Virus-driven evolution of marine *Vibrio*

by

Fatima Aysha Hussain

S.B. Environmental Engineering Science
Massachusetts Institute of Technology (2011)
M.S. Civil and Environmental Engineering
Leland Stanford Junior University (2013)

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Environmental Microbiology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author .....
Department of Civil and Environmental Engineering
May 8, 2020

Certified by .....
Martin F. Polz
Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by .....
Colette L. Heald
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Virus-driven evolution of marine *Vibrio*

by

Fatima Aysha Hussain

Submitted to the Department of Civil and Environmental Engineering
on May 8, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Environmental Microbiology

Abstract

Microorganisms are the most numerous and diverse organisms on the planet and occupy virtually every known habitat. For these microbes, genotypic diversity is intimately linked to their ecology. As one of the main predators of bacteria, bacteriophages (phages) play an important ecological role in regulating the abundance and diversity of bacterial populations. Through highly specific predatory interactions, these bacteria-infecting viruses promote gene turnover in their hosts through frequency-dependent selection. As a result, phages are thought to be key drivers of the immense genetic diversity seen in microbial genomes. However, the impact of phages on bacterial diversification in the wild is poorly understood. This thesis examines virus-driven evolution of environmental microbes using bacteria of the *Vibrio* genus as a model. By isolating and sequencing the genomes of sympatric *Vibrio* strains and the viruses that infect them, we created a unique system to understand how viruses are impacting bacterial genomic evolution in nature. In the first study, we investigated the diversity and dynamics of lysogenic viruses, which integrate into the host genome as prophages, across *Vibrio* populations. By combining comparative genomics and lab-based inductions of lysogenic viruses from natural bacterial strains, we isolated numerous excisable prophages and mobile genetic elements, and found that transfer of prophages is more frequent among related hosts. In the second study, we investigated the evolution of resistance to viruses in bacteria at the resolution of clones. We found that viruses drive the rapid evolutionary turnover of novel phage-defense elements in bacteria, making them one of the strongest, if not the strongest, forces for near-term microbial evolution. Finally, we explored the abundance, diversity, and transfer dynamics of a particular set of lysogenic viruses, related to the newly-discovered *Autolykiviridae*, in *Vibrio*. Together, this work sheds light on the rapid diversification of microbial genomes attributed to viruses and provides an ecologically-grounded perspective on the implications and applications of virus- and microbe-based therapies for environmental and human use.

Thesis Supervisor: Martin F. Polz
Title: Professor of Civil and Environmental Engineering

This doctoral thesis has been examined by a Committee of the Department of Civil and Environmental Engineering as follows:

Professor Sallie (Penny) Chisholm .....
Chair, Thesis Committee
Professor of Civil and Environmental Engineering

Professor Martin F. Polz .....
Thesis Supervisor
Professor of Civil and Environmental Engineering

Professor Otto X. Cordero .....
Member, Thesis Committee
Associate Professor of Civil and Environmental Engineering

Professor Tami Lieberman .....
Member, Thesis Committee
Assistant Professor of Civil and Environmental Engineering

Acknowledgments

To my advisor, Martin Polz, thank you – for your insight, guidance, support, and trust. I am in awe of your ability to simplify the complexity of this world into tangible and meaningful scientific questions. I will forever be grateful to have learned to think like a Polz Lab member.

I extend thanks to my thesis committee: Penny Chisholm, Otto Cordero, Tami Lieberman, and Peter Weigele. My committee meetings were always the perfect balance of challenging and supportive. Thank you for being my biggest fans and harshest critics. I am a better scientist because of you.

In addition to serving as my committee chair, Penny was my undergraduate academic advisor and I owe her thanks for raising me as a scientist. She has always encouraged me to lead with my heart and unapologetically pursue my passions. Most importantly, and both explicitly and by example, she continues to teach me this simple yet critical fact: communication is everything.

In addition to serving on my committee, Otto was my UROP supervisor. Working with Otto made me fall in love with microbial ecology. He taught me how to ask scientific questions – combining creativity, scientific curiosity, and logic. He has been a dear friend and devout cheerleader. Thank you for everything, Otto, especially for bullying me into applying for my first conference talk. It was exactly the push I needed to build confidence in myself and my work mid-PhD.

I want to extend additional thanks to Eric Alm and Daniel Rothman for their engaging lectures, counsel, and espresso, which helped shape my thinking and enriched my graduate school experience.

Many other individuals have contributed significantly to this work and to my development as a scientist. I especially thank Kathryn Kauffman (K2) – my phage guru who has had the greatest influence on my experimental design style, Mikayla Murphy (M2) – an undergraduate wise beyond her years and my right-hand scientist in mass plating and swift coding efforts, and Michael Cutler – our lab dad who has continued to cheer us all on from the comforts of retired life in Western Massachusetts. My lab brothers: Phil Arevalo, Dave VanInsberghe, and Joseph Elsherbini were always generous with their time, helping me to sharpen my computational skills. Towards the end, I am extra indebted to Joseph for holding my hand across the finish line.

I am grateful to Javier Dubert and Fabiola Miranda Sanchez for joining the lab when they did. They taught me how to rein in my projects and reminded me how much fun science can be with a work hard/play hard mentality. Javier raised the work presented in Chapter 3 to a higher level with his genetic talents and Fabiola was my partner in mentoring undergrads during the heroic prophage isolation efforts which are a part of Chapter 2.

To the rest of Polz Lab, past and present: Diana Chien, Manoshi Datta, Heidi Li, Joy Yang, Chris Corzett, Annie Yu, Stefan Thiele, Jan-Hendrik Hehemann, Nate Cermak, Bruno Kotska, Estelle Clerc, Natalie Woods, Kerrin Steensen, Hayley Gadol, Hannah Gavin, Annika Gomez, Denise Stewart, and Ryan Guillemette – thank you for your friendship and you feedback on anything and everything along the way.

I am indebted to my mentors, collaborators, and colleagues for their expert advice, time, and support: Phil Gschwend, Heidi Nepf, Jesse Kroll, Mick Follows, Dianne Newman, George O'Toole, John MacFarlane, Libusha Kelly, Vanja Klepac-Ceraj, Frédérique Le Roux, Kyle Costa, Hans Wildschutte, Kyle Peet, Steve Biller, Gabriel Leventhal, Jesse Shapiro, Peter Chen, Allison Coe, Thomas Hackl, Keven Dooley, Raphael Laurenceau, Jessie Berta-Thompson, Ali Ebrahimi, Leonora Bittleston, Shaul Pollak, Avery Normandin, and B.B. Cael.

And I have deep appreciation for the entire Parsons and CEE community for their dedication to environmental science, education, outreach, and activism – especially for the Parsons and CEE backbone (past and present) who make it all happen: Vicki Murphy, Eileen Covey, Kris Kipp, Sheila Frankel, Jim Long, Markus Buehler, and Kiley Clapper.

Finally, to those who have stood by me through it all. Sean Kearney and Alison Takemura have showered me with love and believed in me with unwavering confidence from the very beginning of this journey. The Maseeh house team and the residents of M4 have gifted me a home away from home, defined by fierce loyalty and extreme silliness. My forever friends: Aaron Thom, Adam Bockelie, Deepa Rao, Emily Kloc, Kelly Rockwell, Mariel Rubin, Marilu Corona, Mary Kate Healey, Lily Berger, Sara Barnowski, Tracey Hayse, and David Blank, who help me keep life in perspective. (David in particular does a phenomenal job – reminding me that taking "squiggles" so seriously is quite absurd.) And, of course, my family. My parents, Ajaz and Naaz, who have always put my education first and instilled in me a love of school and learning. My dearest Dadi, Zubeida, who raised me to fight for what I believe in (and cook for my people with that same level of devotion). And my siblings and their families: Tahseen, Rashid, Waseem, Sana, Aleeza, and Zaid, who make sure home is always a vibrant and exciting place. I love, appreciate, and admire you all so much.

***

This research was funded by the National Science Foundation's Biological Oceanography Division and the Simons Foundation, as well as the National Science Foundation Graduate Research Fellowship Program and the Martin Family Society of Fellows for Sustainability. Additional funding was provided by the Abdul Latif Jameel World Water and Food Security Lab and the Undergraduate Research Fellowship Program at MIT.

Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

Foreword

I defended my thesis on Friday the 13th of March, 2020 in the midst of an unprecedented global pandemic. Viruses, at least the kinds that infect humans, were at the forefront of our collective consciousness. Never in my lifetime had these entities, smaller than the wavelength of visible light, wrought havoc at such a scale: causing severe illness and death, making us retreat into our homes, and halting the global economy. It felt as though, overnight, the study of virus evolution had transformed from an esoteric topic into one with urgent public health implications. Although the work I present here focuses on bacterial viruses, this experience stands as a stark reminder that all organisms are subject to the enigmatic, and often unpredictable, whims of virus evolution.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

Microbes are the most numerous and diverse organisms on the planet, where they carry out processes that have shaped the history of life since their emergence. Assemblages of microbes underlie activities ranging from immunity and nutrient absorption for animal hosts (Rooks and Garrett, 2016), to biogeochemical cycles driving the flow of nutrients in the global oceans (Azam et al., 1994). Understanding and, to an extent, exploiting the diverse functions of microbes has implications for food and energy production, agriculture, and human health. However, our ability to predict and manipulate microbes for broader societal applications hinges upon understanding how microbes live and evolve in complex, dynamic ecosystems.

An important part of the microbial environment is viral predation. Viruses are present in all ecosystems, and bacteria's viral predators – bacteriophages (phages) – often outnumber their prey by an order of magnitude (Wommack et al., 2015). In the coastal ocean, every milliliter of seawater is home to 10^6 bacteria and 10^7 viruses, most of which are thought to be phages (Wommack et al., 2015). By lysing their hosts, phages turn over an estimated 20% of the bacterial biomass in the ocean per day (Suttle, 2007), ultimately impacting the global carbon cycle. Not only are phages abundant, they are also incredibly diverse (Paez-Espino et al., 2016; Roux et al., 2016). And through specific interactions with their hosts, phages influence the diversity and dynamics of microbial populations, shifting community composition and driving evolution by leaving lasting impacts on the genomes of microbes (Stern and

Sorek, 2011).

By studying the dynamics of virus-host interactions in wild populations of microbes, we can begin to uncover the role of these interactions in shaping trajectories of microbial evolution. In this introduction, I review the nature of bacteriophage infections and discuss the importance of a combined ecological and evolutionary approach for understanding the influence of phages on microbial genomic diversity. I describe the significance and benefits of studying marine *Vibrio* communities and outline the unique model system used in this work to investigate microbial evolution in nature. Finally, I outline the aims of each thesis chapter that follows.

1.1 How do phages infect bacteria?

To understand how viruses affect microbial evolution, we must start with how viruses infect bacteria. Phages are passive particles that rely on random diffusion to encounter their hosts. Outer membrane proteins and appendages on the cell surface such as flagella, pili, and lipopolysaccharides serve as common attachment sites for phages (Silva et al., 2016). Binding to appropriate receptors triggers injection of the virus's genetic material into the cell. Upon entry, phage infection primarily takes one of two avenues towards reproduction – lysis or lysogeny (Ackermann and DuBow, 1987; Weinbauer, 2004). Lytic viruses immediately redirect host metabolism to produce progeny phages that are released upon cell lysis (Young, 1992). Lysogenic phages, on the other hand, take a more patient approach; they integrate their DNA into the host genome, becoming prophages and linking their reproduction to that of their host's (Ackermann and DuBow, 1987; Lwoff, 1953). In some cases, lysogenic phages persist in bacteria as plasmids (Mobberley et al., 2008; Ravin, 2011; Signer, 1969). Prophages remain integrated until they are induced, either naturally or as a result of a specific signal like DNA damage, and begin replicating. Induced lysogenic viruses behave as lytic viruses, initiating transcription and translation of viral genes using the host's machinery. In all infections, once viruses are assembled, the cell bursts, releasing the new generation of viruses into the environment to encounter subsequent

hosts and begin the cycle again.

In the environment, high nutrient conditions, under which hosts are rapidly growing, tend to favor lytic viral infections. In contrast, nutrient-depleted and high host density conditions are thought to bias towards lysogeny (Knowles et al., 2016). For example, in the ocean, the fraction of the total virus community consisting of lytic or lysogenic viruses can vary (Paul, 2008), but in productive environments such as coastal water, lytic phages are thought to typically dominate (Paul, 2008; Wilcox and Fuhrman, 1994). Lysogeny has been described as a form of mutualism because it both provides a survival strategy for phages living at low host densities and can confer selective advantages to hosts by providing novel functions including virulence, antibiotic resistance, and defense against further phage infection (Canchaya et al., 2003; Paul, 2008). While many viruses are purely lytic, many appear to be capable of exploiting either lysis or lysogeny, making viruses both predators and symbionts of bacteria.

1.2 How do bacteria defend against phages? And what is the viral response?

Viral predation selects for resistance in bacterial hosts. Bacteria have evolved to combat viral attack at each step of the infection cycle. Hosts can prevent phage adsorption by modifying surface receptors, obstructing receptors through the production of an extracellular matrix, or blocking receptors with competitive inhibitors (Labrie et al., 2010; Samson et al., 2013) . In a co-evolutionary experiment with phage λ and *E. coli* cells, single nucleotide polymorphism (SNP) accumulation in receptor genes and receptor gene loss both gave rise to host resistance (Meyer et al., 2012). Similar results have been seen across the phylogenetic tree of bacterial hosts. For example, in *Prochlorococcus*, both allelic changes and gene content diversity of viral-attachment genes in genomic islands have been shown to yield resistance to specific viruses. Such changes in the environment could effectively reduce the size of susceptible bacterial

populations and allow for coexistence of both predator and prey (Avrani et al., 2011).

For phages that manage to attach to a host and inject their DNA into the cell, bacteria have evolved a wide variety of resistance mechanisms to impede infection (Dy et al., 2014; Hampton et al., 2020; Labrie et al., 2010; Seed, 2015). To inhibit phage DNA replication, hosts can selectively degrade foreign DNA using CRISPR-Cas or restriction-modification (R-M) systems. Present in microbial genomes as cassettes of clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) genes, CRISPR-Cas systems provide hosts with an adaptive immunity (Wiedenheft et al., 2012). Upon surviving a phage infection, hosts acquire a short DNA fragment from the invading virus, which is used for sequence-specific degradation of closely related phage DNA during subsequent infections. CRISPR-Cas systems, however, are not only present in hosts; a phage-encoded CRISPR-Cas system used to combat *Vibrio cholerae* adaptive immunity was recently discovered (Seed et al., 2013). Hosts can also target incoming phage DNA using R-M systems that encode for restriction enzymes, which selectively degrade un-methylated phage DNA (Labrie et al., 2010; Samson et al., 2013). Viruses may evade degradation if the host methylase can outcompete the restriction enzyme and protect the phage DNA though methylation, resulting in a host-induced modification of the phage host range (Luria, 1953). In this case, the surviving progeny viruses are immune to analogous R-M systems in hosts harboring the same system. RM-based defense is often imperfect as the methylase enzyme can accidentally modify the DNA of the virus, effectively protecting it and allowing for infection. In addition, many phages, including coliphage T4, encode for proteins that counter or interfere with restriction enzymes (Labrie et al., 2010). In recent years, our knowledge of phage defense mechanisms has expanded greatly. Phage inducible chromosomal islands (PICIs) have been shown to act as defense mechanisms in *Vibrio cholerae* (Seed et al., 2013). The increasing amount of available genomic data has also allowed for computational searches to hypothesize and test novel putative phage defense systems (Doron et al., 2018). Finally, even where a phage is able to complete the infection cycle and kill its host, the cell may still be able to inhibit the infection from spreading to sister strains. At

the population level bacteria can acquire resistance through abortive infection (Abi) mechanisms where the host's targeted inhibition of lytic propagation results in the death of the infected host, thus allowing nearby uninfected hosts to survive (Labrie et al., 2010; Samson et al., 2013). Abi mechanisms constitute single protein clusters, often encoded by prophages, which are used to target phage protein-DNA complexes. Virulent phages can escape Abi mechanisms through point mutations and selection or, in the case of Lactococcus spp., by exchanging large portions of their genome with inducible prophages through homologous or illegitimate recombination (Labrie and Moineau, 2007; Samson et al., 2013). While the co-evolution of bacteria and phages results in reciprocal adaptations and evolution, the focus of this thesis is on bacterial evolution in response to phages. Phage counter defense strategies are an important implication of the work presented here and are highlighted and discussed in (Koonin and Krupovic, 2020; Pawluk et al., 2018; Samson et al., 2013).

1.3 How do phages influence bacterial population dynamics?

As lytic predators, viruses drive the evolution of resistance mechanisms in bacteria, which in turn select for reciprocal escape adaptations in phages (Bohannan and Lenski, 2000; Labrie et al., 2010; Lindell et al., 2007; Samson et al., 2013; Woolhouse et al., 2002). The result of this co-evolutionary arms race is an increase in genetic diversity of bacteria and phage populations, often described as the Red Queen hypothesis (Avrani et al., 2012; Betts et al., 2018). This paradigm can be observed in the genomes of bacteria, as many of the genes that occur at low frequency within species are hypothesized to be associated with phage recognition sites (Cordero and Polz, 2014; Rodriguez-Valera et al., 2009). Phage genomes also show evidence of co-evolution through the presence of functional host metabolic genes that optimize infections by redirecting host metabolism to favor virus propagation (Lindell et al., 2005, 2007; Thompson et al., 2011). In addition to driving genetic diversity through

exchange, phages are also thought to drive microbial community diversity through negative frequency-dependent selection, often described using the "Kill-the-winner" hypothesis (Thingstad, 2000; Winter et al., 2010). Assuming viral specificity and an ecological trade-off with resistance, kill-the-winner predicts viruses will selectively remove the fastest growing microbe from a community, allowing diverse strains to co-exist. While widely discussed, it does not account for the co-evolution understood to be taking place between phages and hosts, hosts with multiple unique viral predators, as well as the rapid rates of horizontal gene transfer and recombination seen in wild microbial populations. For example, Cordero and Polz hypothesized that the HGT of receptor genes can make host range of viruses a dynamic property, shifting specificity as a function of the mobile gene pool of receptors (Cordero and Polz, 2014). The host range of viruses may be a function of the frequency of specific receptor genes available in the population. How viruses drive changes in the diversity of bacteria at the community, population, or strain level is an open question which we begin to explore in Chapter 3.

1.4 Why use marine *Vibrio* to study virus-driven evolution?

Virus-host interactions in the environment are limited by encounter rates, suggesting that large, fast growing, and highly abundant host populations will be disproportionately susceptible to viral predation. While the average abundance of most bacterial taxa in the ocean is low, transient blooms of opportunistic genera can occur in response to increases in substrate availability (Teeling et al., 2012). *Vibrio* are widely recognized as one such genus in marine environments (Giovannoni et al., 2005). For example, over the course of a year-long study, the largest bacterial bloom observed was *Vibrio*-dominated, showing an increase in *Vibrio* abundance from background levels of 0-2% to 54% of the bacterial community (Teeling et al., 2012). Because of their bloom-bust dynamics, *Vibrio* have been historically used to study virus-host

interactions (Comeau et al., 2006). Additionally, the diversity of Vibriophages is generally high, with Myo-, Sipho- and Podoviridae all represented (Comeau et al., 2006), and varied in host range (Kauffman, 2014).

1.4.1 *Vibrio* as a model system for microbial evolution

Marine *Vibrio* have served as a model for the ecology and evolution of bacterial populations and are a fitting model system to overlay with viral interactions in order to investigate the underlying mechanisms driving this ecology. Work done in the Polz lab has found that *Vibrio* form cohesive population structures, defined by gene flow, and that they are ecologically and genetically diverse, with different populations showing different distributions of particle size habitats and seasonality (Arevalo et al., 2019; Hunt et al., 2008; Preheim et al., 2011; Shapiro and Polz, 2014). Speciation in *Vibrio* has also been tracked in the lab, making this the ideal collection to study the impact of viruses on ecology and evolution. We have found that gene-, rather than strain-, specific selective sweeps occur, meaning that these are highly recombining populations with the possibility for fast gene turnover (Shapiro et al., 2012). Studies of horizontal gene transfer by plasmids and episomes have also been investigated in *Vibrio* populations. Finally, *Vibrio* are genetically tractable and exhibit fast growth rates, allowing for hypothesis-testing lab experiments.

1.4.2 The Nahant Collection: a coastal ocean time-series

To investigate the impact of viral infection on bacterial evolution, this thesis takes advantage of the Nahant Collection, a coastal ocean time-series of marine *Vibrio* and co-occurring viruses. This is the largest to date culture and genome collection of co-occurring phages and hosts (Kauffman et al., 2018). Creation of the Nahant Collection was spearheaded by Kathryn Kauffman as part of her PhD work in the Polz Lab in 2010. The goal of the collection was to create a sampling scheme that captured the diversity, dynamics, and demographics of phage-host interactions using natural isolates. Samples were taken daily for 93 days, from July 24th (ordinal date 204) to

October 23rd (ordinal date 296) from Canoe Cove beach in Nahant, Massachusetts. Daily sampling included total community DNA, total community glycerol stocks, water samples, active viral concentrates, and an assortment of relevant metadata, all described in detail in her dissertation (Kauffman, 2014). On each of three days, from the beginning (ordinal date 222), middle (ordinal date 261), and end (ordinal date 286), approximately 1,000 *Vibrio* strains were isolated on selective media and approximately 480 strains from each day were tested for phage susceptibility using the viral concentrates from the respective days plated along the purified hosts in agar overlays. Hosts that were sensitive to viruses, i.e. those host lawns that exhibited active plaque formation due to specific viral killing ("plaque positive"), were then included in an assay to gauge host-range, challenging every host with every phage; the infectivity of each isolated virus from each plaque positive host was tested by plaque formation in agar overlays. The final dimension of the Nahant Collection is its genetic resolution; to date over 1,300 hosts and over 300 phage genomes have been sequenced, creating a highly resolved phage-host interaction network that can then be used to ask the questions presented in this work.

1.5 Thesis overview

The impact of both lysogenic and lytic phages on bacterial genome evolution is explored in this thesis. Two overarching questions motivated this work: What are the abundances and dynamics of lysogenic viruses in wild microbial populations? (Chapter 2), and How does lytic viral predation drive the near-term evolution of bacterial hosts? (Chapter 3).

In Chapter 2, we combine computational methods and lab-based techniques to survey over 300 genomes of marine *Vibrio* isolates for existing prophages. Using three different published prophage search algorithms on the genomes, and by inducing and sequencing prophages from the isolates, we create a dataset to determine the diversity and dynamics of excising prophages.

In Chapter 3, to determine how lytic viruses drive the genetic diversity of natural

microbial populations, we focus on two nearly clonal populations of bacteria that differ in viral predation. Using a combination of comparative genomics and molecular genetics, we uncover mechanisms of viral resistance underpinning the observed differences in viral killing.

Finally, Appendix A highlights our discovery of prophages related to the recently described *Autolykiviridae* (Appendix B) are abundant in *Vibrio* genomes and, when induced to replicate and excise from their hosts, contribute to the non-tailed majority of viruses observed in the ocean.

Chapters are formatted for publication in specific journals.

Bibliography

Ackermann, H. W. and DuBow, M. S. (1987). *Viruses of Prokaryotes, Volume 1, General Properties of Bacteriophages*. CRC Press, Boca Raton, Florida.

Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J., and Polz, M. F. (2019). A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. *Cell*, 178(4):820–834.e14.

Avrani, S., Schwartz, D. A., and Lindell, D. (2012). Virus-host swinging party in the oceans. *Mobile Genetic Elements*, 2(2):88–95.

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., and Lindell, D. (2011). Genomic island variability facilitates Prochlorococcus-virus coexistence. *Nature*, 474(7353):604–608.

Azam, F., Smith, D. C., Steward, G. F., and HagstrÃűm, Ã. (1994). Bacteria-organic matter coupling and its significance for oceanic carbon cycling. *Microbial Ecology*, 28(2):167–179.

Betts, A., Gray, C., Zelek, M., MacLean, R. C., and King, K. C. (2018). High parasite diversity accelerates host adaptation and diversification. *Science*, 360(6391):907–911.

Bohannan, B. and Lenski, R. (2000). Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. *Ecology Letters*, 3(4):362–377.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and BrÃijssow, H. (2003). Prophage Genomics. *Microbiology and Molecular Biology Reviews*, 67(2):238–276.

Comeau, A. M., Chan, A. M., and Suttle, C. A. (2006). Genetic richness of vibriophages isolated in a coastal environment. *Environmental microbiology*, 8(7):1164–1176.

Cordero, O. X. and Polz, M. F. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Reviews Microbiology*, 12(4):263–273.

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, page eaar4120.

Dy, R. L., Richter, C., Salmond, G. P., and Fineran, P. C. (2014). Remarkable Mechanisms in Microbes to Resist Phage Infections. *Annual Review of Virology*, 1(1):307–331.

Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., Bibbs, L., Eads, J., Richardson, T. H., Noordewier, M., RappÃĺ, M. S., Short, J. M., Carrington, J. C., and Mathur, E. J. (2005). Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science*, 309(5738):1242–1245.

Hampton, H. G., Watson, B. N. J., and Fineran, P. C. (2020). The arms race between bacteria and their phage foes. *Nature*, 577(7790):327–336.

Hunt, D. E., David, L. A., Gevers, D., Preheim, S. P., Alm, E. J., and Polz, M. F. (2008). Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton. *Science*, 320(5879):1081–1085.

Kauffman, A. K. M. (2014). *Demographics of Lytic Viral Infection of Coastal Ocean Vibrio*. PhD thesis, Massachusetts Institute of Technology.

Kauffman, K. M., Brown, J. M., Sharma, R. S., VanInsberghe, D., Elsherbini, J., Polz, M., and Kelly, L. (2018). Viruses of the Nahant Collection, characterization of 251 marine Vibrionaceae viruses. *Scientific Data*, 5:180114.

Knowles, B., Silveira, C. B., Bailey, B. A., Barott, K., Cantu, V. A., CobiÃąn-GÃijemes, A. G., Coutinho, F. H., Dinsdale, E. A., Felts, B., Furby, K. A., George, E. E., Green, K. T., Gregoracci, G. B., Haas, A. F., Haggerty, J. M., Hester, E. R., Hisakawa, N., Kelly, L. W., Lim, Y. W., Little, M., Luque, A., McDole-Somera, T., McNair, K., Oliveira, L. S. d., Quistad, S. D., Robinett, N. L., Sala, E., Salamon, P., Sanchez, S. E., Sandin, S., Silva, G. G. Z., Smith, J., Sullivan, C., Thompson, C., Vermeij, M. J. A., Youle, M., Young, C., Zgliczynski, B., Brainard, R., Edwards, R. A., Nulton, J., Thompson, F., and Rohwer, F. (2016). Lytic to temperate switching of viral communities. *Nature*, 531(7595):466–470.

Koonin, E. V. and Krupovic, M. (2020). Phages build anti-defence barriers. *Nature Microbiology*, 5(1):8–9. Number: 1 Publisher: Nature Publishing Group.

Labrie, S. J. and Moineau, S. (2007). Abortive Infection Mechanisms and Prophage Sequences Significantly Influence the Genetic Makeup of Emerging Lytic Lactococcal Phages. *Journal of Bacteriology*, 189(4):1482–1487.

Labrie, S. J., Samson, J. E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nature Reviews Microbiology*, 8(5):317–327.

Lindell, D., Jaffe, J. D., Coleman, M. L., Futschik, M. E., Axmann, I. M., Rector, T., Kettler, G., Sullivan, M. B., Steen, R., Hess, W. R., Church, G. M., and Chisholm, S. W. (2007). Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature*, 449(7158):83–86.

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M., and Chisholm, S. W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, 438(7064):86–89.

Luria, S. E. (1953). Host-Induced Modifications of Viruses. *Cold Spring Harbor Symposia on Quantitative Biology*, 18:237–244.

Lwoff, A. (1953). Lysogeny. *Bacteriological Reviews*, 17(4):269.

Meyer, J. R., Dobias, D. T., Weitz, J. S., Barrick, J. E., Quick, R. T., and Lenski, R. E. (2012). Repeatability and Contingency in the Evolution of a Key Innovation in Phage Lambda. *Science*, 335(6067):428–432.

Mobberley, J. M., Authement, R. N., Segall, A. M., and Paul, J. H. (2008). The Temperate Marine Phage ?HAP-1 of Halomonas aquamarina Possesses a Linear Plasmid-Like Prophage Genome. *Journal of Virology*, 82(13):6618–6630.

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., and Kyrpides, N. C. (2016). Uncovering EarthâĂŹs virome. *Nature*, 536(7617):425–430. Number: 7617 Publisher: Nature Publishing Group.

Paul, J. H. (2008). Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *The ISME Journal*, 2(6):579–589.

Pawluk, A., Davidson, A. R., and Maxwell, K. L. (2018). Anti-CRISPR: discovery, mechanism and function. *Nature Reviews Microbiology*, 16(1):12–17. Number: 1 Publisher: Nature Publishing Group.

Preheim, S. P., Boucher, Y., Wildschutte, H., David, L. A., Veneziano, D., Alm, E. J., and Polz, M. F. (2011). Metapopulation structure of Vibrionaceae among coastal marine invertebrates. *Environmental Microbiology*, 13(1):265–275.

Ravin, N. V. (2011). N15: The linear phageâĂŞplasmid. *Plasmid*, 65(2):102–109.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., PaÅąiÄĞ, L., Thingstad, T. F., Rohwer, F., and Mira, A. (2009). Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology*, 7(11):828–836.

Rooks, M. G. and Garrett, W. S. (2016). Gut microbiota, metabolites and host immunity. *Nature Reviews Immunology*, 16(6):341–352. Number: 6 Publisher: Nature Publishing Group.

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., Poulos, B. T., Solonenko, N., Lara, E., Poulain, J., Pesant, S., Kandels-Lewis, S., Dimier, C., Picheral, M., Searson, S., Cruaud, C., Alberti, A., Duarte, C. M., Gasol, J. M., VaquÃľ, D., Bork, P., Acinas, S. G., Wincker, P., and Sullivan, M. B. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 537(7622):689–693. Number: 7622 Publisher: Nature Publishing Group.

Samson, J. E., MagadÃąn, A. H., Sabri, M., and Moineau, S. (2013). Revenge of the phages: defeating bacterial defences. *Nature Reviews Microbiology*, 11(10):675–687.

Seed, K. D. (2015). Battling Phages: How Bacteria Defend against Viral Attack. *PLOS Pathogens*, 11(6):e1004847.

Seed, K. D., Lazinski, D. W., Calderwood, S. B., and Camilli, A. (2013). A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature*, 494(7438):489–491.

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., SzabÃş, G., Polz, M. F., and Alm, E. J. (2012). Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science*, 336(6077):48–51.

Shapiro, B. J. and Polz, M. F. (2014). Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology*, 22(5):235–247.

Signer, E. R. (1969). Plasmid Formation: a New Mode of Lysogeny by Phase Îż. *Nature*, 223(5202):158–160.

Silva, J. B., Storms, Z., and Sauvageau, D. (2016). Host receptors for bacteriophage adsorption. *FEMS Microbiology Letters*, page fnw002.

Stern, A. and Sorek, R. (2011). The phage-host arms race: Shaping the evolution of microbes. *BioEssays*, 33(1):43–51.

Suttle (2007). Marine viruses âĂŤ major players in the global ecosystem. *Nature Reviews Microbiology*, 5(10):801–812.

Teeling, H., Fuchs, B. M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C. M., Kassabgy, M., Huang, S., Mann, A. J., Waldmann, J., Weber, M., Klindworth, A., Otto, A., Lange, J., Bernhardt, J., Reinsch, C., Hecker, M., Peplies, J., Bockelmann, F. D., Callies, U., Gerdts, G., Wichels, A., Wiltshire, K. H., GlÃűckner, F. O., Schweder, T., and Amann, R. (2012). Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. *Science*, 336(6081):608–611.

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

Thompson, L. R., Zeng, Q., Kelly, L., Huang, K. H., Singer, A. U., Stubbe, J., and Chisholm, S. W. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences*, 108(39):E757–E764.

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiology Reviews*, 28(2):127–181.

Wiedenheft, B., Sternberg, S. H., and Doudna, J. A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. *Nature*, 482(7385):331–338.

Wilcox, R. M. and Fuhrman, J. A. (1994). Bacterial viruses in coastal seawater: lytic rather than lysogenic production. *Marine Ecology-Progress Series*, 114:35–35.

Winter, C., Bouvier, T., Weinbauer, M. G., and Thingstad, T. F. (2010). Trade-Offs between Competition and Defense Specialists among Unicellular Planktonic Organisms: the âĂIJKilling the WinnerâĂİ Hypothesis Revisited. *Microbiology and Molecular Biology Reviews*, 74(1):42–57.

Wommack, K. E., Nasko, D. J., Chopyk, J., and Sakowski, E. G. (2015). Counts and sequences, observations that continue to change our understanding of viruses in nature. *Journal of Microbiology*, 53(3):181–192.

Woolhouse, M. E. J., Webster, J. P., Domingo, E., Charlesworth, B., and Levin, B. R. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genetics*, 32(4):569–577.

Young, R. (1992). Bacteriophage lysis: mechanism and regulation. *Microbiological Reviews*, 56(3):430–481.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Eco-evolutionary dynamics of prophages in natural microbial populations

Authors: Fatima Aysha Hussain, Joseph Elsherbini, Fabiola Miranda-Sanchez, Heidi Li, Natalie Woods, and Martin Polz

2.1 Short title

The ecology and evolution of wild prophages

2.2 Abstract

Bacterial viruses (phages) are important members of all microbial ecosystems and help shape the diversity and dynamics of their hosts. Unlike lytic phages that control bacterial population dynamics solely through predatory interactions, lysogenic phages first lie dormant in microbial genomes as prophages, often providing the host with beneficial functions, before excising and lysing the cell. Prophages occur commonly, with the majority of sequenced genomes harboring at least one, if not multiple,

putative prophages (Arndt et al., 2016). However, this is likely an underestimation as the sequence and compositional diversity of lysogenic phages makes identifying them challenging. While multiple computational algorithms exist to predict and locate putative prophages in assembled genomes and metagenomes, the predictions of these algorithms often depend on comparing query sequences to databases of known virus genes, ultimately limiting detection to known, well-studied viruses. We compare three such algorithms by searching a collection of 341 genomes of marine *Vibrio* strains for putative prophages. Additionally, we take an untargeted sequencing-based approach to identify all actively excising elements, focusing on prophages, in the same set of 341 marine *Vibrio* spanning 17 ecologically and genetically differentiated populations. The sequenced supernatants from the putative lysogens were used to construct a database of actively excising prophages and other mobile genetic elements (MGEs). The database of prophage-like elements is diverse, containing both tailed and non-tailed viruses as well as many novel prophage-like elements that are missed by all three tested prophage search algorithms. Many of the additional MGEs only harbor hypothetical genes and some elements contain a completely unique set of genes from all others. We find that the transfer of closely related elements is biased towards strains from the same population, while more distantly related groups of elements are found across the *Vibrio* phylogeny. This implies prophage infection dynamics may be phylogenetically limited while prophage families transfer across populations. Together, our results lay the foundation for new prophage discovery and shed light on how prophage distributions may shape gene transfer networks in wild microbes.

2.3 Introduction

Lysogenic viruses lie dormant in microbial cells as prophages and can account for over 20% of the DNA in a given bacterial genome (Arndt et al., 2019a). By providing novel, and often beneficial, traits to their hosts, prophages can influence the ecology of bacteria in addition to driving their genomic evolution. Cholera Toxin phage (CTXΦ), for example, contributes heavily to changing the phenotype of its

host, converting it from a predominantly environmental bacterium into one capable of eliciting gastro-intestinal disease in humans (Waldor and Mekalanos, 1996). Prophages can also provide resistance to other viral invaders for their hosts, help regulate host metabolism, and diversify host populations by providing new functions through horizontal gene transfer (Canchaya et al., 2003; Paul, 2008).

Upon their induction, lysogenic viruses become lytic and kill the host cell to propagate to new hosts. In the case of chronic infections, temperate phages can be shed into the environment without host lysis. Only *Innoviridae*, like CTXΦ, are known to exhibit shedding, but it is difficult to observe this phenomenon, therefore, we are likely underestimating its significance in natural populations (Howard-Varona et al., 2017). Whether propagating through lysis or consistent shedding, lysogenic phages contribute to the abundant viral populations in all ecosystems.

In the ocean, a single milliliter of water can contain as many as 10 million viral particles, but the proportion of lytic-to-lysogenic viruses in marine ecosystems can vary. Bulk measures of lysogeny, through induction of the SOS response using mitomycin C, have demonstrated that the lytic-to-lysogenic ratio have seasonal and geographic dependencies, with cooler weather favoring lysogeny (Paul, 2008; Jiang and Paul, 1998; Wommack and Colwell, 2000). Cold weather, and associated slow growth rates, correlates with low abundances of microbial hosts, and is thus thought to set a preference for lysogeny.

Case studies of phage-host pairs have found the lytic-lysogenic duality is a function of both viral and bacterial abundance. For example, when host concentrations are low, phages often lysogenize, allowing them to survive until hosts become abundant. In one case, phages have been shown to use a quorum-sensing-like system that employs a signaling molecule to regulate their lytic-lysogenic "decision" as a function of the molecule's concentration in the extracellular milieu (Erez et al., 2017). Quorum-sensing molecules produced by the host have also been shown to regulate phage lytic-lysogeny decisions (Silpe and Bassler, 2018; Ghosh et al., 2009). Prophages can excise upon detecting host-produced small molecules when hosts are abundant, increasing their chances for propagation and survival. Finally, bacteria may conduct

"remote-controlled" killing by taking advantage of lysogenic phages in other cells to increase their own survival (Selva et al., 2009). In one example, *Staphylococcus pneumoniae* was shown to produce nonlethal levels of hydrogen peroxide, which were still sufficient to induce the SOS response and drive temperate prophages in competing *Staphylococcus aureus* cells to lyse their hosts. This interaction was shown to be specific, as *S. pneumoniae* prophages are not under SOS control and thus sensitive to hydrogen peroxide.

These case studies show that there are multiple mechanisms regulating the lytic-lysogenic decisions of prophages and that these decisions can impact host dynamics. When considering multiple strains within a genus, others' data have suggested phage-induced killing may be used as a form of competitive exclusion between bacteria. For example, a study testing the lytic capabilities of vaginal *Lactobacillus* prophages found that all isolated prophages were infective and exhibited a broad host-range (Kilic et al., 2001), suggesting that prophage-mediated killing can structure intraspecies competition in microbial communities. In another case, prophages from specific *Vibrio* strains were more likely to kill closely-related strains than distantly related ones (Wendling Carolin C. et al., 2018), perhaps exhibiting interspecies competition. However, the distribution and diversity of prophages in naturally co-occurring microbial populations is unknown.

The usual approach for assessing prophage distribution is to use bioinformatic tools that take an input genome and predict regions that look like prophage. These tools typically either use gene annotations of known phage genes, sequence features such as GC content or kmer differences from background, or a combination of the two. It isn't known how well bioinformatic approaches do in predicting active prophage - that is prophage which can actually excise and package themselves to leave the host cell. It also isn't known how well they can predict novel prophages in the genome when there are no gene markers present in the database.

In this work, we wanted to answer the scientific question of (1) What is this distribution and diversity of prophages in naturally co-occurring microbial populations, and the technical questions of (2) how well do state-of-the-art prophage prediction

tools agree with each other and (3) how well do these tools predict active prophage? We answered these questions by taking an ecological- and population genomics-based approach, using a collection of sympatric *Vibrio* strains, ranging in relatedness from nearly clonal to distantly related. This model system is well-suited for answering these questions because not only have the genomes of all the strains been sequenced, but subsets of this system have previously been studied to better understand similar influences on microbial evolution, including those by plasmids and in response to lytic viruses. (Xue et al., 2015; Kauffman et al., 2018)(Appendices A and B). We take advantage of these previous studies in many ways in the present work. For example, while investigating the eco-evolutionary dynamics of plasmids in *Vibrio*, we discovered a set of prophages that reside in their hosts as linear plasmids rather than inserting their genomes into the host's chromosome. We included these hosts in this work as positive controls and to ask if the distribution of such prophage-plasmids differs from other types of prophages in the collection. While investigating the demographics of lytic viral predators of *Vibrio*, we found that the supernatant from lytic infections often contain prophage DNA, either because the prophage were induced by the infection or from background levels of natural induction.Because we deeply sequenced lysates of lytic infections for many hosts, we were also able to capture any induced prophages from the different hosts used to produce those lysates. In addition to the "residual" reads from lytic infections, we used sequencing from supernatants of mitomycin C induction for many hosts and, for fewer hosts, no inducing agent was used for "natural induction" controls. In all these cases, the supernatant was treated with DNAse, so sequencing represents protected molecules of DNA from the host genome.

Using our collection of 341 genome-sequenced isolates, we find that the three bioinformatic tools for prophage prediction from genomes disagree with each other to a large extent, with 73% of the 6,853 total predicted prophage only being predicted by one of the tools. We find experimentally that many putative mobile genetic elements (MGEs) are not predicted by any of the three tools. These putative MGEs are highly diverse, many containing dozens of genes not seen in any of the other elements we find.

These results add to the database of known prophage genomes and suggest that a combination of bioinformatic tools should be used when trying to predict prophages from bacterial genomes with high confidence. This set of MGEs with many novel sequences captured from co-occurring natural populations is a valuable and unique resource for further study.

2.4 Results

2.4.1 Sequence search algorithms differ in their predictions of putative prophages in *Vibrio* genomes

We searched 341 genomes of environmental *Vibrio* strains for putative prophages using three different algorithms (PHASTER (Arndt et al., 2019b), VirSorter (Roux et al., 2015), and PhiSpy (Akhter et al., 2012)), which revealed that each genome has at least one putative prophage identified by at least one algorithm (Figure 1A). Furthermore, the algorithm that finds the most predicted prophages differs genome to genome. The three algorithms vary significantly in the number of predicted prophages per genome, each finding unique putative prophages the others do not (Figure 1B). PhiSpy finds 3671 putative prophages the other two algorithms do not detect, while VirSorter uniquely finds 1197 and PHASTER uniquely finds 180 putative prophages. Of the 824 prophages identified by two or more methods, VirSorter finds all but nine. Our findings suggest all three search algorithms are useful for identifying non-overlapping putative prophages and should be used together to increase the recovery of putative prophages in a given genome.

The differences in prophage detection of the tested algorithms are most likely attributed to differences in their design. VirSorter identifies putative prophages in bacterial genomes by using a sliding window approach, looking for the presence of "hallmark" genes, an enrichment of virus-like genes found in viral databases, stretches of unannotated and uncharacterized genes, short genes, and co-transcribed genes. PHASTER is an updated version of PHAST (Zhou et al., 2011). Both rely on phage

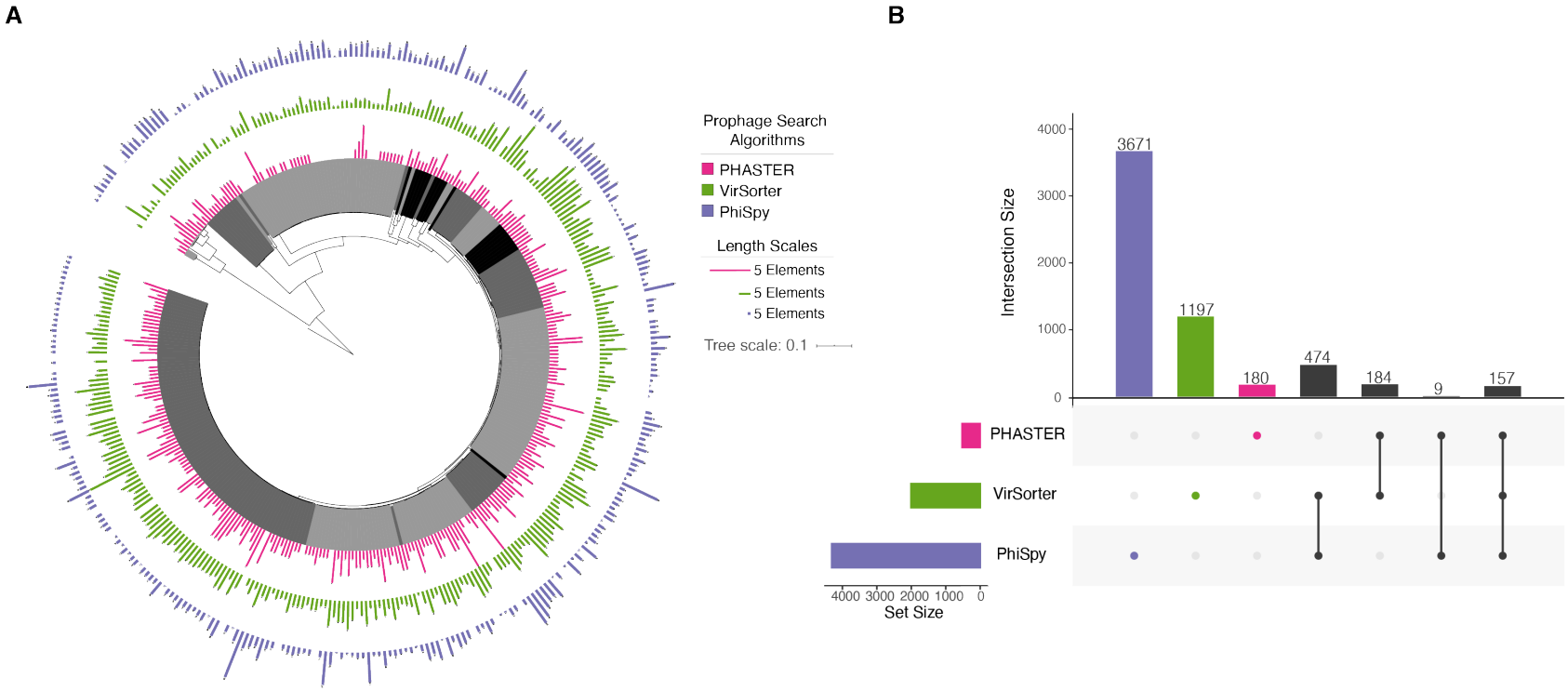


Figure 2-1: **Figure 1:** Predicted prophages in *Vibrio* genomes are inconsistent across prophage finding algorithms. (A) A phylogenetic tree of concatenated ribosomal proteins overlaid with predicted number of prophages identified using three different algorithms: PHASTER, VirSorter, and PhiSpy. (B) Comparison of three prophage search algorithms across all genomes. (top) Number of prophages found uniquely by each algorithm and by combinations of algorithms indicated by (center) linked dots below the horizontal axis. (left) Total numbers of prophages identified by each individual algorithm.

sequence matches to a custom database, found using BLAST (Camacho et al., 2009). PHASTER also searches for tRNAs, attachments sites and targeted sequence annotations. Finally, PhiSpy aims to be the least biased algorithm, using a random forest classifier, incorporating AT skew, gene size, transcription strand orientation, and phage-like 12-mers defined by a user-provided training set.

However, given these differences, the reason behind the skew in prophage calls was not obvious. For example, PHASTER is typically the most conservative in its calls, and PhiSpy should be the most liberal, but for certain strains, the number of prophages found by PHASTER exceeds those found by VirSorter or PhiSpy. The data support the assumption that most, if not all, of the environmental strains presented are putative lysogens. With these predictions in hand, we attempted to verify them experimentally.

2.4.2 Novel prophages and mobile genetic elements are found through induction and sequencing

Given that prophages have been notoriously difficult to identify, and most search algorithms rely on known sequences of viruses and prophages to search for new ones, we took a lab-based approach to identify novel prophages. To find new prophages, we sequenced excised prophages and prophage-like elements from bacterial strains grown under different inducing conditions, including those from cultures treated with mitomycin C, and those from untreated, late stationary phase cultures. Untreated cultures capture natural as well as quorum-sensing induced lysis. We combined these data with existing sequencing data generated from lysates of different strains after being infected by a lytic virus. By removing any reads that mapped to the lytic virus, we were left with residual reads presumably from other excised elements, including prophages.

Using the combined data, we mapped the curated sequencing reads to the bacterial genomes of interest and extracted regions with at least 10x coverage over the background. In doing so, we recovered a total of 473 different putatively mobile genetic

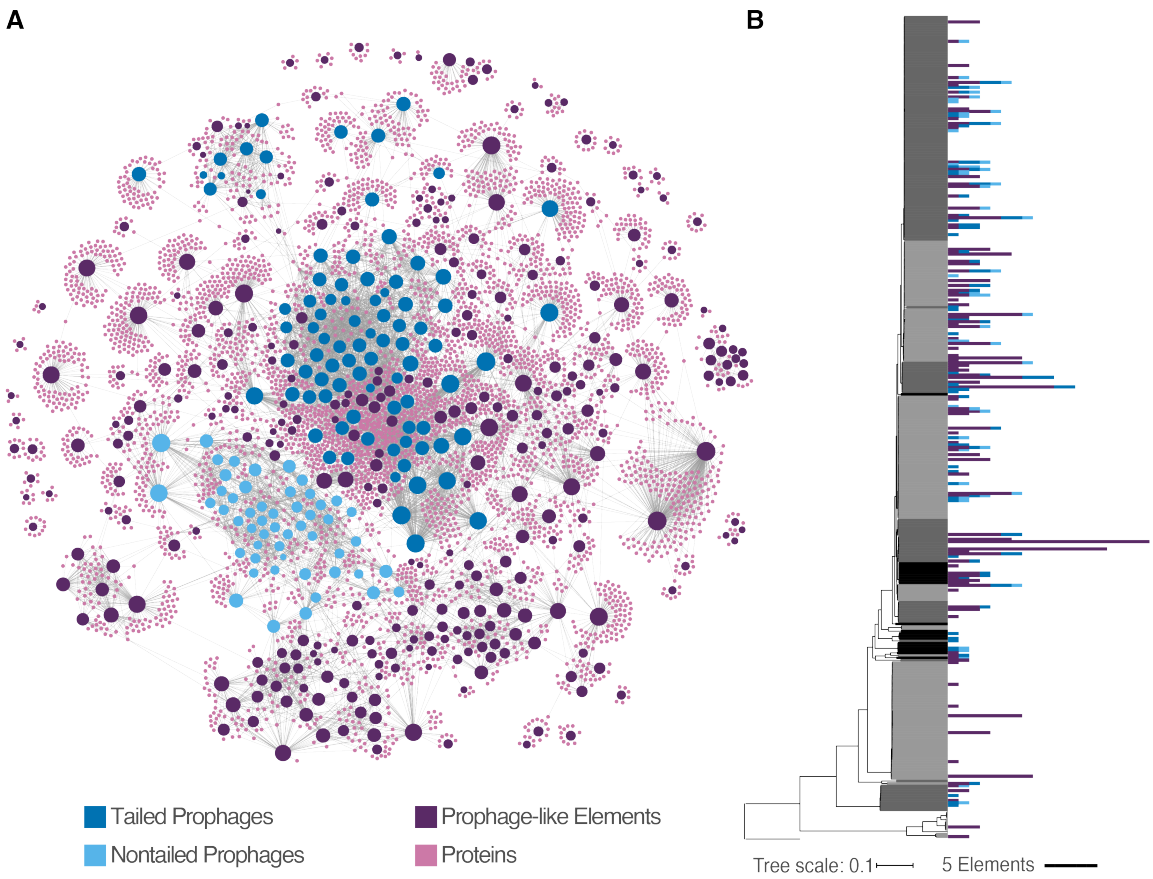


Figure 2-2: **Figure 2:** Induced prophages and prophage-like elements across *Vibrio* populations. (A) Network of prophages and prophage-like elements. Prophages and their proteins are nodes and edges represent shared proteins at 40% similarity determined by MMseqs. (B) Number of each prophage type from (A) across the *Vibrio* phylogeny.

elements (MGEs) over 3 kb in length.We classified these putative MGEs as tailed, non-tailed, or other types of elements based on marker gene searches to known tailed or non-tailed virus proteins, including the major capsid protein, portal proteins, and terminases. In order to investigate the diversity of these elements, we clustered genes present on the elements at 30% protein identity to find they form a relatively sparse network, with many elements not clustering with any other elements and with the vast majority of genes being unique to single elements. (Figure 2A).

The distribution of the elements was patchy across the hosts (Figure 2B). All populations harbor MGEs, but some more than others. In *Vibrio breoganii*, for example, only 10 out of the 49 tested strains had any active MGEs. In *Vibrio tasmaniensis*, on the other hand, 12 of 13 strains have active MGEs. *Vibrio breoganii* have relatively streamlined genomes and have evolved to predominantly live on plant-based particles (Corzett et al., 2018), while Vibrio tasmaniensis strains tend to have larger genomes and have a free-living lifestyle or live on smaller sized particles (Hunt et al., 2008). We investigated whether the size-fraction of the isolated genomes (which tracks with a free-living or particle attached lifestyle) was predictive of the number of putative protected MGEs as it had been for episomes (Xue et al., 2015), but we saw no strong association (Figure S1).

The experimental approach taken here resulted in the discovery of >100 novel elements. When we compared the genomes of the excised prophages with the prophage search algorithms, we found that 353 (75%) elements are found using at least one of the three algorithms, but 120 (25%) go undetected (Figure 3).

2.4.3 The distribution of closely related elements is constrained to closely related hosts

When examining the diversity of elements at the sequence level, we find that the most closely related elements that we observe often occur in closely related taxa (Figure 4A). These elements may be vertically inherited or may indicate recent transfer events. For many elements, the most closely related element is quite dissimilar, even

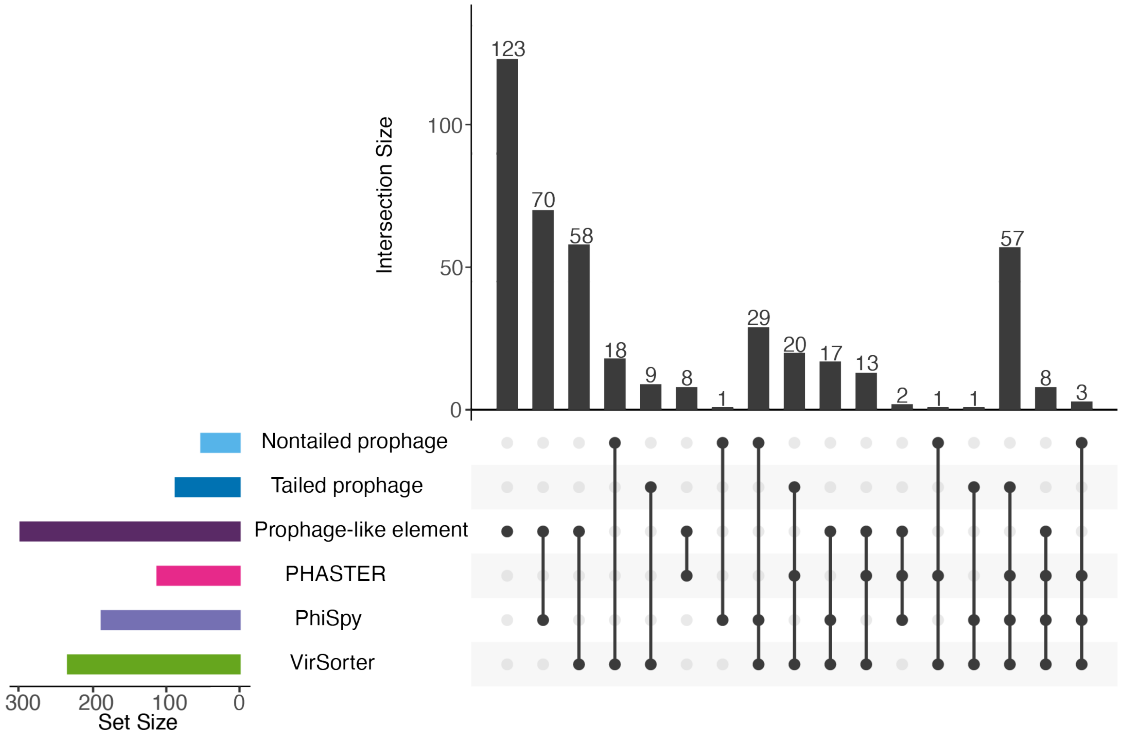


Figure 2-3: **Figure 3:** Comparison of three prophage search algorithms in finding induced elements by type. (top) Number of each type of induced prophages found uniquely by each method and by combinations of methods indicated by (center) linked dots below the horizontal axis. (left) Total numbers of induced nontailed prophages, tailed prophages and other prophage-like elements, and prophages identified by each individual algorithm.

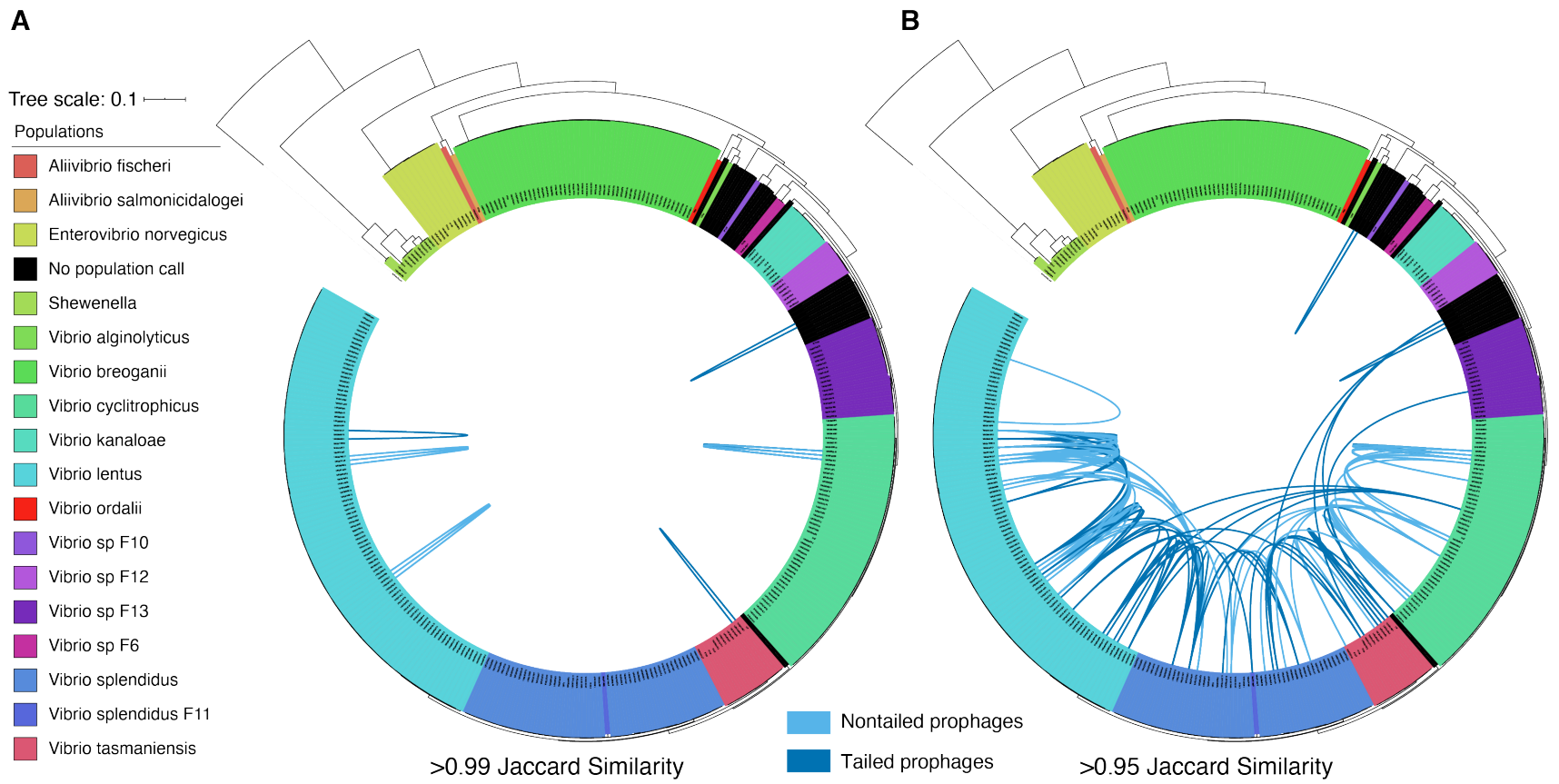


Figure 2-4: **Figure 4:** Putative active MGE network across *Vibrio* hosts. Sharing of putative elements overlaid on an inverted phylogenetic tree of concatenated ribosomal proteins (same as in Fig.1). Sharing of elements was determined using a kmer-based comparison of full nucleotide sequences. Putative elements were compared to one another at two different identity thresholds >0.99 Jaccard similarity (A) and >0.95 Jaccard similarity (B). A connecting line between two genomes represents one element shared between the two.

when found in a relatively closely related genome (Figure S2). When looking at more distantly related pairs of putative MGEs which may represent similar classes of elements, we find membership is sometimes spread amongst populations (Figure 4B).

Finally, there are stark differences in the elements at the gene-level. Figure S3 shows a subset of nine different elements at varying levels of similarity to one another. Examples 1 and 2 have no genes in common with any of the other elements, while 8 and 9 are nearly identical. There is also evidence of recombination among the elements. For example, element 7 shares about half of its genes with element 8 and 9, but the other half of the element shares no genes with any other element in the subset.

2.5 Discussion

In exploring the distribution and dynamics of prophages in environmental *Vibrio* populations, we have uncovered possible ecological roles prophages play in driving the evolutionary dynamics of their hosts. For example, if prophages serve as a tool for competitive exclusion, does prophage-mediated killing of conspecifics dominate within or between populations? We find that many prophages and MGEs are unique, meaning that their frequency in the population is low. This suggests that the turnover of prophages is quite high. We also see that closely related prophages are often found in closely related hosts. This could be attributed to vertical descent, however, comparing to recent findings (Chapter 3) looking at the rapid turnover of phage-defense elements in clonal strains of *Vibrio lentus*, it has been observed that prophage turnover can actually be faster than that of defense elements. This would imply prophages too have a significant role to play in shaping near-term microbial evolution.

2.6 Conclusion and Significance

In exploring the distribution and dynamics of prophages in environmental *Vibrio* populations, we have uncovered possible ecological roles prophages may play in driving

the evolutionary dynamics of their hosts. For example, if participating in competitive exclusion, are bacteria more likely to use prophages to kill within or between their own populations?

Separately, because prophages are difficult to detect, and current algorithms rely on existing databases of known virus genes to identify new prophages, the data presented here will add to these databases and improve current prophage search algorithms.

Finally, here we find that many prophages are unique, meaning that their frequency in the population is low. This suggests that the turnover of prophages is quite high. We also see that closely related prophages are often found in closely related hosts. This could be attributed to vertical descent, however, comparing to previous findings (Chapter 2) looking at the rapid turnover of phage-defense elements in clonal strains of *Vibrio lentus*, it has been observed that prophage turnover can actually be faster than that of defense elements. This would imply prophages too have a significant role to play in shaping near-term microbial evolution.

2.7 Materials and Methods

2.7.1 Prophage induction

Prophages were induced from lysogens both naturally, with the addition of mitomycin C, and as a byproduct of lytic phage lysis.

(1) Select strains were inoculated into 1.2 mL of 2216MB from single colonies and grown in duplicate in 48-well culture blocks, shaking at room temperature. After cultures reached an OD600 of approximately 0.4, Mitomycin C (MMC) was added to one replicate at a final concentration of 0.5 μg/mL. Cultures with and without MMC were grown for a total of 24 hours shaking at room temperature and then pelleted via centrifugation at 5 000 x g for 20 minutes. All supernatants from each treatment were pooled together to form two final "pseudo metagenome" samples. Samples were incubated at room temperature with Turbo DNase for 4 hours, refreshed

with additional DNase, and then incubated again for a total of 24 hours to remove any free DNA. Encapsidated DNA, packaged in vesicles and phages, was then extracted and prepared for sequencing. Genome libraries were prepared for sequencing using the Nextera DNA Library Preparation Kit (Illumina) with 1-2 ng input DNA per isolate, as previously described (Baym et al., 2015). Two libraries were made for each sample and sequenced separately as 4 total samples, multiplexed on one Illumina MiSeq lane.

(2) Individual MMC inductions were not pooled, rather barcoded and then multiplexed for sequencing. Strains were grown from single colonies overnight in 4 mL of 2216MB and then diluted 1:100 the next morning into a final volume of 20 mL of 2216. Cultures were grown in glass test tubes, shaking at room temperature, while OD600 was monitored using an analogue spectrophotometer. Once an OD600 of at least 0.6 was reached, MMC was added to a final concentration of 0.5 μg/mL. Cultures were then allowed to continue growing at room temperature for a total of 24 hours. At that point, cultures were centrifuged at 5 000 x g for 20 minutes in 50 mL Falcon tubes, 0.2 μm-filtered using a Sterivex syringe filter, and placed in ethanol-sterilized centrifugation tubes for ultracentrifugation. Samples were ultra-centrifuged at 32 000 rpm for 2 hours in batches of six. After the first run, the supernatant was discarded, and the sample was rinsed with 0.2 μm -filter sterilized artificial sea water (ASW) and centrifuged again. This rinse was repeated twice and the final palette was allowed to resuspend in 500 μL of elution buffer overnight. The samples were then treated with Turbo DNase per the manufactures protocol in batches of 200 μL each and then extracted. DNA was prepared for sequencing at the MIT BioMicroCenter using the mosquito prep for 96 samples, and ran on a Illumina MiSeq lane.

(3) Residual samples were identified as those inadvertently sequenced as a result of phage isolation. Raw reads for phage preps were mapped to sequenced phage genomes and removed.

For each of the data types described, final reads were mapped to the host genome to identify any regions that were excised and encapsidated using Bowtie2 (Langmead and Salzberg, 2012). Regions greater than 3 000 bp in length with over 10x coverage

above the background were visualized and sorted through by hand to identify regions with significant and consistent spikes in coverage. Regions passing this threshold and within 1 000 bp from one another were stitched together and considered to be one region.

2.7.2 Prophage identification in genomes

Previously published prophage identification algorithms, PhiSpy (Akhter et al., 2012), VirSorter (Roux et al., 2015), and PHASTER (Arndt et al., 2016), were used together to search for putative prophages in select *Vibrio* genomes.

2.7.3 Network of prophages and mobile genetic elements

For each putative prophage-like element we used Prodigal version 2.6.2 [24] to predict open reading frames in "anon" mode. These genes were then clustered at 30% identity using mmseqs2 v11.e1a1c (Steinegger and SÃűding, 2017). We then created a presence-absence network of these gene clusters amongst all the putative prophage-like elements. In this network, the nodes are either a putative prophage-like element or a protein cluster, and edges go from elements to all the protein clusters present in the element. This was visualized using gephi version 0.9.2 (Bastian et al., 2009).

2.7.4 Prediction of prophage element numbers by habitat fraction

We fit a bayesian poisson model to the prophage count data with phylogenetic multilevel pooling using brms (Bürkner et al., 2017). We used only the genomes from the Nahant collection for which we had size fraction isolation information. 100 predictions for the model were plotted for all the genomes.

2.7.5 Transfer and distribution of prophages

For each pair of putative prophage-like elements we used Mash v 2.2.2 (Ondov et al., 2016) to estimate the pairwise similarity. Briefly, Mash finds the set of kmers for each sequence and then estimates the Jaccard index which is a similarity metric for binary data. This metric is normalized to the total kmer set size, so longer sequences that share a high percentage of kmers would have a higher similarity than smaller sequences. For highly similar sequences (which may be the same element transferred or vertically desenced) we used a cutoff of 0.99 Jaccard similarity and for similar sequences (which may be a more distantly vertically inherited elements or elements of a similar type) we used a cutoff of 0.95 similarity.

2.8 Acknowledgements

We extend thanks to Kathryn Kauffman, Michael Cutler, Sara Schwartz for advice and assistance with preliminary methods development.

2.9 Conflicts of Interest

The authors declare that they have no conflict of interest.

Bibliography

Akhter, S., Aziz, R. K., and Edwards, R. A. (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Research*, 40(16):e126.

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research*, 44(W1):W16–W21.

Arndt, D., Marcu, A., Liang, Y., and Wishart, D. S. (2019a). PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. *Briefings in Bioinformatics*, 20(4):1560–1567. Publisher: Oxford Academic.

Arndt, D., Marcu, A., Liang, Y., and Wishart, D. S. (2019b). Phast, phaster and phastest: Tools for finding prophage in bacterial genomes. *Briefings in bioinformatics*, 20(4):1560–1567.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Third International AAAI Conference on Weblogs and Social Media*.

Baym, M., Kryazhimskiy, S., Lieberman, T. D., Chung, H., Desai, M. M., and Kishony, R. (2015). Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. *PLOS ONE*, 10(5):e0128036.

Bürkner, P.-C. et al. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80(1):1–28.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, 10:421.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and BrÃijssow, H. (2003). Prophage Genomics. *Microbiology and Molecular Biology Reviews*, 67(2):238–276.

Corzett, C. H., Elsherbini, J., Chien, D. M., Hehemann, J.-H., Henschel, A., Preheim, S. P., Yu, X., Alm, E. J., and Polz, M. F. (2018). Evolution of a Vegetarian Vibrio: Metabolic Specialization of V. breoganii to Macroalgal Substrates. *Journal of Bacteriology*, pages JB.00020–18.

Erez, Z., Steinberger-Levy, I., Shamir, M., Doron, S., Stokar-Avihail, A., Peleg, Y., Melamed, S., Leavitt, A., Savidor, A., Albeck, S., Amitai, G., and Sorek, R. (2017). Communication between viruses guides lysisâĂŞlysogeny decisions. *Nature*, 541(7638):488–493.

Ghosh, D., Roy, K., Williamson, K. E., Srinivasiah, S., Wommack, K. E., and Radosevich, M. (2009). Acyl-Homoserine Lactones Can Induce Virus Production in Lysogenic Bacteria: an Alternative Paradigm for Prophage Induction. *Applied and Environmental Microbiology*, 75(22):7142–7152.

Howard-Varona, C., Hargreaves, K. R., Abedon, S. T., and Sullivan, M. B. (2017). Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *The ISME Journal*, 11(7):1511–1520.

Hunt, D. E., David, L. A., Gevers, D., Preheim, S. P., Alm, E. J., and Polz, M. F. (2008). Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton. *Science*, 320(5879):1081–1085.

Jiang, S. C. and Paul, J. H. (1998). Significance of lysogeny in the marine environment: studies with isolates and a model of lysogenic phage production. *Microbial ecology*, 35(3-4):235–243.

Kauffman, K. M., Brown, J. M., Sharma, R. S., VanInsberghe, D., Elsherbini, J., Polz, M., and Kelly, L. (2018). Viruses of the Nahant Collection, characterization of 251 marine Vibrionaceae viruses. *Scientific Data*, 5:180114.

Kilic, A. O., Pavlova, S. I., Alpay, S., Kilic, S. S., and Tao, L. (2001). Comparative Study of Vaginal Lactobacillus Phages Isolated from Women in the United States and Turkey: Prevalence, Morphology, Host Range, and DNA Homology. *Clinical and Diagnostic Laboratory Immunology*, 8(1):31–39.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132.

Paul, J. H. (2008). Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *The ISME Journal*, 2(6):579–589.

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

Selva, L., Viana, D., Regev-Yochay, G., Trzcinski, K., Corpa, J. M., Lasa, Ã., Novick, R. P., and PenadÃl's, J. R. (2009). Killing niche competitors by remote-control bacteriophage induction. *Proceedings of the National Academy of Sciences*, 106(4):1234–1238.

Silpe, J. E. and Bassler, B. L. (2018). A Host-Produced Quorum-Sensing Autoinducer Controls a Phage Lysis-Lysogeny Decision. *Cell*.

Steinegger, M. and SÃűding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028.

Waldor, M. K. and Mekalanos, J. J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science*, 272(5270):1910–1914.

Wendling Carolin C., Goehlich Henry, and Roth Olivia (2018). The structure of temperate phageâĂŞbacteria infection networks changes with the phylogenetic distance of the host bacteria. *Biology Letters*, 14(11):20180320.

Wommack, K. E. and Colwell, R. R. (2000). Virioplankton: Viruses in Aquatic Ecosystems. *Microbiology and Molecular Biology Reviews*, 64(1):69–114.

Xue, H., Cordero, O. X., Camas, F. M., Trimble, W., Meyer, F., Guglielmini, J., Rocha, E. P. C., and Polz, M. F. (2015). Eco-Evolutionary Dynamics of Episomes among Ecologically Cohesive Bacterial Populations. *mBio*, 6(3):e00552–15.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). PHAST: A Fast Phage Search Tool. *Nucleic Acids Research*, page gkr485.

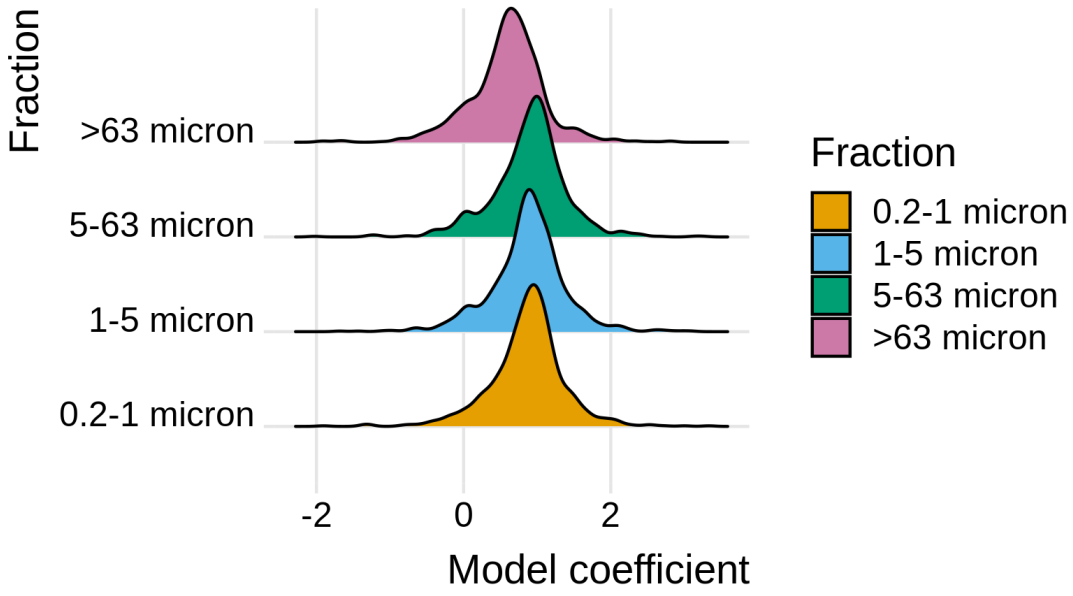


Figure 2-5: **Figure S1:** Model fits of size fraction to predict putative active MGE number. Using a phylogenetic regression, we fit a poisson model to the total number of putative MGEs for each genome given the size-fraction of isolation. Shown here are the log-link coefficients for 1000 posterior draws of the model, which indicate no strong differences amongst the size fractions after accounting for phylogeny.

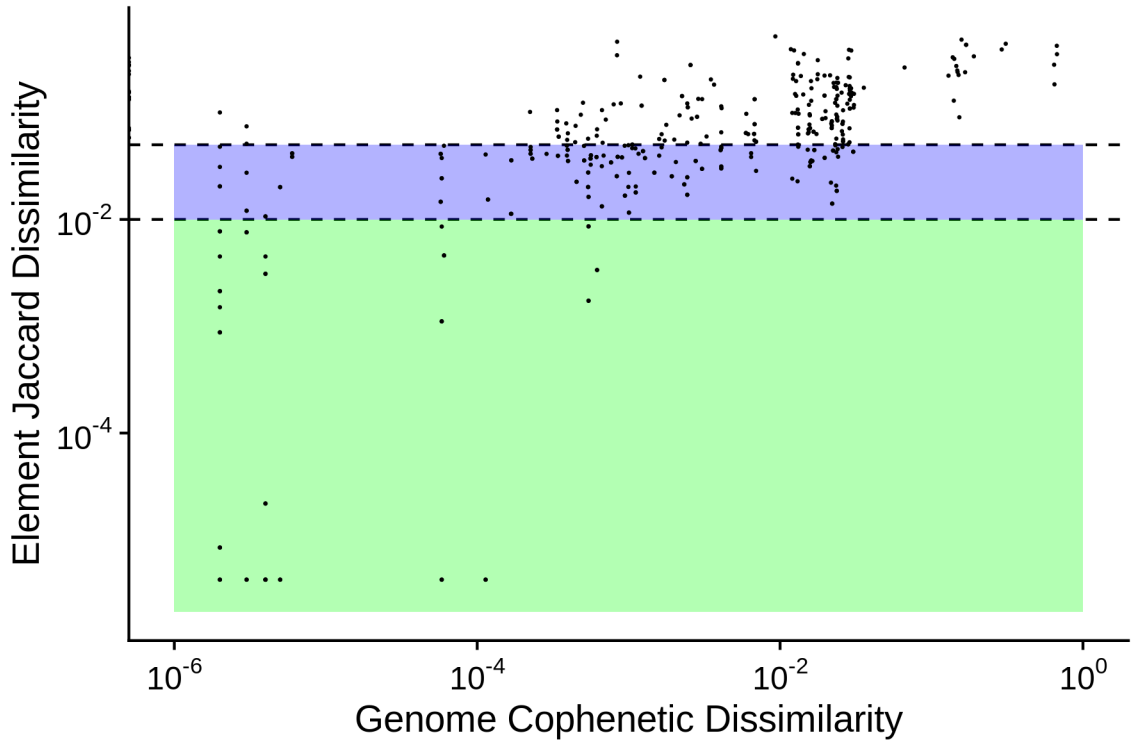


Figure 2-6: **Figure S2:** Genome relatedness compared to element relatedness. For each putative MGE the closest other MGE was found and plotted against the relatedness of the isolate genomes. The phylogeny of isolate genomes was turned into a pairwise dissimilarity by calculating the cophenetic coefficient, and the similarity of elements was found using the Jaccard index on the kmer content of each element. In green is the region of elements which are 0.01 or less in their Jaccard index which are depicted in Figure 4A and in blue are the additional elements up to 0.05 in their Jaccard index depicted in Figure 4B.

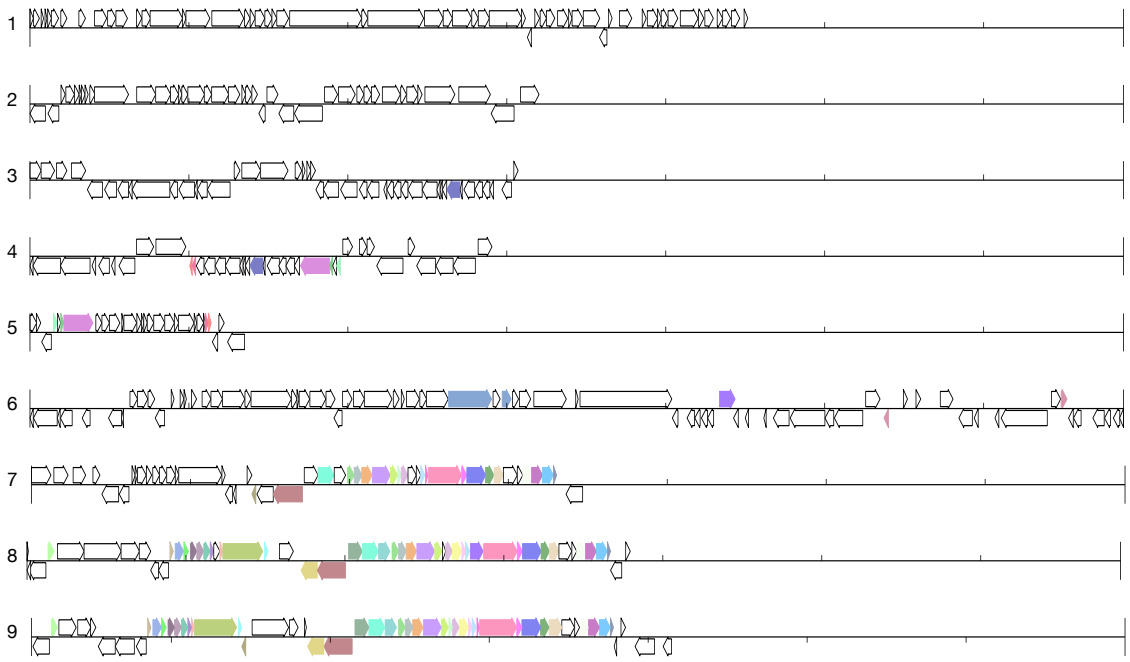


Figure 2-7: **Figure S3:** Gene diagrams of randomly selected representatives of elements. Genes with the same color share 65% protein sequence identity.

Chapter 3

Rapid evolutionary turnover of mobile genetic elements drives microbial resistance to viruses

Notes: The contents of this chapter are in review for publication at the time of writing under the following title: Hussain FA, Dubert J, Elsherbini J, Murphy M, VanInsberghe D, Arevalo P, Kauffman K, Kotska Rodino-Janeiro B, and Polz MF. "Rapid evolutionary turnover of mobile genetic elements drives microbial resistance to viruses." In review, March 2020.
Main-text and supplementary figures and tables are located at the end of the chapter and any referenced supplementary data are available at:
https://github.mit.edu/fatimah

3.1 Abstract

Although it is generally accepted that viruses (phages) drive bacterial evolution, how these dynamics play out in the wild remains poorly understood. Here we show that the arms race between phages and their hosts is mediated by large and highly diverse mobile genetic elements. These phage-defense elements display exceedingly fast evolutionary turnover, resulting in differential phage susceptibility among clonal bac-

terial strains while phage receptors remain invariant. Protection afforded by multiple elements is cumulative, and a single bacterial genome can harbor as many as 18 putative phage-defense elements, overall accounting for 90% of the flexible genome amongst closely-related strains. The rapid turnover of these elements demonstrates that phage resistance is unlinked from other genomic features and that resistance to phage therapy might be as easily acquired as antibiotic resistance.

3.2 One Sentence Summary

For wild microbes, horizontal gene transfer of phage-defense elements decouples resistance to phages from host core function.

3.3 Report

Bacterial viruses (phages) are ubiquitous across Earth's biosphere and control microbial populations through predatory interactions (Wommack and Colwell, 2000; Breitbart and Rohwer, 2005; Hampton et al., 2020). Because successful killing depends on molecular interaction with the host, phages display higher specificity than most other microbial predators – this being an important reason for renewed interest in the clinical use of phages as alternatives to antibiotics (Kortright et al., 2019). Like antibiotics, however, phage killing exerts strong selection for resistance in bacterial hosts (Labrie et al., 2010). Consequently, how bacteria naturally acquire resistance has important implications for understanding microbial community dynamics as well as the long-term success of phage therapy. Laboratory co-evolution studies have consistently identified phage-receptor mutations as key drivers of resistance (Avrani et al., 2011; Meyer et al., 2012; Bohannan and Lenski, 2000), with genetic analyses suggesting secondary contributions by restriction-modification (RM) (Arber and Dussoix, 1962; Wilson and Murray, 1991; Bertani and Weigle, 1953) and abortive infection (Abi) systems (Molineux, 1991; Snyder, 1995). However, because phages target important surface structures such as the lipopolysaccharide (LPS) and

membrane proteins as receptors, it is questionable whether mutations in receptors represent primary adaptive strategies in complex microbial communities since such modifications frequently incur significant fitness costs (Avrani et al., 2011). Indeed, many additional defense mechanisms have recently been discovered, including clustered regularly interspaced short palindromic repeats (CRISPR) systems (Barrangou et al., 2007; Andersson and Banfield, 2008), and several other yet-to-be mechanistically characterized defense mechanisms (Doron et al., 2018; Novick et al., 2010). Importantly, genes encoding both receptors and defense systems have been shown to occur frequently in variable genomic islands (Avrani et al., 2011; Makarova et al., 2011; Rodriguez-Valera et al., 2009), or associated with mobile genetic elements (McDonald et al., 2019; McKitterick and Seed, 2018; Koonin et al., 2019). However, neither the predominant mechanisms of resistance to phages nor the dynamics of resistance gain and loss are well understood for microbes in the wild, limiting our knowledge of how phage predation structures diversity and drives evolution of microbial populations.

Here we combine population genomic and molecular genetic approaches to determine how phage resistance arises in microbial populations evolving in nature. We recently created the largest genomically-resolved phage-host cross-infection network, using sympatric environmental isolates, allowing us to ask at what genetic divergence and by what genetic mechanisms host resistance to phages arises in the wild (Kauffman et al., 2018b,a). This system – the Nahant Collection – was established in the context of a 93-consecutive-day coastal-ocean time series, and comprises over 1,300 strains of marine *Vibrio* ranging in relatedness from near-clonal to species-level divergence (Martin-Platero et al., 2018). *Vibrio* hosts isolated on three different days were used as "bait" to isolate 248 co-occurring lytic viruses by quantitative plaque assays, and each of 245 plaque positive hosts were subsequently challenged with all phage isolates to establish an all-by-all cross-infection matrix (Fig. S1A). Phages tend to be highly specific, as indicated by the general sparsity of the matrix. Additionally, even the most closely related hosts, differentiated by a few single nucleotide polymorphisms (SNPs) across the genome, are preyed upon by different phages (Fig. S1B). Furthermore, this extends to broader host-range phages that, although capable

of infecting multiple hosts, are typically limited to a single strain within a host clade (Fig. S1C).

Because the observation that nearly clonal bacterial strains are subject to differential predation suggests extremely rapid evolution of phage resistance, we sought to identify the responsible mechanism in an exemplary set of 19, nearly clonal isolates of *Vibrio lentus*. These strains share identical nucleotide sequences in 52 ribosomal proteins (Fig. 1A, left) yet form two groups of 4 and 15 strains subject to differential infection by two groups of siphovirus phages consisting of 4 and 18 isolates ("purple" and "orange" in Fig. 1A, Fig. S2A-D). While the phage groups are so divergent that their genomes cannot be aligned (Fig. 1A, top; Fig. S2EF), the two host groups are so recently diverged that they differ only by 14 SNPs across their entire core genome (Fig. S3, Table S1). Re-assaying all pairwise interactions between the phages and hosts in this set over a wide range of phage-to-host ratios (multiplicities of infections, MOIs) revealed that representatives of both phage groups can also attach to strains originally scored as non-hosts and kill them at high concentrations, albeit without production of viable phage progeny (Fig. S4). This effect, termed "lysis from without" (DelbrÃijck, 1940), along with observed adsorption of all phages to all hosts (Fig. S5), led us to hypothesize that all phages in these two groups can recognize receptors on all 19 bacterial strains and that differential resistance among these strains is mediated by intracellular mechanisms rather than by receptor modification.

Supporting our hypothesis, we find that while the two phage types use different receptors, all bacterial genomes encode both sets of receptors. Transposon mutagenesis data suggests that the most likely candidate for the "orange" phage receptor is the Type II secretion system pseudopilus, while the "purple" phages likely use LPS as a primary receptor and a sodium transporter as a secondary receptor (Table S2). Despite the two phage groups using distinct receptors, every identified gene involved in phage entry is identical in nucleotide sequence across the two host types, and therefore part of their core genome. Additionally, SNPs identified upon selecting for laboratory evolved resistance in the host strains following phage exposure corroborated these results: SNPs in the same receptor loci were identified in the spontaneously resistant

strains, verifying the receptor identification, and suggesting that the mode of SNP-based resistance evolution observed in the lab is not always representative of that in the wild (Table S3). All of this evidence supports our hypothesis that receptors do not drive phage specificity in these strains in the environment, leading us to explore potential intracellular mechanisms of host resistance by a combination of comparative genomics and molecular genetics.

Clustering the pangenome of *Vibrio lentus* isolates using high-quality genomes generated by hybrid assemblies of long and short reads (Fig. 1A, right) reveals that each host group harbors a set of putative phage-defense genes and that all of these genes are housed on large, genomically-integrated mobile genetic elements rather than merely being clustered in variable genomic regions as previously suggested for other defense genes (Rodriguez-Valera et al., 2009). Three and two defined genomic regions specific to the "orange" and "purple" phage-susceptible hosts could be identified, respectively (Fig. 1B). These regions appear to be mobile genetic elements, which are likely transferred via site-specific recombination, because each element contains a defined insertion site on the host chromosome, potentially allowing for circularization, and all elements contain integrases and transposases on their periphery (Fig. 1B, Data S1). Similar elements can be found in very distantly related strains, suggesting their common horizontal acquisition and loss (Fig. S6). Because each element carries at least one unique known phage defense system, we refer to them as phage-defense elements (PDEs). PDE1, PDE4, and PDE5 have Type 1 RM systems, and PDE2 and PDE3 have putative toxin-antitoxin (TA) systems, some of which have been shown to act as Abi systems, killing the host upon phage infection (Dy et al., 2014). The PDEs range in size from approximately 10 to 60 kbp, and aside from defense systems and mobile element proteins, most remaining genes on the elements are unannotated. A striking feature of the PDEs is that their insertion does not appear to disrupt host functions. For example, PDE1 inserts into "orange" hosts' 5'-deoxynucleotidase nucleosidase (Yfbr) gene, thereby truncating it, but encodes its own distinct copy of the same gene with a divergent amino acid sequence. Similarly, although PDE2 disrupts a thiol peroxidase gene upon insertion, it encodes for a second peroxidase copy in the

middle of the element (Fig. 1B, Data S1).

That multiple putative PDEs cluster with phage resistance phenotypes poses the question: To what extent do each of these elements contribute to the observed resistance? All three PDE-encoded RM systems appear active, as methylome data show that the distinct sequence motifs corresponding to each RM system are methylated only within "orange" and "purple" phage-host sets – that is, only in the bacterial genomes in which the RMs occur, and in the phages that can kill those hosts (Table S4). To further characterize the contribution to resistance of a full set of PDEs in the host populations, we systematically knocked out the putative defense portions of the three PDEs hypothesized to drive resistance of a representative "orange" host strain (10N.261.55.C8) to "purple" phages (Fig. S7A), and then challenged the knockouts with a representative "purple" phage (1.281.O) (Fig. 2A, Fig. S7B). This genetic analysis supports that in addition to the RM-encoding PDE specific to the "orange" host population, both PDEs encoding putative Abi systems are also active and contribute to phage resistance in a complex, cumulative manner. The complexity of the interaction among all PDEs leading to full resistance is illustrated by deleting them in all possible combinations. Knocking out RM-containing PDE1 alone increased killing tenfold (from 10^{-1} to 10^{-2} phage dilution), but still did not yield viable phage progeny (Fig. 2B). This phenotype is consistent with the expected host-killing of Abi systems (Lindahl et al., 1970) and led us to hypothesize that there may be a multi-level defense structure involving the remaining two PDEs as well. Knocking out Abi-system containing PDE2 and PDE3 independently resulted in no change from the wild type (WT) phenotype. However, knocking out PDE2 and PDE3 together allowed for both killing and propagation of the phage at a wider range of high MOIs ($\sim 10^{-5}$ phage dilution). This phenotype is typical of an RM system-based defense strategy which is inherently imperfect (Kruger and Bickle, 1983) – at high MOIs, we expect higher numbers of co-infecting phages and thus an increase in the probability that an inadvertent phage-DNA methylation will occur at the target motif and allow a phage genome to escape restriction and replicate successfully. Knocking out PDE1 and PDE2 together and PDE1 and PDE3 together resulted in the same tenfold increase

in killing observed when deleting PDE1 alone, indicating PDE2 and PDE3 provide a certain level of redundancy in protection. Finally, knocking out PDE1 and PDE3 together yields killing at much lower MOIs (10^{-5} phage dilution), suggesting PDE3 is a stronger resistance element than PDE2. However, the observed killing still did not consistently yield viable phage propagation (Fig. 2B). Finally, knocking out all three PDEs simultaneously resulted in the "orange" strain becoming just as susceptible to the "purple" phage as "purple" WT host strains. Therefore, we conclude all three elements are needed for full, WT-level defense.

Expanding our genomic comparisons to additional closely-related isolates in our collection reveals that PDEs comprise the vast majority of the flexible gene content. Furthermore, PDEs account for a large fraction of the unannotated genes therein, thereby addressing the general open question of the function of the pan-genome (Rocha, 2018). In addition to the 19 clones, we included 4 strains in our collection that are closely related but exhibit alternative phage sensitivity profiles. Using a k-mer-based approach to conduct all-by-all pair-wise genome comparisons, we identified a total of 30 unique putative PDEs, totaling 862,000 bp in length, shared by different subsets of the 23 strains analyzed (Fig. S8). The number of PDEs ranges between 10 and 18 in each strain, and collectively, the PDEs account for >90% of the flexible genomic regions, which can hence be given a tentative annotation (Fig. 3). Even if only known defense genes (not entire PDEs) are considered, 20% of the flexible gene content is accounted for (Fig. 3). A similar range of 12-21% of flexible gene content is observed when we assayed the fraction of known defense genes in other diverse *Vibrio* species in our collection (Fig. S9). Importantly, because this measure only comprises the defense genes and not the entire PDEs, this suggests that a major portion, and perhaps the majority, of the pan-genome across these species is involved in phage defense and suggests a path forward for annotating this enigmatic genetic repertoire.

Defense being entirely relegated to PDEs confirms theoretical considerations that resistance genes should be mobile because the cost of resistance limits their utility under changing predation pressures (Koonin et al., 2019). Our findings demonstrate

that the rate of turnover can be surprisingly fast, with only 14 SNPs accumulating across the entire genome per 5 PDE transfer events (gains or losses). This finding is likely more general as other *Vibrio* species in our collection, for which clonal isolates are available, also differ in their phage predation profiles, and even among the bacterial pathogens *Listeria* and *Salmonella* that follow a primarily non-recombinogenic mode of evolution, we observe similarly rapid turnover of putative PDEs (Fig. S10, Fig. S11). Thus comparative genomics of near clonal isolates combined with phage host-range data is a fruitful method to discover novel phage-defense mechanisms in an unbiased way. However, we emphasize that high quality genomes assembled using long reads are essential since, in our experience, PDEs assemble poorly when using short read data alone, in part due to AT richness and high density of variable repeat regions.

Because receptor genes are invariant across the two near clonal host groups challenges the notion that receptor variation is primarily responsible for resistance (Avrani et al., 2011; Rodriguez-Valera et al., 2009), we asked to what extent the receptors identified are variable across more diverse populations. Populations are defined here as gene flow clusters that also represent ecological units (Arevalo et al., 2019; Shapiro et al., 2012), and the recognition of population boundaries is key for interpretation of gene or allele frequency in light of selective forces. Surprisingly, looking across 107 *Vibrio* isolates, spanning 10 populations, all putative receptors are highly monomorphic at the population level and possibly under purifying selection. The two genes identified as putative receptors in *V. lentus* are identical, or nearly so, at the nucleotide level within most of the diverse populations and thus well below the average diversity of core genes (Fig. S12A). This is corroborated by phylogenetic trees showing that all members of each population carry the same or highly similar gene variants, a pattern consistent with recent gene-specific selective sweeps, with the notable exception of the pseudopilin gene of the Type II secretion system, which is more diverse in two of the populations (Fig. S12B). Finally, the LPS, which frequently serves as primary receptor for many phages, also appears similar at the population level since the genes responsible for synthesis display population-specific presence/absence

patterns suggesting the synthesis pathway is conserved (Fig. S12C). Thus, although putative receptors can reside in variable regions when more divergent genomes are compared (Rodriguez-Valera et al., 2009), their evolution appears much more constrained when population structure is considered. This constraint may arise because, in wild populations, these surface structures are optimized for ecological interactions, and indicates a key difference from the lab where receptor mutations frequently arise in phage-host co-cultures (Westra et al., 2015). This observed invariance thus suggests that other selective forces that compete with phage resistance play an important role in receptor evolution in the wild, and is consistent with predictions that intracellular defenses should be important under such conditions (Zborowsky and Lindell, 2019). It is possible that receptor-mediated defenses may only be advantageous under extreme predation regimes, or under regimes of low effective diversity such as a clonal infection.

The rapid turnover of PDEs implies that phage resistance is essentially unlinked from other traits within bacterial populations. Low linkage means that in complex microbial communities, bacterial core genomes can be maintained over the long-term even in the face of phage predation, while flexible genome content involved in shielding against phages is highly dynamic. In particular, our results question whether phage predation can increase microbial population diversity at the strain level by virtue of Kill-the-Winner type dynamics, which postulates that fitter genotypes are prevented from outcompeting all others within a population since they are disproportionally affected by phage predation (Winter et al., 2010). Instead, such dynamics are likely limited to acting at the resolution of mobile genetic elements and flexible genes, with limited consequences for the longer-term dynamics of bacterial population core genome diversity. This means that other factors, aside from phage predation, must drive the diversity observed in wild microbial populations. Similarly, rapid transfer of PDEs implies resistance to phage therapy may be easily acquired and quickly spread through bacterial populations, just as the connection to mobile genetic elements (primarily plasmids) has led to an unanticipated rise in antibiotic resistance. Together, our findings suggest that phage resistance is an important, if not the most important,

selective force determining clonal bacterial diversity, with phage-defense elements potentially explaining a very large portion of the previously enigmatic bacterial flexible genome.

3.4 Materials and Methods

3.4.1 Bacteria and phage isolation

Bacteria and phages were obtained in a previous study from coastal seawater collected from Canoe Cove, Nahant, MA, USA, on August 22 (ordinal day 222), September 18 (261), and October 13 (286), 2010 (Kauffman et al., 2018b). *Vibrio* bacteria were isolated using a size fractionation approach, followed by plating on selective media, as described previously (Hunt et al., 2008). Briefly, to capture bacteria associated with large particles and zoo- and phytoplankton, seawater was filtered through a 63 μm average pore size plankton net. To capture bacteria in smaller size fractions, including small particles and smaller zoo- and phytoplankton as well as bacteria occurring in the free-living fraction, water pre-filtered through a 63 μm net was serially passed through 5 μm, 1 μm, and 0.2 μm polycarbonate filters. To isolate vibrios from each fraction, material captured in the plankton net and on filters was resuspended in artificial seawater (ASW; Sea Salts from Sigma-Aldrich), and the suspensions passed through polyethersulfone 0.2 μm filters. These final filters were placed directly on agar plates of MTCBS (Difco Thiosulfate-Citrate-Bile-Sucrose Agar amended with 10 g/L of NaCl to final concentration of 2% w/v) to allow for selective growth of *Vibrio* colonies. Colonies were purified by serial passaging on agar plates of first, TSB2 (Tryptic Soy Broth, 1.5% Difco Bacto Agar, amended with 15 g NaCl to 2% w/v); second, MTCBS, and third TSB2. Colonies were inoculated into 1 mL of Difco 2216 Marine Broth (2216MB) in 96-well 2 mL culture blocks and allowed to grow, shaking at room temperature, for 48 hours. Glycerol stocks for preservation at -80°C were prepared by combining 100 μL of culture with 100 μL of 50% glycerol (50% water) in 96-well microtiter plates. The naming of each strain reflects isolation

location, day, and size fraction: 10N refers to the 2010 collection of samples from Nahant; 222, 261, 286 are the ordinal dates of the year; 54, 55, 56 are three replicates of the 63 μm fraction; 51, 52, 53, are three replicates of the 5 μm fraction; 48, 49, 50 are three replicates of the 1 μm fraction; and 45, 46, 47 are three replicates of the 0.2 μm or free-living fraction. The final portion of the name is the original storage well in a 96-well plate. Note that the "orange" isolates were, with one exception, collected on day 261 and distributed between the 1 μm, 5 μm, and 63 μm fraction, while all "purple" isolates were collected on day 286 from the 63 μm fraction. Therefore, the dynamics described in this work are likely occurring in particle-attached bacterial hosts in the ocean.

For phage isolation, 4 L of seawater was collected in triplicate on each day in the time series and separately filtered through a Sterivex 0.22 μm barrel filter into a sterile 4 L collection bottle using a peristaltic pump. Phages were directly concentrated from this filtrate using an iron flocculation and filtering method described previously (John et al., 2011). Briefly, iron (III) chloride, which is spiked into the sample, precipitates phages from the solution, and then the precipitates are collected onto 90 mm 0.2 μm polycarbonate filters using a glass cup-frit system. Precipitates are finally dissolved in 4 mL of oxalate solution to yield a quantitative concentration of 1,000x from the original 4 L. The final phage concentrate was stored at 4°C in the dark until used to isolate specific viruses for different bacterial hosts.

Vibrio isolates were used as "bait" to obtain phages from the concentrates using direct plating in soft agar overlays. Plaques from the bait assay were archived frozen in 2216MB and glycerol and phages for use in the host range assay were subsequently randomly selected from archives for each host and purified by triple serial passage using tube-free agar overlays (Kauffman et al., 2018a; Kauffman and Polz, 2018) on their hosts of isolation. Phages were amplified on their hosts of isolation using primary small-scale liquid cultures inoculated with plaque plugs from their final serial passage in agar overlays, followed by plating of primary lysates into agar overlays to achieve"at confluence" (saturated with plaques but not completely cleared) were harvested into 2216MB, centrifuged at 5,000 x g for 20 minutes, and filtered through Sterivex 0.22

μm barrel filters to generate the lysates used for the all-by-all host range cross test as well as for phage DNA extraction and sequencing (Kauffman et al., 2018a), as well as methylation profiling (described below).

3.4.2 Phage host-range matrix

Phage host-range was determined in a previous study (Kauffman et al., 2018b). Briefly, all *Vibrio* strains for which at least one phage was found ("plaque positive") were used in the host-range assay and challenged with all phages purified as described above. Bacterial hosts were plated in agar overlays in large 150 mm plates and stamped with phage lysates arranged in triplicate in 96-well arrays using 96-spot blotters (BelArt, Bel-blotter 96-tip replicator, 378760002). Clearing seen in at least 2/3 replicates was scored as a positive kill (Kauffman, 2014). The concentration of each lysate in the original assay was not normalized to allow for higher throughput but the assay was repeated for select hosts at a range of concentrations (see methods for "varying phage concentrations" below).

To organize bacterial hosts in the matrix by phylogeny, concatenation of ribosomal proteins and *hsp60* sequences was used to construct a phylogenetic tree reflecting the relationship of the core genome (Fig. S1A). When genome sequences were available, we used HMMER (Eddy, 2011) to find ribosomal proteins, and aligned the sequences with MAFFT (Katoh and Standley, 2013). Amino acid sequences of *hsp60* proteins were also extracted from genomes via HMMER using pfam PF00118. The *hsp60* sequences were aligned using the mafft-fftnsi algorithm. When genomes were not available, hsp60 sequences that were Sanger-sequenced were added to this alignment using the mafft-fftnsi algorithm with the -addfragments option. The *hsp60* alignment and the ribosomal protein alignment were concatenated and used to create the phylogenetic tree in Figure S1A using RAxML (options: -q, -m GTRGAMMAX) (Stamatakis, 2014).

3.4.3 Phage characterization

Circular representations of the previously sequenced phage genomes (Fig. S2CD) were generated using BRIG (Alikhan et al., 2011) with publicly available NCBI GenBank files; annotations were made based on manual review of GenBank predictions and supplemented with Phyre2 (Kelley et al., 2015) and EggNog-Mapper (Huerta-Cepas et al., 2017, 2019) annotation. Genome diagrams (Fig. S2EF) were generated using the GenoPlotR package in R (Guy et al., 2010) with predicted protein coding genes indicated as arrows colored to correspond to protein sequence clusters, as defined using default settings of MMseqs2 (Steinegger and SÃűding, 2017); and with "orange" (Fig. S2E) and "purple" (Fig. S2F) phages clustered and identified as two separate genus-level groups using the D6 amino acid OPTSIL clustering algorithm in the VICTOR classifier (Meier-Kolthoff and GÃűker, 2017) with whole genome concatenated protein sequences.

3.4.4 Hybrid assemblies of bacterial genomes

Because we noticed that PDEs assemble poorly when genomes were sequenced with short read technology, we used Illumina short reads and Pacific Biosciences (PacBio) long reads in combination to assemble high quality (nearly closed) genomes. For short read sequencing, bacterial isolates were grown overnight from a single colony in 1.2 mL of 2216MB in deep-well blocks and processed in bulk. Genome libraries were prepared for sequencing using the Nextera DNA Library Preparation Kit (Illumina) with 1-2 ng input DNA per isolate, as previously described (Baym et al., 2015). Genomes were sequenced on 100 bp pared-end sequencing runs using Illumina HiSeq, with 50-60 samples multiplexed per lane. When available, Illumina HiSeq short reads from previous work (Kauffman et al., 2018b) were used, otherwise new short read data were generated for this study.

High quality bacterial genomic DNA for PacBio sequencing was prepared separately. A single colony was inoculated into a 250 mL Erlenmeyer flask with 50 mL of 2216MB and grown shaking at room temperature for 24 hours. The fresh culture was

pelleted by centrifugation at 5,000 x g for 20 minutes and then immediately processed for DNA extraction when possible, or frozen at -20°C for short term storage. The Qiagen Genomic Tip 500x kit was used following the manufactures guidelines, and the final DNA was collected by spooling using a glass rod rather than centrifugation to avoid shearing. DNA was stored in 500 μL of elution buffer at 4°C for 24-48 hours to allow for full resuspension before sequencing at either the Yale Center for Genome Analysis (PacBio RS II, without multiplexing) or the BioMicroCenter at MIT (Sequel, with multiplexing).

A custom hybrid assembly pipeline was designed to process the data. Briefly, Pacbio reads were filtered at different length cutoffs using Filtlong (noa, a) and then assembled using Flye (Kolmogorov et al., 2019) to create a set of reference genomes. The reference genomes were visualized using Bandage (Wick et al., 2015) and the best genome was selected, based on completion and coverage, to be used as a reference in the final assembly. For the final assembly, Illumina reads were trimmed using Trim-Galore (noa, b), Pacbio reads were quality filtered with the trimmed Illumina reads using Filtlong, and both sets of processed reads, together with the best Flye assembly, were used as inputs for the Unicycler (Wick et al., 2017) assembler.

Genomes were annotated using Prodigal 2.6 (Hyatt et al., 2010) for Open Reading Frame (ORF) prediction. Predicted ORFs were annotated using InterProScan5 (Jones et al., 2014) using the iprlookup, goterms, and pathways options. InterProScan5 matches against 13 databases by default, which are listed here: https://github.com/ebi-pf-team/interproscan/wiki/HowToRun#included-analyses. Two optional databases were included for this analysis: TMHMM for predicted transmembrane proteins and SignalP for predicted signal peptide cleavage sites.

3.4.5 Host relationships

Phylogenetic relationships among the 19 "orange" and "purple" clones were estimated by comparison of (i) concatenated alignments of ribosomal proteins and *hsp60* as described above in methods for, "Phage host-range matrix" (Fig. S1A), (ii) nucleotide sequences of 52 core ribosomal proteins (Fig. 1A, left) (Yutin et al., 2012), and (iii)

all shared genes (Fig. S3). For ribosomal protein comparisons, we searched for ribosomal proteins in the different genomes using HMMER (Eddy, 2011), filtered the hits using custom python scripts, aligned the hits using MAFFT (Katoh and Standley, 2013), concatenated the alignment using custom python scripts, and constructed the tree using RAxML (parameters raxmlHPC-PTHREADS -f a -x 26789416 -m GTRGAMMAX -p 218957 -# 100) (Stamatakis, 2014). For estimation of whole genome relationships, we used the Parsnp program (Treangen et al., 2014) with the recombination flag (-x) to construct whole genome SNP trees. Then, HarvestTools was used to convert from a ggr format to a snp fasta file, and finally, IQ-Tree was used to optimize the final tree (Fig. S3) (Nguyen et al., 2015). SNPs in the core genome were located using custom python scripts and verified by visualizing on Ginger, the Harvest graphic user interface. SNP details are given in Table S1.

3.4.6 Host range assays at varying phage concentrations

In order to determine the host-ranges of the specific phages used in this work at a higher resolution, we re-assayed a subset of hosts and phages of interest using a range of concentrations. Bacterial hosts were grown in 5 mL of 2216MB overnight from single colonies streaked on 1.5% Bacto Agar plates supplemented with 2216MB. Phage lysates were prepared as described above and diluted in 2216MB to form a ten-fold dilution series from 10^0 to 10^{-7}. Five μL drop spots of each dilution were pipetted onto bacterial host lawns made using a tube-free agar overlay method (Kauffman and Polz, 2018) and incubated at room temperature for 24 hours before evaluating phage entry and efficiency of plating at the varying concentrations (Fig.1A). Plates with different killing were imaged using a flatbed scanner (Epson Perfection V800 Photo Scanner - Product No. B11B22320) and captured using VueScan Software by Hamrick (Fig. S4). Lysis from without was only observed when using this higher resolution assay, thus we recommend performing such an assay when evaluating phage host-range whenever possible.

3.4.7 Phage adsorption assay

In order to determine if all phages were attaching to all hosts, for "orange" phage 1.143.O and "purple" phage 1.281.O, we compared the number of free phages remaining in solution after exposure to "orange" host 10N.261.55.C8, "purple" host 10N.286.54.F7, an unrelated Vibrio (outgroup) control 10N.261.49.C11, and a no-host negative control (2216MB). Three different colonies of each bacterial strain were inoculated in 3 mL of 2216MB and grown shaking at 25°C for 4 hours. Bacterial concentration was estimated at optical density measured at 600 nm wavelength (OD600), and each replicate was normalized to OD600 of 0.3 followed by 100-fold dilution. One mL of each diluted culture was aliquoted into individual wells of a 96-well culture block and bacteria were grown shaking at room temperature for another 3.5 hours to reach mid-exponential phase. Twenty μL of phage lysate was added to each well at varying concentrations (ranging from 0.001 to 10 phages/bacteria on average) and staggered in time to achieve an adsorption time of 30 minutes (Fig. S5). After allowing phages to adsorb, 200 μL of the phage and bacteria mixture was filtered using a 96-well filter system (Millipore MultiScreen Vacuum Manifold) to remove bacteria and any infecting or adsorbed phages. Five μL of a ten-fold dilution series of each well was then drop spotted onto a fresh lawn of a sensitive host (orange strain 10N.261.55.C8 was used for experiments with 1.143.O and 10N.286.54.F7 was used for experiments with 1.281.O) made in rectangular petri dishes (1-well Nunc Rectangular Dishes, Polystyrene, Sterile by Thermo Scientific - Supplier No. 267060). Plates were incubated at room temperature for 18-24 hours and then imaged using a flatbed scanner as described above (Fig. S5). Phage adsorption was estimated by comparing the number of plaque forming units (PFUs) in each dilution series to the no host control. For example, in Figure S5A the same order of magnitude of PFUs is evident in both the outgroup and the no-host control, meaning there is no phage adsorption for the outgroup. Yet, there are an order of magnitude more PFUs in both controls compared to the "purple" and "orange" hosts, implying equal adsorption is seen on the "purple" and "orange" hosts.

3.4.8 Bacterial strain selection and growth conditions for transposon mutagenesis and gene deletions

Because the adsorption assays indicated that both orange and purple phages adsorbed to both host groups, we chose one strain, 10N.261.55.C8 (Orange WT, hereafter C8-WT), for mapping of receptors for both host groups. For receptor mapping, we took advantage of the lysis from without phenotype where phages can effect lysis if hosts possess a specific receptor even if no viable phage are produced. Accordingly, at high phage titer, cells of both host groups are lysed by both phages (Fig. S4), allowing for testing of receptors using a "purple" phage on an "orange" host. The same C8-WT strain was used for characterization of resistance determinants of the "orange" host group by gene deletion (see below).

C8-WT was routinely grown at 25°C in 2216MB or TSB2. The *Escherichia coli* strains were grown in BD Difco Miller Luria-Bertani broth (LB) at 37°C and supplemented for auxotroph strain *E. coli* π3813 with thymidine (0.3 mM), and for strains *E. coli* β3914 and MFDpir with diaminopimelic acid (dapA) (0.3 mM). Antibiotics were used at the following concentrations: erythromycin (Erm) 200 μg/mL, kanamycin (Km) 50 μg/mL and chloramphenicol (Cm) at 5 or 25 μg/mL for *Vibrio* and *E. coli*, respectively.

3.4.9 Receptor identification using transposon mutagenesis

To map phage receptors, transposon mutagenesis was carried out using suicide delivery of a mariner transposon. C8-WT served as recipient and the dapA deficient strain *E. coli* MFDpir as donor with the suicide conjugative plasmid pSC189-Cm (FerriÃÍres et al., 2010) (Table S5). The delivery plasmid (pSC189) can be mobilized via RP4-mediated transfer and it carries the hyperactive C9 mariner transposase (Chiang and Rubin, 2002). Conjugation was carried out by mating assays as described previously (Le Roux et al., 2007) with some modifications. First, donor:recipient ratio was adjusted to 1:3. Overnight cultures were diluted 1:100 in fresh media and grown up to an OD600 of ∼0.4. One mL was separately pelleted at 5,500 x *g* for 2 minutes

and washed in pre-warmed mating media broth MMB-1 (TSB supplemented with 1% NaCl plus dapA) to remove antibiotics and/or residual media. This wash step was repeated twice. Washed pellets were subsequently mixed in the same tube with 500 μL of MMB-1, pelleted and resuspended in a mating spot (20 μL) on mating media agar plates and incubated at 25°C for 18 hours. Mating spots were collected using a Nunc 10 μL sterile plastic inoculation loop and resuspended in 500 μL of ASW. Then, 100 μL of this suspension were spread onto TSB2 plates supplemented with Cm and incubated at 25°C for 48 hours. Finally, the mutant library (totaling 26,662 mutants) was archived in 500 μL aliquots with ASW supplemented with Cm and 25% glycerol (v/v), quickly frozen in a dry ice bath for 10 minutes, and stored at -80°C until testing.

Resistant mutants were selected by challenging the mutant library with high titers of phages. Four aliquots of the mutant library were defrosted, centrifuged by pelleting at 5,000 x g for 5 minute, and then washed with 2216MB to remove any residual glycerol. This wash step was repeated twice, and the washed pellets were then resuspended in their original tube with 1 mL of fresh 2216MB supplemented with Cm. C8-WT served as positive control and was treated equivalently, except for the addition of Cm. The washed mutant library and C8-WT control were both diluted 1:10 in 2216MB and then incubated at room temperature with vigorous shaking (250 rpm) for 1 hour until the cultures reached early exponential phase. To select for phage-resistant mutants, lysates was serially diluted 10-fold and mixed with the mutant library and C8-WT cultures. Aliquots of host-phage culture were mixed into 750 μL of 2216MB top agar (with and without Cm as needed) and spread on large 2216MB bottom agar plates (with and without Cm as needed) following the agar overlay protocol described previously (Kauffman and Polz, 2018). After incubating at room temperature for 48 hours, $\sim$100 phage-resistant colonies were selected at random and serially re-streaked three times on 2216MB agar plates (with and without Cm as needed). Glycerol stocks of each mutant were archived and all mutants were then re-tested for phage susceptibility. The re-test was always done with two "orange" and two "purple" phages, one of each always being the original phage used to isolate

resistant colonies. In all cases, resistance to one "orange" phage yielded resistance to all "orange" phages and resistance to one "purple" phage yielded resistance to all "purple" phages. Cross resistance to opposite or both phage groups was never seen, further supporting the finding that each group of phages uses a different receptor.

Arbitrary PCR (Das et al., 2005) was used to map the transposon insertions in resistant strains. Genomic DNA from each phage-resistant mutant was extracted with Lyse-n-Go direct PCR reagent (Thermo Scientific), and 1 μL of the lysate served as template in arbitrary PCR. This method involved two rounds of PCR amplification (Das et al., 2005): in the first round, genomic DNA was amplified with a fully degenerate primer SS9arb2 (Table S6) containing a 5' tail of known sequence to be used for specific amplification in the second round of PCR (Lauro et al., 2008), paired with primer Mar4 (Table S6) that binds the end of the transposon TnSC189 (Jiao et al., 2005). Optimized conditions for the first round PCR consisted of the following reagent concentrations and amplification parameters: primers SS9arb2 and Mar4 were at 0.5 mM and 0.2 mM, respectively; GoTaq G2 HotStart (Promega) was used with MgCl2 at 2 mM; initial heating for 2 minutes at 95°C, followed by 6 cycles of 30 seconds at 95°C, 30 seconds at 30°C, and 1 minute and 30 seconds at 72°C; 30 cycles of 30 seconds at 95°C, 30 seconds at 55°C and 1 minute and 30 seconds at 72°C, with a final extension for 5 minutes at 72°C. In the second round of PCR amplification, 2.5 μL of the first round PCR product was used as template, combined with a nested primer within the amplified fragment of TnSC189 (Mar4_int2) and primer (Arb3) with sequence identity to the 5' tail of the SS9arb2 (Table S6). For the second round, PCR reagents were used as described above but using both primer concentrations were 0.2 mM, and the PCR was run under the following conditions: 2 minutes at 95°C, 30 cycles of 30 seconds at 95°C, 30 seconds at 58°C and 1 minute and 30 seconds at 72°C, with a final extension time of 5 minutes at 72°C. PCR products were verified by electrophoresis, purified by spin-column using QIAquick PCR Purification (Qiagen) and then Sanger-sequenced. Finally, amplicons were trimmed and mapped to the C8-WT genome to identify transposon insertion locations using a custom python script and BLASTn (Camacho et al., 2009). Hits are present in Table

S2.

3.4.10 Receptor verification using re-sequencing of spontaneously resistant isolates

As an independent method to transposon mutagenesis, to identify phage receptors, we re-sequenced spontaneously resistant mutants from co-cultures of orange host C8-WT and high titer phages. C8-WT was streaked out from glycerol stock onto 2216MB agar plates, inoculated into 5 mL of 2216MB liquid media, gown shaking overnight at room temperature, and plated as a lawn in a soft agar overlay. Five μL drop spots of a phage dilution series were plated on top of the agar, and after 24 hours, resistant colonies that grew in the presence of high phage concentrations were re-streaked three times and archived. For each phage, 10 colonies were archived. Resistant strains were re-streaked and re-tested to verify resistance. The re-test was always done with two "orange" and two "purple" phages, one of each always being the original phage used to isolate resistant colonies just as in the transposon experiments. The results were also consistent with the transposon mutagenesis experiments: In all cases, resistance to one "orange" phage yielded resistance to all "orange" phages and resistance to one "purple" phage yielded resistance to all "purple" phages. Cross resistance to opposite or both phage groups was never seen, further supporting the finding that each group of phages uses a different receptor. Six to seven strains verified in this way were sequenced on an Illumina HiSeq as described in the "Hybrid genome assemblies" section. Reads were trimmed and mapped to the hybrid assembly reference genome using CLC work bench 9. Single nucleotide polymorphisms (SNPs) and indels were identified using a custom pipeline made for CLC work bench 9 and are presented in Table S3. These SNPs were cross-referenced to receptor identification using transposon mutagenesis (see above).

3.4.11 Identification and annotation of putative PDEs in the flexible genome

In order to determine the differences in the flexible genome amongst the 19 "orange" and "purple" strains, we created a multiple alignment using Mugsy (Angiuoli and Salzberg, 2011), and performed a hierarchical clustering of the alignment blocks, greater than 500 bp, by length in R using hclust (R Core Team, 2019) (Fig.1A, right). The two groups clustered by their phage predation profile. This clustering was completely driven by the presence of 5 alignment blocks, three of which were exclusive to the "orange" strains, and two of which were exclusive to the "purple" strains. Upon further investigation of the alignment blocks by hand, we discovered that the alignment blocks corresponded to putative PDEs (Fig. 1B). Gene annotations of the PDEs were performed manually using the consensus obtained from HHPred (Zimmermann et al., 2018), InterProScan5 (Jones et al., 2014), Phyre2 (Kelley et al., 2015), and BLASTp (Altschul et al., 1997) databases tools (Data S1). The search was performed using default options except that HHPred search was performed against COG-KOG 1.0 and Pfam-A_v32.0 databases and that BLASTp was performed using the protein-protein BLAST option. Up to ten significant pfam and COG ($p < 0.05$) from HHPred search where used to compare each gene with pfam-COG accession numbers of phage defense systems from supplementary Table 1 from Doron et al. (Doron et al., 2018). To further characterize the distribution of the PDEs identified among the 19 strains in a larger collection of *Vibrio* genomes (Arevalo et al., 2019), we used BLASTn (Camacho et al., 2009) and custom python scripts to identify the distribution of the mobile elements (Fig. S6). Because when comparing long mobile elements BLASTn will often return multiple overlapping ranges of identity, our BLAST parsing script merges overlapping sequence ranges to avoid over-counting regions within a genome. A PDE was considered present in a genome if at least 80% of the element was present at over 95% identity.

3.4.12 Methylation profiling

In order to discover if the restriction modification systems identified on the PDEs in the "orange" and "purple" strains are active, we determined the methylation sites in both phage-host pairs as outlined in Murray et al. (Murray et al., 2012). Briefly, we submitted the host genomes to REBASE, a well-curated database of restriction modification systems which allows motif prediction based on comparisons to known enzyme-motif pairs (Roberts et al., 2015). Then, we combined the motif data with the methylome data generated using the Base Modification Detection and Motif Analysis pipelines available on the single molecule real-time (SMRT) sequencing portal from Pacific Biosciences. Summary data is presented in Table S4.

3.4.13 Phage defense element knockouts using two-step allelic exchange

To test whether putative PDEs were responsible for phage resistance in C8-WT, we knocked out large portions of each PDE containing genes annotated as being related to phage defense (Fig. S6), in all possible combinations. We found it was possible to knock out nearly all of PDE1 (93.5%), but had to leave in part of the elementâĂŹs Yfbr gene as it replaces the host Yfbr gene upon insertion. For PDE2, we found that knocking out the entire element was not possible in a single step, likely because of the toxicity effects of deleting the entire TA system at once. We therefore proceeded to make a partial deletion (58.8%) from the toxin gene to the 5' end of the element, leaving the antitoxin intact, along with other genes predicted to play roles in insertion/mobilization of the element (integrases, transposases, and recombinases) as indicated by structure and function annotations using HHpred (SÃűding et al., 2005), InterProScan5 (Jones et al., 2014), and Phyre2 (Kelley et al., 2015) (Fig. S6, Data S1). Noting PDE3 also has a putative TA system, we followed the same approach as for PDE2 and made a partial deletion (68.8%) from the toxin gene to the 5' end of the element. The details of the approach are summarized in Figure S6.

Site-directed mutagenesis was used for all deletions. Cloning was carried out using

the New England Biolabs Gibson Assembly Master Mix according to the manufacturer's protocol. Fragments upstream and downstream of the portion of the element to be deleted were separately PCR-amplified using primers specified in Table S6. A third PCR reaction was carried out to amplify the backbone of the plasmid pSW7848T (Val et al., 2012) (Table S5) with primer pairs pSW_F&R (Table S6).These amplicons were cut with Dpn1 (2 hours at 37°C) to inactivate the plasmid template before setting up the Gibson assembly reaction. In all cases, PCR products were verified by gel electrophoresis, purified by spin-column as described above, and DNA concentration was determined using Nanodrop 2000 (Thermo Scientific). Subsequently, 0.03 pmol of each vector was assembled with 0.07 pmol of its specific downstream and upstream DNA fragments at 50°C for 60 minutes. After completion of this reaction, DNA was desalted by dialysis on a 0.0025 μM filter (Millipore) before electroporation into *E. coli* π3813, which was used as a plasmid host for cloning (Roux et al., 2007). Finally, the plasmid DNA was purified, verified by Sanger sequencing, and electroporated into *E. coli* β3914 to be uses as a plasmid host for conjugation (Roux et al., 2007) (Table S5). Conjugation was carried out in a mating spot as described above for the transposon mutagenesis but with some modifications: donor:recipient ratio was changed to 3:1, the mating media was altered to MMB-2 (TSB supplemented with 2% NaCl plus dapA), and the mating spot was incubated at 30°C. Counter-selection of ΔdapA donor was performed by plating on TSB2 agar plates without dapA but supplemented with Cm and glucose 1% (w/v). Antibiotic-resistant colonies are due to the integration of the entire plasmid (CmR) in the chromosome by a single crossover. Colonies were picked, re-grown in liquid media (TSB2) supplemented with Cm and glucose 1% (w/v) to late logarithmic phase and spread on BD Bacto TSB without Dextrose plates supplemented with 2% NaCl (w/v) and 0.2% arabinose. To verify deletions in the single PDE mutants ΔPDE1, ΔPDE2 and ΔPDE3 (Table S5) PCR products generated using primers flanking externally the different regions targeted (ΔPDE1/F&R; ΔPDE2/F&R and ΔPDE3/F&R) (Table S6) were sequenced by Sanger. This procedure was also used to construct double (ΔΔPDE12; ΔΔPDE13; ΔΔPDE23) and triple mutants (ΔΔΔPDE123) but using a single or double mutant as final recipient

during the conjugation step (Table S5).

3.4.14 Phage susceptibility assay

In order to test the susceptibility of the "orange" PDE deletion mutants to "purple" phages, we challenged each mutant with representative "purple" phage 1.281.O in agar overlays (Fig. 2) and in liquid culture (Fig. S7). For the mutant testing in agar overlays, we used the same protocol outlined in the "Host range assays at varying phage concentrations" section above, with one additional step: after the plaques were imaged, we re-streaked the phages from the highest concentration drop spot onto fresh bacterial (host or mutant) lawns to test for phage propagation (Fig. 2B). For the liquid assay, we streaked out each mutant onto 2216MB agar plates, allowed 48 hours for large colonies to form, and inoculated each mutant into 3 mL of 2216MB in triplicate. After growing the cultures shaking at 25°C for 4 hours, we normalized 1 mL of the culture to an OD600 of 0.3, diluted it 100x into a final volume of 5 mL, and aliquoted 200 μL into 12 wells each of a 96-well clear bottom Micro-titer plate (Falcon). A Tecan Microplate Reader with Spark software was used to maintain the cultures shaking at 25°C, monitoring OD600 every 15 minutes. Once OD600 reached 0.3, "purple" phage 1.281.O was added in triplicate at different concentrations to reach the desired multiplicities of infection (Fig. S7B), after which the cultures were returned to the plate reader for the remainder of a 24 hour run. This experiment was run with two mutants, C8-WT, and "purple" host 10N.286.54.F7 each time until all mutants had been tested. Identification of putative PDEs from comparison of closely related genomes of different bacteria We extended the search for novel putative PDEs by searching identify nearly clonal genomes of *Vibrio*, *Salmonella*, and *Listeria*. For the 23 *Vibrio* strains, we selected only those with identical ribosomal proteins. For *Salmonella* and *Listeria*, we selected genomes within the same ribotype using the ribosomal MLST database (https://pubmlst.org/rmlst/), filtered to only include NCBI assemblies, and downloaded genomes from ribotypes with more than 20 members. We used ribotype 8354 for *Salmonella* and MLST strain type 5 for *Listeria* to assay putative PDE distribution in an exemplary manner. In all three

cases, the final set of genomes was run though a custom kmer-based comparative genomic pipeline to identify flexible regions. All pairs of genomes were compared. First, each genome was split into 31-mers using Jellyfish (MarÃğais and Kingsford, 2011), then shared kmers between the genomes being compared were removed and only unique kmers were mapped back to the reference genomes they originated from using Bowtie2 (Langmead and Salzberg, 2012). Any unique region greater than 1,000 bp was kept and a gap of 3,000 bp was allotted to account for genes that may have been shared between the two genomes splitting a complete region. Regions were checked for duplication and the largest region of any overlapping regions was saved. We then used Mash (Ondov et al., 2016) to compare all the unique regions to each other and clustered any region greater than 5kbp with minimum Jaccard similarity of 0.95. We visualized the clustering using Gephi (Bastian et al., 2009) and then chose one representative from each cluster by hand to make a final list of unique regions. We then used BLASTn (Camacho et al., 2009) and custom python scripts to determine which genomes harbored which elements with >95% identity and >80% length considered a match. We removed any element that appeared in all genomes. Finally, we used HMMER (Eddy, 2011) to search each element for known defense genes using Supplementary Table 1 in Doron et al. (Doron et al., 2018). Any unique region with one or more hits was considered to be a putative PDE and depicted in Figure S8, Figure S10, and Figure S11 for *Vibrio*, *Listeria*, and *Salmonella*, respectively.

3.4.15 Proportion of known defense genes in the flexible genome across diverse *Vibrio* populations

To determine the proportion of known defense genes in other *Vibrio* species, we used the species and population designations from Arevalo et al., (Arevalo et al., 2019). We based our identification of flexible genes on the method described in Arevalo et al. ORFs were identified with Prodigal 2.6 (Hyatt et al., 2010) and orthologous genes were clustered using MMseqs2 (Steinegger and SÃűding, 2017). Flexible genes for a given population were defined as orthologs which were present in at least one

member but not present in all members of the population. Flexible genes from each genome were then used as a database which we searched for known defense genes using Supplementary Table 1 in Doron et al. (Doron et al., 2018) using HMMER (Eddy, 2011). Total length of all flexible genes summed for each species and the proportion of genes (by length) with a hit to a known defense gene is shown in Figure S9.

3.4.16 Distribution of putative receptor genes across diverse *Vibrio* populations

To test how diverse receptor genes identified by mutant analysis in *V. lentus* are across other vibrios, we used the species and population designations from Arevalo et al., (Arevalo et al., 2019). In cases where a species was composed of multiple populations, we chose to only analyze the population with the most members. We based our identification of core genes on the method described in Arevalo et al. ORFs were identified with Prodigal 2.6 (Hyatt et al., 2010) and orthologous genes were clustered using MMseqs2 (Steinegger and SÃűding, 2017). Core genes for a given population were defined as orthologs which were present in a single copy in all members of that population. We aligned all genes within orthologous clusters using MUSCLE (Edgar, 2004) and calculated the amino acid diversity between e-identical amino acids divided by the alignment length. Average pairwise amino acid diversity was obtained by taking the mean diversity across all pairs of genes within an orthologous cluster. The amino acid sequences of each receptor gene were aligned using mafft-linsi (Katoh and Standley, 2013) and average pairwise amino acid diversity was calculated as described above. Phylogenetic relationships among receptors identified using genetic approaches were determined across a collection of diverse *Vibrio* isolates

3.5 Acknowledgments

We thank D. Newman, K. Costa, H. Wildschutte, and F. Le Roux for advice and guidance with mutagenesis experiments. We thank M. Cutler for experimental support. We thank B. Cervantes for technical assistance with plate imaging. We thank S. W. Chisholm and L. Kelly for thoughtful comments. We thank O. X. Cordero, S. Kearney, and A. F. Takemura for valuable suggestions throughout. Funding: This work was supported by the Simons Foundation (Life Sciences Project Award-572792), the National Science Foundation Division of Ocean Sciences (OCE-1435868), and an MIT J-WAFS seed grant. Support for F.A.H. was provided by the NSF GRFP and the MIT Martin Society of Fellows for Sustainability. Support for J.D. was provided by a postdoctoral fellowship from Xunta de Galicia (ED481B 2016/032). Author contributions: F.A.H. and M.P. conceived of the project, designed the study, and wrote the paper with contributions from all coauthors. F.A.H. performed the experiments with assistance from J.D. and M.M.. F.A.H., J.D., and B.R. designed the genetic knockouts and conducted manual annotations. J.D. made the knockout mutants. F.A.H. curated the data, conducted formal analyses, and prepared the figures with feedback from all coauthors. F.A.H., J.E., M.M., P.A., and D.V. wrote and/or optimized code for the bioinformatic pipelines. K.K. isolated and performed initial characterization of the lytic viruses. M.P. supervised the project and secured funding. Competing interests: Authors declare no competing interests. Data and materials availability: New genomes used in this work have been deposited under the NCBI BioProject with accession number PRJNA328102. All data, code, and materials are available upon request.

Bibliography

https://github.com/rrwick/Filtlong.

http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST

Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, 12(1):402.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Andersson, A. F. and Banfield, J. F. (2008). Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. *Science*, 320(5879):1047–1050.

Angiuoli, S. V. and Salzberg, S. L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics (Oxford, England)*, 27(3):334–342.

Arber, W. and Dussoix, D. (1962). Host specificity of DNA produced by Escherichia coli: I. Host controlled modification of bacteriophage Îż. *Journal of Molecular Biology*, 5(1):18–36.

Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J., and Polz, M. F. (2019). A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. *Cell*, 178(4):820–834.e14.

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., and Lindell, D. (2011). Genomic island variability facilitates Prochlorococcus-virus coexistence. *Nature*, 474(7353):604–608.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., and Horvath, P. (2007). CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science*, 315(5819):1709–1712.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Third International AAAI Conference on Weblogs and Social Media*.

Baym, M., Kryazhimskiy, S., Lieberman, T. D., Chung, H., Desai, M. M., and Kishony, R. (2015). Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. *PLOS ONE*, 10(5):e0128036.

Bertani, G. and Weigle, J. J. (1953). HOST CONTROLLED VARIATION IN BACTERIAL VIRUSES. *Journal of Bacteriology*, 65(2):113–121.

Bohannan, B. and Lenski, R. (2000). Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. *Ecology Letters*, 3(4):362–377.

Breitbart, M. and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in Microbiology*, 13(6):278–284.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, 10:421.

Chiang, S. L. and Rubin, E. J. (2002). Construction of a mariner-based transposon for epitope-tagging and genomic targeting. *Gene*, 296(1):179–185.

Das, S., Noe, J. C., Paik, S., and Kitten, T. (2005). An improved arbitrary primed PCR method for rapid characterization of transposon insertion sites. *Journal of Microbiological Methods*, 63(1):89–94.

DelbrÃijck, M. (1940). THE GROWTH OF BACTERIOPHAGE AND LYSIS OF THE HOST. *The Journal of General Physiology*, 23(5):643–660.

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, page eaar4120.

Dy, R. L., Richter, C., Salmond, G. P., and Fineran, P. C. (2014). Remarkable Mechanisms in Microbes to Resist Phage Infections. *Annual Review of Virology*, 1(1):307–331.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, 7(10).

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

FerriÃĺres, L., HÃľmery, G., Nham, T., GuÃľrout, A.-M., Mazel, D., Beloin, C., and Ghigo, J.-M. (2010). Silent Mischief: Bacteriophage Mu Insertions Contaminate Products of Escherichia coli Random Mutagenesis Performed Using Suicidal Transposon Delivery Plasmids Mobilized by Broad-Host-Range RP4 Conjugative Machinery. *Journal of Bacteriology*, 192(24):6418–6427.

Guy, L., Roat Kultima, J., and Andersson, S. G. E. (2010). genoPlotR: comparative gene and genome visualization in R. *Bioinformatics*, 26(18):2334–2335.

Hampton, H. G., Watson, B. N. J., and Fineran, P. C. (2020). The arms race between bacteria and their phage foes. *Nature*, 577(7790):327–336.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., and Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, 34(8):2115–2122.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., HernÃąndez-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., and Bork, P. (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–D314.

Hunt, D. E., David, L. A., Gevers, D., Preheim, S. P., Alm, E. J., and Polz, M. F. (2008). Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton. *Science*, 320(5879):1081–1085.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119.

Jiao, Y., Kappler, A., Croal, L. R., and Newman, D. K. (2005). Isolation and Characterization of a Genetically Tractable Photoautotrophic Fe(II)-Oxidizing Bacterium, Rhodopseudomonas palustris Strain TIE-1. *Applied and Environmental Microbiology*, 71(8):4487–4496.

John, S. G., Mendez, C. B., Deng, L., Poulos, B., Kauffman, A. K. M., Kern, S., Brum, J., Polz, M. F., Boyle, E. A., and Sullivan, M. B. (2011). A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environmental Microbiology Reports*, 3(2):195–202.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., and Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.

Katoh, K. and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780.

Kauffman, A. K. M. (2014). *Demographics of Lytic Viral Infection of Coastal Ocean Vibrio*. PhD thesis, Massachusetts Institute of Technology.

Kauffman, K. M., Brown, J. M., Sharma, R. S., VanInsberghe, D., Elsherbini, J., Polz, M., and Kelly, L. (2018a). Viruses of the Nahant Collection, characterization of 251 marine Vibrionaceae viruses. *Scientific Data*, 5:180114.

Kauffman, K. M., Hussain, F. A., Yang, J., Arevalo, P., Brown, J. M., Chang, W. K., VanInsberghe, D., Elsherbini, J., Sharma, R. S., Cutler, M. B., Kelly, L., and Polz, M. F. (2018b). A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature*, 554(7690):118–122.

Kauffman, K. M. and Polz, M. F. (2018). Streamlining standard bacteriophage methods for higher throughput. *MethodsX*, 5:159–172.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10(6):845–858.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5):540–546.

Koonin, E. V., Makarova, K. S., Wolf, Y. I., and Krupovic, M. (2019). Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nature Reviews Genetics*, pages 1–13.

Kortright, K. E., Chan, B. K., Koff, J. L., and Turner, P. E. (2019). Phage Therapy: A Renewed Approach to Combat Antibiotic-Resistant Bacteria. *Cell Host & Microbe*, 25(2):219–232.

Kruger, D. H. and Bickle, T. A. (1983). Bacteriophage survival: multiple mechanisms for avoiding the deoxyribonucleic acid restriction systems of their hosts. *Microbiological Reviews*, 47(3):345–360.

Labrie, S. J., Samson, J. E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nature Reviews Microbiology*, 8(5):317–327.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

Lauro, F. M., Tran, K., Vezzi, A., Vitulo, N., Valle, G., and Bartlett, D. H. (2008). Large-Scale Transposon Mutagenesis of Photobacterium profundum SS9 Reveals New Genetic Loci Important for Growth at Low Temperature and High Pressure. *Journal of Bacteriology*, 190(5):1699–1709.

Le Roux, F., Binesse, J., Saulnier, D., and Mazel, D. (2007). Construction of a Vibrio splendidus Mutant Lacking the Metalloprotease Gene vsm by Use of a Novel Counterselectable Suicide Vector. *Applied and Environmental Microbiology*, 73(3):777–784.

Lindahl, G., Sironi, G., Bialy, H., and Calendar, R. (1970). Bacteriophage Lambda; Abortive Infection of Bacteria Lysogenic for Phage P2. *Proceedings of the National Academy of Sciences*, 66(3):587–594.

Makarova, K. S., Wolf, Y. I., Snir, S., and Koonin, E. V. (2011). Defense Islands in Bacterial and Archaeal Genomes and Prediction of Novel Defense Systems. *Journal of Bacteriology*, 193(21):6039–6056.

Martin-Platero, A. M., Cleary, B., Kauffman, K., Preheim, S. P., McGillicuddy, D. J., Alm, E. J., and Polz, M. F. (2018). High resolution time series reveals cohesive but short-lived communities in coastal plankton. *Nature Communications*, 9(1):266.

MarÃğais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770.

McDonald, N. D., Regmi, A., Morreale, D. P., Borowski, J. D., and Boyd, E. F. (2019). CRISPR-Cas systems are present predominantly on mobile genetic elements in Vibrio species. *BMC Genomics*, 20(1):105.

McKitterick, A. C. and Seed, K. D. (2018). Anti-phage islands force their target phage to directly mediate island excision and spread. *Nature Communications*, 9(1):2348.

Meier-Kolthoff, J. P. and GÃűker, M. (2017). VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics*, 33(21):3396–3404.

Meyer, J. R., Dobias, D. T., Weitz, J. S., Barrick, J. E., Quick, R. T., and Lenski, R. E. (2012). Repeatability and Contingency in the Evolution of a Key Innovation in Phage Lambda. *Science*, 335(6067):428–432.

Molineux, I. J. (1991). Host-parasite interactions: recent developments in the genetics of abortive phage infections. *The New biologist*, 3(3):230–236.

Murray, I. A., Clark, T. A., Morgan, R. D., Boitano, M., Anton, B. P., Luong, K., Fomenkov, A., Turner, S. W., Korlach, J., and Roberts, R. J. (2012). The methylomes of six bacteria. *Nucleic Acids Research*, page gks891.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Novick, R. P., Christie, G. E., and PenadÃĺs, J. R. (2010). The phage-related chromosomal islands of Gram-positive bacteria. *Nature Reviews Microbiology*, 8(8):541–551.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roberts, R. J., Vincze, T., Posfai, J., and Macelis, D. (2015). REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research*, 43(Database issue):D298–299.

Rocha, E. P. C. (2018). Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. *Molecular Biology and Evolution*, 35(6):1338–1347.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., PaÅąiÄĞ, L., Thingstad, T. F., Rohwer, F., and Mira, A. (2009). Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology*, 7(11):828–836.

Roux, F. L., Binesse, J., Saulnier, D., and Mazel, D. (2007). Construction of a Vibrio splendidus Mutant Lacking the Metalloprotease Gene vsm by Use of a Novel Counterselectable Suicide Vector. *Applied and Environmental Microbiology*, 73(3):777–784.

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., SzabÃş, G., Polz, M. F., and Alm, E. J. (2012). Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science*, 336(6077):48–51.

Snyder, L. (1995). Phage-exclusion enzymes: a bonanza of biochemical and cell biology reagents? *Molecular Microbiology*, 15(3):415–420.

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

Steinegger, M. and SÃűding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028.

SÃűding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 33(Web Server issue):W244–W248.

Treangen, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, 15(11):524.

Val, M.-E., Skovgaard, O., Ducos-Galand, M., Bland, M. J., and Mazel, D. (2012). Genome Engineering in Vibrio cholerae: A Feasible Approach to Address Biological Issues. *PLOS Genetics*, 8(1):e1002472.

Westra, E., vanÂăHoute, S., Oyesiku-Blakemore, S., Makin, B., Broniewski, J., Best, A., Bondy-Denomy, J., Davidson, A., Boots, M., and Buckling, A. (2015). Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Current Biology*, 25(8):1043–1049.

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, 13(6):e1005595.

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20):3350–3352.

Wilson, G. G. and Murray, N. E. (1991). Restriction and Modification Systems. *Annual Review of Genetics*, 25(1):585–627.

Winter, C., Bouvier, T., Weinbauer, M. G., and Thingstad, T. F. (2010). Trade-Offs between Competition and Defense Specialists among Unicellular Planktonic Organisms: the âĂIJKilling the WinnerâĂİ Hypothesis Revisited. *Microbiology and Molecular Biology Reviews*, 74(1):42–57.

Wommack, K. E. and Colwell, R. R. (2000). Virioplankton: Viruses in Aquatic Ecosystems. *Microbiology and Molecular Biology Reviews*, 64(1):69–114.

Yutin, N., PuigbÃš, P., Koonin, E. V., and Wolf, Y. I. (2012). Phylogenomics of Prokaryotic Ribosomal Proteins. *PLoS ONE*, 7(5).

Zborowsky, S. and Lindell, D. (2019). Resistance in marine cyanobacteria differs against specialist and generalist cyanophages. *Proceedings of the National Academy of Sciences*, page 201906897.

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., KÃijbler, J., Lozajic, M., Gabler, F., SÃűding, J., Lupas, A. N., and Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology*, 430(15):2237–2243.

3.6 Figures

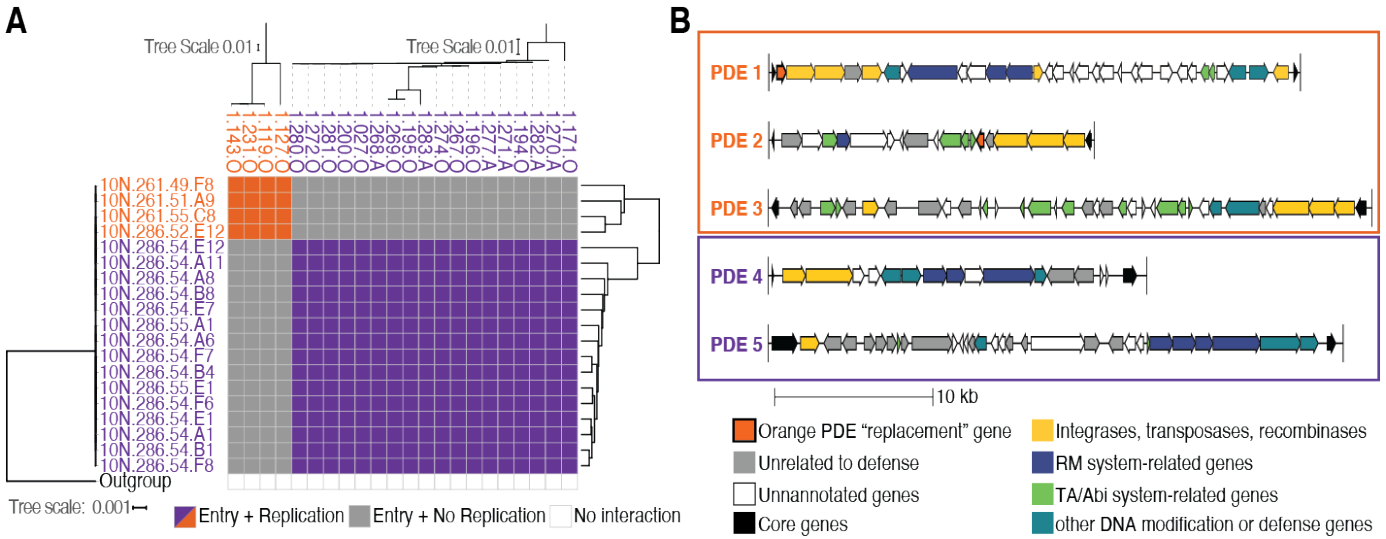


Figure 3-1: **Fig.1: Near-clonal strains of *Vibrio lentus* differ in sensitivity to phage predation and differ in the carriage of mobile genetic elements encoding for phage-defense genes.** (A) Phage host-range matrix with rows representing bacterial strains and columns representing phages. (A-left) Phylogenetic tree of 52 concatenated ribosomal protein sequences, (A-top) whole genome tree of viruses, (A-right) hierarchical clustering of whole genome alignments of bacterial hosts. (B) Gene diagrams of mobile genetic elements specific to the two host groups (as indicated by orange or purple outlines).

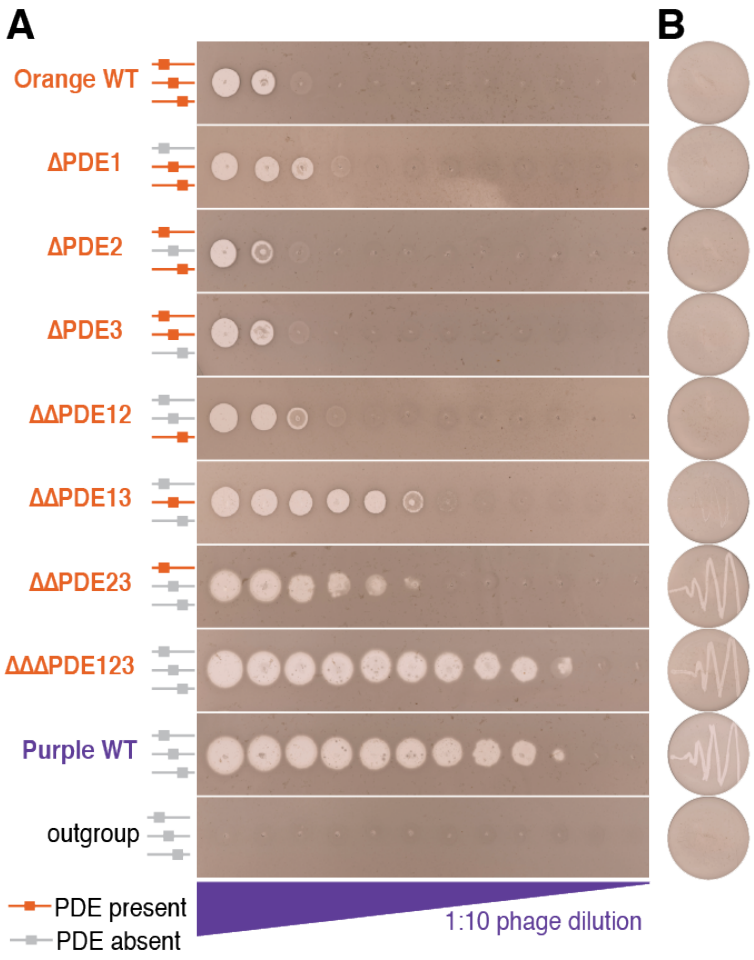


Figure 3-2: **Fig.2: Changes in susceptibility to phage killing observed for phage-defense element (PDE) markerless deletions.** (A) Lawns of bacterial hosts with drop spots of a 1:10 dilution series of "purple" phage (1.281.O). Cartoons on left indicate the presence or absence of different PDEs in each strain. From top to bottom: "orange" wild type host (10N.261.55.C8), Δ PDE1, Δ PDE2, Δ PDE3, ΔΔ PDE12, ΔΔ PDE13, ΔΔ PDE23, ΔΔΔ PDE123, "purple" wild type host (10N.286.54.F7, positive control), outgroup (10N.261.49.C11, negative control). (B) Re-streak test for propagation of phage progeny from drop spot clearings. Only infections of ΔΔ PDE23, ΔΔΔ PDE123, and "purple" wild type hosts produce viable phages, indicated by secondary clearing on the re-streak plates.

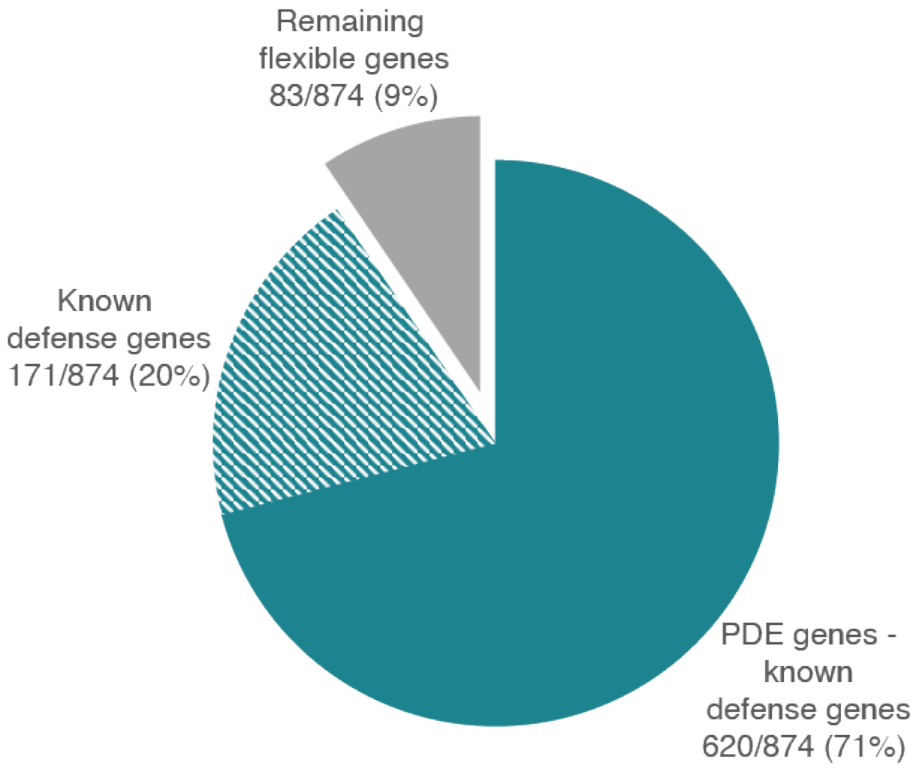


Figure 3-3: **Fig.3: Fraction of the bacterial flexible genome attributed to phage defense.** Amongst the 23 clones, an all-by-all genomic comparison shows 91% of flexible regions greater than 5kbp are putative PDEs. Only 20% of the PDEs match known defense genes while the remining are other PDE-specific genes, many of which are unannotated (71%).

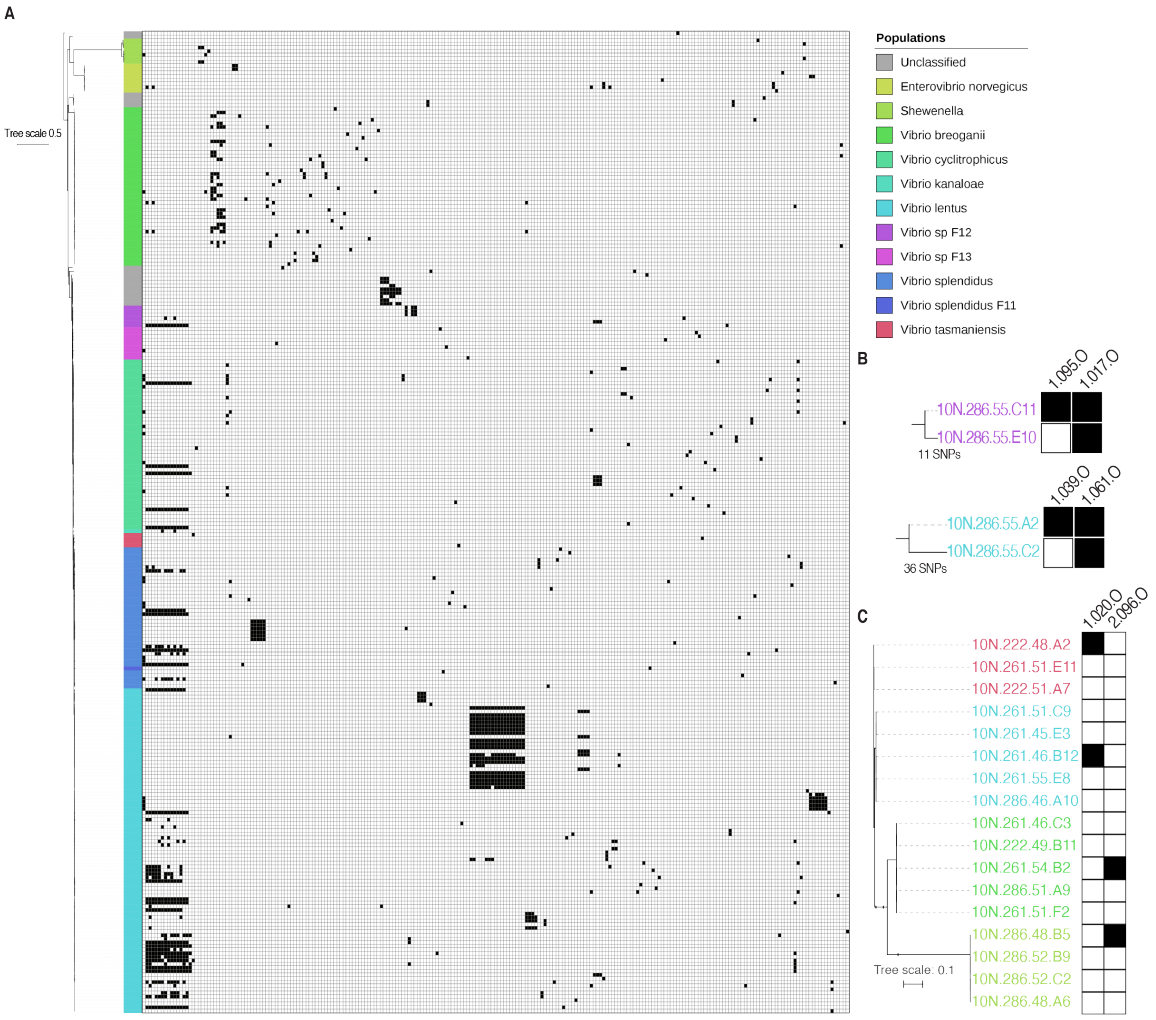


Figure 3-4: **Fig.S1: Phage host-range established using an exhaustive cross-test matrix.** (A) Full matrix with rows depicting bacterial hosts organized by the phylogeny of their ribosomal protein and *hsp60* gene sequences (proxy for core genome), and columns depicting phages ordered by protein similarity identity [modified from Figure 2 in (Kauffman et al., 2018b)]. (B) Closest bacterial relatives differing in phage sensitivity profiles can be distinguished by only few SNPs across their entire core genomes. Trees represent full genome alignments, phage identification codes written above columns, black boxes indicate positive infection determined by plaque assay. (C) Broad host-range phages, defined as host ranges spanning different species, remain strain-specific within different species. Phylogenetic tree constructed using same alignment of core genes as in A, and infection representation analogous to that in B.

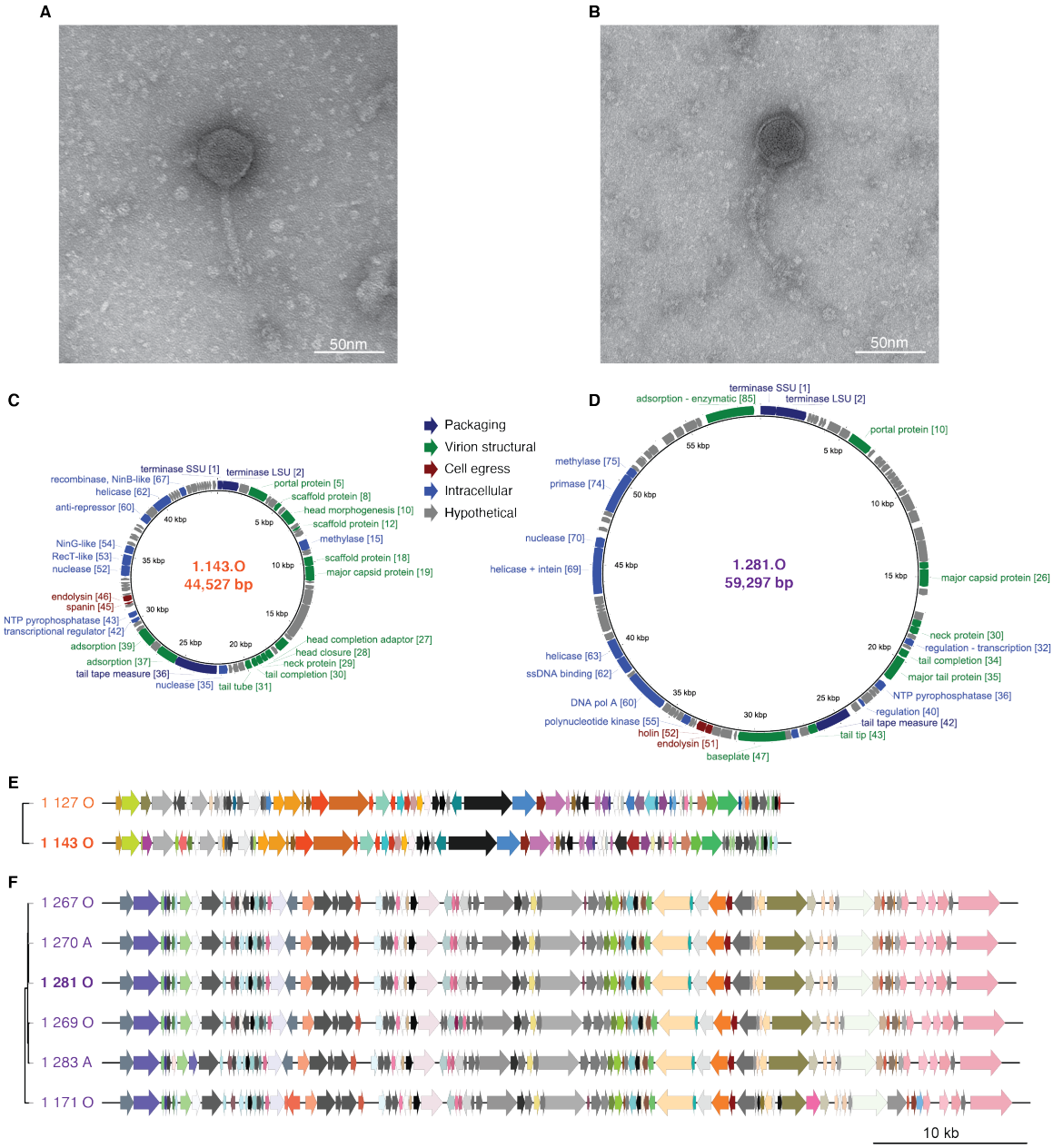


Figure 3-5: **Fig.S2: "Orange" and "purple" phages represent divergent groups of siphoviruses.** (A, B) Electron microscopy of phages representative of "orange" and "purple" groups suggests that both are siphoviridae, with long non-contractile tails. (C, D) Genome characterization of phages representative of "orange" and "purple" groups, respectively, shows that they differ in size by nearly 15 kbp; numbers adjacent to annotations reflect GenBank locus tag. (E, F) Clustering and alignment of phage genomes show that they represent two distinct genus-level groupings. While within each group gene synteny and content are conserved, no gene clusters are shared between groups.

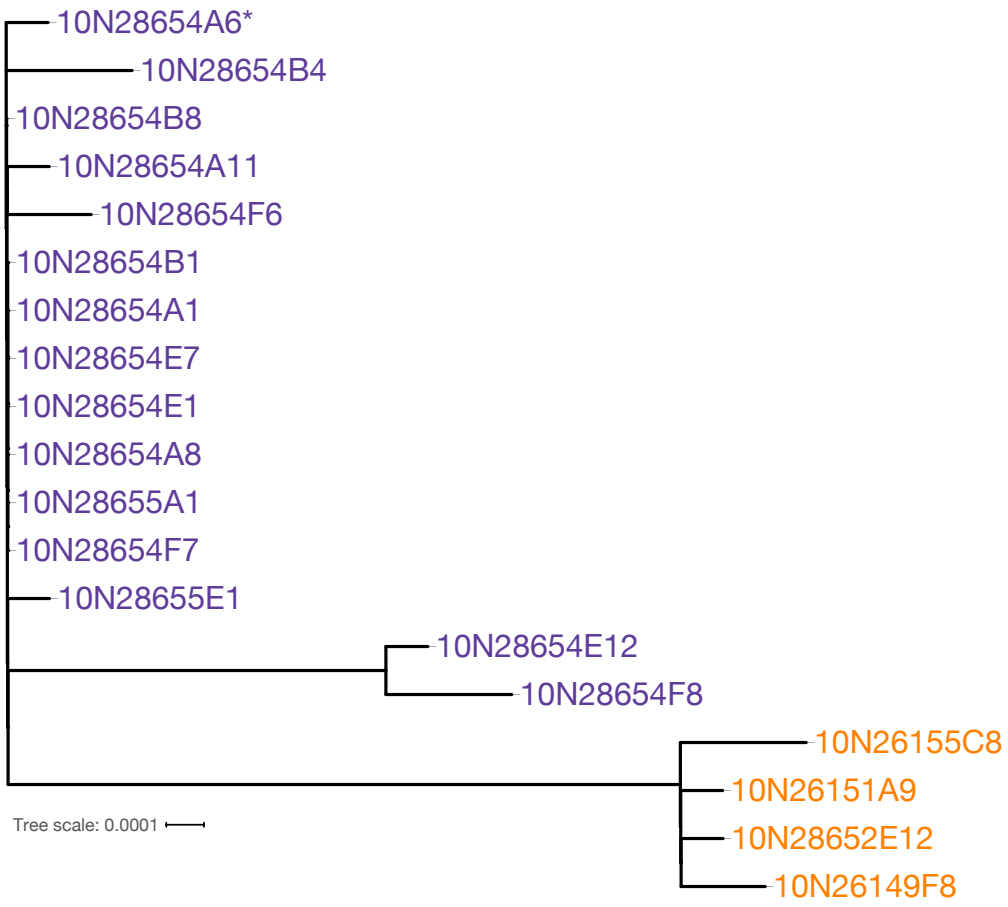


Figure 3-6: **Fig.S3:** Unrooted maximum likelihood tree for core genomes of all the nineteen "orange" and "purple" clonal hosts. The strain chosen by the Parsnp program as a reference is indicated by *. 44 SNPs were identified in the total alignment and 14 SNPs differentiate the "orange" and "purple" subsets (see Table S1 for a full list of SNP locations and descriptions).

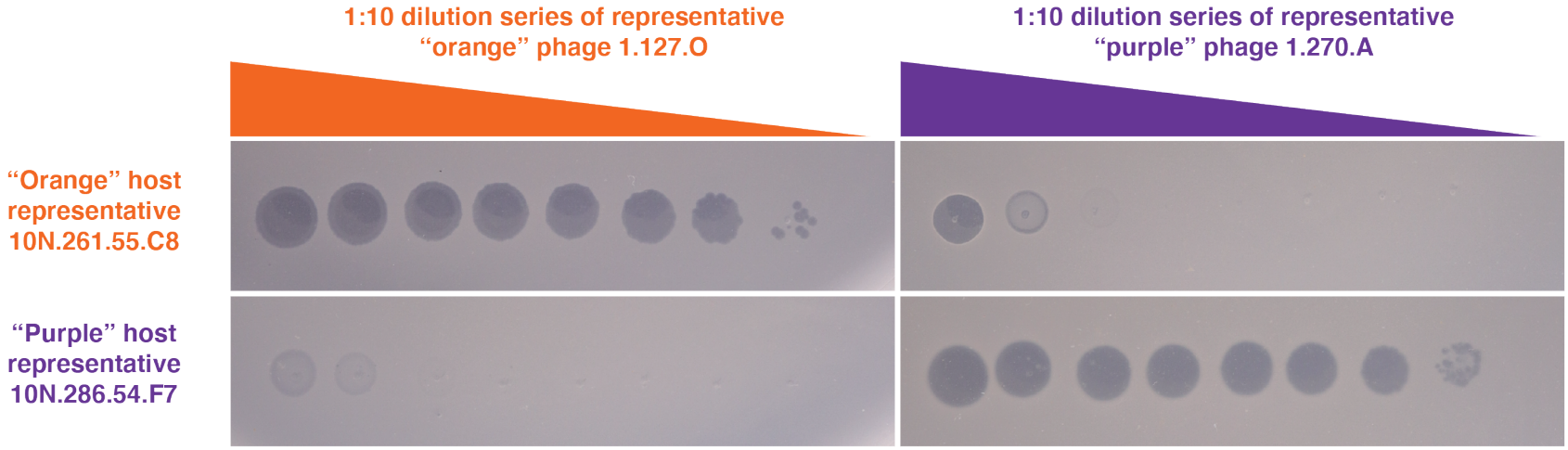


Figure 3-7: **Fig.S4:** Efficiency of plating assay demonstrating effect of differing phage concentrations on host killing. At high concentrations, phage can effect lysis even of non-hosts but without production of viable progeny (“lysis from without”) indicating that phage can attach and enter the cell, but that replication is prevented internally.

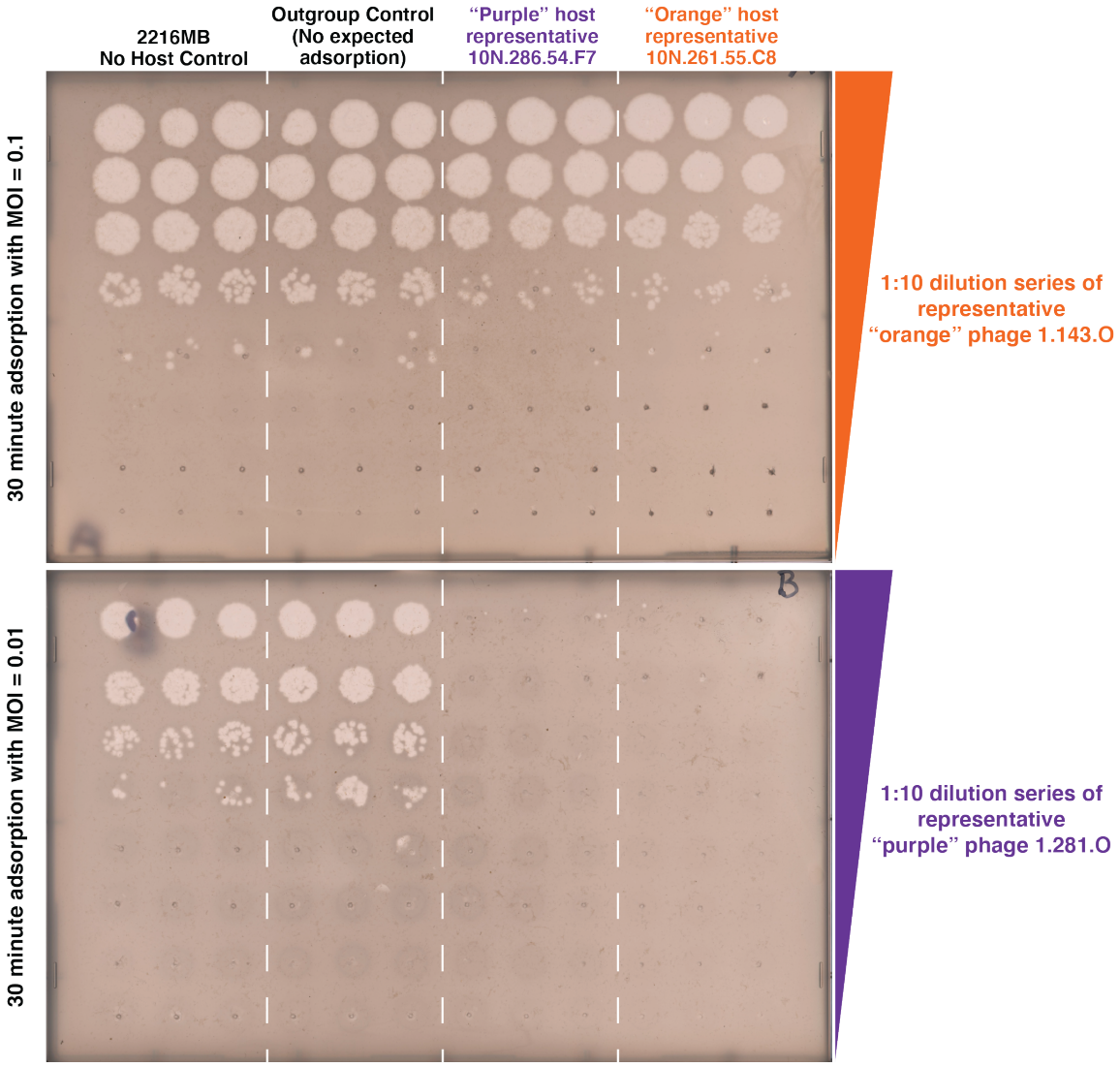


Figure 3-8: **Fig.S5:** Phage adsorption assay showing that phages can adsorb to both "orange" and "purple" strains irrespective of whether those bacterial strains can serve as hosts for viable phage production. After allowing a fixed concentration of phages to adsorb to different bacterial strains, free phages that remained unattached were plated with sensitive hosts to quantify adsorption as the difference to no-host controls (see methods). Both "orange" and "purple" phages were found to adsorb to "orange" and "purple" hosts, but not to an outgroup control. In the top panel, "orange" phage 1.143.O shows the same adsorption phenotype to both "orange" host 10N.261.55.C8 and "purple" host 10N.286.54.F7: the number of free phages decreased by ten-fold. In the bottom panel, "purple" phage 1.281.O shows the same adsorption phenotype to both "orange" host 10N.261.55.C8 and "purple" host 10N.286.54.F7, attaching with full efficiency. In both cases, no attachment is observed for a *Vibrio* outgroup host (10N.261.49.C11) as indicated by the same level of phages as in no host controls.

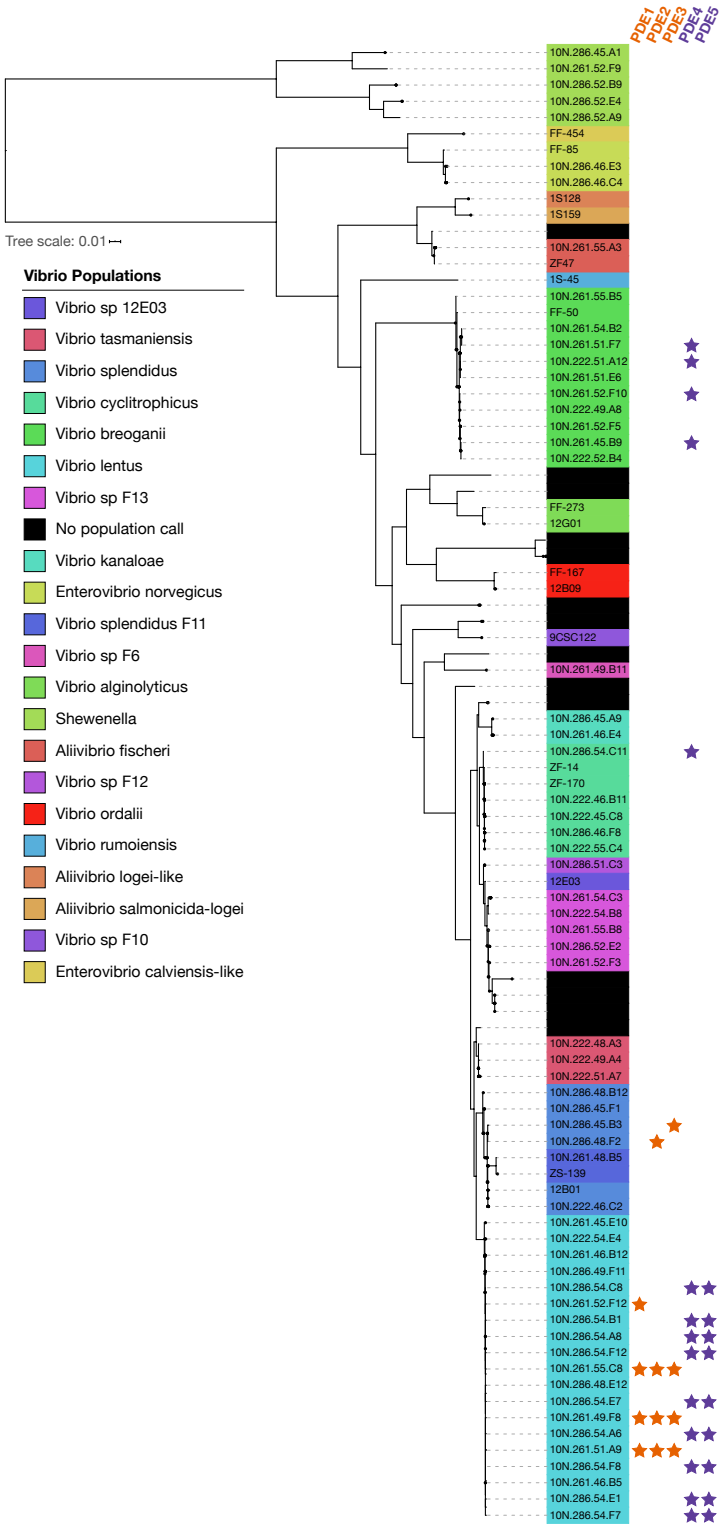


Figure 3-9: **Fig.S6:** Presence of the same phage defense elements (>95% nucleotide identity over >90% of the total element length) in divergent genomic backgrounds suggests their movement via horizontal gene transfer. Pruned tree from Figure S1 depicting the phylogeny of ribosomal protein and *hsp60* gene sequences (proxy for core genome) of each *Vibrio* host.

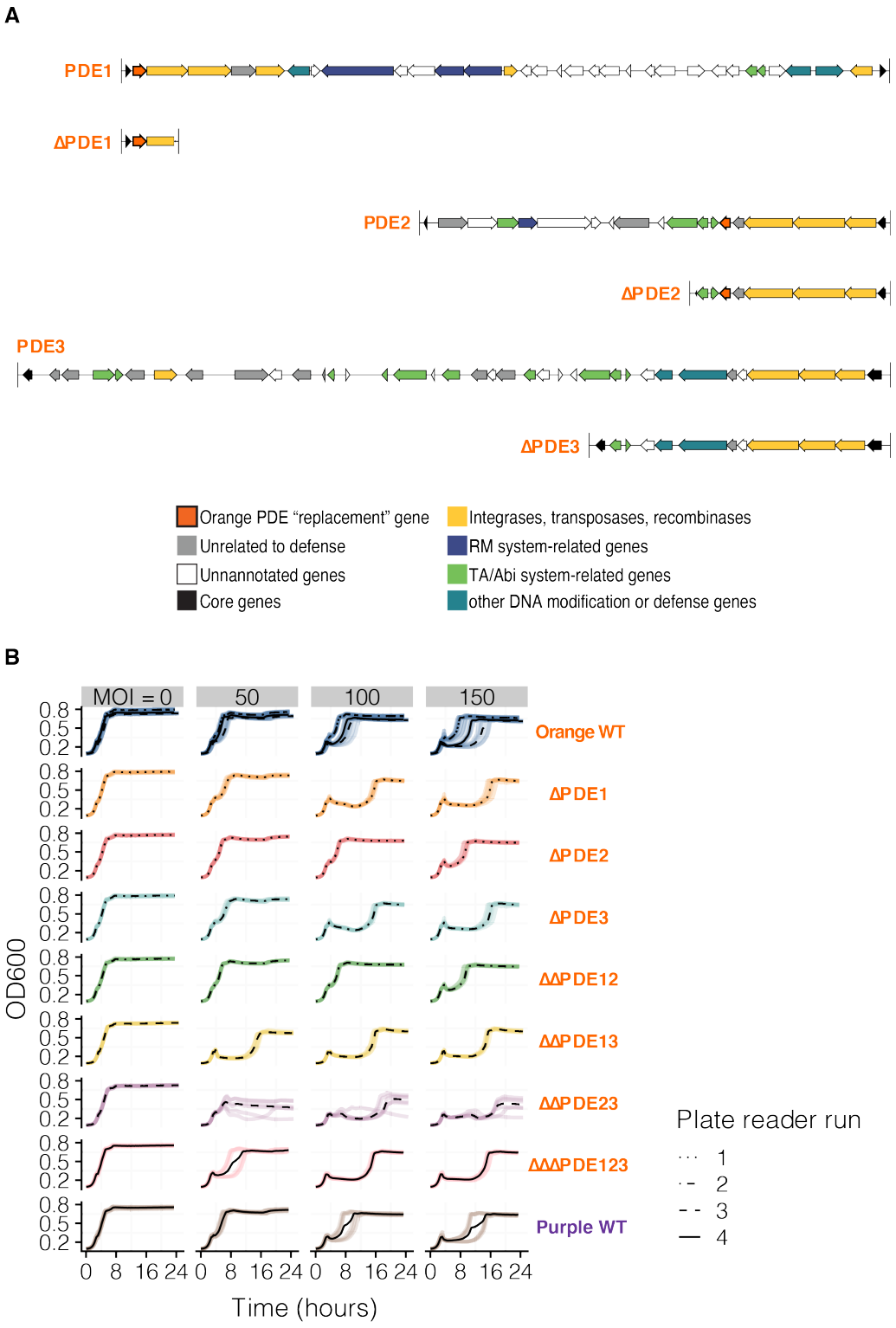


Figure 3-10: **Fig.S7:** PDE deletions and phage susceptibility testing. (A) Genetic knockout diagrams for each phage defense element in the "orange" strains, and (B) growth curves of each combination of knockouts grown to mid-exponential phase and then challenged with "purple" phage 1.281.O at varying concentrations.

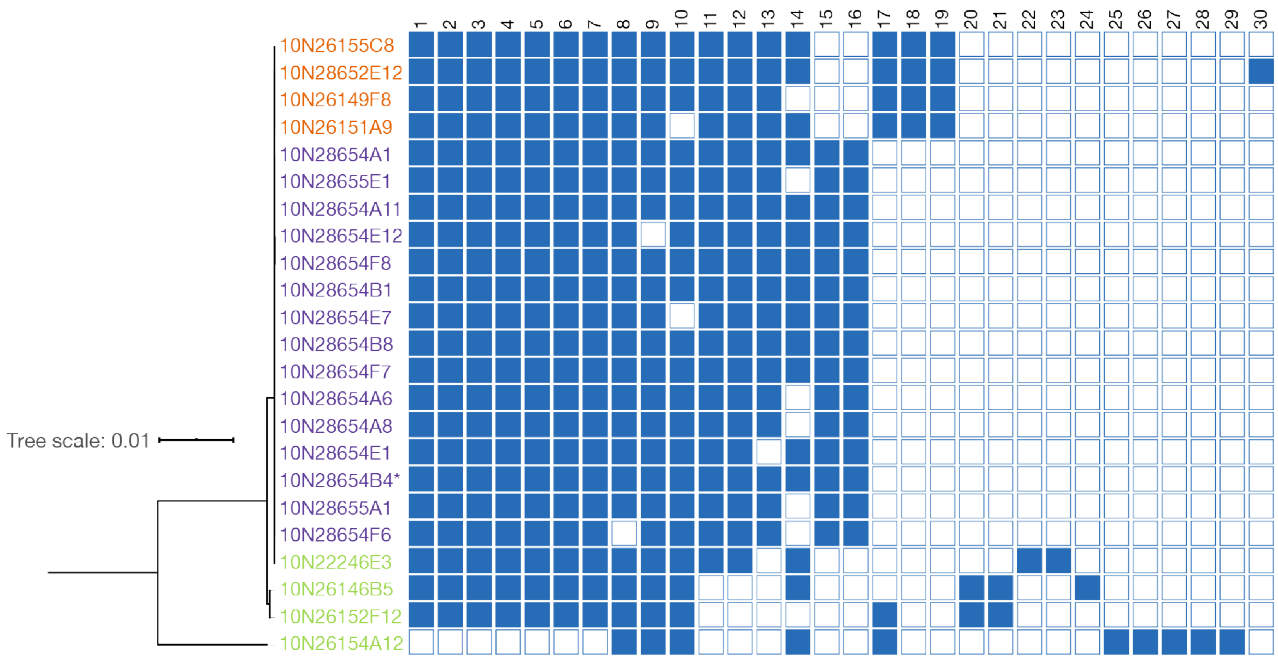


Figure 3-11: **Fig.S8:** Distribution of all putative PDEs in *Vibrio lentus* clones. Bacterial hosts are arranged by core genome tree. Accompanying gene diagrams, identified hits to known defense genes and full annotations are available on: https://github.mit.edu/fatimah

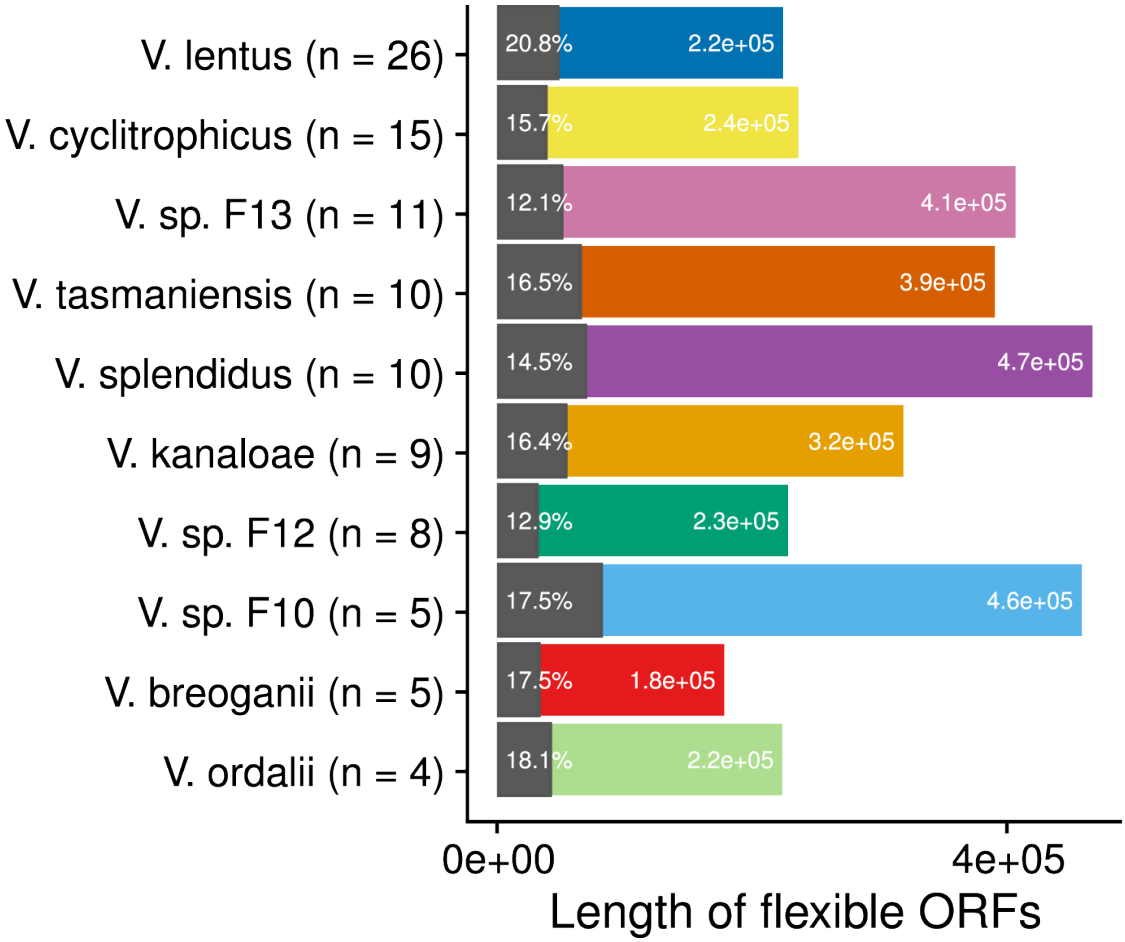


Figure 3-12: **Fig.S9:** Proportion of known phage defense genes by length in the flexible genomes of diverse *Vibrio* species. Between 12-21% of the flexible gene content of ten different species, represented as populations defined as gene flow clusters (Arevalo et al., 2019), can be attributed to known phage defense genes.

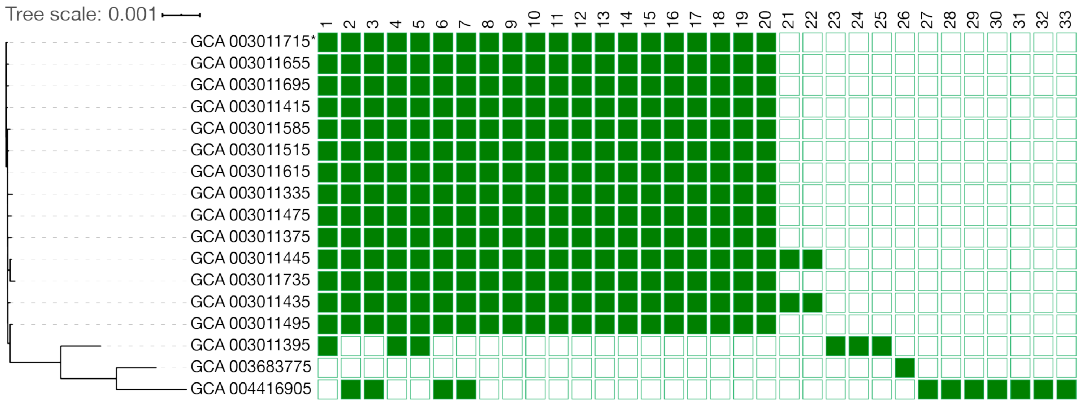


Figure 3-13: **Fig.S10:** Distribution of all putative PDEs in *Listeria* genomes. Bacterial hosts are arranged by core genome tree. Accompanying gene diagrams, identified hits to known defense genes and full annotations are available on: https://github.mit.edu/fatimah

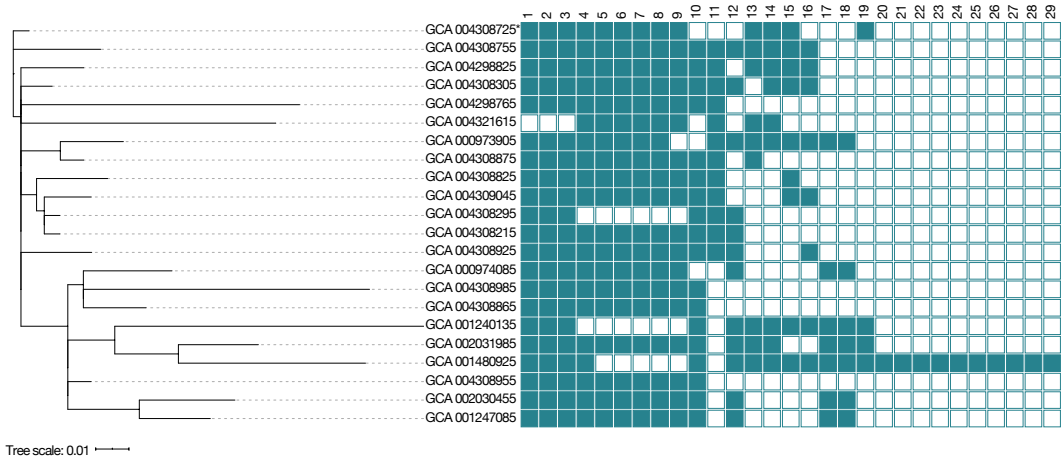


Figure 3-14: **Fig.S11:** Distribution of all putative PDEs in *Salmonella* strains with accompanying gene diagrams. Bacterial hosts are arranged by core genome tree. Accompanying gene diagrams, identified hits to known defense genes and full annotations are available on: https://github.mit.edu/fatimah

Table 3.1: **Table S1:** SNPs in the core genome of 19 "purple" and "orange" clones (matching Fig. S3).

SNP #	Nucleotide change P-->O	ORF in 10N28654F7	Annotation	AA change P-->O	Notes
1	G-->T	ORF_0_1115	Cytochrome O ubiquinol oxidase subunit III	Val-->Phe	
2	C-->T	ORF_0_1604	Unannotated	NA	Asp-->Asp; Near end of a PDE
3	C-->T	ORF_1_1456	Pesticidial crystal protein cry6Aa	Gln-->Stop	
4	G-->A	ORF_1_917	Aerobic cobaltochelatase CobS subunit	Ala-->Val	
5	G-->A	ORF_1_912	TRAP-type C4-dicarboxylate transport system, periplasmic component	NA	Gly-->Gly
6	A-->G	ORF_1_365-366	Unannotated	Stop-->Gln	
7	C-->T	ORF_0_3133	Phosphoenolpyruvate carboxylase	NA	Gly-->Gly
8	C-->T	ORF_0_1978	IcmF-related protein	Ala-->Val	
9	A-->G	ORF_1_87	Unannotated	Thr-->Ala	
10	A-->G	intergenic	NA	NA	upstream of ORF_1_86: putative orphan protein; putative membrane protein
11	C-->G	ORF_1_687	RND multidrug efflux transporter; Acriflavin resistance protein	Gln-->Glu	
12	G-->A	ORF_0_1847	Formate dehydrogenase -O, gamma subunit	Met-->Ile	
13	C-->A	intergenic	NA	NA	upstream of ORF_0_138:tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA
14	G-->A	ORF_1_1585	Probable MFS transporter	NA	Ala-->Ala

Table 3.2: **Table S2:** Receptor identification by transposon mutagenesis.

Host	Phage	Gene ID	Annotation	Tn Hits
10N26155C8	1.143.O	10N26155C8_2_159	General secretion pathway protein H	81
10N26155C8	1.143.O	10N26155C8_2_164	General secretion pathway protein C	9
10N26155C8	1.143.O	10N26155C8_2_162	General secretion pathway protein E	1
10N26155C8	1.143.O	10N26155C8_2_156	General secretion pathway protein K	1
10N26155C8	1.281.O	10N26155C8_2_94	dTDP-4-dehydrorhamnose reductase (EC 1.1.1.133)	58
10N26155C8	1.281.O	10N26155C8_2_93	dTDP-4-dehydrorhamnose 3,5-epimerase (EC 5.1.3.13)	8
10N26155C8	1.281.O	10N26155C8_2_96	dTDP-glucose 4,6-dehydratase (EC 4.2.1.46)	6
10N26155C8	1.281.O	10N26155C8_0_669	Phosphoglucomutase (EC 5.4.2.2)	3
10N26155C8	1.281.O	10N26155C8_0_485	ABC-type multidrug transport system, ATPase and permease component	1
10N26155C8	1.281.O	10N26155C8_0_2482	Na(+)-translocating NADH-quinone reductase subunit B (EC 1.6.5.-)	1
10N26155C8	1.281.O	10N26155C8_0_2480	Na(+)-translocating NADH-quinone reductase subunit D (EC 1.6.5.-)	1

Table 3.3: **Table S3:** Receptor identification by sequencing of spontaneous resistant mutants.

Host	Phage	Replicate	Contig_Position	Type	Length	Gene	AA change	Annotation
10N26155C8	1.119.O	1	2_171942	Deletion	1	ORF_2_160	ORF_2_160:p.Lys7fs	General secretion pathway protein G
10N26155C8	1.119.O	2	2_173977	SNV	1	ORF_2_162	ORF_2_162:p.Lys267*	General secretion pathway protein E
10N26155C8	1.119.O	3	2_1671-167305	Deletion via CA	109	ORF_2_154	NA	General secretion pathway protein M
10N26155C8	1.119.O	4	2_167254	SNV	1	ORF_2_154	ORF_2_154:p.Trp63*	General secretion pathway protein M
10N26155C8	1.119.O	5	2_173221	Deletion	1	ORF_2_161	ORF_2_161:p.Gly18fs	General secretion pathway protein F
10N26155C8	1.119.O	6	2_1719-172018	Deletion via CA	83	ORF_2_160	NA	General secretion pathway protein G
10N26155C8	1.119.O	6	2_166050-166174	Deletion via CA	125	ORF_2_153	NA	General secretion pathway protein N
10N26155C8	1.281.O	1	2_2804974	SNV	1	ORF_0_2483	ORF_0_2483:p.Ser302*	Na(+)-translocating NADH-quinone reductase subunit A (EC 1.6.5.-)
10N26155C8	1.281.O	2	2_68299	Insertion	1	ORF_2_61	ORF_2_61:p.Tyr187fs	Unannotated
10N26155C8	1.281.O	3	2_72111-72115	Deletion via CA	5	NA	NA	intergenic, upstream of 10N26155C8_2_65 (Unannotated)
10N26155C8	1.281.O	3	2_72128-72134	Deletion via CA	7	NA	NA	intergenic, upstream of 10N26155C8_2_65 (Unannotated)
10N26155C8	1.281.O	4	2_68299	Insertion	1	ORF_2_61	ORF_2_61:p.Tyr187fs	Unannotated
10N26155C8	1.281.O	5	2_108233	SNV	1	ORF_2_96	ORF_2_96:p.Gln332*	dTDP-glucose 4,6-dehydratase (EC 4.2.1.46)
10N26155C8	1.281.O	6	2_68845-68906	Deletion via CA	62	ORF_2_61&62	NA	end of 10N26155C8_2_61 and beginning of 10N26155C8_2_62
10N26155C8	1.281.O	7	2_107805	SNV	1	ORF_2_95	ORF_2_95:p.Asn109Tyr	Glucose-1-phosphate thymidylyltransferase (EC 2.7.7.24)

Table 3.4: **Table S4:** Predicted motifs of RM systems on PDEs and methylation fraction in genome.

RM PDE	PDE 1	
Predicted Motif	TCA*BN(4)RTRTC	
	Host/Phage	Fraction Methylated
	10N26155C8	599/600
	1.119.O	1/1
	1.127.O	1/1
	1.143.O	1/1
	1.231.O	1/1
	10N28654F7	0/599
	1.283.A	0/7
	1.281.O	0/8
	1.196.O	0/8
RM PDE	PDE 4	
Predicted Motif	CCA*GN(6)TAA	
	Host/Phage	Fraction Methylated
	10N26155C8	0/646
	1.119.O	0/3
	1.127.O	0/4
	1.143.O	0/3
	1.231.O	0/3
	10N28654F7	635/638
	1.283.A	5/5
	1.281.O	4/4
	1.196.O	4/4
RM PDE	PDE 5	
Predicted Motif	GA*GN(6)GGC	
	Host/Phage	Fraction Methylated
	10N26155C8	0/1795
	1.119.O	0/17
	1.127.O	0/15
	1.143.O	0/17
	1.231.O	0/17
	10N28654F7	1770/1779
	1.283.A	8/8
	1.281.O	10/10
	1.196.O	10/10

Table 3.5: **Table S5:** Strains and plasmids used in transposon mutagenesis and gene deletions.

Strain or plasmid	Description	Reference
Strains		
E. coli		
β3914	(F-) RP4-2-Tc::Mu Δ*dapA*::(erm-pir) *gyrA462 zei-298*::Tn10 (Km_R Erm_R Tc_R)	Le Roux et al., 2007
Π3813	*lacI*$_q$ *thi-1 supE44 endA1 recA1 hsdR17 gyrA462 zei298::Tn10* Δ*thyA::(erm-pir-116)* (Tc_R Erm_R)	Le Roux et al., 2007
MFD*pir*	*E. coli* MG1655 RP4-2-Tc::[ΔMu1::*aac(3)IV*-Δ*aphA*-Δ*nic35*-ΔMu2::*zeo*] Δ*dapA*::(erm-pir) Δ*recA* ($Apra_R$ Zeo_R Erm_R)	Ferrières, et al., 2010
V. lentus		
10N.261.55.C8	Representative "orange" strain (C8-WT)	This study
DPDE1	C8-WT with ΔPDE1: in frame partial deletion of PDE1 (31909/34140 bp)	This study
DPDE2	C8-WT with ΔPDE2: in frame partial deletion of PDE2 (12186/20725 bp)	This study
DPDE3	C8-WT with ΔPDE3: in frame partial deletion of PDE3 (25697/37369 bp)	This study
DDPDE12	C8-WT with ΔPDE1 and ΔPDE2	This study
DDPDE13	C8-WT with ΔPDE1 and ΔPDE3	This study
DDPDE23	C8-WT with ΔPDE2 and ΔPDE3	This study
DDDPDE123	C8-WT with ΔPDE1, ΔPDE2 and ΔPDE3	This study
Plasmids		
pSC189-Cm	*oriT RP4* Π-dependent *oriV* R6K *mariner*-based transposon TnSC189 Δ*kan::cat* (Cm_R Ap_R)	Ferrières, et al., 2010
pSW7848T	*oriV R6K ; oriT RP4; araC-*P_{BAD}*ccdB* Cm_R	Val, et al., 2012
pSWδR-1	pSW7848T::ΔPDE1	This study
pSWδR-2	pSW7848T::ΔPDE2	This study
pSWδR-2	pSW7848T::ΔPDE3	This study

Table 3.6: **Table S6:** Primers used in transposon mutagenesis and gene deletions.

Primer	Sequence 5´-3´	Reference
SS9arb2	GACCACGAGACGCCACACTNNNNNNNNNNACTAG	Lauro et al., 2008
Mar4	TAGGGTTGAGTGTTGTTCCAGTT	Jiao et al., 2005
Mar4_int2	GTCATCGTCATCCTTGTAATCG	This study
Arb3	GACCACGAGACGCCACACT	Lauro et al., 2008
ΔPDE1/F1	GTCGACGGTATCGATAAGCTTGATATCGAATTCCTGCATCATGGCTTGGGTCACTCG	This study
ΔPDE1/R1	GAAACTGGGTGCAAATGTCGTACAGTCTGGTGGGCCTGAG	This study
ΔPDE1/F2	CTCAGGCCCACCAGACTGTACGACATTTGCACCCAGTTTC	This study
ΔPDE1/R2	CCGTCAAGTTGTCATAATTGGTAACGAATCAGACAATTTTGTACCCTAGCGAACATTCTG	This study
ΔPDE1/F	GCCTACAGGTTGCTTTCGTC	This study
ΔPDE1/R	CAGCGCGTATTCTCTCGTTG	This study
ΔPDE2/F1	TAAGCTTGATATCGAATTCCTGCAGGTTGCCATCATTCTATTCGG	This study
ΔPDE2/R1	TGTTAAGGAAGTGGCAAAGTGAATGCACCAAGACTCACCACGAAG	This study
ΔPDE2/F2	AAACCACTTCGTGGTGAGTCTTGGTGCATTCACTTTGCCACTTCC	This study
ΔPDE2/R2	AATTGGTAACGAATCAGACAATTTTGTGAGAAGTACGGTGTTTGG	This study
ΔPDE2/F	TCGCTGAGGTTTGCTCTAC	This study
ΔPDE2/R	ATTACGATGAAGCTCAAAGCC	This study
ΔPDE3/F1	GCTTGATATCGAATTCCTGCAATTGCTAACCTACTGCCTTAC	This study
ΔPDE3/R1	GGAAGTGGCAAAGTGAATGCTGGAAACTCACTCACTCACTC	This study
ΔPDE3/F2	GAGTGAGTGAGTGAGTTTCCAGCATTCACTTTGCCACTTCC	This study
ΔPDE3/R2	CATAATTGGTAACGAATCAGACAATTGATGCTTATCGTGCGGTAAATG	This study
ΔPDE3/F	GCGTAATGTCAGTTTGATTTCGATG	This study
ΔPDE3/R	CAAGATCACTATGCAGGAACAGG	This study
pSW_ F	AATTGTCTGATTCGTTACCAATTATG	This study
pSW_ R	TGCAGGAATTCGATATCAAGC	This study

Chapter 4

Conclusions and Outlook

Ecological interactions are thought to drive the fine-scale diversity observed in wild microbial populations. For example, heterotrophic bacteria of the genus *Vibrio*, the model system used in this work, have been shown to differentiate into genetically and ecologically cohesive populations that are distinguished by the unique niche space and microenvironments they occupy (Hunt et al., 2008; Preheim et al., 2011; Shapiro et al., 2012; Szabo et al., 2013; Wildschutte et al., 2010).

Studying the genomes of microbes such as these has enabled us to begin to learn how bacteria evolve in nature. High rates of recombination are hallmarks of these bacterial populations, resulting in low linkage of genes across the genomes and leading to the observation of gene-specific sweeps that drive adaptation (Shapiro et al., 2012; Shapiro and Polz, 2014). Using this framework, we can begin to ask targeted questions about how ecological and social interactions drive frequency-dependent selective pressures on specific genes of interest in bacterial populations (Cordero et al., 2012; Cordero and Polz, 2014) .

In 2014, at the beginning of my PhD, Otto Cordero and Martin Polz wrote a review paper elegantly describing how to think about, "microbial genomic diversity in light of evolutionary ecology." In that paper, Cordero and Polz hypothesized that, through negative-frequency dependent selection, phage predation should drive high turnover of genes that encode receptors mediating viral attachment. In broad genomic and metagenomic surveys, receptor genes are commonly found in genomic islands (Avrani

et al., 2011; Rodriguez-Valera et al., 2009), and so, they argued, the horizontal transfer of these genes between bacteria may dynamically protect individual genotypes from specific viral predation. This shuffling of genes would be selected for and ultimately establish microdiversity at the population level. This argument established the framework for interrogating the relationships between phages and their microbial hosts and heavily influenced the approaches and hypotheses that developed into my thesis.

Here, we discovered that genes occurring at low frequencies in *Vibrio* populations are indeed phage related. The genes exhibiting rapid evolutionary turnover are prophages (Chapter 2) and, instead of genes encoding phage receptors, phage-defense elements (PDEs) that drive host resistance to viruses (Chapter 3).

4.1 Overview of thesis chapters and next steps

In Chapter 2, I described a survey we conducted to uncover the abundance and diversity of prophages in an ecologically cohesive population of marine microbes. By inducing prophages from putative lysogens and then sequencing the prophage genomes, we created a database of diverse, excisable prophages and other mobile genetic elements. We found that many elements go undetected by current prophage search algorithms, and the same elements are typically only shared among closely related strains within the same population. Our results advocate the use of an experimental approach for novel prophage discovery and shed light on how the distribution of prophages in microbial genomes may reflect ongoing gene transfer networks in wild microbial populations.

Similar prophages were most often among bacteria within the same population, implying that prophages act in a population-specific manner, killing or lysogenizing co-occurring, closely related strains. That is, prophages residing in bacteria may be used for competitive interference against microbes looking to occupy the same niche. Or, the opposite may be true: prophages could exhibit lysogenic behavior within certain populations while acting as lytic predators in other populations. One way

to test these hypotheses would be to induce out and isolate prophages from a set of ecologically cohesive bacteria, like the ones used here, and test the host ranges of those viruses. If prophages can be used as mechanisms of competitive exclusion of sister strains, then we would expect prophage infections to have a phylogenetic bias towards killing within, rather than across, populations. Entry of such population-specific prophages could be regulated by the presence of conserved receptors across the population, as we observed for the lytic phage receptors identified in Chapter 3. Some evidence for this hypothesis exists in *Lactobacillus* strains that live in the vagina (Kilic et al., 2001) and in environmental *Vibrio* strains (Wendling Carolin C. et al., 2018), but such a study using sympatric microbes ranging in relatedness has not been conducted. In contrast, if phages are lysogenic within some populations, and lytic in others, they may drive antagonistic interactions between populations, much like antibiotics (Cordero et al., 2012).

In addition to inducing and identifying prophages, the methods used in Chapter 2 led us to find other excisable mobile genetic elements. These elements, which are secreted by induced cultures and resilient to nuclease digestion, may play an important part in driving the evolution of the flexible genome. Future work should investigate the physical properties and genetic content of these uncharacterized elements. For example, it is possible these elements are carried in vesicles. While preliminary studies of vesicles have shown that they carry DNA fragments from across the host genome (Biller et al., 2014), this phenomenon has not been systematically investigated in marine heterotrophs. It is also possible that the conditions under which our samples were collected — stress due to DNA damage and lytic phage infection — results in differential expression of genes in the genome, and thus, bias in the genes and transcripts packaged upon excision.

Not all putative prophages in the *Vibrio* genomes were induced using our methods. This lack of induction may be because some prophages are no longer active. However, an alternative explanation is that not all potentially active prophages are induced under the tested conditions. For example, while Mitomycin C has been commonly used as an inducing agent (Otsuji et al., 1959), it is not commonly present

in marine environments, and furthermore, may not work on all prophages. In addition, there are many other known chemical and physical inducers, including UV (Castellazzi et al., 1972), quorum sensing molecules (Ghosh et al., 2009; Silpe and Bassler, 2018), cigarette smoke (Pavlova and Tao, 2000), and environmental pollutants (Cochran and Paul, 1998). Future work testing a panel of known inducers will be valuable in determining the extent of the active inducible prophage pool in these strains. Finally, the inductions conducted in Chapter 2 were done using monocultures of isolates. Any prophage inductions determined by microbe-(different) microbe or microbe-environment interactions were missed. Testing for prophage inductions in mixed communities with increasing levels of complexity (pairs, triples, etc.) would help to elucidate how microbial interactions drive prophage dynamics in complex microbial communities.

We are just beginning to understand the diversity, abundance, and dynamics of prophage-host interactions. The more microbial genomes we sequence, the more prophages we will find. This diversity extends to many environments, including the human microbiome and its associated phageome (Manrique et al., 2016; Modi et al., 2013). Prophages are thought to play an important role in the gut and vaginal microbial communities, and probably most other microbial communities in the body. As we move to design probiotics and prebiotics as therapeutics, we will want to know how existing prophages, both in the microbial therapeutics and in the communities they target, will respond to a changing environment. The development of microbial therapeutics without regard for the latent phages in their genomes and the stimuli they respond to could lead to unintended consequences. For instance, induced prophages may kill or lysogenize beneficial conspecifics, promoting the growth of unwanted taxa. Prophages are known to carry toxin and antibiotic resistance genes. The inadvertent spread of such prophages or the recombination of such phages with other existing prophages in the genomes of microbial therapeutics could cause wide spread of such undesirable phenotypes. Indeed, with proper consideration, the inclusion of prophages that respond to clinically relevant stimuli could become a valuable component of prebiotic design and success. Studying prophage-host interactions and

co-evolution in diverse host systems will be key moving forward. Moreover, culturing bacterial strains in the lab will be valuable to identify those prophages missed by sequence searches alone.

In Chapter 3, we used population genomics and molecular genetics to study the rapid evolutionary changes driven by lytic phages in natural populations of bacteria. We found nearly clonal strains differ in their carriage of phage-defense elements (PDEs), and as a result, exhibit distinct phage predation profiles when challenged with co-occurring lytic viruses. We found there is a cumulative defense structure, with multiple different PDEs acting together to yield protection. These elements are incredibly diverse, with the majority of their genes unannotated, and abundant, accounting for a substantial proportion of the flexible genome.

One reason PDEs have been difficult to identify computationally is because they often break in genome assemblies. High quality genomes, sequenced using long reads from PacBio and Oxford Nanopore, surmount this difficulty and will soon be commonplace. Sequencing the genomes of multiple bacterial strains from the same environment using these technologies in the future will likely lead to the discovery of more PDEs using computational methods alone.

Further genetic characterization of the different genes in the PDEs discovered in this work will help us understand more about their unique biology. Studying phage-defense systems has proven to be valuable in the past. For example, CRISPR and restriction enzymes function natively to protect bacteria from phage infection, but are now indispensable components of modern molecular biology (Salmond and Fineran, 2015).

Although there is a strong track record for repurposing the unique enzymatic activities present on PDEs, we have a limited understanding of how they structure ecological interactions in microbial populations. In one strain, we found that a minimum of three PDEs were necessary to yield resistance to the tested phage family. From our bioinformatic search, we found that a single strain can harbor as many as 18 unique PDEs. Key questions that remains are: How do these different elements exclude one another, and how do they work in synergy? For example, do some PDEs

offer broad protection against viruses, while others have narrow specificity? Is there a tradeoff between these strategies; that is, are PDEs offering broad protection that is less effective than more specific PDEs for defense against individual viruses? From our work, we see that phage defense and phage susceptibility are not binary, but rather a spectrum. To answer these questions, we need lab-based experiments in genetically tractable model organisms to enable us to see how carriage of various combinations of PDEs alters host susceptibility to a panel of viruses.

Additionally, we find PDEs are fully responsible for phage defense, while receptors remain invariant. This discovery emphasizes that evolution in the wild differs significantly from evolution observed in the lab, particularly in experimental systems where growth conditions do not mimic those in nature, and those not open to gene flow. Bacteria living and evolving in complex, dynamic environments with other microbes and viral predators behave differently than they do growing in isolation in nutrient rich constant conditions in the lab. Diverse populations of environmental bacteria rely on receptors for nutrient uptake, attachment, or other essential uses. Growth in nutrient rich media likely removes the purifying selective pressure on receptor genes to maintain optimized function as phage predation becomes the only immediate threat when nutrients are plentiful. When we evolve strains in the lab in the presence of viral predators, they are quick to evolve resistance through allelic changes in receptor genes. When we observe how bacteria have evolved resistance to phages in the wild, we see conservation in the receptor genes and resistance evolution being driven via rapid evolutionary turnover of PDEs.

However, this dichotomy may be specific to evolutionary modality. Under different relative strengths of selection and frequency of horizontal gene transfer, these observations might not hold. For example, populations of pathogens are often clonal, exhibiting lower rates of recombination. Receptor variation may be a main driver of resistance to phages for such microbes. Furthermore, it is possible that in a different set of organisms, living in a different ecosystem, purifying selection on the receptors might not be as strong. If bacteria were able to modify their receptors with a less severe tradeoff, or if viruses targeted receptors that were interchangeable with others,

then bacteria might be able to combine receptor modification with the gain and loss of PDEs to defend against their viral predators. Recent work has shown that broad host-range viruses may drive the development of receptor-based defenses while narrow host-range viruses may drive the evolution of internal defenses like PDEs (Zborowsky and Lindell, 2019). Investigating the genomes of bacteria targeted by both specialist and generalist viruses in nature with a similar approach as done in Chapter 3 would complement this work well. Furthermore, in the wild, multiple viruses may use multiple different receptors to infect a given host. Co-infections may result in an increased or altered need for phage defense, which also begs further investigation.

Due to their killing potential and specificity, phages are rapidly being developed as therapeutics to treat infections of antibiotic-resistant pathogens. When used to treat clonal pathogenic strains, phage therapy can serve as a life-saving marvel. However, applying phages in wider applications – for example pre-treating leafy greens with phages to prevent foodborne pathogens, or using phages to control pathogenic outbreaks in aquaculture systems open to the environment – may lead to unintended consequences. In complex environmental systems where diverse microbes readily exchange genes, we would expect to see the horizontal gene transfer, and subsequent selection, of PDEs lead to widespread phage resistance, much as antibiotic resistance, primarily associated with plasmids, has become more prevalent through the increased use of antibiotics. As phage therapy efforts become more widespread, studying the evolution of resistance to phages in natural systems will help us anticipate and proactively implement countermeasures to design effective phage therapeutics.

The rapid transfer of PDEs found in this work is critical because it implies that negative frequency-dependent selection by viruses yields high turnover of defense genes, unlinking the defense phenotype from the rest of the host's core phenotypes. This combination of selection and gene flow ultimately decouples viral specificity from the bacterial strain and links it to the presence or absence of particular mobile genetic elements within a given strain. This means that phage-host specificity is incredibly dynamic, and that (1) hosts can alter resistance to phages and (2) phages can shift host specificity at remarkable rates. Further work in modeling eco-evolutionary dy-

namics of phage-host interactions given our findings may help us understand different evolutionary trajectories in different ecologies.

The work presented here begins to tackle the question of how bacterial viruses shape the diversity and evolution of microbial populations in the wild. Not only are bacteria and viruses highly diverse, but their interactions and, consequently, their genomes are also highly dynamic. In the wild, bacteria and viruses are continually evolving and co-evolving. Rampant recombination and heavy selective pressure from both lytic and lysogenic phages causes the remodeling of bacterial genomes on faster timescales than we have previously appreciated.

Bibliography

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., and Lindell, D. (2011). Genomic island variability facilitates Prochlorococcus-virus coexistence. *Nature*, 474(7353):604–608.

Biller, S. J., Schubotz, F., Roggensack, S. E., Thompson, A. W., Summons, R. E., and Chisholm, S. W. (2014). Bacterial Vesicles in Marine Ecosystems. *Science*, 343(6167):183–186.

Castellazzi, M., George, J., and Buttin, G. (1972). Prophage induction and cell division in E. coli. *Molecular and General Genetics MGG*, 119(2):153–174.

Cochran, P. K. and Paul, J. H. (1998). Seasonal Abundance of Lysogenic Bacteria in a Subtropical Estuary. *Applied and Environmental Microbiology*, 64(6):2308–2312.

Cordero, O. X. and Polz, M. F. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Reviews Microbiology*, 12(4):263–273.

Cordero, O. X., Wildschutte, H., Kirkup, B., Proehl, S., Ngo, L., Hussain, F., Roux, F. L., Mincer, T., and Polz, M. F. (2012). Ecological Populations of Bacteria Act as Socially Cohesive Units of Antibiotic Production and Resistance. *Science*, 337(6099):1228–1231.

Ghosh, D., Roy, K., Williamson, K. E., Srinivasiah, S., Wommack, K. E., and Radosevich, M. (2009). Acyl-Homoserine Lactones Can Induce Virus Production in Lysogenic Bacteria: an Alternative Paradigm for Prophage Induction. *Applied and Environmental Microbiology*, 75(22):7142–7152.

Hunt, D. E., David, L. A., Gevers, D., Preheim, S. P., Alm, E. J., and Polz, M. F. (2008). Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton. *Science*, 320(5879):1081–1085.

Kilic, A. O., Pavlova, S. I., Alpay, S., Kilic, S. S., and Tao, L. (2001). Comparative Study of Vaginal Lactobacillus Phages Isolated from Women in the United States and Turkey: Prevalence, Morphology, Host Range, and DNA Homology. *Clinical and Diagnostic Laboratory Immunology*, 8(1):31–39.

Manrique, P., Bolduc, B., Walk, S. T., Oost, J. v. d., Vos, W. M. d., and Young, M. J. (2016). Healthy human gut phageome. *Proceedings of the National Academy of Sciences*, 113(37):10400–10405. Publisher: National Academy of Sciences Section: Biological Sciences.

Modi, S. R., Lee, H. H., Spina, C. S., and Collins, J. J. (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature*, 499(7457):219–222.

Otsuji, N., Sekiguchi, M., Iijima, T., and Takagi, Y. (1959). Induction of Phage Formation in the Lysogenic Escherichia coli K-12 by Mitomycin C. *Nature*, 184(4692):1079–1080.

Pavlova, S. I. and Tao, L. (2000). Induction of vaginal Lactobacillus phages by the cigarette smoke chemical benzo[a]pyrene diol epoxide. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 466(1):57–62.

Preheim, S. P., Boucher, Y., Wildschutte, H., David, L. A., Veneziano, D., Alm, E. J., and Polz, M. F. (2011). Metapopulation structure of Vibrionaceae among coastal marine invertebrates. *Environmental Microbiology*, 13(1):265–275.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., PaÅąiÄĞ, L., Thingstad, T. F., Rohwer, F., and Mira, A. (2009). Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology*, 7(11):828–836.

Salmond, G. P. C. and Fineran, P. C. (2015). A century of the phage: past, present and future. *Nature Reviews Microbiology*, 13(12):777–786.

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., SzabÃş, G., Polz, M. F., and Alm, E. J. (2012). Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science*, 336(6077):48–51.

Shapiro, B. J. and Polz, M. F. (2014). Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology*, 22(5):235–247.

Silpe, J. E. and Bassler, B. L. (2018). A Host-Produced Quorum-Sensing Autoinducer Controls a Phage Lysis-Lysogeny Decision. *Cell*.

Szabo, G., Preheim, S. P., Kauffman, K. M., David, L. A., Shapiro, J., Alm, E. J., and Polz, M. F. (2013). Reproducibility of Vibrionaceae population structure in coastal bacterioplankton. *The ISME Journal*, 7(3):509–519.

Wendling Carolin C., Goehlich Henry, and Roth Olivia (2018). The structure of temperate phageâĂŞbacteria infection networks changes with the phylogenetic distance of the host bacteria. *Biology Letters*, 14(11):20180320.

Wildschutte, H., Preheim, S. P., Hernandez, Y., and Polz, M. F. (2010). O-antigen diversity and lateral transfer of the wbe region among Vibrio splendidus isolates. *Environmental Microbiology*, 12(11):2977–2987.

Zborowsky, S. and Lindell, D. (2019). Resistance in marine cyanobacteria differs against specialist and generalist cyanophages. *Proceedings of the National Academy of Sciences*, page 201906897.

Appendix A

Autolykiviridae-like prophages are widespread in marine *Vibrio* and contribute to the nontailed viral majority

A.1 Overview

The surface ocean is dominated by nontailed viral morphotypes (Brum et al., 2013), yet tailed viruses are most abundant in viral culture collections. Previous bioinformatic analyses identified putative Corticovirus-like prophages in Proteobacteria (Krupovic and Bamford, 2007). Here, we identify related prophages in diverse *Vibrio* genomes and demonstrate that they readily excise from their hosts as nontailed viruses. The widespread distribution and continuous release of prophages related to lytic nontailed viruses suggests that these prophages play a significant role in the ecology of marine *Vibrio* and contribute to the nontailed viral majority.

A.2 Results Summary

A.2.1 Lysogenic nontailed viruses are prevalent and widely distributed in diverse *Vibrio* genomes.

Nontailed prophages show homology to PM2 in their major capsid protein and packaging ATPase. When we used these genes to search 758 fully sequenced *Vibrio* genomes, we found 248 (33%) harbor a Corticovirus-like prophage, and of these, 28% contain more than one prophage (Figure 1).

Genetically similar elements are not necessarily found in closely related hosts, suggesting that the elements have a broad host range. Major capsid protein distribution suggests prophage mobility and association with both plasmids and the host chromosomes (Figure 2). Additionally, prophages with identical MCP sequences are found in distantly related hosts (Figure 3). Sequencing data suggest the prophages use a site-specific recombinase to integrate into the host genome directly upstream to a tRNA dihydrouridine synthase, and the surrounding genes, upstream and downstream of the integration site, may determine host specificity. Additionally, for several genomes with the element, we identify a partner strain that has identical upstream and downstream regions, but is missing the prophage, leading us to believe these may be potential hosts for the corresponding prophage.

A.2.2 Prophages excise from genomes naturally as nontailed viruses.

By concentrating, nuclease treating, and sequencing the supernatants of putative lysogenic cultures, we found that active prophages excise during both mid-exponential and late-stationary phases, indicating continuous release during host growth. By running viral concentrates along a density gradient, we find excised viruses have a lower bouyant density than tailed viruses, similarly to the Autolykiviridae (Figure 4, Appendix B). The lower buoyancy may be attributed to an internal lipid membrane. Thin-section transmission electron microscopy provides evidence that the morphol-

ogy of the excised elements resembles nontailed viruses with an internal lipid membrane, morphologically analogous to their lytic counterparts, the Corticovirus PM2 and the recently discovered Autolykiviridae. Also similarly to the Autolykiviridae, nontailed prophages are difficult to image with negative staining transmission electron microscopy alone and require thin-section preparation to visualize (Figure 5).

A.3 Materials and Methods

A.3.1 Abundance, diversity, and transfer of prophages

In order to identify nontailed prophages in *Vibrio* genomes, we used MAFFT (Katoh and Standley, 2013) to align two conserved genes – the major capsid protein (MCP) and the packaging ATPase – from previously identified putative prophages and used the alignments to search for prophages in our in-house genome database of over 750 *Vibrio* genomes using HMMER (Eddy, 2011). The insertion site of a subset of elements was determined by manually searching for conserved regions upstream and downstream of inserted prophages and then searching for said regions in other genomes using BLAST (Altschul et al., 1990). To determine the diversity of the prophages, we used the MCP protein sequence as a proxy.

A.3.2 Sequence-based approach to identify naturally induced prophages

In order to verify prophages were in fact actively excising, we sequenced the supernatant of putative lysogens. Putative lysogens were grown up for 1 week in 1L batch cultures. Then, cultures were filtered to remove cells and remaining supernatant was concentrated using PEG. Samples were purified using iodixanol-based density ultracentrifugation, and fractions containing prophages were determined using PCR primers specific to the major capsid protein. PCR positive fractions were DNase-treated, extracted, and sequenced using Illumina HiSeq Rapid. Finally, reads were mapped back to the putative lysogen genome to determine the full prophage sequence

and insertion cite using CLC Genomics Workbench. Details of this method are described in Appendix B.

A.3.3 Imaging excised prophages

A representative lysogen (5S149) was grown up for 1 week in a 1L batch culture and then pelleted via centrifugation at 1,000 xg for 1 hour. The supernatant was filtered using a 0.22 μm vacuum filtration system and then concentrated using 100 kDa Ultrasette tangential flow filter and re-filtered through a 0.2 μm syringe filter as previously described (Biller et al., 2014). The concentrate was pelleted by ultracentrifugation at 32,000 rpm for 2 hours at 4C using a Beckman Coulter centrifuge with an SW32Ti rotor. The sample was chloroform treated to remove vesicles and then imaged as is and as thin-sections at the Whitehead Microscopy facility at MIT by Nicki Watson.

Bibliography

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Biller, S. J., Schubotz, F., Roggensack, S. E., Thompson, A. W., Summons, R. E., and Chisholm, S. W. (2014). Bacterial Vesicles in Marine Ecosystems. *Science*, 343(6167):183–186.

Brum, J. R., Schenck, R. O., and Sullivan, M. B. (2013). Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *The ISME Journal*, 7(9):1738–1751.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, 7(10).

Katoh, K. and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780.

Krupovic, M. and Bamford, D. H. (2007). Putative prophages related to lytic tailless marine dsDNA phage PM2 are widespread in the genomes of aquatic bacteria. *BMC Genomics*, 8(1):236.

A.4 Figures

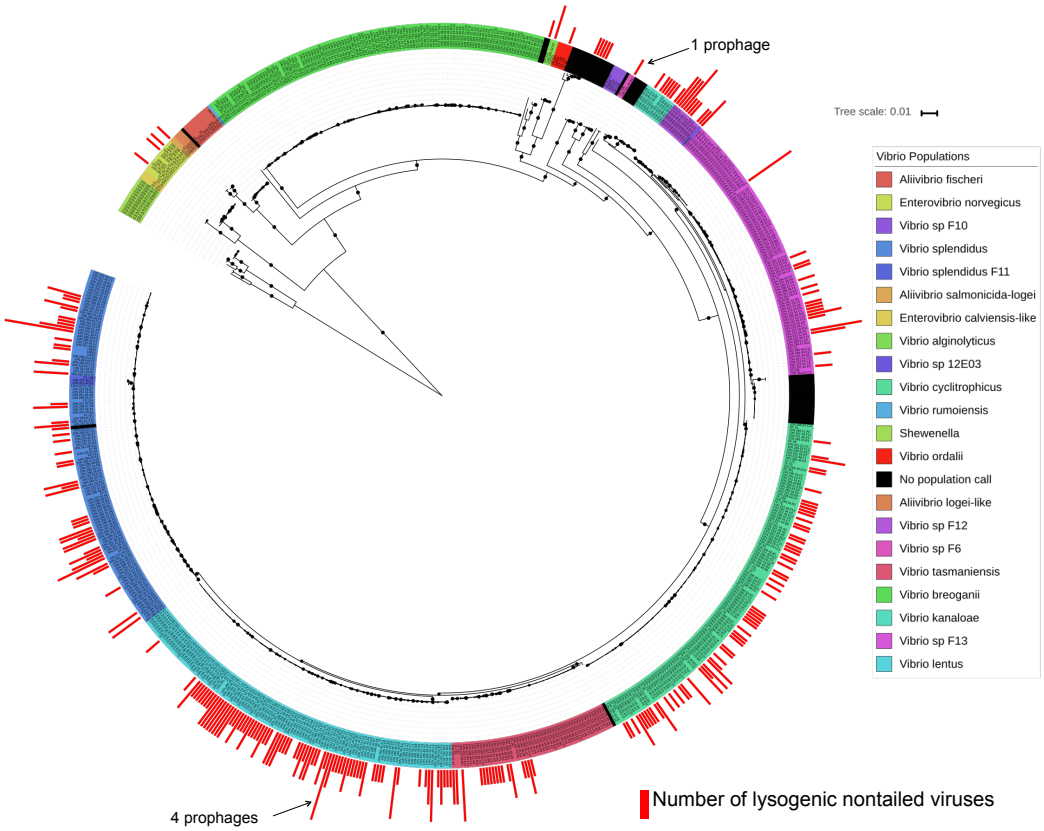


Figure A-1: **Fig.1:** A concatenated ribosomal protein tree of 758 *Vibrio* genomes shows 1/3 harbor greater than or equal to 1 lysogenic nontailed prophage.

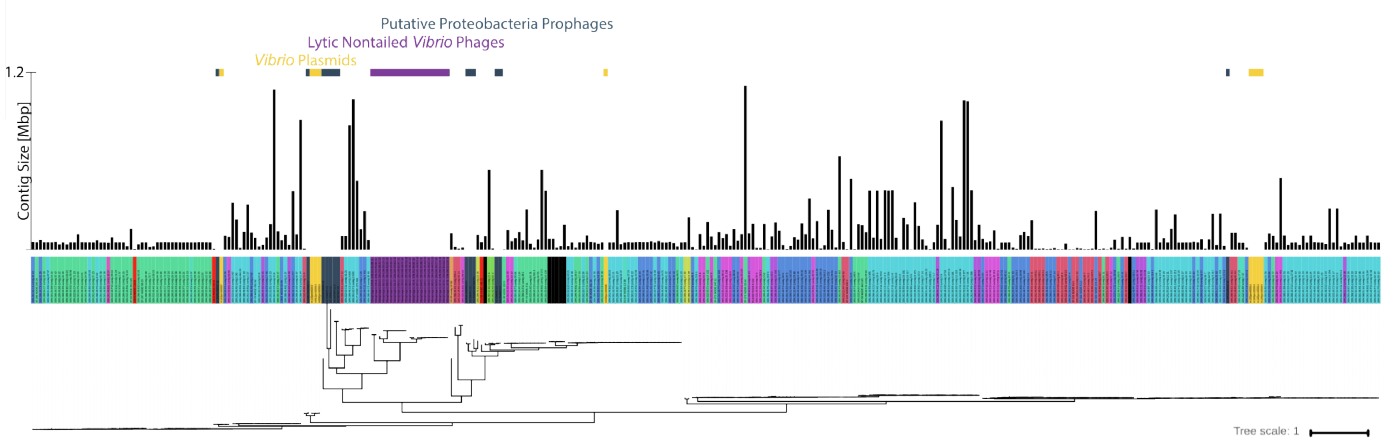


Figure A-2: **Fig.2:** Phylogenetic tree constructed using the aminoacid sequences of the major capsid protein (MCP). Colored annotations represent host populations (Same as Figure 1). Bar graph along the top displays the length of the genome contig upon which each element is found. Most noticeable is a set of closely related prophages which appear to insert into a putative 50kb plasmid.

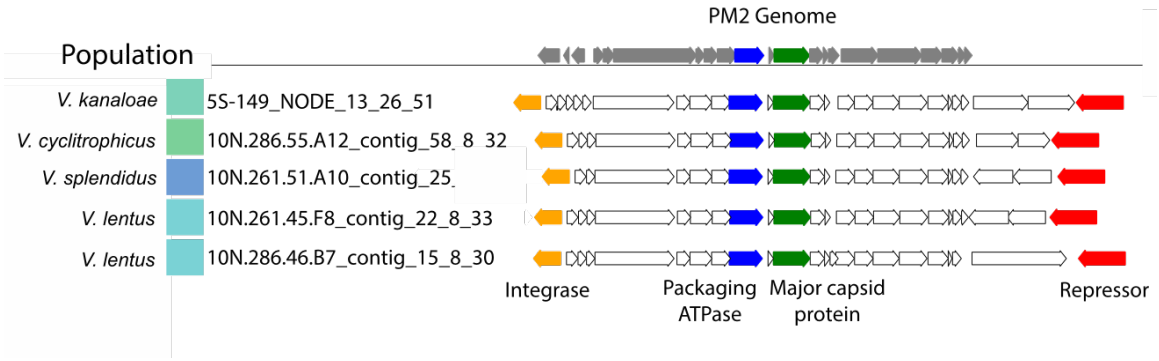


Figure A-3: **Fig.3:** Gene diagrams of nontailed prophages with identical major capsid protein sequences. Labels on the left show the host population in which each is found.

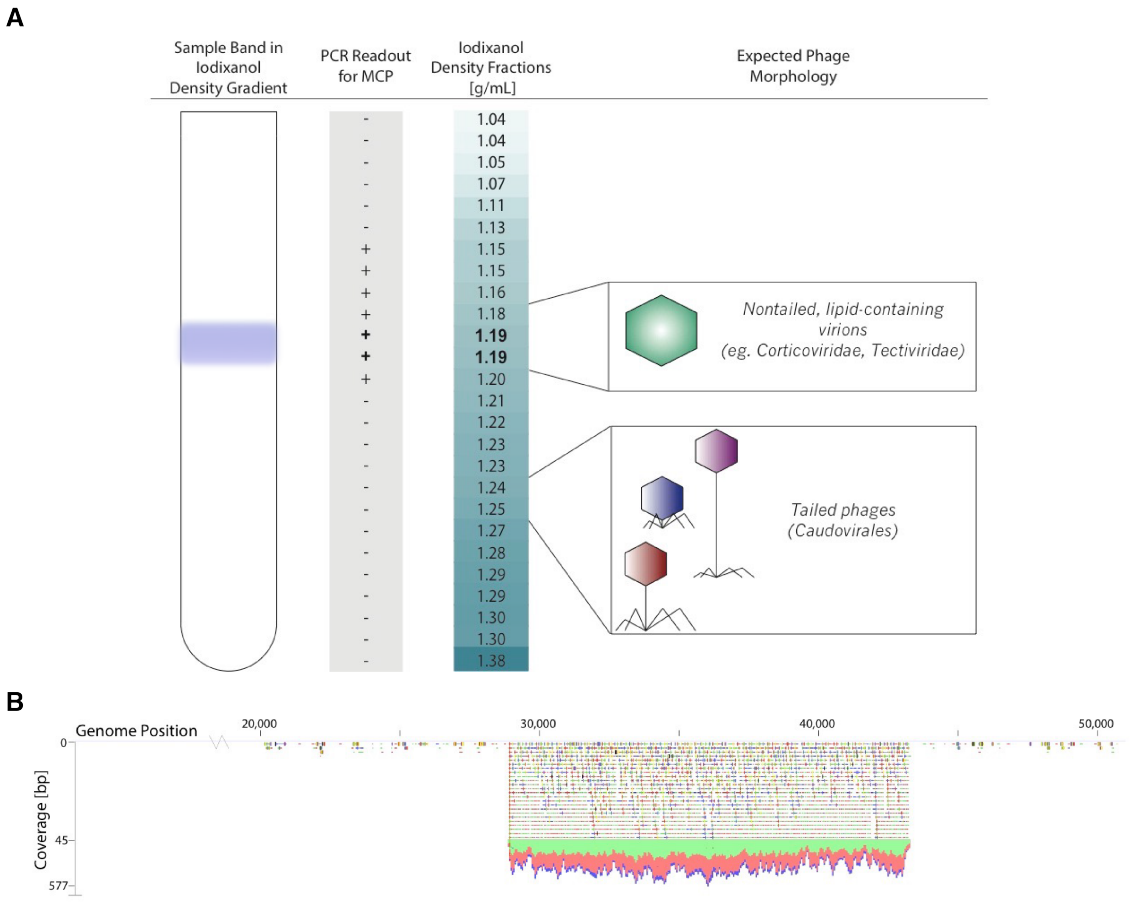


Figure A-4: **Fig.4:** (A) Representation of a density gradient depicting buoyant fraction for tailed and nontailed prophages. nontailed prophages equilibrate at a density of 1.18-1.19 g/mL in an iodixanol density gradient, verified using PCR targeting the major capsid protein. Tailed viruses typically equilibrate to density fractions in the range of 1.24-1.25 g/mL (extrapolated, see methods in Appendix B) (B) Mapping of excised virus reads to bacterial host's genome. Reference genome coordinates are displayed across the horizontal axis and coverage of viral reads are graphed on the vertical axis.

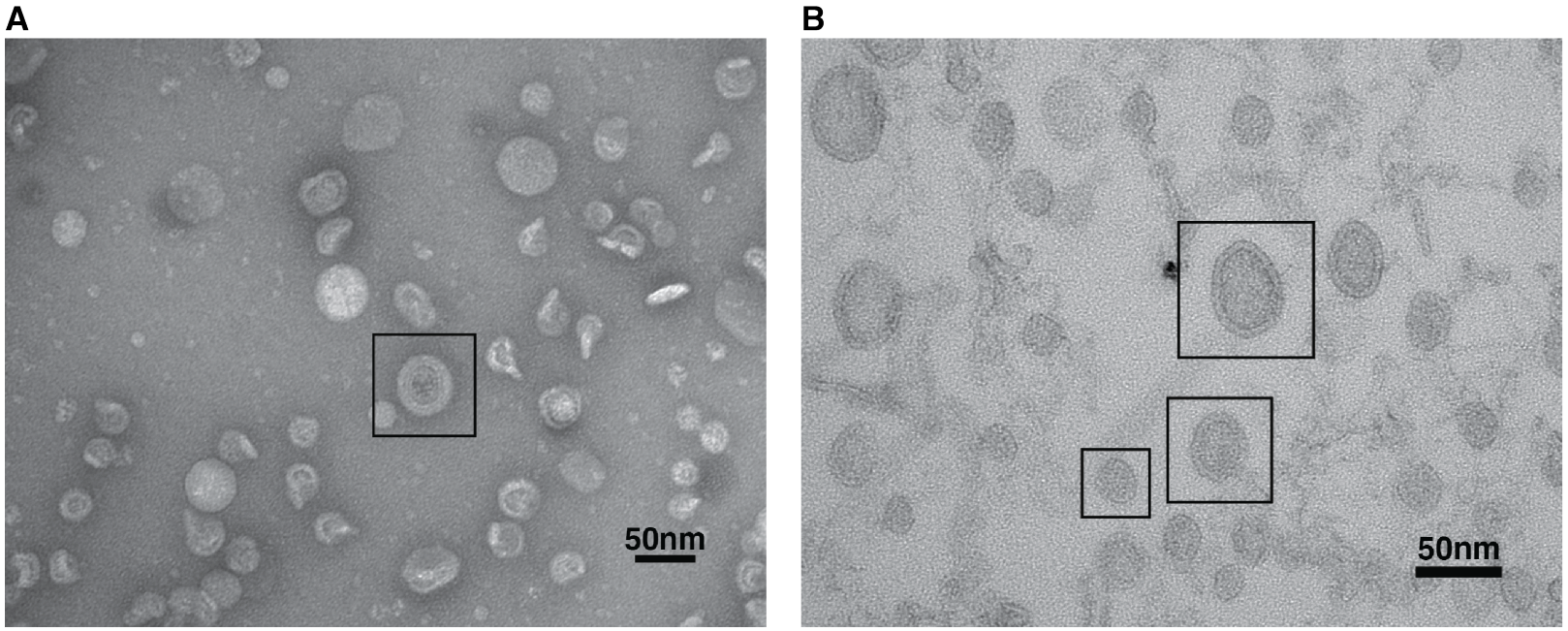


Figure A-5: **Fig.5:** Transmission electron microscopy imaging of nontailed prophages. (A) Negative staining of chloroform-treated supernatant containing nontailed prophages. A single nontailed prophage is seen (boxed) along with damaged vesicles. (B) Thin-section preparation of the same sample in (A). Various cross-sections of nontailed viruses are boxed. Variations in size are attributed to placement of thin-section slice along the center axis of the viruses.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix B

A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria

Kauffman, K. M., **Hussain, F. A.**, Yang, J., Arevalo, P., Brown, J. M., Chang, W. K., VanInsberghe, D., Elsherbini, J., Sharma, R. S., Cutler, M. B., Kelly, L., and Polz, M. F. (2018). A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. Nature, 554:118
doi:10.1038/nature25474

LETTER

doi:10.1038/nature25474

A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria

Kathryn M. Kauffman[1], Fatima A. Hussain[1], Joy Yang[1], Philip Arevalo[1], Julia M. Brown[2]†, William K. Chang[2], David VanInsberghe[1], Joseph Elsherbini[1], Radhey S. Sharma[1]†, Michael B. Cutler[1], Libusha Kelly[2] & Martin F. Polz[1]

The most abundant viruses on Earth are thought to be double-stranded DNA (dsDNA) viruses that infect bacteria[1]. However, tailed bacterial dsDNA viruses (*Caudovirales*), which dominate sequence and culture collections, are not representative of the environmental diversity of viruses[2,3]. In fact, non-tailed viruses often dominate ocean samples numerically[4], raising the fundamental question of the nature of these viruses. Here we characterize a group of marine dsDNA non-tailed viruses with short 10-kb genomes isolated during a study that quantified the diversity of viruses infecting Vibrionaceae bacteria. These viruses, which we propose to name the *Autolykiviridae*, represent a novel family within the ancient lineage of double jelly roll (DJR) capsid viruses. Ecologically, members of the *Autolykiviridae* have a broad host range, killing on average 34 hosts in four *Vibrio* species, in contrast to tailed viruses which kill on average only two hosts in one species. Biochemical and physical characterization of autolykiviruses reveals multiple virion features that cause systematic loss of DJR viruses in sequencing and culture-based studies, and we describe simple procedural adjustments to recover them. We identify DJR viruses in the genomes of diverse major bacterial and archaeal phyla, and in marine water column and sediment metagenomes, and find that their diversity greatly exceeds the diversity that is currently captured by the three recognized families of such viruses. Overall, these data suggest that viruses of the non-tailed dsDNA DJR lineage are important but often overlooked predators of bacteria and archaea that impose fundamentally different predation and gene transfer regimes on microbial systems than on tailed viruses, which form the basis of all environmental models of bacteria–virus interactions.

The dsDNA viruses consist of two ancient major lineages, both of which are proposed to have evolved from viruses that infect bacteria, and both include members that infect all three domains of life[5–8]. These two lineages emerged from ancestors with distinct folds in their major capsid proteins, the HK97 fold[9] and the DJR fold[10], and among the dsDNA bacterial viruses, these two groups are recognizable as 'tailed' and 'non-tailed' viruses, respectively. However, despite the DJR being the second most common capsid fold among all viral taxa[11], with the single jelly roll fold being the most common, bacterial DJR viruses are essentially missing from culture and sequence collections, which are instead dominated by HK97-lineage tailed viruses[12]. Whereas there are 215 described genera of tailed viruses[13], with 1,993 *Caudovirales* genomes in the NCBI RefSeq database (as of 3 October 2017)[14], there are only three described genera of non-tailed DJR bacterial and archaeal viruses, and 8 NCBI RefSeq genomes. Notably, only one of these sequenced DJR non-tailed viruses, the corticovirus PM2, which was isolated 50 years ago, is of marine origin[15]. This is particularly puzzling, given that electron microscopy-based surveys have revealed that non-tailed viruses comprise 51–92% of viruses observed in global surface oceans[4,16] and dsDNA viruses are thought to represent the majority of marine viruses[17], suggesting that non-tailed dsDNA viruses should be abundant. Directed efforts have led to the discovery that non-tailed RNA viruses that infect eukaryotes can be abundant[18], and that non-tailed single-stranded DNA (ssDNA) viruses that infect bacteria[19] are also diverse, although with a low abundance[20], in marine systems. However, it remains unresolved whether non-tailed dsDNA viruses, such as those in the ancient and diverse DJR capsid lineage, are contributors to the enigmatic non-tailed majority of viruses that dominate the global ocean.

In a large survey of viruses that infect the ubiquitous marine bacterial family Vibrionaceae, we recovered a diverse collection of non-tailed viruses from a quantitative assay that exposed 1,334 Vibrionaceae isolates to concentrates of co-occurring viruses (Methods). We used a quantitative isolation approach that enabled the capture of all viruses

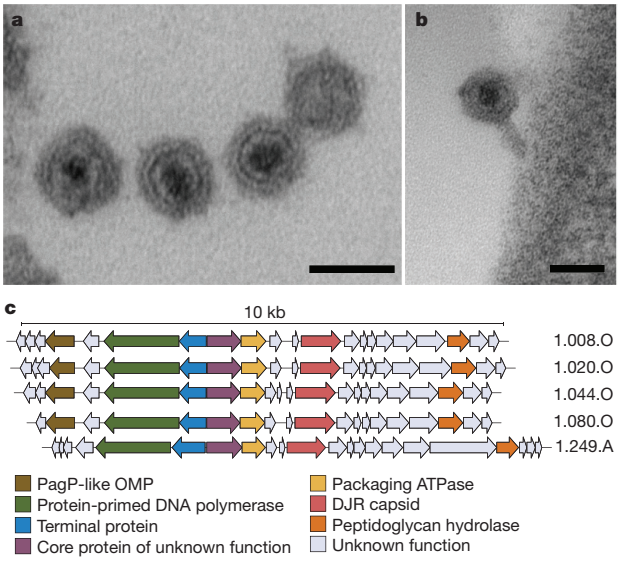


Figure 1 | *Autolykiviridae* is a new family of non-tailed dsDNA viruses in the DJR capsid lineage. a, Thin-section electron microscopy of autolykivirus plaques shows non-tailed virions with inner cores similar to those of the lipid-bilayer-containing non-tailed corticovirus PM2 (see Methods for experimental details and references). **b**, Rare virions show a tectivirus-like tail-tube-like structure when adjacent to cell membrane. Scale bars, 50 nm. **c**, Alignment of five genomes representing autolykivirus diversity, open reading frames are represented by block arrows. The linear 10-kb autolykivirus genomes have inverted terminal repeats and are shorter than those of the tailed viruses described here, which range from 21.7–348.9 kb (median = 47 kb) and encompass the range of tailed *Vibrio* virus genomes in GenBank.

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, New York 10461, USA. †Present addresses: Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine 04544, USA (J.M.B.); Department of Environmental Studies, Bioresources & Environmental Biotechnology Laboratory, University of Delhi, Delhi 110007, India (R.S.S.).

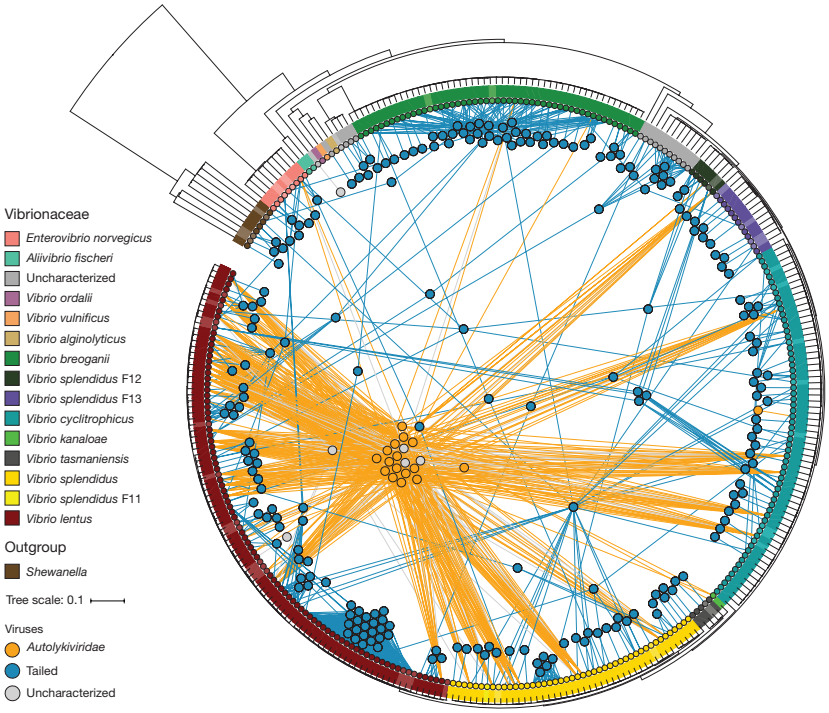


Figure 2 | Autolykiviruses dominate the lytic viral infection network of marine *Vibrio*. Inverted phylogenetic tree showing relationships among all 318 bacterial strains assayed based on concatenated alignments of *hsp60* and ribosomal protein genes, and using a partitioned model in RAxML[31] to allow placement of 40 strains for which only the *hsp60* gene sequence was available. Isolates are predominantly non-clonal. Leaves represent bacterial isolates coloured by species. Nodes represent 247 viruses described as *Autolykiviridae* ($n = 17$), tailed ($n = 224$) or uncharacterized ($n = 6$; no genome sequence). The edges represent infections coloured by viral type.

capable of lytic growth and colony (plaque) formation, mixing viral concentrates from co-occurring water samples and incubating them in solid-phase agar overlay for two weeks. Sequencing of 241 viruses that were randomly selected from each of 239 different plaque-positive hosts indicated that 18 of these viruses were a novel type that had small genomes (approximately 10 kb). Electron microscopy revealed that these viruses were non-tailed (Fig. 1a), although we also observed rare virions that showed tail-tube-like structures when in contact with cell membranes (Fig. 1b and Extended Data Fig. 1), consistent with known formation of such tubes during infection by other non-tailed viruses, including the dsDNA *Tectiviridae* (PRD1)[21] and the ssDNA *Microviridae* (PhiX174)[22]. Notably, the capsid size of a representative member (mean ± s.d. diameter, 49 ± 2 nm; Methods) of these novel viruses was closely aligned to the most abundant viral capsid size (mean ± s.d. 54 ± 12 nm) observed in the surface ocean by electron microscopy[4]. This size is similar to the size of the only described non-tailed marine dsDNA and RNA isolates of bacterial viruses, PM2 and 06 N-58P, respectively, which both have 60-nm diameter capsids[4], but is different from the size of the six described non-tailed ssDNA isolates of bacterial viruses, which have bimodal capsid diameter distributions centred around 31 nm and 72 nm (ref. 4). These observations suggest that these new viral isolates are representatives of the non-tailed viral majority.

Genome sequences and phylogenetic analyses of the non-tailed dsDNA *Vibrio* viruses show that they represent a new family of bacterial viruses, which we propose to name *Autolykiviridae*, in reference to Autolykos, a character in Greek mythology notable for being difficult to catch. Genome alignments of autolykivirus isolates reveal that they are diverse at the nucleotide level (Extended Data Fig. 2a), with whole-genome nucleotide identity as low as 31% (Extended Data Fig. 2b), yet display high synteny overall—sharing a core of six of their approximately 20 proteins, with additional proteins shared among subsets of the isolates (Fig. 1c, Extended Data Fig. 3 and Supplementary Data 1). Phylogenetic analyses reveal that members of the *Autolykiviridae* are most closely related to the corticovirus PM2 in their major capsid protein (21–25% amino acid identity, Extended Data Fig. 4a, b), are poorly resolved in their packaging ATPase (12–16% amino acid identity to *Corticoviridae* and *Turriviridae* viruses, Extended Data Fig. 4c, d) and are most closely related to members of the *Tectiviridae* in their protein-primed DNA polymerase (36–37% amino acid identity, Extended Data Fig. 4e, f). The high sequence divergence of autolykiviruses in these core genes, in addition to their divergent phylogenetic association with previously described virus families, supports their identification as a family-level lineage.

To characterize the potential ecological impact of autolykiviruses, we conducted a large-scale host-range assay, and found that they commonly killed hosts in multiple species, whereas the majority of tailed viruses killed only few and closely related hosts. We used a collection of Vibrionaceae viruses that were isolated by quantitative direct plating and tested infectivity of 241 viruses on 318 bacterial isolates. We found that the autolykiviruses were disproportionate contributors to lysis, responsible for 38% of killings although representing only 7% of all tested viruses (Fig. 2). Notably, despite the high genomic diversity of the autolykiviruses and of the hosts they infect, these viruses share extensively overlapping host-range profiles (Methods and Extended Data Fig. 5). This pattern is similar to that observed for members of the *Tectiviridae*, which infect hosts in multiple Gram-negative genera in a plasmid-dependent manner[23]. The finding that the autolykiviruses more commonly infect diverse species within a genus than tailed viruses suggests that these two groups may have fundamentally different impacts on microbial community ecology and evolution.

Biochemical and phenotypic characterization of members of the *Autolykiviridae* revealed several properties that make them subject to systematic loss in studies of viral diversity. Firstly, we found that

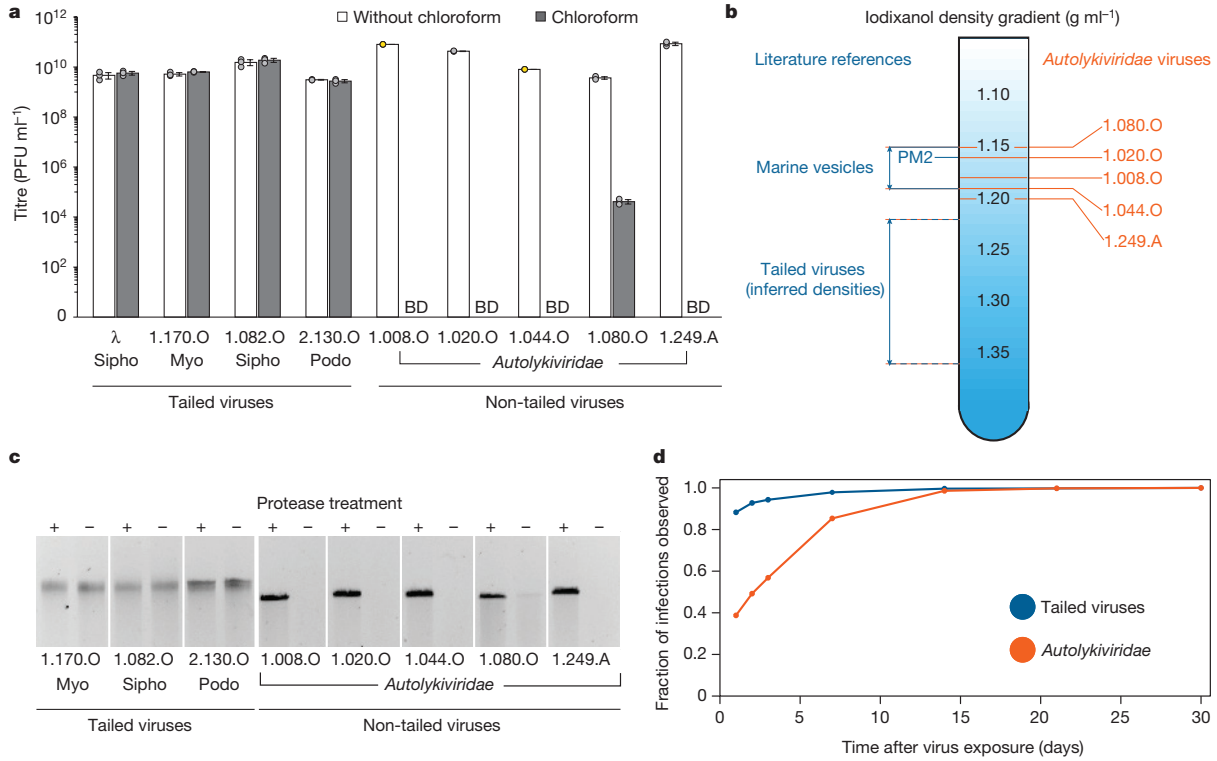


Figure 3 | Recovery of autolykiviruses is subject to multiple methodological biases. a, Comparison of chloroform sensitivity of tailed viruses and representative autolykiviruses (Extended Data Fig. 2), measured by plaque-forming assay after chloroform exposure. Data are mean ± s.d. of three independent replicates, data points in yellow represent lower-bound values. BD, below the detection limit of 199 plaque forming units (PFU) per ml. **b**, Buoyant density of autolykiviruses in iodixanol, in relation to previously reported densities for the lipid-containing non-tailed corticovirus PM2 and marine (outer membrane) vesicles (solid lines) in iodixanol, and inferred range of caesium-chloride-targeted tailed viruses (dashed lines) on the basis of linear extrapolation from PM2 (see Methods). **c**, Comparison of tailed virus and autolykivirus genome recovery with and without protease treatment. Protease-treated sample loading volumes normalized to 50 ng, equal volumes of untreated partner samples in adjacent lanes. The cropped gel image is representative of three independent experiments (gel source data are shown in Supplementary Fig. 1). **d**, Comparison of the cumulative proportion of observed tailed virus and autolykivirus killing over time. Infections ($n = 498$ and 844, autolykiviruses and tailed viruses, respectively) assayed as drop-spot clearings in large-scale host-range assay.

chloroform (Fig. 3a), a reagent that is commonly added to viral preparations to kill contaminating bacterial cells[24], reduced infectivity of autolykiviruses to below the level of detection. Secondly, we observed lower buoyant densities for autolykiviruses than those inferred for tailed viruses, probably owing to the presence of a lipid bilayer within their capsid, placing them outside the range that is commonly targeted in density gradient-based preparations of bacterial viruses from environmental samples[24,25] (Fig. 3b and Methods). Thirdly, owing to the presence of covalently bound proteins that alter DNA partitioning, the genomes of autolykiviruses require treatment with protease to enable efficient DNA extraction; this is not a standard component of extraction protocols targeting tailed viruses (Fig. 3c). Additional features, such as time to detection and decay rate, may also contribute to recovery bias (Methods, Fig. 3d and Extended Data Fig. 6). That *Autolykiviridae*-like viruses have not been definitively described for *Vibrio*, which have served as a major model of host–virus interactions and have been used to isolate viruses for nearly 100 years[26], suggests that the impact of these biases is severe and that related viruses that infect a diverse range of other bacteria are also likely to be systematically lost as a result of the same biases. We therefore suggest that, except for studies that specifically target subsets of viruses, viral concentrates for isolation and metagenomics are prepared: (1) without chloroform, (2) without density gradients and (3) with protease treatment during extraction.

Using combined cultivation and bioinformatic approaches, we show that DJR elements also exist as actively mobilizing prophages and episomes in *Vibrio*. Genome-integrated elements that have previously been identified as widespread putative corticovirus-like prophages[27] are active and naturally excise to produce nuclease-protected extracellular particles (Extended Data Fig. 7a, b). Furthermore, a set of broad host-range plasmids that have previously been identified as non-transmissible[28] encode DJR capsid proteins and associated packaging ATPases and are thus also DJR elements (Extended Data Fig. 7c, d). These findings suggest that DJR-encoding mobile elements that have been identified in cellular sequence databases, either as plasmids or integrated prophages, contribute to the pool of environmental non-tailed viruses.

We next investigated how diverse DJR elements are among bacteria and archaea, as well as in the marine environment. Considering the paucity and high divergence of reference sequences, we used a two-phase iterative hidden Markov model-based search approach to first generate a broad panel of DJR capsid sequences associated with putative prophages of bacterial and archaeal genomes, and to then search nine cellular and viral metagenomes that represent marine sediment- and water-column-derived samples (Methods and Extended Data Table 1), as well as NCBI environmental metagenomes.

Our searches reveal that the diversity and host associations of DJR viruses far exceeds the level that is currently recognized, as putative DJR prophage capsids were identified in 13 bacterial and archaeal phyla and metagenomic sequences, suggesting the existence of at least 13 additional novel lineages with unknown hosts. Whereas DJR viruses and prophages had previously been shown to infect two archaeal[6] and two bacterial phyla[23,27], the first phase of our search revealed the presence

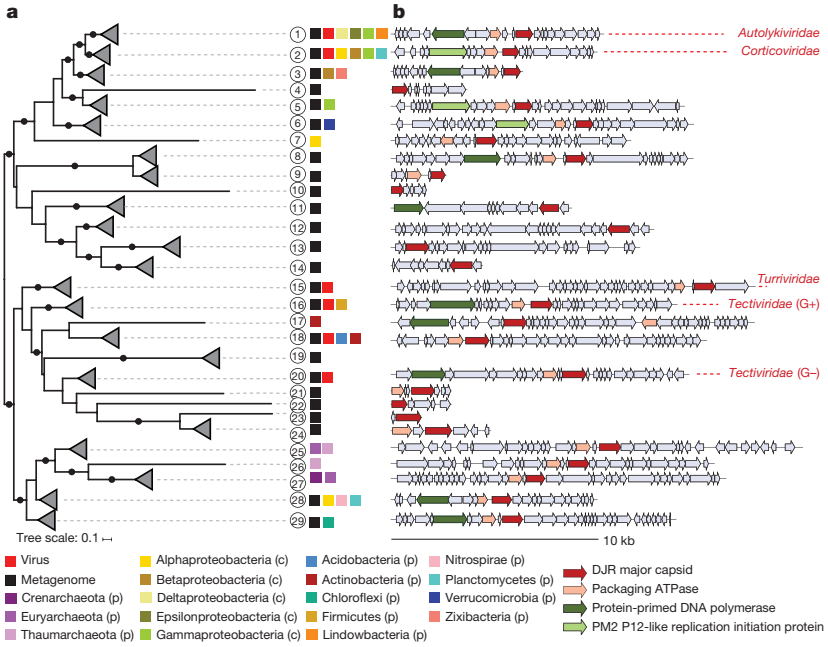


Figure 4 | DJR capsid viruses are far more diverse than the three currently recognized families, and include hosts in diverse bacterial and archaeal phyla. **a**, Phylogeny of 442 bacterial and archaeal DJR virus capsid proteins (sequences in Supplementary Data 1), including representatives of three previously described DJR virus families and sequences newly identified here; group numbers are assigned to each branch for reference, coloured blocks indicate hosts, black circles on branches indicate approximate likelihood-ratio test (aLRT) branch support ≥0.9. **b**, Element gene diagrams from each group show prophage host genome neighbourhoods and metagenome contigs often contain additional genes common to DJR elements (contig information in Extended Data Table 2). G+, Gram-positive; G−, Gram-negative.

of DJR virus capsids in genomes of nine additional phyla, including the two most abundant groups in the marine environment, the Alphaproteobacteria and the Thaumarchaeota (Fig. 4a and Methods). Moreover, analyses of marine metagenomes reveal that the environmental diversity of bacterial and archaeal DJR capsids exceeds that of our reference panel by several-fold (Fig. 4a). To specifically and conservatively evaluate the diversity of bacterial and archaeal DJR viruses, we selected only sequences with strong support for structural similarity to these groups for further analyses, omitting a large cluster of putative eukaryotic DJR viruses and sequences with no detectable similarity to known proteins (Extended Data Fig. 8). DJR genomic neighbourhoods encompass other viral proteins, and carriage of the protein-primed polymerase, which is associated with the presence of covalently bound terminal proteins, is common across deeply divergent lineages (Fig. 4b). Members of these groups would thus also be subject to the protease-dependent extraction bias (Fig. 3c).

The discovery of the autolykiviruses provides insight into the nature of the non-tailed viruses that dominate the global surface ocean, and suggests that dsDNA bacterial and archaeal DJR viruses have been systematically excluded from discovery. By providing genome-sequenced isolates and optimized approaches for targeted recovery of additional diverse representatives, we address a major challenge for metagenomic surveys—the paucity of viral reference genomes necessary for the interpretation of the uncharacterized majority of sequence diversity and function[29]. The extensive sequence diversity that we find among bacterial and archaeal DJR elements suggests that additional, culture-based reference sequences will be required to assess their true environmental diversity. The distinctively broad host ranges of members of the *Autolykiviridae* and related DJR elements also suggest that, if such viruses are capable of packaging host DNA, they may have an even more important role in facilitating the observed gene transfers between highly divergent bacteria in microbial communities[30] than the highly specific tailed viruses. Finally, the recovery of the non-tailed autolykiviruses represents a first step in revealing extensive missed diversity in one of the two major ancient lineages of dsDNA bacterial viruses and suggests that their ecological and evolutionary importance for microbial systems is far greater than is currently recognized.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 August 2016; accepted 28 December 2017.
Published online 24 January 2018.

1. Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64,** 69–114 (2000).
2. Krishnamurthy, S. R. & Wang, D. Origins and challenges of viral dark matter. *Virus Res.* **239,** 136–142 (2017).
3. Krupovic, M., Prangishvili, D., Hendrix, R. W. & Bamford, D. H. Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiol. Mol. Biol. Rev.* **75,** 610–635 (2011).
4. Brum, J. R., Schenck, R. O. & Sullivan, M. B. Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J.* **7,** 1738–1751 (2013).
5. Benson, S. D., Bamford, J. K. H., Bamford, D. H. & Burnett, R. M. Does common architecture reveal a viral lineage spanning all three domains of life? *Mol. Cell* **16,** 673–685 (2004).
6. Krupovič, M. & Bamford, D. H. Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. *Virology* **375,** 292–300 (2008).
7. Pietilä, M. K. *et al.* Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story. *Proc. Natl Acad. Sci. USA* **110,** 10604–10609 (2013).
8. Koonin, E. V., Dolja, V. V. & Krupovič, M. Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* **479–480,** 2–25 (2015).
9. Wikoff, W. R. *et al.* Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* **289,** 2129–2133 (2000).
10. Krupovič, M. & Bamford, D. H. Virus evolution: how far does the double β-barrel viral lineage extend? *Nat. Rev. Microbiol.* **6,** 941–948 (2008).
11. Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Natl Acad. Sci. USA* **114,** E2401–E2410 (2017).
12. Iranzo, J., Krupovic, M. & Koonin, E. V. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *MBio* **7,** e00978–16 (2016).
13. International Committee on Taxonomy of Viruses. ICTV Master Species List v.1.3 https://talk.ictvonline.org/files/master-species-lists/m/msl/6776 (2016).

14. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **43,** D571–D577 (2015).
15. Espejo, R. T. & Canelo, E. S. Properties of bacteriophage PM2: a lipid-containing bacterial virus. *Virology* **34,** 738–747 (1968).
16. Wommack, K. E., Hill, R. T., Kessel, M., Russek-Cohen, E. & Colwell, R. R. Distribution of viruses in the Chesapeake Bay. *Appl. Environ. Microbiol.* **58,** 2965–2970 (1992).
17. Andrews-Pfannkoch, C., Fadrosh, D. W., Thorpe, J. & Williamson, S. J. Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. *Appl. Environ. Microbiol.* **76,** 5039–5045 (2010).
18. Steward, G. F. *et al.* Are we missing half of the viruses in the ocean? *ISME J.* **7,** 672–679 (2013).
19. Labonté, J. M. & Suttle, C. A. Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J.* **7,** 2169–2177 (2013).
20. Roux, S. *et al.* Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **4,** e2777 (2016).
21. Peralta, B. *et al.* Mechanism of membranous tunnelling nanotube formation in viral genome delivery. *PLoS Biol.* **11,** e1001667 (2013).
22. Sun, L. *et al.* Icosahedral bacteriophage ΦX174 forms a tail for DNA transport during infection. *Nature* **505,** 432–435 (2014).
23. Saren, A.-M. *et al.* A snapshot of viral evolution from genome analysis of the *Tectiviridae* family. *J. Mol. Biol.* **350,** 427–440 (2005).
24. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4,** 470–483 (2009).
25. Castro-Mejía, J. L. *et al.* Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. *Microbiome* **3,** 64 (2015).
26. D'Herelle, F. Studies upon Asiatic cholera. *Yale J. Biol. Med.* **1,** 195–219 (1929).
27. Krupovič, M. & Bamford, D. H. Putative prophages related to lytic tailless marine dsDNA phage PM2 are widespread in the genomes of aquatic bacteria. *BMC Genomics* **8,** 236 (2007).
28. Xue, H. *et al.* Eco-evolutionary dynamics of episomes among ecologically cohesive bacterial populations. *MBio* **6,** e00552–e15 (2015).
29. Brum, J. R. & Sullivan, M. B. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13,** 147–159 (2015).
30. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480,** 241–244 (2011).
31. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. King, P. Weigele, J. Daily and J. Chodera for comments and suggestions; T. Soni and members of the Polz laboratory for assistance with sampling; S. Labrie for guidance in viral genome extractions and sequencing library preparation, and C. Haase-Pettingell for assistance with density gradients; N. Watson for electron microscopy; and R. Ratzlaff for discussions and the suggestion of electron microscopy of virus plaques in agar overlay. This work was supported by grants from the National Science Foundation OCE 1435993 to M.F.P. and L.K., the NSF GRFP to F.A.H. and the WHOI Ocean Ventures Fund to K.M.K.

Author Contributions K.M.K., F.A.H., L.K. and M.F.P. designed the study and planned experiments and analyses. K.M.K., L.K. and M.F.P. wrote the paper with contributions from all authors. K.M.K. conducted field sampling, isolations and experimental characterizations of lytic viruses. J.Y. conducted the statistical analyses of the viral decay experiment and wrote the scripts to visualize the infection matrix as a phylogeny-anchored network, which was based on the host ribosomal protein tree generated by P.A. W.K.C. and L.K. performed the quantification of significance of host sharing. F.A.H. performed isolation and characterization of active *Vibrio* DJR prophages. Bacterial genome sequencing libraries were prepared by M.B.C., assembled by P.A., and curated and annotated by P.A. and J.E. The viral genome sequencing libraries were prepared by K.M.K. and R.S.S., assembled by J.M.B. and K.M.K., and annotated and curated by J.M.B., K.M.K., J.E., W.K.C. and L.K. The viral metagenome sequencing libraries were prepared by K.M.K., and assembled and curated by P.A. and L.K. The bioinformatic analyses of microbial genomes and metagenomes for DJR capsid elements were performed by L.K. and K.M.K., and the visualization of the DJR network was performed by D.V. M.B.C. provided field and laboratory technical support throughout. Although specific contributions are highlighted for each author, all authors contributed in additional ways through contributions to figures, analyses, discussion of results and comments on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to M.F.P. (mpolz@mit.edu) or L.K. (libusha.kelly@einstein.yu.edu).

Reviewer Information *Nature* thanks J. Fuhrman, E. V. Koonin and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Isolation, culturing and sequencing of bacteria and viruses. Bacteria and viruses were collected from the littoral marine zone at Canoe Cove, Nahant, Massachusetts, USA, on 22 August (ordinal day 222), 18 September (261) and 13 October (286) 2010.

Bacteria were collected using previously described size-fractionation and selective-medium cultivation-based methods[32]. Bacterial genome libraries were prepared for sequencing using a tagmentation-based approach and 1–2 ng input DNA per isolate, as previously described[33]. Genomes were sequenced in multiplexed pools of 50–60 samples per Illumina HiSeq lane. Accession numbers for all bacterial genomes are provided in Supplementary Data 3 and are included under NCBI BioProject PRJNA328102.

Bacterial phylogenetic relationships were determined by concatenation of ribosomal proteins and *hsp60* sequences. For all strains with available genome sequences (278), ribosomal proteins were extracted from genomes with hmmsearch[34] and aligned with MAFFT[35] as previously described[36]. Full-length *hsp60* sequences were also extracted from these genomes using hmmsearch with default parameters and the Cpn60 hmm (PF00118) from Pfam[37]. The *hsp60* sequences were aligned using the mafft-fftnsi algorithm. Sanger-sequenced *hsp60* fragments from 40 strains that lacked genome sequences were added to this alignment using the mafft-fftnsi algorithm with the –addfragments option. The *hsp60* alignment was concatenated to the ribosomal protein alignment and used to create a phylogeny using RAxML under a partitioned general time reversible (GTR) model (options: –q, -m GTRGAMMAX)[31]. Shimodaira–Hasegawa (SH)-like supports were calculated using RAxML and taxonomy was assigned by manual inspection[36].

Viruses were collected using a previously described iron flocculation approach[38], using 4-l sample volumes, 0.2-μm pre-filtration to remove bacteria, 0.2-μm filters for floc capture, and oxalate solution for resuspension to maintain virus viability. Isolation of viruses was performed by directly plating virus concentrates in agar overlays on hosts from each of the same days, as follows. Iron-oxalate concentrate volumes equivalent to 15 ml of seawater were mixed with 150 μl of overnight host culture and 2 ml molten top agar to form host lawns in overlay and allow for plaque formation (top agar: 52 °C, 0.4% agar, 5% glycerol, in 2216 Marine Broth (2216MB); bottom agar: 1% agar, 5% glycerol, 125 ml l^{-1} of chitin supplement (40 g l^{-1} coarsely ground chitin, autoclaved, 0.2-μm filtered) in 2216MB). After incubation for two weeks, all plaques from plates containing fewer than approximately 25 plaques were archived, and a random subsample of each distinct plaque morphotype was archived for plates with more plaques. Plaque plugs were first eluted in 200 μl of 2216MB, a subsample of 150 μl was then filtered to remove bacteria for storage of virions at 4 °C, and the remainder was supplemented with 50% glycerol for storage at −20 °C. For purification and amplification, one archived plaque was randomly selected from all available plaques for each host and serially re-passaged at least three times, with stocks preferentially recovered from −20 °C archives. In a small number of cases, multiple plaques were purified for a given host and these are identifiable by the nomenclature described below.

Sequencing and genome analysis of viruses was as follows. In brief, high-titre plate lysates of serially purified viruses were concentrated using 30-kDa centrifugal filter units (Millipore, Ultracel 30K, UFC903024) and washed with 1:100 2216MB to reduce salts for nuclease treatment. Concentrates were brought to approximately 500 μl using 1:100-diluted 2216MB and then treated with DNase I and RNase A for 65 min at 37 °C to digest unencapsidated nucleic acids. Nuclease-treated viral lysates were extracted by addition of 1:10 final volume of SDS mix (0.25 M EDTA, 0.5 M Tris-HCl (pH 9.0), 2.5% sodium dodecyl sulphate), 30 min incubation at 65 °C; addition of 0.125 volumes 8 M potassium acetate, 60 min incubation on ice; addition of 0.5 volumes phenol–chloroform; and recovery of nucleic acids from aqueous phase by isopropanol and ethanol precipitation. Genomes were fragmented by sonication, libraries sequenced in multiplexed pools using Illumina MiSeq and HiSeq technologies, assembled using CLC Genomics Workbench v6.5.1 and v8.5.1 and CLC assembly cell v4.4.2.133896, and manually curated to standardize genome start positions for the *Caudovirales*. Bioinformatic analyses indicate that all sequenced viruses, except autolykiviruses, are members of the *Caudovirales*.

The viral-strain naming convention is described using the example of 1.008.O._10N.286.54.E5, with specific identifiers separated by a full stop. The first position (here '1') represents a unique identifier for each independent plaque isolated from a given host from the initial exposure of that host to an environmental virus concentrate. The second position (here '008') represents a unique working ID for a host strain. The third position (here 'O') indicates a unique sublineage generated from a single plaque during viral serial purification, for example, owing to the emergence of multiple plaque morphologies. Following the underscore is the full strain ID of the host of isolation. Viral genome accession numbers are provided in Supplementary Data 3 and are included under NCBI BioProject PRJNA328102.

Characterization of virions of autolykiviruses. Morphology of the autolykiviruses was determined by thin-section electron microscopy (TEM) of a representative member, 1.008.O (Fig. 1a, b and Extended Data Fig. 1a–c). TEM was performed once on a single agar overlay. Viruses were visualized by generating plaques in the agar overlay and then fixing the overlay for 14 h by addition of fixative (2.5% glutaraldehyde, 3% paraformaldehyde with 5% sucrose in 0.1 M sodium cacodylate buffer (pH 7.4)). The top agar was collected into 0.1 M sodium cacodylate buffer (pH 7.4), pelleted and washed in sodium cacodylate buffer, soaked overnight in 1% OsO_4 in veronal-acetate buffer, stained en bloc overnight with 0.5% uranyl acetate in veronal-acetate buffer, dehydrated and embedded in Embed-812 resin. Ultrathin sections were prepared on a Leica Ultracut E microtome with a Diatome diamond knife, stained with 2% uranyl acetate and lead citrate. Sections were examined using an FEI Tecnai Spirit electron microscope at 80 kV and photographed with an AMT CCD camera. The selection of a region containing both uninfected and lysed cells allowed for capture of multiple stages of infection without the need for optimization of infection course timing. Capsid measurements were made using ImageJ[39]. Ten virus particles in a single image (magnification of 98,000×) were each measured at three different cross sections and the average calculated for each virus was used to determine the overall mean and standard deviation. Observations of inner-core and tail-tube-like structures are consistent with those previously described for the lipid-bilayer-containing non-tailed corticovirus[40] and the non-tailed tectivirus PRD1[21], respectively.

Ultracentrifugation in Optiprep iodixanol gradient medium (Sigma D1556) was used to determine the density of representative autolykiviruses. This medium was selected on the basis of previous demonstrations of sensitivity of some viruses, including the related corticovirus PM2, to caesium chloride and sucrose gradients[24]. Use of iodixanol also allowed for direct culture-based assay of viral activity in density gradient fractions without prior dialysis, as required for other density gradient media. Density gradients were prepared using artificial seawater (ASW, Sigma S9883, 40 g l^{-1}) diluent, as follows: eight density layers spanning 20% to 54% iodixanol were manually laid in Seton 7030 tubes and loaded with 500 μl of polyethylene glycol (PEG)-precipitated viral concentrate resuspended in ASW. Samples were centrifuged for 10 h at 20 °C and 35,000 r.p.m. in a SW41 swinging-bucket rotor in a Beckman L8M ultracentrifuge. Density gradient fractions were collected using the Biocomp Piston Gradient Fractionator (BioComp Instruments Inc.) and densities determined as mass per 100 μl volume, using a standard laboratory pipette[41]. Densities for each virus were defined as the fraction with greatest plaque-forming activity in an agar-overlay assay of density fractions collected from a single density column. Data shown in Fig. 3b are from a single experiment that included five representative autolykiviruses (1.008.O, 1.18 g ml^{-1}; 1.020.O, 1.16 g ml^{-1}; 1.044.O, 1.19 g ml^{-1}; 1.080.O, 1.15 g m l^{-1}; and 1.249.A, 1.20 g ml^{-1}), processed together in a common centrifugation run. This test set also included the lipid-containing PRD1 tectivirus as an internal control, however, this tectivirus showed a bimodal distribution of peak infectivity (1.15 g ml^{-1} and 1.21 g ml^{-1}) associated with distinct plaque morphologies, observations consistent with the presence of both the expected lower-density PRD1 and a contaminating higher-density tailed prophage in the stock. A subsequent independent iodixanol density gradient centrifugation experiment that included both an autolykivirus (1.044.O) and a tailed virus (1.255.O) also showed a lower density for the autolykivirus (1.17 g ml^{-1}) than for the tailed virus (1.22 g ml^{-1}).

The buoyant densities of corticovirus PM2[42] and marine vesicles[43] in iodixanol (1.16 g ml^{-1} and 1.15–1.19 g ml^{-1}, respectively) indicated in Fig. 3b are based on literature values, and the range for tailed viruses is inferred from the literature value for PM2 assuming a linear correspondence between the density of any virus in iodixanol and its density in caesium chloride. Using values from a study[42] in which densities for PM2 were determined in both iodixanol (1.16 g ml^{-1}) and caesium chloride (1.28 g ml^{-1}), we infer that the tailed viruses targeted at the 1.35–1.50 g ml^{-1} interface in caesium chloride would span densities from 1.22–1.36 g ml^{-1} in iodixanol. We note that whereas there is extensive data showing that lipid-containing viruses are outside of the range commonly targeted for tailed viruses in caesium chloride, data on tailed virus densities measured in both caesium chloride and iodixanol are lacking. We therefore caution that our approximation is a guideline, and that future studies attempting targeted isolation from iodixanol density gradient media should first ensure that iodixanol also yields the desired separation to the same extent as caesium chloride.

Chloroform sensitivity of members of the *Autolykiviridae* was assessed using a 0.2-volume ratio of chloroform to virus-containing solution, as commonly applied to viral concentrates to eliminate bacterial contamination[24]. The test set of viruses (Fig. 3a) included five representative autolykiviruses (1.008.O, 1.020.O, 1.044.O, 1.080.O, 1.249.A; all with genome size of approximately 10 kb) and four representative tailed viruses, including the *Escherichia coli* siphovirus Lambda, and three representative Vibrionaceae viruses: a siphovirus (1.082.O; approximately

36 kb), a myovirus (1.170.O; approximately 134 kb) and a podovirus (2.130.O; approximately 76 kb). All viruses were tested in three independent replicates, and each replicate included a pair of samples, with and without chloroform exposure. Chloroform-treatment samples were mixed with chloroform, all samples were gently vortexed for 6 s, incubated at room temperature, and mixed twice by finger-flick over 2 h. Samples were then centrifuged at 5,000 *g* for 5 min and the activity was assessed using a dilution series of drop spots (5 µl) on host agar-overlay lawns, including no-virus controls to allow detection of chloroform carry-over.

The effect of protease treatment on recovery of nucleic acids from tailed and autolykivirus viruses was assessed using a method commonly applied to generate marine viral metagenomes[44,45]. The test set of viruses (Fig. 3c) included five representative autolykiviruses (1.008.O, 1.020.O, 1.044.O, 1.080.O, 1.249.A; all approximately 10 kb) and three representative tailed Vibrionaceae viruses: a siphovirus (1.082.O; around 36 kb), a myovirus (1.170.O; about 134 kb) and a podovirus (2.130.O; approximately 76 kb). Each viral concentrate was extracted three times independently, each time with samples in a different order, as follows. Lysates were nuclease-treated in 50 µl reactions containing 1× Turbo DNase buffer, 1 µl Turbo DNase (Ambion AM2239), 0.5 µl RNase A (Thermo Scientific EN0531) and incubated at 37 °C for 60 min. Nuclease reactions were halted by addition of EDTA to 100 mM and heat inactivation for 10 min at 75 °C. Samples treated with protease received 0.5 µl of proteinase K (Epicentre MPRK092) per 62.5 µl reaction; all samples were incubated at 65 °C for 20 min. Four sub-aliquots for each virus sample and treatment were pooled to a final volume of 250 µl, mixed with 1 ml Wizard PCR Preps DNA Purification Resin (Promega A718A), loaded onto Wizard Minicolumns (Promega A721B), washed with 2 ml 80% isopropanol and collected by centrifugal wash with 80 °C TE buffer. dsDNA concentrations were quantified by fluorescence using the Quant-iT PicoGreen kit as per the manufacturers' protocol (ThermoFisher Scientific P7589). Products were visualized by agarose gel electrophoresis (0.7% agarose, 0.5× TBE, 90 V for 90 min) with load volumes normalized to 50 ng across proteinase K-treated samples and all corresponding non-proteinase K-treated replicates loaded at equal volume to their treated replicates from the same independent experiment. The included gel image (Fig. 3c, gel source data are shown in Supplementary Fig. 1) is a representative experiment that contains a single replicate of each virus, for this representative experiment the kruskal. test implementation of the Kruskal–Wallis test in R (v3.3.0) was used to test for differences between autolykiviruses and tailed viruses in the fold difference in DNA recovered with protease compared to without protease, measured by PicoGreen fluorescence (two-sided Kruskal–Wallis rank-sum test; $\chi^2 = 5$, d.f. $= 1$, $P = 0.02535$; $n = 5$ and 3, median $= 8.35$ and 0.97, for autolykiviruses and tailed viruses, respectively).

Evaluation of decay rates of autolykiviruses and tailed viruses. Decay of eight viruses (five autolykiviruses and three tailed) was monitored in ASW in borosilicate vials at room temperature in the dark over 34 days. Activity was measured by drop-spot plating of three independent serial dilutions of each of four replicate samples of each virus on days 0, 1, 10, 20 and 34. A linear mixed model for the decay data was fit using the lme4 package[46,47] in R[48], with $\log_{10}$ of the PFU counts as the response variable, an intercept (starting PFU) and slope over time (decay rate) as fixed effects, and intercept and slope for each virus as well as intercept and slope for each bottle nested in each virus as random effects. Decay rates measured in log loss per day (t) over the observation period were variable among viruses and substantially higher for two of the autolykiviruses (1.008.O, $-0.03t$; 1.020.O, $-0.08t$) than for the other autolykiviruses (1.044.O, $-0.02t$; 1.080.O, $-0.01t$; and 1.249.A, $-0.01t$) and the tailed viruses (podovirus 2.130.O, $-0.01t$; myovirus 1.170.O, $0t$; siphovirus 1.082.O, $0t$); 95% conditional predictive intervals of the autolykiviruses showed no overlap with the myovirus or siphovirus, nor did those for autolykiviruses 1.008.O and 1.020.O show overlap with the other autolykiviruses.

Annotation of DJR element genomes, contigs and genomic neighbourhoods. Open reading frames for all virus, plasmid, prophage and metagenomic contigs were identified using Prodigal[49] v2.6.1 with the -p meta option. Elements not sequenced as part of this study were recovered as follows: parent nucleotide sequences of all DJR proteins with accession numbers were downloaded manually through NCBI Batch Entrez; where DJR proteins occurred in microbial genome backbones, regions of 20 kb centred around the DJR were downloaded. Proteins called *de novo* during this work from metagenomic contigs were re-associated with their parent contigs. Protein sequences derived from the OM-RGC were linked back to their metagenomic assemblies using the *Tara* Oceans companion website tsv table (ftp://ftp.sra.ebi.ac.uk/vol1/ERA412/ERA412970/tab/OM-RGC_seq.release.tsv.gz)[50]. Metagenomic assemblies were downloaded from the European Nucleotide Archive (http://www.ebi.ac.uk/ena/about/tara-oceans-assemblies)[50,51] and scaffolds associated with each hit were extracted from the assemblies. Clusters of homologues were identified by performing an all-by-all BLASTp, requiring a minimum bitscore of 50, and clustering all pairs unweighted using Markov cluster algorithm (MCL)[52] v14.137 with an inflation parameter set to 1.4. Structural annotations were performed using the Phyre2[53] webportal and, for a subset of proteins, HHpred[54] through the MPI Bioinformatics Toolkit[55]. Sequence similarity-based annotations were performed using BLASTp searches against NCBI RefSeq Virus genome, the NCBI Batch Web Conserved Domain[56] search tool, and EggNOG-Mapper[57]. Sequences were also annotated with InterProScan[58] v5.17-56.0 using the iprlookup, goterms and pathways options, and including two optional databases, TMHMM and SignalP[59], in addition to the 13 default databases. All annotations and cluster information are provided in Supplementary Table 1. Genome diagram figures were prepared using the GenoPlotR[60] package in R and refined in Adobe Illustrator.

Detailed analyses of the gene of unknown function adjacent to the gene encoding the protein-primed DNA polymerase (pDNApol) in autolykivirus genomes suggest that it encodes the terminal protein that is necessary for the protein-primed DNA polymerases to initiate replication. The observations that: (1) this is a core gene shared by all autolykiviruses; (2) the secondary structure predictions for the encoded protein are consistent with other terminal proteins; and (3) it is located adjacent to the pDNApol, as is the gene for the terminal protein in PRD1 and phi29, are strong evidence to suggest that this orthologous cluster in the autolykiviruses represents a novel terminal protein.

Construction of alignments and phylogenetic trees. To evaluate nucleotide diversity among members of the *Autolykiviridae*, we performed whole-genome alignments using the EMBL–EBI implementation of Clustal Omega[61–63] and PhyML[64] with SMS[65] v1.8.1 (Extended Data Fig. 2). To evaluate the relationship of the autolykiviruses to known DJR viruses, and the support for their establishment as a new viral family, we evaluated gene trees for three conserved genes representative of the structural and replication functions of these viruses, the major capsid protein and packaging ATPase, and pDNApol, respectively. We included only bacteria- and archaea-infecting viruses, excluding eukaryote-infecting DJR viruses, and defined membership in each of the gene trees as described in ref. 12, with the exception that we excluded members of the *Caudovirales* from the pDNApol tree. When protein sequences were available in the pVOG database[66], these were used, otherwise sequences were downloaded from NCBI RefSeq[14]. Included in the major capsid protein tree were members of the *Tectiviridae*, *Corticoviridae* and *Turriviridae*; of these only the Gram-positive bacteria-infecting members of the *Tectiviridae* were included in a pVOG (VOG0339), with the others acquired from NCBI RefSeq. Included in the packaging ATPase tree were viruses from the *Tectiviridae*, *Corticoviridae* and *Turriviridae* families, and *Sphaerolipoviridae*; the pVOGs (VOG4814, VOG0337) for this gene did not include the *Corticoviridae* or *Turriviridae*, which were acquired from NCBI RefSeq. Included in the pDNApol tree were viruses in the *Tectiviridae* and *Ampullaviridae* families, and Salterprovirus (VOG0334). All sequences, including those of the *Autolykiviridae* virus representatives, were clustered using usearch (usearch -cluster_fast query.fasta -sort length -id 0.9 -centroids nr.fasta -uc clusters.uc) and representative members of each cluster were selected for consistency across gene trees where possible[67]. All alignments and phylogenetic trees were constructed using the alignment, curation and maximum likelihood tree-building pipeline workflow referred to as eggNOG41 in the ETE v3.0.0b36 tree-building tool[68], implementing Clustal Omega[61], Muscle[69], MAFFT v5[70], M-Coffee[71], trimAI[72] and PhyML 3.0[64], and executed as: ete3 build -a my_sequences.fasta -w eggnog41 -o results/[73].

Characterization of the host range of autolykiviruses. Host ranges of the autolykiviruses and tailed viruses were characterized using drop-spot assays, and a host panel that included all hosts of isolation of the purified viruses. Viruses were applied to agar-overlays of host lawns as triplicate randomized-position spots in 150-mm Petri dishes using 96-spot blotters (BelArt, Bel-blotter 96-tip replicator, 378760002). Activity was monitored for all spots on days 1, 2, 3, 7, 14, 21 and 30 by marking boundaries of clearings on the Petri dish. At the termination of the experiment, all positives were called, blind to corresponding replicates, and sizes of clearings at each time point were recorded. Potential for cross-contamination was assessed by visual inspection and considered in final conservative manual curation of 'positive' infection calls. As a result, some cases with 3/3 positive replicates were discarded due to high probability of cross-contamination and some cases with 2/3 positive replicates were included when, for example, these were the only positives on a test plate.

The infection dataset, which was curated as described above, including only viruses that infected their host of isolation again in the host range assay and derived from independent plaques in the original isolation, included 247 viruses (Fig. 2). For statistical comparisons of infections of autolykiviruses and tailed viruses, only the 241 sequenced viruses were included. Four autolykiviruses were excluded from infection analyses, because they either represent genomically identical sublineages of a member included in the analyses (1.107.A, 1.107.B and 1.249.B) or because they did not infect their original host of isolation in the large-scale host range assay (1.095.O).

The kruskal.test implementation of the Kruskal–Wallis test in R (v3.3.0) was used to test for differences between autolykiviruses and tailed viruses in number of hosts (two-sided Kruskal–Wallis rank-sum test; $\chi^2 = 38.9724$, d.f. = 1, $P = 4.298 \times 10^{-10}$; $n = 17$ and 224, median = 34 and 2, for *Autolykiviridae* and tailed viruses, respectively) and number of host species (two-sided Kruskal–Wallis rank-sum test; $\chi^2 = 94.9497$, d.f. = 1, $P < 2.2 \times 10^{-16}$; $n = 17$ and 224, median = 4 and 1, for autolykiviruses and tailed viruses, respectively); for the test of the number of host species, assignments were based on the species defined in Fig. 2. For comparisons of average genome identity of hosts infected by the autolykiviruses and the tailed viruses, only infections between fully sequenced bacteria and viruses with >1 host were included (two-sided Kruskal–Wallis rank-sum test; $\chi^2 = 26.1429$, d.f. = 1, $P = 3.171 \times 10^{-7}$; $n = 16$ and 106, median = 93.04% and 99.97%, for autolykiviruses and tailed viruses, respectively). Evaluation of the time to detection of plaques in the host range assay also showed that, on average, the autolykiviruses required three times longer than tailed viruses to become detectable in culture ($n = 498$ infections by 17 autolykiviruses, $n = 844$ infections by 224 tailed viruses; median = 3 days and 1 day, for autolykiviruses and tailed viruses, respectively; two-sided Kruskal–Wallis rank-sum test; $\chi^2 = 374.7938$, d.f. = 1, $P < 2.2 \times 10^{-16}$; Fig. 3d and Extended Data Fig. 6).

In order to visualize the infection network with reference to the host phylogeny, we used iTOL[74] to generate an inverted circular representation of the host phylogeny and combined this with a Gephi-based ordered infection network representation generated using custom scripts in R and the packages igraph and rgexf[75,76]. In Gephi, all nodes were connected to a dummy hub node at the centre of the network, host nodes were fixed to the periphery and ordered to match the iTOL tree, and virus nodes were connected to the hosts that they were able to infect. The Force Atlas 2 layout was used to adjust the position of the virus nodes in the network.

Quantification of the significance of host sharing. For each pair of viruses, X and Y, that share at least one host, the significance of the overlap in host range was calculated as follows. Assuming that Y infects K hosts out of a population of N hosts, and X infects n randomly selected hosts, the probability that X and Y will coinfect k or more hosts is given by $P = f(k; N, K, n)$, where f is the probability mass function of the hypergeometric distribution. We set k, N, K and n to their empirically observed values and take the negative log of P as the significance of coinfection between X and Y.

Characterization of active DJR prophages in *Vibrio*. DJR prophages were isolated and sequenced from *V. kanaloae* (5S-149; contig_10: 28913-43245) and *V. cyclitrophicus* (10N.286.55.C7; contig_73: 31709-46046) as follows: 1 ml of overnight host culture grown in 2216MB was inoculated into a 2-l baffle-flask containing 1 l of fresh 2216MB. Cultures were grown with shaking at room temperature for seven days to allow for natural induction. Cells were removed using centrifugation (spun in sterilized 1-l polypropylene canisters at 5,000*g* for 15 min at 20 °C using a JLA-8.1000 rotor in a Beckman Coulter Avanti J-20 XP centrifuge) followed by filtration of the supernatant through a 0.2-μm vacuum filter (Corning 1,000 ml sterile Vacuum Filter/Storage Bottle System, 0.22-μm PES Membrane). Cell-free 0.2-μm filtrate was concentrated using PEG precipitation, as follows: 10% w/v of PEG 8000 (Sigma–Aldrich) was added to 700 ml of the filtrate at 0.6 M NaCl, solution was incubated with shaking at room temperature until PEG was visibly dissolved (3 h), incubated overnight at 4 °C, after which the solution was centrifuged at 8,000*g* for just under 4 h at 20 °C. The pellet was then collected with a sterile transfer pipette, resuspended in a final volume of 4 ml 0.02-μm-filtered ASW (ASW, 40 g l^{-1} Sigma Sea Salt solution prepared in sterile water) and stored at 4 °C. A total of 0.7 ml of the PEG-concentrated sample was purified using iodixanol-based density ultracentrifugation (density gradient 20–54% iodixanol (OptiPrep) in ASW, centrifuged in a Beckman L8M centrifuge in an SW41 rotor for 10 h at 20 °C at 35,000 r.p.m.). Gradients were unloaded as 26 fractions using a Biocomp Piston Gradient Fractionator (BioComp Instruments). Densities for each fraction were determined as mass per volume using a standard laboratory pipette[41]. Aliquots of each fraction were DNase-treated in 50-μl reaction volumes with 1× TURBO DNase buffer and 1 μl TURBO DNase and incubated at 25 °C overnight, followed by addition of fresh TURBO DNase (1 μl) and further incubation at 25 °C for 2.5 h. DNase treatment was validated using gel electrophoresis of treated and untreated genomic DNA controls in comparable iodixanol solutions. DNA extractions were carried out as follows: 0.02-μm-filtered ASW was added to reach a final volume of 100 μl; nuclease activity was halted by addition of 1/10 final volume of hot SDS mix, incubated at 75 °C for 10 min, then at 65 °C for 20 min; proteins were degraded by addition of 1 μl proteinase K per 100 μl of reaction volume and incubated at 65 °C for 20 min; DNA was recovered by addition of 1:1 ratio of Agencourt AMPure XP beads (Beckman Coulter) with standard ethanol washes and elution in 20 μl 0.2-μm-filtered Elution Buffer (EB, Qiagen). Density fractions for sequencing were selected on the basis of a PCR assay using major capsid protein-specific primers for each element: extracted DNA from fraction 12 (density = 1.19 for 5S-149 and 1.18 for 10N.286.55.C7) exhibited the brightest PCR band, suggesting the highest prophage concentration.

Final DNA extraction concentrations were quantified using a NanoDrop (5S-149 fraction 12 = 92.5 ng μl^{-1} and 10N.286.55.C7 fraction 12 = 24.9 ng μl^{-1}). Major capsid gene-specific primers were ordered from IDT, with sequences as follows:

5S-149_MCP_F2, 5′-ACAGTTCACACAAGCGGGTC-3′; 5S-149_MCP_R2, 5′-AGTTCGCTGTGATAACGCCTA-3′; 10N.286.55.C7_MCP_F2, 5′-TCTTTTACGGGGACGGGCTA-3′, 10N.286.55.C7_MCP_R2, 5′-CGCATATCTTCAAGCGCACG-3′.

Sample libraries were prepared for sequencing using the same tagmentation-based approach used for the bacterial genomes and ultimately multiplexed along with bacterial genomes on a single Illumina HiSeq lane. Sequenced reads were quality trimmed and mapped back to the reference genome of each lysogen to identify the full prophage region using CLC Genomics Workbench v8.5.1. *De novo* assemblies of reads also assembled the entire prophage into a single contig, which revealed the circular topology of the excised elements.

Metagenome preparation. An environmental sample was collected for metagenome preparation on 26 October 2014 (ordinal day 299) at Nahant, Massachusetts, USA. Eight replicate 4-l samples were collected and pre-filtered using 0.2-μm Sterivex filters; the filtrate was iron-chloride flocculated, collected on 0.2-μm Isopore polycarbonate filters (Millipore, GTTP09030) and resuspended in 4 ml oxalate solution, as described in ref. 38. For metagenome preparation, 1-ml subsamples from each of the eight replicates were pooled and PEG-concentrated (mixed: 8 ml pooled replicate subsamples, 0.8 g PEG, 8 ml 0.02-μm-filtered 1 M NaCl; dissolved at room temperature for 2.75 h; incubated overnight at 4 °C; centrifuged at 8,000*g* for 40 min at room temperature; the supernatant was then removed and the pellet resuspended in 600-μl 0.02-μm-filtered ASW; incubated at 4 °C); the sample contained abundant white precipitate. Virus activity in pre- and post-concentration samples was compared using agar-overlay plating and plaque counts with the indicator host 10N.261.45.B10 to assess potential losses due to precipitation and recoveries were found to be 79% ($n = 3$). Nuclease activity was confirmed in samples diluted 1:1 with 0.02-μm ASW.

A metagenome (14N.299.NahantUnfrac) was prepared from the concentrated sample as follows. To remove unencapsidated nucleic acids, the concentrated sample was pelleted to remove precipitates, a 100-μl subsample was removed and diluted 1:1 with 0.02-μm-filtered ASW, supplemented with 2 μl Turbo DNase and 2 μl RNase and incubated for 45 min at room temperature, pelleted to remove additional precipitates, and supplemented with an additional 2 μl Turbo DNase and incubated for an additional 85 min. Next, to inactivate nucleases, the sample was supplemented with 0.5 M EDTA to a final concentration of 15 mM EDTA and incubated at 75 °C for 20 min. The sample was then extracted using the MasterPure DNA extraction kit (EpiCentre MPRK092) with proteinase K following the manufacturers' recommended protocol, with the exception of including an extended overnight ethanol precipitation. PicoGreen quantitation showed a final concentration of 75.1 ng μl^{-1} in 20 μl, representing an original volume of 1,333 ml of seawater. Sequencing libraries were prepared using the Nextera Tagmentation approach as previously described[33], with an input concentration of 2 ng. Libraries were sequenced on a full NextSeq lane with 76 by 76 paired-end reads, at the MIT BioMicro Center.

A low buoyant density metagenome (14N.299.NahantLF) was prepared from the pooled replicates by density fractionating the PEG-concentrated virus sample and pooling subsamples of three low buoyant density fractions for extraction, as follows. First, 350 μl of PEG-concentrated viruses (equivalent to 4,666 ml of original seawater) was loaded onto an iodixanol (OptiPrep) density step-gradient (20–54% iodixanol in ASW buffer), and centrifuged in a Beckman L8M centrifuge with an SW41 rotor for 10 h at 20 °C at 35,000 r.p.m. (this procedure yielded precipitates upon addition of the sample to the density gradient). Then, gradients were unloaded as 26 fractions using a Biocomp Piston Gradient Fractionator (BioComp Instruments Inc.). Densities for each fraction were determined as mass per volume using a standard laboratory pipette[41]. Density fractions for metagenome preparation were conservatively selected on the basis of activity on a host infected by most autolykiviruses in the collection, 10N.261.45.B10 (fractions 9, 10, 11, with densities of 1.15, 1.16, 1.17 g ml^{-1}, respectively), these size fractions were conservatively selected and are known to exclude some members of the *Autolykiviridae* (Fig. 3b) as well as members of the excising DJR prophages of *Vibrio* (see Characterization of active DJR prophages in *Vibrio*). Selected iodixanol fractions were pooled (975.8 μl), nuclease-treated (1× TURBO DNase buffer, 2 μl TURBO DNase per 100 μl final volume, 1 μl RNase A per 100 μl final volume), incubated at 25 °C for 3.25 h in 50-μl reaction volumes, after which the nuclease activity was halted by addition of 1/10 final volume of hot SDS mix and incubation at 75 °C for 10 min, 65 °C for 20 min. The sample was then treated with proteinase K with addition of 1 μl

per 100 μl of reaction volume, incubated at 65 °C for 20 min and the DNA was recovered by addition of 0.5 volumes of Agencourt AMPure XP beads (Beckman Coulter) with standard ethanol washes and elution in 20 μl PCR-grade water. The 14N.299.NahantLF extract contained 8 ng μl^{-1} DNA as determined by fluorescence. Sequencing libraries were prepared as previously described[33], using 12 replicate reactions that each had 1.13 ng input DNA, with the following modifications: input DNA extract was enriched for larger fragments with a 0.6× bead-based size selection, extension time in the second PCR in the protocol was increased to 60 s, and bead-based size selection was used to enrich for ~615-bp-length fragments following pooling of all 12 reactions. Libraries were sequenced on a full Illumina MiSeq lane with 250 × 250 paired-end reads, at the MIT MicroBio Center.

Reads for both the 14N.299.NahantUnfrac and the 14N.299.NahantLF were prepared as follows. Quality-trimmed paired and unpaired reads were assembled using the clc_assembler command (v4.4.2.133896) in the CLC Assembly Cell (CLC bio) with default parameters. Open reading frames were called with Prodigal v2.6.1 using the -p meta flag and otherwise default parameters. This protocol yielded 239,907 and 642,418 total genes for the 14N.299.NahantUnfrac and 14N.299.NahantLF metagenomes, respectively. Accession numbers for both metagenomes associated with this study are provided in Supplementary Data 3 and are included under NCBI BioProject PRJNA328102.

Identifying additional diverse bacterial and archaeal virus DJR capsid sequences. In order to evaluate DJR viruses in metagenomes, we first generated a reference panel of diverse bacterial and archaeal DJR virus capsid sequences that could then be used in metagenomic searches. To achieve this, we combined manual and iterative hidden Markov model (HMM)-based sequence searches of public databases, with structural and phylogenetic analyses of 'hit' sequences to generate a high-confidence, extensively curated and diverse bacterial and archaeal DJR virus capsid reference sequence set.

Our searches were initialized with a seed set of 24 DJR reference sequences, including four autolykiviruses, one corticovirus, ten corticovirus-like putative prophages[27], one Gram-negative-infecting tectivirus, three Gram-positive-infecting tectiviruses, two turriviruses, the two excising *Vibrio* prophages described here, and one *Vibrio* plasmid identified here as a DJR element. Jackhmmer[77] (https://www.ebi.ac.uk/Tools/hmmer/search/jackhmmer) searches against UniProt[78] were used to generate HMMs for further searches, as well as to identify additional diverse DJR candidates, as revealed in the taxonomy view. We manually curated each round of HMM building and stopped the iterative search before eukaryotic viral proteins, primarily phycodnaviruses, were included in the HMM. This step was taken to skew our search towards bacterial and archaeal representatives of the DJR capsids. A subset of 12 HMMs was next used to search against NCBI bacterial (21,476) and archaeal (772) genomes (GenBank[79] Genomes, May 2017, ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/). These HMMs included: diverse representatives of the seed set (one each of the *Autolykiviridae*, the *Corticoviridae*, corticovirus-like putative prophages, Gram-positive-infecting viruses of the *Tectiviridae*, Gram-negative-infecting viruses of the *Tectiviridae* and the *Turriviridae*), additional recovered sequences confirmed to be virus capsid-like DJRs by curation with Phyre2[53] and the MPI Bioinformatics Toolkit[55] implementation of HHpred[54] (one each from genomes of *Magnetospirillum*, Opitutaceae, *Sulfobacillus*, *Nitrososphaera* and *Alcanivorax*), and one eukaryotic Chlorella virus DJR. Protein sequences for all downloaded microbial genomes were generated using Prodigal with the -p meta flag and otherwise default parameters and searches performed using the hmmsearch[34] tool (hmmer v3.1b2). These searches yielded 818 combined total unique hits, which were reduced to 196 by automatic screening to first require: (1) a size of 200–400 amino acid residues, the expected bacterial/viral DJR capsid size; (2) no hits to repeat domains. Next, manual trimming was applied to remove proteins with other functional domain annotations and the remaining sequences were then curated for a DJR-capsid-like structure, as described above.

This starting dataset enabled us to identify additional sequences in groups of particular interest, such as the alphaproteobacteria and other bacterial viruses, using manual blastp[80] searches against the GenBank non-redundant protein database[79]. All additional hits identified manually were curated using Phyre2[53] and HHpred[54,55] to identify sequences related to DJR protein structures from the Protein Data Bank (PDB)[81]. The sequences that were retrieved represent diverse phyla of archaea and bacteria, including Euryarchaeota, Crenarchaeota, Thaumarchaeota, Proteobacteria (alpha, beta, delta, epsilon and gamma representatives), Acidobacteria, Actinobacteria, Chloroflexi, Firmicutes, Lindowbacteria, Nitrospirae, Planctomycetes, Verrucomicrobia and Zixibacteria. One additional phage sequence was also identified from an unpublished *Rhodococcus* phage, and was described as a tectivirus although identified as a siphovirus by the NCBI taxonomy identifier. These putative DJR capsid proteins, plus the seed set of DJR bacterial and phage capsid proteins, a total of 179 unique sequences (Supplementary Table 2; marked as 'Reference' in Extended Data Fig. 8a), comprise our reference set and were next used to search ten bacterial and viral metagenomes.

Identifying potential DJR capsid proteins in metagenomes. Using our expanded reference set of sequences, we took a two-pronged approach to identify DJR capsid proteins in metagenomes. All proteins in each of ten metagenomes representing marine bacterial and viral fractions from environmental samples were analysed as follows (Extended Data Table 1). First, we ran jackhmmer (hmmer v3.1b2)[34] with default parameters for each sequence in each metagenome and extracted hits with a full-sequence score >20. This analysis identifies sequences that are closely related to each of our individual DJR proteins. Second, we built a HMM out of the DJR reference protein sequence alignment using hmmbuild and then used hmmsearch to screen all proteins in each metagenome iteratively for five iterations and extracted hits with a full-sequence score >20. This second analysis potentially identifies more distantly related DJR sequences. These approaches together yielded 43,734 total potential DJR sequences.

Identifying relationships between potential DJR capsid proteins. We next wanted to identify clusters of proteins that might represent novel environmentally relevant groups of DJR capsid-containing elements associated with bacteria and archaea. We therefore combined a series of annotation and curation approaches to focus on proteins with strong support for associations with either bacterial or archaeal hosts.

First, we screened all sequences using the NCBI Batch Web Conserved Domain[56] search tool (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) with default parameters[56,82]. We retained only metagenomic sequences with either no hits to any conserved domains, or hits to known DJR capsid superfamilies, specifically: Capsid_N (cl25189), Capsid_NCLDV (cl04526) and Phage_Capsid_P3 (cl20087). Next, we used psiblast[80] to compare each sequence to the PDB[81] protein structure database with an *e* value cut-off off 1×10^{-4} and retained only metagenomic sequences that either had no hits to any structures, or hits to known DJR virus capsids. The DJR PDB IDs used were: 1hx6, 2vvf, 1m3y, 2bbd, 5j7o, 3sam, 3j31, icjd, 4il7, 1m4x and 1j5q. Together, these screens narrowed our set to 25,874 potential DJR sequences.

To increase confidence, we next clustered all proteins and curated these clusters on the basis of both confirmed structural similarity to DJR capsids and sequence similarity to known viruses, as follows. First, we performed an all-by-all BLASTp search with a bitscore cut-off of 50 or better and clustered all proteins using MCL with unweighted BLAST matches and inflation value of 1.5 (Supplementary Table 2). We next annotated all proteins by whether they could be identified as a DJR through the Conserved Domain search, psiblast, Phyre2[53] or HHpred[54,55] (only performed for a small subset of sequences). We then screened the around 26,000 sequences against the NCBI RefSeq Viral database (https://www.ncbi.nlm.nih.gov/genome/viruses/) and annotated all sequences with a best bitscore ≥50 to a *Caudovirales* virus sequence as spurious. Combining these annotations, we identified all clusters for which the number of sequences annotated as DJR was greater than those annotated as spurious and retained only these clusters for additional curation. Next, all retained clusters were evaluated for evidence of false positives as identified by Phyre2[53] structural similarity searches, with a requirement for length >250 amino acids, 95% confidence identification, and 75% alignment coverage, and any clusters with >5% of sequences with false positives were discarded. These additional curations together yielded a total of 14,666 passing proteins, which were retained for network visualization (Extended Data Fig. 8) along with two additional protein sequences that were among our references and structurally confirmed as DJR sequences (GenBank accessions: AOI82551.1 and WP_060243308.1) but were captured in an MCL cluster that was discarded due to abundant hits to sequences with *Caudovirales* virus taxonomy identifiers. Notably, although these sequences had very high confidence assignments to DJR major capsid proteins by both Phyre2 (PDB hits for both sequences to corticovirus PM2 capsid 2w0c; 100% confidence, 25–26% identity, 93% alignment coverage) and HHpred (PDB hits for both sequences to corticovirus PM2 major capsid protein 2vvf, 100% probability, $e < 2 \times 10^{-47}$, target coverage 97%), they both had BLASTp bitscores of 45.8 against large proteins in tailed cyanophage (GenBank accessions: YP_007675165.1 and YP_009325074.1).

To ensure that proteins selected for subsequent phylogenetic analyses (Fig. 4) were strongly supported as being associated with viruses of bacterial and archaeal hosts, we next evaluated protein clusters on the basis of similarity to known DJR sequences and structures. All DJR hits identified by Conserved Domain search, psiblast, Phyre2 or HHpred were classified as either eukaryotic or bacterial and archaeal. Clusters with structurally annotatable sequences were dominated by either bacterial- and archaeal- or eukaryotic-associated virus DJR assignments, therefore, if the sum of bacterial- and archaeal-associated DJR hits (PDB identifiers: 1cjd, 1gw7, 1gw8, 1hb5, 1hb7, 1hb9, 1hqn, 1hx6, 1w8x, 2bbd, 2vvf, 2w0c and 3j31; Conserved Domain identifier: Phage_Capsid_P3 superfamily) was greater than the number of eukaryote- or virophage-associated hits (PDB identifiers: 1j5q, 1m3y, 1m4x, 3j26, 3kk5, 3sam, 4il7 and 5j7o; Conserved Domain identifier: Capsid_N superfamily, Capsid_NCLDV superfamily) then the cluster was classified as

bacterial- and archaeal-associated (42 clusters; 788 proteins); if the reverse was true the cluster was classified as eukaryotic-associated (32 clusters; 12,998 proteins); and if there were no identified matches to any DJRs, the cluster was classified as unknown (474 clusters, including singletons; 882 proteins). The network diagram (Extended Data Fig. 8) was generated using the Python package NetworkX and visualized using Gephi v0.9.1. The network structure was generated using the ForceAtlas 2 force directed layout method, with the option to prevent node overlap.

All clusters identified as bacterial- and archaeal-associated were then further curated to identify clusters for which, despite inclusion of some members with hits to bacterial and archaeal virus DJRs, there was a prevalence of sequences with no matches to DJR structures despite both Phyre2 and HHpred annotation. Finally, all proteins passing these filters were required to be unique and at least 200 amino acids in length for inclusion in the final DJR major capsid tree; this yielded a final set of 442 proteins (Fig. 4 and Supplementary Data 2, with ten procedural duplicate sequences identified in 'Notes' column).

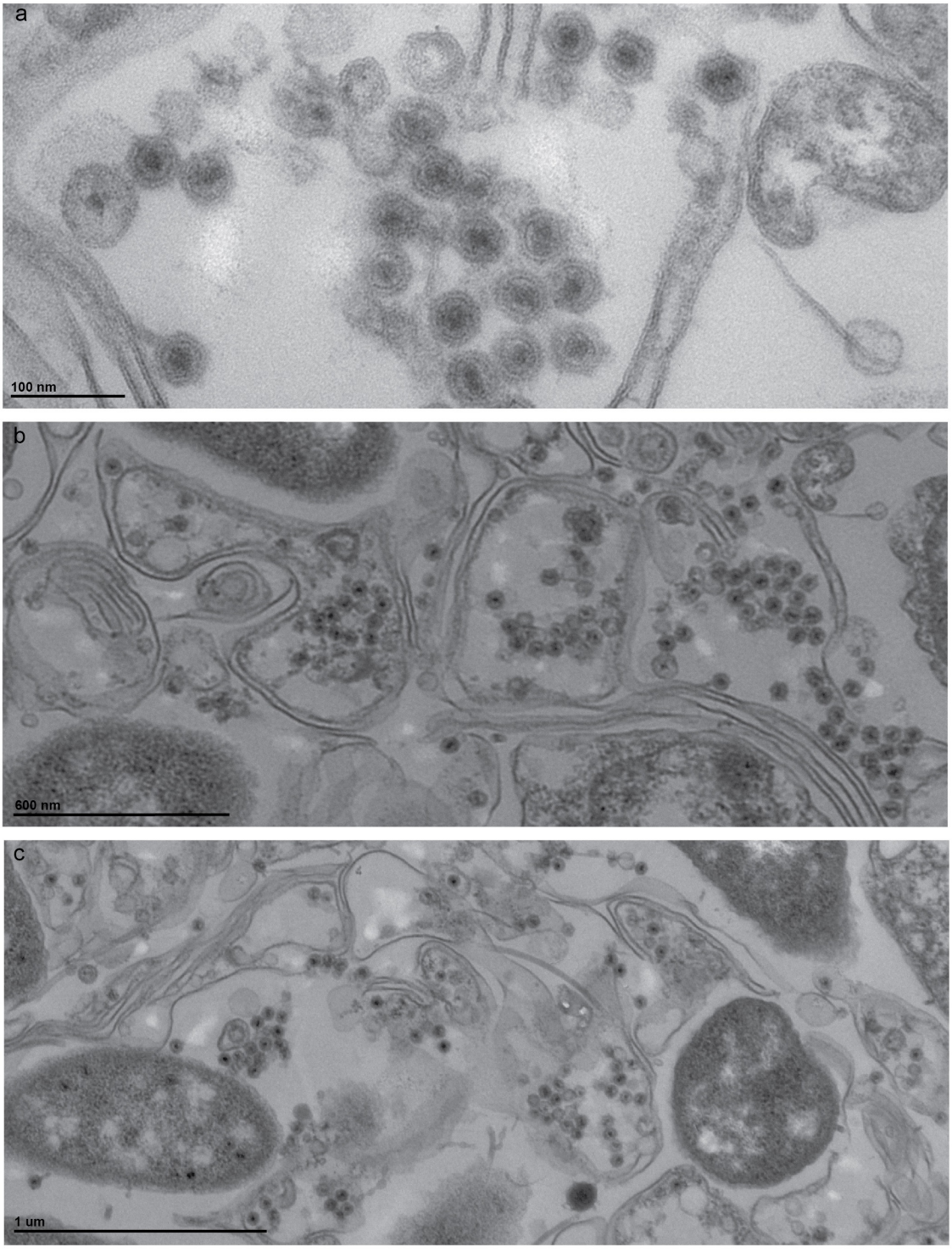
To build the phylogenetic tree from these 442 sequences, we executed the ETE toolkit[68] eggNOG41 phylogenetic workflow, as described above. The eggNOG41 gene tree workflow is used to construct trees in the EggNOG orthology database[73] and is therefore appropriate to construct a tree for related but very diverse sequences, as we have with our DJR protein set. In brief, this workflow incorporates comparison of several multiple alignment tools, an alignment trimming step that removes columns with >10% gaps, and protein model selection before constructing a tree in PhyML. In PhyML, the workflow optimizes the topology, the branch lengths and rate parameters (transition/transversion ratio, proportion of invariant sites, gamma distribution parameter). Equilibrium amino-acid frequencies are estimated using frequencies defined by the substitution model (in this case, the JTT model), four substitution rate categories and aLRT branch supports are used to construct the final tree. The tree was visualized in iTOL, collapsed on the basis of average internal branch length of 2.0, and exported for figure preparation in Adobe Illustrator. To provide an overview of the genomic neighbourhoods of the putative DJR capsid proteins in our phylogenetic tree, we identified a representative virus, genome-neighbourhood or metagenomic contig for each of the 29 major branches or clades (Fig. 4b and Extended Data Table 2) and annotated these (Fig. 4b, Supplementary Data 1) as described above for the autolykiviruses.

Code availability. All custom codes associated with this work are available from the authors upon request.

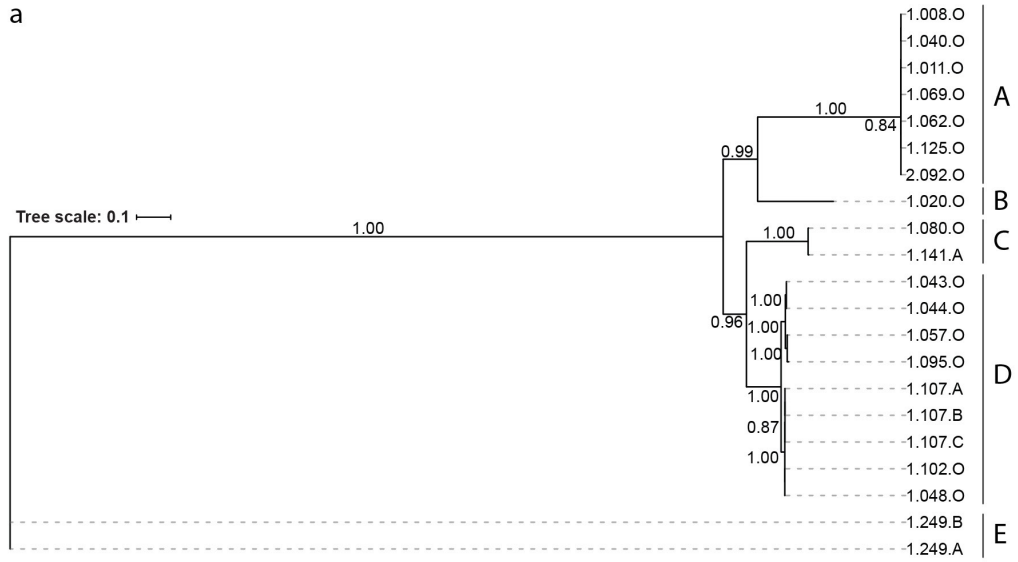
Data availability. Annotation information for the autolykiviruses and elements shown with genome diagrams is provided in Supplementary Data 1. Accession numbers, taxonomy and annotation of DJR capsid proteins included in the trees and network are provided in Supplementary Data 2. GenBank accession numbers for newly obtained sequences and previously published genomes included in Fig. 2 and Extended Data Fig. 6 are provided in Supplementary Data 3, with all new sequences associated with this work included under the Nahant Collection of NCBI BioProject with accession number PRJNA328102. Metagenomes used in this study are listed with citations in Extended Data Table 1 and include: *Tara* Oceans, viromes, ftp://ftp.imicrobe.us/projects/197/TOV_43_all_contigs_predicted_proteins.faa.gz; *Tara* Oceans, ocean microbiome reference gene catalogue, ftp://ftp.sra.ebi.ac.uk/vol1/ERA412/ERA412970/tab/OM-RGC_seq.release.tsv.gz; methane seep sediment, BioProject accession PRJNA290197; Rifle sediment, BioProject accession PRJNA288027; Mediterranean Sea virome, GenBank accessions, AP013358–AP014505; Mediterranean Sea metagenome, GenBank accessions, GU942957:GU943153; Chesapeake Bay virome, Sequence Read Archive accession, SRR4293227; NCBI environmental metagenomes, ftp://ftp.ncbi.nlm.nih.gov/blast/db/env_nr*.tar.gz; and two metagenomes generated in this study, Nahant light fraction viral metagenome, (deposited at GenBank under accession PDMW00000000; the version described here is PDMW01000000); and Nahant Viral Metagenome (deposited at GenBank under accession PDMX00000000, the version described here is PDMX01000000). All other data are available from the authors upon reasonable request.

32. Hunt, D. E. *et al.* Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320,** 1081–1085 (2008).
33. Baym, M. *et al.* Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS ONE* **10,** e0128036 (2015).
34. Eddy, S. R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7,** e1002195 (2011).
35. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
36. Hehemann, J.-H. *et al.* Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. *Nat. Commun.* **7,** 12860 (2016).
37. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–D285 (2016).
38. John, S. G. *et al.* A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.* **3,** 195–202 (2011).
39. Rasband, W. S. *ImageJ* (U.S. National Institutes of Health, 1997).
40. Silbert, J. A., Salditt, M. & Franklin, R. M. Structure and synthesis of a lipid-containing bacteriophage. 3. Purification of bacteriophage PM2 and some structural studies on the virion. *Virology* **39,** 666–681 (1969).
41. Lawrence, J. E. & Steward, G. F. in *Manual of Aquatic Viral Ecology* (eds Wilhelm, S. W., Weinbauer, M. G. & Suttle, C. A.) 166–181 (ASLO, 2010).
42. Kivelä, H. M., Männistö, R. H., Kalkkinen, N. & Bamford, D. H. Purification and protein composition of PM2, the first lipid-containing bacterial virus to be isolated. *Virology* **262,** 364–374 (1999).
43. Biller, S. J. *et al.* Bacterial vesicles in marine ecosystems. *Science* **343,** 183–186 (2014).
44. Hurwitz, B. L., Deng, L., Poulos, B. T. & Sullivan, M. B. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.* **15,** 1428–1440 (2013).
45. Henn, M. R. *et al.* Analysis of high-throughput sequencing and annotation strategies for phage genomes. *PLoS ONE* **5,** e9083 (2010).
46. Bates, D. M. lme4: Mixed-effects modeling with R. http://lme4.0.r-forge.r-project.org/lMMwR/lrgprt.pdf (2010).
47. Bates, D. & Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67,** 1–48 (2015).
48. R Core Team. R: A Language and Environment for Statistical Computing. http://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, Austria, 2016).
49. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
50. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348,** 1261359 (2015).
51. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532,** 465–470 (2016).
52. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).
53. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10,** 845–858 (2015).
54. Hildebrand, A., Remmert, M., Biegert, A. & Söding, J. Fast and accurate automatic structure prediction with HHpred. *Proteins* **77,** 128–132 (2009).
55. Alva, V., Nam, S.-Z., Söding, J. & Lupas, A. N. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res.* **44,** W410–W415 (2016).
56. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45,** D200–D203 (2017).
57. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34,** 2115–2122 (2017).
58. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30,** 1236–1240 (2014).
59. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8,** 785–786 (2011).
60. Guy, L., Kultima, J. R. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26,** 2334–2335 (2010).
61. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).
62. Li, W. *et al.* The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.* **43,** W580–W584 (2015).
63. McWilliam, H. *et al.* Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res.* **41,** W597–W600 (2013).
64. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).
65. Lefort, V., Longueville, J.-E. & Gascuel, O. SMS: smart model selection in PhyML. *Mol. Biol. Evol.* **34,** 2422–2424 (2017).
66. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* **45,** D491–D498 (2017).
67. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
68. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33,** 1635–1638 (2016).
69. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
70. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33,** 511–518 (2005).
71. Wallace, I. M., O'Sullivan, O., Higgins, D. G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **34,** 1692–1699 (2006).
72. 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).
73. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44** (D1), D286–D293 (2016).

74. Letunic, I. & Bork, P. Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23,** 127–128 (2007).
75. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJ.* **Complex Systems,** 1695 (2006).
76. Vega Yon, G., Fabrega Lacoa, J. & Kunst, J. B. rgexf: Build, Import, and Export GEXF Graph Files. https://cran.r-project.org/web/packages/rgexf/index.html (2015).
77. Finn, R. D. *et al.* HMMER web server: 2015 update. *Nucleic Acids Res.* **43,** W30–W38 (2015).
78. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).
79. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41,** D36–D42 (2013).
80. 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).
81. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
82. Marchler-Bauer, A. *et al.* CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* **39,** D225–D229 (2011).
83. Adriaenssens, E. & Brister, J. R. How to name and classify your phage: an informal guide. *Viruses* **9,** 70 (2017).
84. Clerissi, C. *et al.* Unveiling of the diversity of prasinoviruses (*Phycodnaviridae*) in marine samples by using high-throughput sequencing analyses of PCR-amplified DNA polymerase and major capsid protein genes. *Appl. Environ. Microbiol.* **80,** 3150–3160 (2014).
85. Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348,** 1261498 (2015).
86. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541,** 353–358 (2017).
87. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7,** 13219 (2016).
88. Anantharaman, K. *et al.* Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum. *PeerJ* **4,** e1607 (2016).
89. Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E. & Ghai, R. Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9,** e1003987 (2013).
90. Ghai, R. *et al.* Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J.* **4,** 1154–1166 (2010).
91. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44,** D67–D72 (2016).



Extended Data Figure 1 | Members of the *Autolykiviridae* are non-tailed viruses that may form tail tubes on contact with cells. Thin-section electron microscopy of an agar overlay containing plaques of representative *Autolykiviridae* virus 1.008.O (see Methods for experimental details). **a**, Virus particles in contact with cell membranes are observed to occasionally possess tail-tube-like structures, whereas those not in contact with cells do not. **b**, Lower magnification of same field of view as **a** shows that tail-tube-free virions are more common than those with tail tubes. **c**, Lower magnification view of virion in Fig. 1b also shows that the presence of the tail tube is associated with cell contact and is not observed in nearby virions.

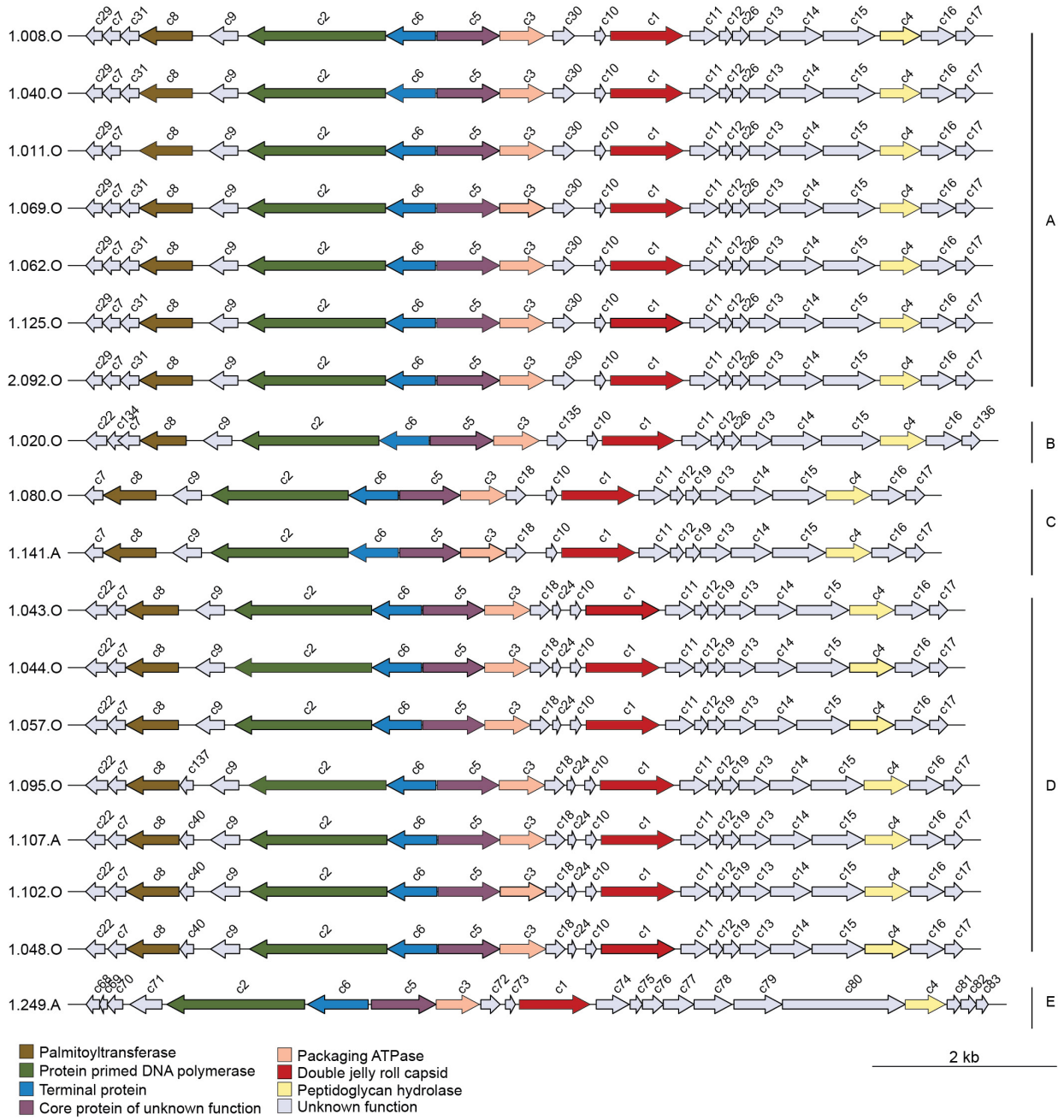


b

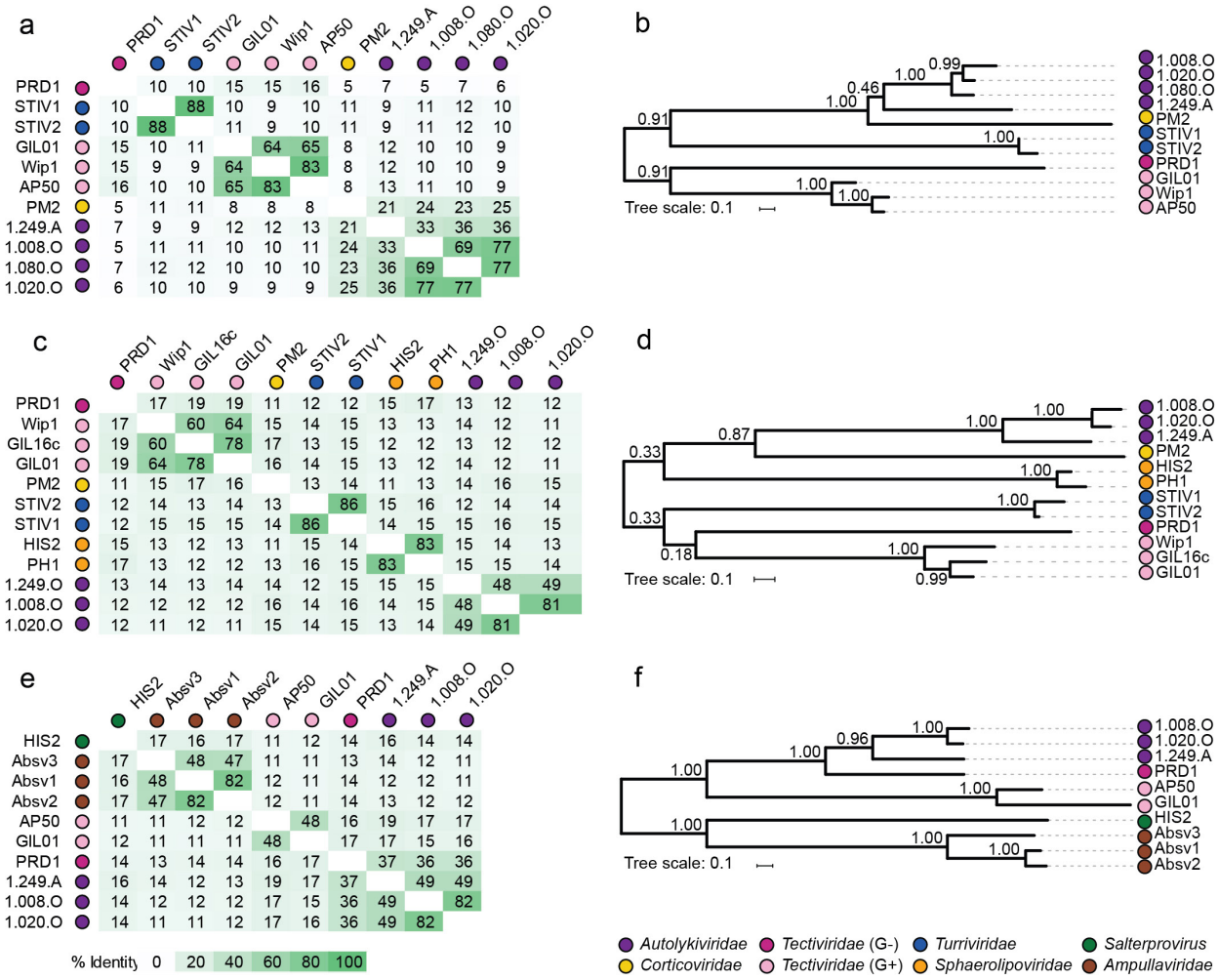
	1.008.O	1.040.O	1.011.O	1.069.O	1.062.O	1.125.O	2.092.O	1.020.O	1.080.O	1.141.A	1.043.O	1.044.O	1.057.O	1.095.O	1.102.O	1.107.A	1.107.B	1.107.C	1.048.O	1.249.A	1.249.B
1.008.O		100	99.9	100	100	99.9	99.9	60	59	59	60	60	60	60	61	61	61	61	61	32	32
1.040.O	100		99.9	100	100	99.9	99.9	60	59	59	60	60	60	60	61	61	61	61	61	32	32
1.011.O	99.9	99.9		99.9	99.9	99.8	99.9	60	59	59	60	60	60	60	61	61	61	61	61	32	32
1.069.O	100	100	99.9		100	99.9	99.9	60	59	59	60	60	60	60	61	61	61	61	61	32	32
1.062.O	100	100	99.9	100		99.9	100	60	59	59	60	60	60	60	61	61	61	61	61	32	32
1.125.O	99.9	99.9	99.8	99.9	99.9		99.9	60	59	59	60	60	60	60	61	61	61	61	61	32	32
2.092.O	99.9	99.9	99.9	99.9	100	99.9		60	59	59	60	60	60	60	61	61	61	61	61	32	32
1.020.O	60	60	60	60	60	60	60		63	63	65	65	65	65	65	65	65	65	65	33	33
1.080.O	59	59	59	59	59	59	59	63		99.6	77	77	78	77	77	77	77	77	77	32	32
1.141.A	59	59	59	59	59	59	59	63	99.6		77	77	78	77	77	77	77	77	77	32	32
1.043.O	60	60	60	60	60	60	60	65	77	77		100	99	97	96	96	96	96	96	32	32
1.044.O	60	60	60	60	60	60	60	65	77	77	100		99	97	96	96	96	96	96	32	32
1.057.O	60	60	60	60	60	60	60	65	78	78	99	99		98	96	96	96	96	96	32	32
1.095.O	60	60	60	60	60	60	60	65	77	77	97	97	98		97	97	97	97	97	31	31
1.102.O	61	61	61	61	61	61	61	65	77	77	96	96	96	97		100	100	100	100	31	31
1.107.A	61	61	61	61	61	61	61	65	77	77	96	96	96	97	100		100	100	100	31	31
1.107.B	61	61	61	61	61	61	61	65	77	77	96	96	96	97	100	100		100	100	31	31
1.107.C	61	61	61	61	61	61	61	65	77	77	96	96	96	97	100	100	100		100	31	31
1.048.O	61	61	61	61	61	61	61	65	77	77	96	96	96	97	100	100	100	100		31	31
1.249.A	32	32	32	32	32	32	32	33	32	32	32	32	32	31	31	31	31	31	31		100
1.249.B	32	32	32	32	32	32	32	33	32	32	32	32	32	31	31	31	31	31	31	100	

% Identity 0 20 40 60 80 100

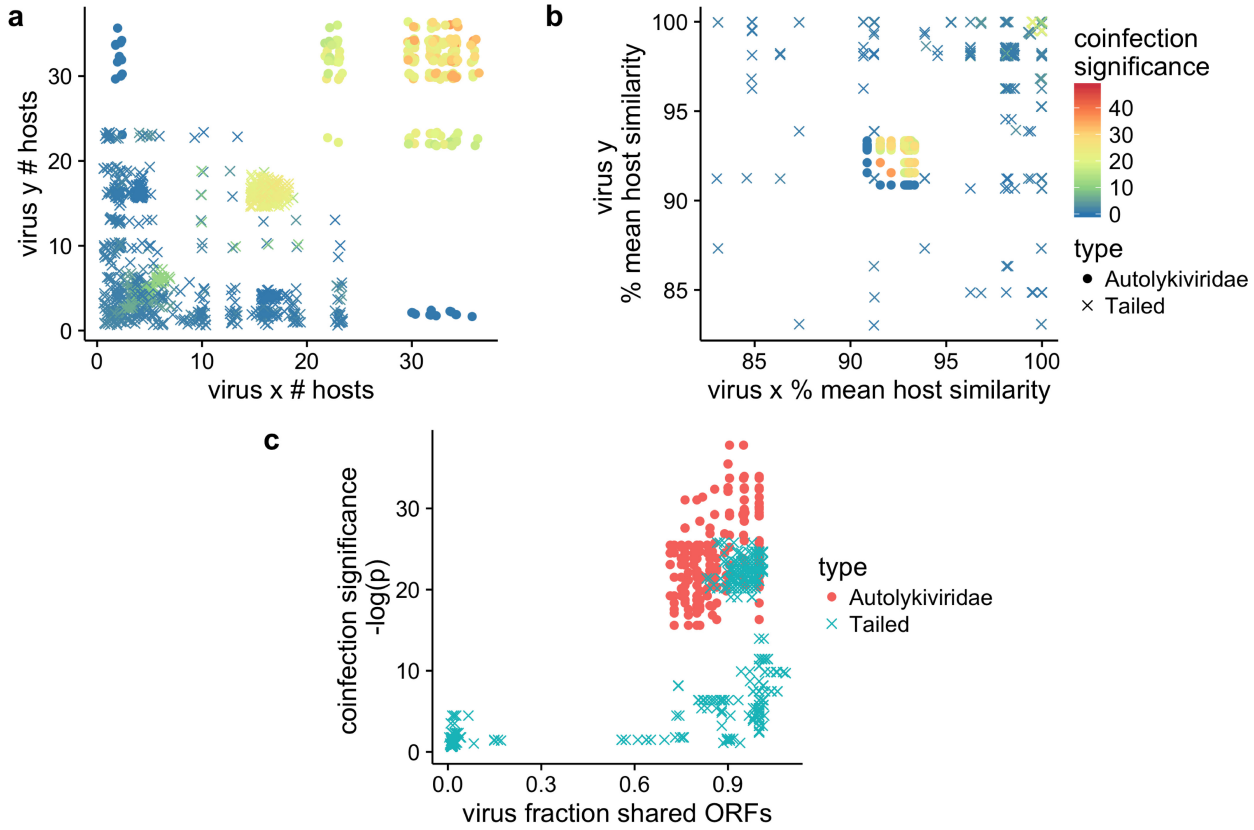
Extended Data Figure 2 | Whole-genome alignments show that the family *Autolykiviridae* consists of five major sequence diversity clusters. a, Maximum Likelihood phylogeny of whole-genome nucleotide alignments of 21 autolykiviruses. Alignments were made with Clustal Omega and the phylogenetic tree was generated with PhyML-SMS with aLRT branch supports. Scale bar, substitutions per base. **b**, Percentage of whole-genome nucleotide identities among 21 autolykivirus genomes on the basis of the Clustal Omega alignment. Assumptions of 50% and 95% identity for genus and species classifications[83], respectively, suggest that these viruses represent two genera (groups A, B, C, D and group E) and five species. Two viruses with identical genomes were isolated at time points 39 days apart (1.048.O and 1.102.O), viruses with the same number and different letter suffixes represent lineages derived from a single plaque that gave rise to variable morphotypes during serial purification.



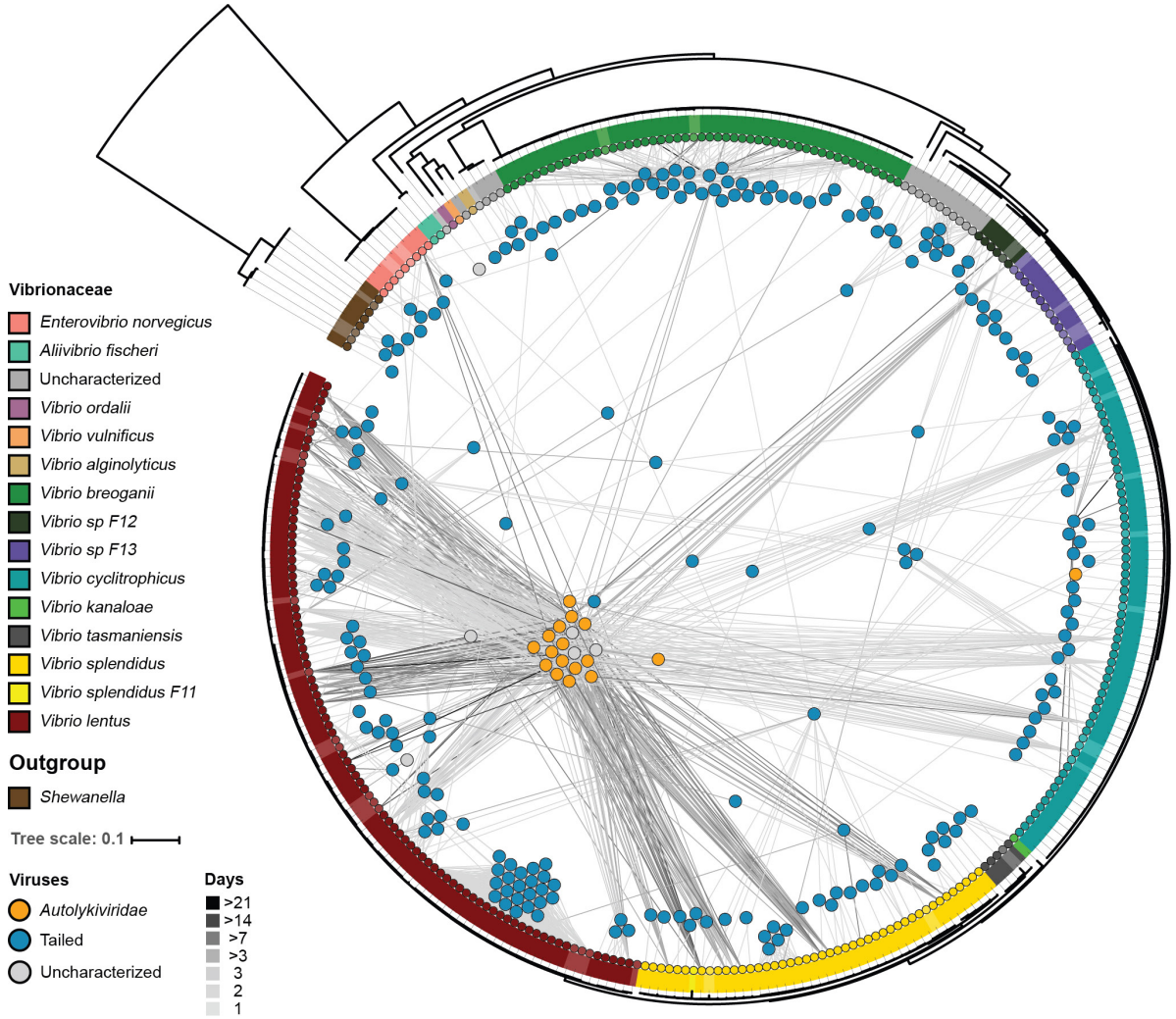
Extended Data Figure 3 | Genomes of members of the *Autolykiviridae* are syntenic despite extensive diversity at the nucleotide level. Virus genomes are grouped by nucleotide similarity (as identified in Extended Data Fig. 2). Homologous proteins were identified by performing an all-by-all BLASTp, requiring a minimum bitscore of 50, and clustering all pairs unweighted, using MCL with an inflation parameter set to 1.4 (Methods), cluster membership is identified by the label over the block arrows in the genome diagram. Protein clustering reveals that in addition to the six proteins identifiable by sequence similarity as core to all characterized autolykiviruses, additional protein clusters are shared among various subsets of the identified viral genome groups. For example, in the region of the genome to the right of the major capsid protein, 17 out of 18 viruses (genome groups A, B, C and D) share a set of seven protein clusters of unknown function (c11, c12, c13, c14, c15, c16 and c17); among these viruses, two additional proteins are shared only within subsets of the genomes (c26 in genome groups A and B; c19 in genome groups C and D).



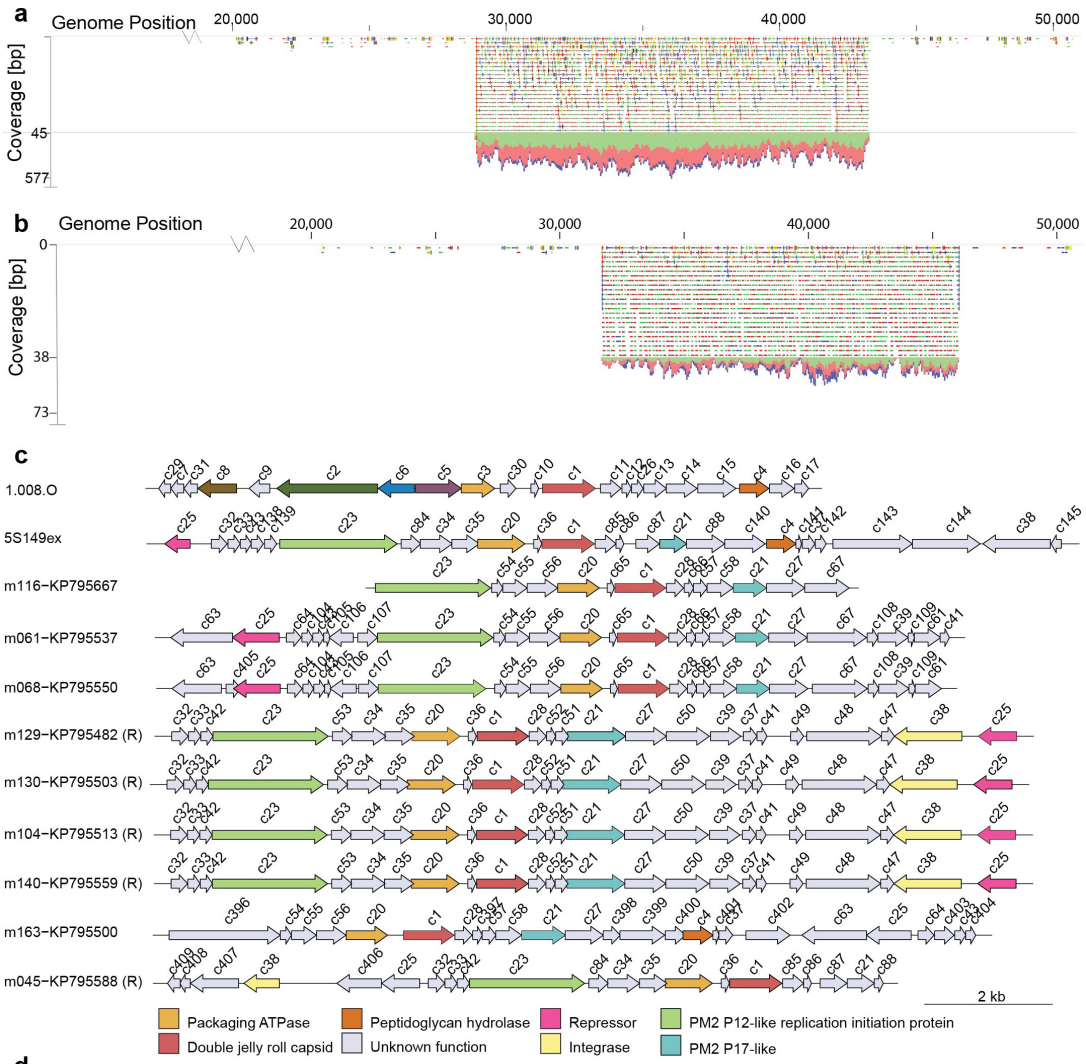
Extended Data Figure 4 | Packaging and replication protein-sequence phylogenies of autolykiviruses are incongruent with respect to other known families of non-tailed dsDNA viruses. Autolykiviruses are most similar to the corticovirus PM2 in their major capsid protein, poorly resolved in their packaging ATPase, and most similar to the tectiviruses in their protein-primed DNA polymerase. Pairwise identities and phylogenies of the protein sequences of the DJR major capsid protein (**a** and **b**), packaging ATPase (**c** and **d**) and protein primed DNA polymerase (**e** and **f**). Members of the *Tectiviridae* infecting Gram-positive and Gram-negative hosts are shown separately as G+ and G−, respectively. All alignments were performed using the ETE3 Toolkit with workflow eggNOG41. All trees are maximum-likelihood trees with aLRT branch supports.



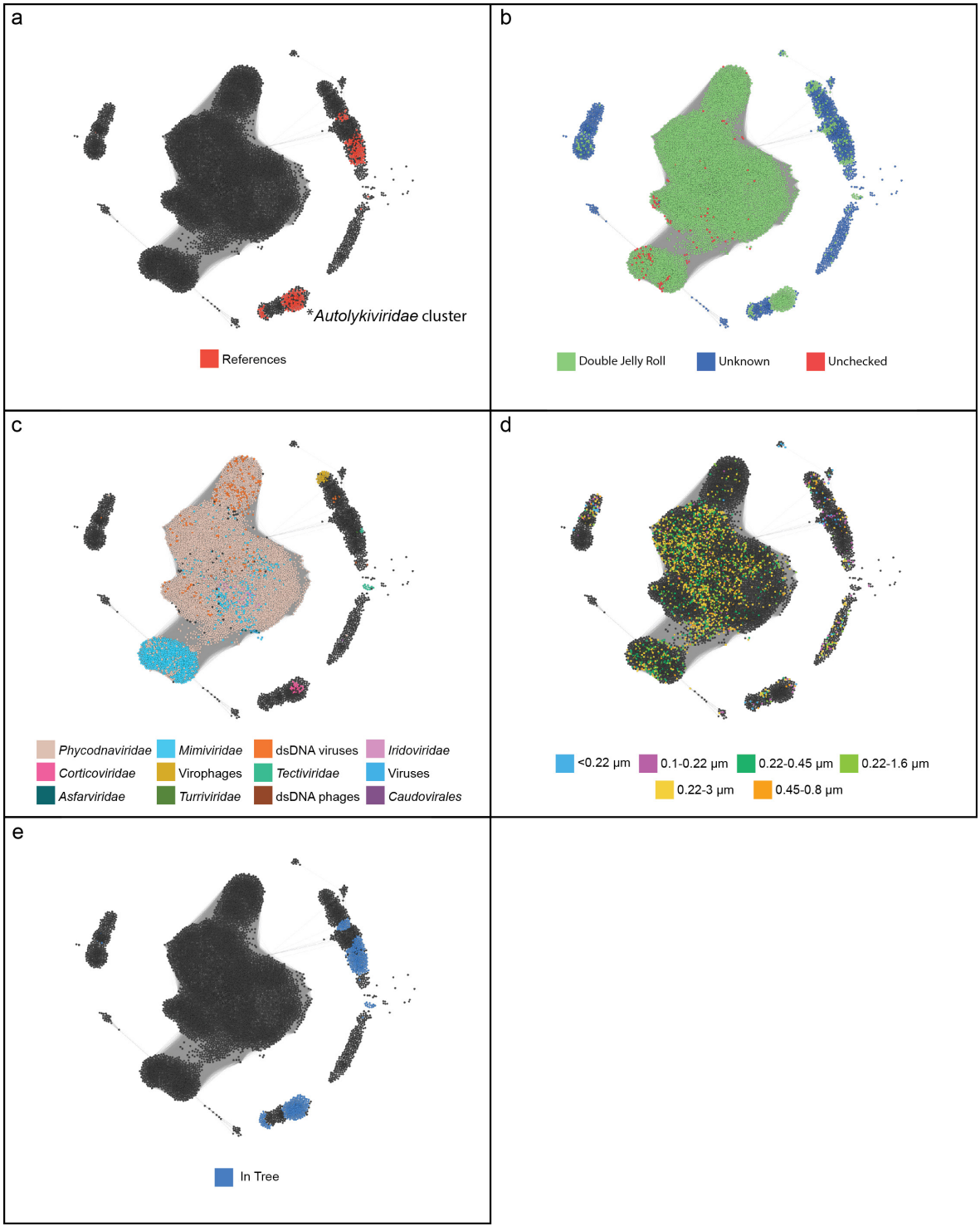
Extended Data Figure 5 | Sequence-diverse autolykiviruses share extensively overlapping host ranges that include diverse hosts. **a**, Pairwise coinfection significance by host count. Autolykiviruses exhibit highly significant host sharing. **b**, Pairwise coinfection significance compared to mean pairwise genomic similarity of the host. Autolykiviruses exhibit more significant host sharing than tailed phages of comparable host diversity. **a**, **b**, Coinfection significance as defined in Methods. **c**, Pairwise coinfection significance compared to viral genomic similarity measured as a fraction of shared open reading frames (ORFs). Autolykiviruses exhibit more significant host sharing than tailed viruses of comparable genomic similarity. A total of 998 reciprocal pairs of tailed viruses and 236 reciprocal pairs of autolykiviruses are shown, representing all pairs of viruses within each group (141 unique tailed, 16 unique autolykiviruses) that share at least one host.



Extended Data Figure 6 | Autolykiviruses show delayed host lysis compared with other viruses. Inverted phylogenetic tree showing the relationships among all 318 assayed bacterial strains on the basis of the concatenated alignments of the *hsp60* and ribosomal protein genes, and using a partitioned model in RaxML to allow placement of 40 strains for which only the *hsp60* gene sequence was available (Methods). Isolates are generally non-clonal. Leaves represent Vibrionaceae isolates and are coloured by population (Methods). Nodes represent viruses and are coloured by morphotype, as defined by major capsid protein or genome composition (Methods; non-tailed in orange, tailed in blue, unsequenced viruses in grey); edges represent infections with intensity increasing with increased time required for observation of plaques. Whereas 94% of tailed virus infections were detected within three days in host range assays, only 57% of autolykivirus infections were detected in that time, with 15% requiring more than seven days to be detected.



Extended Data Figure 7 | DJR elements in Vibrionaceae include naturally excising integrated prophages and broad host-range plasmids. Prophages of representative group 5 DJR elements (Fig. 4) naturally excise from their *Vibrio* hosts during growth in culture. Sequencing of nuclease-treated cell-free culture supernatants reveals sharply delineated regions of high coverage read mapping with respect to host genome background, indicating the presence of extracellular nuclease-protected prophage DNA. **a**, *V. kanaloae* 5S-149 DJR prophage. **b**, *Vibrio* 10N.286.55.C7 DJR prophage. **c**, Genome diagrams of the excising 5S-149 DJR prophage and the nine Vibrionaceae plasmids[28] that are identified here as DJR elements show that they are syntenic and all share the DJR capsid protein, packaging ATPase and the corticovirus PM2 P17-like protein. MCL clustering of proteins on the basis of the BLASTp sequence similarity reveals that additional proteins, including integrases, repressors, peptidoglycan hydrolases and replication initiation genes, are common but not universal within these elements. **d**, Pairwise percentage of whole-genome nucleotide identities between 5S-149 DJR prophage and the DJR Vibrionaceae plasmids show that these elements are highly diverse at the nucleotide level and that 100% nucleotide-identical 13.6-kb plasmids are found in hosts in multiple species.



Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Network of DJR virus capsids identified in bacterial and archaeal genomes and marine metagenomes. Iterative HMM-based searches of marine metagenomes, on the basis of a reference panel of autolykiviruses and previously identified DJR capsid bacterial and archaeal viruses, yield approximately 15,000 proteins following stringent quality control filtering of the initial approximately 45,000 sequences that were recovered. Network visualization reflects MCL clustering of BLASTp-based similarities among sequences. **a**, Placement of reference panel sequences within the network. **b**, Characterization of proteins as DJRs on the basis of sequence- and structural-similarity-based annotation. **c**, Best BLASTp matches to RefSeq viruses, bitscore requirement of 50. **d**, Association of *Tara* Oceans-derived sequences to size fraction of isolation. **e**, Subset of sequences selected for phylogenetic analyses (Fig. 4) on the basis of membership in protein clusters strongly supported as bacterial and archaeal virus DJR capsids and requiring a length of $\geq$200 amino acids (Methods). We note that this selection is conservative, given the greater number and diversity of sequences recovered by our HMM-based search that passed all quality controls and show no structural- or sequence-based similarity to any other proteins, and thus were excluded from further analyses. The observed dominance of eukaryotic virus DJR capsids in this search is predicted to reflect four major aspects of our approach. First, inclusion of cellular metagenomes allows capture of large viruses such as the *Mimiviridae* (>400 nm), *Iridoviridae* (120–350 nm) and *Phycodnaviridae* (100–220 nm). Second, some *Phycodnaviridae* have been shown to encode up to eight sequence-diverse copies of their DJR major capsid gene[84]. Third, <0.22 μm viral metagenomes are biased against recovery of bacterial and archaeal DJR viruses, as described here. And fourth, the sequence content of HMMs using iterative searches is defined by the search space, such that if eukaryotic virus DJR capsid sequences are well represented, as they are in the larger size-fraction sequence databases used here, they will drive searches towards increased detection of similar sequences.

Extended Data Table 1 | Metagenomes used in this study

Metagenomic dataset	Data Source
Tara Oceans Viromes[85]	ftp://ftp.imicrobe.us/projects/197/TOV_43_all_contigs_predicted_proteins.faa.gz
Tara Oceans OM-Reference Gene Catalog[50]	ftp://ftp.sra.ebi.ac.uk/vol1/ERA412/ERA412970/tab/OM-RGC_seq.release.tsv.gz
Methane Seep Sediment	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA290197
Rifle Sediment[86-88]	https://www.ncbi.nlm.nih.gov/bioproject/?term=288027
Mediterranean Sea Virome[89]	GenBank Accessions: AP013358:AP014505
Mediterranean Sea Metagenome[90]	GenBank Accessions: GU942957:GU943153
Chesapeake Bay Virome	https://www.ncbi.nlm.nih.gov/sra/SRR4293227/
NCBI environmental metagenomes[91]	ftp://ftp.ncbi.nlm.nih.gov/blast/db/env_nr*.tar.gz
Nahant Light Fraction Viral Metagenome	This paper, GenBank Accession: PDMW00000000
Nahant Viral Metagenome	This paper, GenBank Accession: PDMX00000000

Description and data sources[50,85–91] for each of the metagenomic datasets used in this study.

Extended Data Table 2 | Contigs of DJR elements

Group	Type	Host phylum	Accession Number	Element or Contig Name
1	*Autolykiviridae*	Proteobacteria (c: gamma)	In Supplementary Data File 3	1.008.O
2	*Corticoviridae*	Proteobacteria (c: gamma)	NC_000867.1	PM2
3	Host-associated	Zixibacteria	MEWP01000034.1 (R)	Zixibacteria bacterium RBG_16_53_22 RBG_16_scaffold_15471
4	Metagenomic Contig		LAZR01011096.1	LCGC14_contig011102
5	Host-associated	Proteobacteria (c: gamma)	AJYX02000003.1: 28913-43245	Excised *Vibrio kanaloae* 5S-149 DJR prophage
6	Metagenomic Contig		CEUD01119479.1	TARA_137_MES_0.22-3_scaffold209516_1
7	Host-associated	Proteobacteria (c: alpha)	JPUR01000104.1 (R)	Marinosulfonomonas sp. PRT-SC04 contig_12486
8	Metagenomic Contig		newLF_contig_7959 (R)	Nahant-LF_contig7959
9	Metagenomic Contig		LF_contig_41867	Nahant-LF_contig41867
10	Metagenomic Contig		LAZR01031772.1	LCGC14_contig031806
11	Metagenomic Contig		CEPX01414959.1	TARA_037_MES_0.1-0.22_C20701301_1
12	Metagenomic Contig		CEPX01063969.1	TARA_037_MES_0.1-0.22_scaffold154415_1
13	Metagenomic Contig		LAZR01001928.1	LCGC14_contig001928
14	Metagenomic Contig		LAZR01007575.1	LCGC14_contig007576
15	*Turriviridae*	Crenarchaeota	NC_005892.1	STIV1
16	*Tectiviridae* (G-)	Firmicutes	NC_011523.1	AP50
17	Host-associated	Actinobacteria	CP006261.1	*Streptomyces collinus* Tu 365 plasmid pSCO2
18	Host-associated	Acidobacterium	MEKI01000026.1	Acidobacteria bacterium RBG_13_68_16 RBG_13_scaffold_1666
20	*Tectiviridae* (G-)	Proteobacteria (c: gamma)	NC_001421.2	PRD1
21	Metagenomic Contig		LAZR01015278.1 (R)	LCGC14_contig015288
22	Metagenomic Contig		CEPX01030804.1 (R)	TARA_037_MES_0.1-0.22_scaffold77974_1
23	Metagenomic Contig		CBAY_603174 (R)	Chesapeake Bay Virome contig_603174
24	Metagenomic Contig		CBAY_52461	Chesapeake Bay Virome contig_52461
25	Host-associated	Thaumarchaeota	CP007174.1: 2808000-2828000	*Nitrososphaera evergladensis* SR1
26	Host-associated	Thaumarchaeota	CP002408.1: 258549-274228	*Nitrososphaera gargensis* Ga9.2
27	Host-associated	Crenarchaeota	CP003317.1	*Pyrobaculum oguniense* TE7 extrachromosomal element
28	Metagenomic Contig		lcl_contig_23053 (R)	Chesapeake Bay Virome contig_23053
29	Metagenomic Contig		CEPX01198169.1 (R)	TARA_037_MES_0.1-0.22_scaffold345758_1

Source information for contigs of DJR group representatives presented in Fig. 4b. Additional notations in 'Accession number' columns include: (1) coordinate information if a contig represents an extraction from a larger sequence; (2) an R if the contig is presented in the reverse orientation with respect to annotations provided in Supplementary Table 1.

nature research

Corresponding author(s): Martin Polz, Libusha Kelly

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Experiments comparing Autolykiviridae and tailed viruses included either all available members meeting quality control requirements (Infection assay presented in Fig. 2, Fig. 3D, and Extended Data Fig. 6) as limited by recovery during initial sampling protocol (described in the the Methods in the section "Isolation, culturing, and sequencing of bacteria and viruses"), or one selected representative (Chloroform Assay, Fig. 3A; Density, Fig. 3B; Protease Assay, Fig. 3C) of each of the major diversity groups of the Autolykiviridae and the three tail morphotypes for the Caudovirales. No statistical methods were used to predetermine sample sizes.

2. Data exclusions

Describe any data exclusions.

The criteria for inclusion & exclusion of viruses and hosts in the presented infection analyses are described in the methods section on pages 25-26, in the section headed "Characterization of Autolykiviridae host range". The infection dataset presented in Fig. 2 and Extended Data Fig. 6 includes 247 viruses, excluded were viruses from the original dataset that did not infect their host of isolation again in the large scale host range assay or that did not derive from independent plaques in the original isolation. For statistical comparisons of infections of Autolykiviridae and tailed viruses, the 241 sequenced viruses were included, with four sequenced Autolykiviridae excluded from infection analyses because they represent either genomically-identical sublineages of a member included in the analyses (1.107.A, 1.107.B, and 1.249.B), or because they did not infect their original host of isolation in the large-scale host range assay (1.095.O).

3. Replication

Describe whether the experimental findings were reliably reproduced.

Large scale infection assay (Fig. 2, Extended Data Fig. 6) - The large scale infection assay included 3 replicates of each interaction; the entire experiment was performed once as described. Observations in subsequent smaller-scale host range assays with members of these collections have been consistent with those described here.

Chloroform assay (Fig. 3a) - The chloroform assay included 3 replicates of each interaction and the 5 Autolykiviridae included in these experiments are representative of the diversity of this group and represent biological replicates; the entire experiment was performed once as described in the manuscript. Observations in subsequent similar experiments are consistent with those described here.

Density gradient (Fig. 3b) - The density gradient determination was performed once as described in the manuscript, the 5 Autolykiviridae included in these experiments are representative of the diversity of this group and represent biological replicates. Observations in subsequent similar experiments are consistent with those described here.

Protease treatment (Fig. 3c) - The protease treatment comparisons were performed in three separate experiments, each with a single replicate of each virus and treatment, a representative gel from one of these experiments is shown; the 5

Autolykiviridae included in these experiments are representative of the diversity of this group and represent biological replicates.

Infection timing (Fig. 3d) - The large scale infection assay included 3 replicates of each interaction; the entire experiment was performed once as described. Observations in subsequent similar experiments are consistent with those described here.

Decay assay (inline in Main & Methods) - The decay assay included 4 replicates of each virus; the entire experiment was performed once as described.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

All experiments comparing Autolykiviridae and tailed viruses included either all available members meeting quality control requirements (Infection assay presented in Fig. 2, Fig. 3D, and Extended Data Fig. 6), or one selected representative (Chloroform Assay, Fig. 3A; Density, Fig. 3B; Protease Assay, Fig. 3C) of each of the major diversity groups of the Autolykiviridae and the three tail morphotypes of the Caudovirales. Sample order was haphazardly assigned for each of the three independent replicates of the protease assay (Fig. 3C), position in each of three plate sectors was haphazardly assigned for each of three virus lysate replicates in the large scale infection assay (Fig. 2, Extended Data Fig. 6), otherwise samples were not randomized for the experiments.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Large scale infection assay (Fig. 2, Extended Data Fig. 6) - Results of the host range assay were performed blinded insofar as 1) they were recorded without reference to position of the three haphazardly assigned replicates on a given assay plate, and 2) a large number of assays with different sets of viruses were recorded at the same time reducing likelihood of pattern detection. This is briefly indicated in the methods.
In no other experiments were investigators blinded to group allocation during data collection or analyses.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

Commercial and publicly available open-source software were used to analyze data in this study, these are indicated in their associated sections in the Methods, and include: BLASTp v.2.2.29+, CLC Assembly Cell v.4.4.2.133896, CLC Genomics Workbench v.8.5.1, Clustal Omega (EMBL-EBI web portal), EggNOG-Mapper v.4.5.1, ETE v.3.0.0b36 (implementing Clustal Omega, trimAl, MUSCLE, PhyML

v.3.0, MAFFT v5, M-Coffee, Gephi v.0.9.1, HHpred (MPI Bioinformatics Toolkit webportal), hmmer v.3.1b2, ImageJ, InterProScan v.5.17-56.0, iTOL v.4, MAFFT, MCL v.14.137, NCBI Batch Web Conserved Domain search tool, PhyML v.3.0 with SMS v.1.8.1, Phye2 webportal, Prodigal v.2.6.1 and v.2.6.3, Python with package NetworkX v.1.1.10, RAxML, R v.3.3.0 with packages GenoPlotR, data.table, ggplot2, cowplot, igraph, rgexf, lme4.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	Bacteria and virus strains described as isolated in this work are available from the authors upon request.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	The study did not involve human research participants.