assigned by the

combined classical approaches were considered the true hosts of the genomes in Test Set 3. A) F1-score of methods when considering all predictions regardless of score at multiple taxonomic levels. B) Association between score cutoff and recall for each taxonomic level. C) Association between score cutoff and precision for each taxonomic level. D) Associations between precision and recall in function of score cutoff. Figure S8 depicts the association between precision and score cutoff of VirHostMatcher-Net for score values above the 75th percentile.

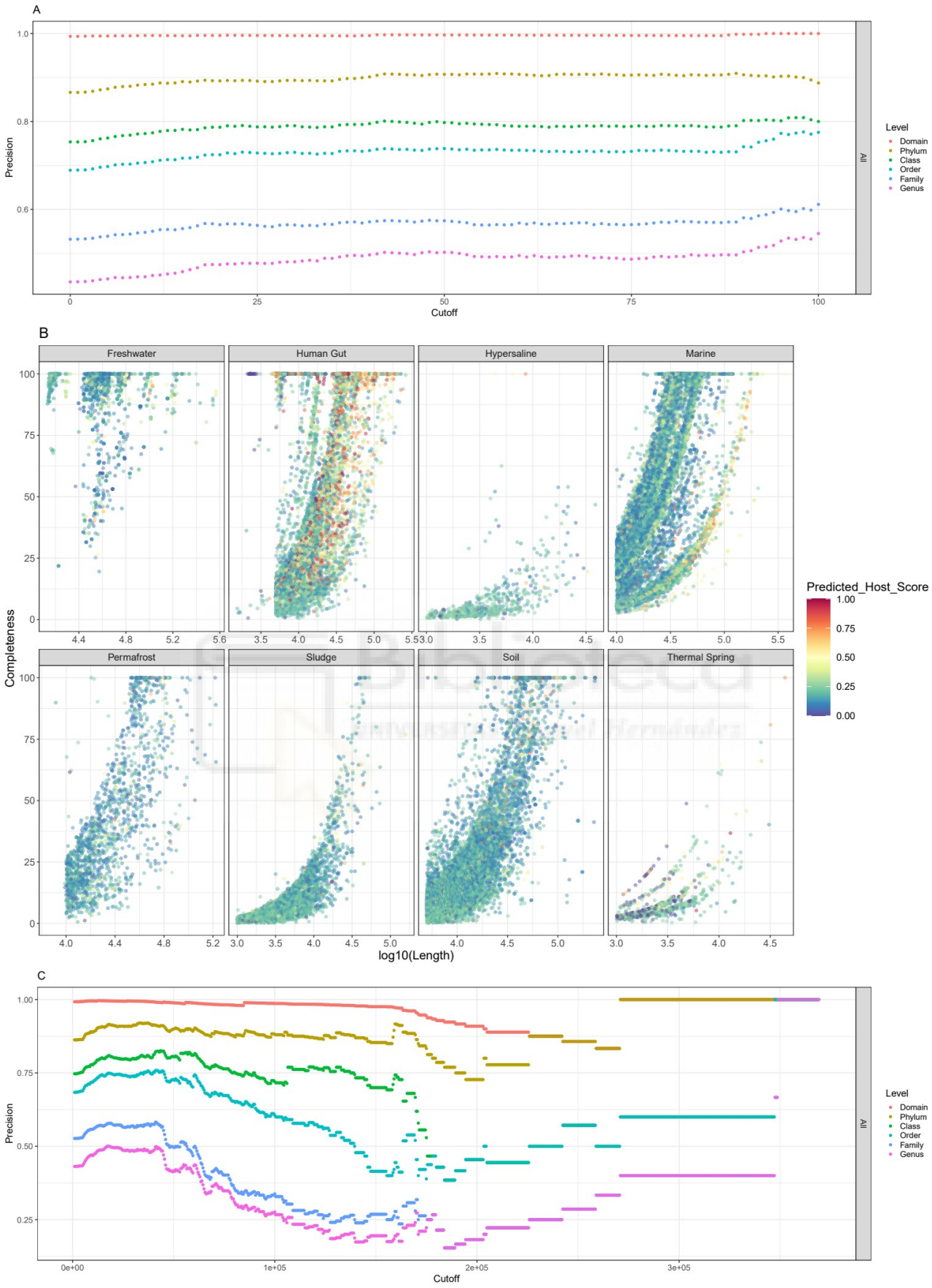


Figure S5: Associations between the performance of RaFAH and genome length/completeness on Test Set 3 genomes. A) Scatterplot displaying the cutoff for genome completeness (X axis) and

precision of RaFAH (y axis). B) Association between genome legngth (X axis) and genome completeness (Y axis) estimated with CheckV across 8 ecosystems (Panels). C) Scatterplot displaying the cutoff for genome length (X axis) and precision of RaFAH (y axis).

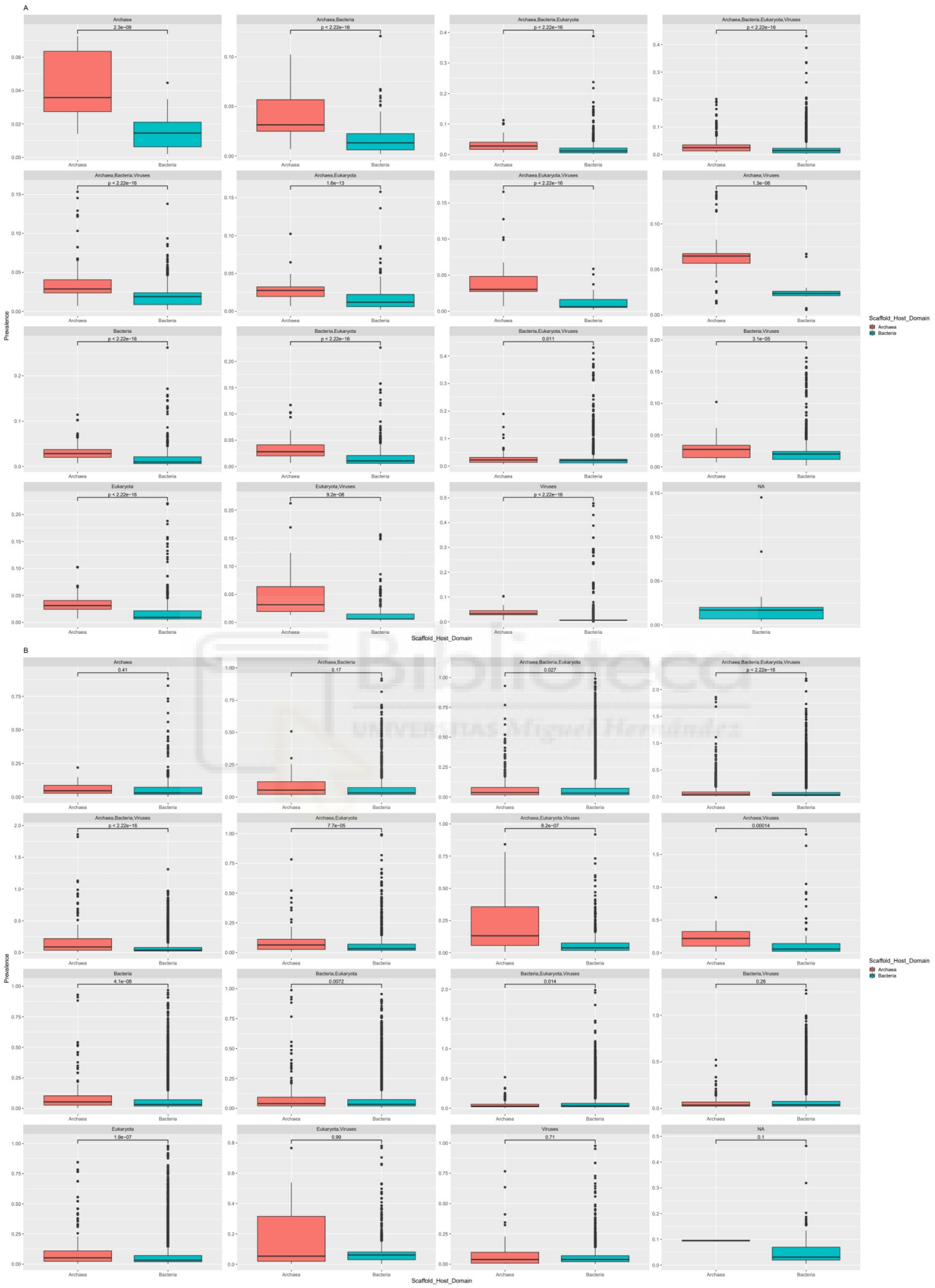


Figure S6: Prevalence of Pfam domains among viruses. Pfam domains were grouped according to their expected taxonomic ranges (depicted above each panel). Only values derived from scaffolds

with at least 5 CDS are shown to reduce noise. A) Comparisons of Pfam domain prevalence between RefSeq viruses of Archaea and Bacteria. The *p* values of each comparison obtained with the Mann-Whitney test are depicted above bars. B) Pfam domain prevalence between RefSeq viruses of Archaea and of Bacteria from TestSet3. Notice the different y axes on each panel.

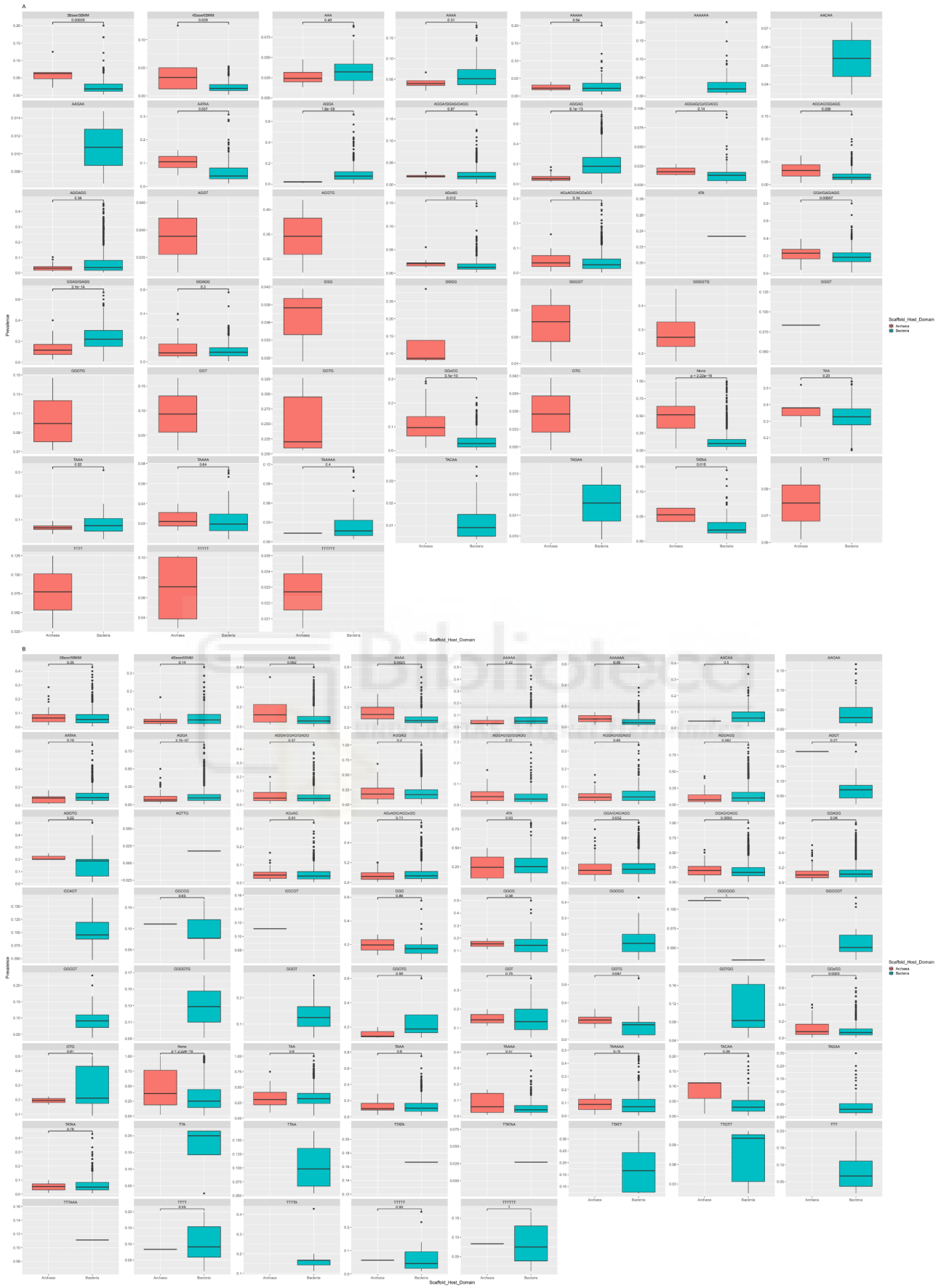


Figure S7: Prevalence of ribosomal binding site (RBS) motifs among viruses. Only values derived from scaffolds with at least 5 CDS are shown to reduce noise. A) Comparisons of RBS motif

prevalence between RefSeq viruses of Archaea and Bacteria. The *p*-values of each comparison obtained with the Mann-Whitney test are depicted above bars. B) Comparisons RBS motif prevalence between RefSeq viruses of Archaea and Bacteria and viruses from TestSet3.

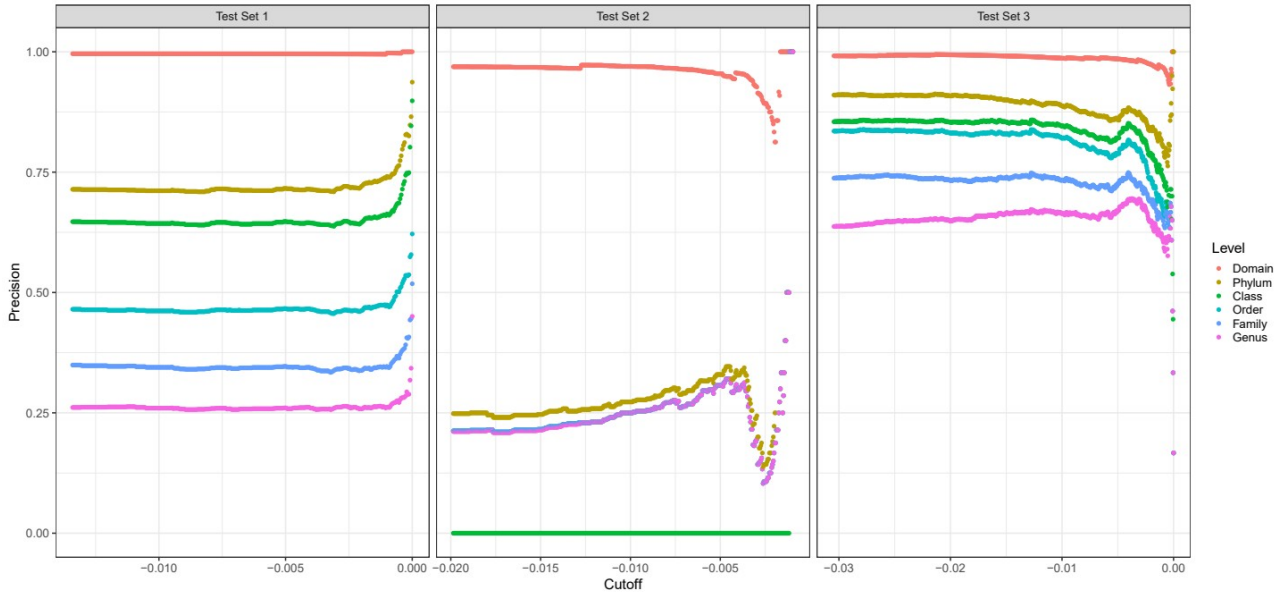


Figure S8: Association between precision and score cutoff for VirHostMatcher-Net in Test Sets 1, 2 and 3. All scores below the 75th percentile value were excluded from this analysis.

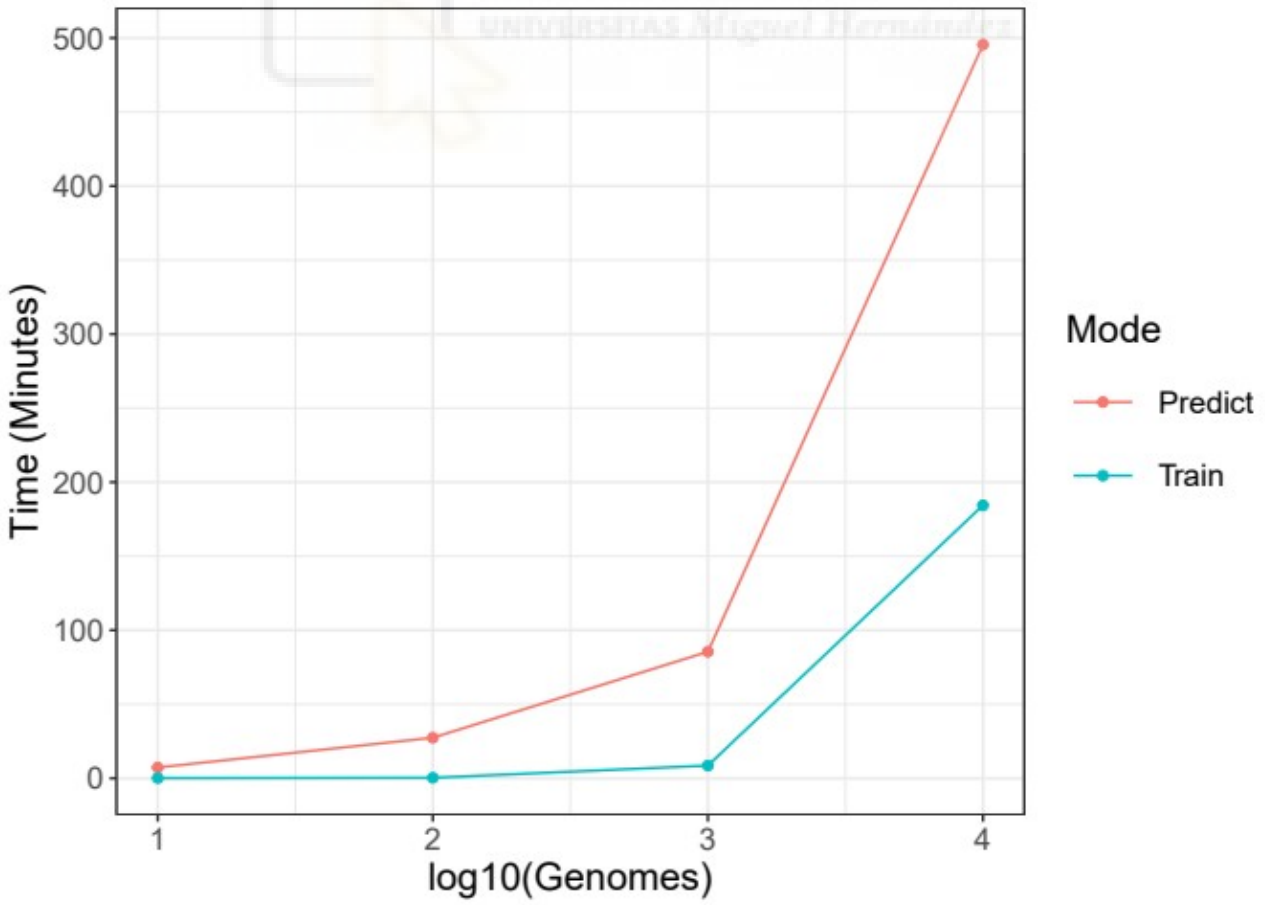


Figure S9: Timing of RaFAH computation on Training and Prediction modes (Y axis) as a function of the number of input genomes (X axis). Calculations were performed using randomly selected genomes of Test Set 3 on an Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz machine using 24 threads.

Supplementary Data 1: Fasta file containing the nucleotide sequences from NCBI RefSeq and GLUVAB viruses that made up Training Sets 1, 2, and 3. Due to file size Supplementary Data 1 is available on Figshare with DOI: https://doi.org/10.6084/m9.figshare.14208500

Supplementary Data 2: Fasta file containing the nucleotide sequences from NCBI RefSeq that made up Testing Set 1. Due to file size Supplementary Data 2 is available on Figshare with DOI: https://doi.org/10.6084/m9.figshare.14210591

Supplementary Data 3: Fasta file containing the nucleotide sequences from SAG derived viruses that made up Testing Set 2. Due to file size Supplementary Data 3 is available on Figshare with DOI: https://doi.org/10.6084/m9.figshare.14208506

Supplementary Data 4: Fasta file containing the nucleotide sequences from metagenome derived viruses that made up Testing Set 3. Due to file size Supplementary Data 4 is available on Figshare with DOI: https://doi.org/10.6084/m9.figshare.14210612

AGRADECIMIENTOS

Tras cinco años y una pandemia, me encuentro en la línea de meta de mi doctorado y, tras recapitular todo lo vivido, me doy cuenta de que acabo esta etapa en mucho mejor estado (como científico y como persona) de cómo la empecé. Me gustaría dedicar unas palabras a las personas que me han ayudado a llegar hasta aquí:

En primer lugar, más que nada porque sin él nada de esto hubiera sido posible, me gustaría agradecer a Ricardo Mallavia su ayuda desinteresada con aquél estudiante recién llegado de Glasgow que quería empezar un doctorado en la UMH, aunque fuera sin financiación. Sin su ayuda, Paco jamás habría encontrado mi CV y no habría tenido la gran oportunidad de trabajar en su laboratorio.

Después, querría agradecer a mis dos directores de tesis, Paco y Mario, por su infinita paciencia y sus sabios consejos. Todo lo que he aprendido estos últimos años os lo debo a vosotros. Como último doctorando de Paco, me uno a una larguísima estela de microbiólogos formados en su laboratorio. Sólo espero poder continuar investigando y por tanto, contribuyendo a su legado como mentor.

Por supuesto, no puedo olvidarme de mis compañeros de laboratorio, con los que he pasado penurias y alegrías y que se han demostrado ser verdaderos amigos, tanto dentro como fuera del laboratorio. Josema y PedroJ me han dado tantos consejos que casi podrían ser mi tercer director de tesis; Felipe y Rafa, compañeros fagólogos; Juanjo, siempre dispuesto a echar una mano; Ricardo, César, Aurelia, Raquel… Muchas gracias por todo!

Let me switch to English for a bit, just in case Rohit Ghai reads this. Thank you so much for letting me stay in your lab for three months! It was short but I did learn a lot. I have not forgotten about those freshwater phages, I promise! I also want to thank Dina, Paul, Vinicius and all the other people from České Budějovice for making my stay there as memorable as it was.

Y finalmente, pero no por ello menos importantes, quiero agradecer a las personas que siempre han estado ahí. Gracias a mis padres por mantenerme, tanto económicamente como emocionalmente. Gracias a Celia, que aunque nos llevemos ocho años es la mejor hermana que se podría imaginar. Gracias a Alberto, compañero de penurias durante la carrera y que incluso dedicó un verano a ayudarme a acabar un programa del laboratorio. Gracias a mi grupo de Amigos de alicante: Jose, Josué, Carlos, Albert, Andrei y Sergi, que son el mejor grupo de amigos que he tenido nunca. Y, sobre todo, gracias a Inés, que siempre ha estado conmigo de una forma o de otra. Gracias por las charlas hasta bien entrada la madrugada, los viajes en verano y tu capacidad para comprenderme y apoyarme en mis momentos más bajos.