Unraveling the Eco-evolutionary Complexity of Uncultivated Bacteriophages in the Biosphere

Alaina R. Weinheimer

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Biological Sciences

Frank O. Aylward
Roderick Jensen
Ann Stevens
Liqing Zhang

November 8, 2022
Blacksburg, VA

jumbo bacteriophages, marine phage ecology, bacteriophage evolution, phage complexity, bacteriophage metagenomics

Unraveling the Eco-evolutionary Complexity of Uncultivated Bacteriophages in the Biosphere

Alaina R. Weinheimer

Academic Abstract

Bacteriophages, or phages, have historically been distinguished by their small sizes and relatively simple genomes compared to cellular life. Discoveries over recent decades, however, have uncovered remarkably large phages, called jumbo phages, which are defined by having genomes over 200 kilobases and contain virion sizes comparable to small bacteria. These exceptionally large phages prompt questions on how such complexity emerges and persists in the virosphere, when being simple is so successful with shorter replication times and larger burst sizes. This dissertation aims to address these knowledge gaps by examining the evolutionary and ecological contexts of genomic and community-level complexity of phages using a variety of metagenomic datasets, namely from marine environments. Toward understanding the coexistence of jumbo phages among smaller phages, Chapter 1 provides a literature review on jumbo phage diversity, associated fitness tradeoffs of largeness, and predictions on which environments or ecological conditions may be enriched in jumbo phages. Chapter 2 assesses the evolutionary context giving rise to complex phages, by examining a group of phages that encode a multi-subunit DNA-dependent RNA polymerase homologous to that of cells. This gene fortuitously enabled phylogenetic analyses of phages with cellular life and revealed that these phages likely emerged prior to the divergence of bacteria and archaea, rather than acquiring the gene from their hosts more recently. Chapter 3 examines the biogeography of genomic complexity in the ocean by identifying and comparing groups of jumbo phages in seawater metagenomes of the global ocean. This work revealed that jumbo phages with distinct replication

machinery also have distinct distributions, with some groups more common in surface waters than deeper waters and vice versa. Chapter 4 compares drivers of phage complexity at the community level (based on diversity) with the drivers of prokaryotic community diversity by examining seawater metagenomes from contrasting ecosystems off the coasts of the Isthmus of Panama. Despite phages' requiring their hosts to replicate, the results show that factors increasing phage and prokaryotic diversity do not always align. This discrepancy highlights the role the environment also plays in governing virus-host interactions, such as impacting dispersal ranges and adsorption efficiency. Collectively, this dissertation addresses how, what, and where complexity in the virosphere occurs using culture-independent methods and contributes to our growing understanding of the breadth of viral diversity and ecology.

Unraveling the Eco-evolutionary Complexity of Uncultivated Bacteriophages in the Biosphere

Alaina R. Weinheimer

General Audience Abstract

While many viruses cause disease and threaten animal and plant health, most viruses on Earth infect microbes, which are tiny, single-celled organisms like bacteria. These viruses can be used to kill harmful bacteria, like certain *Escherichia coli* (*E. coli*), and they impact the movement of nutrients in ecosystems because microbes like algae form the basis of food webs in the sea. While most known viruses are very tiny, larger viruses have been recently discovered over recent decades. Being a big virus can be very costly, as it takes more resources for these viruses to replicate or reproduce. Despite these costs, big viruses can be found in many environments around the world, such as the human intestine and the deep sea, which suggests that being large as a virus might be useful in some circumstances. This dissertation aims to uncover how, why, and where being large as a virus is most successful. This research specifically focuses on a group of viruses called phages, which are viruses that infect microbes called bacteria and archaea. Larger phages, those with genomes four times the size of most other phages and twenty times the size of the COVID19 virus, are called jumbo phages. Chapter 1describes the diversity of jumbo phages, what advantages they may have over smaller phages and which environments these advantages may be most helpful. Chapter 2 examines how complex phages evolved by analyzing a group of phages that have a special gene that is also found in all cellular life (microbes, plants, and animals). The evolutionary history of this gene suggests that phages possessed this gene prior to the emergence of major cellular groups (bacteria, archaea, and eukarya), rather than stealing this gene from their host more recently. Chapter 3 uncovers where

different types of jumbo phages are most prevalent in the ocean; some are more common in surface waters, and some are more common in deeper waters. Finally, Chapter 4 aims to understand the complexity of phage communities in terms of where phages are most diverse. We found they are more diverse in habitats where bacterial diversity is lower, which is unexpected but shows that the environment plays a major role in virus-host interactions. Overall, this dissertation uncovers the diversity, distribution, and origins of complexity in phages and phage communities, so that we can better understand how they impact the environment and affect microbes that power ecosystems.

Dedication

This work is dedicated to my family, in particular my parents who have inspired, supported, and encouraged me to pursue my passions, work hard, and believe in myself – from wanting to be an interior designer, actress, and art historian to now. I would also like to mention my adventurous grandmothers, Janet Grady and Lynn Gillespie (who passed away just months ago), as they have inspired me to continue learning, traveling, and exploring throughout life.

Acknowledgements

Firstly, I would like to thank my advisor Dr. Frank Aylward for his unwavering support, encouragement, and endless patience over the past few years. Working in his lab has been a massive privilege, and his exceptional mentorship and wisdom has greatly developed my abilities and confidence to think creatively, tackle challenging questions, and of course troubleshoot my numerous code errors. He fostered a supportive, friendly lab environment, and I am grateful for the stimulating conversations and helpful feedback over the years from current and former lab members: Carolina Martinez Gutierrez, Dr. Mohammad Moniruzzaman, Nitin Niar, Sangita Karki, Dr. Anh Ha, Roxanna Farzad, and Paula Erazo. I am grateful for having an insightful and encouraging committee of Dr. Ann Stevens, Dr. Liqing Zhang, and Dr. Rick Jensen to guide me through these diverse projects. I am particularly grateful for having gotten to further learn from the wisdom of Dr. Stevens as a professor of one of my courses and through the Microbiology Club. I would like to thank the Biological Science department as a whole and especially the staff and faculty who have helped make my grad student experience more comfortable, such as Dr. Jeff Walters, Dr. Ignacio Moore, Rebecca Zimmerman, and Dr. Robert Cohen. I already miss chats with Dr. Cohen and seeing Socrates on my walks home, and I greatly appreciate his support in activities of the Biology Grad Student Association which strengthened the strong feeling of community within the department. Additionally, I would like to thank Dr. Mary Kasarda, Mike Ervine, Shelley Johnson, and others of ICTAS for providing me both financial and academic support over the years.

I would also like to thank the Global Change Center and the Interfaces of Global Change community, especially Dr. Bill Hopkins and Jessica Zielske. Being a part of the IGC has led to many lasting friendships, has broadened my perspective of science and society, and has led to some exciting collaborations during my time at Virginia Tech that have truly made my graduate

experience a joy amidst the inherent stress. I particularly appreciate the support from Dr. Hopkins in encouraging my scientific pursuits and sharing my appreciation for tailgating and little beers, as well as his enthusiastic support for the numerous social activities I proposed.

I would like to thank Dr. Jarrod Scott and Dr. Matthieu Leray at the Smithsonian Tropical Research Institute for their guidance and support through our collaboration on Chapter 4 and the development of new project ideas. I have always left our meetings inspired and invigorated to tackle new questions.

Finally, I would like to thank my family and friends who have helped me stay grounded and have fun throughout graduate school. I am so lucky to have grown so close to so many lovely humans, especially living with "Good Enough" House pod: Chloe Moore, Dr. Sarah Kuchinsky, Lisa Tabor, Earl Gilbert, and Travis Vorhees through the height of the pandemic. I'd like to thank the IGCheers crew: Korin Rex, Isaac Van Diest, Chloe Moore, Camilo Alfonso, and Devin Hoffman for their incredible friendship and numerous bizarre but entertaining discussions at Rivermill (How many is a few? Three? Five?) that have reminded me that science is as fun as it is serious. I'd also like to thank the Bio Babes group for many fun evenings, ladies nights, and vent sessions that have kept morale high through graduate school. I'd particularly like to thank Carolina for her friendship and encouragement as we've weathered the highs and lows of a PhD together, from many delicious breakfast vent sessions to navigating ISME together. I'd also like to thank my best friend Savannah for keeping my perspective throughout the PhD and the many laughs and adventures we've had in the beautiful Appalachia. I'd like to thank my parents and siblings for their support over the years that included visiting me on multiple occasions which meant the world to me. Finally, I'd like to thank my partner Keir (and Penguin!) for his love and support throughout the PhD.

Table of Contents

Attributions

Chapter 1: Literature review on jumbo phage diversity, evolution, and fitness tradeoffs

This chapter has one additional author:

Frank O. Aylward, Department of Biological Sciences, Blacksburg, VA, USA

Chapter 2: A distinct lineage of Caudovirales that encodes a deeply branching multi-subunit RNA polymerase

Permission to reproduce this article has been granted by the journal *Nature Communications*, online version: https://doi.org/10.1038/s41467-020-18281-3

This chapter has one additional author:

Frank O. Aylward, Department of Biological Sciences, Blacksburg, VA, USA

Chapter 3: Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages

Permission to reproduce this article has been granted by the journal *ISME,* online version: https://doi.org/10.1038/s41396-022-01214-x

This chapter has one additional author:

Frank O. Aylward, Department of Biological Sciences, Blacksburg, VA, USA

Chapter 4: Contrasting drivers of abundant phage and prokaryotic communities in tropical, coastal ecosystems across the Isthmus of Panama

This chapter has three additional authors:

Frank O. Aylward, Department of Biological Sciences, Blacksburg, VA, USA

Matthieu Leray, Smithsonian Tropical Research Institute, Balboa, Ancon, Republic of Panama

Jarrod Scott, Smithsonian Tropical Research Institute, Balboa, Ancon, Republic of Panama

Chapter 1: Literature review on jumbo phage diversity, evolution, and fitness tradeoffs

Alaina R. Weinheimer and Frank O. Aylward

Historical context and motivation

In 2002, Harald Brüssow and Roger Hendrix published in *Cell,* "Phage genomics: small is beautiful," which celebrated the vast diversity and possibilities of discoveries through phages afforded by their small genomes relative to cells[1]. Phages are viruses of bacteria and archaea, and most known phages belong to the class Caudoviricetes, which have double stranded DNA genomes and are distinguished from other phages by their tailed capsids (Figure 1a). They use this tail to attach to hosts and inject their DNA into the host cell to begin an infection, during which they either begin replicating their genomes and making progeny particles immediately ultimately leading to the lysis of their host to release their offspring in the lytic cycle (Figure 1b), or they first integrate into the host genome in the lysogenic cycle and only excise from the host genome to being their replication. Because phages can rely on their hosts' machinery for their replication, they need few genes, with most Caudoviricetes having genomes of 50 kilobases, or 100 times smaller than a typical bacterium.

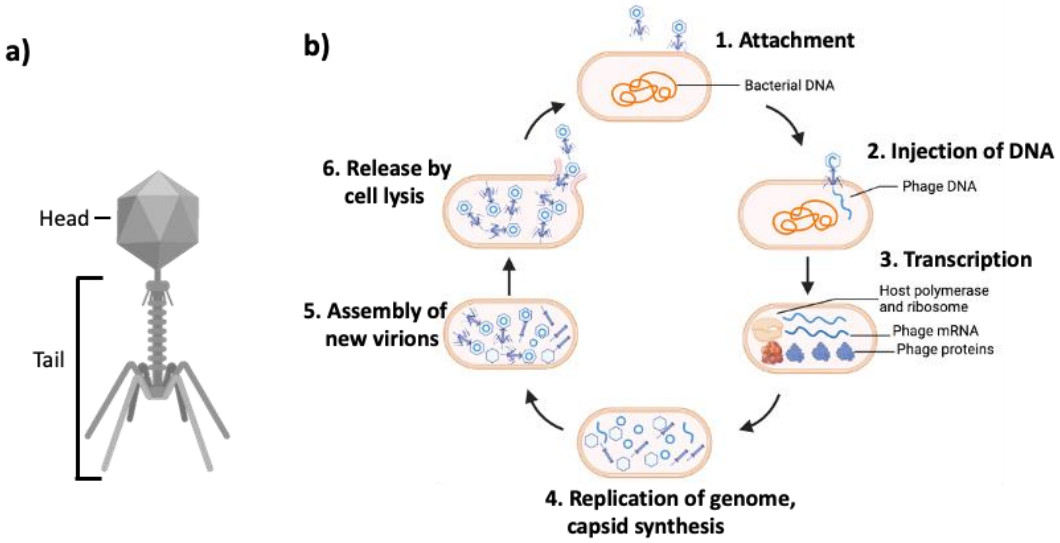


Figure 1. (a) Morphology of a Caudoviricetes capsid with the head and tail designated. **(b)** Diagram of phage replication in the lytic cycle. (Icons courtesy of BioRender).

Research on these simple entities over the past century has led to major breakthroughs in molecular biology, such as the findings that DNA is the heritable material between generations and that three nucleotides code for a single amino acid[2]. While scientific research has benefited from the compact size and simplicity of phage genomes, evolutionary pressures may have enforced smallness in the virosphere, as phages with smaller genomes replicate quicker and produce more progeny per infection than larger phages[3]. Despite the apparent fitness advantages of being small, however, larger and more complex phages have been uncovered through both isolation and culture-independent sequencing approaches used in a range of environments and on a diversity of hosts[4]. One group of these larger phages are called jumbo phages, which belong to the class Caudoviricetes and are solely distinguished from other Caudoviricetes by having genomes over 200 kilobases[5]. With larger genomes, jumbo phages have been found to encode a variety of functions beyond the essentials, such as tubulin proteins that form nucleus-like

compartments during infections to protect their replication from host defenses like CRISPR-Cas[6,7]. The remarkable complexity of jumbo phages and their increased isolation, along with their growing application in phage therapy treatments have revived interest in jumbo phages over the past decade[8–10].

Jumbo phages, however, are much less frequently isolated than smaller phages, representing approximately 2.3% of cultured phages available on the NCBI's Viral Genome Portal in July 2021. Their apparent scarcity may not only have resulted from methodological biases that favor the enrichment of smaller phages (e.g. filter pore sizes, agar concentrations)[11,12], but also be due to fitness drawbacks of being big, such as longer replication periods and smaller burst sizes[3]. Nevertheless, the global distribution of jumbo phages[8,10] indicates that the benefits of having a larger genome and thus more genes outweigh the costs in diverse conditions. In this review, I examine the ecological contexts that would bolster the existence of the complex jumbo phages and develop predictions on which environments may be enriched in jumbo phages. To inform these predictions, I first summarize the diversity of jumbo phages and how they might have emerged. I then assess fitness tradeoffs of largeness. Next, I discuss the hosts and current known distributions of jumbo phages, as these may relate to their fitness tradeoffs, despite methodological biases. Finally, I predict the ecological contexts that favor largeness in the form of environmental stability, productivity, and microbial diversity. This work will not only inform viral ecology and evolution, but also assist in the development of phage therapies and other biotechnological applications of jumbo phages.

Jumbo phage diversity

While jumbo phages, and phages in general, are incredibly diverse, jumbo phages do share a number of features. As members of the class Caudoviricetes, they are tailed, double stranded

DNA viruses that infect bacteria[5]. Jumbo phages have yet to be isolated with archaea[13]. Because genome size correlates allometrically with capsid size for dsDNA phages[14], jumbo phages tend to have larger capsid sizes, with some reaching the size of small bacteria[15]. The morphologies of cultured jumbo phages fall into the siphovirus and myovirus groups[13] (Figure 2a,b). Siphoviruses have long tails, while myoviruses have medium-sized, contractile tails[2]. None have been isolated with a podovirus morphology in which the tail is much shorter than the capsid head (Figure 2c). This may be due to methodological biases or an unknown physical constraint that a shorter tail imposes during the injection of phage DNA into the host upon infection.

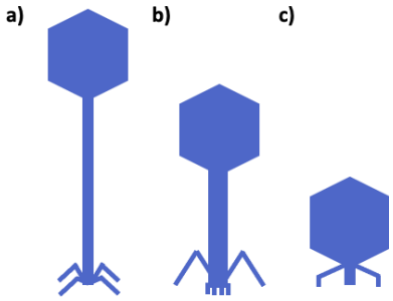


Figure 2. Morphologies of siphovirus (**a**), myovirus (**b**), and podovirus (**c**) capsids.

Regarding their replication, all cultured jumbo phages are strictly lytic[10] (Figure 1b). Although most lytic phages rely heavily on the temporal organization of their genes to regulate transcription, jumbo phage genomes are highly mosaic, and all encode transcription factors to compensate[5]. Notably, while DNA polymerases can be found in some smaller phages, all jumbo phages encode at least one DNA polymerase[16]. The type of DNA polymerase, however, varies between jumbo phages and includes classical T7-type DNA polymerase, a bacterial-like DNA polymerase III, and a divergent family B DNA polymerase, among others. Finally, another common feature of jumbo phages is the presence of tRNAs in their genomes. Some smaller phages also carry tRNAs, but all cultured jumbo phages carry at least one and some up to 22 tRNAs[16]. In fact, a metagenomic survey on jumbo phages predicted up to 67 tRNAs in one of the

complete, assembled genomes[8]. In short, jumbo phages primarily all have large genome sizes, mosaic genomes, at least one DNA polymerase, at least one transcription factor, and at least one tRNA.

Despite these common features of jumbo phages, a recent comparative genomic study by Iyer et al (2021)[16] uncovered three distinct groups of jumbo phages based on shared genes that primarily function in replication. Briefly, Group 1 jumbo phages encode many genes of the *Pseudomonas aeruginosa* phage PhiKZ, such as a divergent family B polymerase, multisubunit RNA polymerase, and the tubulin proteins mentioned that create the remarkable nucleus-like structure it creates during infection to shield phage replication from host defenses. Group 2 jumbo phages share many genes with T4-like phages, such as a T4 UsvW/poxviral A18 type DNA helicases, and a gp23 major capsid proteins. These jumbo phages also encode a classical family B DNA polymerase and their own sigma factors. Group 3 jumbo phages include two subgroup groups with one that is distinguished by a coliphage T7-type DNA polymerase and the other distinguished by a DNA polymerase III type similar to bacteria. A recent survey that examined jumbo phages in metagenomes of seawater from the global ocean suggests that these groups even have distinct biogeography, with Group 1 and Group 3 jumbo phages found in deeper water and Group 2 jumbo phages more present in surface waters (See Chapter 3, Weinheimer and Aylward 2022[12]). Hundreds of uncultivated jumbo phages have been detected in other recent metagenomic surveys[8,17], suggesting there may be even more, novel replication strategies to distinguish groups of jumbo phages. Notably, roughly 86% of the genes encoded by marine jumbo phages in Weinheimer and Aylward 2022 had no hits to databases despite using very sensitive approaches, highlighting the vast untapped diversity and novel genes to uncover from examining jumbo phages.

Jumbo phage emergence

Because phages lack a high resolution, phylogenetic marker gene analogous to the small subunit ribosomal RNA (rRNA) of cells, reconstructing the evolutionary history of phages, and thus jumbo phages, poses many challenges. Nevertheless, several lines of evidence support multiple, independent origins of jumbo phages. To start, jumbo phages have both sipho- and myovirus morphologies, which are thought to correspond to disparate evolutionary groups[16]. Furthermore, a phylogenetic study using concatenated alignments of conserved genes for cultured phages[18], in addition to a phylogenetic analysis from a study of metagenomic phages that only used the terminase large subunit gene[8], found that jumbo phages were present within clades of smaller phages. Also, the jumbo phage groups of the Iyer et al study included smaller non-jumbo phages, namely with genomes of at least 180 kb. The processes or events that could lead to largeness in phages also support the possibility of distinct, parallel origins of jumbo phages from smaller phages.

A popular mechanism to explain the emergence of jumbo phages is the ratchet model, proposed by R. W. Hendrix in 2009[5]. In this model, a mutation that results in the expansion of the capsid also leads to more genetic material being loaded into the capsid for phages, which could eventually mutate into unique, functional genes. While a mutation could also occur that shrinks that capsid size, adding less DNA to a capsid could result in the loss of essential genes. Thus, the upward trajectory of genome size gives the model its name of the "ratchet model" because genome size ratchets up. Headful packaging is a mechanism used by diverse phages, and thus, this phenomenon could occur independently in different lineages of phages. In support of this model, a study in capsid size by Hua et al 2017[19] reported much higher percent terminal redundancy packaged into the jumbo phage capsids than in the non-jumbo phage capsids. I performed a T-test using the Table 2 of Hua et al 2017, which revealed non-jumbo phages had

significantly lower percentages of terminal redundancy than the jumbo phages, with a mean of 4% compared to the jumbo phages with a mean of 18.83%, which I plotted in the boxplots of Figure 3a (*p value* < 0.05). Interestingly, however, genome length did not significantly correlate with percent terminal redundancy upon my analysis and my plotting of terminal redundancy versus genome length (Figure 1b; p value > 0.05, Pearson correlation). The variation in terminal redundancy between the jumbo phages may signify different time points of capsid expansion. As suggested by Hua et al 2017, perhaps a higher terminal redundancy resulted from a recent expansion in the capsid size as the additional genes have not had sufficient time to functionally diverge. Taken together, the clustering of jumbo phages within groups or clades of smaller phages, along with the proposed ratchet model of genome size expansion, point to multiple, independent origins of jumbo phages.

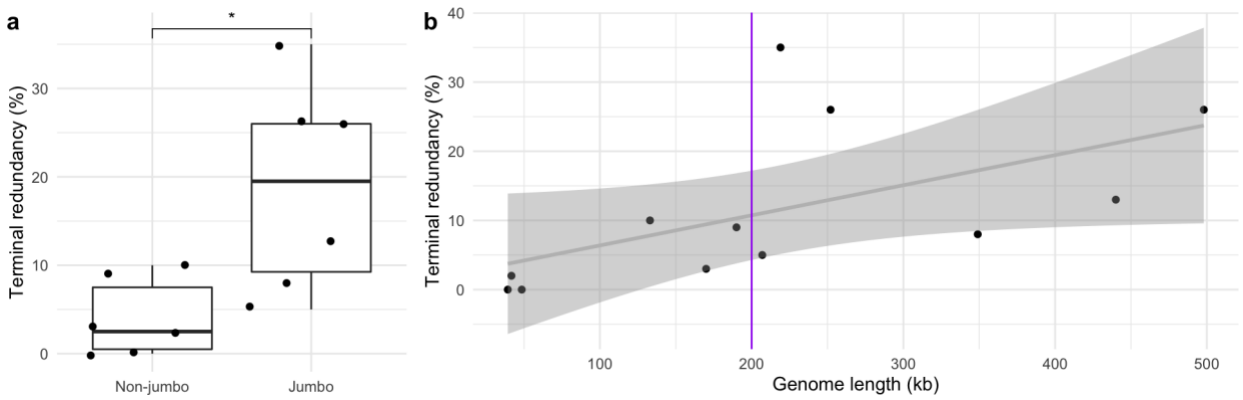


Figure 3. Percent terminal redundancy and genome length in phages. (**a**) Boxplot comparing terminal redundancy percentages between non-jumbo phages and jumbo phages (* = p value < 0.05). (**b**) Scatterplot of percent terminal redundancy versus genome length. Gray line corresponds to the linear regression calculated with pearson correlation (p value 0.053) and standard error shaded in gray. Purple line drawn at the jumbo phage mark of 200 kb. Data used from Hua et al 2017.

Not only have jumbo phages likely evolved on multiple, independent occasions, but they also may be quite ancient relative to today's cellular lineages. Some jumbo phages have been found to encode proteins that are more similar to archaeal and eukaryotic versions rather than to bacterial versions, such as a peptidyl tRNA hydrolase (PTH2)[20]. Furthermore, a phylogenetic analysis of DNA-dependent multi-subunit RNA polymerases homologous to that of cells found in metagenomic phages that likely correspond to jumbo phages have also suggested the divergence of jumbo phages prior to the splitting of the archaeal and bacterial lineages (See Chapter 2 for Weinheimer and Aylward 2020)[21]. Additionally, some jumbo phages have been found to use uracil in place of thymine of their DNA, which may be a remnant of the RNA world thought to have preceded the evolution of cells which all have thymine in their DNA[22]. Taken together, largeness in the virosphere is likely an ancient strategy that emerged on multiple, distinct occasions.

Fitness tradeoffs of jumbo phages

The large diversity in phage genome sizes points to the different components of viral fitness that counterselect for smaller and larger viruses. Viral fitness, as in reproductive success, has been calculated as a function of several core components formulated by Edwards et al 2021[3] based on models from Levin et al 1977[23]. Briefly, this equation relates increased viral fitness with higher rates of effective adsorption, larger burst sizes, lower decay rates, lower host mortality rates, and shorter latency periods. Effective adsorption refers to the successful attachment of viruses to their hosts that leads to an infection. Burst size refers to the number of phages produced per infection. Latency period refers to the length of time between injecting DNA into the cell and bursting the cell to release progeny. As this review focuses on jumbo phages, I will refer to these fitness components from the perspective of jumbo phages. Much of the benefits discussed below

overlap with the extensive synthesis of Edwards et al 2021 which reviewed genome size tradeoffs of aquatic viruses that included eukaryotic viruses, such as the Nucleocytoviricota. While the general benefits apply to both prokaryotic and eukaryotic viruses, the underlying biology of these benefits of course differ which result in disparate ecological conditions where each group is favored (See Section below *Jumbo phage distribution)*.

Fitness drawbacks of jumbo phages

The fitness disadvantages of jumbo phages largely boil down to their having smaller burst sizes, longer latency periods, and potentially lower effective adsorption rates. Jumbo phages require more resources per progeny virus compared to their smaller counterparts[24], which can result in fewer progeny per infection. While some phages may have metabolic genes to compensate for this[16], jumbo phage burst sizes may also be constrained by the physical size of their hosts, as their larger capsids occupy more volume[14]. Additionally, the smaller burst size further compounds when resources are limited for the host, such as under nutrient starvation[25,26]. Latency periods are longer for larger viruses[3,27] likely because replicating a longer genome and assembling a larger particle can take longer, and this also elongates under nutrient stress[25,26]. Nevertheless, some larger phages have developed strategies to compensate for these disadvantages, Edwards et al 2021 found latency periods are only 2 -3 times longer for phages that are 10-fold larger. They suggest that larger viruses are more efficient at capitalizing on their hosts' resources, and they calculated that nucleotide production rates of the largest viruses are roughly 100 times faster than that of the smallest viruses. The extensive replication genes of larger phages[16] may account for this efficient extraction, elaborated in the *Fitness advantages of jumbo phages* section below.

Finally, the impact of largeness on effective adsorption rates remains unclear and is likely context dependent. Effective adsorption rates depend on both (i) the contact rate between a phage and its host and (ii) the frequency of adsorption or the irreversible binding between the phage and its host that leads to infection[3]. Because larger viruses have fewer progeny released per infection (lower burst sizes), and their progeny are released less frequently (longer latency periods) compared to their smaller counterparts, they may be less prevalent in the environment and thus have lower contact rates. Furthermore, larger virions tend to have lower diffusivity than smaller viruses[3], which could reduce contact rates even amidst higher jumbo phage densities. Some jumbo phage capsids contain elaborate protein complexes and/or more capsid proteins that allow them bind to their hosts and stay attached to their hosts more efficiently compared to smaller phages[10], potentially resulting in jumbo phages' having higher adsorption or affinity to their hosts than smaller phages. Analysis of compiled adsorption rates by Edwards et al 2021, however, found that adsorption rates were not correlated with genome size for phages, despite a strong correlation for eukaryotic viruses. Thus, the extent that these capsid features meaningfully compensate for the lower prevalence and lower diffusivity of jumbo phages to increase effective adsorption rates remains unclear and potentially depends on the environment (i.e. high turbulence vs. low turbulence [28]).

Fitness advantages of jumbo phages

While mo' genes mean mo' problems, sometimes mo' is mo'. For jumbo phages, their fitness advantages can primarily be attributed to having more genes. These extra genes have been grouped into three broad categories for the purposes of this review: capsid or structural genes, replication and other metabolic processes, and defenses, summarized in Table 1. Specifics on

each category are detailed below. Jumbo phages are known to have broader host ranges than smaller phages, and this is likely due to a combination of added genes in all categories which is elaborated on in its own section below.

Capsid genes: increase stability, increase adsorption probability

Virion decay has been shown to be associated with physical stressors (e.g. UV light), adsorption to non-hosts or other materials, and grazing by organisms, among other factors[29–32]. Several studies have found that larger virions generally persist longer in the environment. For instance, Heldal and Bratbak 1991[33] observed that virions above 60 nm decayed slower than smaller virions, and an analysis by Edwards et al 2021 on data from De Paepe and Taddei 2006[34] on phage multiplication rates and decay rates found that decay rate decreased with increasing virus diameter. The mechanisms underlying this persistence, however, remain unknown. Reduced genome density and higher capsid surface mass of larger viruses have been proposed to support virion stability, namely for dsDNA eukaryotic viruses[3]. Bacteriophages are less likely to benefit from this as they require an optimal internal pressure in order to inject their DNA into the host cell, which may be why genome density was not observed to reduce much with increased capsid size[3]. Alternatively, jumbo phages may carry relevant enzymes within their capsids, as has been found in large eukaryotic viruses of the Nucelocytoviricota that carry a photolyase which may enhance virion persistence amidst environmental stress[35]. While some jumbo phages have been found to carry RNA polymerases in their capsids[36], enzymes related to capsid persistence have yet to be found. Finally, exterior features of jumbo phage capsids, afforded by extra structural genes, may enhance their persistence. Jumbo phages are known to contain unique features on

their capsids[10], but these could serve functions beyond capsid stability, such as enhancing attachment.

Although no studies have quantitatively compared virion stability and virion size, several studies have examined other life history traits that are related to virion size, enabling inferences on largeness and virion stability. A study on marine *Vibrio* phages found that those with longer latency periods better endured thermal stress and chloroform treatment than those with shorter replication times[37]. Interestingly, however, a study on *Rhizobium* phages found that the phages with the longer latency periods withstood UV radiation better than the fastest replicator, but the fastest replicator had higher thermal tolerance[38]. These mixed findings on virion persistence in relation to longer latency periods indicate the pervasive role that ecology may play in governing fitness tradeoffs of largeness. Future studies on capsid size and stability under a variety of environmental conditions will help inform this and may assist the development of phage therapy treatments, as stable virions are needed to ensure their delivery to the target host.

As mentioned in the previous subsection, jumbo phages have been noted for their elaborate capsids with decorations and other proteins that putatively enhance their attachment to hosts[10]. For instance, two jumbo phages recently isolated on *Tenacibaculum maritimum* from seawater were found to have tail fibers which were proposed to assist these phages in finding and attaching to their host[39]. Another putative benefit of an elaborate capsid is the exclusion of co-infecting phages. With more capsid proteins to attach to receptors on the host, the fewer host receptors are available to other phages. A final possibility is that these fancy capsid proteins may reduce non-specific attachment to non-hosts, but both these benefits are speculations and detailed analyses on the ecology of phage capsid proteins will help inform their explicit function.

Replication machinery and auxiliary metabolic genes: increase replication efficiency and reduce reliance on host

While many viruses use their host's machinery to replicate, jumbo phages encode a diverse repertoire of enzymes, tRNAs, and other components for their replication[8,16]. These have been suggested to contribute to the efficiency of jumbo phage infections, despite requiring much more resources per infection than smaller phages[3]. Some jumbo phages also encode genes in other metabolic processes, such as the carbon cycle and photosynthesis machinery, which are thought to augment the host cell's energetics and lead to more efficient replication[40,41]. Furthermore, some metabolic genes are purported to protect the host from unfavorable conditions, such as phage-encoded D1 proteins, part of the photosystem II complex that oxidizes water, persisting amidst photoinhibition of host-encoded D1 proteins[42]. The presence of these extraneous genes can also help phages subvert host defenses. To prevent the spread of a phage, host cells often shutdown their translational machinery[43]. Jumbo phages have been shown to overcome this by encoding their own tRNAs, RNA repair genes, and other translational genes[16,44]. Not all jumbo phages encode such metabolic genes; in fact, some lack an RNA polymerase[16]. The variable distribution in jumbo phage genomes of replication and metabolic genes specific to a host niche not only points to the diverse origins of jumbo phages, but also suggests different selection conditions have led to these strategies (elaborated in Jumbo phage distribution).

Defenses: counteract host defenses and silence co-infecting phages

Akin to predator-prey dynamics, phages and their hosts have been shown to engage in an evolutionary arms race in which the emergence of resistance or a defense against infection by phage is met with a counter-defense from the phage[45,46]. Such counter-defenses have been

uncovered in phages of all sizes[46,47]. Jumbo phages, however, have particularly been noted for having broad diversity of counter defenses, many reviewed in Iyer et al 2021[16]. One remarkable strategy found in jumbo phages is the formation of a nucleus-like compartment during infections. This compartment protects the viral infection from defenses, such as CRISPR-Cas enzymes, as well as other restriction enzymes, that can inhibit future infections [6,7]. It is likely that much of the unknown functions in phage genomes and bacterial genomes relate to this phage-host arms race, which will be of particular interest as phage therapy continues to develop and requires susceptible hosts[47]. Jumbo phages are not only subject to host defenses, however. They are also prone to co-infecting phages, and some jumbo phages have been found to encode genes that silence or prevent infection of other phages. For instance, Iyer et al. 2021 proposed that the lipopolysaccharide biosynthesis genes encoded by some jumbo cyanophages may modify the exterior of their host cells during infection to prevent attachment and infection by other phages. Additionally, a metagenomic survey found that jumbo phages have co-opted most known types of CRISPR-Cas systems and appear to target structural and regulatory of other phages[8]. The longer latency periods of jumbo phages may make their hosts more prone to secondary infections, particularly as co-infecting phages have been shown to capitalize on different stages of replication of each other[48]. At present, however, no co-infection experiments have been performed with jumbo phages. While there are also no quantitative analyses on the number of defense genes and genome size of diverse phages, some jumbo phages have been found to coexist with their hosts longer than smaller phages[27], which may be attributed to the large arsenal of defenses carried by jumbo phages. A culture-independent approach to examine defense gene enrichment and genome size may be the analysis of single cell genomic data from microbes in

natural communities to uncover co-infection frequencies, virus genome sizes, and encoded defenses.

Broaden host range: persist in variable host availability

A phage's host range refers to the variety of hosts that a phage can infect. The host range of larger phages have been found to be wider than smaller phages from studies on a variety of bacterial groups, such as *Xanthomonas citri*[49], *Cellulophaga baltica*, *Pseudoalteromonas*[3], and *Vibrio* species[50,51]. Observed broadened host range of jumbo phages may be due a variety of factors that relate to the mentioned benefits of largeness. To start, with more capsid or structural proteins, jumbo phages can potentially bind to more diverse hosts. Such has been hypothesized for the broad host range *Pseudomonas* phage PA5oct which was found to recognize host LPS and type IV fimbriae, and potentially others[52]. The increased replication machinery of jumbo phages may reduce the reliance of jumbo phages on specific host machinery, which could further expand a jumbo phage's host range, as suggested by the enrichment of nucleotide metabolism genes and broader host ranges of the *Salmonella* phage pSal-SNUABM-04[53] and enrichment in translational genes and tRNAs linked to a broader host range of a jumbo phage of *Xanthomonas citri*[49]. Finally, the extensive anti-host defenses found in diverse jumbo phage genomes[16] may further broaden jumbo phage host ranges by equipping them to evade the defenses from a diversity of hosts. Regardless of underlying reason, a broad host range is a common feature of jumbo phages and may enable their persistence among microbial communities with highly variable composition by being able to subsist on a diversity of hosts[54,55].

Jumbo phage hosts

From soils to wastewater to the human intestine, jumbo phages have been isolated from a variety of ecosystems [10,56], albeit in low frequencies and on a limited set of hosts compared to their smaller counterparts[57]. In addition to the methodological biases against jumbo phage isolation previously mentioned, the isolation of jumbo phages, and all phages for that matter, is biased by which hosts are available in culture. These hosts are primarily pathogens of humans or agricultural organisms, which are by no means the dominant microbes in nature[58]. Despite these biases, one would expect hosts that have more phages isolated on them to have proportionally more jumbo phages isolated on them. However, analysis by Cook et al[57] of available phages on GenBank found that the most prevalent hosts of cultured phages do not align with the most prevalent hosts of jumbo phages, suggesting some constraints or preferences of jumbo phages on certain hosts. Only thirteen of the top thirty genera of phage hosts were also in the top thirty genera of jumbo phage hosts. These included *Erwinia, Aeromonas, Synechococcus, Ralstonia, Yersinia, Salmonella, Vibrio, Pseudomonas, Bacillus, Klebsiella, Escherichia, Staphylococcus, and Acinetobacter*. Notably, the most common host genera of all cultured phages, *Mycobacterium*, lacks any isolated jumbo phages, which may be attributed to their tiny size of 300-500 nm[59], being even smaller than some jumbo phages. Further suggesting constraints of jumbo phages to infect certain host genera, there was marked variation in the proportion of jumbo phages infecting different host genera. For instance, *Escherichia* was the second most common host of phages (1,075 total phages) in the dataset examined by Cook et al[57], but the 28th most common host for jumbo phages (with roughly ~1% of phages being jumbo). In contrast, *Erwinia* is the 26th most common phage host genera with 78 phages and the fifth most common jumbo phage host genera, with ~40% of its phages as jumbo. Most strikingly, the only phage isolated on *Photobacterium* in GenBank is a jumbo phage and three of the four phages of

Tenacibaculum are jumbo phages. These could be a result of chance but may also relate to the biology of these hosts as marine organisms with relatively large cell sizes[60,61]. The seeming enrichment of jumbo phages on certain hosts is likely constrained by host features, such as size and motility. Because jumbo phage virions comprise more volume and require more energy to form than smaller phages[14], they cannot infect bacteria with cell sizes smaller than themselves. Edwards et al. 2021[3] suggested that larger viruses would preferentially infect motile hosts because these viruses have lower diffusivity and abundance than smaller viruses, making it hard to contact hosts. Quantitative analysis on sizes of cultured jumbo phages versus common smaller phages could test this constraint. Aside from size and motility, the environments where jumbo phages are favored may limit which hosts jumbo phages infect.

Jumbo phage distribution

From wastewater to the human intestine, jumbo phages have been isolated from a wide range of environments[10], as well as detected in metagenomes from diverse ecosystems[8]. Nevertheless, because of the diffusion limitations of their large size, jumbo phages have been proposed to be enriched in aquatic environments[10]. Upon examination of the environments of the 51 jumbo phages compiled by Yuan and Gao 2017, jumbo phages are most commonly isolated from aquatic environments (14 phages) followed by wastewater (8) and soil (8), and then sediment (3), plant matter (3), and animals (2). The remaining thirteen have unknown environments as they are likely lab pests. Culture-independent metagenomic surveys has also shown an enrichment of jumbo phages in marine samples, as well as in animal-associated samples, but this likely reflects the biased availability of metagenomes from the ocean and human gut. Furthermore, metagenomic surveys such as this often assemble to the contig level, which often correspond to fragmented, incomplete genomes and may miss jumbo phages[62]. Compounding this, most

metagenomic surveys of viruses use fractions below 0.22 μm, which can further exclude jumbo phages[12,63].

Prior to the advent of inexpensive, high throughput sequencing, other culture-independent methods of characterizing viral diversity were often used and included pulsed-field gel electrophoresis (PGGE) separated viral genomes based on length[64–66]. Though many of these studies filtered samples through 0.22 μm, jumbo phages were often detected. Of particular note, a study on diverse marine regions (Pacific, Adriatic, Antarctic, Arctic) found that virus genomes over 200 kilobases were more common in temperate waters than the polar regions examined, despite the polar samples being collected during the summers of their respective hemispheres[67]. This may point to a thermal limitation in jumbo phage distribution. When examining viral genome size diversity in the same region of an annual time series, Wommack et al (1991) found viruses of 200 kilobases were absent or reduced in July and August, which have moderate stratification of the water column, opposed to months with higher or lower stratification. High mixing, or low stratification, may bring larger phages in contact with hosts better. At the same time, high stratification may result in stability of host communities that evolve resistance to phages more quickly and favor phages with more defense strategies such as jumbo phages[68]. Outside of the ocean, a study on phages in sheep intestines detected jumbo phages in all samples, regardless of feeding schedule or pen enclosure, and included some phages over 400 kilobases[64,69]. This potentially points to the energy requirements of jumbo phages over small phages, considering the gut is likely a more eutrophic environment than the ocean. However, these hypotheses would need direct testing. Future work could include the revival of PGGE approaches to better uncover the biogeography of jumbo phages.

Predictions on jumbo phage biogeography

Identifying where jumbo phages are common is not only useful to better understand viral ecology and evolution, but also helpful for identifying phage therapy agents, as jumbo phages are particularly useful for their broad host range and persistence[49,52]. While we lack a synoptic survey of jumbo phage distribution, the principles of life history ecology can be useful to make predictions on where they may be found most, which brings us to the culmination of this review as this relates to the fitness benefits of jumbo phages.

The range of phage genome sizes can be considered analogous to the spectrum of life history strategies between *r* and *K*-selected organisms [68,70], such as flies versus elephants in the animal kingdom. Suttle 2007 plotted a rank abundance curve of viruses from metagenomic data and proposed that the most abundant phages are *r*-strategists, with rapid replication, high burst sizes, and high virulence, and may include podoviruses and microviruses. Meanwhile, less abundant phages are *K*-strategists, with lower burst sizes, slower replication, lower virulence, and higher persistence, such as larger or temperate viruses. These suggested features, included those listed in jumbo phage advantages and disadvantages, are outlined in Table 1 with potential environmental conditions that would favor the listed feature. Environmental conditions were categorized into three groups: stability (and complexity), productivity, and microbial diversity. Stability refers to the level of a disturbance an environment experiences. For instance, rainforest soil or urban waterways may have low stability, as sporadic storm input can drastically change the conditions. Stability is also proxied for complexity here because a jumbo phage in a stable environment or a complex environment will experience the similar challenges (e.g. finding hosts) in these conditions. Productivity refers to the growth of bacteria in an environment which generally relates to its trophic state (eutrophic versus oligotrophic). An example of a highly productive environment is the human intestine, and desert soil being a lower productive

environment. Finally, microbial diversity is considered as a final metric because this results from a combination of conditions mentioned, such as instability enhancing microbial diversity in sediments[71], but also can directly impact competition among phages and the extent that a broad host range or high persistence may help a jumbo phage find a suitable host.

Feature	Stability	Productivity	Host diversity
Capsid persistence	Low	Low	Low
Adsorption efficiency	Low	NA	High
Replication efficiency	Low	Low	High
Host metabolic expansion	Low	NA	NA
Host defense evasion	High	Low	High
Co-infection exclusion	NA	NA	High
Broadened host range	Low	Low	High
Smaller burst sizes	NA	High	Low
Longer latency	NA	High	Low

Table 1. Features of jumbo phages associated with fitness tradeoffs and the ecological conditions (Stability, Productivity, Host diversity) that would favor these features.

Stability

Almost all benefits of jumbo phages would favor their presence in unstable or complex environments (Table 1). To start, a sturdy capsid may enable jumbo phages to persist in stochastic exposure to harsh conditions, such as drought periods in soil. The higher adsorption efficiency of jumbo phages could also help them attach effectively to their hosts amidst high shear, which may accompany tidal shifts, storm events, etc. Enhanced replication efficiency is particularly useful when resource availability to hosts becomes limited. Similarly, by encoding additional metabolic genes, jumbo phages can ensure their replication if conditions become hostile to their hosts, such as the phage-encoded photosynthesis genes reducing their host's photoinhibition under high UV light[42]. Finally, a broad host range can especially help when the

available microbial community varies due to fluctuating local conditions, according to optimal foraging theory[54]. Conversely, additional anti-host defenses carried by jumbo phages may be more useful in stable environments, under which the predominant microbes have evolved resistance to virulent phages and thus may remain susceptible to jumbo phages[68]. Aside from defenses, the numerous advantages jumbo phages have in unstable environments over smaller phages, suggest their enrichment in these conditions. Similarly, these mentioned benefits may help jumbo phages in highly complex environments, as hosts and phages may move between diverse conditions despite the overall environment being stable, such as the human gut or layers of soil.

Productivity (or trophic status)

The productivity of an environment can relate to the density of hosts, the size of available microbes, and the competition for hosts. Edwards et al 2021 proposed larger viruses would be enriched in less productive, or oligotrophic, environments because microbial (mainly protist) densities are lower here, favoring the benefits of larger viruses such as a broad host range, high capsid stability, and elevated replication efficiency. Furthermore, protists tend to be more motile in oligotrophic conditions[72] to scavenge for resources, which could increase contact rates between large viruses that have low diffusivity. Their proposition of low density environments favoring larger viruses is more likely applicable for eukaryotic giant viruses than for jumbo phages because (i) the distribution of eukaryotic cells is more pronounced than that of prokaryotes along trophic gradients[72,73] and (ii) the ten-fold larger size of eukaryotic cells over prokaryotic cells may be less constraining for large viruses to infect.

The minimum volume size of hosts for jumbo phages may thus restrict jumbo phages to environments where larger cells can thrive. As prokaryotic cell size is largely limited by the diffusion of nutrients, eutrophic environments likely favor larger cells, such as animal intestines or coastal ecosystems[74]. The infrequency of jumbo phages infecting *Prochlorococcus* over *Synechococcus* (3 versus 53 in GenBank as of September 2021[57]) may support this supposition. *Prochlorococcus* and Synehococcus are genera of photosynthetic bacteria found throughout the global ocean. *Prochlorococcus* bacteria, however, dominate the oligotrophic, open ocean and have some of the smallest cell and genome sizes. Meanwhile, *Synechococcus* species have a greater range of cell sizes and are more prominent in coastal waters[75]. Furthermore, an survey of both metagenomic and cultured jumbo phage distribution in the open ocean by Weinheimer and Aylward 2022 (Chapter 3) found that most jumbo phages were found in surface waters, which are more productive due to photosynthetic bacteria than deeper, darker waters[73]. Intriguingly, this study found that two groups of jumbo phages distinguished by replication machinery were preferentially abundant in deeper waters, where productivity is lower. One group encoded many genes of PhiKZ-like phages including those to form a nucleus-like compartment during infections. A defining feature of the other group included a T7-like RNA polymerase. These disparate distributions could be a result of chance, as only a few genomes from these groups were detected in these studies, or the sinking of these phages from the surface. In any case, the presence of jumbo phages in lower productive environments invites intriguing questions on host-phage ecology and evolution.

Microbial diversity

The diversity of available hosts will likely impact whether specialist or generalist phages, with broad host ranges like jumbo phages, are favored[54]. A higher community diversity will likely favor jumbo phages, as they are more likely to find a suitable host. Environments that contain high microbial diversity have been found to include different combinations of stability and productivity. For instance, high levels of disturbance, or instability, have been found to increase microbial diversity and production in marine sediments[71]. In contrast, a study on microbial communities along a disturbance gradient in a bay off the coast of Brazil found that although the bacterial production rates were highest in the most nutrient-rich, disturbed sites, the bacterial diversity was lowest there. Contrasting in productivity, microbial diversity has been found to be elevated in less productive environments (i.e. mesopelagic ocean versus the surface ocean[76]), potentially due to less steady or stable nutrient sources. In any case, one might expect viral diversity to increase with microbial diversity. Findings of Weinheimer et al in Chapter 4 on phage and prokaryotic communities across the Isthmus of Panama, however, found that microbial diversity only positively correlates with viral diversity when spatial scales and environmental gradients are most pronounced. In fact, there may even be a negative correlation between prokaryotic diversity and phage diversity. For instance, they found that phage communities were more diverse in the Atlantic coast compared to the Pacific coast, but prokaryotic communities were more diverse in the Pacific coast. They propose that broader host ranges of phages may decouple the relationship of phage-host diversity, which could mean that jumbo phages are more prevalent in the Atlantic over the Pacific. All in all, it is likely that jumbo phages are more prevalent where microbial diversity is highest, but this relationship may be dampened in cases when the microbial diversity is driven by scarce resources that would

result in smaller available host cells or when production rates are higher and disturbance is high that also make opportunistic phages more competitive.

Synthesis

Considering that jumbo phages likely thrive in environments where stability is low, productivity is high, or combinations of the two that would result in high microbial diversity, jumbo phages would presumably be most prevalent in nutrient-rich, complex environments like animal digestive tracts, sewage effluent, and urbanized coastal ecosystems. To test these hypotheses, future work could include applying PGGE studies on genome size diversity of these environments. Because certain conditions, such seawater depth in the ocean, might also impact the types of jumbo phages present (Weinheimer and Aylward 2022), the future work could also include detecting certain hallmarks of these jumbo phage groups in diverse metagenomes and comparing where they are found.

Conclusion

Jumbo phages are increasingly recognized for their unique biology in helping us understand the gap between prokaryotes and eukaryotes[77], replication strategies of ancient replicons predating cells[16], ecological and evolutionary trajectories of viral life history strategies, along with their use in phage therapeutics assisted by their broad host ranges and persistence[49,52]. Despite this growing interest, much about the biology and distribution of jumbo phages remains unknown, as few are in culture and current short-read sequencing methods limit their detection in metagenomes[10,12]. Nevertheless, evaluating potential fitness tradeoffs of large viruses can be helpful to inform where we might expect to find these genomic peculiarities, as such has been conducted for aquatic giant viruses of eukaryotes[3]. Here, the diversity, emergence, fitness

tradeoffs, currently known hosts, and current known distribution of jumbo phages to arrive at predictions on jumbo phage biogeography were reviewed. The great diversity of jumbo phages suggests that different environments will play to different benefits of jumbo phages[12]. Nevertheless, the broader host range, greater capsid stability, enhanced replication efficiency, broader metabolic capacity, and increased defense repertoire of jumbo phages suggest that they will be most abundant in environments of low stability, high productivity, or combinations of the two that result in high microbial diversity. At the same time, available host cell sizes likely limit jumbo phage distributions to nutrient-rich conditions, where larger cells are favored[74]. Thus, bearing all of these environmental features in mind, jumbo phages are likely enriched in the animal digestive tracts, sewage outfalls, and marine sediments. Because these predictions are based on current distributions of jumbo phages and ecological theory, future work to test these hypotheses could include surveys of genome size ranges with PGGE in diverse environments, mesocosm experiments on genome size variation in which nutrient availability and disturbance levels are manipulated, and sequencing surveys that target the inclusion of larger phages to uncover both the general distribution of jumbo phages and the specific distribution of certain groups of jumbo phages. Furthermore, the distribution of jumbo phages can also further our understanding of why jumbo phages emerged. Because even though "small is beautiful"[1], big is also beautiful.

References

1. Brüssow, H. & Hendrix, R. W. Phage genomics: small is beautiful. *Cell* **108**, 13–16 (2002).
2. Clokie, M. R., Millard, A. D., Letarov, A. V. & Heaphy, S. Phages in nature. *Bacteriophage* **1**, 31–45 (2011).
3. Edwards, K. F., Steward, G. F. & Schvarcz, C. R. Making sense of virus size and the

tradeoffs shaping viral fitness. *Ecol. Lett.* **24**, 363–373 (2021).

4. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat. Commun.* **11**, 1710 (2020).
5. Hendrix, R. W. Jumbo bacteriophages. *Curr. Top. Microbiol. Immunol.* **328**, 229–240 (2009).
6. Mendoza, S. D. *et al.* A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature* **577**, 244–248 (2020).
7. Malone, L. M. *et al.* A jumbo phage that forms a nucleus-like structure evades CRISPR–Cas DNA targeting but is vulnerable to type III RNA-based immunity. *Nature Microbiology* **5**, 48–55 (2019).
8. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425–431 (2020).
9. Rai, P. *et al.* Characterisation of broad-spectrum phiKZ like jumbo phage and its utilisation in controlling multidrug-resistant *Pseudomonas aeruginosa* isolates. *Microb. Pathog.* **172**, 105767 (2022).
10. Yuan, Y. & Gao, M. Jumbo bacteriophages: an overview. *Front. Microbiol.* **8**, 403 (2017).
11. Serwer, P., Hayes, S. J., Thomas, J. A. & Hardies, S. C. Propagating the missing bacteriophages: a large bacteriophage in a new class. *Virol. J.* **4**, 21 (2007).
12. Weinheimer, A. R. & Aylward, F. O. Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages. *ISME J.* **16**, 1657–1667 (2022).
13. Nazir, A., Ali, A., Qing, H. & Tong, Y. Emerging aspects of jumbo bacteriophages. *Infect. Drug Resist.* **14**, 5041–5055 (2021).

14. Cui, J., Schlub, T. E. & Holmes, E. C. An allometric relationship between the genome length and virion volume of viruses. *J. Virol.* **88**, 6403–6410 (2014).
15. Ageno, M., Donelli, G. & Guglielmi, F. Structure and physico-chemical properties of bacteriophage G. II, The shape and symmetry of the capsid. *Micron* **4**, 376–403 (1973).
16. Iyer, L. M., Anantharaman, V., Krishnan, A., Maxwell Burroughs, A. & Aravind, L. Jumbo phages: a comparative genomic overview of core functions and adaptions for biological conflicts. *Viruses* **13,** 63 (2021).
17. Yahara, K. *et al.* Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria. *Nat. Commun.* **12**, 27 (2021).
18. Low, S. J., Džunková, M., Chaumeil, P.-A., Parks, D. H. & Hugenholtz, P. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nat Microbiol* **4**, 1306–1315 (2019).
19. Hua, J. *et al.* Capsids and Genomes of Jumbo-Sized Bacteriophages Reveal the Evolutionary Reach of the HK97 Fold. *MBio* **8**, e01579-17 (2017).
20. Powers, R. *et al.* Solution structure of *Archaeglobus fulgidis* peptidyl-tRNA hydrolase (Pth2) provides evidence for an extensive conserved family of Pth2 enzymes in archea, bacteria, and eukaryotes. *Protein Sci.* **14**, 2849–2861 (2005).
21. Weinheimer, A. R. & Aylward, F. O. A distinct lineage of Caudovirales that encodes a deeply branching multi-subunit RNA polymerase. *Nat. Commun.* **11**, 4506 (2020).
22. Nagy, K. K., Skurnik, M. & Vértessy, B. G. Viruses with U-DNA: New avenues for biotechnology. *Viruses* **13**, 875 (2021).
23. Levin, B. R., Stewart, F. M. & Chao, L. Resource-limited growth, competition, and predation: A model and experimental studies with bacteria and bacteriophage. *Am. Nat.* **111**,

3–24 (1977).

24. Edwards, K. F. & Steward, G. F. Host traits drive viral life histories across phytoplankton viruses. *Am. Nat.* **191**, 566–581 (2018).
25. Maat, D. S. & Brussaard, C. P. D. Both phosphorus- and nitrogen limitation constrain viral proliferation in marine phytoplankton. *Aquat. Microb. Ecol.* **77**, 87–97 (2016).
26. Wilson, W. H., Carr, N. G. & Mann, N. H. The effect of phosphate status on the kinetics of cyanophage infection in the oceanic cyanobacterium *Synechococcus sp.* WH78031. *Journal of Phycology* **32**, 506–516 (1996).
27. Fujiwara, A. *et al.* Biocontrol of *Ralstonia solanacearum* by treatment with lytic bacteriophages. *Appl. Environ. Microbiol.* **77**, 4155–4162 (2011).
28. Thomas, W. E., Trintchina, E., Forero, M., Vogel, V. & Sokurenko, E. V. Bacterial adhesion to target cells enhanced by shear force. *Cell* **109**, 913–923 (2002).
29. Abedon, S. T. Impact of phage properties on bacterial survival. *Contemporary Trends in Bacteriophage Research* 217–235 (2009).
30. Suttle, C. A. & Chen, F. Mechanisms and rates of decay of marine viruses in seawater. *Appl. Environ. Microbiol.* **58**, 3721–3729 (1992).
31. Noble, R. T. & Fuhrman, J. A. Virus decay and its causes in coastal waters. *Appl. Environ. Microbiol.* **63**, 77–83 (1997).
32. Brown, J. M. *et al.* Single cell genomics reveals viruses consumed by marine protists. *Front. Microbiol.* **11**, (2020).
33. Heldal, M. & Bratbak, G. Production and decay of viruses in aquatic environments. *Mar. Ecol. Prog. Ser.* **72**, 205–212 (1991).
34. De Paepe, M. & Taddei, F. Viruses' life history: towards a mechanistic basis of a trade-off

between survival and reproduction among phages. *PLoS Biol.* **4**, e193 (2006).

35. Fischer, M. G., Kelly, I., Foster, L. J. & Suttle, C. A. The virion of *Cafeteria roenbergensis* virus (CroV) contains a complex suite of proteins for transcription and DNA repair. *Virology* **466**, 82–94 (2014).
36. Ceyssens, P.-J. *et al.* Development of giant bacteriophage ϕKZ is independent of the host transcription apparatus. *J. Virol.* **88**, 10501–10510 (2014).
37. Okano, S., Yoshikawa, T. *et al*. Characterization of *Vibrio harveyi* bacteriophages isolated from aquaculture tanks. *Mem. Fac. Fish. Kagoshima Univ.* **56**, 55–62 (2007).
38. Jaiswal, S. K. & Dhar, B. Morphology and general characteristics of phages specific to Lens culinaris rhizobia. *Biol. Fertil. Soils* **46**, 681–687 (2010).
39. Kawato, Y. *et al.* A novel jumbo *Tenacibaculum maritimum* lytic phage with head-fiber-like appendages. *Arch. Virol.* **165**, 303–311 (2020).
40. Naknaen, A., Suttinun, O., Surachat, K., Khan, E. & Pomwised, R. A novel jumbo phage PhiMa05 inhibits harmful *Microcystis sp. Front. Microbiol.* **12**, 660351 (2021).
41. Yamada, T. *et al.* A jumbo phage infecting the phytopathogen *Ralstonia solanacearum* defines a new lineage of the Myoviridae family. *Virology* **398**, 135–147 (2010).
42. Bailey, S., Clokie, M. R. J., Millard, A. & Mann, N. H. Cyanophage infection and photoinhibition in marine cyanobacteria. *Res. Microbiol.* **155**, 720–725 (2004).
43. Bitton, L., Klaiman, D. & Kaufmann, G. Phage T4-induced DNA breaks activate a tRNA repair-defying anticodon nuclease. *Mol. Microbiol.* **97**, 898–910 (2015).
44. Yang, J. Y. *et al.* Degradation of host translational machinery drives tRNA acquisition in viruses. *Cell Syst* **12**, 771–779.e5 (2021).
45. Hall, A. R., Scanlan, P. D., Morgan, A. D. & Buckling, A. Host-parasite coevolutionary

arms races give way to fluctuating selection. *Ecol. Lett.* **14**, 635–642 (2011).

46. Safari, F. *et al.* The interaction of phages and bacteria: the co-evolutionary arms race. *Crit. Rev. Biotechnol.* **40**, 119–137 (2020).
47. Hussain, F. A. *et al.* Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science* **374**, 488–492 (2021).
48. Brewster, L., Langley, M. & Twa, D. Co-infection of C3000 *Escherichia coli* with bacteriophages MS2 and, T7 or ΦX-174 results in differential cell lysis patterns. *JEMI* **16**, 139-143 (2012).
49. Yoshikawa, G. *et al. Xanthomonas citri* jumbo phage XacN1 exhibits a wide host range and high complement of tRNA genes. *Sci. Rep.* **8**, 4486 (2018).
50. Kim, S. G. *et al.* Isolation and characterisation of pVa-21, a giant bacteriophage with anti-biofilm potential against *Vibrio alginolyticus*. *Sci. Rep.* **9**, 6284 (2019).
51. Matsuzaki, S., Tanaka, S., Koga, T. & Kawata, T. A broad-host-range vibriophage, KVP40, isolated from sea water. *Microbiol. Immunol.* **36**, 93–97 (1992).
52. Olszak, T. *et al. Pseudomonas aeruginosa* PA5oct jumbo phage impacts planktonic and biofilm population and reduces its host virulence. *Viruses* **11**, (2019).
53. Kwon, J. *et al.* Isolation and characterization of *Salmonella* jumbo-phage pSal-SNUABM-04. *Viruses* **13**, (2020).
54. Guyader, S. & Burch, C. L. Optimal foraging predicts the ecology but not the evolution of host specialization in bacteriophages. *PLoS One* **3**, e1946 (2008).
55. Dekel-Bird, N. P., Sabehi, G., Mosevitzky, B. & Lindell, D. Host-dependent differences in abundance, composition and host range of cyanophages from the Red Sea. *Environmental Microbiology* **17**, 1286–1299 (2015).

56. Devoto, A. E. *et al.* Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nature microbiology* **4**, 693–700 (2019).

57. Cook, R. *et al.* INfrastructure for a PHAge REference Database: Identification of large-scale biases in the current collection of cultured phage genomes. *Phage (New Rochelle)* **2**, 214–223 (2021).

58. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).

59. Parte, A. *et al. Bergey's Manual of Systematic Bacteriology: Volume 5: The Actinobacteria*. (Springer New York, 2012).

60. Piñeiro-Vidal, M., Riaza, A. & Santos, Y. *Tenacibaculum discolor sp*. nov. and *Tenacibaculum gallaicum sp*. nov., isolated from sole (*Solea senegalensis*) and turbot (*Psetta maxima*) culture systems. *Int. J. Syst. Evol. Microbiol.* **58**, 21–25 (2008).

61. de W Blackburn, C. *Food Spoilage Microorganisms*. (Woodhead Publishing, 2006).

62. García-López, R., Vázquez-Castellanos, J. F. & Moya, A. Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. *Front Bioeng Biotechnol* **3**, 141 (2015).

63. Palermo, C. N., Shea, D. W. & Short, S. M. Analysis of different size fractions provides a more complete perspective of viral diversity in a freshwater embayment. *Appl. Environ. Microbiol.* **87**, (2021).

64. Klieve, A. V. & Swain, R. A. Estimation of ruminal bacteriophage numbers by pulsed-field gel electrophoresis and laser densitometry. *Appl. Environ. Microbiol.* **59**, 2299–2303 (1993).

65. Wommack, K. E., Ravel, J., Hill, R. T., Chun, J. & Colwell, R. R. Population dynamics of Chesapeake bay virioplankton: total-community analysis by pulsed-field gel electrophoresis.

Appl. Environ. Microbiol. **65**, 231–240 (1999).

66. Maruyama, A., Oda, M. & Higashihara, T. Abundance of virus-sized non-DNase-digestible DNA (coated DNA) in eutrophic seawater. *Appl. Environ. Microbiol.* **59**, 712–717 (1993).
67. Steward, G. F., Montiel, J. L. & Azam, F. Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol. Oceanogr.* **45**, 1697-1706 (2000).
68. Suttle, C. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801-812 (2007).
69. Swain, R. A., Nolan, J. V. & Klieve, A. V. Natural variability and diurnal fluctuations within the bacteriophage population of the rumen. *Appl. Environ. Microbiol.* **62**, 994–997 (1996).
70. Keen, E. C. Tradeoffs in bacteriophage life histories. *Bacteriophage* **4**, e28365 (2014).
71. Galand, P. E. *et al.* Disturbance increases microbial community diversity and production in marine sediments. *Frontiers in Microbiology* **7**, 1950 (2016).
72. Edwards, K. F. Mixotrophy in nanoflagellates across environmental gradients in the ocean. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 6211–6220 (2019).
73. Li, W. K. W., Head, E. J. H. & Glen Harrison, W. Macroecological limits of heterotrophic bacterial abundance in the ocean. *Deep Sea Res. Part I* **51**, 1529–1540 (2004).
74. Schulz, H. N. & Jorgensen, B. B. Big bacteria. *Annu. Rev. Microbiol.* **55**, 105–137 (2001).
75. Partensky, F., Blanchot, J. & Vaulot, D. Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters: a review. *Bulletin-Institut Oceanographique Monaco-Numero Special* 457–476 (1999).
76. Signori, C. N., Thomas, F., Enrich-Prast, A., Pollery, R. C. G. & Sievert, S. M. Microbial

diversity and community structure across environmental gradients in Bransfield Strait, Western Antarctic Peninsula. *Front. Microbiol.* **5**, 647 (2014).

77. Brüssow, H. Huge bacteriophages: bridging the gap? *Environ. Microbiol.* **22**, 1965–1970 (2020).

Chapter 2: A distinct lineage of *Caudovirales* that encodes a deeply branching multi-subunit RNA polymerase

Weinheimer, A. R., & Aylward, F. O. (2020). A distinct lineage of Caudovirales that encodes a deeply branching multi-subunit RNA polymerase. Nature communications, 11(1), 1-9.

Reproduced from the journal *Nature Communications*.

Abstract

Bacteriophage play critical roles in the biosphere, but their vast genomic diversity has obscured their evolutionary origins, and phylogenetic analyses have traditionally been hindered by their lacking universal phylogenetic marker genes. In this study we mine metagenomic data and identify a clade of *Caudovirales* that encodes the β and β' subunits of multi-subunit RNA polymerase (RNAP), a high-resolution phylogenetic marker which enables detailed evolutionary analyses. Our RNAP phylogeny revealed that the *Caudovirales* RNAP forms a clade distinct from cellular homologs, suggesting an ancient acquisition of this enzyme. Within these multimeric RNAP-encoding *Caudovirales* (mReC), we find that the similarity of major capsid proteins and terminase large subunits further suggests they form a distinct clade with common evolutionary origin. Our study characterizes a clade of RNAP-encoding *Caudovirales* and suggests the ancient origin of this enzyme in this group, underscoring the important role of viruses in the early evolution of life on Earth.

Keywords: Bacteriophage, *Caudovirales*, Tree of Life, viral origins, multi-subunit RNA Polymerase

Abbreviations: RNAP: multi-subunit DNA-directed RNA polymerase; TOL: Tree of Life; NCLDV: Nucleo-Cytoplasmic Large DNA viruses; bp: base pairs, kb: kilobases

Introduction

The creation of a Tree of Life (TOL) that encompasses all life forms on Earth has been a central goal of Biology ever since the concept of evolution was introduced by Darwin and Wallace in the 19th Century[1]. In recent decades, efforts toward the creation of a TOL has progressed considerably due to advances in molecular phylogenetic methods, the increased availability of whole-genome sequencing datasets, and the identification of highly-conserved marker genes that are present in all cellular life and can be readily compared in molecular phylogenetic trees[2–4]. Although the TOL has been a useful concept for studying cellular lineages, it has proven more problematic for viruses, however, which lack the necessary phylogenetic marker genes that would allow their inclusion into typical molecular phylogenies of cellular life[5]. Indeed, current frameworks for dealing with viruses generally group them together as capsid-encoding organisms in order to distinguish them from cellular ribosome-encoding organisms, a framework that explicitly classifies them due to their lack of the ribosomal genes that are commonly used for constructing molecular phylogenies[6].

There are high-resolution phylogenetic marker genes other than ribosomal genes that can provide insight into deep evolutionary relationships, however, such as the multi-subunit DNA-directed RNA polymerase (RNAP). RNAP is an ancient enzyme present in all Bacteria, Archaea, and Eukarya, and it has often been used to provide high-resolution phylogenies of divergent microbial lineages[7–10]. Importantly, although viruses lack ribosomal genes, some viruses encode their own copy of RNAP that can be used to evaluate their evolutionary relationships to cellular life; for example, one recent study focusing on Nucleo-Cytoplasmic Large DNA Viruses (NCLDV) analyzed RNAP phylogenies and found evidence that this viral group emerged prior to modern Eukaryotes[11].

Multi-subunit RNAP is composed of two core subunits referred to as the β and β' subunits in Bacteria, RPB1 and RPB2 subunits in Eukarya, and B and A in Archaea, respectively. These two subunits are homologous and likely originate from an ancient duplication event[12]. Several bacteriophages have been found to encode RNAPs that have likely been acquired in distinct ways. For example, the recently-discovered crAss-phage, which are widespread in the human microbiome, encode an RNA polymerase enzyme in which the β and β' subunits are fused into one protein[13], but this enzyme is highly divergent from cellular RNAP subunits and sequence homology is not easily identified. Moreover, some bacteriophage have been shown to encode a single-subunit YonO protein which shares homologous motifs with the β' subunit of RNAP and functions as a DNA-dependent RNA polymerase[14], but these enzymes are also highly divergent compared to cellular homologs. Additionally, a recent large-scale metagenomic analysis identified the presence of multi-subunit RNAP homologs in contigs from environmental bacteriophage, suggesting that the prevalence of this enzyme in viruses may be broader than previously thought[15].

In this study, we surveyed multiple large DNA sequence datasets to identify the occurrence of RNAP in bacteriophage genomes and examine the evolutionary links between these enzymes and cellular homologs. Because the vast majority of viral diversity remains uncultivated, we focused our analysis on viral sequences present in metagenomic datasets and ultimately identified 97 bacteriophage-encoded RNAP that we used for subsequent analysis. Phylogenetic analyses of the RNAP encoded by these bacteriophages suggests that they are distinct from cellular RNAP and are the result of an ancient acquisition. Moreover, analysis of other marker genes suggest these viruses belong to a lineage of *Caudovirales*, which we refer to as multi-subunit RNAP-encoding *Caudovirales* (mReC).

Results and Discussion

Detection of RNAP-encoding Caudovirales

We analyzed 1,545 previously assembled metagenomic datasets[16] and 760,453 viral sequences available in the online viral sequence repository IMG/VR[17] (see Methods). We compared all encoded proteins in these genomes and contigs against Hidden Markov Models constructed from cellular homologs of the β and β' RNAP subunits (COG0085 and COG0086[18]) so that we could identify enzymes that have not diverged so far from their cellular homologs as to prevent robust sequence alignment and phylogenetic analysis (see Methods for details). In total, we identified 266 viral metagenomic contigs that encode both the β and β' subunits of RNAP. In diagnostic phylogenies, 97 contigs encoded RNAP subunits which clustered separately from homologs in cells and eukaryotic viruses (NCLDV) in a distinct deep-branching clade (Supplementary Fig. 1). These contigs also encoded viral signatures such as capsid, terminase, baseplate, wedge, portal, and tail proteins, indicating that they derive from *Caudovirales* (Figure 1, Supplementary Dataset 1). Moreover, 3 of these contigs were >200 kbp in length, suggesting they belong to "jumbo bacteriophage". These contigs were identified in metagenomes that were sequenced from a variety of different aquatic, host-associated, and engineered environments, further suggesting they are widespread in the biosphere (Supplementary Fig. 2). These results are consistent with a recent large-scale metagenomic survey of viruses, which identified homologs of RNAP in several environmental bacteriophage sequences[15]. Given the unusual presence of multi-subunit RNAP in these *Caudovirales* contigs, which we refer to as multi-subunit RNAP-encoding *Caudovirales* (mReC), we focused on them for purposes of this study.

The mReC branch deeply within an RNAP Tree of Life

We constructed an unrooted phylogenetic tree of the mReC RNAP sequences with representative Archaea, Bacteria, Eukarya, and NCLDV (Supplementary Dataset 2) using maximum likelihood analyses in IQ-TREE[19], with amino acid substitution model LG+C60+F+Γ4, a site-heterogeneous approach, which is particularly effective for estimating ancient divergences[20], corrects for long-branch attraction between divergent lineages, and is commonly used in deep phylogenetic studies[21,22]. The resulting tree revealed a distinct clustering of bacteriophage sequences on a separate branch from all other lineages (Figure 2), with 100 ultrafast bootstrap support and an Internode Certainty (IC) of 1 for the monophyly of the mReC clade. The clustering of Archaea, Eukaryota, and NCLDV together and on a distinct branch from Bacteria is also consistent with previous studies[11].

To ensure our unrooted phylogeny was based on a high quality alignment of homologous RNAP regions, we manually-inspected the β and β' subunit alignments and identified eight highly conserved regions (Figure 3, Supplementary Fig. 3). These highly conserved regions were discerned based on both alignment conservation and quality (see Methods). Many of these regions corresponded to known conserved motifs in RNAP; within the β subunit, these structures included the DNA-binding site, double-psi beta barrels, and the connector to the β' subunit[23]. Within the β' subunit, conserved regions included double-psi beta barrel structures and the catalytic core[23]. This catalytic core hosts the active site which coordinates a magnesium ion and contains the highly conserved NADFDG motif. Upon visualization in the structure of the yeast *Saccharomyces cerevisiae* RNAP II crystal structure (PDB ID: 2e2i; chain A of RPB1 and B RPB2 corresponding to the β' and β subunits, respectively)[24], we observed that highly conserved regions tended to be at the interface of the two subunits (Figure 4a). This is consistent with

selective pressure against mutations that interfere with the association and binding of the core subunits of the protein complex[12].

In addition to using the LG+C60+F+Γ4 site-heterogeneous model for phylogenetic construction, we also used several other approaches to correct for possible artefacts introduced by increased substitution rates in viral lineages, which could potentially result in long branch attraction. Firstly, we constructed phylogenies from trimmed alignments using varying levels of stringency (positions with 10%, 30%, 50%, and 70% of gaps removed, Supplementary Fig. 4; see Methods). We then evaluated the resulting alignment quality based on conservation, identity, and quality (Supplementary Dataset 3, see Methods). Secondly, in addition to varying levels of trimming stringency, we removed up to 50% of fast-evolving sites in the RNAP alignment using the TIGER software[25] which groups alignment sites based on their substitution rates. Removal of fast evolving sites consistently maintained both the known monophyly of the clade grouping NCLDV, archaea and eukaryotic RNAP and the distinct clustering of mReC RNAP from all cellular RNAP (Supplementary Fig. 5). Lastly, we also performed phylogenetic analysis using the PhyloBayes software[26] to ensure that the results of maximum-likelihood and Bayesian approaches were consistent; using the alignment with 30% of gaps removed, we once again recovered a topology consistent with our other methods (Supplementary Fig. 6; see Methods for details). Thus, by using a combination of different alignment quality checks and phylogenetic reconstruction methods, we provide evidence that the mReC RNAP sequences belong to a distinct lineage that likely have a common origin.

The mReC Form a Distinct Clade Within the Caudovirales

To further investigate the monophyly of the mReC contigs, we performed phylogenetic analysis on other phage marker genes to ascertain if they supported the monophyly of mReC. We detected 8 major capsid proteins (MCP) and 31 large terminase subunit proteins (TerL) in these mReC (Supplementary Dataset 1). All MCP proteins had best matches to the same VOG and Pfam family (VOG11186 and PF07068, respectively), suggesting they have common evolutionary origins. For the TerL proteins, 27 of the 31 that had matches to the VOG database were classified to the same family (VOG01069); only 17 of these proteins had hits to TerL proteins in the Pfam database, all of which matched to the same family (PF03237). We reconstructed phylogenies of the MCP and TerL proteins which showed that those proteins found within mReC tended to cluster together compared to available references in IMG/VR and Viral RefSeq, further suggesting they have common evolutionary origins (Figure 5). The mReC proteins clade together with some references in IMG/VR that do not encode RNAP, but this is expected given that the contigs in this database are fragmented and may belong to complete genomes that encode RNAP. There were some exceptions to the trend of placement of mReC MCP and TerL proteins in the same region of these trees, such as a divergent TerL that is evident in our tree of VOG01069 references (Figure 5c). Given that the mReC likely comprise a diverse clade of *Caudovirales*, it is likely that horizontal gene transfer has occurred within this group and other *Caudovirales* lineages at some point, which may explain this exception. Together with the observed distinct clustering of mReC RNAP, these results suggest that at least the majority of the mReC derive from a distinct clade of *Caudovirales* with a common evolutionary origin.

Rooting Analysis for the RNAP Tree of Life

Although the unrooted RNAP phylogenetic tree suggests that mReC RNAP originate from an ancient divergence from cellular homologs, the precise nature of these origins remain ambiguous as long as the tree remains unrooted. Previous studies have rooted the tree of life using paralogous protein families that are the product of an ancient duplication that predates the divergence of the primary domains. In these approaches, which have variously used elongation factors, ATPase subunits, and aminoacyl-tRNA synthetases[27–29], one gene family effectively serves as the outgroup for its paralog. Because the β and β' subunits of RNAP are paralogous and originate from an ancient duplication[30], we sought to use this approach to estimate a rooted tree.

First, we aligned the β and β' subunits with each other and identified conserved regions (Supplementary Fig. 7). This alignment had markedly lower quality than alignments generated using only β and β' individually (Supplementary Dataset 3), which is expected given these subunits arose from an ancient duplication event that likely took place before the divergence of Bacteria and Archaea. Nevertheless, distinct conserved regions were identified. Similar to the conserved regions found within the individual subunits, the regions shared between the β and β' subunits appeared to be at the interface of the two subunits (Figure 4b). We then reconstructed a phylogeny using the LG+C60+F+Γ4 amino acid substitution model in IQ-TREE, which we refer to as the β/β' paralog tree. This tree revealed a topology in which Bacteria and mReC RNAP are sister clades in both the β and β' subunits regions of the tree (Figure 6a). We proceeded to remove quickly-evolving sites to evaluate the stability of this topology; we found that the Bacteria-mReC RNAP sister relationship was relatively stable (Figure 6b). The monophyly of the Archaea-Eukarya-NCLDV branch was used as a control to assess when so many positions had been removed such that phylogenetic estimation became unreliable; the bootstrap and

internode certainty of this node remained high in almost all alignments. Interestingly, upon removing up to 45% of the fastest evolving sites, the most well-supported topology shifted in the β' subunit such that Bacteria appear to have emerged prior to all other lineages (Figure 6b), though this pattern was not shared in the β subunit. This analysis provides some evidence that mReC RNAP were acquired at or near the diversification Bacteria, but results should be interpreted cautiously given they are derived from an alignment of highly divergent β and β' subunits; future analysis using additional sequences or incorporating structural information would potentially provide further insight.

One explanation for our observed results is that mReC acquired RNAP from cellular lineages in the distant past, potentially even prior to the diversification of the major bacterial phyla (i.e., from a proto-bacterial lineage). This interpretation must be made cautiously, however; while our concatenated β and β' RNAP tree indicates that mReC RNAP forms a distinct clade separate from cellular lineages, the rooted β/β' paralog tree is based on the alignment of highly divergent sequences and does not provide a definitive root. One may postulate an alternative scenario in which the mReC RNAP have an accelerated evolutionary rate that obfuscates phylogenetic analyses and potentially renders the resulting trees unreliable; indeed, other viruses encode YonO proteins or other homologs to RNAP subunits that have diverged considerably and cannot be robustly aligned to cellular homologs[13,14]. RNAP homologs from mReC still retain highly conserved regions that are readily alignable to cellular homologs, however (Figure 3), suggesting that accurate phylogenetic assessment may still be possible for these proteins. Regardless, further analyses will be needed to definitively trace the evolutionary origin of these divergent *Caudovirales*-encoded RNAP genes.

Conclusion

Here we provide phylogenetic evidence for the ancient acquisition of RNAP in a clade of *Caudovirales*, which is an important step in understanding the ancient evolution of this enzyme as well as the dynamic gene exchange between viruses and microbial life in the distant past. Although we originally suspected that multiple acquisitions of *Caudovirales* RNAP from their hosts would be the most likely explanation for the presence of these genes in bacteriophage, similar to what has recently been shown for phage-encoded ribosomal proteins[31,32], the deep-branching placement of *Caudovirales* RNAP in our phylogenies implicates a single ancient acquisition within a distinct *Caudovirales* clade (Figure 2), which we refer to as multi-subunit RNAP-encoding *Caudovirales* (mReC). Phylogenies of capsid and terminase proteins in the mReC further supports their common evolutionary history. Using the paralogy of the β and β' subunits of RNAP, we assess the possibility that the divergence of these mReC sequences from cellular life occurred near the time of the divergence of Bacteria and the branch leading to Archaea, Eukarya, and NCLDV. Deep-branching nodes in our rooting analysis remain highly uncertain due to the highly divergent nature of the paralogous β and β' alignment, however, and the results of analyses that rely on alignments of β and β' subunits remain speculative. Further work will be necessary to provide more insight into the precise timing at which these *Caudovirales* RNAP were acquired from cellular lineages.

It is likely that other bacteriophage groups have independently acquired RNAP from cellular lineages. For example, the human gut-associated crAssphage also encode a single protein that bears sequence motifs consistent with the fusion of β and β' subunits[33]; we were unable to identify any recognizable sequence homology of crAssphage RNAP to the COG0085 and COG0086 HMMs we used in this study, however, indicating that the crAssphage enzyme is highly divergent and acquired independently from the mReC. Moreover, other phage have been

found to encode a single-subunit YonO protein with similarity to the β' subunit of RNAP[14], though once again the highly divergent nature of these proteins hinders detailed phylogenetic analysis. It is not surprising that many divergent enzymes with either sequence or structural homology to RNAP subunits are present in the viral world considering the antiquity of this enzyme; indeed, structural homology has even been noted between RNAP subunits and eukaryotic RNA-dependent RNA polymerase and archaeal replicative DNA polymerase[34]. Evolutionary analysis of ancient enzyme complexes such as multi-subunit RNAP can therefore yield important insight into ancient events in the evolutionary history of both cellular lineages and viruses.

Methods

Dataset selection and RNAP detection. Because the majority of viruses in nature remain uncultured, we searched for bacteriophage RNAP in metagenomic nucleotide sequences from 1,545 curated metagenomes[16] (contigs > 10 kb) and IMG V/R release July 1, 2018 version 4 (contigs > 10 kb; 418,506 contigs)[17], in addition to all cultured viral genomes with bacterial hosts available in viral RefSeq release 96[35]. We downloaded the nucleotide sequences from these datasets and predicted protein sequences with Prodigal[36] (version 2.6.3). Default parameters were used for the RefSeq genomes, and the -p meta option was used for the metagenomic sequences. Amino acid (aa) sequences were searched using HMMER3[37] (v. 3.2.1) against HMM profiles of RNAP β and RNAP β' from the Clusters of Orthologous Groups (COG) protein family database (2014 update)[18], corresponding to COG0085 and COG0086, respectively (E-value 1e-5). Metagenomic contigs and genomes retained for downstream analysis encoded both high quality COG0085 and COG0086 matches that met the following criteria: minimum score of 80,

minimum length of 800 aa, presence of the conserved 'NADxDGD' motif in COG0086, presence of a predicted stop codon, and the absence of any 'X' characters.

Phylogenetic reconstruction and phage classification. To construct an RNAP phylogeny, we selected a diverse array of references (Supplementary Dataset 2). For eukaryote representation, genes corresponding to the β and β' subunits of RNAP II (RPB2 and RPB1 in the eukaryote nomenclature, respectively) were included, as this enzyme's function most closely matches that of Bacteria and Archaea[30]. For initial diagnostic phylogenetic trees, these amino acid sequences were then input to the ete3[38] (version 3.1.1) workflow in which a concatenated alignment was performed with Clustal Omega[39](v. 1.2.3), and a tree was inferred with FastTree[40] (v. 2.1) using the standard_fasttree workflow and sptree_fasttree_all supermatrix. The tree was then visualized on the webserver iTOL[41] (version 4.0, Supplementary Fig. 1). No RNAP sequences in bacteriophage genomes in viral RefSeq encoded RNAP subunits that met our criteria. Sequences encoding RNAP that did not cluster with RNAP of cells or eukaryotic viruses were considered belonging to putative bacteriophage. These sequences were then confirmed as viral based on presence of viral marker genes and enrichment of viral genes relative to cellular genes using the tool ViralRecall (contig mode, minimum score 0, https://github.com/faylward/viralrecall) (Supplementary Dataset 1). Additionally, 48 of the 97 contigs encoded at least one phage hallmark gene (major capsid protein, terminase, baseplate wedge, tail, and portal proteins), which were detected by searching against HMM profiles of these proteins in EggNog 5.0[42], Pfam release 32[43], and VOG (vogdb.org, downloaded April 14, 2020) databases (E-value 1e-3) (full annotations can be found in Supplementary Dataset 1). Plots of the open reading frames of the largest 10 contigs and their hits to the VOG and Pfam databases (using genoplotR[44] (version

0.8.9) in R (version 3.5.1)[45] using Rstudio (version 1.1.456)[46] and Inkscape (v 0.92), Figure 1) revealed that the RNAP genes were typically surrounded by Caudovirales hallmark proteins. To remove redundancy and lower computational load, 65 mReC from across the diversity of their encoded RNAP were selected to serve as a subset for subsequent phylogenetic analyses and alignment visualizations. Phylogenies of the β and β' concatenated alignment were reconstructed using maximum likelihood in IQ-TREE with the LG+C60+F+Γ4 amino acid substitution model because mixed substitution rate models have been shown to be useful for phylogenetic reconstruction of divergent sequences[20] and have been used in other studies for constructing deep phylogenetic relationships[21,22]. To estimate branch support, each tree was reconstructed with a 1,000-replicate, ultra-fast bootstrap approximation and RaXML[47] to calculate absolute and relative internode certainty values. Internode certainty has recently been proposed as a useful alternative to bootstrap support[48], and these values give a measure of the support for a given topology compared to other well-supported alternatives.

Alignment quality control. To improve the alignment for phylogenetic reconstruction, we trimmed the concatenated alignment of the β and β' subunits for positions containing gaps in over 10%, 30%, 50%, and 70% of the sequences using trimAl[49] (version 1.2rev59). Alignments of each threshold were visualized in JalView[50] and searched for regions of known conserved functions. Quality was assessed based on overall identity and conservation, which is calculated in JalView based on both identity and physio-chemical properties[51]. Additional metrics considered were output by the Alignment Manipulation and Summary (AMAS) tool[52] with the summary command (Supplementary Dataset 3). Furthermore, we constructed concatenated β and β' phylogenies with all of these trimming stringencies to assess their effect on results; ultimately

we found that all trimming stringencies reliably recovered the deep branching mReC RNAP clade (Supplementary Fig. 4). We report the results of 30% gaps removed in main text Figures 2, 3, 4, and 6; and other trimming results are provided in Supplementary Fig. 4.

Identification of highly conserved regions. To ensure confidence in the alignment quality and that the sequences of all taxa compared were homologs, we searched for regions of sequences conserved among all taxa. Sequences of the subunits β and β' were aligned separately, and the alignment of each subunit was trimmed with a 30% gap-threshold. Highly conserved regions within each subunit were manually distinguished based on consecutive positions of increased conservation and identity as calculated in the annotations file of the alignment output by JalView[50] (version 2.11.1.0) the software in which the alignments were visualized (Figure 3, Supplementary Fig. 3). The average identity of these regions ranged from 66.039% to 81.832%, and conservation ranged from 7.000 to 7.679 (Supplementary Dataset 3). The location of these regions was linked to function, when possible, based on the structural annotations of Iyer et al. 2003[12] and Sauguet 2019[23]. Residues within the conserved regions were visualized on the structure of *Saccharomyces cerevisiae* RNAP II (PDB: 2e2i chain A, chain B corresponding to the β' homolog of RPB1 and β homolog of RPB2, respectively)[24] using the graphical software PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC) (Figure 4a).

Topology testing with removal of fast-evolving sites. Because viruses can have fast evolutionary rates, we reconstructed phylogenies after removing fast-evolving positions in the concatenated alignment that might have obscured the initial topology via long branch attraction. To identify fast-evolving sites in our concatenated β and β' alignment, we used TIGER[25] (version 1.02),

which categorizes positions in an alignment into bins based on substitution rates. We set TIGER to bin each position in the alignment into 1000 different rates of substitution. We then reconstructed phylogenies after removal of fast-evolving sites; we did this in increments of 5% of full alignment length up to the fastest-evolving 50% positions. Consistent with our initial methods, these trees were reconstructed using the LG+C60+F+Γ4 amino acid substitution model and 1,000-replicates for ultra-fast bootstrap approximation with IQ-TREE 1.6.11. Support for different topologies were assessed based on both ultrafast bootstrap values[53] and internode certainty values that were calculated with RAxML[47] (version 8.2.12). These branch support values were plotted against alignment length in R with the package ggplot2[54] (version 3.1.1), and results are presented in Supplementary Figure 5.

Long branch attraction assessment. In addition to alignment quality control and removal of fast-evolving sites, we performed the following analyses to ensure our results were robust and to limit the possible effect of Long Branch Attraction, which can manifest when highly divergent sequences are included in phylogenetic reconstructions.

First, we examined the similarity of the mReC RNAP β' subunit to that of other proteins known to have common motifs, which included eukaryotic RNA-dependent RNA polymerase (RdRp)[12] and crAssphage RNA polymerase genes[13]. We performed an HMM search (E-value cutoff 10^{-3}) of the RdRp proteins used in the study by Iyer at al. 2003[12] and the crAssphage RNAP proteins in the study Yutin et al. 2018[13] against COG0085 and COG0086 HMM profiles, and found no detectable sequence homology. Similarity between the eukaryotic RdRp was further examined by aligning the RdRp sequences with the COG0086 sequences (same as those used in Figure 2) with Clustal Omega. The alignment was trimmed for 30% gapped regions using trimAl and a

phylogeny was constructed using the LG+C60+F+Γ4 amino acid substitution model and 1000 ultrafast bootstrap replicates in IQ-TREE. Branch support was estimated with RAxML for absolute and relative internode certainty values. The resulting tree shows a clear case of LBA, with long branches of the RdRp clade within the NCLDV (Supplementary Fig. 8a).

As another test of LBA, we constructed another β' phylogeny in which we included 12 bacteriophage sequences from Viral RefSeq that hit to COG0086, but did not meet our initial sequence filtering criteria, as this protein was typically fragmented into different genes which resulted in low bitscores (details on the sequences included here can be found in Supplementary Dataset 2). We concatenated these fragmented sequences and aligned them with the RdRp and COG0086 proteins of taxa specified in Figure 2 using Clustal Omega. We then reconstructed a tree with maximum likelihood in IQ-TREE using the LG+C60+F+Γ4 amino acid substitution model and 1000 ultrafast bootstrap replicates, which yielded a tree that grouped the Viral RefSeq proteins with the RdRp (Supplementary Fig. 8b). The unstable branching of both RdRp and Viral RefSeq proteins suggests that LBA is a major issue for these sequences, and provides justification for their exclusion from subsequent analyses.

To further assess if the mReC monophyly held when only the mReC RNAP and Bacteria were compared alone, we constructed a concatenated alignment of the mReC and Bacteria RNAP in COG0085 and COG0086 amino acid sequences with the ete3 standard workflow. The resulting alignment was trimmed for 30% gapped positions, and we then constructed the phylogeny with maximum likelihood using the LG+C60+F+Γ4 amino acid substitution model and 1000 ultrafast bootstrap replicates in IQ-TREE. Bootstrap and Internode Certainty support values were calculated with RaXML. The resulting tree (Supplementary Fig. 9) once again recovered distinct clades of mReC and bacterial RNAP. This further suggests that the distinct clustering of mReC

sequences is not due to long branch repulsion away from both bacterial and archaeal/eukaryotic sequences.

To test if our results were maintained when using a different phylogenetic reconstruction method, we inferred the tree of Figure 2 with Bayesian approaches using PhyloBayes 4.1c[26] in which we ran two independent chains with the mixture model CAT+GTR, and the heterogeneity of site evolutionary rates were modeled using a gamma distribution with 4 categories. Supermatrices were recoded with the Dayhoff 6 scheme. The chains ran until convergence (maxdiff <0.3) which was assessed with bpcomp using 1000 burn-in trees and checking every 10 trees to calculate posterior consensus. The consensus tree (Supplementary Fig. 6) maintained the topology observed in the maximum likelihood reconstruction.

Rooting analysis. Toward resolving a root in our RNAP phylogeny, we performed a rooting analysis based on the paralogy of the β and β' subunits. Leveraging the ancient gene duplication history of the β and β' subunits, one subunit can serve as the outgroup of the other subunit. First, we aligned the sequences of each subunit individually with Clustal Omega. We then trimmed this alignment with 30% gap threshold with trimAl. Next, we aligned these alignments of each subunit to each other with the profile-profile alignment option in Clustal Omega. To ensure the β and β' subunits are indeed paralogous and contain enough similarity for phylogenetic use, we visualized the alignment with JalView and identified regions conserved between the subunits and all taxa based on identity and conservation (Supplementary Dataset 3, Supplementary Fig. 7). One of these regions corresponded to the double-psi beta barrel structures conserved among both subunits[12,23,54]. All conserved regions were visualized on the structure of *Saccharomyces cerevisiae* RNAP II (PDB: 2e2i chain A, chain B)[24] in PyMOL (Figure 4b).

To confirm the observed topology, as performed with the unrooted analysis, we removed the fastest evolving sites belonging to up to 50% of the positions in the alignment in increments of 5% with TIGER and trimAl. We then reconstructed phylogenetic trees of alignments with sites belonging to the fastest evolving rates incrementally with IQ-TREE using the LG+C60+F+Γ4 amino acid substitution models and 1000-replicates for ultra-fast bootstrap approximation. Branch support for different topologies was inferred based on ultra-fast bootstrap values and internode certainty values calculated with RAxML (Figure 6a). These branch support values were recorded and plotted against alignment length in R with the package ggplot2[54] (Figure 6b).

Major capsid protein and terminase diversity assessment. To explore the diversity of major capsid proteins (MCPs) and large subunits of the terminase protein (TerL) in the mReC relative to other bacteriophage, we performed HMM searches of protein sequences from all mReC contigs, all contigs in IMGVR 2.0, and viral RefSeq genomes that had bacterial hosts against HMM profiles of bacteriophage MCP and TerL in the VOG and Pfam databases (Supplementary Dataset 1). Eight of the mReC contigs encoded MCPs and 31 encoded TerL. All mReC MCP proteins had best hits to the same VOG and Pfam families (VOG11186 and PF07068, respectively). Of all 17 mReC proteins with hits to a known TerL in the Pfam database, all had best hits to the same family (PF03237). Of 31 total mReC proteins had hits to a TerL family in VOG, 27 had best hits to VOG01069, and the remaining 4 had hits to different VOG TerL families (Supplementary Dataset 1). Phylogenetic trees including both mReC proteins and reference proteins with best matches to the same protein family were constructed to evaluate if mReC proteins tended to cluster within the same clade. Due to the large number of reference proteins in IMGVR that had matches to the same Pfam and VOG protein families as the mReC

MCP and TerL proteins, we randomly selected 500 reference proteins from the total hits in the IMGVR and RefSeq databases using seqtk subseq command[55]. We then generated alignments using Clustal Omega and phylogenetic trees using IQ-TREE (best model selected using the ModelFinder tool[56], 1000 ultrafast bootstraps used, Figure 5).

Data Availability. Sequences used included RefSeq: NCBI Reference Sequence Database release 96 (https://www.ncbi.nlm.nih.gov/refseq/) with accession numbers specified in Supplementary Dataset 2, Integrated Microbial Genomes / Virus release January 2018 (version 4) (https://genome.jgi.doe.gov/portal/IMG_VR/IMG_VR.home.html). Contigs used in this study are listed in Supplementary Dataset 2. Protein family HMM profiles were downloaded from the following databases with their version or release number in parentheses: Clusters of Orthologous Groups (COG, 2003, 2014 update;https://www.ncbi.nlm.nih.gov/COG/), eggNOG (v5.0; http://eggnog5.embl.de/#/app/downloads), Pfam (release 32; ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/Pfam-A.hmm.gz), and Virus Orthologous Groups (VOG, downloaded April 14, 2020;vogdb.org). Amino acid sequences, alignments, phylogenetic trees, and tree branch support values are available at https://github.com/scubalaina/Bacteriophage_RNAP.

Code Availability. All software used was publicly available and cited with options reported in the Methods.

References

1. Darwin, C. On the origin of species. (1871) doi:10.5962/bhl.title.28875.
2. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–

740 (1997).

3. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 4576–4579 (1990).
4. Forterre, P. The universal tree of life: an update. *Front. Microbiol.* **6**, 717 (2015).
5. Brüssow, H. The not so universal tree of life or the place of viruses in the living world. *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 364 2263–2274 (2009).
6. Raoult, D. & Forterre, P. Redefining viruses: lessons from Mimivirus. *Nat. Rev. Microbiol.* **6**, 315–319 (2008).
7. Klenk, H.-P., Palm, P. & Zillig, W. DNA-Dependent RNA Polymerases as Phylogenetic Marker Molecules. *Systematic and Applied Microbiology* vol. 16 638–647 (1993).
8. Roux, S., Enault, F., Bronner, G. & Debroas, D. Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems. *FEMS Microbiol. Ecol.* **78**, 617–628 (2011).
9. Walsh, D. A., Bapteste, E., Kamekura, M. & Doolittle, W. F. Evolution of the RNA polymerase B' subunit gene (rpoB') in Halobacteriales: a complementary molecular marker to the SSU rRNA gene. *Mol. Biol. Evol.* **21**, 2340–2351 (2004).
10. Klenk, H.-P., Zilllg, W., Lanzendorfer, M., Grampp, B. & Palm, P. Location of Protist Lineages in a Phylogenetic Tree Inferred from Sequences of DNA-dependent RNA Polymerases. *Archiv für Protistenkunde* vol. 145 221–230 (1995).
11. Guglielmini, J., Woo, A., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes.

doi:10.1101/455816.

12. Iyer, L. M., Koonin, E. V. & Aravind, L. Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struct. Biol.* **3**, 1 (2003).
13. Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol* **3**, 38–46 (2018).
14. Forrest, D., James, K., Yuzenkova, Y. & Zenkin, N. Single-peptide DNA-dependent RNA polymerase homologous to multi-subunit RNA polymerase. *Nat. Commun.* **8**, 15774 (2017).
15. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
16. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).
17. Paez-Espino, D. *et al.* IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Res.* **45**, D457–D465 (2017).
18. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–9 (2015).
19. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
20. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
21. Lax, G. *et al.* Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature* **564**, 410–414 (2018).

22. Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6670–6675 (2015).
23. Sauguet, L. The Extended 'Two-Barrel' Polymerases Superfamily: Structure, Function and Evolution. *J. Mol. Biol.* **431**, 4167–4183 (2019).
24. Wang, D., Bushnell, D. A., Westover, K. D., Kaplan, C. D. & Kornberg, R. D. Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* **127**, 941–954 (2006).
25. Cummins, C. A. & McInerney, J. O. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Syst. Biol.* **60**, 833–844 (2011).
26. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**, 611–615 (2013).
27. Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. & Miyata, T. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9355–9359 (1989).
28. Brown, J. R. & Doolittle, W. F. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 2441–2445 (1995).
29. Zhaxybayeva, O., Lapierre, P. & Gogarten, J. P. Ancient gene duplications and the root(s) of the tree of life. *Protoplasma* **227**, 53–64 (2005).
30. Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nat. Rev. Microbiol.* **9**, 85–98 (2011).

31. Mizuno, C. M. *et al.* Numerous cultivated and uncultivated viruses encode ribosomal proteins. *Nat. Commun.* **10**, 752 (2019).
32. Yutin, N., Wolf, Y. I. & Koonin, E. V. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology* vols 466-467 38–52 (2014).
33. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat Microbiol* **4**, 1727–1736 (2019).
34. Koonin, E. V., Krupovic, M., Ishino, S. & Ishino, Y. The replication machinery of LUCA: common origin of DNA replication and transcription. *BMC Biol.* **18**, 1–8 (2020).
35. Brister, J. R., Rodney Brister, J., Ako-adjei, D., Bao, Y. & Blinkova, O. NCBI Viral Genomes Resource. *Nucleic Acids Research* vol. 43 D571–D577 (2015).
36. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
37. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
38. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution* vol. 33 1635–1638 (2016).
39. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
40. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
41. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
42. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*

47, D309–D314 (2019).

43. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
44. Guy, L., Kultima, J. R. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).
45. R Core Team. *A language and environment for statistical computing*. (2018).
46. RStudio Team. *RStudio: Integrated Development for R*. (RStudio Team, 2020).
47. Stamatakis, A. Using RAxML to Infer Phylogenies. *Current Protocols in Bioinformatics* 6.14.1–6.14.14 (2015) doi:10.1002/0471250953.bi0614s51.
48. Salichos, L., Stamatakis, A. & Rokas, A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol. Biol. Evol.* **31**, 1261–1271 (2014).
49. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
50. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
51. Livingstone, C. D. & Barton, G. J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9**, 745–756 (1993).
52. Borowiec, M. L. AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ* **4**, e1660 (2016).
53. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

54. Wickham, H. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics* vol. 3 180–185 (2011).
55. Li, H. Seqtk Toolkit for processing sequences in FASTA/Q formats. *Seqtk GitHub* https://github.com/lh3/seqtk (2013).
56. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* vol. 14 587–589 (2017).

Acknowledgements

We thank members for the Aylward Lab for helpful discussions. ARW was supported by a Doctoral Scholarship from the Virginia Tech Institute for Critical Technology and Applied Science. We acknowledge the National Science Foundation (NSF-IIBR 1918271) and a Simons Early Career Grant in Maine Microbial Ecology and Evolution to FOA. We acknowledge the use of the Advanced Research Computing resources at Virginia Tech.

Author Contributions. ARW and FOA designed the experiment, ARW performed the research, and ARW and FOA wrote the manuscript.

Competing Interests. The authors declare no competing interests.

Figures

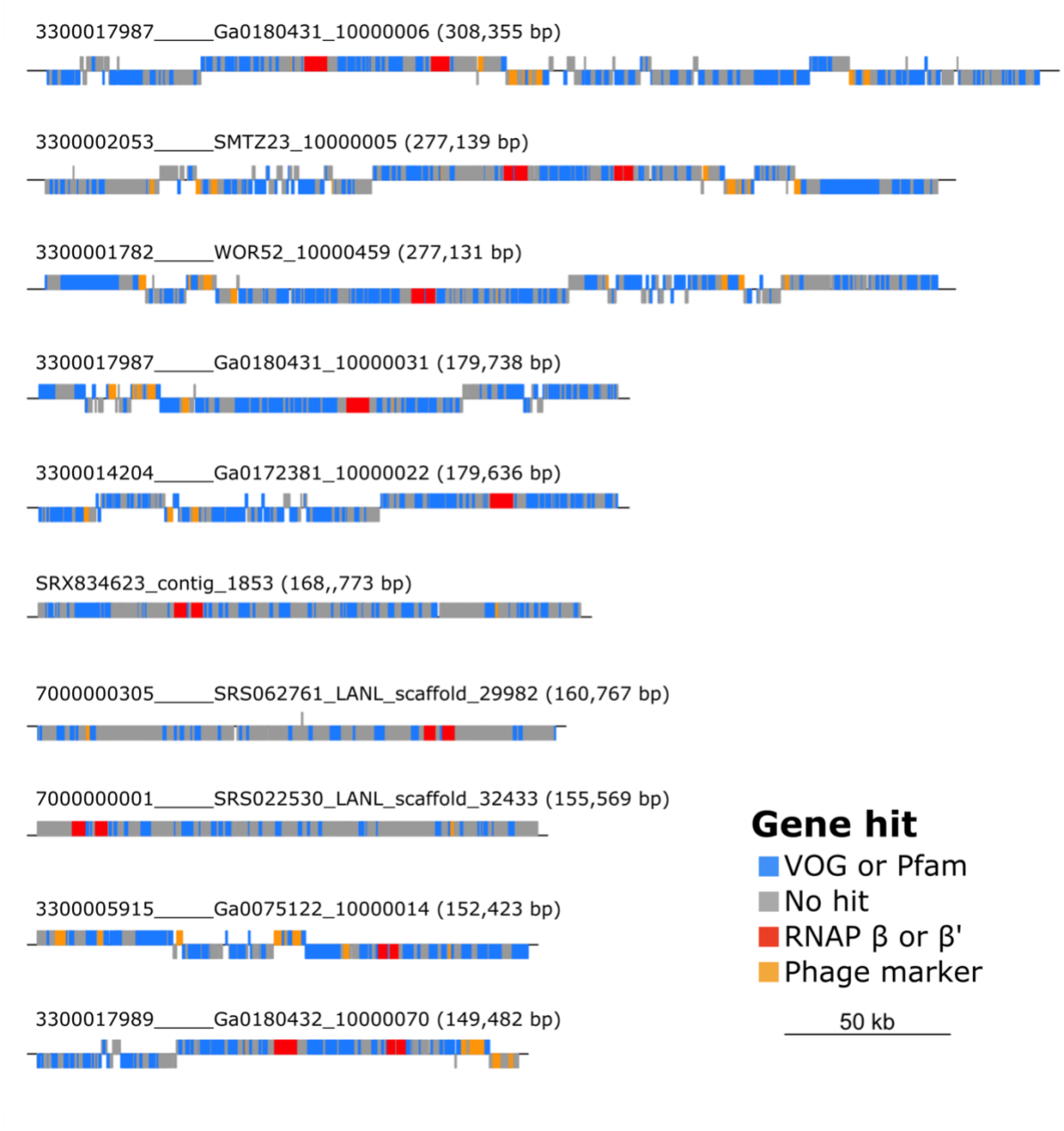


Figure 1. Genome plots of the ORFs of the ten longest mReC contigs. Above each plot is the contig name with its length in parentheses. Color corresponds to gene or database. Phage marker genes include baseplate wedge, portal proteins, capsids, terminases, and tail proteins. Scale bar corresponds to genome length of 50 kb.

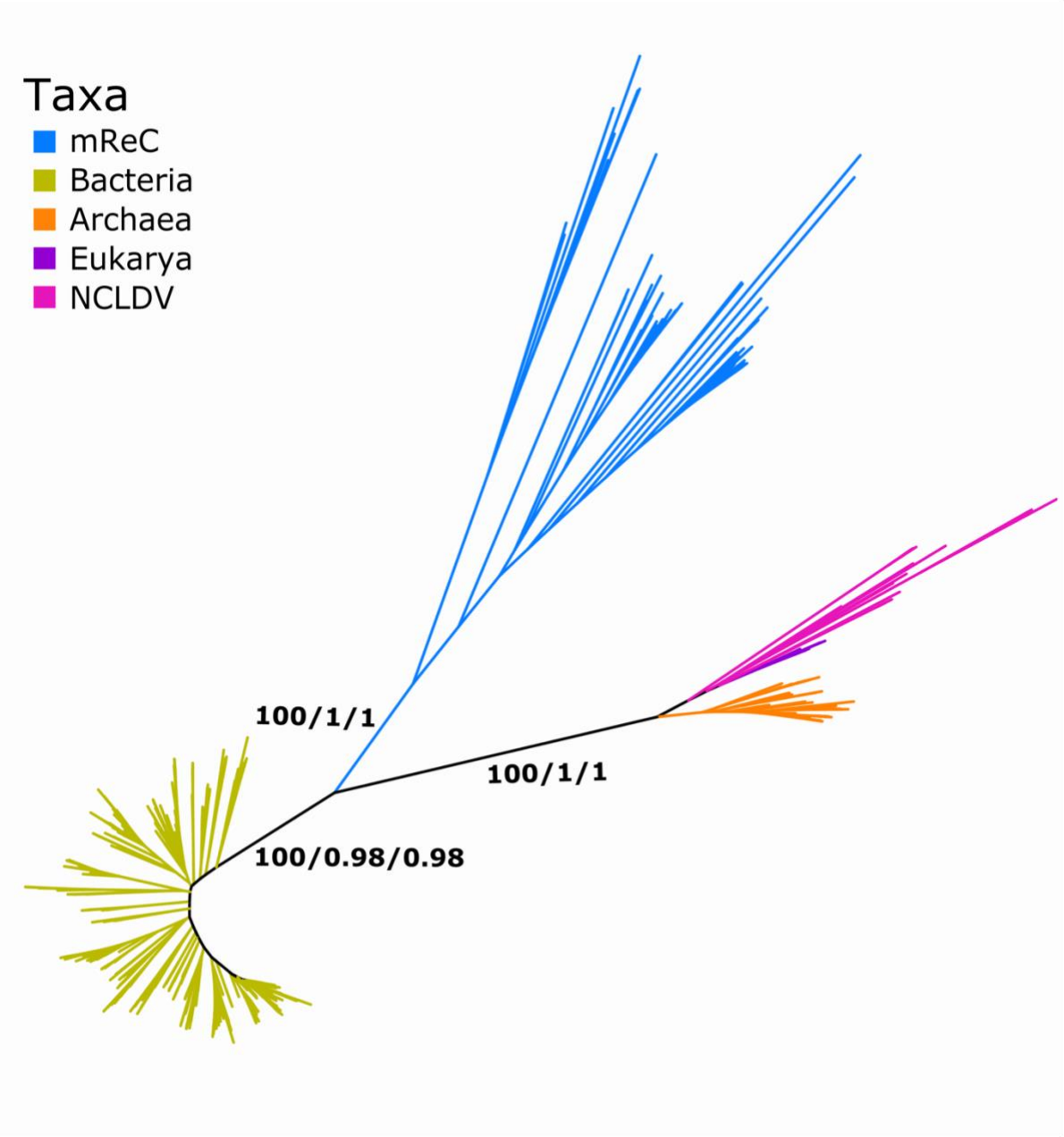


Figure 2. Unrooted phylogeny of concatenated RNAP β and RNAP β' amino acid sequences. Homologs of β and β' used for Eukarya were RPB2 and RPB1, and homologs B and A in Archaea, respectively. Phylogeny was constructed from the concatenated alignment of RNAP β and β' of 589 taxa constructed using maximum likelihood with the amino acid substitution model LG+C60+F+Γ4. Branch color corresponds to taxonomic group. Branch

support values are from left to right: ultrafast bootstrap from 1000 replicates reported as a proportion out of 100, relative Internode Certainty (IC) out of 1, absolute IC out of 1. Instructions for accessing full trees with support values can be found in the Data Availability section.

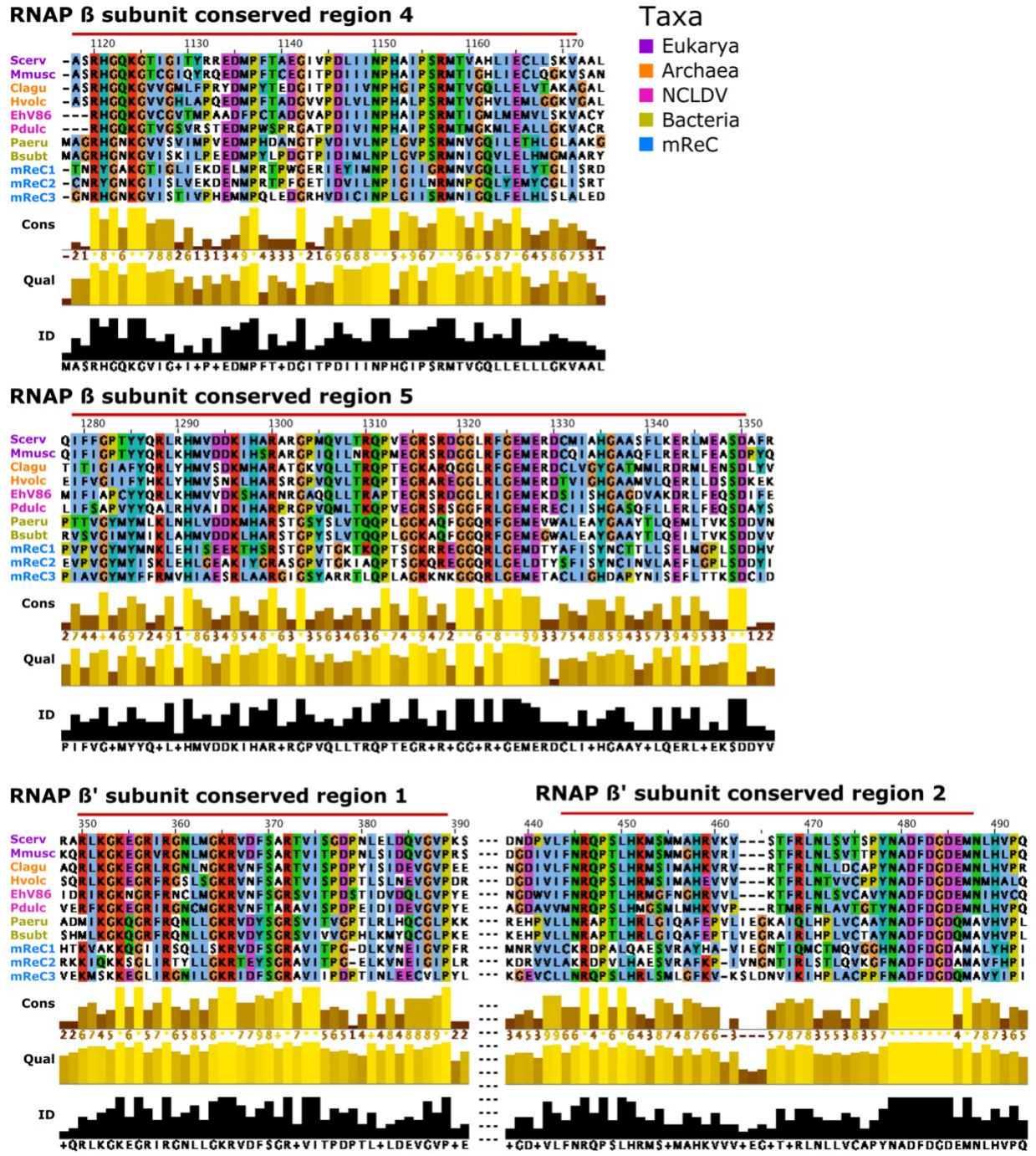


Figure 3. Highly conserved regions of the RNAP β and β' subunits alignment conserved across taxa. Eukarya homologs RPB2 and RPB1, Archaea homologs B and A, were used for β and β', respectively. Scerv (*S. cerevisiae*), Mmusc (*Mus* musculus), Clagu (*Caldisphaera lagunensis*), Hvolc (*Haloferax volcanii*), Pdulc (*Pandoravirus dulcis*), Paeru (*Pseudomonas aeruginosa*), Bsubt (*Bacillus subtillis*), mReC 1, 2, and 3 (multisubunit RNAP-encoding *Caudovirales* 1, 2, and 3).

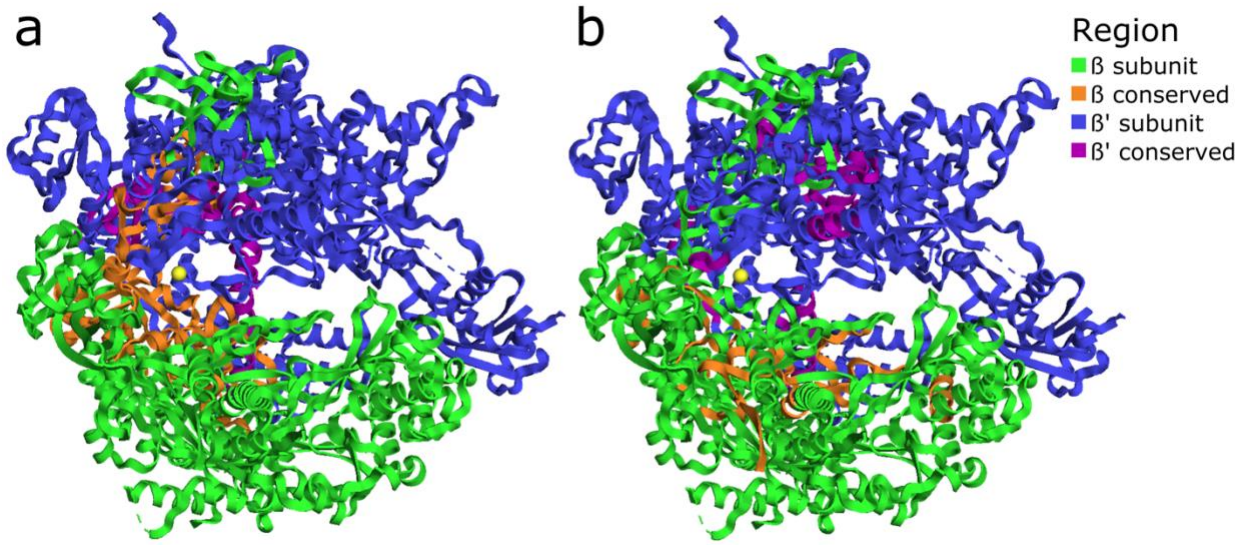


Figure 4. Conserved regions visualized in yeast RNAP protein structure. Image of RNAP β (RPB2 subunit in eukaryotes) and RNAP β' (RPB1 subunit in eukaryotes) subunit structures of *S. cerevisiae* (PDBid: 2e2i)[24]. Colors correspond to subunit and conserved regions. (a) Regions conserved across all examined taxonomic groups (Supplementary Dataset 3) when amino acid sequences of β and β' are aligned separately. (b) Regions conserved between β and β' across all examined taxonomic groups when β and β' are aligned to each other (Supplementary Dataset 3).

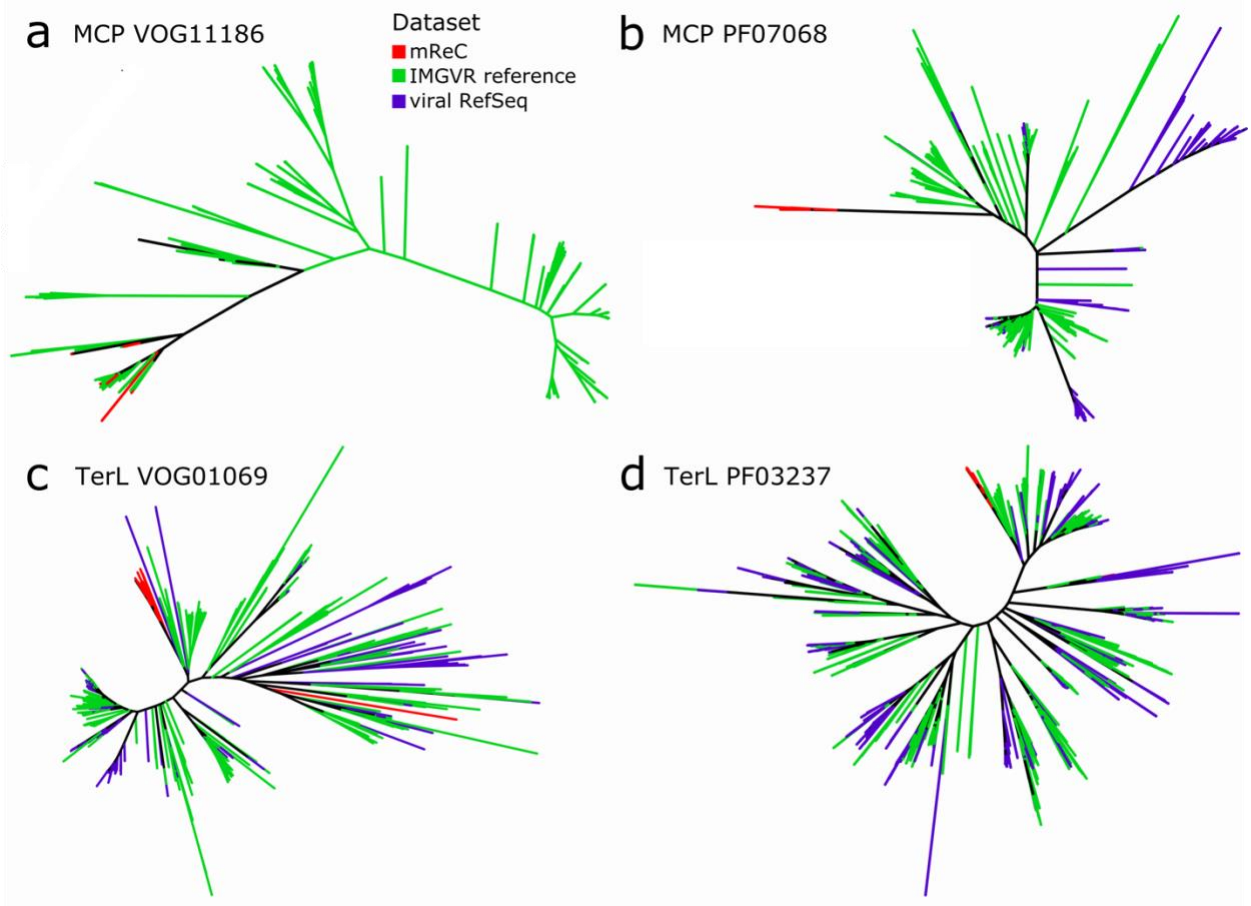


Figure 5. Unrooted phylogenies of mReC capsid and terminase proteins together with references. In panels a and b, phylogenies of the protein sequences for the major capsid protein (MCP), and in panels c and d, terminase large subunit protein (TerL). from the VOG (panels a and c) and Pfam (panels b and d) databases. Tree was inferred with maximum likelihood in IQ-TREE using ModelFinder. Branch color corresponds to the source of the sequences. Instructions for accessing full trees with support values can be found in the Data Availability section.

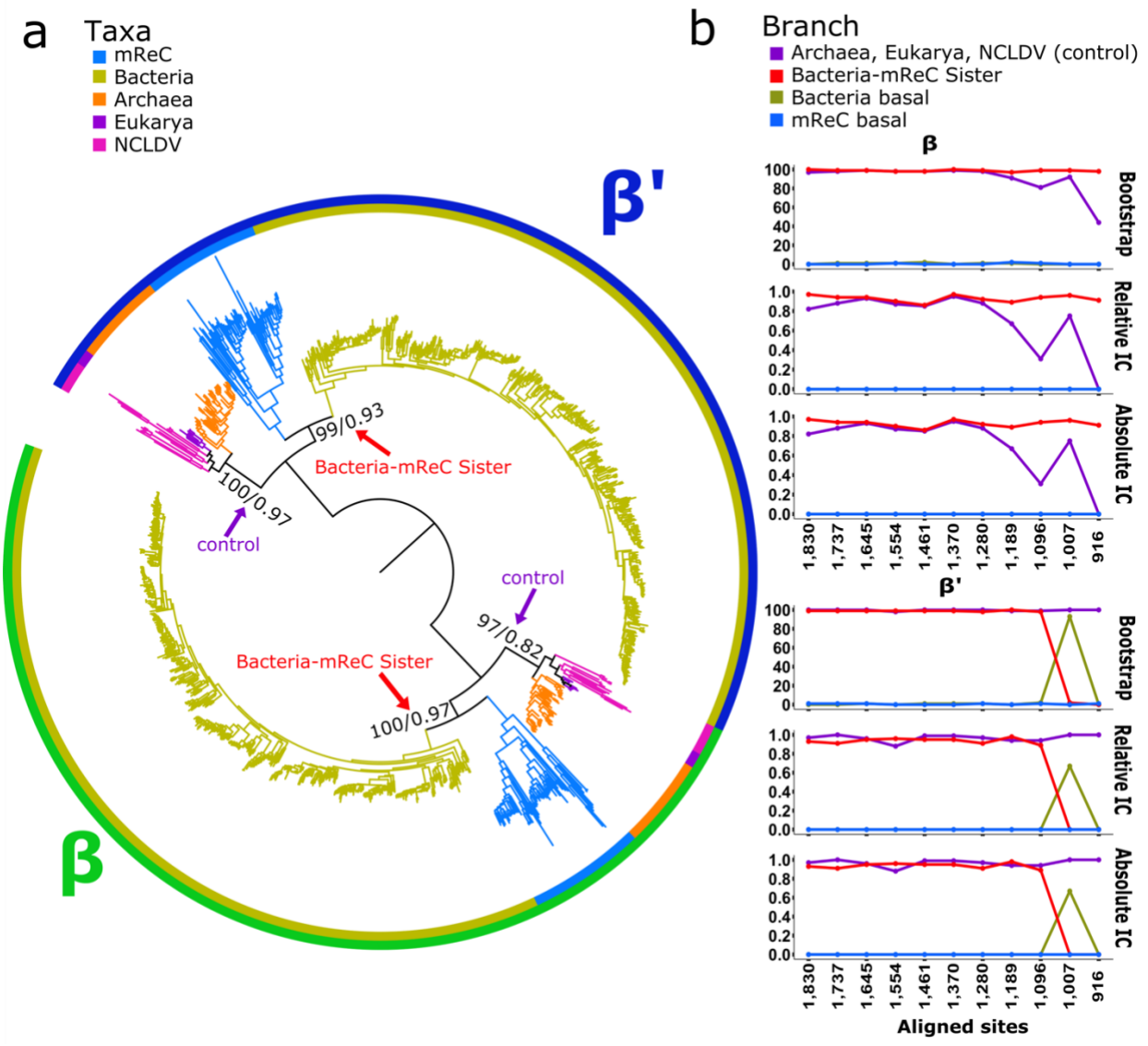


Figure 6. Paralogy-based rooting analysis of the RNAP tree. a) Rooted tree of β and β' subunits. Homologs of the β and β' subunits used for Eukarya were RPB2 and RPB1, and homologs B and A used for Archaea, respectively. Amino acid sequences for each subunit were aligned to each other, and phylogeny was constructed using maximum likelihood with the amino acid substitution model LG + C60 + F + Γ4. Branch color and inner ring color strip corresponds to taxonomic group. Outer ring color strip corresponds to subunit β (green) and β' (blue). At selected branches, first number refers to ultrafast bootstrap support of 1000 replicates reported as a percent out of 100 and the second number refers to relative Internode Certainty (IC) value out

of 1. Arrows point to branches used to assess support of trees as fast evolving sites were removed (See Methods). b) Line plots of branch support corresponding to different hypotheses in the β (upper) and β' (lower) clusters as fast evolving sites are removed in steps of 5% until 50% of the fastest evolving sites are removed from the alignment (See Methods). Purple corresponds to branch support of Archaea, Eukarya, and NCLDV RNAP together. Blue line corresponds to the branch supporting mReC RNAP is basal to all other lineages considered. Yellow line corresponds to the branch supporting bacterial RNAP is basal to all other lineages considered. Red line corresponds to the branch supporting that mReC and bacterial RNAP diverged together prior to other groups. Instructions for accessing full trees with support values can be found in the Data Availability section.

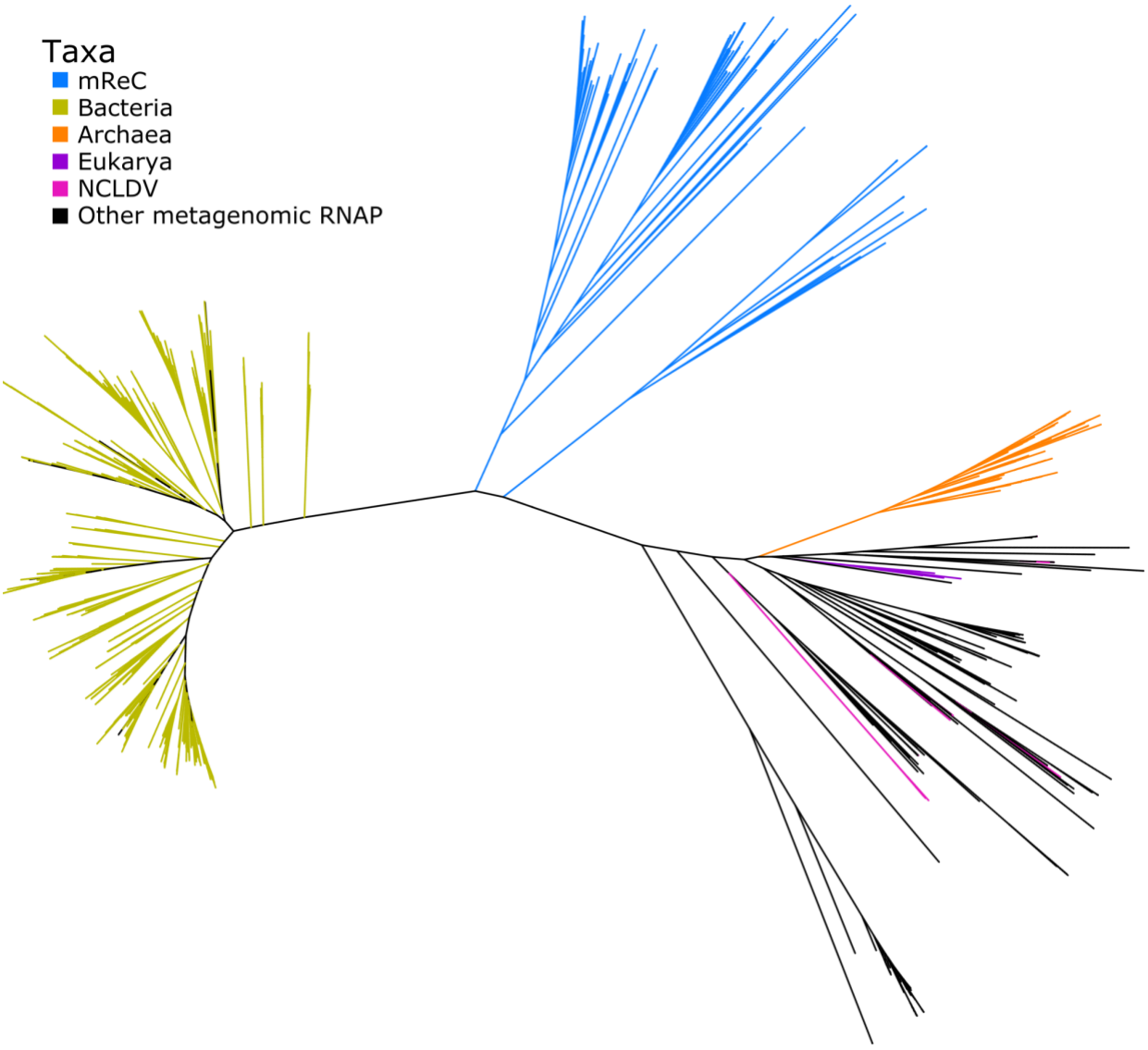
Supplemental Information

Supplementary Datasets. All supplementary datasets from this study can be found on the GitHub link https://github.com/scubalaina/Bacteriophage_RNAP .

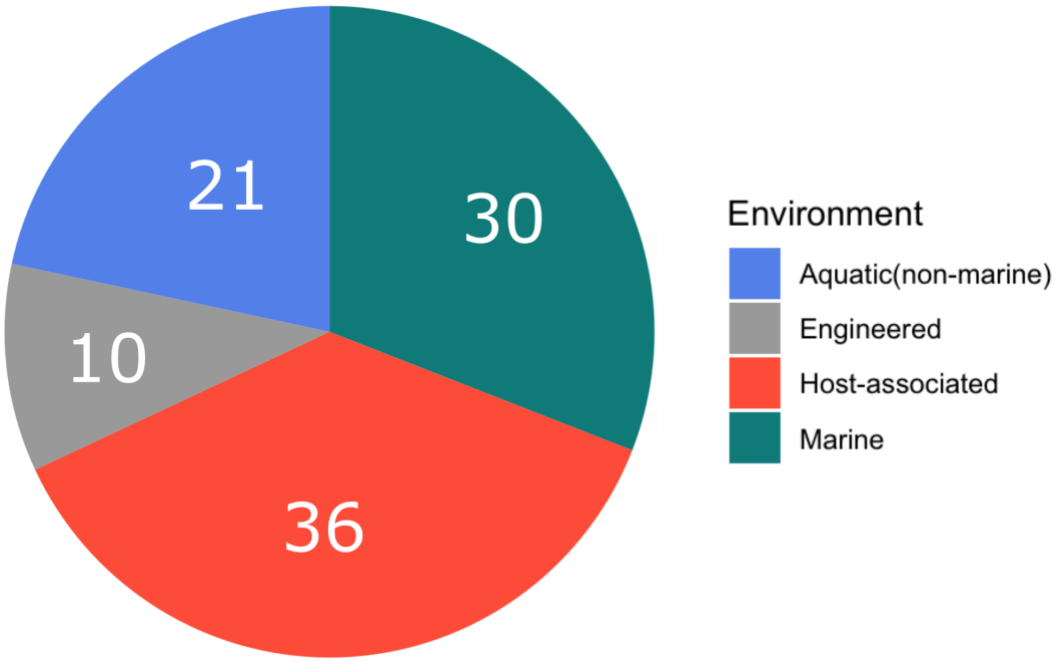
Supplemental Dataset 1: Viral sequence information including ViralRecall output, number of MCP hits, number TerL hits, the HMM output of proteins used for MCP and TerL trees, the full annotation of the mReC contigs, and the environments of the mReC contigs.

Supplemental Dataset 2: Information on RNA polymerase sequences included in phylogenies, as well as RdRp and other viral sequences, parsed HMM output of Viral RefSeq hits to COG0085 and COG0086.

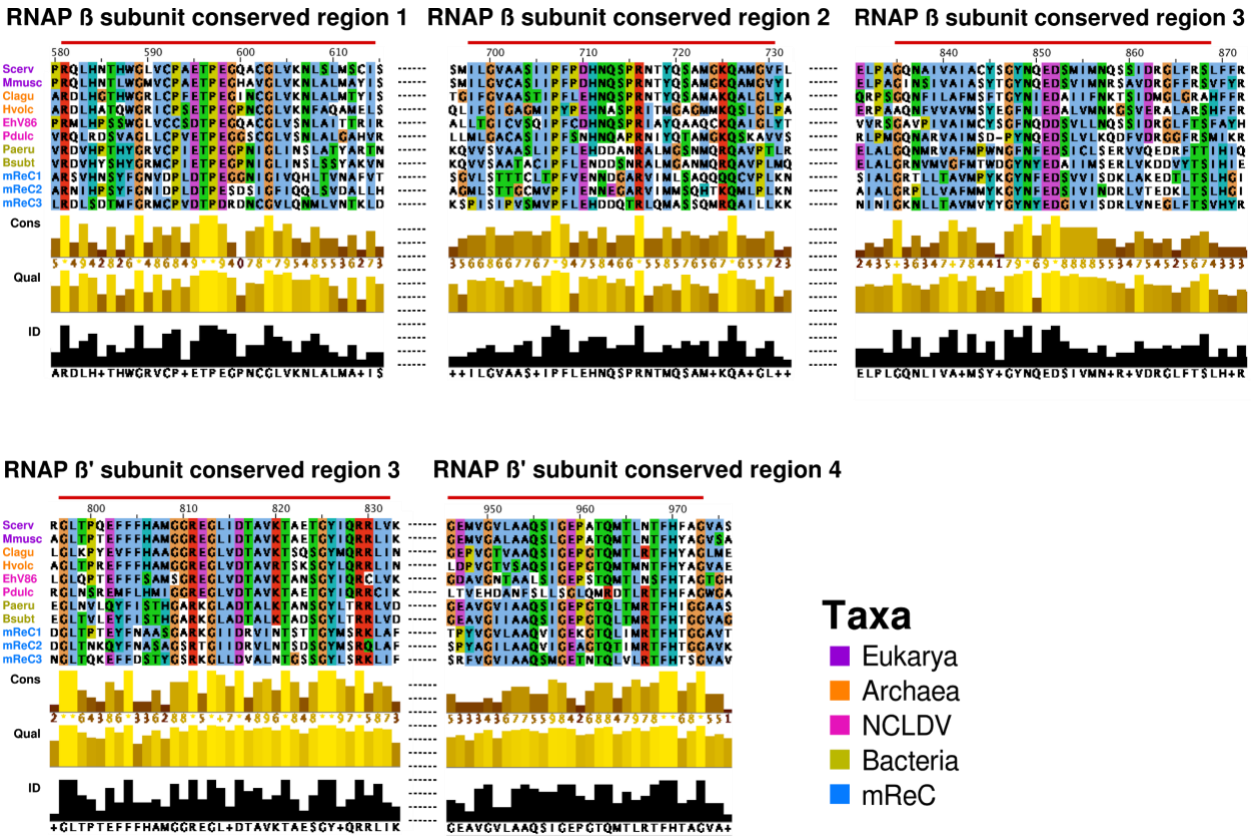
Supplemental Dataset 3: Sequence alignment composition for phylogenies.



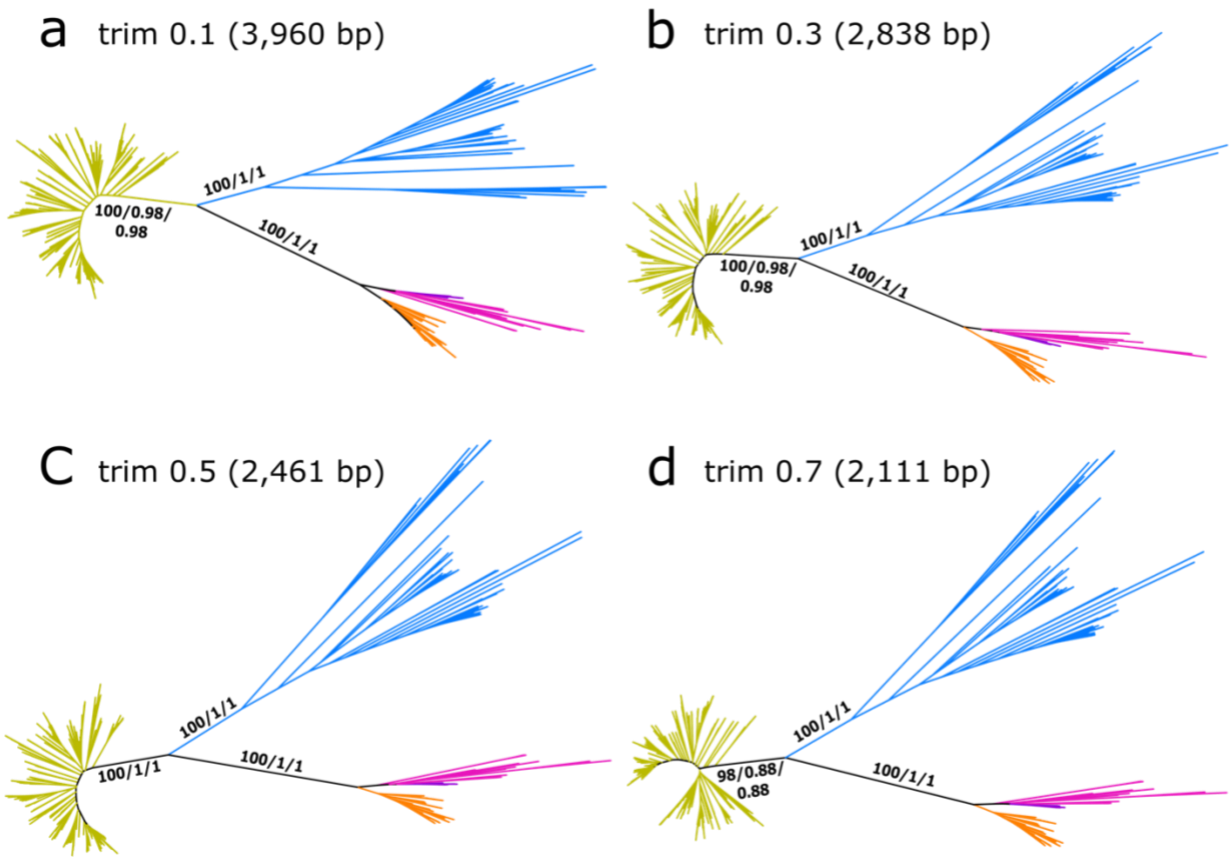
Supplementary Figure 1. Unrooted diagnostic tree constructed with FastTree of the concatenated β and β' amino acid sequence alignment (taxa in Dataset 1) for the initial identification of deep-branching bacteriophage RNAP. Branch color corresponds to taxa.



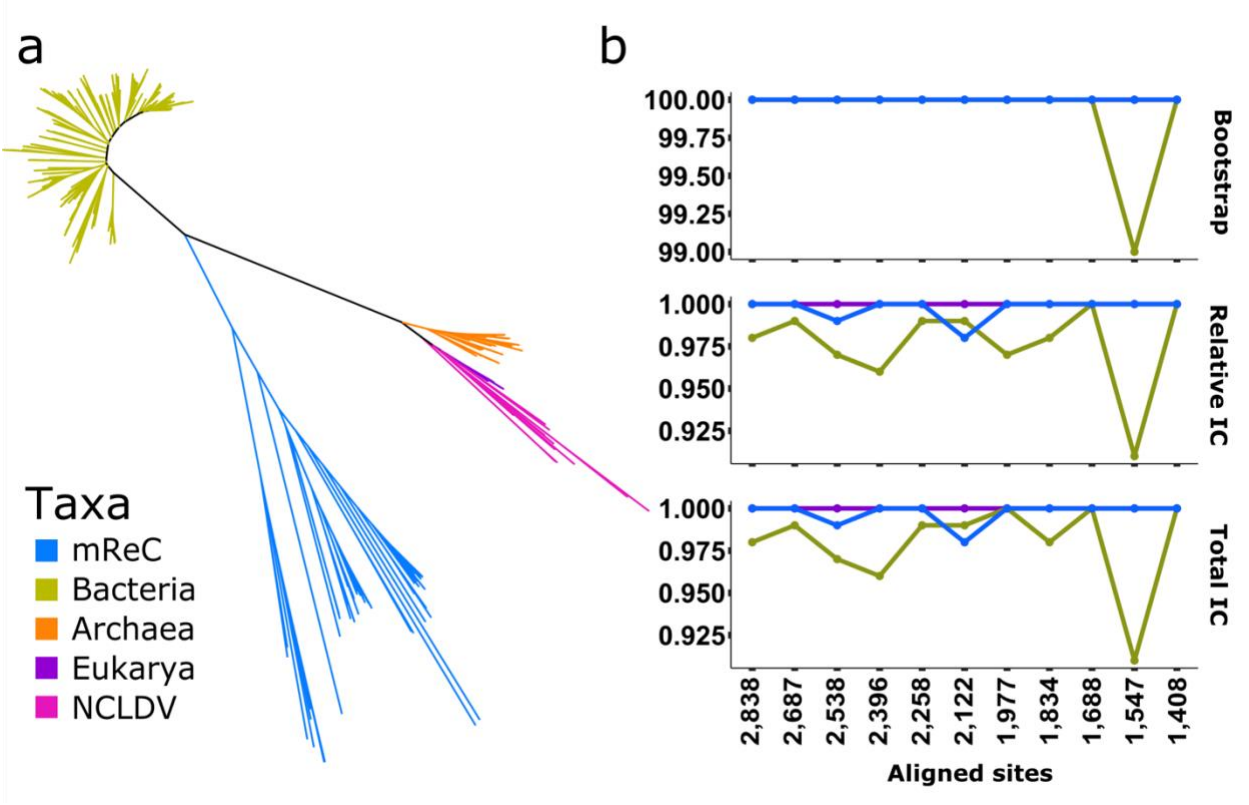
Supplementary Figure 2. Environmental distribution of the metagenomes from which mReC were identified (Dataset S3).



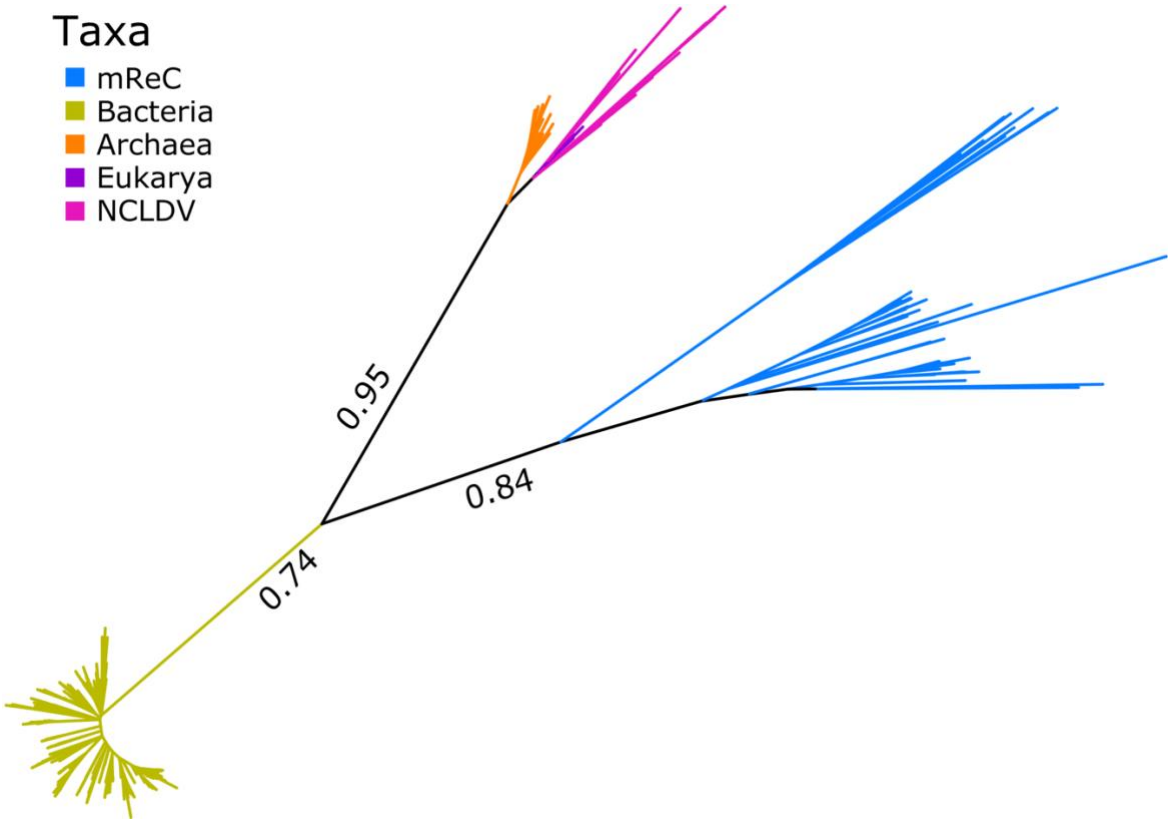
Supplementary Figure 3. Visualization of conserved regions, additional to those in the main text (Figure 2.) within each subunit of representative taxa as displayed in JalView with colors of residues corresponding to biochemical properties according to the Clustalx code. Colors of taxa on left correspond to taxonomic group according to Group table. Scerv (*S. cerevisiae*), Mmusc (*Mus* musculus), Clagu (*Caldisphaera lagunensis*), Hvolc (*Haloferax volcanii*), Pdulc (*Pandoravirus dulcis*), Paeru (*Pseudomonas aeruginosa*), Bsubt (*Bacillus subtillis*), mReC 1, 2, and 3 (multisubunit RNAP-encoding *Caudovirales* 1, 2, and 3).



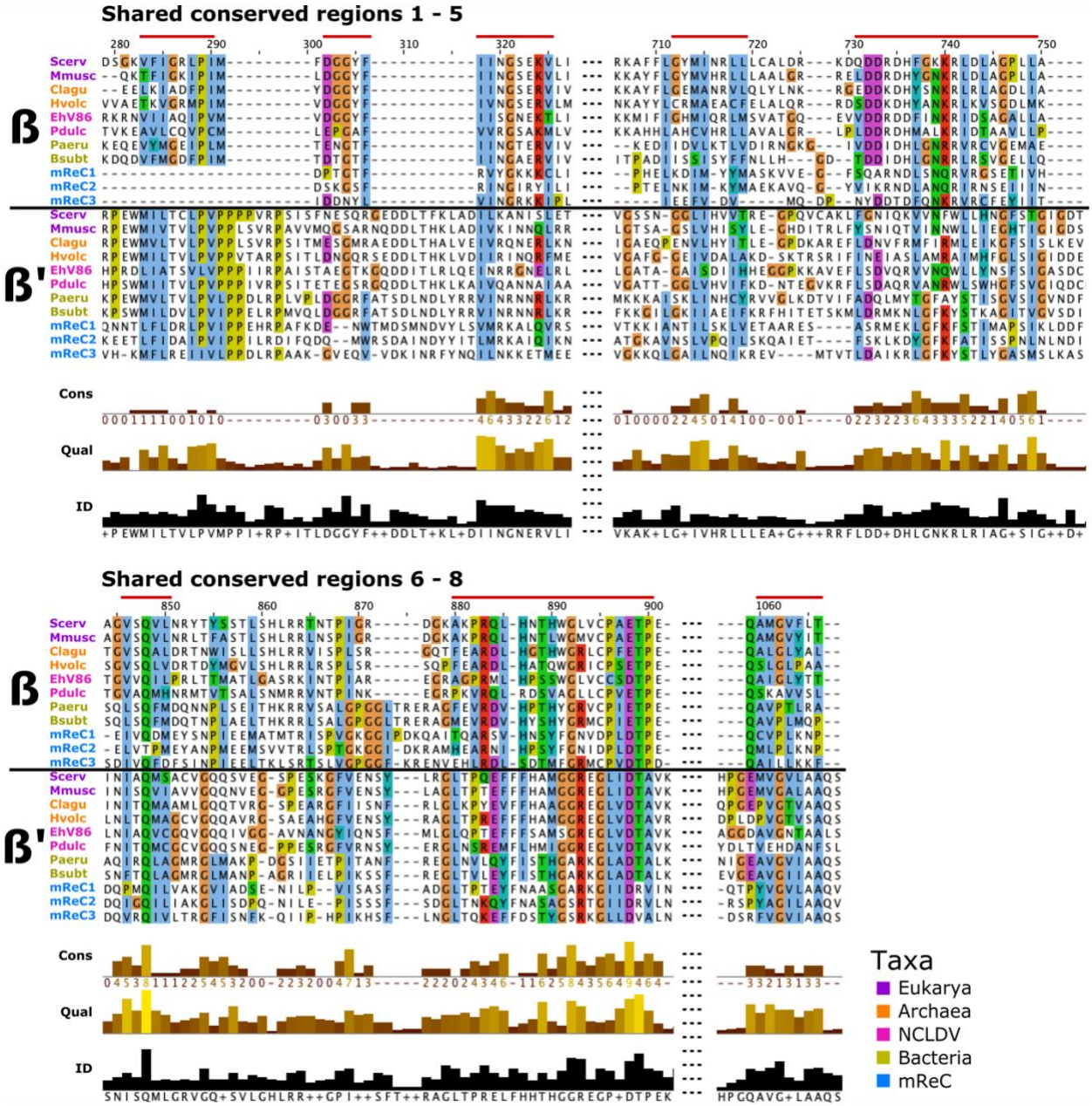
Supplementary Figure 4. Unrooted tree of concatenated β and β' amino acid sequence alignment constructed with maximum likelihood (LG+C60+F+Γ4 model in IQ-TREE). Same taxa as Figure 1. Branch colors correspond to taxonomy (yellow: Bacteria, blue: mReC, pink: NCLDV, purple: Eukarya, orange: Archaea). Each panel is a different gap-trimming stringency as a proportion out of 1 and alignment length is in parentheses. Branch support values from left to right: ultrafast bootstrap of 1000 replicates reported as a proportion out of 100, relative IC out of 1, and absolute IC values out of 1.



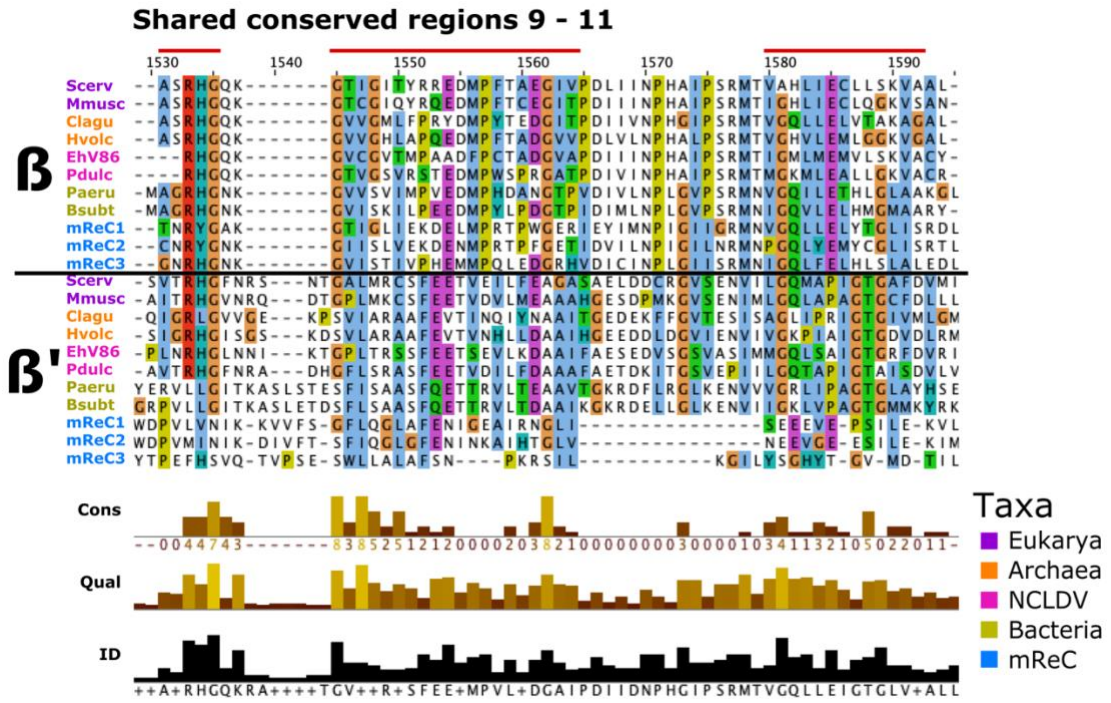
Supplementary Figure 5. a) Unrooted tree of concatenated β and β' amino acid sequence alignment of taxa in Figure 1 constructed with maximum likelihood (LG+C60+F+Γ4 model in IQ-TREE). b) Results of removing fast-evolving sites; plots of branch support values, ultrafast bootstrap of 1000 replicates reported as a percent out of 100, relative IC out of 1, and absolute IC values out of 1, for the Archaeal, Eukaryotic, NCLDV clade (purple), Bacterial clade (yellow), and mReC (blue).



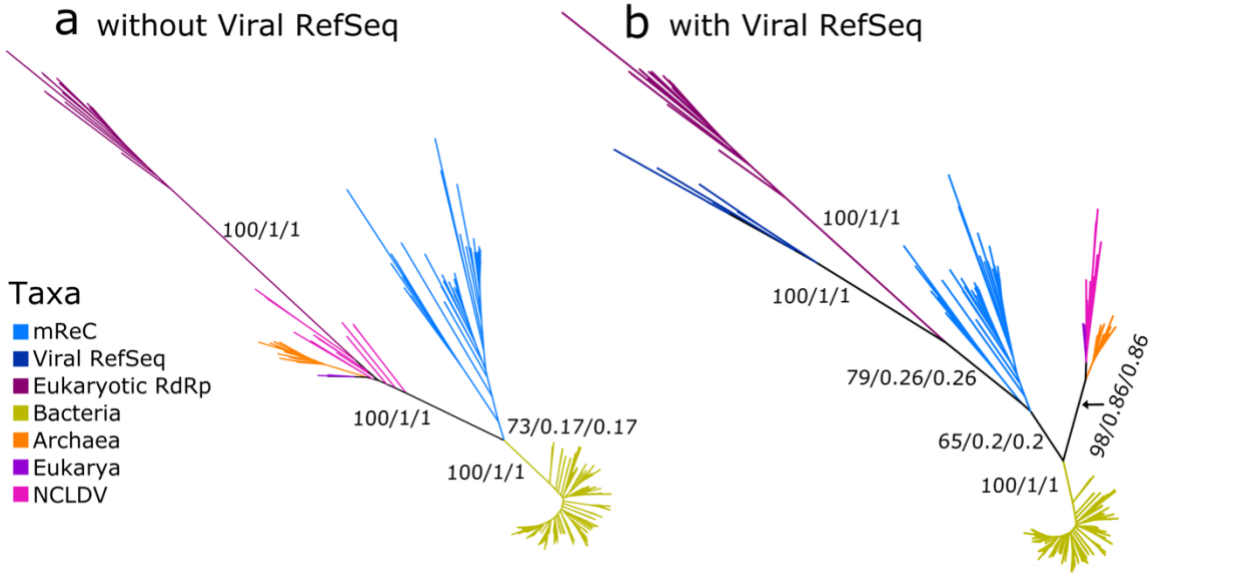
Supplementary Figure 6. Unrooted tree of concatenated β and β' amino acid sequence alignment of taxa in Figure 1 reconstructed with the Bayesian approach implemented in PhyloBayes 4.1c (CAT-GTR-**Γ**4) consensus tree (maxdiff < 0.3). Branch support corresponds to posterior probabilities. Branch color corresponds to taxa.



Supplementary Figure 7. Conserved regions between the aligned β and β' subunits of representative taxa as displayed in JalView Clustalx coloring (see Methods). For taxa abbreviations see Supplementary Fig. 2. Regions continue on next page.

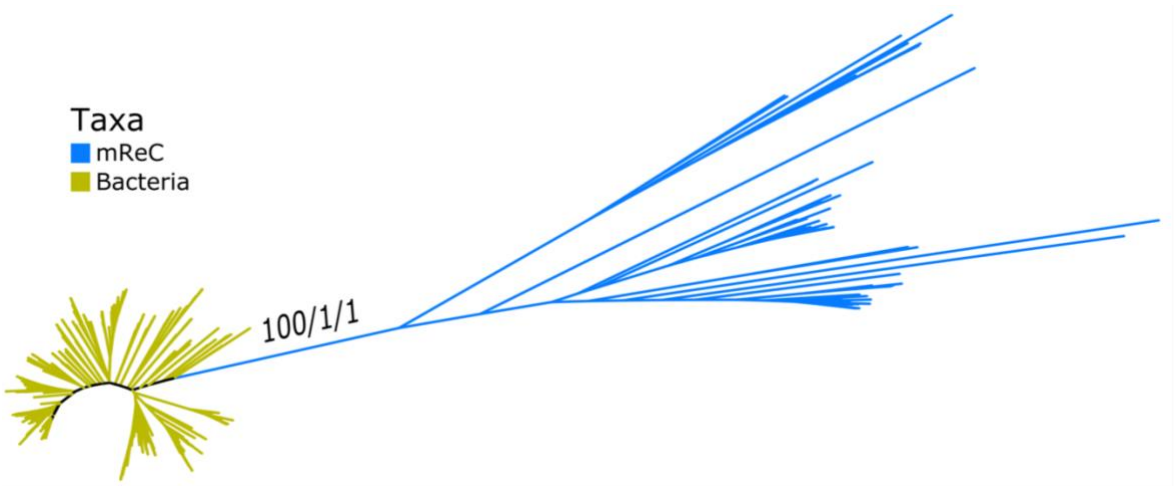


Supplementary Figure 7 continued.



Supplementary Figure 8. a) Unrooted phylogeny of eukaryotic RNA-dependent RNA polymerase (RdRp) aligned with β' amino acid sequences of taxa in Figure 1 constructed with

maximum likelihood (LG+C60+F+Γ4 model in IQ-TREE) with branch support values from left to right: ultrafast bootstrap support of 1000 replicates reported as a proportion out of 100, relative IC out of 1, and absolute IC out of 1. b) Phylogeny of sequences in a) aligned with Viral RefSeq β’ amino acid sequences (Dataset 1).



Supplementary Figure 9. Unrooted tree of concatenated alignment β and βamino acid sequences in Bacteria and mReC constructed with maximum likelihood (LG+C60+F+Γ4 model in IQTree) with branch support values from left to right: ultrafast bootstrap support of 1000 replicates reported as a proportion out of 100, relative IC out of 1, and absolute IC out of 1.

Chapter 3: A distinct lineage of *Caudovirales* that encodes a deeply branching multi-subunit RNA polymerase

Weinheimer, A. R., & Aylward, F. O. (2022). Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages. *The ISME Journal*, 16(6), 1657-1667.

Reproduced from the journal *The ISME Journal*.

Abstract

Recent research has underscored the immense diversity and key biogeochemical roles of large DNA viruses in the ocean. Although they are important constituents of marine ecosystems, it is sometimes difficult to detect these viruses due to their large size and complex genomes. This is true for "jumbo" bacteriophages, which have genome sizes >200 kbp and large capsids reaching up to 0.45 μm in diameter. In this study, we sought to assess the genomic diversity and distribution of these bacteriophages in the ocean by generating and analyzing jumbo phage genomes from metagenomes. We recover 85 marine jumbo phages that ranged in size from 201-498 kilobases, and we examine their genetic similarities and biogeography together with a reference database of marine jumbo phage genomes. By analyzing Tara Oceans metagenomic data, we show that although most jumbo phages can be detected in a range of different size fractions; 17 of our bins tend to be found in those greater than 0.22 μm, potentially due to their large size. Our network-based analysis of gene sharing patterns reveals that jumbo bacteriophages belong to five genome clusters that are typified by diverse replication strategies, genomic repertoires, and potential host ranges. Our analysis of jumbo phage distributions in the ocean reveals that depth is a major factor shaping their biogeography, with some phage genome clusters occurring preferentially in either surface or mesopelagic waters, respectively. Taken together, our findings indicate that jumbo

phages are widespread community members in the ocean with complex genomic repertoires and ecological impacts that warrant further targeted investigation.

Introduction

Although historically noted for their small virion sizes and simple genomes [1], viruses with large particles and elaborate genomes have been discovered in recent decades throughout the biosphere [2–4]. These complex viruses not only invite intriguing evolutionary questions [5–7], but also expand the potential roles viruses have in shaping microbial community structure and biogeochemical cycling [4, 8–10]. One group of these larger viruses are jumbo bacteriophages (jumbo phages), which have traditionally been defined as *Caudovirales* with genomes over 200 kilobases in length [11]. While a recent survey of cultured jumbo phages showed jumbo phages share some universal features and genes, such as encoding DNA polymerases and the terminase large subunit (TerL), these features do not distinguish them from smaller phages, and several lines of evidence suggest that jumbo phages emerged from smaller phages multiple times independently [6]. For example, a recent phylogenetic study of cultured *Caudovirales* used concatenated protein alignments to generate phylogenies and found that the most supported clades within the *Caudovirales* family do not consistently correspond to genome length [12]. Furthermore, a previous study found that jumbo phages cluster with smaller phages based on gene content and are best grouped by replication machinery, among other infection apparati [6]. Taken together, jumbo phages likely form distinct clades within the *Caudovirales*.

Although the first jumbo phages were isolated as early as the 1970s [13], these viruses have remained relatively sparse in culture, representing less than 3% (n=93) of complete *Caudovirales* genomes on NCBI's RefSeq Viral Genome Portal (downloaded July 5, 2020) and 2.2% of the INPHARED database [14]. All cultured jumbo phage capsids have morphologies of myoviruses

or siphoviruses, and they undergo infection cycles that reflect temporal patterns of lytic *Caudovirales* [15,16]. Some jumbo phages are known to stall infections resulting in "pseudolysogeny", however, which has been proposed as a competitive strategy against other phages to prevent superinfection [6]. Jumbo phages that have been studied extensively are primarily investigated for their unusually complex functional capabilities, such as encoding entire transcriptional apparatti [17] or sophisticated anti-CRISPR defense mechanisms [18, 19]. Regarding their ecological range, cultivated jumbo phages have been isolated on both Gram negative and Gram positive bacteria [15], and a recent metagenomic survey uncovered these viruses in diverse, global ecosystems [4].

Despite this apparent broad environmental distribution, common methods for viral isolation and diversity surveys often bias against the inclusion of jumbo phages. Because viruses have historically been considered smaller than cells, many viral diversity surveys specifically examine only small particle sizes. For example, in plaque assays, agar concentrations are often too high for larger phage particles to diffuse through compared to smaller particles [20]. Moreover, filters are often used to remove cells when preparing viral enrichments for metagenomic sequencing [21], which excludes larger viruses [9, 22]. Particularly in marine studies, the <0.22 μm fraction, sometimes even referred to as the "viral fraction" [23], is most commonly examined for viruses [24–26]. Jumbo phages can have particles over 0.45 μm in length (i.e. *Pseudomonas aeruginosa* phage PhiKZ) [13], however, and will therefore be excluded from <0.22 μm size fractions. Lastly, in bioinformatic pipelines, phage sequences are often only assembled to the contig or scaffold level, which is sometimes sufficient for the assembly of most known smaller phage genomes [27], but often leaves larger phage genomes fragmented into multiple contigs and may require additional joining of contigs into bins [28]. Overall, considering these biases and the recently-discovered

broad distribution of these viruses [4], jumbo phages may represent underappreciated components of marine microbial communities and food webs that warrant further examination.

In this study, we examine the diversity and prevalence of jumbo phages in the global ocean. We develop a workflow for generating and validating high-quality jumbo bacteriophage bins from metagenomic data with which we identify 85 bins of jumbo phages. We then compare the genetic content of these jumbo phages with other cultured phages of all sizes and metagenomic jumbo phages from other studies. We find that the jumbo phages of this study group into five distinct clusters that are distinguished by diverse replication machinery and infection strategies, implicating a broad range of potential jumbo phage-host interactions in the ocean. We then assess the distribution of jumbo phages belonging to these genome clusters in the ocean by using metagenomic data from Tara Oceans [29]. Mapping the Tara Oceans metagenomic data onto the jumbo phage sequences reveals that these jumbos phages are collectively widely distributed in the ocean, but vary in biogeography both within and between clusters, with some more enriched in surface waters relative to deeper waters and vice versa. Upon examining the collective presence of jumbo phages in different filter fractions, we also find that most could be detected in a range of different size fractions, although 34 were recovered from only >0.22 μm fractions. Our results support the view that jumbo phages are widespread in the biosphere and may play underappreciated roles in ecosystems around the globe.

Results and Discussion

Detection and validation of high-quality jumbo phage bins

Due to the large size of jumbo bacteriophage genomes, it is likely that they are present in multiple distinct contigs in metagenomic datasets and therefore require binning to recover high-quality metagenome-assembled genomes (MAGs) [28]. This has been shown for large DNA viruses that

infect eukaryotes, where several recent studies have successfully employed binning approaches to recover viral MAGs [2, 3, 30]. Here, we used the same 1,545 high-quality metagenomic assemblies [31] used in a recent study to recover giant viruses of eukaryotes [3], but we modified the bioinformatic pipeline to identify bins of jumbo bacteriophages. These metagenomes were compiled by Parks et al. 2017 [31] and included all available metagenomes on the NCBI's Short Read Archive by December 31, 2015 (see Parks et al. 2017 [31]). This dataset includes a wide variety of marine metagenomes (n=469) with many non-Tara metagenomes (n=165). We focused our benchmarking and distribution analyses on Tara data [29] because of the well-curated metadata and size fractions in this dataset. We first binned the contigs from these assemblies with MetaBat2 [32], which groups contigs together based on similar nucleotide composition and coverage profiles and focused on bins of at least 200 kilobases in total length. We subsequently identified bins composed of bacteriophage contigs through analysis with VirSorter2 [33], VIBRANT [34], and CheckV [35] (see Methods for details).

The occurrence of multiple copies of highly conserved marker genes is typically used to assess the level of contamination present in metagenome-derived genomes of bacteria and archaea [36]. Because bacteriophage lack these marker genes [37], we developed alternative strategies to assess possible contamination in our jumbo phage bins. Firstly, we refined the set of bins by retaining those with no more than 5 contigs that were each at least 5 kilobases in length to reduce the possibility that spurious contigs were put together. Secondly, we assessed the possibility that two strains of smaller phages with similar nucleotide composition may be binned together by aligning the contigs in a bin to each other. Bins that had contigs with high sequence similarity across the majority of their lengths were discarded (Supplemental Figure 1). Thirdly, we discarded bins if their contigs exhibited aberrant co-abundance profiles in different metagenomes (see

Supplemental Methods). To generate these co-abundance profiles, we mapped reads from 225 marine metagenomes provided by Tara Oceans [29] onto the bins. Coverage variation between contigs was benchmarked based on read-mapping results from artificially-fragmented reference genomes present in the samples (See Methods for details). Only bins with coverage variation below our empirically-derived threshold were retained. Using this stringent filtering, we identified 85 bins belonging to jumbo bacteriophages. These bins ranged in length from 202 bp to 498 kbp, and 31 consisted of a single contig, while 54 consisted of 2-5 contigs (Supplemental Figure 2).

To assess global diversity patterns of jumbo bacteriophages, we combined these jumbo phage bins together with a compiled database of previously-identified jumbo phages that included all complete jumbo *Caudovirales* genomes on RefSeq (downloaded July 5th, 2020), the INPHARED database [14], the cultured jumbo phage survey [6], the Al-Shayeb et al. 2020 study [4], and marine jumbo phage contigs from metagenomic surveys of GOV 2.0 [26] (60 jumbo phages), ALOHA 2.0 [38] (8 jumbo phages), and one megaphage MAG recovered from datasets of the English Channel [39]. Ultimately, we arrived at a set of 244 jumbo phages, including the 85 bins, that were present in at least one Tara Oceans sample (min. 20% genome covered, see Methods) or deriving from a marine dataset (i.e. ALOHA, GOV 2.0) which we analyzed further in this study and refer to as marine jumbo phages. Statistics on genomic features can be found in Supplemental Dataset 1.

Marine jumbo phages belong to distinct groups with diverse infection strategies

Because bacteriophages lack high-resolution, universal marker genes for classification, such as 16S rRNA in bacteria, phages are often grouped by gene content [40, 41]. Here, we generated a bipartite network that included the 85 bins of jumbo phages with a dataset of available *Caudovirales* complete genomes in RefSeq (3,012 genomes; downloaded July 5th, 2020) and the

full set of reference jumbo phages described above. To construct the bipartite network, we compared proteins encoded in all the phage genomes to the VOG database, and each genome was linked to VOG hits that were present (Figure 1, Supplemental Dataset 2, see Methods for details). To identify groups of phage genomes with similar VOG profiles, we employed a spinglass community detection algorithm [42] to generate genome clusters. Similar methods have recently been used to analyze evolutionary relationships in other dsDNA viruses [41]. The marine jumbo phages of this study clustered into five groups that included both jumbo and non-jumbo phage genomes (Figure 2a). We refer to these five clusters as Phage Genome Clusters (PGCs): PGC_A, PGC_B, PGC_C, PGC_D, and PGC_E. These PGCs included cultured phages and metagenome-derived jumbo phages found in various environments (i.e. aquatic, engineered) and isolated on a diversity of hosts (i.e. Firmicutes, Proteobacteria, Bacteroidetes) (Figure 2b,c). Of the marine jumbo phages, 135 belonged to PGC_A, 11 to PGC_B, 90 to PGC_C, 7 to PGC_D, and 1 to PGC_E (Figure 1b). In addition to this network-based analysis, we also examined phylogenies of the major capsid protein (MCP) and the terminase large subunit (TerL) encoded by the marine jumbo phages and the same reference phage set examined in the network (Figure 1c, 1d). With the exception of PGC_A, the marine jumbo phages that belong to the same PGC appeared more closely related to each other than those belonging to different clusters. The polyphyletic placement of jumbo phages in these marker gene phylogenies is consistent with the view that genome gigantism evolved multiple times, independently within the *Caudovirales [6]*.

We then compared functional content encoded by the marine jumbo phages in the PGCs to identify functional differences that distinguish these groups. PGC_E was excluded from this analysis because this genome cluster contained only a single jumbo phage. Collectively, most genes of the marine jumbo phages could not be assigned a function (mean: 86.60%, std dev: 7.01%;

Supplemental Dataset 3), which is common with environmental viruses [43, 44]. Genes with known functions primarily belonged to functional categories related to viral replication machinery, such as information processing and virion structure (Figure 3a), and these genes drove the variation between the genome clusters of marine jumbo phages (Figure 3b). A recent comparative genomic analysis of cultivated jumbo phages was able to identify three types of jumbo phages that are defined by different infection strategies and host interactions (referred to as Groups 1-3) [6]. We cross-referenced our PGCs and found that PGCs B, C, and D of this study corresponded to Groups 1, 2, and 3, respectively, suggesting that these genome clusters contain phages with distinct infection and replication strategies. PGC_A corresponded to multiple groups, indicating that this genome cluster contains a particularly broad diversity of phages.

The second largest phage cluster with marine jumbo phages, PGC_B, consists of 238 phages (11 (4.6%) marine jumbo phages, including 10 bins generated here), and included cultured phages of Group 1, which is typified by *Pseudomonas aeruginosa* phage PhiKZ. Supporting this correspondence with Group 1, all marine jumbo phages of PGC_B encoded the same distinct replication and transcription machinery, including a divergent family B DNA polymerase and a multi-subunit RNA polymerase (Figure 3b, Supplemental Dataset 3). These marine jumbo phages also encoded a PhiKZ internal head protein, and they uniquely encoded shell and tubulin homologs which has recently been found in PhiKZ phages to assist in the formation of a nucleus-like compartment during infection that protects the replicating phage from host defenses [18, 19]. Although we could not confidently predict hosts for the 11 metagenomic marine jumbo phages in this PGC_B (Supplemental Dataset 1), the cultured phages of this genome cluster infect pathogenic bacteria belonging to the phyla Proteobacteria (178 phages) and Firmicutes (6 phages) (Figure 2c), implicating a potential host range for marine jumbo phages in PGC_B.

The next largest phage genome cluster, PGC_C, comprised of 156 phages total (90 marine jumbo phages (57.7%); 4 bins generated from this study) and included reference jumbo phages in Group 2 (31, 19.9%) which are typified by Alphaproteobacteria and Cyanobacteria phages. Likewise, the host range of other cultured phages in PGC_C support the Group 2 correspondence, either infecting Cyanobacteria (139 phages) or Proteobacteria (4 phages) (Figure 2c). Furthermore, all 3 marine metagenomic phages in PGC_C for which hosts could be predicted were matched to Cyanobacteria hosts (Supplemental Dataset 1). Functional annotations of PGC_C marine jumbo phages revealed nearly all encode a family B DNA polymerase (97.8% phages) and the photosystem II D2 protein (PF00124, VOG04549) characteristic of cyanophages (90% phages) (Figure 3b). This PGC included the reference *Prochlorococcus* phage P-TIM68 (NC_028955.1), which encodes components of both photosystem I and II as a mechanism to enhance cyclic electron flow during infection [45]. This suggests that an enhanced complement of genes used to manipulate host physiology during infection may be a driver of large genome sizes in this group. Additionally, most of the PGC_C marine jumbo phages encoded lipopolysaccharide (LPS) biosynthesis proteins (76%), which have been found in cyanophage genomes that may induce a "pseudolysogeny" state, when infected host cells are dormant, by changing the surface of the host cell and preventing additional phage infections [6] (Supplemental Dataset 3). Taken together, most marine jumbo phages of PGC_C likely follow host interactions of jumbo cyanophages, such as potentially manipulating host metabolism by encoding their own photosynthetic genes and potentially inducing a pseudolysogenic state.

Finally, phages of PGC_D totaled at 47 phages, of which 7 were marine jumbo phages generated in this study (14.9%). This group included Group 3 jumbo phages (15, 31.9%), which is primarily distinguished by encoding a T7-type DNA polymerase but is not typified by a particular phage

type or host range. Supporting this grouping, all marine jumbo phages in this study encoded a T7 DNA polymerase which belongs to family A DNA polymerases (Figure 3b, Supplemental Dataset 3). Most of the other genes distinctively encoded by the marine jumbo phages in this group included structural genes related to T7 (T7 baseplate, T7 capsid proteins), a eukaryotic DNA topoisomerase I catalytic core (PF01028), and DNA structural modification genes (MmcB-like DNA repair protein, DNA gyrase B). Hosts of metagenomic marine jumbo phages in PGC_D could not be predicted (Supplemental Dataset 1); however, cultured Group 3 jumbo phages in PGC_D all infect Proteobacteria, primarily Enterobacteria and other pathogens.

The largest of the phage genome clusters, PGC_A, contained 469 phages, including 135 marine jumbo phages (63 bins from this study). This genome cluster contained the largest jumbo phages, such as *Bacillus* phage G (498 kb) and the marine megaphage Mar_Mega_1 (656 kb) recently recovered from the English Channel [39]. Unlike other PGCs, PGC_A contained mostly metagenomic phages (401, 85%, Figure 2b,c). Considering PGC_A contains the largest jumbo phages (Figure 1b, 2a), the vast genetic diversity in this PGC might explain why few genes were found to distinguish this group. Of the genes unique to PGC_A, only one was present in at least half of the phages (51.9%), which was a Bacterial DNA polymerase III alpha NTPase domain (PF07733). The host ranges of cultured phages from this PGC further reflect the large diversity of this group and included a variety of phyla and genera that can perform complex metabolisms or lifestyles, such as the nitrogen-fixing Cyanobacteria of the *Nodularia* genus isolated from the Baltic Sea (accessions NC_048756.1 and NC_048757.1) and the Bacteroidetes bacteria *Rhodothermus* isolated from a hot spring in Iceland (NC_004735.1) [46]. Because this group contains an abundance of metagenome-derived genomes that encode mostly proteins with no VOG annotation (Supplementary Dataset 2), it is possible that it includes several distinct lineages that

could not be distinguished using the community detection algorithm of the bipartite network analysis.

Relative abundance of jumbo bacteriophages across size fractions

To explore the distribution of the marine jumbo phages in the ocean, we first examined the size fractions in which the jumbo phages were most prevalent. To remove redundancy, we examined the 244 jumbo phages at the population-level (>80% genes with >95% average nucleotide identity [24]), corresponding to 142 populations (11 unique to this study, corresponding to 47 bins). We then mapped Tara Oceans metagenomes onto the 142 jumbo phage populations, and 102 of these populations could be detected [min. 20% of genome covered], resulting in 74 populations in PGC_A, 2 in PGC_B, 22 in PGC_C, 3 in PGC_D, and 1 in PGC_E. Out of the 225 Tara Oceans metagenomes examined, 213 (94.6%) contained at least one jumbo phage population (median: 7, Supplemental Dataset 4). Jumbo phages were more frequently detected in samples below 0.22 µm (<-0.22 µm, 0.1-0.22 µm) than those above 0.22 µm (0.45-0.8 µm, 0.22-0.45 µm, 0.22-1.6 µm, 0.22-3 µm) (Figure 4A). All samples in the <-0.22 µm fraction and the 0.1-0.22 µm fraction had at least one jumbo phage present, while the larger fractions ranged from 89% to 97%. Interestingly, we detected 34 populations (33.3%) exclusive to samples above 0.22 µm, compared to only one population (0.98%) exclusive to samples below 0.22 µm. A similar disparity in virus detection between size fractions has been reported for large eukaryotic viruses, where roughly 41% of phylotypes were present in the 0.22-3 µm size fraction but absent in fractions below 0.22 µm [9]. In contrast to this study, where certain viral groups were more prevalent in larger size fractions than smaller, a jumbo phage's PGC membership or genome size generally did not affect its

probability of detection at different size fractions (Supplemental Figure 3, Supplemental Figure 4).

We also compared jumbo phage diversity (defined as population richness), relative abundance (calculated in reads per kilobase per million (RPKM)), and community composition between the size fractions (based on Bray-Curtis distance matrices). Collectively, samples of the size fractions below 0.22 µm were significantly more diverse (p value <0.0001, Wilcox test) and had significantly higher relative abundances (p value < 0.0001, Wilcox test) of jumbo phages relative to the size fractions above 0.22 µm. Despite these differences in diversity and relative abundances, jumbo phage community composition did not significantly differ between the >0.22 and <0.22 µm size fractions when comparing samples based on presence/absence data (p value = 0.1082, ANOSIM, presence/absence Bray-Curtis distance matrix, Figure 4d), but did differ when using relative abundance data (p value = 0.0001, ANOSIM, RPKM Bray-Curtis distance matrix, Figure 4e).

To directly test the effect of the 0.22 µm size fraction cut-off on jumbo phage recovery, we examined a subset of the samples that were co-collected at the same station and depth for the fractions below 0.22 µm (<-0.22 or 0.1-022) and above 0.22 µm (0.22-1.6 µm or 0.22-3 µm). The number of detected jumbo phage populations was significantly higher in samples below 0.22 µm than above 0.22 µm (p value = 0.000138, Wilcox test, Supplemental Figure 5a). The relative abundance of jumbo phages was also significantly higher in size fractions below 0.22 µm than above 0.22 µm (p value = 0.00001, Wilcox test, Supplemental Figure 5b). Likewise, community composition significantly differed between samples above and below 0.22 µm (p value = 0.0001, ANOSIM, presence/absence Bray-Curtis distance matrix, Supplemental Figure 5c,d). Taken together, these findings suggest that using size fractions below 0.22 µm to analyze phages

enhances the signal of jumbo phage sequences, relative to samples of larger size fractions, likely due to cellular sequences present in the larger sizes. Notwithstanding, roughly 33% of jumbo phages in this study were exclusive to size fractions above 0.22 μm, indicating that analyzing a range of size fractions is necessary for a more synoptic view of jumbo phages in the environment.

Biogeography of jumbo bacteriophages in the global ocean

Jumbo phage populations varied by depth in different PGCs. Jumbo populations in this study varied in their distribution (Supplemental Figure 6) but were collectively found in all three depths (Surface (SRF), Deep Chlorophyll Maximum (DCM), Mesopelagic (MES)) examined (Figure 5, Supplemental Figure 7; Supplemental Dataset 4). Although jumbo phages were more prevalent in samples of the viral size fractions (<-0.22 or 0.1-0.22 μm), we focused biogeographic analyses on the 0.22-1.6 or 0.22-3 μm size fractions because the most sites were available for these samples enabling comparisons between depths and biomes. For instance, only four MES samples and no Westerlies samples were available in the viral fractions. When applicable, analyses were also completed with viral fraction samples, and results are deposited at the end of the Supplement (Supplemental Figures 16 - 24). Jumbo phage communities differed significantly between depths (p value = 0.0001, ANOSIM based on presence/absence and RPKM Bray-Curtis distance matrices, Supplemental Figure 8), consistent with the dramatic transition in community composition that occurs from surface waters to below the deep chlorophyll maximum [38, 47]. Specifically, the diversity of jumbo phages across depths varied by genome cluster (Figure 5d, Supplemental Figure 10), with PGC_A and PGC_C exhibiting higher prevalence in the epipelagic (SRF and DCM). Although PGC_B and PGC_D had too few populations detected to generalize for these clusters (2 and 3, respectively), our results for these phages showed that PGC_B and PGC_D were typically

less prevalent in SRF samples compared to DCM and MES samples. PGC_C is typified by cyanophages, providing a clear reason why this phage group is enriched in surface waters. Conversely, PGC_B is typified by *Pseudomonas aeruginosa* PhiKZ phages, suggesting these PGC_B marine jumbo phages may be infecting heterotrophic bacteria that are potentially less prevalent in surface waters. Overall relative abundance and diversity of jumbo phages in this study were significantly higher in the epipelagic zone (Supplemental Figure 8), partly because most of these phages are in PGC_A. These collective and PGC-specific patterns held when examining only those samples that were co-collected at all three depths (Figure 5c, Supplemental Figures 8, 9, 10). This general pattern therefore reflects what has been found in previous studies on the depth distribution of viruses and viral protein clusters, where more were unique in the euphotic (i.e. epipelagic) than aphotic depths [9, 26, 48, 49], although this contrasts what has recently been found in the Pacific Ocean, where overall viral diversity increased in the mesopelagic [38].

Jumbo phage biogeography across biomes. Collectively, jumbo phages could be found in all three Longhurst biomes (Coastal, Westerlies, Trades) (Supplemental Figure 11), and jumbo phage communities in this study significantly differed in composition between the biomes (p value < 0.05, ANOSIM, presence/absence Bray-Curtis distance matrix, Supplemental Figure 11c,). However, no biome appeared to be a particular hotspot for jumbo phages, as they were not significantly more diverse in any biome (Supplemental Figure 11a). When looking at depths separately, the relative abundance of jumbo phages in SRF samples was significantly higher in Coastal samples (Supplemental Figure 12b), but no clear biome was a hotspot for phages in DCM and MES samples (Supplemental Figure 11). Likewise, jumbo phage community composition only differed between samples in the SRF and DCM, but not in MES samples (p value < 0.05, ANOSIM, presence/absence Bray-Curtis distance matrices). Upon examining jumbo phages by

group, jumbo phages from PGCs A, C, and D could be detected in all biomes, while PGB_B phages could not be detected in Westerlies samples (Supplemental Figure 13). Similar to their collective results, no PGC was enriched in a single biome (Supplemental Figure 13). A recent global study on marine viruses has found that viral diversity is better explained by ecological zones defined by physicochemical factors like temperature, rather than by Longhurst biomes defined by patterns of chlorophyll *a* concentrations [26], suggesting that Longhurst biomes may not be good predictors of viral diversity in general.

Jumbo phage populations ranged in endemicity. Fifteen populations (14.7%) were detected in only one Tara station, and six (5.6%) were present in over half of the stations (>=34) (Figure 5c). Both the more endemic populations and the more prevalent jumbo phages belonged to PGCs A and C. PGC_A and PGC_C contained the most populations, which likely explains the wide range of endemicity of phages in these clusters (Supplemental Figures 14, 15). Moreover, the cyanobacterial hosts that are known for many of the jumbo phages in PGC_C are widespread in the ocean, which may also explain the prevalence of this group of phages. In general, the heterogeneous distribution and abundance of these jumbo phages is consistent with the seed bank hypothesis, which postulates that viruses are passively dispersed throughout the ocean and viral community structure is shaped by local selective forces [24, 50]. This framework has previously been used to explain why phage distributions range from extremely cosmopolitan to extremely rare, which is a pattern that also appears to hold for jumbo bacteriophages.

Conclusion

Large DNA viruses are becoming increasingly recognized as critical components of the virosphere, notable for their intriguing evolutionary histories [51, 52], vast functional capacities [3, 4], and global distribution [2, 4]. Here, we assess the diversity and ecology of marine jumbo

bacteriophages, which have historically been difficult to study due to biases in filtration and isolation strategies. We employed a binning strategy to generate and quality-check genomes of jumbo phages and used it to identify 85 high quality bins. We employed a conservative approach to genome binning because binning has traditionally not been used for bacteriophages, and as a result these bins likely represent a small fraction of total jumbo phages in these marine samples. We combined these bins together with reference jumbo bacteriophage genomes, and ultimately identified 102 populations that are present in Tara Oceans metagenomes. When compared with other metagenomic jumbo phages and cultured phages of all sizes, we found that marine jumbo phages primarily belong to four phage genome clusters (PGCs) that largely encode distinct replication machinery, biogeography, and potential hosts. For example, marine jumbo phages in PGC_C follow cyanophage infection strategies and ecology, as this cluster included cultured marine cyanophages and encoded classic family B DNA polymerases and photosynthesis enzymes characteristic of cyanophages. Furthermore, we found they are enriched in surface waters relative to the mesopelagic, consistent with the geographic range of their hosts. In contrast, marine jumbo phages of PGC_B included cultured PhiKZ phages of *Pseudomonas aeruginosa* and uniquely encoded multi-subunit RNA polymerases and tubulin, which are thought to play a role in the remarkable nucleus-like structures that these viruses employ as an anti-CRIPSR defense [18, 19]. PGC_B was more often found in mesopelagic waters, suggesting that this complex infection strategy may be more common in the deep ocean. PGC_A contained a large number of metagenome-derived viruses and was not as well-defined as the other clusters; it is possible that this cluster contains several distinct lineages, and more in-depth analyses will be required to assess. Overall, these results suggest that jumbo phages exhibit diverse biology and ecology, consistent

with the view that they are an incredibly diverse set of phages with unique evolutionary histories [6].

The jumbo phages we analyze are collectively widespread throughout the ocean and are typically more diverse and abundant in epipelagic waters, which reflect previous findings that surface waters usually harbor a higher per-sample alpha diversity of viral groups compared to deeper waters [9, 26]. Larger phages therefore appear to coexist in patterns broadly similar to smaller viruses despite the disadvantages of their size, such as smaller burst sizes and lower host contact rates [53]. In eukaryotic giant viruses, it has been hypothesized that these disadvantages are potentially offset by higher infection efficiency, broader host ranges, decreased decay rates, and higher rates of successful attachments compared to smaller viruses [53]. Although some of these advantages to viral gigantism may also apply to jumbo bacteriophages, it is unlikely that they are all applicable. For example, given that they are tailed *Caudovirales* [6], jumbo phages likely possess higher host specificity in part due to their non-phagocytotic mode of infection. Nonetheless, the large genomes of jumbo bacteriophage often encode an expanded complement of genes used to manipulate host physiology during infection, and these may play critical roles in promoting infection efficiency or offsetting host defense mechanisms. The impressive complement of photosynthesis genes in PGC_C is at least partially responsible for the large genomes in this lineage, while the genes involved in anti-CRISPR defense found in PGC_B indicate that a host-virus arms race may be responsible for genome gigantism in this group. Interestingly, the largest number of jumbo phage genomes we identified belong to PGC_A, which is largely uncharacterized and composed of primarily metagenome-derived genomes, suggesting that these viruses have as-yet unidentified infection strategies. Overall, it is likely that the factors leading to and maintaining genome gigantism in each of these genome clusters are distinct. Future work further characterizing the

hosts of these jumbo phages and the details of their infection programs, particularly in PGC_A, will therefore be critical to understanding mechanisms that underlie complexity in the virosphere and maintain diversity. Moreover, future in-depth examination of the genomics and evolutionary histories of jumbo phages will be an important step to integrating these viruses into a meaningful taxonomy and clarifying their evolutionary relationships to other *Caudovirales*.

Methods

Jumbo phage binning and detection. An overview of the pipeline can be found in Supplemental Figure 25. Metagenomic scaffolds were downloaded from 1,545 assemblies by Parks et al. 2017 [31] and binned with MetaBAT2 [32] (-s 200000 --unbinned -t 32 -m 5000 --minS 75 --maxEdges 75) using the coverage files provided by Parks et al. 2017. Bins were retained if they summed to at least 200,000 base-pairs and comprised <= 5 contigs (min. contig size 5kb). Proteins were predicted with Prodigal [54] using default settings on each bin individually. Bins were retained if they lacked more than one ribosomal protein, lacked overlapping regions (via promer and gnuplot [55] with MUMmer 3.0 [56]), had fewer hits to NCLDV than phage (via LASTp [57] against RefSeq r99), lacked more than one NCLDV marker gene (via hmmsearch (hmmer.org) against NCLDV marker gene HMM profiles [58]) and had even read coverage of Tara Ocean metagenomes via coverM [59] (https://github.com/wwood/CoverM). Briefly, "even read coverage" means that the read coverage of each contig in the bin varied between one another below a variation threshold determined by reference mapping results (See Supplemental Methods for details). Jumbo phages were then detected by running the bins through VirSorter2 [33], VIBRANT [34], and CheckV [35]. Bins were considered putative phages if they had at least an average dsDNAphage score of >0.9 from VirSorter2 or a VirSorter2 average score >0.5 and either i) CheckV quality of medium or higher or ii) VIBRANT consensus classification as viral. Ultimately,

85 bins were retained for downstream analyses. Prior to further gene-based analyses, we checked if the jumbo bins used alternative genetic codes with Codetta [60], and all were found to use the standard bacteria code 11; we therefore proceeded with the initial Prodigal predictions. Bins were grouped into populations based on single-linkages with a compiled set of 898 jumbo bacteriophages (RefSeq phages over 200 kilobases (93), phage sequences over 200 kilobases from the INPHARED database (354) [14], Iyer et al 2021 (46, non-overlapping with INPHARED) [6], Al-Shayeb et al. 2020 [4] jumbo phage genomes (336), GOV 2.0 (60) [26], ALOHA 2.0 (8) [38], and a megaphage assembled from the English Channel [39]) based on nucleotide sequences of genes (predicted with prodigal -d flag) aligned with BLASTn [61] (>95% average nucleotide identity, >80% genes) [24]. See Supplemental Methods for details.

Bipartite network and phylogenetic analyses. Because phages lack universal, high resolution phylogenetic marker genes, gene-sharing networks have typically been used to classify phages [41, 62]. Bipartite networks are commonly used to examine evolutionary relationships in divergent viral lineages [3, 41]. To classify the jumbo bins of this study with a bipartite network, reference phage sequences were compiled from RefSeq's *Caudovirales* complete genomes (downloaded July 2020 from NCBI's Virus genome portal; 3,012 genomes) along with the curated jumbo phage set used in the population analysis. Proteins of jumbo bins and this reference set were predicted with Prodigal and searched against the Virus Orthologous Groups (VOGs, vogdb.org) via HMM searches (E value 0.001). A bipartite network was made based on shared VOGs using igraph (graph.incidence) (1.2.5) [63] in R (version 3.5.1) [64] with RStudio (version 1.1.456) [65]. A previous study classified divergent viral lineages via the spinglass community detection algorithm [42], which we used on the bipartite network generated here via igraph (50 spins for 100 iterations).

Final clusters were determined by using those of the iteration with the highest modularity. Plots of the cluster composition from the bipartite network analysis were made with ggplot2 (3.1.1) [66] in R with Rstudio. TerL (terminase large subunit) and MCP (major capsid protein) trees were made with hits to TerL VOG families and MCP VOG families encoded by the jumbo phages and reference hits (hmmsearch, E value < 0.001; See Data Availability). Reference hits were de-replicated with CD-HIT (version 4.8.1 -c 0.9) [67] and filtered for size (See Supplemental Methods). Proteins were then aligned with Clustal Omega [68], trimmed with trimAl (-gt 0.1) [69] and constructed with IQ-TREE [70] (TEST model selection with ModelFinder [71]).

Size fraction and ecological analyses. Metagenomic reads from Tara Oceans were trimmed with trim_galore (--paired --length 50 -e 0.1 -q 50) and subsampled to an even depth of 20 million reads per sample with seqkit sample (-s 1000 -n 20000000 -2). These reads were then mapped onto the population representatives of the jumbo phage set (535 populations). For the mapping, the reference database of the representative jumbo phage sequences was created with minimap2 (minimap2 -x sr -d)[72], and the mapping was carried out with coverM (coverm genome --min-read-percent-identity 95 -m covered_fraction rpkm count variance length -t 32 --minimap2-reference-is-index --coupled)). Mapping results were retained if at least 20% of the phage genome was covered (see Supplemental Methods for benchmarking, Supplemental Figure 26). Relative abundance of a phage in each sample was calculated in reads per kilobase per million (RPKM). Statistical analyses and plots were carried out in R with vegan (2.5-5) [73], ggplot2, maps (3.3.0) [74], and ggpubr (0.2.4) [75] packages. Community composition was compared between variables using ANOSIMs based on Bray-Curtis distances using both presence/absence and RPKM matrices

with a significance p values < 0.05. Statistical tests were carried out with the ggplot2 function stat_compare_means(label="p.signif").

Annotation. Amino acid sequences of genes were annotated with HMM searches (E value < 0.001) against the Pfam [76] (version 32), eggNOG (5.0) [77], and VOG (release 98) databases. Virion structural protein families were identified based on VOG hit descriptions (Supplemental Dataset 3). Consensus annotation was based on Pfam annotations and then the highest bit score between eggNOG and VOG hits. Functions were grouped into larger categories (Supplemental Dataset 3). Clusters from the network analyses were compared for functional composition between marine jumbo phages by averaging the proportion of genes in a functional category (Figure 3a). Genes with the highest variance between PGCs were identified based on the variance in the proportion of genomes in a PGC with that gene. Those with a variance >0.2 and a known function were visualized with pheatmap [78] in R (Figure 3b).

Host prediction. Hosts of the jumbo phage bins were predicted based on matching CRISPR spacers, tRNAs, and gene content. CRISPR spacers were predicted on the Genome Taxonomy Database (release 95) [79], metagenome assembled genomes (MAGs) of bacteria and archaea from the metagenomes that the jumbo phage bins derived (provided by Parks et al. 2017 [31]), and on the jumbo phage bins with minCED (derived from reference [80]). Spacers were aligned with BLASTn (-task blastn-short) and matches were >24 bp with <= 1 mismatches [4]. Only one jumbo phage contained a CRISPR array, but the spacers did not match any other jumbo phages or MAGs. tRNA sequences were predicted with tRNAscan-SE (-bacteria option) [81] on the MAGs and jumbo phage bins. Promiscuous tRNAs [82] were removed (BLASTn hits 100% ID, <= 1

mismatches). Jumbo phage tRNAs were aligned against the MAGs tRNAs and NCBI nr database (BLASTn 100% ID, <= 1 mismatches) [4]. Lastly, hosts were assigned based on the taxonomy of coding sequence matches to the MAGs (BLASTn). Hits to phyla were summed and a putative host phylum had three times the number of hits as the phylum with the next most hits as used in a previous study [4]. If a putative host could be predicted by multiple methods, a consensus host was assigned if all approaches agreed on a phylum. If the methods disagreed at the phylum-level, no putative host was assigned.

Data Availability. Nucleic acid sequences and protein predictions for the 85 bins analyzed in this study and the proteins and files for the phylogenetic analyses (proteins, HMM profiles, treefiles) can be found on FigShare (https://figshare.com/projects/Marine_jumbo_phages/127391).

Competing Interests. The authors declare no competing financial interests.

Acknowledgements

We thank members of the Aylward Lab for helpful feedback. We thank Anh Ha for assistance with read mapping. This work was performed using compute nodes available at the Virginia Tech Advanced Research and Computing Center.

Author Contributions. A.R.W. and F.O.A. designed the experiment. A.R.W. performed the research. A.R.W. and F.O.A. wrote the manuscript.

References

1. Brüssow H, Hendrix RW. Phage Genomics. Cell. 2002;108:13–16.
2. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, et al. Giant virus

diversity and host interactions through global metagenomics. Nature. 2020;578:432–436.

3. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nat Commun. 2020;11:1710.
4. 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.
5. Koonin EV, Yutin N. Origin and Evolution of Eukaryotic Large Nucleo-Cytoplasmic DNA Viruses. Intervirology . 2010;53:284–292.
6. Iyer ML, Anantharaman V, Krishnan A, Burroughs AM, Aravind L. Jumbo Phages: A Comparative Genomic Overview of Core Functions and Adaptions for Biological Conflicts. Viruses. 2021;13.1:63.
7. Raoult D, Forterre P. Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol. 2008;6:315–319.
8. Ha AD, Moniruzzaman M, Aylward FO. High Transcriptional Activity and Diverse Functional Repertoires of Hundreds of Giant Viruses in a Coastal Marine System. mSystems. 2021;6:e00293-21.
9. Endo H, Blanc-Mathieu R, Li Y, Salazar G, Henry N, Labadie K, et al. Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. Nat Ecol Evol, 2020;4:1639–1649.
10. Kaneko H, Blanc-Mathieu R, Endo H, Chaffron S, Delmont TO, Gaia M, et al. Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean. iScience. 2021;24:102002.
11. Hendrix RW. Jumbo bacteriophages. Curr Top Microbiol Immunol. 2009;328:229–240.

12. 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. Nat Microbiol. 2019;4:1306–1315.
13. Krylov VN, Zhazykov IZ. Pseudomonas bacteriophage phiKZ--possible model for studying the genetic control of morphogenesis. Genetika. 1978;14:678–685.
14. Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Stekel DJ, et al. INfrastructure for a PHAge REference Database: Identification of large-scale biases in the current collection of phage genomes. PHAGE. 2021;2.4:214-223.
15. Yuan Y, Gao M. Jumbo Bacteriophages: An Overview. Frontiers in Microbiology. 2017;8:403.
16. Nazir, A, Ali A, Qing H, & Tong Y. Emerging aspects of jumbo bacteriophages. Infect Drug Resist. 2021;14:5041.
17. Ceyssens P-J, Minakhin L, Van den Bossche A, Yakunina M, Klimuk E, Blasdel B, et al. Development of Giant Bacteriophage KZ Is Independent of the Host Transcription Apparatus. J Virol. 2014;88:10501–10510.
18. Mendoza SD, Nieweglowska ES, Govindarajan S, Leon LM, Berry JD, Tiwari A, et al. A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. Nature. 2020;577:244–248.
19. Malone LM, Warring SL, Jackson SA, Warnecke C, Gardner PP, Gumy LF, et al. A jumbo phage that forms a nucleus-like structure evades CRISPR–Cas DNA targeting but is vulnerable to type III RNA-based immunity. Nat Microbiol. 2020;5:48–55.
20. Serwer P, Hayes SJ, Thomas JA, Hardies SC. Propagating the missing bacteriophages: a large bacteriophage in a new class. Virol J 2007;4:21.

21. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. Nat Protoc. 2009;4:470–483.
22. Palermo CN, Shea DW, Short SM. Analysis of different size fractions provides a more complete perspective of viral diversity in a freshwater embayment. Appl Environ Microbiol. 2021;87.11:e00197-21.
23. Sieburth JM, Smetacek V, Lenz J. Pelagic ecosystem structure: heterotrophic compartments of the plankton and their relationship to plankton size fractions 1. Limnol and Oceanogr. 1978;23:1256–1263.
24. 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.
25. 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.
26. 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.
27. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. Benchmarking viromics: an evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ. 2017;5:e3817.
28. García-López R, Vázquez-Castellanos JF, Moya A. Fragmentation and Coverage Variation in Viral Metagenome Assemblies, and Their Effect in Diversity Calculations. Front Bioeng Biotechnol. 2015;3:141.

29. Sunagawa S, Pedro Coelho L, Chaffron S, Roat Kultima J, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348.6237:1261359.
30. Bäckström D, Yutin N, Jørgensen SL, Dharamshi J, Homa F, Zaremba-Niedwiedzka K, et al. Virus genomes from deep sea sediments expand the ocean megavirome and support independent origins of viral gigantism. MBio. 2019;10.
31. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533–1542.
32. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7:e7359.
33. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome. 2021;9.1:1-13.
34. 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;8:1-23.
35. 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.
36. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.
37. Adriaenssens EM, Cowan DA. Using signature genes as tools to assess environmental viral

ecology and diversity. Appl and Environ Microbiol. 2014;80.15:4470-80.

38. 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.
39. Michniewski S, Rihtman B, Cook R, Jones MA, Wilson WH, Scanlan DJ, et al. Identification of a new family of 'megaphages' that are abundant in the marine environment. ISME Comm. 2021;1.1:1-4.
40. 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. Nat Biotechnol. 2019;37:632-639.
41. Iranzo J, Krupovic M, Koonin EV. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. MBio. 2016;7.4:e00978-16.
42. Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E Stat Nonlin Soft Matter Phys 2006;74:016110.
43. 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.
44. Pappas N, Dutilh BE. Finding functional associations between prokaryotic virus orthologous groups: a proof of concept. BMC Bioinform. 2021;22:1–11.
45. Fridman S, Flores-Uribe J, Larom S, Alalouf O, Liran O, Yacoby I, et al. A myovirus encoding both photosystem I and II proteins enhances cyclic electron flow in infected Prochlorococcus cells. Nat Microbiol. 2017;2:1350–1357.
46. Hjorleifsdottir S, Aevarsson A, Hreggvidsson GO, Fridjonsson OH, Kristjansson JK.

Isolation, growth and genome of the Rhodothermus RM378 thermophilic bacteriophage. Extremophiles. 2014;18:261–270.

47. Mende DR, Bryant JA, Aylward FO, Eppley JM, Nielsen T, Karl DM, et al. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nat Microbiol. 2017;2:1367–1373.
48. 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.
49. 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.
50. Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 2005;13:278–284.
51. Brüssow H. Huge bacteriophages: bridging the gap? Environ Microbiol. 2020;22:1965–1970.
52. Koonin EV, Yutin N. Evolution of the large nucleocytoplasmic DNA viruses of eukaryotes and convergent origins of viral gigantism. Adv Virus Res. 2019;103:167–202.
53. Edwards KF, Steward GF, Schvarcz CR. Making sense of virus size and the tradeoffs shaping viral fitness. Ecol Lett. 2021;24:363–373.
54. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 2010;11.1:1-11.
55. Williams T, Kelley C. Gnuplot 5.2 Manual: an interactive plotting program. 2017.
56. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and

open software for comparing large genomes. Genome Biol. 2004;5.2:1-9.

57. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21:487–493.
58. Aylward FO, Moniruzzaman M. ViralRecall - a flexible command-line tool for the detection of giant virus signatures in 'omic data. Viruses. 2021;13.2:150.
59. wwood. GitHub - wwood/CoverM: Read coverage calculator for metagenomics. https://github.com/wwood/CoverM. Accessed 23 Jul 2021.
60. Shulgina Y, Eddy SR. A computational screen for alternative genetic codes in over 250,000 genomes. eLife 2021;10:e71402.
61. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–410.
62. Iranzo J, Koonin EV, Prangishvili D, Krupovic M. Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsidless mobile elements. J Virol. 2016;90.24:11043-11055.
63. igraph – Network analysis software. http://igraph.org. Accessed 23 Jul 2021.
64. R Core Team. R: A Language and Environment for Statistical Computing. 2019. R Foundation for Statistical Computing, Vienna, Austria.
65. RStudio. https://rstudio.com/. Accessed 12 Oct 2021.
66. Wickham H. ggplot2. Wiley Interdisciplinary Reviews: Comput Stat . 2011;3:180–185.
67. 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.
68. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol.

2011;7:539.

69. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–1973.
70. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–274.
71. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–589.
72. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–3100.
73. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003;14:927–930.
74. maps: Draw Geographical Maps. https://CRAN.R-project.org/package=maps. Accessed 22 Jul 2021.
75. Kassambara A. 'ggplot2' Based Publication Ready Plots [R package ggpubr version 0.4.0]. 2020.
76. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2020;49:D412–D419.
77. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47:D309–D314.
78. pheatmap: Pretty Heatmaps. https://CRAN.R-project.org/package=pheatmap. Accessed 22 Sep 2021.

79. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36:996–1004.
80. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinform. 2007;8.1:1-8.
81. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res 2016;44.W1:W54–W57.
82. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering earth's virome. Nature. 2016;536:425–430.

Figures

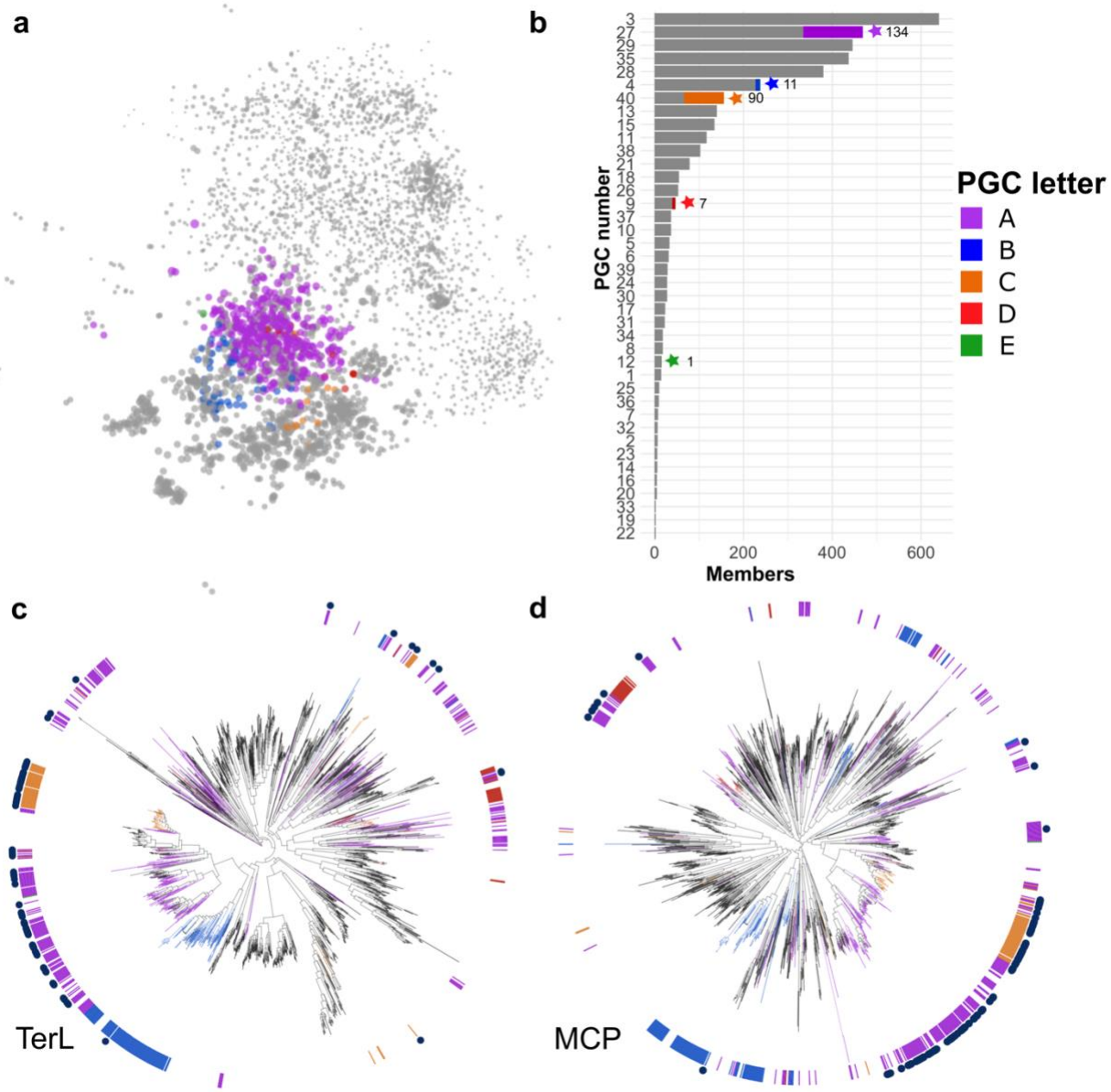


Figure 1. Comparison of bins to reference gene content and evolutionary histories. (a) Network with marine jumbos and references as nodes and edges based on shared VOGs. Marine jumbo phage nodes are colored by PGC as detected with spinglass community detection analysis, other nodes are in gray. Node size corresponds to the natural log of genome length in kilobases. **(b)** Barchart of the number of members in each PGC. PGCs with marine jumbo phages are denoted with a star and the number of marine jumbo phages in that PGC. Proportion of marine

jumbo phages in that PGC is colored. c,d) Phylogenies of TerL **(c)** and MCP **(d)** proteins with references and bins. Inner ring and branches are colored by the 5 PGCs that marine jumbo phages belong to. Navy blue circles in the outer ring denote marine jumbo phages.

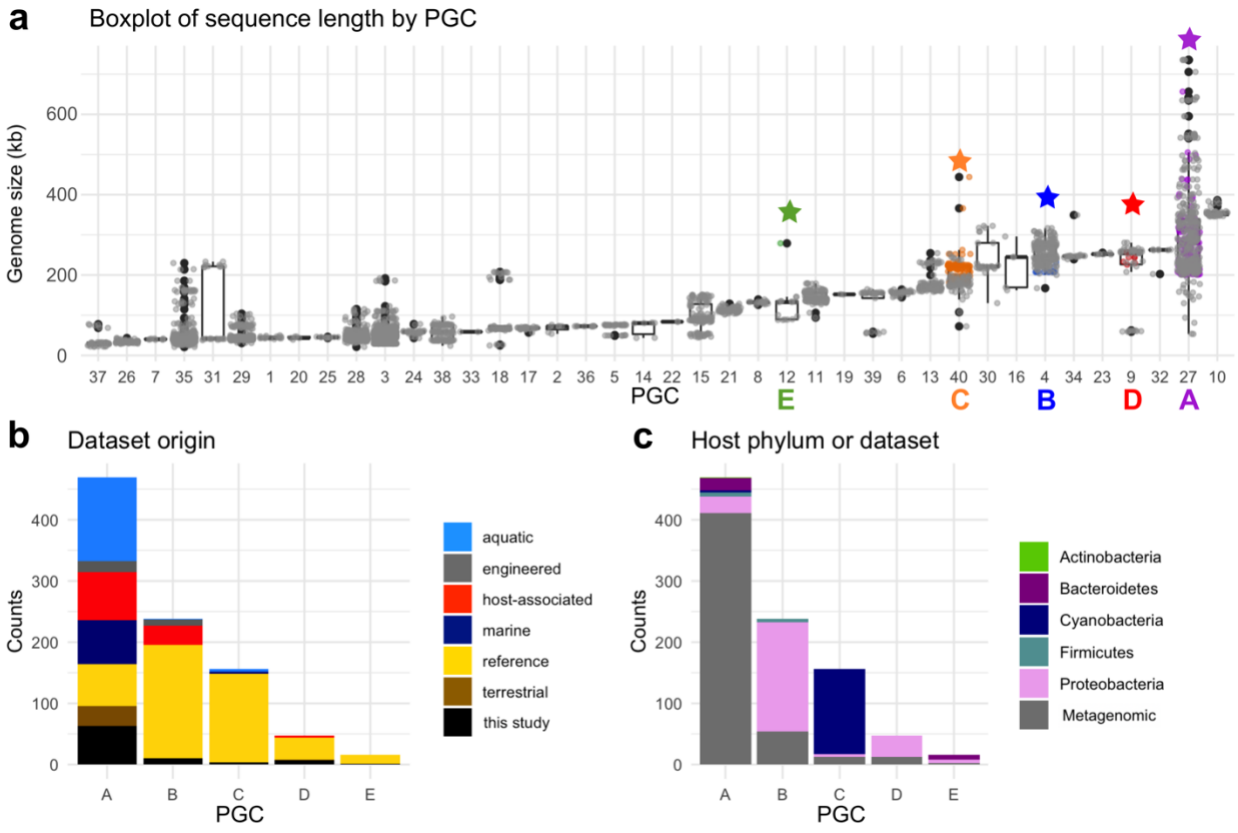


Figure 2. (a) Boxplot of genome length in each network cluster (x-axis is PGC number). Star denotes PGC with a marine jumbo phage and the color matches the PGC letters of Figure 1. **(b)** Stacked barplot of the metagenome environment from which each phage derives from in each PGC (x-axis). Reference (yellow) are cultured phages, in black are the bins of jumbo phages from this study **(c)** Stacked barplot of the host phylum of the RefSeq cultured phages in each cluster; metagenomic phages are in gray.

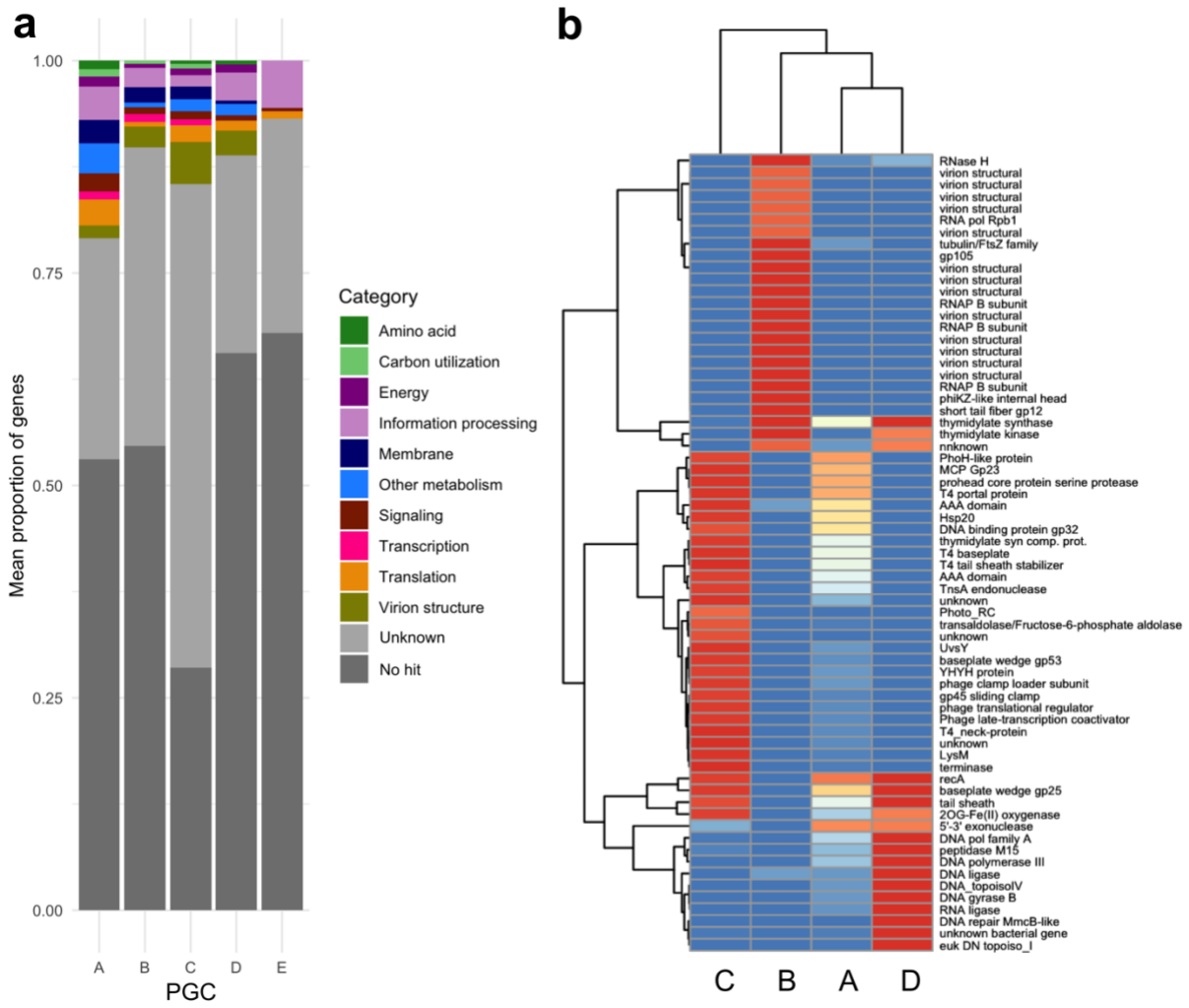


Figure 3 (a) Functional categories for genes encoded by jumbo phages averaged by PGC. **(b)** Heatmap of proportion of genomes in each PGC that contain the listed genes. Listed genes were selected based on containing a known function and variance between the PGCs above 0.2. Dendrogram was generated based on hierarchical clustering in pheatmap.

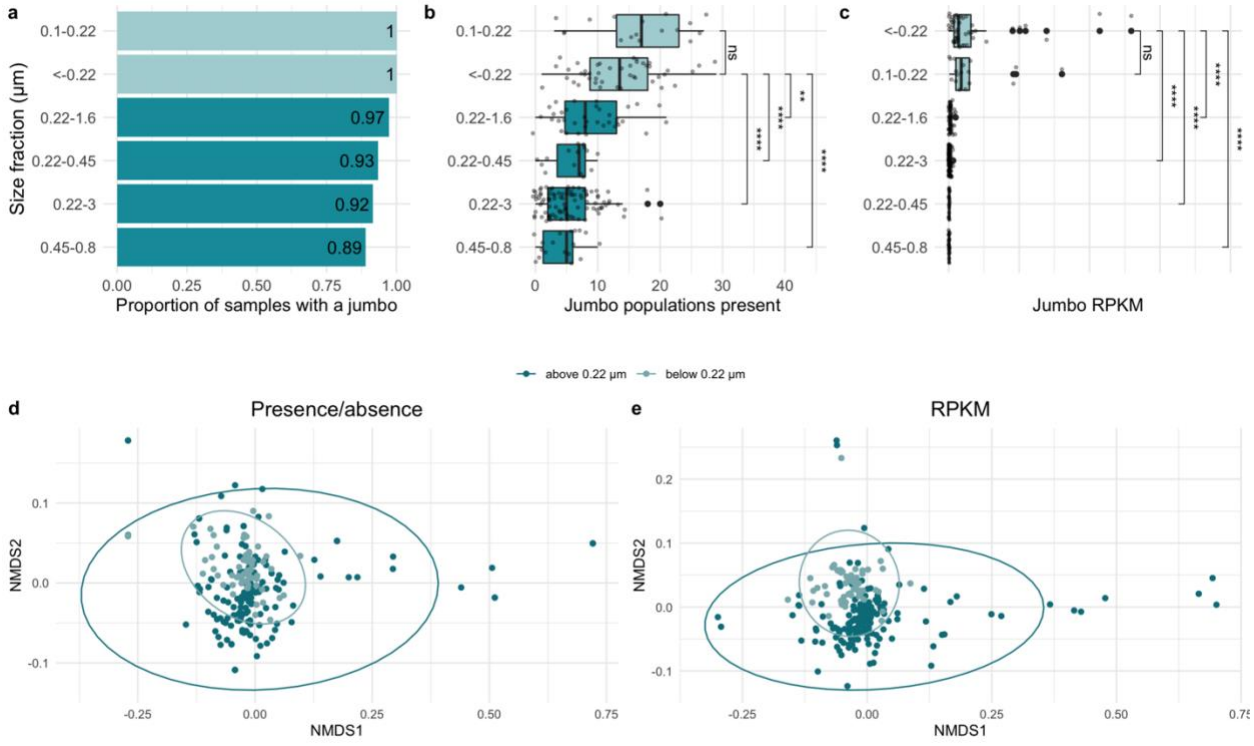


Figure 4. Comparison of jumbo abundance and presence in samples of different filter size fractions. Dark teal are fractions with minimum sizes of 0.22 µm or higher. Light teal are fractions with a maximum size of 0.22 µm or lower. **(a)** Barchart of the proportion of samples with at least one marine jumbo phage (x-axis) by size fraction (y-axis) sorted from highest to lowest. **(b)** Boxplot with x-axis as the number of marine jumbo phages found in a sample with size fraction on the y-axis sorted by median. **(c)** Boxplot with x-axis as the relative abundance of marine jumbo phages found in a sample (RPKM) with size fraction on the y-axis sorted by median. Significance bars in c,d correspond to Wilcox test, with stars corresponding to *p* values < 0.05 and those with *p* values > 0.05 as not significant "ns" (stat_compare_means function). **(d,e)** NMDS plots (Bray Curtis dissimilarity distances) of jumbo phage composition in each sample using presence absence data **(d)** and relative abundance data **(e)**. Samples are colored by size fraction distinction above 0.22 µm (dark teal) and below 0.22 µm (light teal). Ellipses calculated based on multivariate normal distribution.

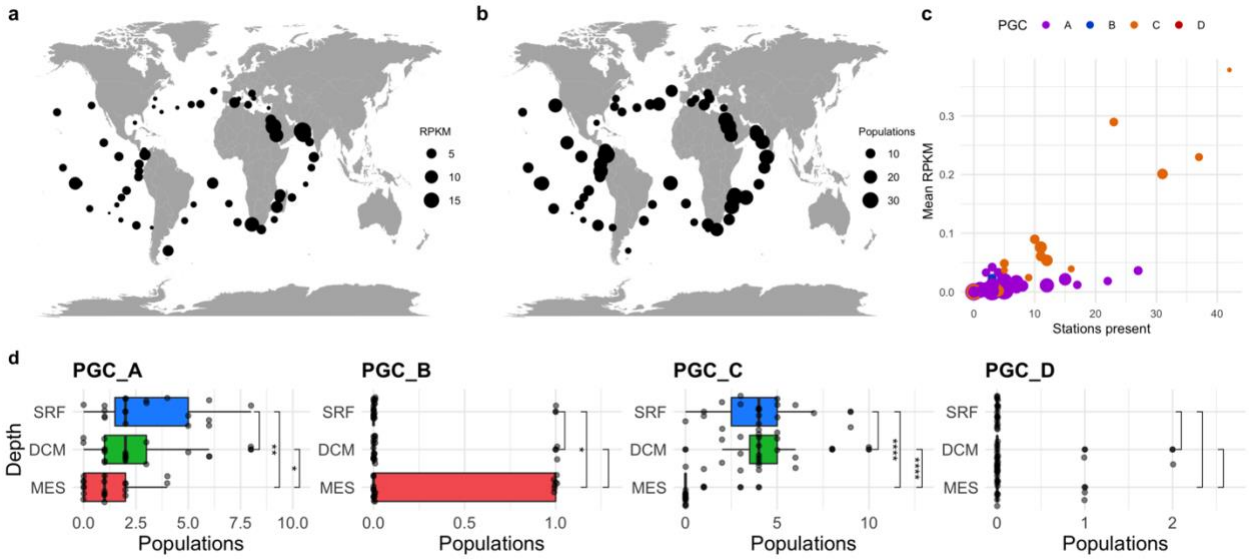


Figure 5. (a,b) Maps of the relative abundance **(a)** of total jumbo phages (in RPKM) and **(b)** total number of jumbo populations present regardless of phage cluster membership in each surface (SRF) sample of the picoplankton size fraction (either 0.22-3 µm or 0.22-1.6 µm depending on availability). Dots sizes are proportional to the number of populations or RPKM. **(c)** Scatterplot of the mean RPKM of a jumbo population in SRF picoplankton samples versus the number of SRF picoplankton stations it was present. Populations are colored by PGC and size corresponds to putative genome length in 100 kilobases. **(d)** Boxplot of the number of jumbo phage populations in samples co-collected at each depth sorted by mean for each PGC. Significance bars correspond to Wilcox test, with stars corresponding to p values < 0.05 (stat_compare_means function).

Supplemental Information

Supplemental Datasets can be found at the link:

https://figshare.com/projects/Marine_jumbo_phages/127391

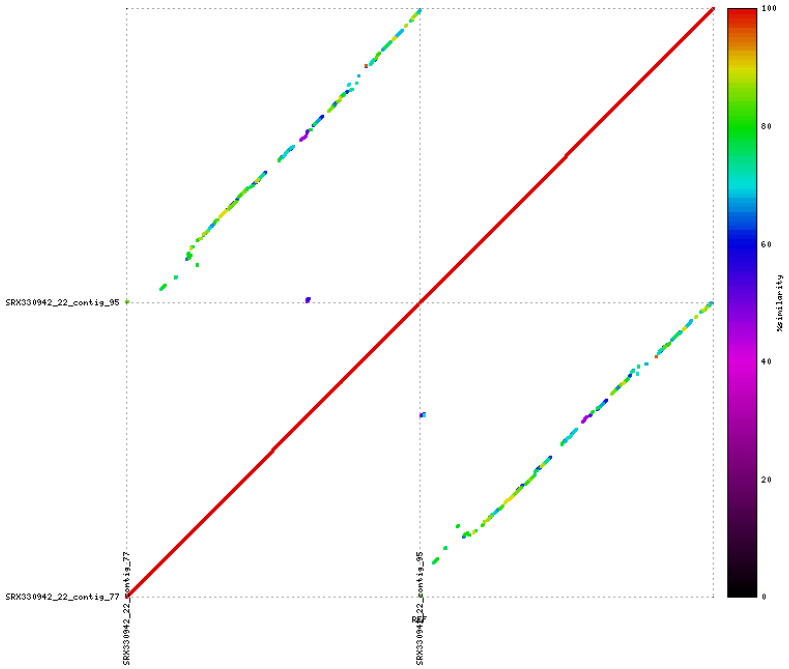
Supplemental Dataset 1: Excel file of marine jumbo phage genomic features (length, contigs, population representatives), host predictions, network cluster membership, Iyer group designation, jumbo phage detection results.

Supplemental Dataset 2: VOG matrix used to generate network and cluster membership of all sequences in network, table features of phage sequences (i.e. cluster membership) used to annotate network.

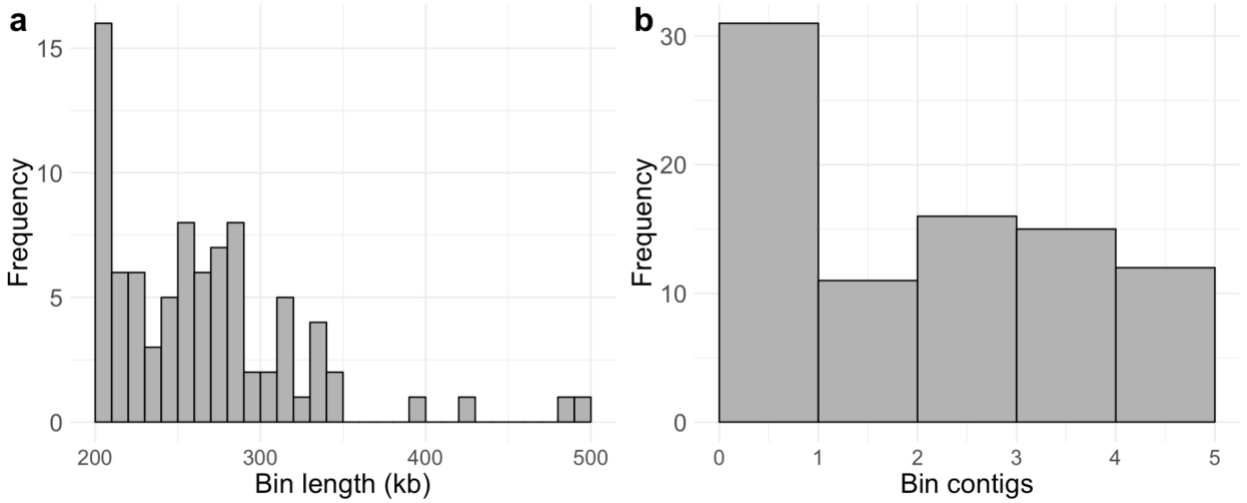
Supplemental Dataset 3: Protein annotation file; sheet 1 contains marine jumbo phage protein hits to EggNOG, VOG,and Pfam; sheet 2 contains category descriptions; sheet 3 specifies virion structure VOGs.

Supplemental Dataset 4: Read mapping results (counts, fraction covered, RPKM, presence/absence); sample metadata (i.e. longitude, latitude, biome); list of genomes used for benchmarking genome coverage threshold.

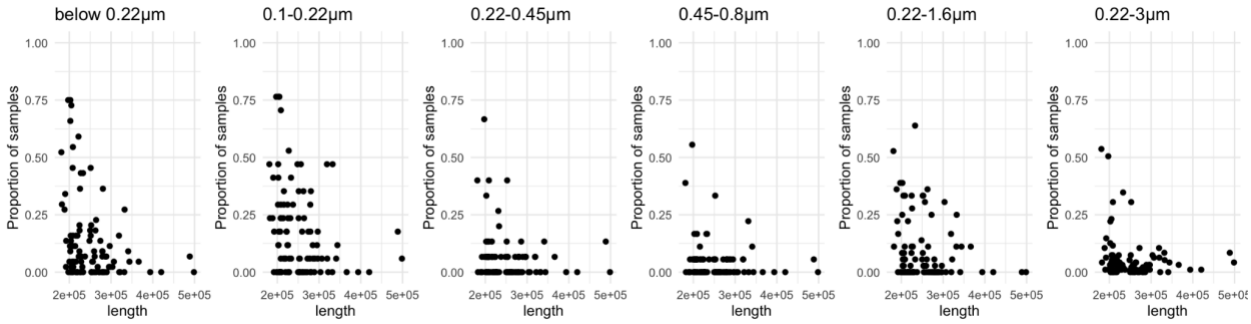
Supplemental Figures below.



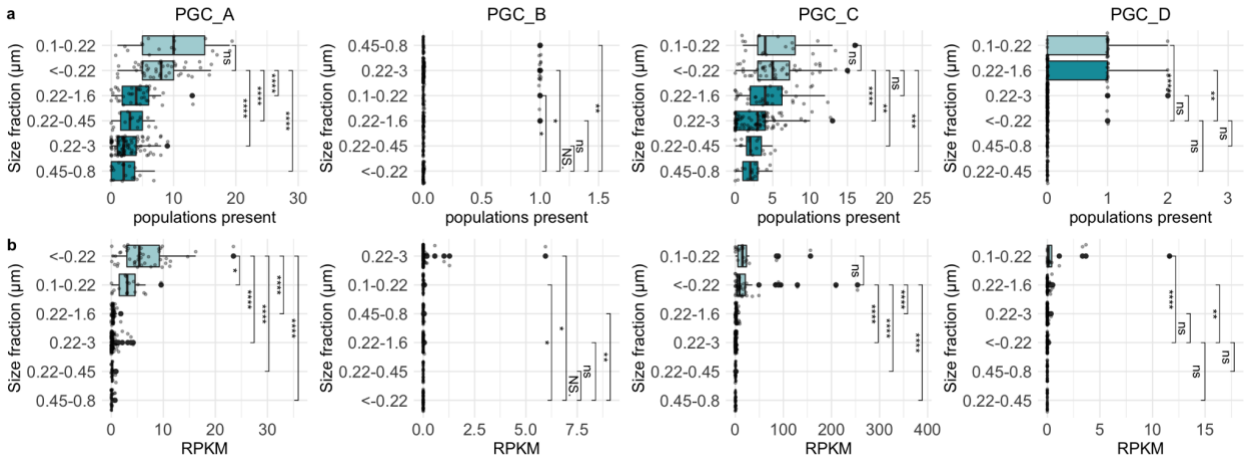
Supplemental Figure 1. Example mummerplot of promer alignment between contigs of a single bin. The top right quadrant shows the alignment of the top contig to the bottom contig and the top left quadrant shows the alignment of the top contig to itself. The color of the line corresponds to percent identity. Diagonal lines in the top left and bottom right quadrant show the two contigs of this bin align to each other across their entire lengths with relatively high percent identity, suggesting these contigs belong to two smaller phages, rather than a single jumbo phage genome.



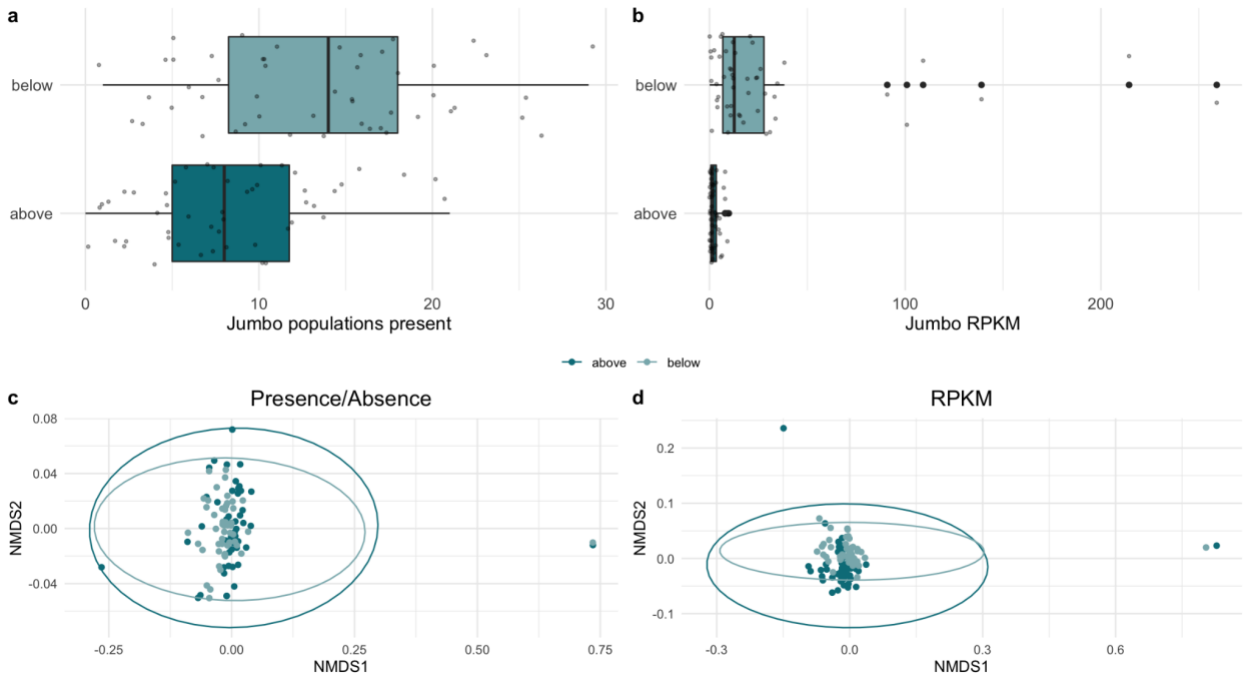
Supplemental Figure 2. (a) Histogram of jumbo bin lengths **(b)** Histogram of the number of contigs in each bin



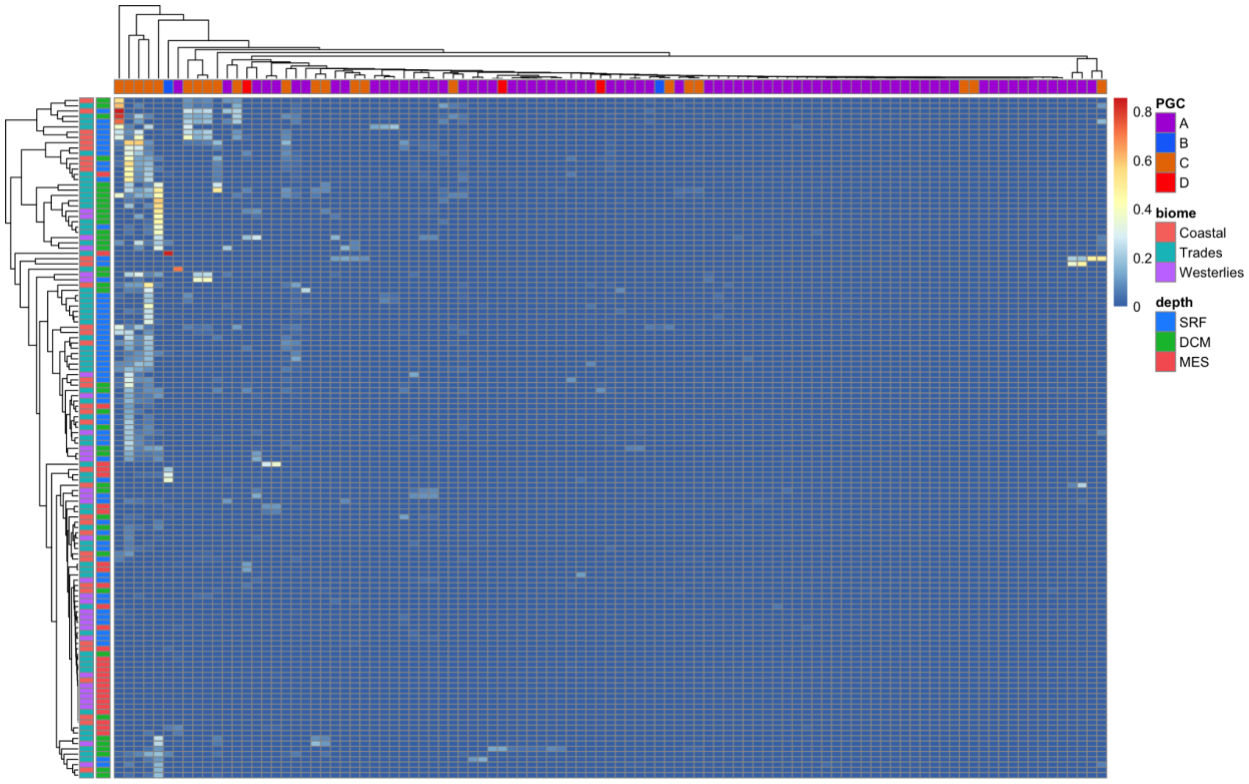
Supplemental Figure 3. Scatterplot with proportion of samples at different size fractions that a jumbo phage is present (y-axis) vs. its length (x-axis)



Supplemental Figure 4. (a) boxplots for each PGC of the number of jumbo phage populations in each sample of different size fractions sorted by mean.**(b)** boxplots for each PGC of the relative abundance of jumbo phages (RPKM) in each sample of different size fractions sorted by median. Significance bars correspond to Wilcox test, with stars corresponding to *p* values < 0.05 and those with *p* values > 0.05 as not significant "ns" (ggplot2 stat_compare_means function).

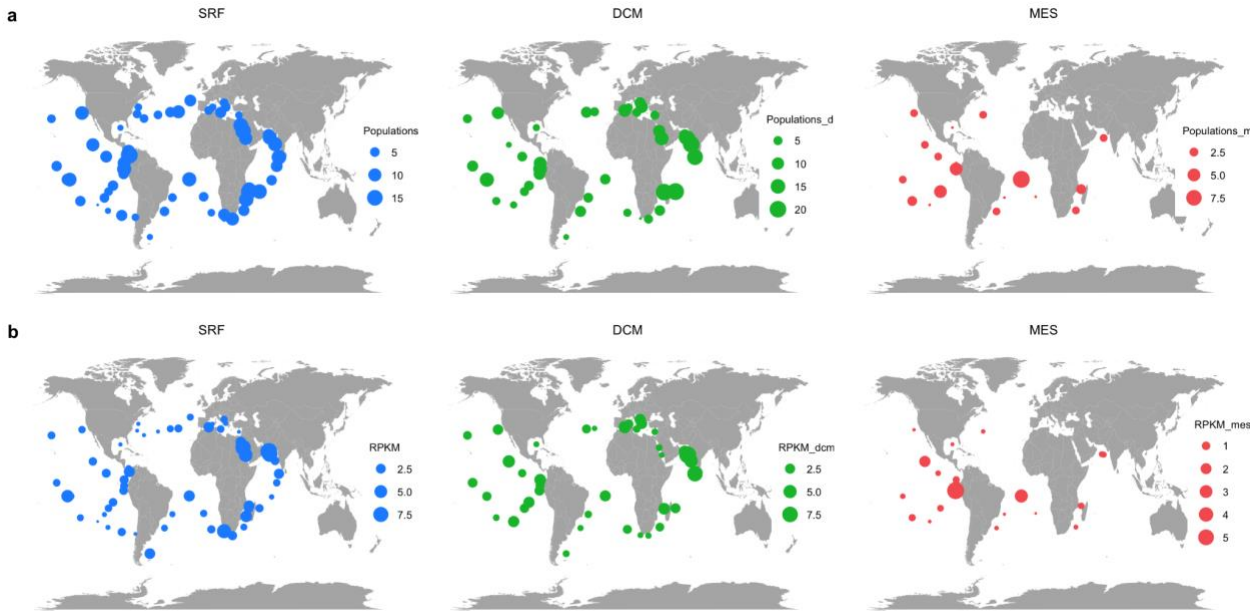


Supplemental Figure 5. (**a**) Boxplot of the number of jumbo phage populations present co-collected at the same station and depth but filtered at below 0.22 µm (size fractions "<-0.22µm" or "0.1-0.22µm") and above 0.22 µm (size fraction "0.22-1.6µm" or "0.22-3µm") (**b**) boxplot of the total RPKM of jumbo phages in these samples. (**c**) NMDS plot of samples based on Bray-Curtis distance matrices of jumbo populations' presence/absence (Richness); communities significantly differed between above and below 0.22 (*p* value = 0.0001, ANOSIM Statistic R 0.1178) (**d**) NMDS plot based on jumbo populations' RPKM; communities significantly differed between above and below 0.22 (*p* value = 0.0001, ANOSIM Statistic R 0.2229). Ellipses calculated based on multivariate normal distribution.

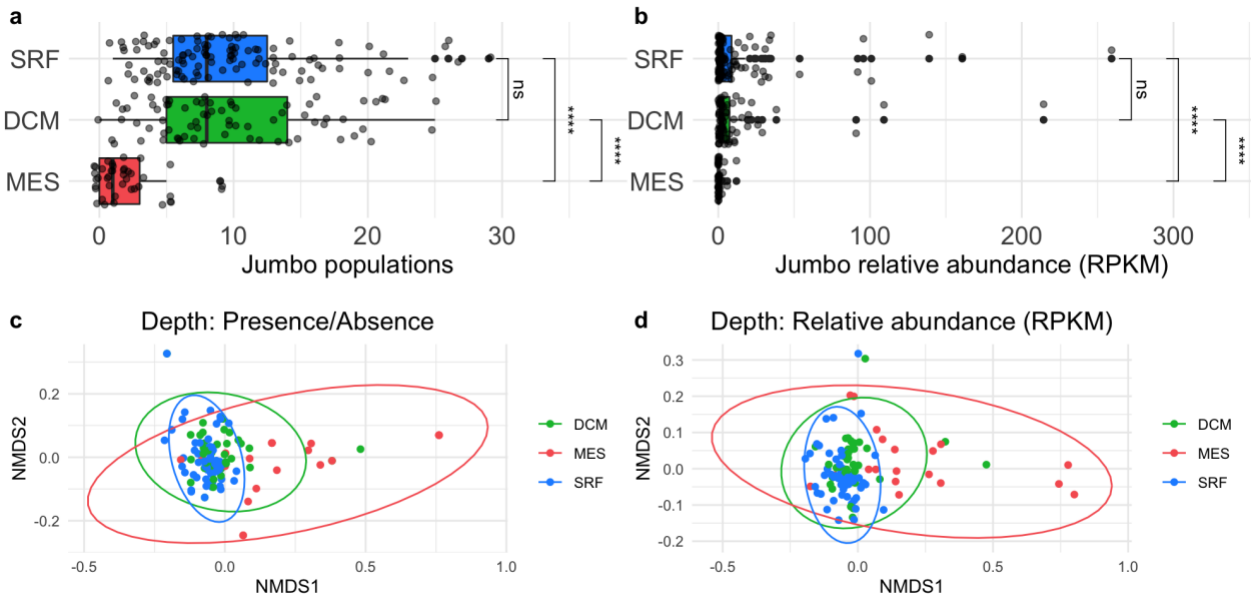


Supplemental Figure 6. Heat map of the log-transformed RPKM (log10(1+RPKM)) of each jumbo phage from this study (columns) in each picoplankton sample (rows). Rows and columns

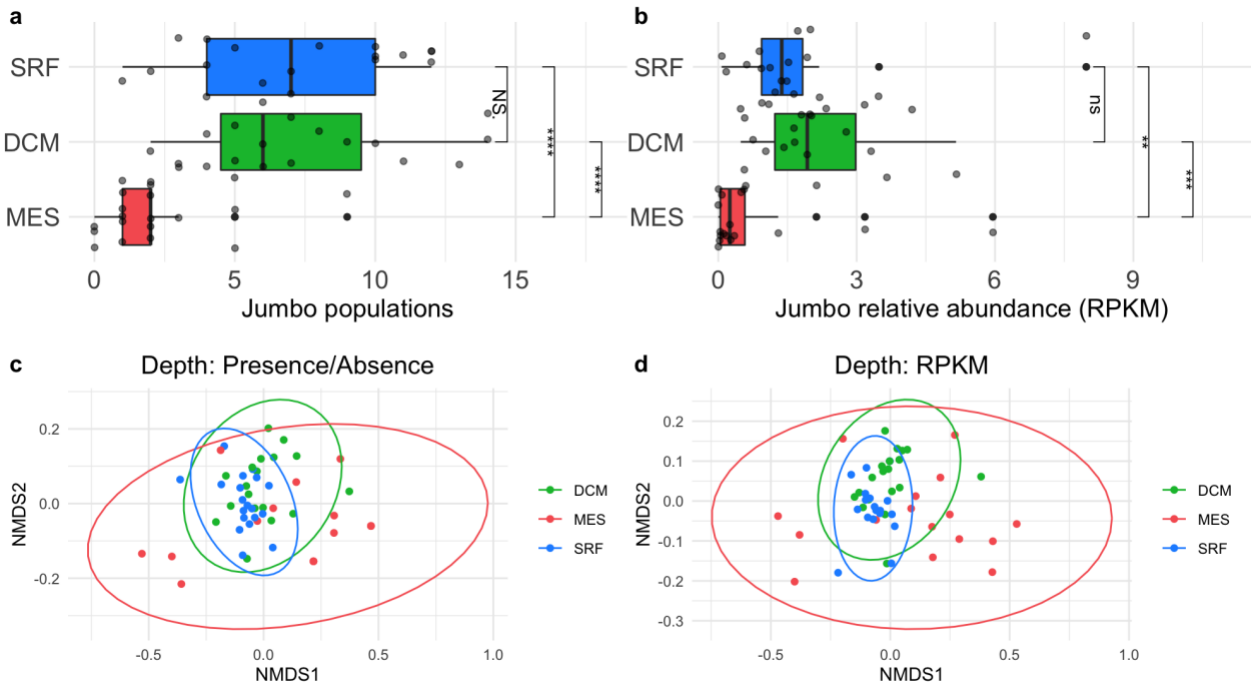
are clustered using hierarchical clustering via pheatmap default settings. Row annotation strip corresponds to each phage's PGC. The outer annotation strip on the columns corresponds to a sample's biome and the inner strip corresponds to a sample's depth.



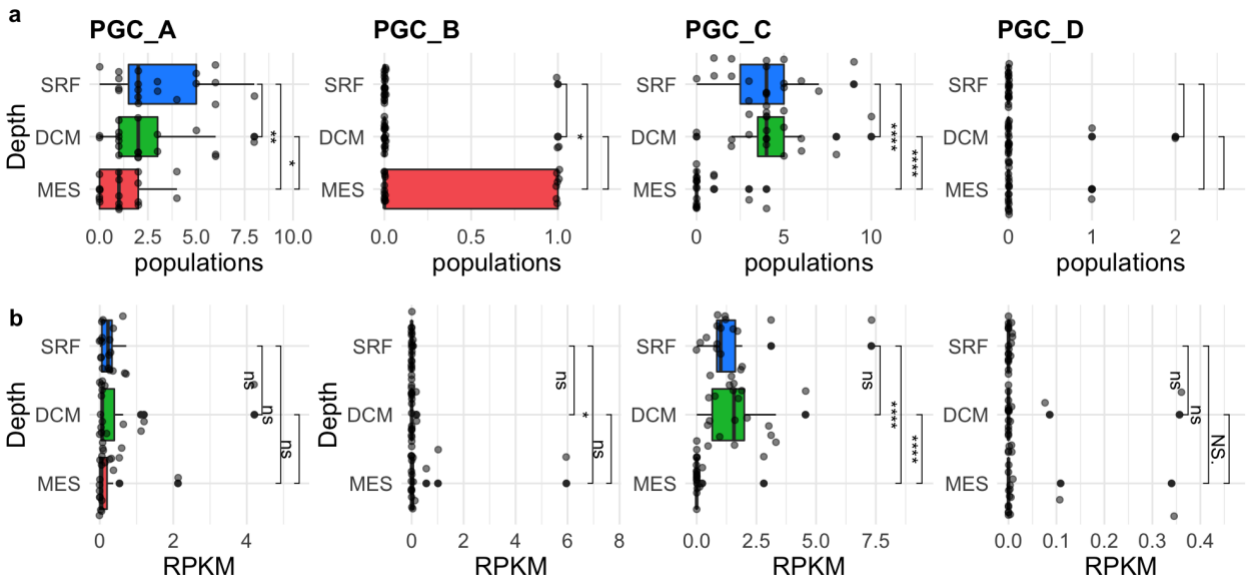
Supplemental Figure 7. Distribution of picoplankton fraction (0.22-1.6 μm or 0.22-3 μm) at each depth. Points are colored by depth (SRF - blue, DCM - green, MES - red). Point sizes in upper row maps correspond to the number of jumbo populations in a sample and point sizes in bottom row maps correspond to jumbo relative abundance (RPKM).



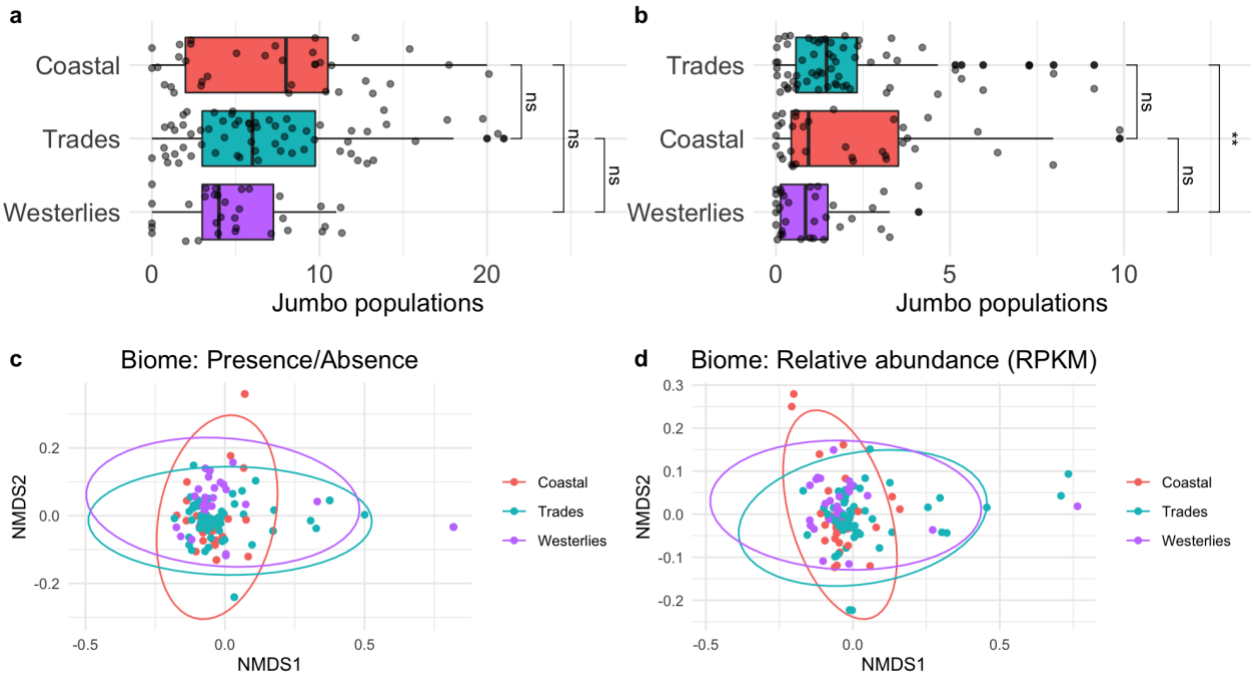
Supplemental Figure 8. (a,b) boxplots of jumbo populations present **(a)** and jumbo abundance in RPKM **(b)** in picoplankton samples by depth sorted by median abundance. Significance bars for a,b correspond to Wilcox test, with stars corresponding to *p* values < 0.05 and those with *p* values > 0.05 as not significant "ns" (ggplot2 stat_compare_means function). **(c,d)** NMDS plots of jumbo composition in those samples based on Bray-Curtis dissimilarity distances using jumbo populations' presence/absence data **(c)** and jumbo population relative abundance in RPKM **(d)** colored by depth. Green - DCM, red - MES, blue - SRF. Ellipses calculated by multivariate normal distribution. Depths were significantly different using ANOSIM (*p* values < 0.01).



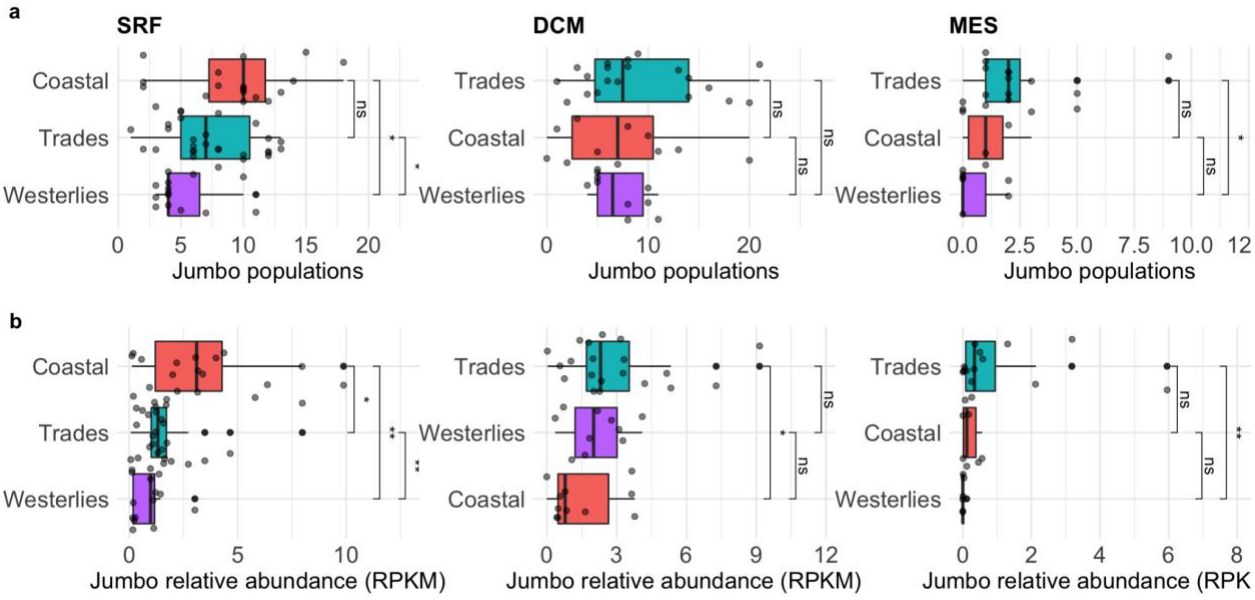
Supplemental Figure 9. (a,b) boxplots of jumbo populations present **(a)** and jumbo abundance in RPKM **(b)** in samples of stations co-collected at all three depths in the picoplankton fraction sorted by median abundance. Significance bars for a,b correspond to Wilcox test, with stars corresponding to *p* values < 0.05 and those with *p* values > 0.05 as not significant "ns" (ggplot2 stat_compare_means function). **(c,d)** NMDS plots of jumbo composition in those samples based on Bray-Curtis dissimilarity distances using jumbo populations' presence/absence data **(c)** and jumbo population relative abundance in RPKM **(d)** colored by depth. Green - DCM, red - MES, blue - SRF. Ellipses calculated by multivariate normal distribution. Depths were significantly different using ANOSIM (*p* values < 0.01).



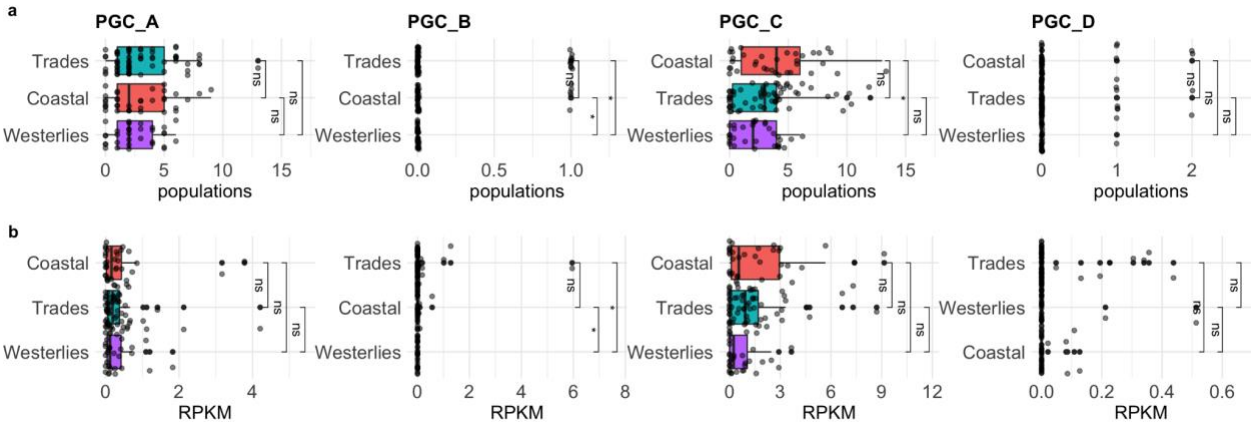
Supplemental Figure 10. (a,b) boxplots for PGCs A-D of jumbo populations present **(a)** and jumbo abundance in RPKM **(b)** in samples of stations co-collected at all three depths in the picoplankton fraction sorted by mean abundance. Significance bars correspond to Wilcox test, with stars corresponding to *p* value < 0.05 and those with *p* values > 0.05 as not significant "ns" (ggplot2 stat_compare_means function).



Supplemental Figure 11. (a,b) boxplots of jumbo abundance in RPKM **(a)** and jumbo population richness **(b)** in samples of the picoplankton fractions, sorted by median abundance at different biomes. **(c,d)** NMDS plots of jumbo composition in those samples based on jumbo population abundance **(c)** and jumbo populations' presence **(d)** colored by biome. pink - Coastal, blue - Trades, purple - Westerlies. Ellipses calculated by multivariate normal distribution. Biomes were significantly different using ANOSIM (p values < 0.05). Significance bars in a,b correspond to Wilcox test, with stars corresponding to p values < 0.05 and those with p values > 0.05 as not significant "ns" (ggplot2 stat_compare_means function).

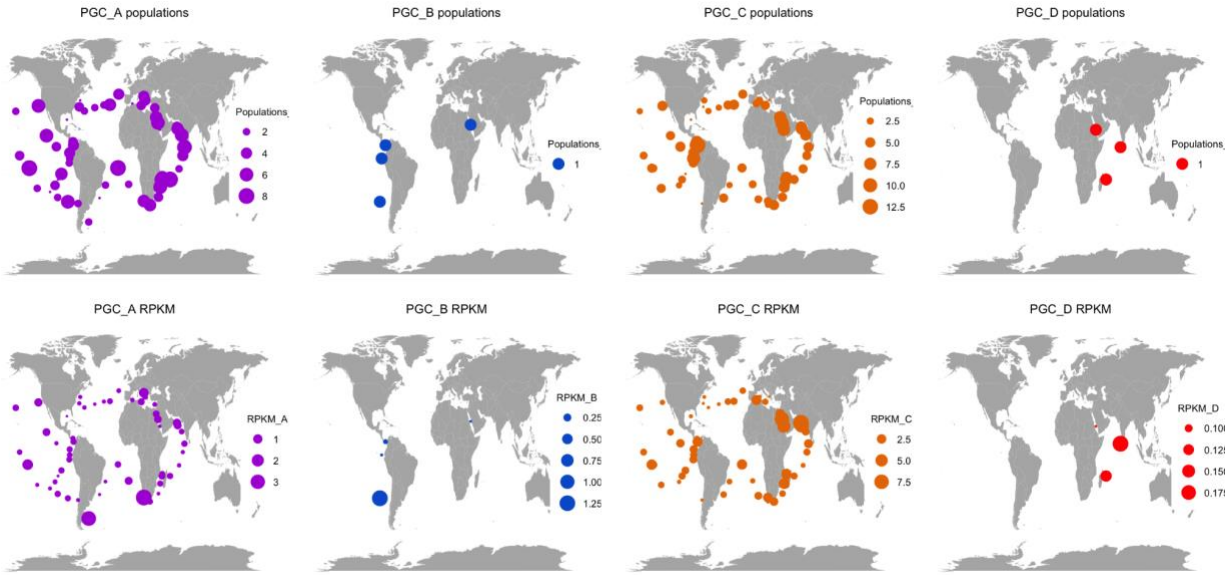


Supplemental Figure 12. (a,b) boxplots of jumbo abundance in RPKM **(a)** and jumbo population richness **(b)** in picoplankton samples of each depth separated by biome and sorted by median abundance in the different biomes. Significance bars correspond to Wilcox test, with stars corresponding to *p* values < 0.05 and those with *p* values > 0.05 as not significant "ns" (ggplot2 stat_compare_means function).

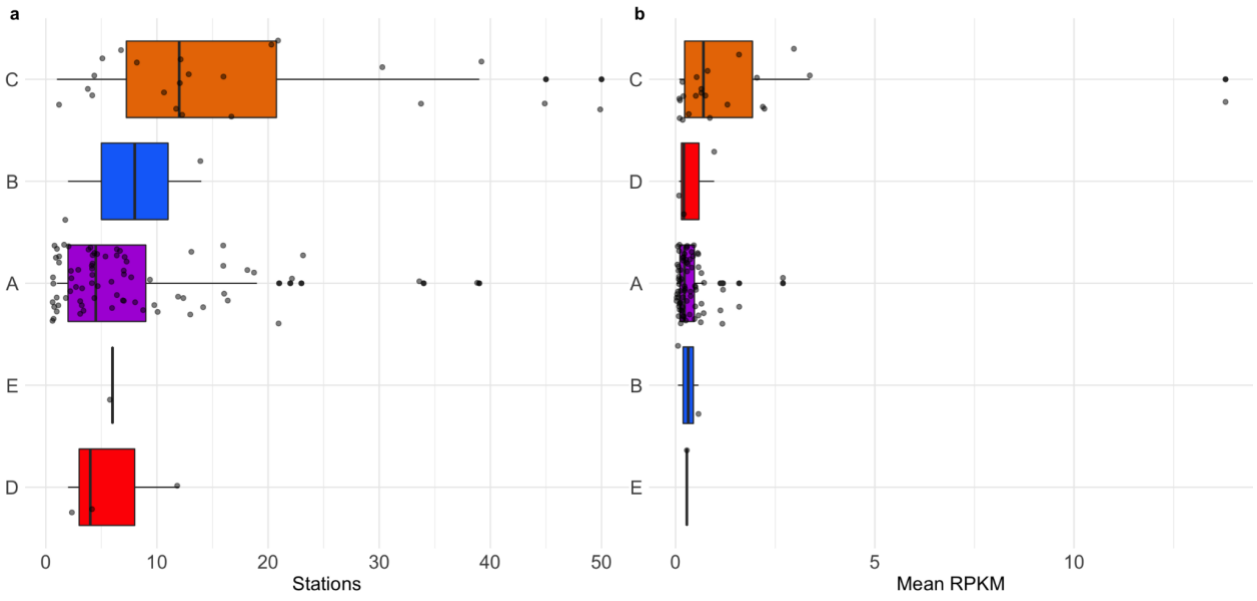


Supplemental Figure 13. (a,b) boxplots of jumbo abundance in RPKM **(a)** and jumbo population richness **(b)** in picoplankton samples of each cluster separated by biome and sorted by

median abundance in the different biomes. Significance bars correspond to Wilcox test, with stars corresponding to *p* values < 0.05 and those with *p* values > 0.05 as not significant "ns" (ggplot2 stat_compare_means function).



Supplemental Figure 14. Maps of the number of jumbo populations (top row maps) and total RPKM (bottom row maps) of jumbo phages in the titled PGC in surface samples of the picoplankton fraction colored by PGC.



Supplemental Figure 15. Boxplots of the total number of stations a marine jumbo phage is present (a) and the mean RPKM that a jumbo phage is present (b) separated by PGC, sorted by median. Colors correspond to PGC.

Figures with viral fraction (<-0.22 or 0.1-0.22) results:

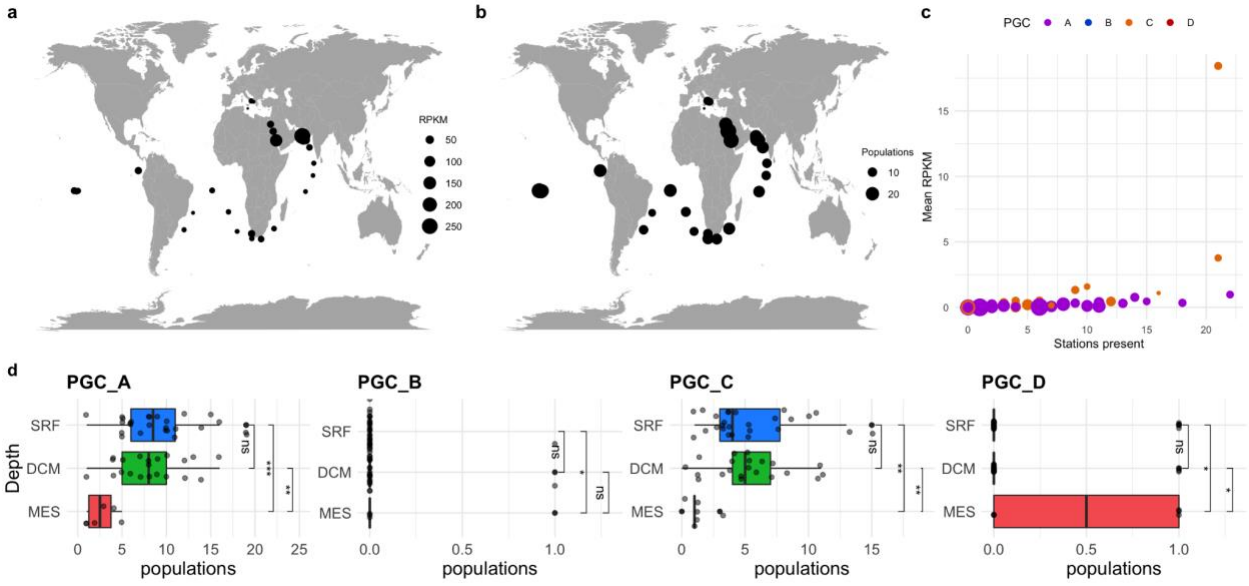
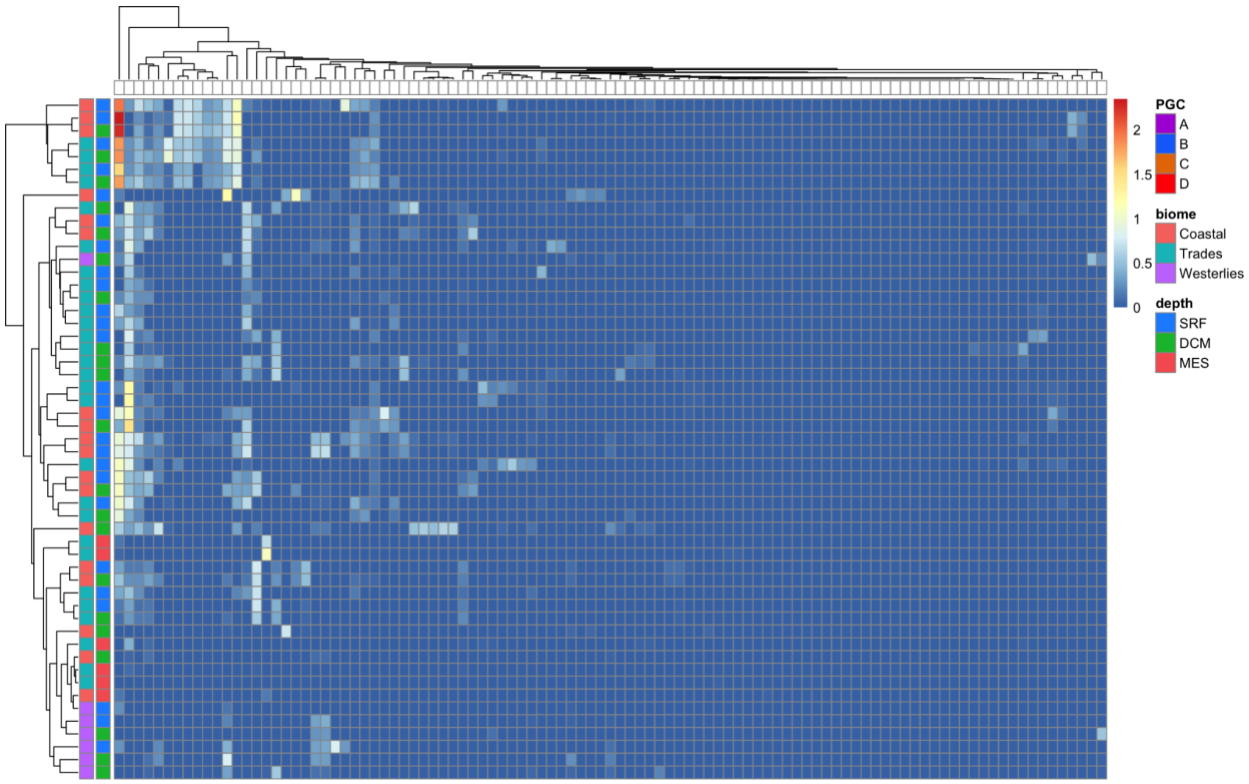
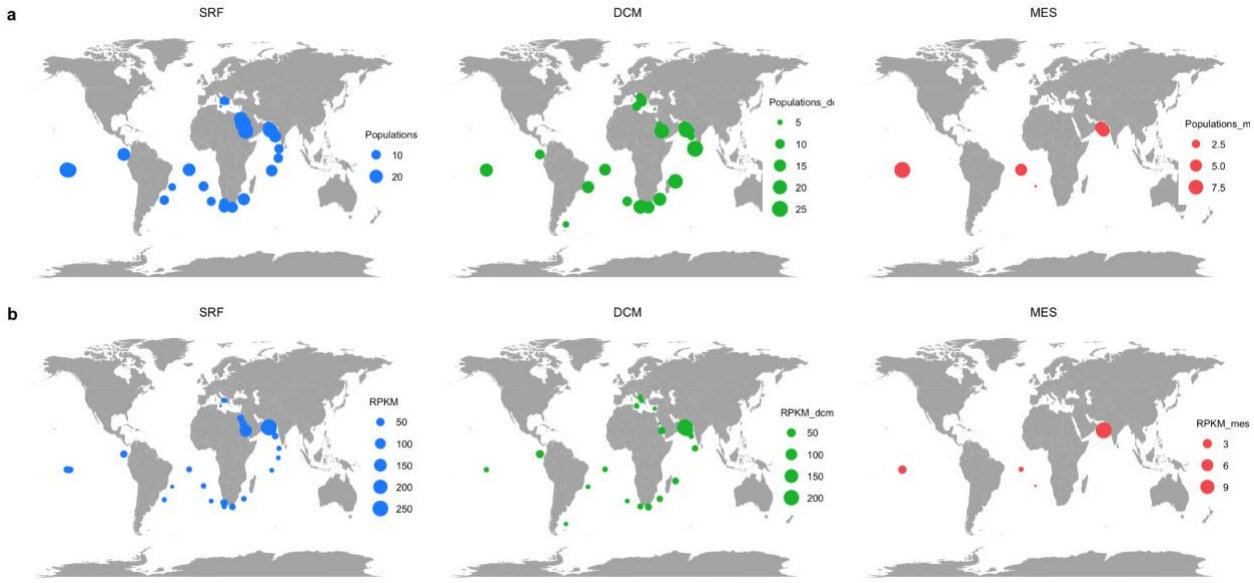


Figure 16. a,b) Maps of the relative abundance (**a**) of total jumbo phages (in RPKM) and (**b**) total number of jumbo populations present regardless of phage cluster membership in each surface (SRF) sample of the virome size fractions (either <-0.22 µm or 0.1-0.22 µm depending on availability). Dots sizes are proportional to the number of populations or RPKM and colored by biome (Coastal - pink, Westerlies - purple, Trades - blue). (**c**) Scatterplot of the mean RPKM of a jumbo population in SRF virome samples versus the number of SRF picoplankton stations it was present. Populations are colored by PGC and size corresponds to putative genome length in 100 kilobases. (**d**) Boxplot of the number of jumbo phage populations in a sample separated by

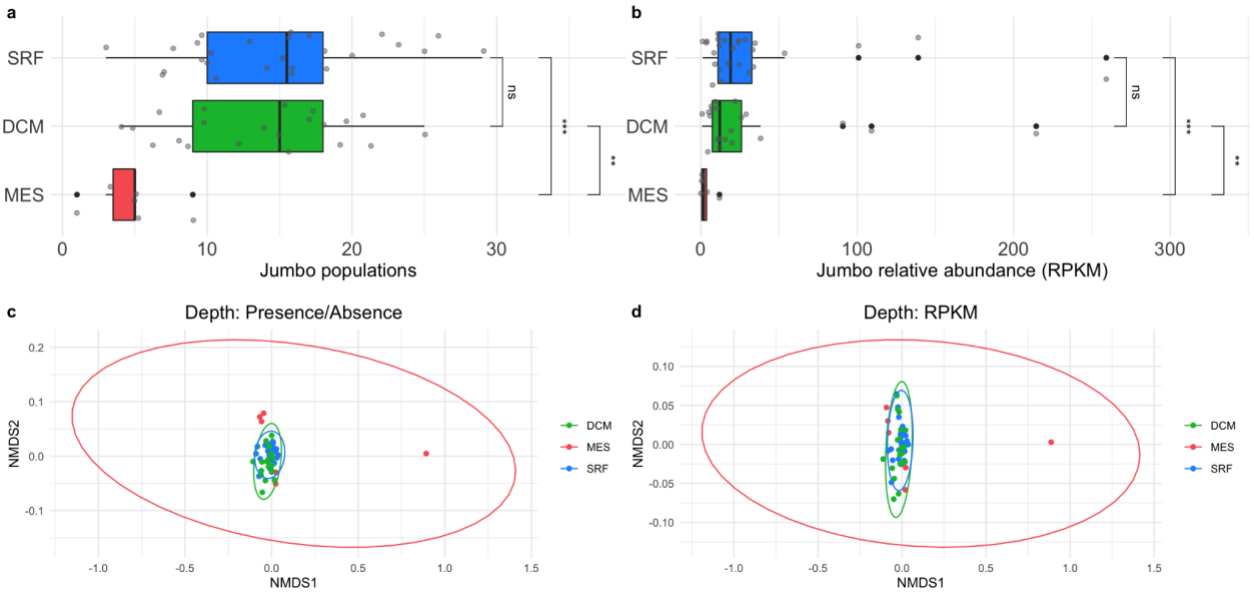
depth sorted by median for each PGC. Significance bars correspond to Wilcox test, with stars corresponding to *p* values < 0.05 (stat_compare_means function)



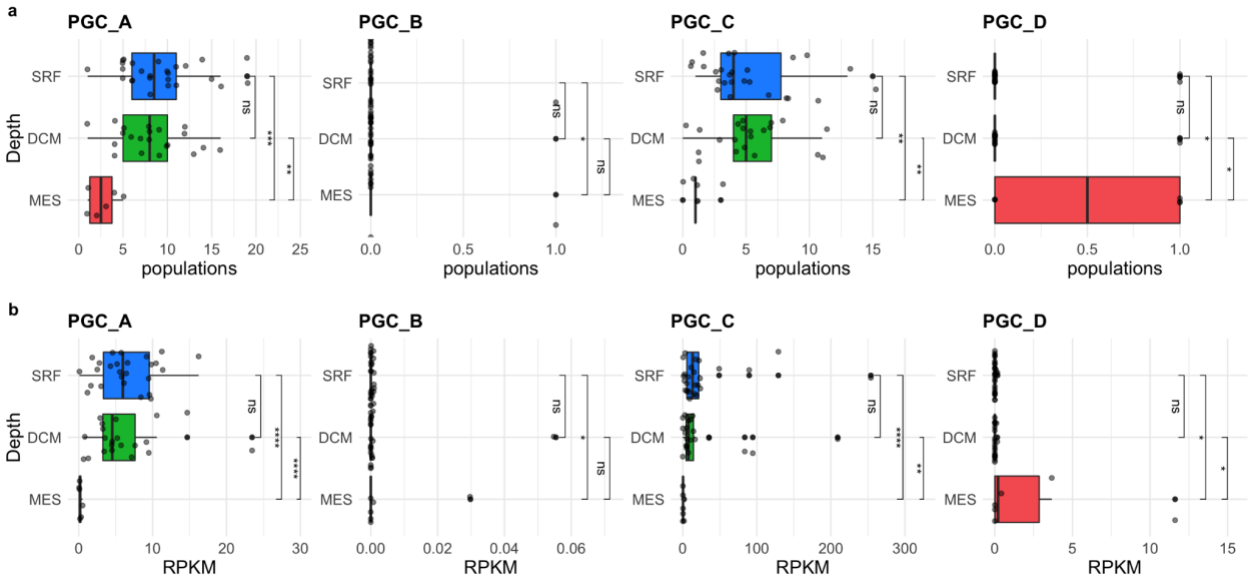
Supplemental Figure 17. Heat map of the log-transformed RPKM (log10(1+RPKM)) of each jumbo phage from this study (columns) in each picoplankton sample (rows). Rows and columns are clustered using hierarchical clustering via pheatmap default settings. Row annotation strip corresponds to each phage's PGC. The outer annotation strip on the columns corresponds to a sample's biome and the inner strip corresponds to a sample's depth.



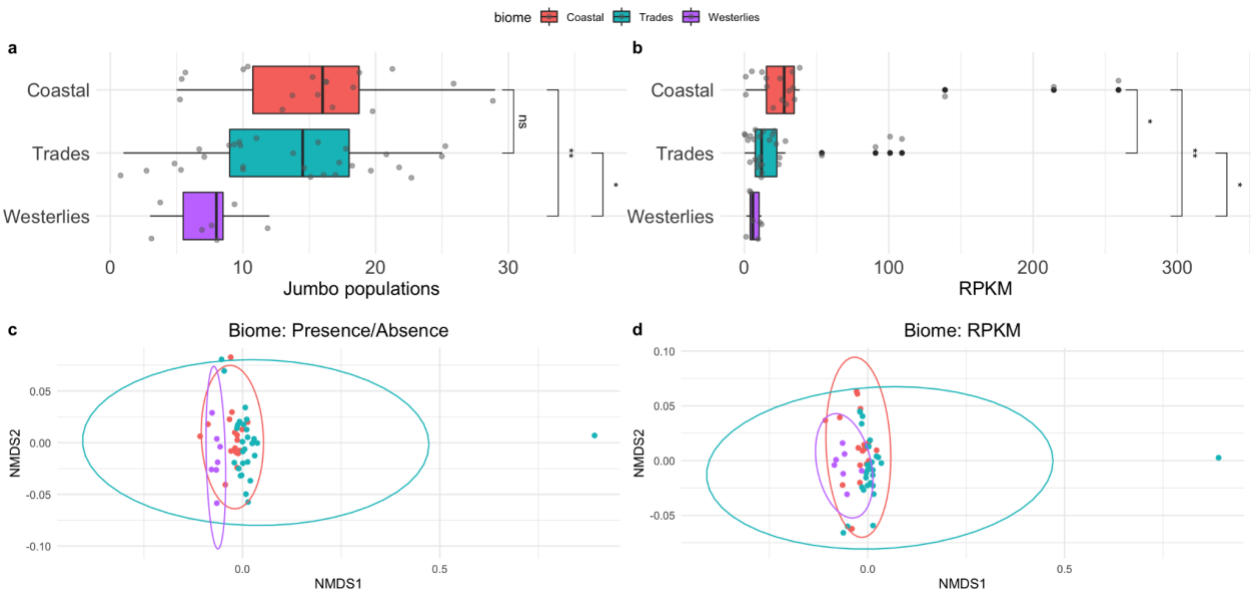
Supplemental Figure 18. Distribution of virome fractions (<-0.22 µm or 0.1-0.22 µm) at each depth. Points are colored by depth (SRF - blue, DCM - green, MES - red). Point sizes in upper row maps correspond to the number of jumbo populations in a sample and point sizes in bottom row maps correspond to jumbo relative abundance (RPKM).



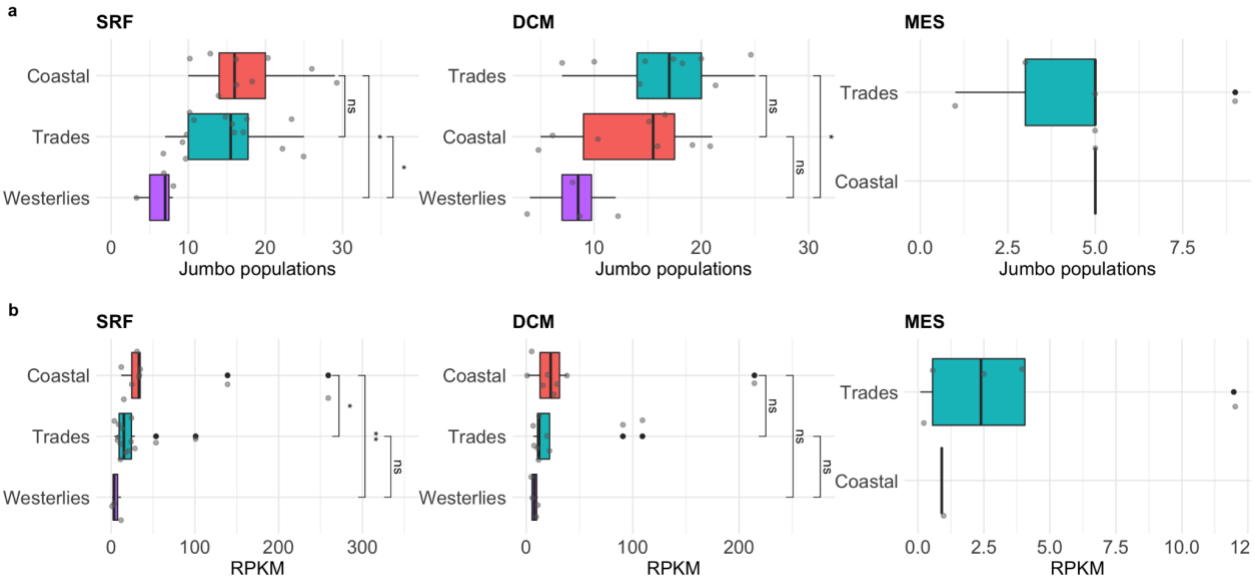
Supplemental Figure 19. (a,b) boxplots of jumbo populations present **(a)** and jumbo abundance in RPKM **(b)** in viral fraction samples by depth sorted by mean abundance. Significance bars for a,b correspond to Wilcox test, with stars corresponding to *p* values < 0.05 and those with *p* values > 0.05 as not significant "ns" (ggplot2 stat_compare_means function). **(c,d)** NMDS plots of jumbo composition in those samples based on Bray-Curtis dissimilarity distances using jumbo populations' presence/absence data **(c)** and jumbo population relative abundance in RPKM **(d)** colored by depth. Green - DCM, red - MES, blue - SRF. Ellipses calculated by multivariate normal distribution. Depths were significantly different using ANOSIM (*p* values < 0.05).



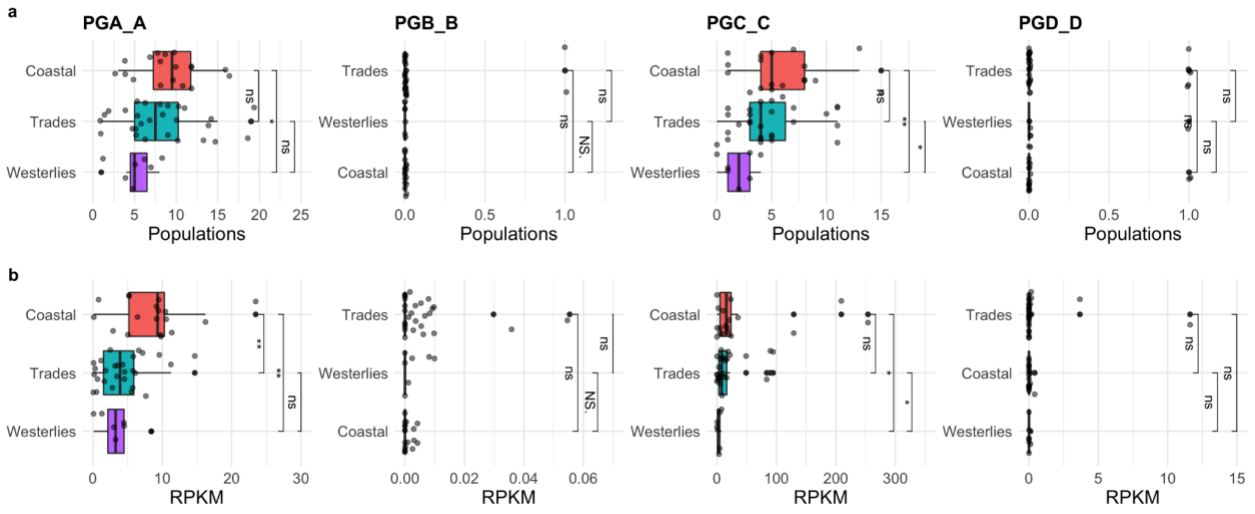
Supplemental Figure 20. (a,b) boxplots for PGCs A-D of jumbo populations present **(a)** and jumbo abundance in RPKM **(b)** in viral fractions samples of stations at all three depths sorted by mean abundance. Significance bars correspond to Wilcox test, with stars corresponding to *p* value < 0.05 and those with *p* values > 0.05 as not significant "ns" (ggplot2 stat_compare_means function).



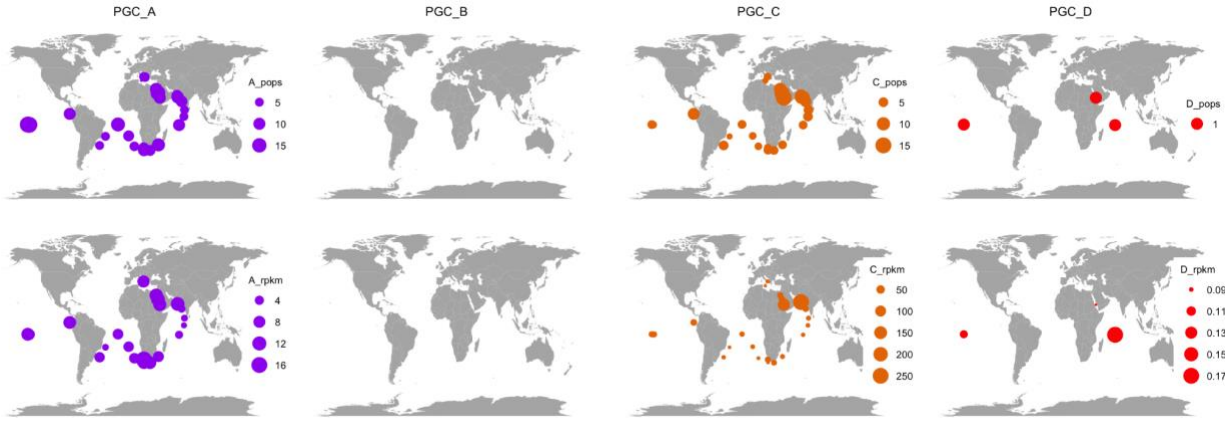
Supplemental Figure 21. (a,b) boxplots of jumbo abundance in RPKM **(a)** and jumbo population richness **(b)** in viral fraction samples sorted by median abundance at different biomes. **(c,d)** NMDS plots of jumbo composition in those samples based on jumbo population abundance **(c)** and jumbo populations' presence **(d)** colored by biome. pink - Coastal, blue - Trades, purple - Westerlies. Ellipses calculated by multivariate normal distribution. Biomes were significantly different using ANOSIM (p value < 0.01, R statistic 0.2). Significance bars in a,b correspond to Wilcox test, with stars corresponding to p values < 0.05 and those with p values > 0.05 as not significant "ns" (ggplot2 stat_compare_means function).



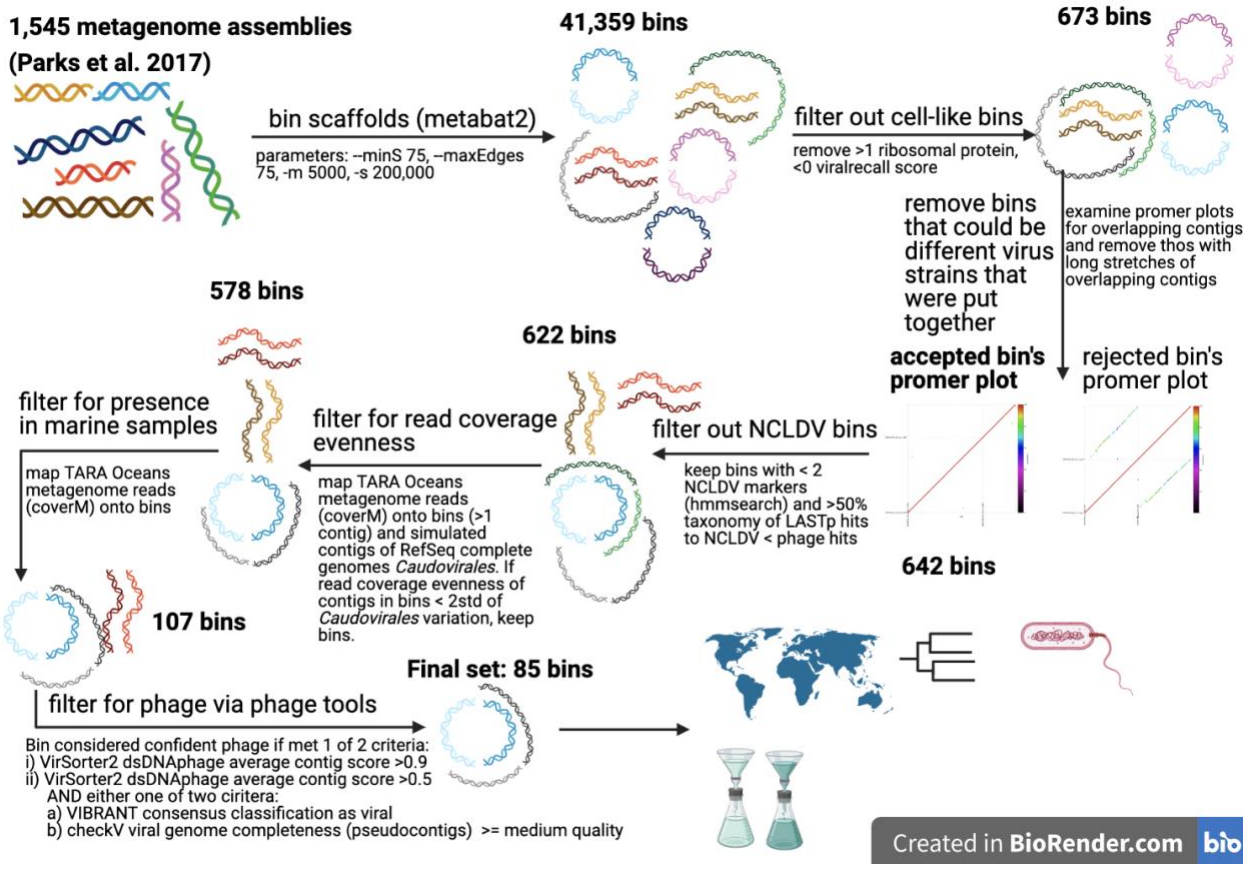
Supplemental Figure 22. (a,b) boxplots of jumbo abundance in RPKM **(a)** and jumbo population richness **(b)** in viral samples of each depth separated by biome and sorted by median abundance in the different biomes. Significance bars correspond to Wilcox test, with stars corresponding to *p* values < 0.05 and those with *p* values > 0.05 as not significant "ns" (ggplot2 stat_compare_means function).



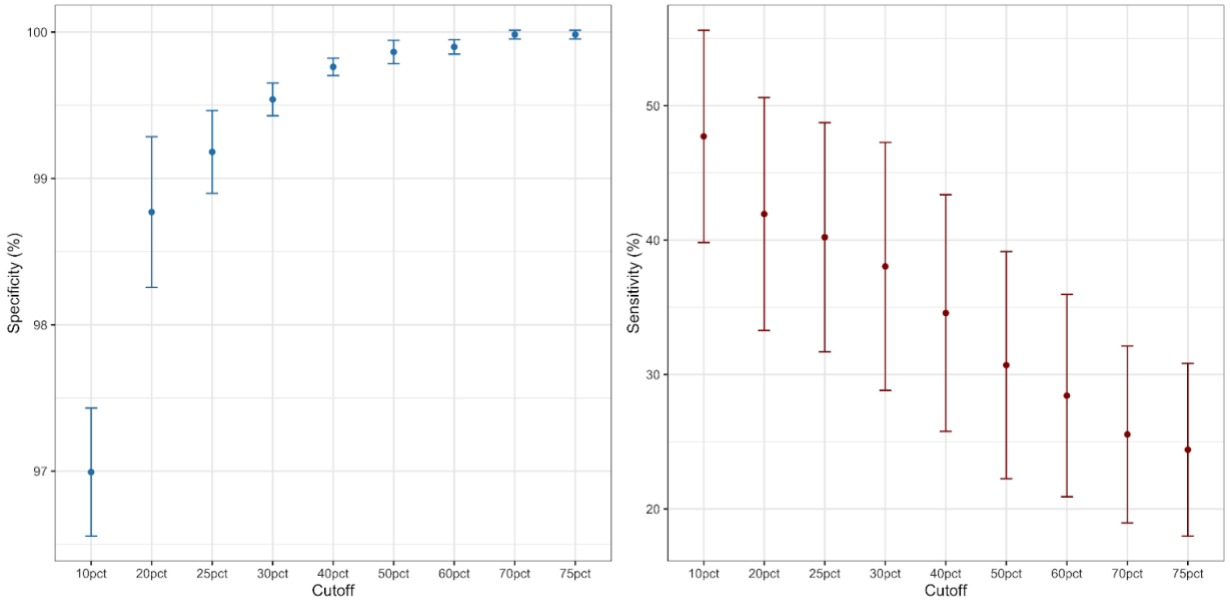
Supplemental Figure 23. (a,b) boxplots of jumbo abundance in RPKM **(a)** and jumbo population richness **(b)** in viral fraction samples of each cluster separated by biome and sorted by median abundance in the different biomes. Significance bars correspond to Wilcox test, with stars corresponding to *p* values < 0.05 and those with *p* values > 0.05 as not significant "ns" (ggplot2 stat_compare_means function).



Supplemental Figure 24. Maps of the number of jumbo populations (top row maps) and total RPKM (bottom row maps) of jumbo phages in the titled PGC in surface samples of the viral fractions colored by PGC.



Supplemental Figure 25. Overview of binning pipeline. Details in Supplemental Methods.



Supplemental Figure 26. Specificity (left) and sensitivity (right) of reference phage identification in simulated metagenomic communities (Roux et al. 2017). Different % covered fraction thresholds are shown on the x-axis. Error bars denote standard error.

Supplemental Methods with References

Supplemental Methods

Binning and screening for non-phage bins.

Contig sequences and coverage information from 1,545 metagenomes were downloaded from Parks et al 2017 (Parks et al. 2017). Contigs were binned with MetaBAT 2 (Kang et al. 2019) with the options --maxEdge 75 --minS 75 -m 5000, and -s 200000, which resulted in 41,359 bins. Bins were then filtered for containing a maximum of 5 contigs (1,456 bins remained). We then predicted the proteins on each bin with prodigal (Hyatt et al. 2010) using default options. To begin filtering out bins potentially belonging to cells, we removed bins that encoded more than 1 ribosomal protein, which were detected with hidden markov model (HMM) searches via HMMER version 3.2.1) (Eddy 2011) (E value 0.001) against 27 Cluster of Orthologous Groups ribosome protein HMM profiles (Galperin et al. 2021) (1,043 bins remained). Next, we ran a beta version of ViralRecall (Moniruzzaman et al. 2020) on the bins to remove bins that had negative scores, which indicate they encode more cellular proteins than viral proteins (673 bins remained).

To address the automated binning complication of strain heterogeneity (cases where contigs binned together based on similar tetranucleotide frequencies and coverage, but actually belong to different viruses), we examined for potential overlapping of conserved regions between contigs by running promer (--maxmatch option) via MUMmer (Kurtz et al. 2004), which compares

sequences to each other. We then examined this output with mummerplot (--color --png options) for cases where contigs contained extended conserved regions with other contigs in the bin and discarded these bins (example in Supplemental Figure 15).

The remaining 642 bins were then screened for Nucleocytoplasmic Large DNA Viruses (NCLDV) by searching the bin proteins for 8 NCLDV markers with an HMM search (E value 0.001). These eight markers included the following with the minimum bitscore cutoff to be considered a hit in parentheses: A32 (200), D5 (200), SFII (200), mcp (200), mRNAc (200), PolB (500), RNR (200), VLTF3 (200). Additionally, a LASTp (Kiełbasa et al. 2011) was run on the proteins of the 642 bins against RefSeq r99 (E value < 0.001). If the taxonomy of hits to phage proteins outnumbered hits to NCLDV proteins or the number of NCLDV markers was below 2, the bin was considered phage (622 bins remained). Bins were then filtered to remove spurious contigs by removing bins with contigs shorter than 5 kb (610 bins remained). Additionally, we removed bins that contained potentially contaminating contigs based on read mapping coverage (see below).

Validation of bins with multiple metagenomic read mapping and detection in marine samples

To further ensure contigs belonging to different phages were not spuriously binned together, we assessed for evenness in contig coverage by mapping reads from different metagenomes to the bins. We used Tara Oceans metagenomes for the mapping (Sunagawa et al. 2015) so these results could also be used to detect marine jumbo phages. Specifically, we focused on results from samples filtered above 0.22 μm to minimize instances of fragmented capsids or free DNA complicating coverage results, which may be more likely in the viral fractions if a capsid is larger than 0.22 μm. Because read mapping evenness can vary in phage genomes due to

conserved regions (Sieradzki et al. 2019), we used mapping results from a reference dataset to benchmark a threshold variation level. For this, we compiled this reference dataset by downloading nucleotide sequences of all complete genomes belonging to the *Caudovirales* order on NCBI's Viral Genomes Portal on July 5, 2020 (referred to as "RefSeq Caudo") and subsetted for jumbo phages ("RefSeq jumbo"); we also included jumbo phage sequences curated by Al-Shayeb et al 2020 (Al-Shayeb et al. 2020). We fragmented these reference jumbo reference sequences with an in-house python script into contigs (1-5) of over 10 kb in length. We then mapped the Tara Oceans metagenomes to this reference set and the bins with multiple contigs (342 bins) with coverM (wwood n.d.) (coverm contig --min-read-percent-identity 95 -m covered_fraction rpkm count variance length -t 32 --minimap2-reference-is-index --coupled; database of phages for mapping was created with minimap2 minimap2 -x sr -d)(Li 2021) and retained phages with at least 10% covered. Next, we calculated the standard deviation of coverage reported in reads per kilobase per million (RPKM) of the different contigs in a bin with a python script. Reads per kilobase per million is calculated by dividing the number of reads mapped to a sequence by the sequence length in kilobases to account for differences in sequence length between genomes and then dividing that by the million number of reads in the sample to account for differences in read depth between samples. The RPKM tables of the bins and references were split by Tara Oceans depth ("env") type (SRF, DCM, MES). In R (3.5.1) (R Core Team 2019) via RStudio (1.1.456). For each depth, we set the maximum standard deviation cutoff to the 95th percentile standard deviation value of the reference RPKM variation (i.e. quantile(reference_srf$std_dev, 0.95). To determine the percentage of samples a bin must have mapped below the reference 0.95 cutoff at each depth, we filtered the bin RPKM table for each depth using percent below the cutoff until the distribution of the standard deviation values

for the bins was not significantly different from the reference distribution using a Wilcox test (*p* value > 0.05). 310 of the 342 bins passed. 268 bins comprised only one contig, totaling the bins at 578. Based on the read mapping results from samples of all size fractions, a jumbo phage was considered present in a sample if at least 10% of its genome was mapped by the sample. Of the 578 bins that passed the validation test, 107 bins were present in marine samples.

Validation of bins as phage with phage-detection tools and population clustering with other jumbo phages

Contigs of the remaining 107 bins were run through VirSorter2 (Guo et al. 2021), VIBRANT (Kieft, Zhou, and Anantharaman 2020), and CheckV (Nayfach et al. 2021). CheckV was also run on pseudocontigs of multi-contig bins that were generated using an in-house python script to join the contigs of the bins together with "N"s. First, bins were retained if the VirSorter2 dsDNAphage score of their contigs averaged above 0.9 (75 bins). Next, bins were retained that had a minimum VirSorter2 dsDNA score average above 0.5 and either had been (i) classified as "virus" by VIBRANT or (ii) considered viral by CheckV with genome quality of medium or above. To further ensure the bins contained non-redundancy between their contigs or contamination, we ran CheckV on the bin's contigs individually and examined the completeness estimation. Only 4 contigs were detected as complete, circular genomes (3 high confidence, 1 low confidence) based on direct terminal repeats (DTRs). These contigs belonged to bins that only contained one contig, suggesting these single-contig bins represent complete jumbo phage genomes. The bins used for subsequent analyses then totaled at 85 bins.

Prior to further gene-based analyses, we checked if the bins of jumbo phages used alternative genetic codes, as has been found for some jumbo phages(Devoto et al. 2019; Al-Shayeb et al.

2020), with Codetta(Shulgina and Eddy 2021) (default options). None clearly used codes other than the standard code 11, and we proceeded with the initial prodigal protein predictions.

We then compared the bins to other jumbo phages and identified those belonging to the same population, defined by sharing over 80% of genes with at least 95% average nucleotide identity with one or more other phages in the population (single-linkage)(Brum et al. 2015). This jumbo phage reference set included those on RefSeq belonging to the *Caudovirales* (93 phages), those prepared by Al-Shayeb et al. 2020 (336 phages) (Al-Shayeb et al. 2020), those available in GenBank compiled by Iyer et al 2021 (Iyer et al. 2021) and Cook et al 2021(Cook et al., n.d.) (400 phages), GOV 2.0 (60 phages) (Gregory et al. 2019), ALOHA 2 (8 phages) (Luo et al. 2020), and one megaphage from the English Channel (Michniewski et al. 2021). These additional jumbo phage sequences and the phage sequences from RefSeq jumbo are referred to as the "jumbo references", totaled at 898 sequences. Nucleotide and amino acid sequences of genes encoded by the 85 jumbo bins and the 898 jumbo references were predicted with prodigal using the default genome setting for each genome individually (-a,-d options). These genes were then aligned to each other with BLASTn. Bins were considered belonging to the same population if 80% of their genes aligned to another bin's genes with an average nucleotide identity of at least 95% (Brum et al. 2015). This analysis resulted in 535 jumbo phage populations, 59 of which contained a jumbo bin generated from this study and 47 populations solely contained bins from this study.

Bipartite network analysis

Jumbo bins were clustered with the jumbo references and *Caudovirales* on RefSeq of all genome sizes and reference jumbo phage set described above based on composition Virus Orthologous Groups (VOG: vogdb.org, downloaded April 14, 2020). Amino acid sequences were searched

against HMM profiles in the VOG database via HMM searches (E-value < 0.001). A matrix of VOG families as columns and phage as rows was generated from the hmm output with an in-house python script (Supplemental Dataset 2). The matrix was loaded into R and an incidence graph was computed with the R library igraph(1.2.5) ("Igraph – Network Analysis Software" n.d.). Clusters were then detected with the spinglass algorithm (Reichardt and Bornholdt 2006) using 50 spins. The spinglass clustering was run 100 times with different seeds. The final clusters were discerned based on the iteration that yielded the highest modularity (seed 544, modularity 0.5856642). Network was visualized with igraph using the Fruchterman-Reingold layout with 5000 iterations (layout.fruchterman.reingold(niter=5000)). Clusters of which the jumbo bins belonged were plotted with ggplot2(3.1.1) (Wickham 2011a) in R for composition of RefSeq phages host phyla and dataset origin; figures were joined with ggpubr(0.2.4) and Inkscape(v 0.92).

MCP and TerL Phylogenies

All major capsid protein (MCP) and terminase large subunit HMM profiles were compiled from vogdb.org (release 98) (see FigShare (https://figshare.com/account/home#/projects/127391)). Proteins of the jumbo bins, jumbo references, and all other *Caudovirales* on RefSeq were searched against these databases with HMM searches (-E 0.001 flag). To reduce the dataset to facilitate phylogenetic analyses and improve the alignment quality, we took only the best hit (highest bitscore) encoded by a phage and then removed protein sequences that were less than two standard deviations below the median length encoded by the references, which was 96 amino acids (aa) for the MCP and 170 aa for TerL. This quality filtering resulted in 74 MCP protein sequences encoded by the bins, 3,193 MCP proteins encoded by the references, 80 TerL

proteins encoded by the bins, and 3,466 TerL proteins encoded by the references. To further reduce the reference dataset, we clustered the reference hits with cd-hit(Fu et al. 2012) using a 90% ID cutoff (-c 0.9 option). This clustering resulted in 1,180 reference MCP and 1,348 reference TerL protein sequences.

In total, 1,254 MCP protein sequences and 1,428 TerL sequences were aligned separately with Clustal Omega (Sievers et al. 2011). The alignments were then trimmed using trimAl (parameter -gt 0.1) (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009). A tree was reconstructed with this alignment using IQ-TREE (Nguyen et al. 2015) with ModelFinder (Kalyaanamoorthy et al. 2017) to select the best fit model according to the Bayesian Information Criterion, which was VT+F+G4 for the MCP alignment and Blosum62+F+G4 for the TerL alignment and 1,000 ultrafast bootstrap replicates. Trees were visualized in iTOL (v5) (Letunic and Bork 2021) and jumbo bins were colored with network cluster. Figures were joined with ggpubr and Inkscape.

Annotation

Amino acid sequences of jumbo phages were searched with HMM searches against HMM profiles of the EggNOG 5.0(Huerta-Cepas et al. 2019) (E-value <0.001), VOG (release 98), and Pfam (Pfam-A, version 32) (Mistry et al. 2020) databases. To identify virion structural proteins, structural proteins in VOG were manually identified (Supplemental Dataset 3). A consensus annotation of a protein was determined based first by Pfam hit because these functions are well-curated (Mistry et al. 2020), then by VOG or EggNOG hit based on bitscore. Pfam does not assign functional categories, so a gene's functional category was based on the category of its EggNOG hit or virion structure designation. These categories were subsequently merged into broader categories (Supplementary Dataset 3). Functions that had multiple EggNOG categories

(i.e. NK) were tallied individually. Stacked barplots of the functional composition of each jumbo phage cluster were based on the average proportion of genes belonging to the category and plotted in R with ggplot2. Genes with known functions that drove variation between clusters A-D of this study were identified by first calculating the proportion of genomes which encoded a given gene in each cluster and then calculating the variance of the proportion of genomes among the clusters in R. Genes with variance above 0.2 and known functions were retained for heatmap visualization made with pheatmap in R.

Group 1.0 jumbo phages, belonging to PGC_B in this study, are known to encode a divergent family B DNA polymerase. As this gene has not been included in the databases examined, we identified the HMM profile of the VOG family corresponding to this divergent family B DNA polymerase by searching a reference sequence for this gene (YP_009153312.1) with an HMM search (-E 0.001) against the VOG database, which was VOG09941 (bitscore > 1000). We then compared the bitscore of genes that hit to this VOG with the classic family B DNA polymerase (PF00136) to identify the occurrence of the divergent family B DNA polymerase in the phages.

Distribution analyses

To examine the distribution of populations of jumbo phages in the ocean, we mapped reads from the Tara Oceans metagenomes used in the bin validation, but excluded Polar samples as there were only 5 available in this set. Reads were trimmed and subsampled to 20 million per sample. They were then mapped onto the representative sequences of the 535 jumbo phage populations as follows. The reference database of the representative jumbo phage sequences was created with minimap2 (minimap2 -x sr -d) and the mapping was carried out with on the jumbo phages with coverM (coverm genome --min-read-percent-identity 95 -m covered_fraction rpkm count

variance length -t 32 --minimap2-reference-is-index --coupled); Mapping results were retained if at least 20% of the phage genome was covered (see *Benchmarking percent coverage for distribution* section below this section). To compare mapping results between phages and samples, reads per kilobase per million (RPKM) was then calculated by dividing the number of reads aligned to a phage by the length of the phage in kilobases and then dividing that by the number of reads in the sample in millions, which accounts for differences in phage sequence length and difference in sample sequencing depth. Statistical tests and plots of the mapping results were carried out in R with the stat_compare_means(label="p.signif") function in ggplot2 to compare samples between biomes, fractions, and depth in richness and abundance of jumbo phage populations. Compositional differences of jumbo phage between samples based on both presence/abundance and RPKM matrices were compared with ANOSIMs in R using the anosim function from the package vegan(2.5-5) (Dixon 2003) (distance="bray",permutations=9999). Maps were plotted in R with the maps ("Maps: Draw Geographical Maps" n.d.) and ggplot2 libraries. Boxplots were plotted in R with ggplot2 and plyr(1.8.4) (Wickham 2011b) to order axes. Non-metric dimensional scaling plots were generated in R based on Bray Curtis dissimilarity matrices (vegdist (method="bray)) using the metaMDS vegan function and visualized with ggplot2; ellipses were calculated with stat_ellipse(type="norm"). Figures were joined with ggpubr in R.

Benchmarking percent coverage for distribution

We considered a jumbo phage to be present in a sample if at least 20% of the genome could be recovered by read mapping with at least 1X coverage (a 20% fraction covered cutoff). To ensure that this was an appropriate cutoff that did not lead to a large number of false-positive

identifications, we benchmarked different cutoffs using three *in silico* viromes generated in a previous study (Roux et al. 2017). We downloaded the trimmed reads from the mock communities labelled Samples 1, 2, and 3 (10 million paired-end reads per sample) and mapped the reads against the complete reference genomes that were used in their construction. The databases used for mapping also included ~2,000 *Caudovirales* genomes selected from the INPHARED database that were added to assess the incidence of false positive phage detection (Supplemental Dataset 4). The additional genomes were selected randomly from a set of *Caudovirales* in INPHARED that had a MASH (Ondov et al. 2016) distances >0.05 when compared to all genomes used to make the mock communities; this was done because the addition of genomes that were closely-related to those used to make the mock communities cannot be considered to be true false positives. We mapped reads with CoverM using the same parameters we used in our jumbo phage work (95% identity), and we then calculated the sensitivity and specificity of different % fraction covered cutoffs (see Supplemental Figure 26). These results revealed that a 20% fraction covered cutoff had a specificity >98%, indicating that it is appropriate for our purposes and that higher values would further decrease sensitivity without a marked increase in specificity.

Host prediction

Hosts were estimated for the bins based on CRISPR spacers, tRNAs, and the taxonomy of genes. CRISPR spacers were predicted for the Genome Taxonomy Database (release 95)(Parks et al. 2018) and metagenome assembled genomes (MAGs) of bacteria and archaea from the metagenomes of which the jumbo phages derived by Parks et al 2017 (Parks et al. 2017), as well of the jumbo bins. All spacers were aligned to the jumbo bins and hits were at least 24 basepairs

in length with <= 1 mismatch (Al-Shayeb et al. 2020) via BLASTn (-task blastn-short). No hosts could be assigned with this approach, and no jumbo bins targeted other jumbo bins. tRNAs were predicted on the jumbo bins and the same MAGs set with tRNAscan-SE (Lowe and Chan 2016) (-bacteria option). Promiscuous tRNAs were downloaded from Paez-Espino et al. 2016(Paez-Espino et al. 2016) and removed based on BLASTn hits (100% ID, <= 1 mismatches). Jumbo tRNAs were then aligned against the MAGs tRNAs with BLASTn and matches were considered with 100% ID and no more than one mismatch. Jumbo phage tRNA sequences were also searched against NCBI's nonredundant database using the BLASTn webserver and matches were retained with the same criteria. Finally, hosts were assigned based on the taxonomy BLASTn hits to the coding sequences of the MAGs. A putative host phylum was considered if a phylum had three times as many hits than the phylum with the next most hits (Al-Shayeb et al. 2020).

References

Al-Shayeb, Basem, Rohan Sachdeva, Lin-Xing Chen, Fred Ward, Patrick Munk, Audra Devoto, Cindy J. Castelle, et al. 2020. “Clades of Huge Phages from across Earth’s Ecosystems.” *Nature* 578 (7795): 425–31.

Brum, Jennifer R., J. Cesar Ignacio-Espinoza, Simon Roux, Guilhem Doulcier, Silvia G. Acinas, Adriana Alberti, Samuel Chaffron, et al. 2015. “Ocean Plankton. Patterns and Ecological Drivers of Ocean Viral Communities.” *Science* 348 (6237): 1261498.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics* 25 (15): 1972–73.

Cook, Ryan, Nathan Brown, Tamsin Redgwell, Branko Rihtman, Megan Barnes, Dov J. Stekel,

Martha Clokie, Jon Hobman, Michael Jones, and Andrew D. Millard. n.d. “INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Phage Genomes.” https://doi.org/10.1101/2021.05.01.442102.

Devoto, Audra E., Joanne M. Santini, Matthew R. Olm, Karthik Anantharaman, Patrick Munk, Jenny Tung, Elizabeth A. Archie, et al. 2019. “Megaphages Infect Prevotella and Variants Are Widespread in Gut Microbiomes.” *Nature Microbiology* 4 (4): 693–700.

Dixon, Philip. 2003. “VEGAN, a Package of R Functions for Community Ecology.” *Journal of Vegetation Science*. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x.

Eddy, Sean R. 2011. “Accelerated Profile HMM Searches.” *PLoS Computational Biology* 7 (10): e1002195.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data.” *Bioinformatics* 28 (23): 3150–52.

Galperin, Michael Y., Yuri I. Wolf, Kira S. Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V. Koonin. 2021. “COG Database Update: Focus on Microbial Diversity, Model Organisms, and Widespread Pathogens.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa1018.

Gregory, Ann C., Ahmed A. Zayed, Nádia Conceição-Neto, Ben Temperton, Ben Bolduc, Adriana Alberti, Mathieu Ardyna, et al. 2019. “Marine DNA Viral Macro- and Microdiversity from Pole to Pole.” *Cell* 177 (5): 1109–23.e14.

Guo, Jiarong, Ben Bolduc, Ahmed A. Zayed, Arvind Varsani, Guillermo Dominguez-Huerta, Tom O. Delmont, Akbar Adjie Pratama, et al. 2021. “VirSorter2: A Multi-Classifier, Expert-Guided Approach to Detect Diverse DNA and RNA Viruses.” *Microbiome* 9 (1): 37.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K. Forslund, Helen Cook, Daniel R. Mende, et al. 2019. "eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses." *Nucleic Acids Research* 47 (D1): D309–14.

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

"Igraph – Network Analysis Software." n.d. Accessed July 23, 2021. http://igraph.org.

Iyer, Lakshminarayan M., Vivek Anantharaman, Arunkumar Krishnan, A. Maxwell Burroughs, and L. Aravind. 2021. "Jumbo Phages: A Comparative Genomic Overview of Core Functions and Adaptions for Biological Conflicts." *Viruses* https://doi.org/10.3390/v13010063.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, and Lars S. Jermiin. 2017. "ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates." *Nature Methods* 14 (6): 587–89.

Kang, Dongwan D., Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. 2019. "MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies." *PeerJ* 7 (July): e7359.

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

Kiełbasa, Szymon M., Raymond Wan, Kengo Sato, Paul Horton, and Martin C. Frith. 2011. "Adaptive Seeds Tame Genomic Sequence Comparison." *Genome Research* 21 (3): 487–93.

Kurtz, Stefan, Adam Phillippy, Arthur L. Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L. Salzberg. 2004. “Versatile and Open Software for Comparing Large Genomes.” *Genome Biology* 5 (2): R12.

Letunic, Ivica, and Peer Bork. 2021. “Interactive Tree Of Life (iTOL) v5: An Online Tool for Phylogenetic Tree Display and Annotation.” *Nucleic Acids Research* 49 (W1): W293–96.

Li, Heng. 2021. “New Strategies to Improve minimap2 Alignment Accuracy.” *Bioinformatics* , October. https://doi.org/10.1093/bioinformatics/btab705.

Lowe, Todd M., and Patricia P. Chan. 2016. “tRNAscan-SE On-Line: Integrating Search and Context for Analysis of Transfer RNA Genes.” *Nucleic Acids Research* 44 (W1): W54–57.

Luo, Elaine, John M. Eppley, Anna E. Romano, Daniel R. Mende, and Edward F. DeLong. 2020. “Double-Stranded DNA Virioplankton Dynamics and Reproductive Strategies in the Oligotrophic Open Ocean Water Column.” *The ISME Journal*. https://doi.org/10.1038/s41396-020-0604-8.

“Maps: Draw Geographical Maps.” n.d. Accessed July 25, 2021. https://CRAN.R-project.org/package=maps.

Michniewski, Slawomir, Branko Rihtman, Ryan Cook, Michael A. Jones, William H. Wilson, David J. Scanlan, and Andrew Millard. 2021. “Identification of a New Family of ‘megaphages’ That Are Abundant in the Marine Environment.” *bioRxiv*. https://doi.org/10.1101/2021.07.26.453748.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, et al. 2020. “Pfam: The Protein Families Database in 2021.” *Nucleic Acids Research* 49 (D1): D412–19.

Moniruzzaman, Mohammad, Alaina R. Weinheimer, Carolina A. Martinez-Gutierrez, and Frank

O. Aylward. 2020. “Widespread Endogenization of Giant Viruses Shapes Genomes of Green Algae.” *Nature* 588 (7836): 141–45.

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

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” *Molecular Biology and Evolution* 32 (1): 268–74.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome biology* 17(1):1-14.

Paez-Espino, David, Emiley A. Eloe-Fadrosh, Georgios A. Pavlopoulos, Alex D. Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N. Ivanova, and Nikos C. Kyrpides. 2016. “Uncovering Earth’s Virome.” *Nature* 536 (7617): 425–30.

Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2018. “A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life.” *Nature Biotechnology* 36 (10): 996–1004.

Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. “Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life.” *Nature Microbiology* 2 (11): 1533–42.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing* (version 3.6.1).

Vienna, Austria: R Foundation for Statistical Computing.

Reichardt, Jörg, and Stefan Bornholdt. 2006. “Statistical Mechanics of Community Detection.” *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 74 (1 Pt 2): 016110.

Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A., & Sullivan, M. B. 2017. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* 5: e3817.

Shulgina, Yekaterina, and Sean R. Eddy. 2021. “A Computational Screen for Alternative Genetic Codes in over 250,000 Genomes.” *eLife* 10 (November). https://doi.org/10.7554/eLife.71402.

Sieradzki, Ella T., J. Cesar Ignacio-Espinoza, David M. Needham, Erin B. Fichot, and Jed A. Fuhrman. 2019. “Dynamic Marine Viral Infections and Major Contribution to Photosynthetic Processes Shown by Spatiotemporal Picoplankton Metatranscriptomes.” *Nature Communications* 10 (1): 1–9.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. “Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega.” *Molecular Systems Biology* 7 (1): 539.

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, et al. 2015. “Ocean Plankton. Structure and Function of the Global Ocean Microbiome.” *Science* 348 (6237): 1261359.

Wickham, Hadley. 2011a. “ggplot2.” *Wiley Interdisciplinary Reviews: Computational Statistics*. https://doi.org/10.1002/wics.147.

———. 2011b. “The Split-Apply-Combine Strategy for Data Analysis.” *Journal of Statistical Software* 40 (1): 1–29.

wwood. n.d. “GitHub - wwood/CoverM: Read Coverage Calculator for Metagenomics.” Accessed July 23, 2021. https://github.com/wwood/CoverM.

Chapter 4: Contrasting drivers of abundant phage and prokaryotic communities in tropical, coastal ecosystems across the Isthmus of Panama

Alaina R. Weinheimer, Frank O. Aylward, Matthieu Leray, and Jarrod J. Scott

Abstract

Phages, or viruses that infect bacteria and archaea, are ubiquitous and abundant members of Earth's ecosystems that impact the flow of nutrients, evolution of microbes, and food web dynamics by selectively infecting and killing their prokaryotic hosts. Because phages can only replicate through their hosts, they are inherently linked to processes impacting their hosts' distribution and susceptibility to infection. Despite these links, phages can also be affected by environmental parameters independent of their hosts, such as pH or salinity which impact cell adsorption or virion degradation. To understand these complex links, in this study, we leverage the unique ecological context of the Isthmus of Panama, which narrowly disconnects the productive Tropical Eastern Pacific (TEP) and Tropical Western Atlantic (TWA) provinces, and compare factors that shape active marine phage and prokaryotic communities. Metagenomic sequencing of seawater from mangroves and reefs of both the TEP and TWA coasts of Panama suggest that pronounced environmental gradients, such as along the TEP mangrove rivers, result in common dispersal and physicochemical parameters shaping both prokaryotic and phage community composition and diversity. Conversely, we find that when environmental conditions are relatively similar across adjacent habitats, such as between the mangroves and reefs in the TWA, prokaryotic communities are more influenced by local abiotic conditions while phage communities are shaped more by dispersal. Collectively, this work provides a framework for addressing the co-variability between viruses and their hosts in marine systems and for

identifying the different factors that drive consistent versus disparate trends in community shifts, which is essential to inform models of these interactions.

Introduction

Microbes are crucial components of Earth's ecosystems, particularly in the ocean, where they form the foundation of food webs, power biogeochemical cycles, and expand the ecological niches of plants and animals[1,2]. Outnumbering even microbes, viruses serve as major top-down control on microbial communities and modulate microbial ecology and evolution through selective killing via infections, horizontal gene transfer via transduction, and metabolic reprogramming during infections[3]. Understanding viral impacts on microbes is critical toward modeling the movement of nutrients through ecosystems[4], the evolution of microbial pathogens[5], and the dynamics of organismal-associated microbiomes[6]. While rapid advances in sequencing and microscopy technologies over the past few decades have begun to unfold the vast diversity, complexity, and breadth of viruses in nature[7–9], major questions remain on which factors shape viral communities and how this relates to concomitant shifts in microbial communities. Because viruses are restricted to reproducing through their hosts, viruses are inherently linked to processes related to their hosts' distribution and susceptibility to infection. Despite these tight links, patterns in the composition and diversity of these two groups can differ depending on the parameter or environment. Showing coupled shifts, for instance, viral diversity and microbial diversity in the ocean has been shown to increase with depth[10,11], and the pH of soils has been shown to co-vary with viral and prokaryotic (bacteria or archaea) diversity[12]. In contrast to this coupling, a study examining soil communities and one on communities in freshwater springs showed that viral communities shifted over spatial scales and environmental parameters that did not always match that of microbes[13,14]. Several possibilities have been suggested to explain these

contrasts, such as broader host ranges of viruses lowering the impact of available host composition on viral community structure[15], or metacommunity dynamics[14] such as the importance of high dispersal versus species local adaptation that may differ between microbes and viruses. Taken together, these studies highlight the necessity to untangle the complexity in the link between viral communities and microbial communities, to better characterize roles of microbes and viruses in the environment.

In this study, we leverage the unique biogeography of the Isthmus of Panama to uncover factors shaping viral and microbial communities across a diverse array of tropical coastal environments in two oceans. The Isthmus of Panama gradually formed and finally completely disconnected the Tropical Western Atlantic ocean (TWA) from the Tropical Eastern Pacific (TEP) ocean approximately 2.8 million years ago[16]. The TWA became oligotrophic, leading to the proliferation of reef-building corals. The TEP remained eutrophic, with patchy coral reefs dominated by fewer species of scleractinian corals. Expansive mangroves thrive adjacent to coral reefs in both the TEP and the TWA. Nonetheless, mangroves of the TWA are influenced by much smaller tidal oscillations than in the TEP. In addition, the TWA supports thinner fringes of mangroves made of shorter trees than in the productive TEP[17]. These contrasting coasts with parallel habitat types of mangroves and coral reefs allow comparisons of viral and microbial communities at two spatial scales, locally between habitat types and globally between oceans[18]. Given the intrinsic link of viruses to their hosts, our null hypothesis was that factors shaping viral communities mirror those of microbial communities, and this similarity would be most visible at global scales between the oceans since the spatial separation and chemical differences between oceans are so large. An alternative hypothesis is that factors shaping viral communities would not match those of the corresponding microbial communities, and these differences would be

most apparent at smaller scales where subtle differences in environmental parameters can influence contact rates of viruses to hosts, growth rates of hosts, and other physical aspects that may decouple viral communities from microbial communities.

To address these hypotheses, we examined both prokaryotic and viral community diversity in seawater metagenomes filtered for the 0.22-0.8 μm size fraction. While most known viruses are smaller than 0.22 μm, the viruses detected in this size fraction correspond to a subset of the viral community that includes larger viruses (e.g. jumbo bacteriophages), actively replicating (pre-lytic) viruses, lysogenic viruses (those integrated in the genomes of hosts) or viruses that have stuck to particles, putatively representing an active or abundant subset of the viral community[19]. Because viruses of bacteria and archaea that belong to the class *Caudoviricetes* are ubiquitous members of ecosystems, we focused our analyses on these viruses that we refer to as phages. For the microbes, we focused on the prokaryotes, as they are the putative host pool of the phages. To directly compare phage and prokaryotic diversity and minimize information loss, we applied a gene-based approach. We selected marker genes for both phages and prokaryotes that we benchmarked against more commonly used contig-based analyses.

Our results reveal a variety of contexts when factors shaping phage and prokaryotic communities align and when they diverge. Supporting our null hypothesis, the phage and prokaryotic communities were both distinct between oceans. The importance of habitat type, however, differed between these groups. Distinctions in phage community composition between mangroves and reefs depended on the ocean, with the mangrove communities being highly distinct from the reef communities in the TEP but not the TWA, likely due to the strong salinity gradient of the mangrove rivers in the TEP. Meanwhile, the prokaryotic communities were equally distinct between the habitat types in both oceans. The lack of separation between the

habitat types of the phage communities in the TWA compared to the prokaryotic communities suggests that changes in environmental parameters influence prokaryotic communities nearly equally, as dispersal limits or physical separation, while phage communities are more structured by dispersal limitations than local conditions. Most strikingly, we found that phage communities were more diverse in the TWA, while prokaryotic communities were more diverse in the TEP, suggesting phage production or breadth of host range may differ in ecosystems of the more productive island archipelago of the TEP than in the oligotrophic coastal bay of the TWA. Overall, these findings highlight the necessity to examine viruses with their potential host community together to better untangle processes driving their interactions with each other and the environment in natural, mixed communities. The contexts when phage and prokaryotic communities do not couple each other is crucial for modeling phage-host interactions as they relate to microbial mortality, and ultimately biogeochemical cycling in ecosystems.

Results and Discussion

Benchmarking methods to assess phage and prokaryotic diversity.

To directly compare phage and prokaryotic diversity and minimize information loss, we benchmarked and employed a novel gene-based approach (see Methods), in which families of the major capsid protein (MCP) and terminase large subunit (TerL) belonging to the class *Caudoviricetes* compiled from the Virus Orthologous Groups database (vogdb.org; Supplemental Dataset 2) were detected within the proteins of the contig assemblies from the metagenomes (Supplemental Dataset 2). Phage contigs were also detected for comparison. Reads from all samples were then mapped on to these sequences for their relative abundances in each sample (Supplemental Dataset 3, see Methods), and ecological statistics held for all three metrics (MCP, TerL, contigs; Supplemental Dataset 4). Results from TerL were reported here as this was the

most prevalent gene (Supplemental Dataset 4) and enabled direct comparison with prokaryotic single-genes (versus metagenome assembled genomes). Prokaryotic diversity was detected with proteins families of three genes from the Clusters of Orthologous Groups (COG) database[20]: RNA polymerase β (COG85), RNA polymerase β' (COG86), and a ribosome-binding ATPase YchF (COG12), which has been used in a previous study[21]. Reads from all samples were then mapped on to these sequences for their relative abundances in each sample (Supplemental Dataset 3), and ecological statistics held for all three metrics (COG85, COG86, COG12; Supplemental Dataset 4). The results of RNA polymerase β (COG85) are reported here as this was the most prevalent gene in the dataset (Supplemental Dataset 3). Details can be found in the Methods to use this approach for other datasets and studies.

Proximity and physicochemical variation determine whether factors shaping phage and prokaryotic community composition align.

When comparing the oceans, both phage and prokaryotic community composition significantly differed between the TEP and TWA (Fig. 2a,b; ANOSIM *p values* < 0.05), and their variation correlated with each other (Mantel test *p value* < 0.05). The phage composition, however, differed to a larger extent than did the prokaryotic composition between the oceans, with only 12% of phages found in both oceans (Fig. 2g) compared to 24% of prokaryotes detected in both oceans (Fig. 2h). Furthermore, more physicochemical parameters varied strongly with phage composition than with prokaryotic composition (Fig. 2a,b). The stronger distinction of phage composition between oceans compared to prokaryotes may result from higher dispersal limitations of most phages in the ocean. Although some phages have been detected globally[22],

this cosmopolitan distribution may be less common for phages than for prokaryotes which could be investigated further in future studies.

Distinctions in community composition between the mangroves and reefs depended on the oceans. In the TEP, both phage and prokaryotic community compositions significantly differed between the habitat types (Fig. 2e,f). In the TWA, only the prokaryotic composition was distinct (Fig. 2d), and the phage composition did not differ (Fig. 2c). In the TEP, the mangrove samples were collected along two rivers, with samples spanning from fully saline to fully freshwater. A set of fully saline samples was also collected along a mangrove channel (Supplemental Dataset 1). Although salinity is known as one the greatest factor limiting species ranges[23,24], the separation of the prokaryotic and phage communities here appear to cluster by river rather than by salinity (Supplemental Fig. 1). Nevertheless, the same physicochemical parameters seemed to vary with the both phage and prokaryotic community composition in the TEP (Fig. 2e,f). This suggests that the phage and prokaryotic communities in the TEP are likely impacted by dispersal and environmental parameters similarly. Meanwhile in the TWA, physicochemical differences between mangroves and reefs were less pronounced than in the TEP (Supplemental Fig. 2), and these habitats were closer in proximity (Fig. 1). Despite the lower variation in physicochemical parameters within and between reef and mangrove habitats, prokaryotic community composition partitions between habitats. This suggests that prokaryotic communities may respond to factors that we have not measured in this study, such as the distribution of dissolved and particulate organic matter. The close proximity of the mangroves and reefs, however, may have resulted in high dispersal of phages between the habitat types leading to lower distinctions in the composition, as most phages were found in both habitat types (65%; Fig. 2g). The dispersal of phages between habitat types can result in a lag or delay in the shifts of

phage community structure to changes in host composition because phages can only replicate upon attaching to and infecting their hosts.

Taken together, these results suggest that phage and prokaryotic community composition align when environmental conditions and spatial scales strongly structure putative host communities such as for the TEP samples (Fig. 2a,e,f). Meanwhile, when these parameters are less variable, as in the TWA here, dispersal forces may structure phage communities more so than for putative host communities (Fig. 2b,c,d). Another explanation between uncoupled patterns of composition in the TWA could be that physicochemical parameters interact differently on the phage and prokaryotes. For instance, pH can impact the adsorption of phages to their hosts, despite the presence of their hosts[25]. These parameters, however, would need to be tested directly.

The most prevalent and influential phages and prokaryotes distinguishing the communities belong to diverse taxa and ecological groups.

To determine which groups of phages and prokaryotes were driving the distinctions in the composition of communities, we classified the sequences using multiple approaches. The phages were classified based on the taxonomy of their putative host estimated by the alignment of the terminase large subunit (TerL) sequences to genes of RefSeq 207 and examining the host of the hits. RNA polymerase beta subunit (RNAP β) sequences used to represent prokaryotic diversity here were classified based on the consensus classification of the contig on which the RNAP β was present (See Methods for details).

Of the top ten most prevalent genera based on average relative abundance across samples, only three genera overlapped for prokaryotes and putative phage hosts: *Synechococcus*, *Prochlorococcus*, and *Pelagibacter* (Fig. 3). These genera are known as dominant members of

the ocean[26,27]; furthermore, because the phage sequences may also correspond to integrated phages of the prokaryotic community, this may have resulted in the co-prevalence of these genera in both phage and prokaryotic communities. Nevertheless, the general lack of overlap in prevalent phage and prokaryotic genera may have resulted from several factors such as technical limitations in classifying both the phages and prokaryotic sequences, or that most viral lysis occurs for rare but highly productive microbes, as has been observed off the coast of British Columbia in Canada[28], which would result in dominant viruses that infect rarer hosts.

In general, the average genus composition of both the putative phage hosts and of the prokaryotes corroborate the compositional distinctions observed above when using sequence diversity (Fig. 2), with the phage communities being highly similar between mangroves and reefs in the WA, but very distinct in the TEP (Fig. 3b), and the prokaryotic communities being distinct between mangroves and reefs in both oceans (Fig. 3d). In both phages and prokaryotes, the enrichment of *Prochlorococcus* in the TEP relative to the TWA highlights the physicochemical features of the ocean, as the TEP sites were more exposed to pelagic waters than the TWA sites and *Prochlorococcus* is known to be more dominant in pelagic waters than coastal waters where *Synechococcus* is prevalent[29]. Notably, a fully freshwater sample (EPM_13A, 0 ppt salinity) only contained *Prochlorococcus* of the top genera in the putative host community for the phages (Fig. 4a). *Prochlorococcus* bacteria are rarely found in brackish or freshwater conditions[30,31], and instead, a *Prochlorococcus*-like bacteria that is larger in cell size than its marine counterpart has been reported in estuaries[31]. Thus, the presence of this phage terminase with homology to that of a *Prochlorococcus* phage in the fully fresh sample suggests that either (i) this phage infects this *Prochlorococcus*-like freshwater bacteria, (ii) that it has a broad host range that enables it to infect marine and freshwater bacteria, (iii) or that its homology is a result of the limitation of the

reference database. Of the prokaryotic community in this freshwater sample, only an unknown genus in the Proteobacteria phylum was found that was also prevalent in the other samples (Fig. 4c), unsurprisingly as diverse Proteobacteria are common in freshwater systems[32]. The remarkable divergence of the genera in this freshwater sample for both the prokaryotes and putative host community of the phages highlights the crucial role of salinity in shaping microbial communities[23,24].

We then examined which phages and prokaryotes drove the most variation between the samples, determined by those that significantly varied the most with variation in the communities (envfit test; *p values* < 0.05; See Methods; Supplemental Dataset 5). When examining all samples of both oceans and habitats, the phage (WA_000000419261_10), whose terminase showed high homology to that of the *Puniceispirillum phage HMO-2011,* drove the most variation followed by ten other equally influential phages that putatively infect a diversity of host genera (*Prochlorococcus, Puniceispirillum, Acinetobacter, Mycobacterium, Kiloniella, Laceyella, Escherichia*) spanning four phyla (Supplemental Dataset 5). Matching the whole community distinctions between oceans and habitats of Fig. 2, all but one of these 11 phages were exclusively detected in one ocean (TEP or TWA), with phages of the TWA mostly present in both mangroves and reefs and most of those exclusive to the TEP found only in mangrove samples. In contrast, variation in the prokaryotic communities was primarily driven by *Synechococcus* bacteria (top 3 most influential; Supplemental Dataset 5), which follows its known distinction between pelagic and coastal conditions[29], such as the TEP between TWA here. The elevated importance of phages predicted to infect chemoheterotrophic versus photoautotrophic bacteria in driving phage community composition compared to the prokaryotic community further highlights that phage lysis predominantly occurs on the most productive

members of the community, which are often heterotrophic bacteria that experience boom-and-bust cycles as nutrients become available[28].

When examining samples of the TWA and TEP separately, the primary genera or putative host genera driving the variation in the prokaryotic and phage communities respectively aligned in trophic niche for the TWA but contrasted for the TEP. This is surprising because the phage and prokaryotic communities did not align in habitat distinction or physicochemical parameters driving their composition in the TWA (Fig. 2c,d) but they did in the TEP (Fig. 2e,f). In the TWA, the two phages that drove most of the variation showed high homology to the terminase of *Pelagibacter* phage HTVC008M and the *Puniceispirillum* phage HMO-2011, host genera that are both heterotrophic bacteria found throughout the global ocean[26,33]. For the prokaryotes, the top genera also belonged to heterotrophic groups with the top prokaryote belonging to an uncultivated genus WTJO01 in the Puniceispirillales order, and the next most influential belonging to an uncultivated genus UBA974 in the Flavobacteriales order. These heterotrophic bacteria are also found throughout the oceans[26,34]. The overlap in trophic niche of these genera for driving the phage and prokaryotic communities, despite the differences in physicochemical and habitat distinctions, highlights the robust conditions that these groups can inhabit which could explain the lack of alignment in the environmental features driving the overall phage and prokaryotic community compositions.

In contrast to the TWA, the putative hosts of phages driving the variation of phage communities within the TEP did not align trophically despite their overlap in significant physicochemical parameters and habitat distinctions (Fig. 2e,f). The most influential phages primarily putatively infect bacteria belonging to the photosynthetic *Synechococcus* genus (seven of the top ten), while the most influential prokaryotes primarily belonged to unknown genera in the Betaproteobacteria

class (Supplemental Dataset 5). While these genera contrast each other in trophic lifestyles, these bacteria are known to be highly influenced by salinity[31,35,36] , which widely varied in the TEP as the mangrove samples were collected along freshwater rivers. These results suggest that while phage and prokaryotic communities vary with salinity, the types of bacteria and phages that are most affected by salinity in these sites do not necessarily align.

High prokaryotic diversity is rarely coupled with high phage diversity.

Through selective killing by phages and resistance mechanisms by prokaryotes, phages and prokaryotes are known to drive each other's evolution and microdiversity[37,38], but how these interactions manifest in generating macrodiversity remains poorly studied. Here, we examined the alpha diversity of samples to uncover which environments contain high phage and prokaryotic sequence diversity. We used the Shannon's Diversity index to measure alpha diversity, as this metric accounts for both richness and evenness[39]. Taxa are proxied here as unique marker sequences (see Methods). When comparing diversity between oceans, surprisingly, phage communities were significantly more diverse in the TWA (Fig. 4a) while prokaryotes were more diverse in the TEP (Fig. 4b). These patterns between the oceans held when comparing mangrove and reef samples separately (Supplemental Fig. 3). The contrasting diversity patterns of phage and prokaryotes between oceans may be a result of several abiotic and biotic factors. Although the TWA is generally more oligotrophic than the TEP, the bay where the samples were collected in this study has historically been subject to high levels of runoff, which has been found to elevate bacterial production and density but result in decreased bacterial diversity compared to nearby pristine sites[40]. Thus, the reduced prokaryotic diversity in the TWA compared to the TEP may be due to the pollutants, while the elevated phage diversity

in the TWA compared to the TEP may be due to increased bacterial production and thus phage replication and release. Alternatively, the higher phage diversity in an environment that has lower prokaryotic diversity could be because a variety of phages infect the same hosts. The ecological conditions that would enable the coexistence of diverse phages that infect the same host may be related to the contact rates with hosts[41] or flux between environments introducing novel phages, which may differ between the TEP and TWA, but these would need to be tested directly.

When comparing diversity between habitat types of each ocean, a different pattern emerged. Prokaryotic diversity did not vary between habitat types in both oceans (Fig. 4e,h) despite significant compositional differences between mangroves in reefs (Fig 1d,f), which highlights that conditions which result in increased alpha and beta diversity do not align[42]. In other words, environments that have the greatest variation in community composition (beta diversity) do not always have the most species, or site-level diversity (alpha diversity). For example, a study by Walters and Martiny (2020)[42] that compared the microbial diversity of across ecosystems found that soil samples have the highest number of microbial species (alpha diversity), but sediment, biofilms, and inland waters had the greatest variation between microbial communities (beta diversity).

While the prokaryotic communities exhibited this lack of alignment between alpha and beta diversity differences, the phage communities exhibited the opposite. In the TWA, phage communities did not significantly differ in composition between habitat types (Fig 1c) and did not significantly differ in Shannon's diversity between habitat types (Fig. 1d), and phage diversity did not correlate with prokaryotic diversity (Figure 4f). In the TEP, phage communities differed significantly in composition between habitat types (Fig 1e), and phage communities

differed significantly in Shannon's diversity between the habitat types, with mangrove samples having lower diversity than the reefs samples. Phage diversity did not correlate with prokaryotic diversity in the reef samples but did in the mangrove samples (Fig 4i). In the TEP, mangroves were significantly less saline than reefs (median 28.93 vs. 30.655 ppt, respectively), likely due to the freshwater rivers; additionally, the mangroves were more acidic than reefs (median pH 7.885 vs 8.09), as was the case in the TWA. While these physicochemical differences between TEP mangroves and reef did not manifest in prokaryotic diversity differences, the phage diversity may have been impacted as pH and salinity are known to impact adsorption rates of phages to their hosts[25,43]. Furthermore, when the two samples with the lowest salinity were removed (0.06 ppt (EPM_12A1) and 0.09 ppt (EPM_13A1)), phage diversity and prokaryotic diversity did not significantly correlate (Supplemental Fig. 4), highlighting the impact of extreme differences in salinity on phage-host interactions.

Overall, these results highlight the nuances in the relationship between the alpha diversity of phages and their prokaryotic hosts. Similar to the mixed patterns of the compositional differences between oceans and habitat types, these diversity patterns suggest that phage and prokaryotic community diversity only match each other when physicochemical conditions or physical separation imposed on the prokaryotic community are so strong that it limits phage contact rates and replication, such as in the TEP mangroves. The decoupling of these diversities could have resulted from a variety of explanations, such as differences in host densities or the host range of phages, that would need direct investigation in future studies.

Conclusion

In this study, we leveraged the unique biogeography of the Isthmus of Panama to compare drivers of phage and prokaryotic diversity at both global scales between oceans and local scales

between habitat types within each ocean by examining mangrove and reef habitats of the TEP and TWA coasts (Fig. 1). We found that drivers of phage and prokaryotic communities align most when physicochemical and spatial scales are sharp, such as between the oceans and between the TEP mangroves and reefs. Meanwhile, these factors diverge when there are subtle physicochemical differences and minimal physical separation in environments, like between the mangroves and reefs of the TWA. In these cases, prokaryotic communities may locally adapt to the minor environmental differences, as we observed distinction between prokaryotic communities of the mangroves and reefs. The phage communities, however, may be influenced more by high dispersal between the environments, overwhelming environmental or habitat distinctions, as we observed no significant difference between mangroves and reefs of the TWA. A similar pattern has been observed in a freshwater spring system of southern Florida, where the prokaryotic communities were distinct between the river, head, and mixed zones, but the phages communities were not distinct between the head and mixed zone, which the authors attributed potentially to high dispersal of phages between these two zones[14].

Despite cases when drivers of phage and prokaryotic community composition align, our results show that putative host genera of phages that drive phage communities differ from prokaryotes in all spatial and physicochemical scales. Very few of the most dominant phage members infect genera of the most dominant prokaryotes. This may be potentially because most phages are infecting the most productive prokaryotes which exhibit boom and bust reproductive cycles, rather than the most stably abundant[28]. The lack of coupling in the genera of the most abundant phages with the most abundant prokaryote genera could also be explained by deviations from the Kill-the-Winner model in which phages rise in abundance to kill the most dominant prokaryote[44]. Deviations have been observed in a freshwater lake where the abundance of some phages have

been found to peak before, during, or after their host's peak in abundance[45]. Likewise, we found that high phage diversity is rarely coupled with high host diversity. The only context when their diversity correlated was in the TEP mangrove samples, which points to the strong role of salinity in shaping both prokaryotic and phage communities[23,43]. Because phage diversity was higher in the TWA than the TEP, we suspect that host production rates may drive phage diversity such that even if there are blooms of a single bacteria, a variety of phages may surface to infect that host. Conversely, a lower phage diversity amidst high prokaryotic diversity may result if phages have broad host ranges, which may be related to host contact rates[41], but future work is needed to test these hypotheses directly.

All in all, this study provides a framework and demonstrates an application for comparing phage and prokaryotic community composition and diversity in a variety of marine environments. We uncover conditions when the tight links of phages and prokaryotes result in similar factors driving their diversity and composition, such as between oceans, and when these tight links are weakened. By understanding when these phage-host links are strengthened or weakened, we can better predict the outcome of interactions between phages and prokaryote populations of different environments to inform models of nutrient cycling mediated by microbes and the release of organic matter through viral lysis of microbes.

Methods

Sample and environmental data collection. Seawater samples were collected ~1 m above the seafloor on coral reefs and mangroves (1-4 m depth) in the TEP and TWA coasts of Panama in 2017 (see Supplemental Dataset 1 for coordinates and collection dates). Seawater samples were collected in sterile Whirl-Pak Bags and kept on ice and in the dark until filtration at either the Smithsonian Tropical Research Institute (STRI) Coiba (TEP) or Bocas del Toro research stations

(TWA), where they were then vacuum filtered through 0.22 µm nitrocellulose membrane filters (Millipore). Filters were then frozen and transported to STRI's molecular facility at Isla Naos Laboratory in Panama City in liquid nitrogen and stored at −80 °C until DNA extractions. DNA was extracted from each filter using a Qiagen Powersoil extraction kit following the manufacturer's protocol with minor modifications to increase the yield46. Metagenomic shotgun libraries were prepared with the Illumina DNA Nextera Flex kit following the manufacturer's protocol. Shotgun metagenomics reads were sequenced on an Illumina Nextseq platform. Dissolved oxygen, temperature, salinity, and pH were measured with a pre-calibrated Professional Plus handheld YSI (Yellow Springs, USA).

Metagenome preparation, sequencing, and assembly. Trimmomatic (v0.39)[47] was used for adapter clipping and initial quality trimming of raw metagenomic data (N = 57). The program anvi'o (v7.1)[48] was used to build to build a Snakemake (v.5.10.0)[49] workflow for co-assembly analysis. In the workflow, the iu_filter_quality_minoche function from the Illumina Utils package (v2.12)[50] was used for additional quality filtering and MEGAHIT (v1.2.9)[51] for co-assembly (–min-contig-len: 1000, –presets: meta-sensitive). Three separate co-assemblies were performed using MEGAHIT based on initial assessment of the metagenomic data. All TWA samples (reef and mangrove) were co-assembled (n = 29); from the TEP, one co-assembly was performed for reef samples (n = 16) and another for mangrove samples (n = 12). Next, the anvi-gen-contigs-database was used to generate a database of contigs. Within the Snakemake workflow, KrakenUniq (v0.5.8)[52] was used for taxonomic classification of short reads against a user-constructed database of archaea, bacteria, viral, fungi, and protozoa reads from RefSeq and

the NCBI nt database. Taxonomic classification of contigs was performed using Centrifuge (v1.0.4_beta)[53], against the bacterial, archaeal, human, and viral genomes database.

Phage marker gene and contig curation. For the marker gene detection, open reading frames (ORFs) were predicted with prodigal[54] (-p meta -a -d) on contigs of all sizes (753,612 EP; 574,304 WA contigs | 2,168,906 EP; 1,756,476 WA ORFs; 3,925,382 total ORFs). Amino acid sequences of the ORFs were then searched against all MCP and TerL HMM profiles available in the Virus Orthologous Group database (vogdb.org) version 208 (Supplemental Dataset 2) using hmmsearch (hmmer.org; E value < 0.00001, bitscores > 41 and > 33, respectively, minimum length of open reading frame >=826 and >= 885 nucleotides, respectively). The threshold bitscores were determined by searching proteins predicted with prodigal (default per genome) from all *Caudovirales* genomes from Viral Genomes Portal downloaded on July 26, 2021 against the MCP and TerL profiles, taking the top hit from each genome and identifying the minimum bitscore required to include at least 98% of hits. After filtering for bitscore, the minimum length of a hit was decided based on containing at least 98% of those reference hits. This resulted in 3,749 MCP genes and 5,369 TerL genes. These were then de-replicated at 100% identity across the entire length of one sequence using BLASTn[55], which resulted in 3,722 representative MCP and 5,350 TerL (See Data Availability).

For the detection of phage contigs, contigs over 10 kilobases (7,619 EP; 10,839 WA) were analyzed with VirSorter2[56] and CheckV[57] as follows. First, contigs over 10 kilobases (EP: 7,619, WA: 10,839;) were run through VirSorter2 (virsorter run --min-score 0.5 all) and retained if they scored over 0.5 for dsDNAphage as their max_group (EP: 1,513, WA: 3,272). These contigs were then run through CheckV (checkv end_to_end) to trim potential host genomes flanking the

contigs. Trimmed provirus and virus sequences were combined and filtered for at least 10 kb (EP: 1,482, WA: 3,203). The trimmed sequences were then run through VirSorter again and retained if they scored over 0.95 or scored at least 0.5 and encoded at least 2 phage hallmark genes. This resulted in 3,885 contigs. Virus detection summary for each contig is in Supplemental Dataset 2.

Prokaryote marker gene curation. The same ORF and amino acid sequences used for the phage marker gene detection were searched against HMM profiles corresponding to genes to the Clusters of Orthologous Groups (COG) protein families of COG0012 (COG12, ribosome-binding ATP-ase), COG0085 (COG85, RNA polymerase β subunit), and COG0086 (COG86, RNA polymerase β' subunit)[20] jointly using hmmsearch (E value < 0.00001, bitscores cutoffs of 210, 200, and 200 respectively[58]. See Supplementary Dataset 4 for the number of hits of each gene.

Distribution detection. Reads from all samples were subset to an even depth to the number of reads in the sample with the fewest reads (2,992,107 reads) with seqkit[59] sample (-s 1000, -2). Reads were then mapped to an index of the phage marker genes, phage contigs, and prokaryote marker genes made with minimap2[60] -x sr. CoverM[61] (https://github.com/wwood/CoverM) was then used for the mapping (coverm contig --min-read-percent-identity 95 -m covered_fraction rpkm count variance length --minimap2-reference-is-index --min-covered-fraction 0 --coupled) and retained with 50% gene covered or 20% of contig covered[62] (Supplemental Dataset 3). See Supplemental Dataset 4 for the number of each sequence type detected in at least one sample.

Visualizations, statistical analyses, and sequence benchmarking. All plots aside from the maps were created in R (version 3.5.1)[63] with Rstudio (version 1.1.456)[64] using vegan[65], ggpubr[66], and ggplot2 (3.1.1)[67]. Maps were created with QGIS (3.24) using the Voyager plug-in for the base and overlaid with sample data. Because statistics and trends held regardless of protein examined per bacteria or phage (Supplemental Dataset 4), we focused on the TerL results to represent phage diversity and COG85 results to represent bacterial diversity, as these genes were the most prevalent in the dataset. Influential sequences and physicochemical parameters were identified by those varying the most with variation in the communities of all samples based on significant vector length (vegan package function envfit, perm=999, na.rm=TRUE; calculated with |NMDS1-NMDS2|; *p values* < 0.01). Community composition of samples were compared and visualized in non-metric dimensional scaling (NMDS) plots using Bray-Curtis distances of relative abundances calculated with reads per kilobase per million (RPKM) using vegan (metaMDS(distance = "bray")). Two outlier samples were excluded in the community compositional analyses as these were highly divergent (WAM_TWN and EPM_13A1) and skewed the results (Supplemental Fig. 5,6). WAM_TWN was sampled in a highly polluted site, and EPM_13A1 was sampled from a completely freshwater sample, which likely resulted in their aberrant community compositions at the genus-level (Fig. 3c,d). Significant distinctions between oceans and habitat types were determined with ANOSIM test (vegan package) based on Bray-Curtis dissimilarity matrices using the RPKM data (anosim(distance="bray",permutations=9999)).

Gene taxonomy. Prokaryotic sequences corresponding to COG85 were classified via Centrifuge[53]. For the phages, amino acid sequences of TerL genes were aligned to RefSeq 207

with LAST[68] (lastal -m 10 -f BlastTab; E value cutoff 10^{-5}), and the taxonomy of the hit's host was reported (i.e. a hit to a *Prochlorococcus* phage meant the taxonomy of *Prochlorococcus* was reported). The top hit was detected based on percent identity. The top 10 genera based on average relative abundance across samples was reported.

Data Availability. Reads from metagenomes will be deposited on the European Nucleotide Archive upon publication. Sequences of marker genes and phage contigs can be found on the FigShare repository upon publication, along with the VOG and COG HMM profiles used for marker gene detection.

Code Availability. Custom scripts used for this study are found in the GitHub repository (https://github.com/scubalaina/panama_phages).

Acknowledgements

We thank members of the Aylward Lab for helpful feedback. We thank the Smithsonian Tropical Research Institute staff at the Bocas del Toro and Naos stations. This work was performed using compute nodes available at the Virginia Tech Advanced Research and Computing Center and on the Smithsonian High-Performance Cluster (SI/HPC), Smithsonian Institution (doi:10.25572/SIHPC). This work was supported by grants from the Gordon and Betty Moore Foundation awarded to STRI and UC Davis (doi:10.37807/GBMF5603), the NSF CAREER award (IIBR-2141862) to FOA and a Simons Early Career Award in Marine Microbial Ecology and Evolution to FOA. ARW was supported by an ICTAS Doctoral Scholars Fellowship. Research permits were provided by the Autoridad Nacional del Ambiente de Panamá.

References

1. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).
2. Wilkins, L. G. E. et al. Host-associated microbiomes drive structure and function of marine ecosystems. *PLoS Biol.* **17**, e3000533 (2019).
3. Brussaard, C. P. D. et al. Global-scale processes with a nanoscale drive: the role of marine viruses. *ISME J.* **2**, 575–578 (2008).
4. Proctor, L. M. & Fuhrman, J. A. Viral mortality of marine bacteria and cyanobacteria. *Nature* **343**, 60–62 (1990).
5. LeGault, K. N. et al. Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. *Science* **373**, (2021).
6. Wahida, A., Tang, F. & Barr, J. J. Rethinking phage-bacteria-eukaryotic relationships and their influence on human health. *Cell Host Microbe* **29**, 681–688 (2021).
7. Sullivan, M. B., Weitz, J. S. & Wilhelm, S. Viral ecology comes of age. *Environ. Microbiol. Rep.* **9**, 33–35 (2017).
8. Aylward, F. O. & Moniruzzaman, M. Viral complexity. *Biomolecules* **12**, (2022).
9. Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E. & Ghai, R. Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9**, e1003987 (2013).
10. Luo, E., Eppley, J. M., Romano, A. E., Mende, D. R. & DeLong, E. F. Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. *ISME J.* **14**, 1304–1315 (2020).
11. Gregory, A. C. et al. Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109–1123.e14 (2019).
12. Lee, S. et al. Soil pH influences the structure of virus communities at local and global scales.

Soil Biol. Biochem. **166**, 108569 (2022).

13. Santos-Medellin, C. et al. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J.* **15**, 1956–1970 (2021).
14. Malki, K. et al. Spatial and temporal dynamics of prokaryotic and viral community assemblages in a lotic system (Manatee Springs, Florida). *Appl. Environ. Microbiol.* **87**, e0064621 (2021).
15. de Jonge, P. A., Nobrega, F. L., Brouns, S. J. J. & Dutilh, B. E. Molecular and evolutionary determinants of bacteriophage host range. *Trends Microbiol.* **27**, 51–63 (2019).
16. O'Dea, A. et al. Formation of the Isthmus of Panama. *Sci Adv* **2**, e1600883 (2016).
17. D'Croz, L. Status and uses of mangroves in the Republic of Panama. *Conservation and sustainable utilization of mangrove forests in Latin America and Africa Regions Part I. ITTO/ISME Mangrove Ecosystems Technical Reports, Okinawa Japan*. 115-127 (1993).
18. Leray, M. et al. Natural experiments and long-term monitoring are critical to understand and predict marine host–microbe ecology and evolution. *PLoS Biol.* **19**, e3001322 (2021).
19. López-Pérez, M., Haro-Moreno, J. M., Gonzalez-Serrano, R., Parras-Moltó, M. & Rodriguez-Valera, F. Genome diversity of marine phages recovered from Mediterranean metagenomes: Size matters. *PLoS Genet.* **13**, e1007018 (2017).
20. Galperin, M. Y. et al. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2021).
21. Mende, D. R. et al. Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nat Microbiol* **2**, 1367–1373 (2017).
22. Breitbart, M., Miyake, J. H. & Rohwer, F. Global distribution of nearly identical phage-

encoded DNA sequences. *FEMS Microbiol. Lett.* **236**, 249–256 (2004).

23. Logares, R. et al. Infrequent marine–freshwater transitions in the microbial world. *Trends Microbiol.* **17**, 414–422 (2009).
24. Cabello-Yeves, P. J. & Rodriguez-Valera, F. Marine-freshwater prokaryotic transitions require extensive changes in the predicted proteome. *Microbiome* **7**, 117 (2019).
25. Binetti, A. G., Quiberoni, A. & Reinheimer, J. A. Phage adsorption to *Streptococcus thermophilus*. Influence of environmental factors and characterization of cell-receptors. *Food Res. Int.* **35**, 73–83 (2002).
26. Morris, R. M. et al. SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810 (2002).
27. Fuhrman, J. A., Cram, J. A. & Needham, D. M. Marine microbial community dynamics and their ecological interpretation. *Nat. Rev. Microbiol.* **13**, 133–146 (2015).
28. Zhong, K. X., Wirth, J. F., Chan, A. M. & Suttle, C. A. Mortality by ribosomal sequencing (MoRS) provides a window into taxon-specific cell lysis. *ISME J.* (2022) doi:10.1038/s41396-022-01327-3.
29. Partensky, F. & Blanchot, J. Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters: a review. *Bull. Inst. Med. Res. Kuala Lumpur*. 457-476 (1999).
30. Vaulot, D., Partensky, F. & Neveux, J. Winter presence of prochlorophytes in surface waters of the northwestern Mediterranean Sea. *Limnology and Oceanography.* **35**, 1156-1164 (1990).
31. Shang, X., Zhang, L. H. & Zhang, J. *Prochlorococcus*-like populations detected by flow cytometry in the fresh and brackish waters of the Changjiang Estuary. *Journal of the Marine*

Biological **3**, (2007).

32. Zwart, G., et al. Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquat. Microb. Ecol*. **28**, 141-155 (2002).
33. Huang, S., Wilhelm, S. W., Harvey, H. R., Taylor, K. & Jiao, N. Novel lineages of *Prochlorococcus* and *Synechococcus* in the global oceans. *ISME J.* **6**, 285-297 (2012).
34. Kirchman, D. The ecology of Cytophaga–Flavobacteria in aquatic environments. *FEMS Microbiol. Ecol.* **39**, 91–100 (2002).
35. Garneau, M. E., Vincent, W. F., Alonso-Sáez, L., Gratton, Y. & Lovejoy, C. Prokaryotic community structure and heterotrophic production in a river-influenced coastal arctic ecosystem. *Aquat. Microb. Ecol.* **42**, 27–40 (2006).
36. Waterbury, J. B. & Valois, F. W. Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophages abundant in seawater. *Appl. Environ. Microbiol.* **59**, 3393–3399 (1993).
37. Hussain, F. A. et al. Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science* **374**, 488–492 (2021).
38. Ahlgren, N. A., Perelman, J. N., Yeh, Y.-C. & Fuhrman, J. A. Multi-year dynamics of fine-scale marine cyanobacterial populations are more strongly explained by phage interactions than abiotic, bottom-up factors. *Environ. Microbiol.* **21**, 2948–2963 (2019).
39. Hill, T. C. J., Walsh, K. A., Harris, J. A. & Moffett, B. F. Using ecological diversity measures with bacterial communities. *FEMS Microbiol. Ecol.* **43**, 1–11 (2003).
40. Vieira, R. P. et al. Relationships between bacterial diversity and environmental variables in a tropical marine environment, Rio de Janeiro. *Environ. Microbiol.* **10**, 189–199 (2008).
41. Guyader, S. & Burch, C. L. Optimal foraging predicts the ecology but not the evolution of

host specialization in bacteriophages. *PLoS One* **3**, e1946 (2008).

42. Walters, K. E. & Martiny, J. B. H. Alpha-, beta-, and gamma-diversity of bacteria varies across habitats. *PLoS One* **15**, e0233872 (2020).
43. Kukkaro, P. & Bamford, D. H. Virus-host interactions in environments with a wide range of ionic strengths. *Environ. Microbiol. Rep.* **1**, 71–77 (2009).
44. Thingstad, T. F. & Frede Thingstad, T. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography* **45**, 1320-1328 (2000).
45. Arkhipova, K. et al. Temporal dynamics of uncultured viruses: a new dimension in viral diversity. *ISME J.* **12**, 199–211 (2018).
46. Nguyen, B. N. et al. Environmental DNA survey captures patterns of fish and invertebrate diversity across a tropical seascape. *Sci. Rep.* **10**, 6729 (2020).
47. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
48. Eren, A. M. et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
49. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **34**, 3600 (2018).
50. Eren, A. M., Vineis, J. H., Morrison, H. G. & Sogin, M. L. A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS One* **8**, e66643 (2013).
51. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

52. Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* **19**, 198 (2018).
53. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
54. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
55. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
56. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
57. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
58. Martinez-Gutierrez, C. A. & Aylward, F. O. Phylogenetic signal, congruence, and uncertainty across bacteria and archaea. *Mol. Biol. Evol.* **38**, 5514–5527 (2021).
59. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, e0163962 (2016).
60. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
61. Woodcroft, B. J. *CoverM: Read coverage calculator for metagenomics*. https://github.com/wwood/CoverM
62. Weinheimer, A. R. & Aylward, F. O. Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages. *ISME J.* **16**, 1657–1667 (2022).
63. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for

Statistical Computing, Vienna, Austria (2019).

64. RStudio. https://rstudio.com. Accessed 12 Oct 2021.
65. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
66. Kassambara, A. 'ggplot2' based publication ready plots [R package ggpubr version 0.4.0]. (2020). https://cran.r-project.org/package=ggpubr.
67. Wickham, H. *et al.* ggplot2: elegant graphics for data analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org (2016).
68. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).

Figures

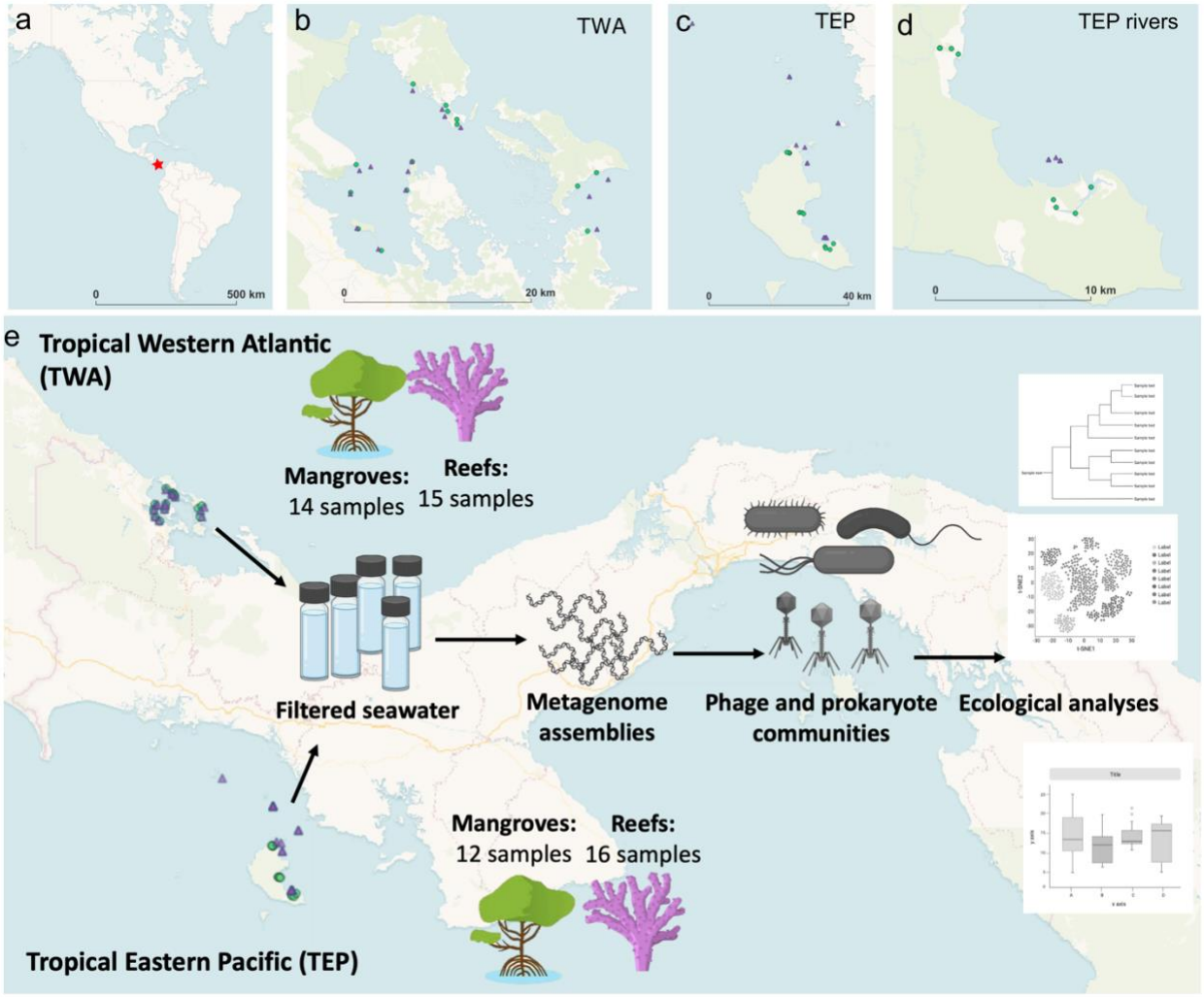


Figure 1. Overview of project design with maps of sample locations. (a) World map with Panama denoted as red star. (b) Map of sample sites from the Tropical Western Atlantic (TWA) coast of Panama. (c) Map of sample sites from Tropical Eastern Pacific (TEP) coast of Panama. (f) Map of TEP mangrove samples zoomed in on those collected along two freshwater rivers and the nearby reef samples. (e) Graphical abstract of project approaches. *Green triangles are mangrove samples. Pink circles are reef samples.*

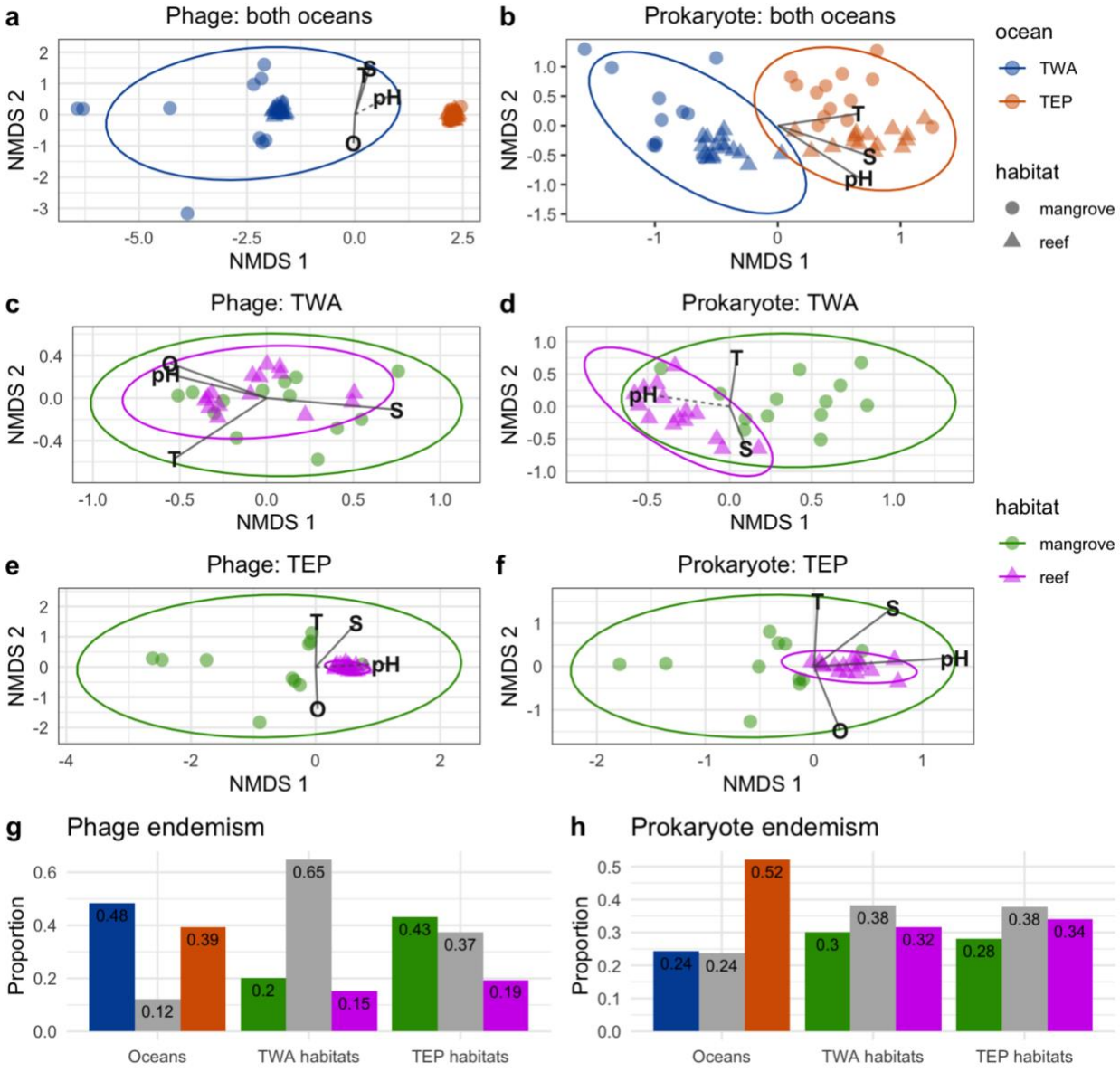


Figure 2. Comparisons of phage and prokaryotic community composition and endemism in marine habitats of the Tropical Western Atlantic (TWA) and Tropical Eastern Pacific (TEP). NMDS plots of samples based on phage or prokaryote community composition (Bray-Curtis distance), overlaid with environmental parameters significantly varying with community variation. Solid lines correspond to p-values below 0.01 and dashed below 0.05. Bottom barcharts compare the proportion of phages endemic to an environment or shared between them (in gray). (O - dissolved oxygen, S - salinity, T - temperature). Salinity represents total dissolved

solids and specific conductivity, as these variables directly correlated with each other in this dataset.

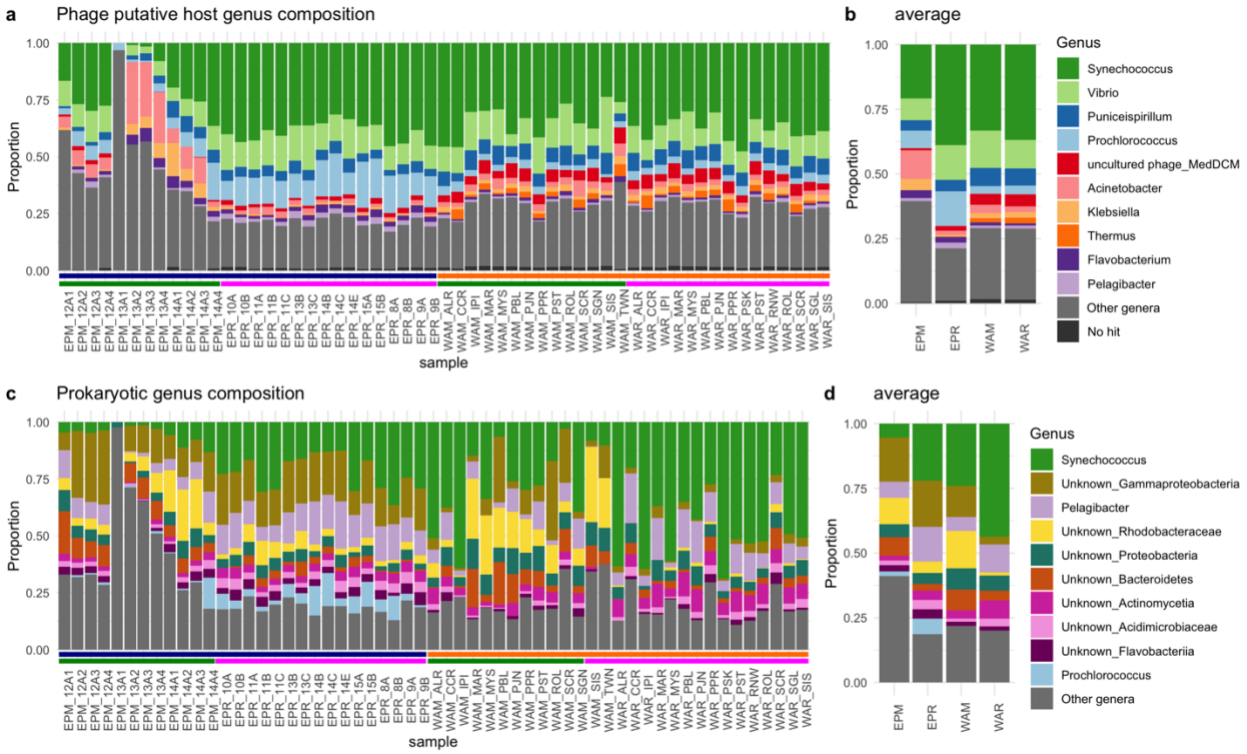


Figure 3. Genus composition of phage putative host communities and prokaryote communities. Stacked barplots of the phage putative host communities (a) or prokaryotic communities (c) colored and sorted by top ten average most abundant genera. Color strips on bottom indicate ocean and habitat where samples were collected: top row: navy = TEP; orange = TWA; green = mangroves; pink = reefs. (b,d) the average putative host genera composition (b) or prokaryotic genera composition (c) of ocean and habitat type combination EPM = TEP mangrove, ERP = TEP reef, WAM = TWA mangrove, WAR = TWA reef.

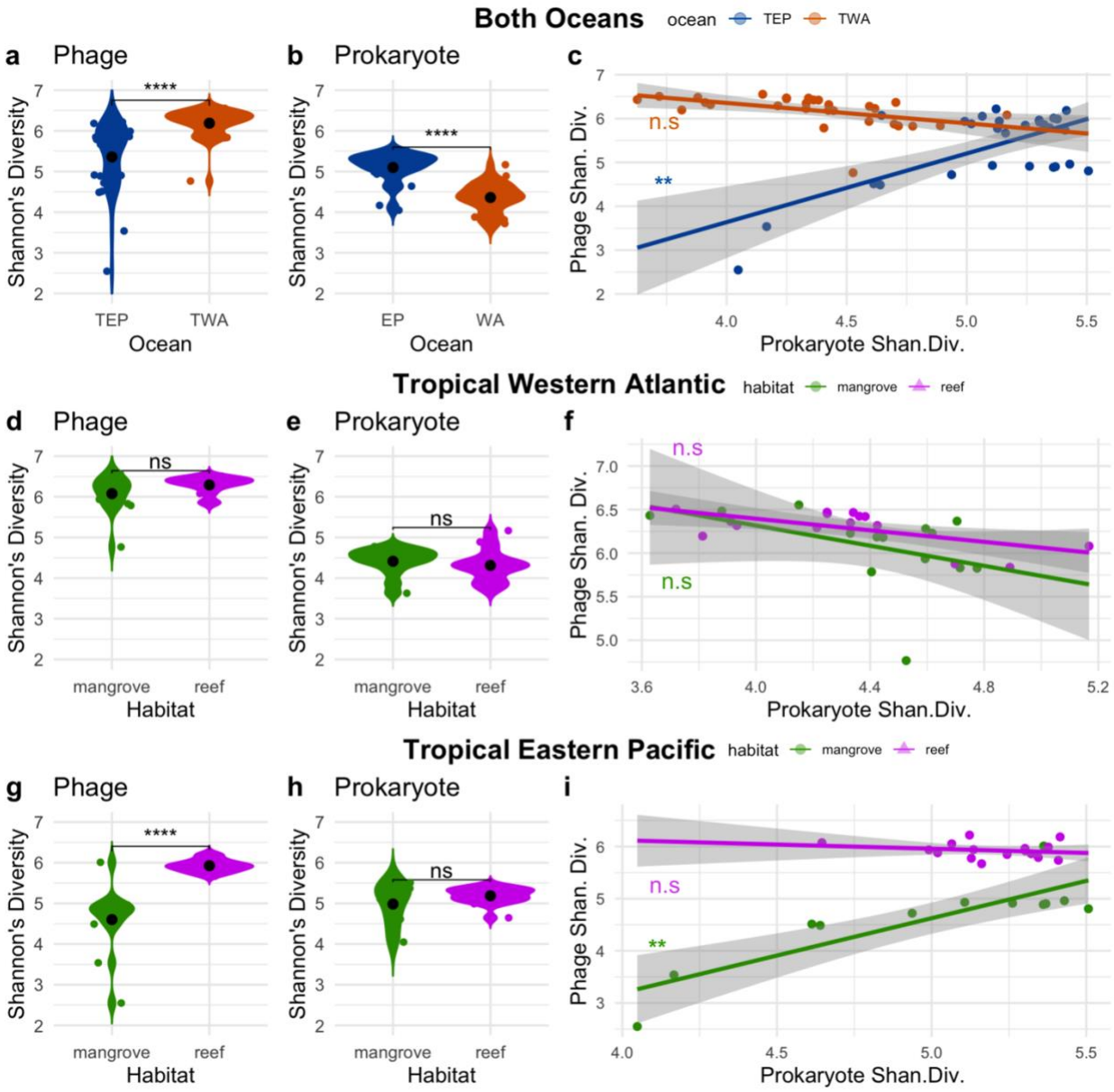


Figure 4. Phage and prokaryotic Shannon's Diversity in marine habitats of the Tropical Western Atlantic and the Tropical Eastern Pacific. 2a,b,d,e,g,h are violin plots of Shannon's Diversity of phages and prokaryotes in different samples. 2c,f,i are scatterplots of phage Shannon's Diversity plotted against prokaryotic Shannon's Diversity in a sample, with linear regression lines drawn and standard deviations shaded in gray. (Shan. Div. = Shannon's Diversity).

Supplementary Information

Supplemental Datasets can be found at the GitHub link:

https://github.com/scubalaina/panama_phages

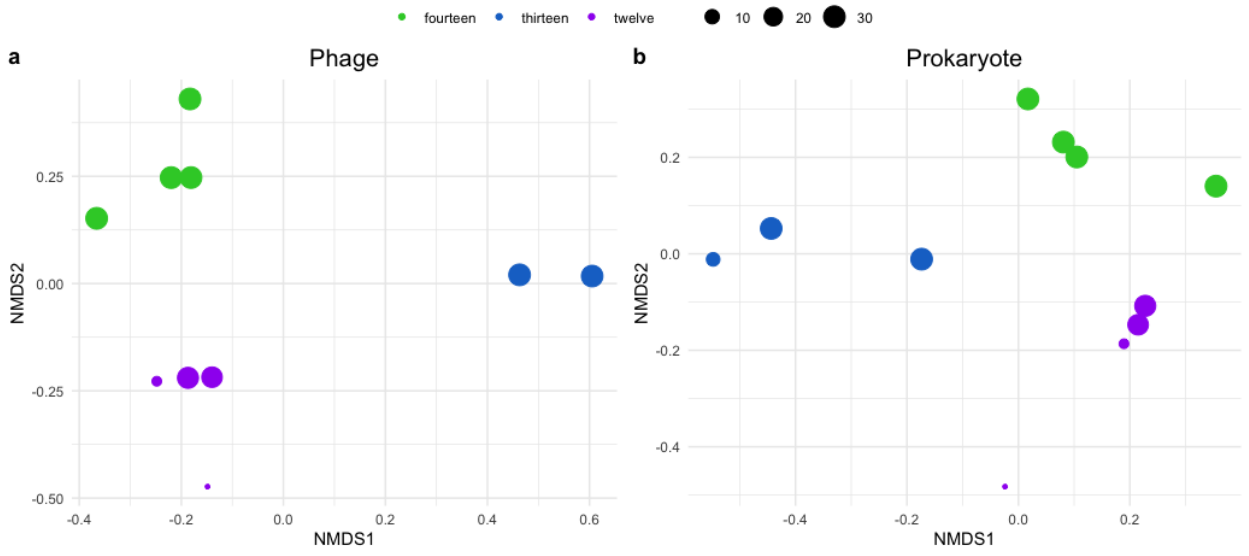
Supplemental Dataset 1. Information on each sample's coordinates, physicochemical parameters, ocean, habitat type, and diversity of sequences detected, among other features. Used sample metadata for all figures. Used diversity data for Figure 4.

Supplemental Dataset 2. Information of the VOG profiles used to detect the phage marker sequences and the virus detection tool outputs.

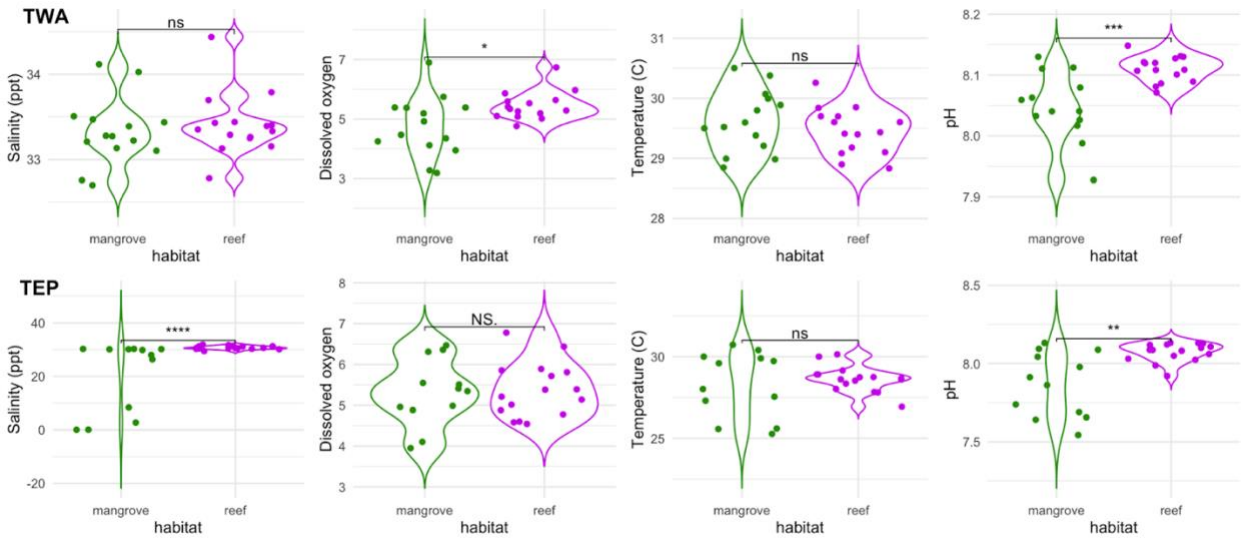
Supplemental Dataset 3. RPKM tables of reads mapped from each sample for the phage and prokaryotic sequences. Used for Figure 2.

Supplemental Dataset 4. Information on the number of each sequence present in a file and the ecological statistics reported and the manuscript (ANOSIMs, Mantel tests, correlations) used to benchmark the sequences.

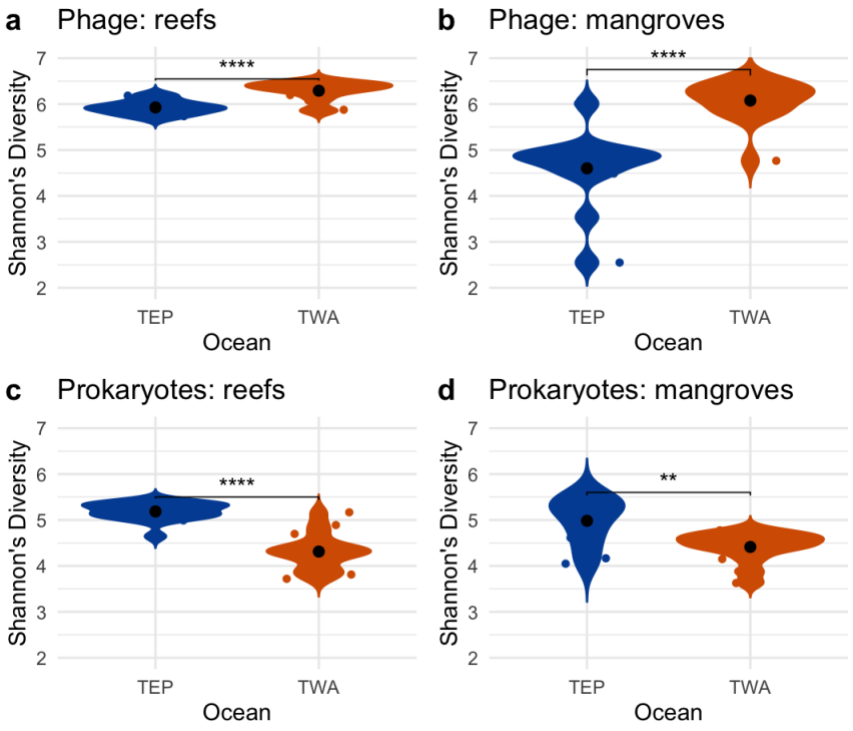
Supplemental Dataset 5. Sequence information for TerL and RNAP B on which samples they were found, their classifications, and their envfit test results. Used for Figure 3.



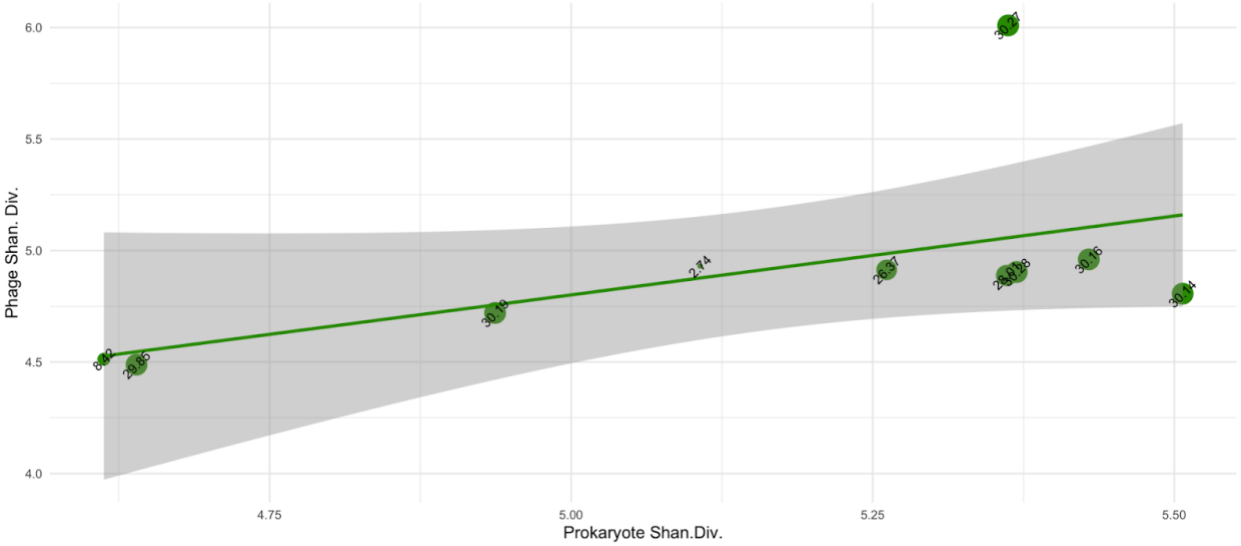
Supplemental Figure 1. NMDS plot of samples based on bray-curtis distances of phage (**a**) and prokaryote (**b**) relative abundance with data points colored by river and size of point scaled to salinity (ppt).



Supplemental Figure 2. Violin plots of the physicochemical parameters measured for the WA samples (top row) and EP samples (bottom row) between mangroves and reefs. Stars correspond to significance (*=0.05,**=0.01,***=0.001, ns or NS = not significant)

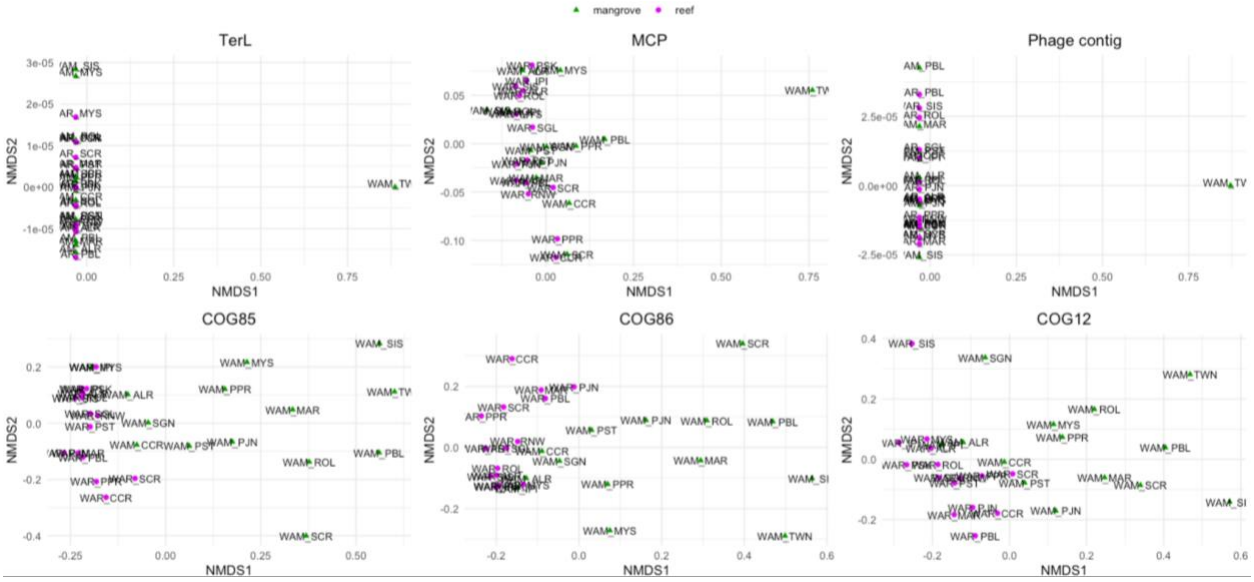


Supplemental Figure 3. Shannon's Diversity of phages (a,b) and prokaryotes (c,d) in reefs (a,c) and mangrove (b,d) samples examined separately. Stars correspond to significance (*=0.05,**=0.01,***=0.001, ns or NS = not significant).

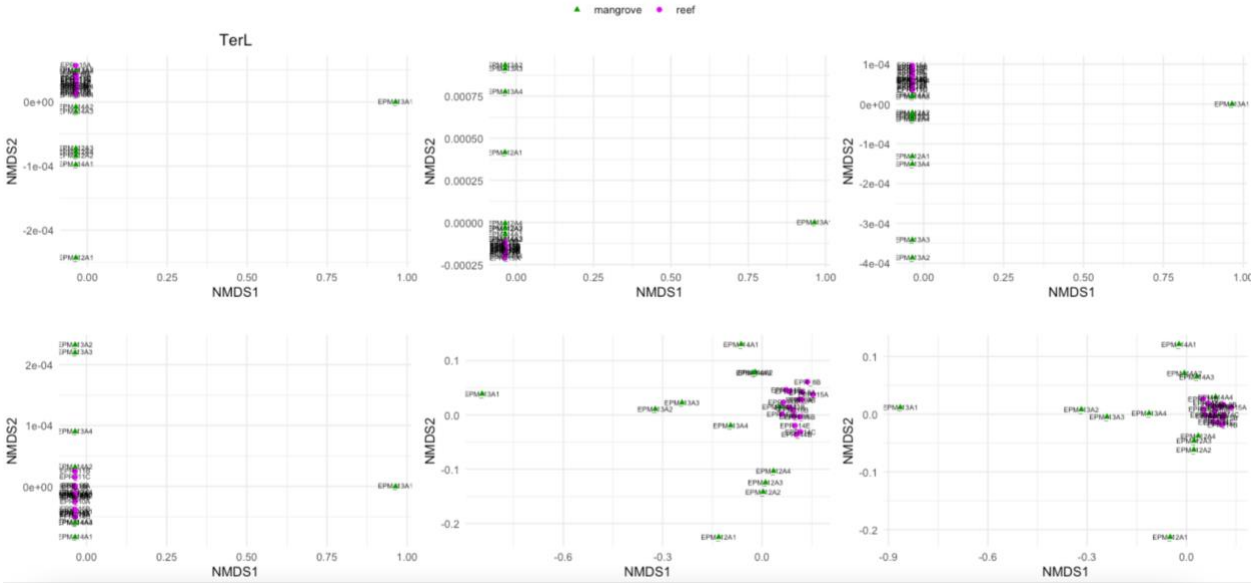


Supplemental Figure 4. Phage Shannon's Diversity (Shan. Div.) versus prokaryotic Shannon's Diversity of TEP mangrove samples with the two freshest samples (EPM_13A1 and

EPM_12A1) removed. No significant correlation. Text labels refer to salinity (ppt). Point sizes correspond to salinity.



Supplemental Figure 5. NMDS plots of TWA samples based on phage sequences (top row) or prokaryotic sequences (bottom row) color by habitat type that included WAM_TWN.



Supplemental Figure 6. NMDS plots of TEP samples based on phage sequences (top row) or prokaryotic sequences (bottom row) color by habitat type that included EPM_13A1.

Summary and Outlook

The discovery of remarkably complex viruses in recent decades has begged questions on how and why such complexity emerges. In this dissertation, I examine sequence data toward understanding the evolution, diversity, and ecology of genomic and community-level complexity among bacteriophages. One group of complex phages, called jumbo phages, are a major focus of two chapters (Chapters 1 and 3). These phages are defined as having genomes over 200 kilobases. The literature review of Chapter 1 aimed to address why genomic complexity emerges among phages by providing an overview of known jumbo phages, associated fitness tradeoffs, and hypotheses on where jumbo phages may be enriched based on these tradeoffs. Chapter 2 examined how and when complex phages emerged by investigating the evolutionary history of a group of phages that uniquely encode an RNA polymerase homologous to that of cells. By including phages in the Tree of Life with this gene, we find that this group of phages likely emerged alongside the divergence of archaea and bacteria, suggesting that complexity is an ancient strategy. Chapter 3 examined where and what types of jumbo phages are found throughout the ocean by employing a novel bioinformatic pipeline to detect jumbo phages in seawater metagenomes. This study revealed that jumbo phage groups distinguished by diverse replication strategies also have different biogeographies. For instance, one group that encodes photosynthesis machinery resembling T4 phages are enriched in surface waters, while a group that encodes proteins that form nucleus-like structures to protect their replication from host cells like that of PhiKZ phages are enriched in deeper waters, which shows that different routes to complexity contain distinct ecological niches. Finally, Chapter 4 aimed to uncover drivers of phage community complexity by characterizing phage and prokaryotic communities off the coasts of Panama to compare several environments at once. This work revealed phage diversity is rarely correlated with prokaryotic diversity, but rather potentially with microbial

productivity. Collectively, these chapters provide foundational information on the why, how, what, and where genomic and community-level phage complexity emerges and persists.

Future work that would greatly further our understanding of viral complexity includes the development of methods that target jumbo phages. Presently, many methods in virology favor smaller particles, such as filter pore sizes, agar concentrations, and bioinformatic pipelines. The development of approaches to enrich for larger viral particles from environmental samples for sequencing, microscopy, and culturing could include the use of larger filter pore sizes (e.g. 0.45 μm) followed by density gradient separation to exclude small cells. Furthermore, improvements in long-read sequencing could greatly enhance our ability to reconstruct larger phage genomes, particularly as many may encode repeats or hypervariable regions that complicate current short-read assemblies. By improving our characterization of jumbo phages from nature, we can both identify novel and potentially biotechnological useful genes encoded by these large phages, as well as improve our understanding on why and how these complex phages emerged.

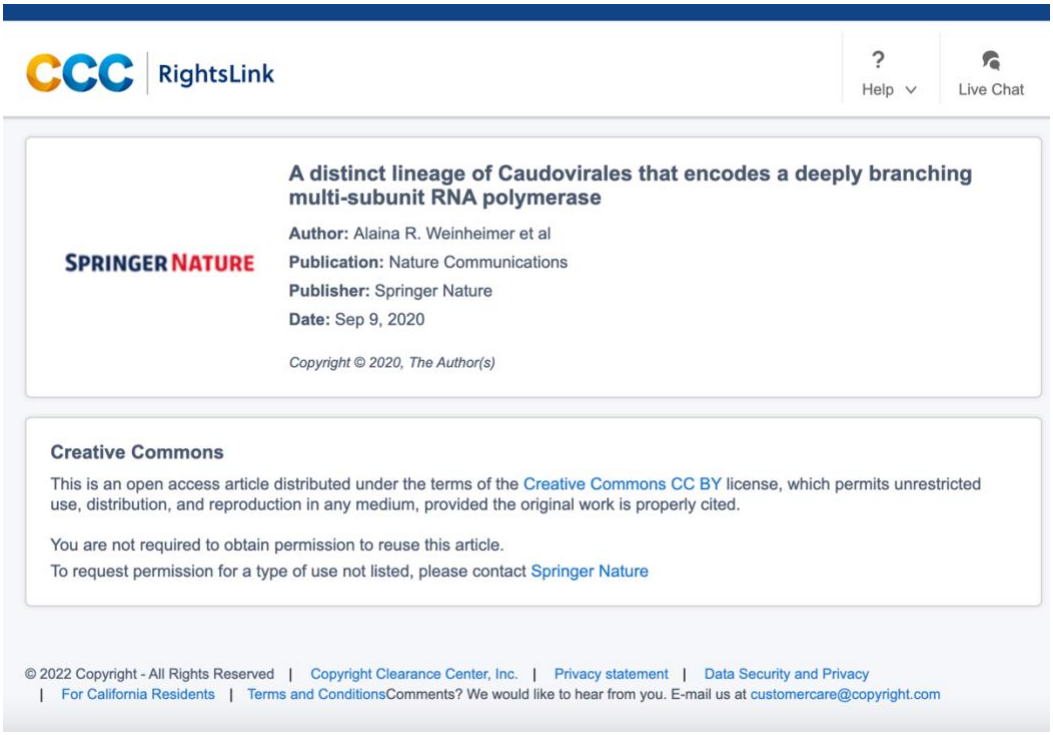
More generally, the study of phage evolution and diversity would be greatly enhanced with a clearer taxonomic framework. Within the time of my writing this dissertation the longstanding viral order of the *Caudovirales* and its families *Myoviridae*, *Podoviridae*, and *Siphoviridae* have been disbanded. Tailed, dsDNA phages are now classified under that class Caudoviricetes, and new approaches to defining families have been proposed that involve both genomic distances and phylogenies of selected hallmark genes. In any case, there still lacks a unifying backbone to phage classification. An ongoing project of myself and Dr. Aylward aims to provide this unifying phylogeny by grouping phages based on shared genes that do not involve de novo clustering like current methods. We hope this work reveals defining features of both family-level and high order groupings of phages that could enable biologically meaningful

classification of novel phages, particularly those with incomplete genomes which are frequently recovered from current sequence data.

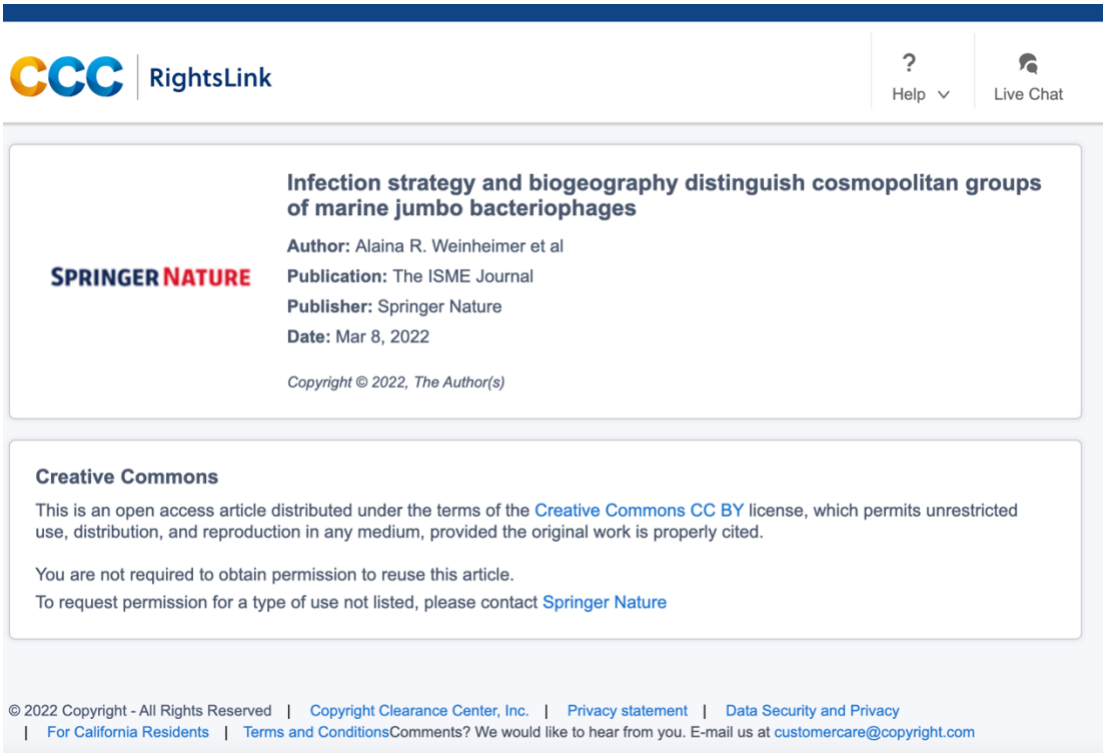
The inclusion of large, complex phages in our characterization of phage communities and in culture-based studies, along with a unifying phage taxonomy, will greatly enhance our understanding of the breadth and diversity of phages in the biosphere. The past decades have led to a renaissance in phage biology, as culture-independent methods have resulted in the discovery of the ubiquity and abundance of viruses in the environment. I suspect the exploration of the untapped diversity of phages will lead to some of the major breakthroughs in modern biology, particularly the complex viruses that have expanded and challenged the definition of life.

Appendix

A.1. Chapter 2 Publisher Copyright



A.2. Chapter 3 Publisher Copyright



RightsLink

Help

Live Chat

Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages

SPRINGER NATURE

Author: Alaina R. Weinheimer et al

Publication: The ISME Journal

Publisher: Springer Nature

Date: Mar 8, 2022

Copyright © 2022, The Author(s)

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact Springer Nature