DISSERTATION | DOCTORAL THESIS

Titel | Title

Diversity and Evolution of Mobile Genetic Elements in Vibrio

verfasst von | submitted by

Kerrin Steensen B.Sc. M.Sc.

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 794 685 437
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Biologie
Betreut von | Supervisor:	Univ.-Prof. Dr. Martin F. Polz

Acknowledgements

First, I would like to thank Martin Polz for being an outstanding mentor and teacher. I highly appreciated the countless hours of scientific discussion, your guidance on how to ask the right questions, and your support and kindness throughout. To this day, I have been intrigued by your idealistic (and sometimes artistic) way of doing science.

I would also like to thank my committee members Kathryn Kauffman, Thomas Boettcher, Matthias Horn, and Mart Krupovic for their advice and helpful discussions during my PhD. In addition to being on my committee, Kathryn: you have been an inspiration to me in many ways. Thank you for your support and for speaking up when you felt things needed to be said.

I am deeply thankful to my friends and colleagues from the Polz lab for their invaluable support and contributions to this thesis:
Nina Bartlau, for being such a lighthearted and reliable friend, and for preventing any homesickness with your Northern German attitude. Xiaoqian Annie Yu, for all the coffee walks, scientific discussions, and for making me feel connected when COVID hit. Anna Lopatina, for all the fun moments and for sharing your seemingly endless knowledge of phages. Ben Roller, for being the most supportive colleague in difficult situations - you are truly contributing to creating a fairer world. Cathrine Hellerschmied, for all the fun moments we shared. Annika Gomez, for all the insightful and fun conversations we had. Thank you to Michael Wolfram, Jacob Bobonis, Paul Damian Jimenez Nuno, Nir Shmuel Even, Edo Danilyan, Fabienne Kreutter, Katja Nafe, Sonja Frere, Sebastiano Muscia and Yun-Wei Lien for providing helpful feedback and creating a wonderful lab culture.

Thank you to those who have shared their knowledge, worked with me, or have collaborated on entire research projects: Joana Séneca, Shaul Pollak, Julie Chen, Marton Palatinszky, Shubham Joge, Lovro Trgovec-Greif, Jay Osvatic, Petra Pjevac, Otto Cordero, Eduard Fadeev, Katharina Sures, Rachel Szabo, and Craig Herbold. I am also grateful to my past mentors - Rolf Daniel, Dominik Schneider, Fatima Hussain, and Bernd Wemheuer - for shaping me as a scientist. Furthermore, I would like to thank the members of the Pollak group for the scientific discussions.

A huge thanks goes to the best doc school coordinators Nadja Holzleitner, Franziska Bauchinger, and Katharina Klaushofer, as well as to the heads Thomas Rattei and Stephan Krämer. I would like to thank the entire JMF team for scientific support, collaboration, and sequencing. Thank you also to the LiSC team, who provided an amazing computational infrastructure. Thank you to the admin team, the staff scientists, and to the TA team for making my life easier in many ways.

I would like to thank all the colleagues and friends I met at DoME/CeMess. Patrick, for your friendship, your critical thinking, the many enjoyable math and writing sessions (and the after-beers) and most importantly - for always providing a different point of view. Lukas, thank you for always cheering me up and for all the fun times we had together. Johanna, thanks for being such a good listener during all our enjoyable baking sessions. Sarah, for all the thoughtful and fun discussions. Thank you to Bela, Julia K., Geli, Franzi, Chiara, Sara, Julia R., Bettina, Nicole, Eva and Sean for many great moments. I would especially like to thank Laura, Dani, Christina, Anouk, and Flo for being the most considerate and wonderful office mates.

There are so many more people that I would like to mention that have made my PhD nicer every day. I had a great support network, and I am incredibly thankful to all of you who accompanied me along the way. Thank you also to my peers at the doc:muv program for providing such a valuable source of inspiration and mental support. I also want to thank all members of the Posaunenchor in Vienna - especially Jil, Johanna, Jan, and Anke - for being my main source of fun and distraction over the past years.

Thank you to Jonathan, for your love and support over the past few years, and for helping me navigate all the ups and downs of this journey.

Finally, I would like to thank my family for their unconditional love and support. My parents and bonus-parents Bärbel, Hans-Joachim, Birgit, and Reinhard, for creating homes filled with books, fun and love, and for being there when I needed you most. I want to specifically thank my mother for being the most inspiring role model a daughter could wish for, and my father for supporting every adventure I embarked on. Thank you, Henrik for being the best brother on Earth. Thanks to my grandparents, and specifically to my grandmother, Elisabeth, for teaching me math, strategic thinking, and how to appreciate the beauty of nature from an early age - you laid the foundation for my career as a natural scientist. I love you all so much.

Table of Contents

This thesis underwent language and grammar improvement with Chat-GPT-4-turbo following the AI policy of the University of Vienna.

"*Maybe you are not studying phages; maybe you are studying phage houses.*"

Kathryn Kauffman, Year One of my PhD.

Abstract

The exchange of mobile genetic elements (MGEs) drives bacterial evolution, as evidenced by the highly variable element content observed among closely related bacterial strains. There are different types of MGEs, such as viruses, transposons, or plasmids, each with its own mechanism of movement, accessory genes, and impact on bacterial hosts. Despite the ongoing discovery of novel MGEs, the full extent of their diversity and impact on the bacterial hosts remains largely unknown. In this dissertation, I explore the diversity and evolution of MGEs in a model system consisting of *Vibrio* bacteria sampled from the same coastal ocean environment. In the first chapter, I investigate the diversity and distribution of prophages, which are viral genomes integrated into their bacterial hosts' chromosome and are among the first MGEs discovered. Historically, research on prophages mainly concentrated on tailed bacteriophages – bacterial viruses that employ a classical head-and-tail morphology – which now constitute the majority of bacteriophages in public databases. However, newer evidence suggests that other prophage types also integrate into *Vibrio* genomes. By applying *de novo* predictions, which are unaffected by database biases favoring tailed bacteriophages, I reveal that tailless and filamentous prophages outnumber tailed prophages. Tailless and filamentous bacteriophage genomes are integrated into conserved genomic sites thereby preventing functional gene disruption, and they frequently encode accessory genes involved in defense against other bacteriophages. In the second chapter, I investigate the diversity of MGEs beyond prophages in the *Vibrio* model system. I observe a large number and diversity of MGEs, a total of 2,600 elements infecting 127 bacteria sampled from a single environment, indicating a key role of MGEs in the evolution of their bacterial hosts. Furthermore, I show that MGEs are the main contributors to the flexible gene pool of bacteria, which consists of genes that vary among individuals within a bacterial species. I demonstrate that the diversity of MGEs is mediated by a remarkable variety of MGE types and subtypes, as well as highly variable accessory genes within MGE types that were often encoded in specific locations between blocks of conserved genes or at the MGE's border. This dissertation highlights the large, vastly unexplored diversity and complex interactions of MGEs and their major importance for the genetic variation observed within bacterial species in nature.

Zusammenfassung

Der Austausch mobiler genetischer Elemente spielt eine zentrale Rolle in der bakteriellen Evolution, der sich besonders in der starken Variabilität solcher Elemente zwischen sehr nah verwandten bakteriellen Stämmen äußert. Verschiedene mobile genetische Elemente, wie beispielweise Prophagen, Transposons oder Plasmide, können sich untereinander in der Mobilität, im Gen-Pool und in ihren Auswirkungen auf den bakteriellen Wirt stark unterscheiden. Obwohl immer mehr neue mobile genetische Elemente beschrieben werden, ist ihre gesamte Diversität noch weitgehend unbekannt. Diese Dissertation befasst sich mit der Diversität und Evolution mobiler genetischer Elemente in einem Modellsystem bestehend aus Bakterien der Gattung *Vibrio*, die gemeinsam in einem küstennahen, marinen Ökosystem vorkommen. Das erste Kapitel behandelt die Diversität und die Verbreitung von Prophagen – Viren, die ihre DNA in das bakterielle Wirtsgenom integrieren können. In der Vergangenheit fokussierte sich die Prophagen-Forschung hauptsächlich auf sogenannte Schwanz-Viren. Inzwischen gibt es jedoch Hinweise darauf, dass auch andere Arten von Viren ihr Genom in *Vibrio* Bakterien integrieren können. Da Schwanz-Viren in den aktuellen Datenbanken vorherrschend sind, werden andere Viren in der computergestützten Analyse häufig übersehen. Deshalb wurde die Identifizierung der Prophagen in dieser Arbeit weitgehend unabhängig von Datenbankinformationen, also „de novo“, durchgeführt. Auf diese Weise konnte ich zeigen, dass Prophagen in unserem Modellsystem deutlich häufiger zu den schwanzlosen und filamentösen Viren gehören als zu den besser untersuchten Schwanz-Viren. Diese schwanzlosen und filamentösen Viren zeichnen sich durch wenige, spezifische Integrationsstellen im Genom aus, ohne die Funktion der betroffenen Stelle langfristig zu beeinträchtigen. Schwanzlose und filamentöse Viren enthalten häufig einen variablen Genabschnitt, der auch im Zusammenhang mit Virus Abwehrfunktionen steht. Im zweiten Kapitel der Dissertation erweitere ich die Suche nach mobilen genetischen Elementen über Prophagen hinaus, um deren Vielfalt systematisch zu erfassen. Das Ergebnis umfasste eine große Anzahl und Vielfalt von insgesamt 2.600 mobilen genetischen Elementen aus 127 Isolaten mariner Bakterien, was auf eine Schlüsselrolle mobiler genetischer Elemente in der Evolution von Bakterien in der Umwelt hinweist. Dabei konnte ich zeigen, dass mobile genetische Elemente einen Großteil des flexiblen Genpools ausmachen. Ein Teil der Diversität in mobilen genetischen Elementen ist auf hochvariable Abschnitte in ihrem Genom zurückzuführen, in dem nicht-essenzielle Gene oder kleinere mobile genetische Elemente kodiert sind. Diese Dissertation zeigt die enorme Anzahl und Vielfalt mobiler genetischer Elemente sowie ihre komplexen Interaktionen auf und belegt, dass sie maßgeblich zur Diversifizierung von Bakterien in der Natur beitragen.

Introduction

Because bacteria reproduce asexually, each offspring should be an exact copy of its parent. Their evolution was therefore long thought to have been mainly driven by the accumulation of single nucleotide mutations over time (1). However, research over the past decades has demonstrated that horizontal acquisition of genetic material is a major contributor to bacterial evolution (1,2). Bacteria quickly acquire different traits, such as antibiotic resistance or toxins (3,4), which can spread through bacterial populations. While large amounts of DNA can be gained in a single horizontal gene transfer event, this material can be lost just as rapidly. Understanding the evolutionary dynamics and the ecological implications of these genetic exchanges is a central aim in microbial biology.

A large fraction of horizontal gene transfer is mediated by mobile genetic elements (MGEs). These DNA segments can self-mobilize or rely on other MGEs to mobilize within or between genomes (5,6). There are many types of MGEs that carry different genes involved in the MGE's maintenance or additional functions. Because of the fast gain and loss of MGEs, even closely related bacteria can exhibit large genomic variation (7,8). The presence of individual MGEs can have profound consequences for the bacterial host, as MGEs significantly contribute to the host's gene pool and influence its metabolic and defensive repertoire (5,6). They can also impose fitness costs on the host, such as causing cell lysis to accomplish their own replication and spread (9).

The discovery of the first MGEs was based on observations of changes in host physiology that were not explainable by sporadic mutations in the host chromosome. To this day, the discoveries of bacteriophages (also phages), plasmids and transposons remain milestones in biology, as they have fundamentally shaped our current understanding of evolution (10–12). Although modern sequencing techniques now facilitate the computational search for MGEs (13,14), the large variety of MGE types and their rapid diversification complicate MGE classification and limit our understanding of their evolutionary history (6).

Rapid genomic changes mediated by MGEs, and their strong impact on their bacterial hosts make MGEs major players of bacterial ecology and evolution. Yet, the enormous diversity of MGEs confronts us with a multitude of both known and unknown element types whose abundances and complex interactions remain far from comprehensively described. Therefore, systematic investigation of the diversity of both well-known and previously overlooked MGEs in natural environments is key to better understanding their role in bacterial biology and evolution.

Bacteria as hosts of mobile genetic elements

Although nearly all organisms on Earth are thought to contain mobile elements (15), bacteria are particularly shaped by the horizontal acquisition of genetic material making them a model for studying horizontal gene transfer (also lateral gene transfer) (1,2). The full extent of the impact that horizontal gene transfers have on bacterial evolution only became evident when sequencing costs decreased, resulting in an increase in available genome sequences over the last decades (16). Bacteria within the same species were observed to exhibit higher variation in their genomes than expected, an insight that led to the concept of the bacterial pangenome (17): Genes shared by all bacteria within species boundaries were defined as core genome, while less frequent genes represented the flexible genome (17,18). This genomic variability within species is now predominantly attributed to horizontal gene transfers (19).

Mechanistically, horizontal gene transfer has been assigned to three different categories: transformation, conjugation or transduction. Transformation is the uptake of free DNA, while conjugation is the transfer of DNA relying on cell-to-cell contact, and transduction is the injection of DNA by bacteriophages (13). Conjugation and transduction are considered the primary mechanisms of horizontal gene transfer employed by MGEs. The host and its environment impact the rate at which MGEs are being transferred. The best-studied example is lysogenic phages, viruses that are able to

integrate their genomes into the bacterial host genome as so-called prophages. These viruses carry genes for viral particle assembly which are usually inactive during integration (20). Prophages can be activated (induced) upon specific cell programs such as the SOS response (21) or by quorum sensing (22) ultimately leading to host lysis and spread of phage particles (9). Similarly, conjugation has been shown to increase during inflammatory responses of the host (23). Whether transferred MGEs remain successfully incorporated also depends on the recipient: Bacteria have co-evolved with MGEs to prevent infection or reduce the costs of MGE carriage. For example, bacterial hosts need to carry specific receptors allowing for phage attachment to the cell surface (24). Bacteria can also gain immunity from phage infection by encoding phage defense systems (25). Furthermore, bacteria have been found to contain so-called hotspots of integration (26) - specific genomic sites that are frequently occupied by MGEs – thereby reducing the risk of the host's core genes being disrupted. Whether a successfully transferred MGE remains in the host's genome or is eliminated is always dependent on selection, which often favors the loss of inserted MGEs (27).

Mobility of mobile genetic elements

There are many different mobile genetic element types, and they employ diverse strategies for intra- and intercellular mobility (6). MGEs such as plasmids and integrative and conjugative elements (ICEs, formerly also known as "conjugative transposons") often encode the conjugation machinery needed for horizontal transfer themselves (28,29). Similarly, prophages carry several genes for the replication and packaging of their DNA into viral particles, which can be released and infect novel hosts. Other MGEs primarily rely on self-mobilizable MGEs to be transferred between organisms. For instance, phage satellites are small MGEs that hijack prophages by packaging their own DNA into the capsid of the activated prophage (30). Certain MGEs such as integrons might transfer along with conjugative plasmids or ICEs (29,31).

Once an MGE is inside the bacterial cell, it may remain episomal or insert into the genome through transposition or site-specific recombination (13). Transposition is mediated by transposase genes, which facilitate the movement or copying of MGEs (so-called transposons) into other genomic locations (32). Other MGEs encode site-specific integrases that mediate insertion into more frequently populated genomic hotspots (33). There are also many MGEs which do not encode the required mobility genes themselves. For example, short recognition sequences can be sufficient for an MGE to be mobilized by exploiting other recombinases present in the cell, such as terminal inverted repeats that are recognized by transposases (32).

Overall, the genetic mobility features encoded on MGEs dictate their mode and frequency of intra- and intercellular movement and lead to dependencies among MGE types and the host. Investigating these complex interactions is key to understanding the pace and circumstances under which MGE transfers take place.

Mobile genetic elements as reservoirs of genetic diversity

While many MGE types have a conserved set of genes and genetic signatures mediating their mobility and maintenance, MGEs can also carry accessory genes providing a fitness advantage to the host (34). For instance, increased horizontal gene transfer rates under stress conditions might help the host to adapt quickly to environmental changes (5,35). The accessory genes carried by an MGE are thought to be related to the MGE type (36). For example, conjugative plasmids and mobilizable integrons are well known to carry and spread antibiotic resistance genes (37,38). Many MGE types have recently been shown to protect against infecting phages by carrying phage defense genes (39–41). Furthermore, MGEs inserted into the genomes of pathogenic bacteria often carry virulence factors involved in disease development. One prominent example is the filamentous phage CTXphi encoding for the cholera toxin in *Vibrio cholerae* (3). However, in many cases the function of accessory genes encoded by MGEs remains unknown, as functional information in current databases is limited (42).

Bacterial hosts experience a tradeoff between the additional functional repertoire and the fitness costs conferred by MGEs. The maintenance of an MGE alone can decrease bacterial fitness (43), while other MGEs might disrupt relevant host genes during insertion (44), and prophages can even kill the host upon induction (9). Nonetheless, MGEs are very abundant in bacterial genomes, suggesting that their functional repertoire is either essential for survival, such as toxin-antitoxin systems that induce cell death upon loss, or provides a selective advantage under specific conditions, such as defense genes protecting against phage infection. However, which functions different MGE types provide and how these functions are acquired remain poorly understood.

Mobile genetic element identification

In the last century, the discovery of MGEs solely relied on experimental work. Advances in sequencing techniques offered the opportunity to identify MGEs through computational prediction (14). However, the range of highly diverse MGE types complicates the computational identification of MGEs. Certain MGE types, such as plasmids, transposons, prophages and integrons, have been discovered in the last century and encompass many characterized examples that facilitate their computational detection. While there are current efforts to identify different MGEs (14,44,45), many tools and databases are highly specialized for a single MGE type (46–48). Although many novel MGE types are being discovered (30,40,49–51), the number of experimentally characterized examples remains small. Thus, the computational identification of MGEs becomes limited because it largely relies on the availability of known sequences.

Database biases can further complicate the discovery of MGEs. If most characterized examples of an MGE originate from a specific subtype or a single bacterial clade, other variants of this MGE may go undetected. For example, the majority of publicly available prophage genomes belong to a diverse set of tailed prophages from the class *Caudoviricetes* (47). Although there are a number of studies reporting prophages derived from tailless *Vinavirales* and filamentous *Tubulavirales* phages (Figure 1) (52–54), their lower representation in the databases biases against the discovery of those underrepresented groups.

Another challenge using computational approaches is the assignment of the genomic boundaries between the host and its integrated MGE. As many identification tools rely on detecting marker genes, which are not necessarily located at the MGE's borders, predictions often suffer from incorrectly assigned MGE boundaries (55). However, uncertainty in assigned MGE boundaries can complicate the characterization of the predicted MGEs and hinder the assessment of the evolutionary history of nearby genes. The accurate prediction of MGE boundaries still requires additional experimental data (55) or comparative genomics, which is limited to closely related bacteria.

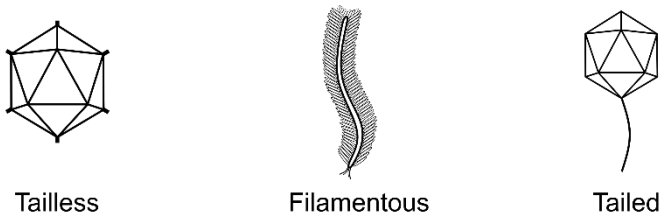

Figure 1 - Schematic Representation of Tailless, Filamentous and Tailed Bacteriophages.

Marine *Vibrionaceae* as a model system for mobile genetic elements

Bacteria of the family *Vibrionaceae* have served as a model system to study MGEs such as plasmids, bacteriophages and even phage satellites (30,54,56,57). *Vibrionaceae* representatives are easy to isolate, fast-growing, well suited for DNA extraction, and genetically tractable, making them a good system for the experimental testing of hypotheses in the lab. To study MGEs in their natural context, I

use one of the largest collections of bacteria from the family *Vibrionaceae*. The bacteria were isolated from coastal ocean water samples collected in the North American Atlantic during several projects performed in the Polz lab over the past 20 years (8,57,58). The collection includes over 900 sequenced isolates of marine *Vibrionaceae*. These isolates have been shown to form ecologically and genetically cohesive units (59), with genomes ranging from nearly clonal to genetically diverse. The collection encompasses several deeply sampled *Vibrio* species represented by hundreds of genetically distinct isolates. In certain cases, these isolates differ by only a few single nucleotide polymorphisms (SNPs) within their clonal frame, i.e. the recombination-free part of the core genome, showcasing the vast microdiversity within bacterial populations.

Studying MGE diversity and the computational prediction of MGEs in natural environments face several challenges that can be circumvented by using this model system of marine *Vibrionaceae*. First, a comprehensive set of MGEs can be predicted *de novo* using comparative genomics without introducing database biases, because the close relationship of bacterial strains in the *Vibrionaceae* collection enables the precise location of genetically variable regions, consisting of the flexible genome which encompasses MGEs. Second, the evolutionary changes of MGEs can be investigated, as the diversification of MGEs is expected to be relatively small in very closely related bacterial strains. Third, the bacteria in this collection originate from a shared environment in which gene flow is not restricted. Thus, the investigation of recent MGE transfers is possible, including the identification of specific donor-recipient pairs.

Thesis outline

The chapters presented in this thesis aim to explore the diversity and the impact of MGEs infecting bacteria from a collection of marine *Vibrio*. Throughout this thesis, I aim to include those MGEs that have previously been largely overlooked. The two chapters were driven by the following questions: 1) What is the abundance, diversity and distribution of prophages? and 2) What is the diversity of MGEs and to what extent does this diversity contribute to the flexible gene pool of their bacterial hosts?

In chapter I, I demonstrate that tailless and filamentous prophages are predominant in marine *Vibrio*, even though tailed bacteriophage genomes from the *Caudoviricetes* family constitute the majority of phage sequences in public databases. By combining comparative genomics and sequence data collected from prophage-induced cultures, I predict prophages while avoiding database biases, thereby including prophages that were previously underappreciated. I demonstrate that tailless and filamentous prophages are adapted to very few integration sites, as opposed to the more diverse tailed phages. Furthermore, I show that tailless and filamentous prophages are highly diverse, with their diversity partly driven by variation in their accessory genes. This chapter highlights that tailless and filamentous prophages are, like tailed prophages, abundant and diverse drivers of the ecology and evolution of bacteria in nature.

In chapter II, I show that MGEs originating from bacteria isolated from a single environment are numerous and highly diverse, suggesting that previous large-scale studies may have underestimated the overall diversity of MGEs. In fact, I demonstrate that MGEs represent the main contributors to the flexible gene pool of their bacterial hosts. This diversity is driven by both a variety of different MGE types and subtypes, and highly variable gene content encoded in distinct genomic locations across a wide range of different MGEs, which harbor accessory genes or even entire MGEs. This chapter showcases the high abundance and large diversity of MGEs in closely related bacteria indicating the key role of MGEs in the evolution and adaptation of bacteria in the environment and demonstrates that MGEs are the major source of the variation observed in bacterial pangenomes.

References

1. Gogarten JP, Doolittle WF, Lawrence JG. Prokaryotic Evolution in Light of Gene Transfer. *Mol Biol Evol*. 2002 Dec 1;19(12):2226–38.

2. Smith JM, Smith NH, O'Rourke M, Spratt BG. How clonal are bacteria? *Proc Natl Acad Sci USA*. 1993 May 15;90(10):4384–8.

3. Fullner KJ, Boucher JC, Hanes MA, Haines GK, Meehan BM, Walchle C, et al. The Contribution of Accessory Toxins of *Vibrio cholerae* O1 El Tor to the Proinflammatory Response in a Murine Pulmonary Cholera Model. *JEM.* 2002 Jun 3;195(11):1455–62.

4. Davies J, Davies D. Origins and Evolution of Antibiotic Resistance. *Microbiol Mol Biol Rev*. 2010 Sep;74(3):417–33.

5. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol*. 2005 Sep;3(9):722–32.

6. Haudiquet M, De Sousa JM, Touchon M, Rocha EPC. Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Phil Trans R Soc B*. 2022 Oct 10;377(1861):20210234.

7. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP. Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat Commun*. 2014 Nov 19;5(1):5471.

8. Hussain FA, Dubert J, Elsherbini J, Murphy M, VanInsberghe D, Arevalo P, et al. Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science*. 2021;374(6566):488–92.

9. Lwoff A. Lysogeny. *Bacteriol rev*. 1953;269–337.

10. Salmond GPC, Fineran PC. A century of the phage: Past, present and future. *Nat Rev Microbiol*. 2015;13(12):777–86.

11. Helinski DR. A Brief History of Plasmids. *EcoSal Plus*. 2022;10(1).

12. Feschotte C. Transposable elements: McClintock's legacy revisited. *Nat Rev Genet*. 2023 Nov;24(11):797–800.

13. Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol*. 2022 Apr;20(4):206–18.

14. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. *Nat Biotechnol*. 2024 Aug;42(8):1303–12.

15. Koonin EV. Viruses and mobile elements as drivers of evolutionary transitions. *Phil Trans R Soc B*. 2016 Aug 19;371(1701):20150442.

16. Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol*. 2005 Sep 1;3(9):679–87.

17. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae* : Implications for the microbial "pan-genome." *Proc Natl Acad Sci USA*. 2005 Sep 27;102(39):13950–5.

18. Young JPW, Crossman LC, Johnston AW, Thomson NR, Ghazoui ZF, Hull KH, et al. The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol*. 2006 Apr 26;7(4):R34.

19. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. *Nat Microbiol*. 2017 Mar 28;2(4):17040.

20. Owen SV, Canals R, Wenner N, Hammarlöf DL, Kröger C, Hinton JCD. A window into lysogeny: revealing temperate phage biology with transcriptomics. *Microb Genom*. 2020 Feb 1 ;6(2).

21. Baharoglu Z, Mazel D. SOS, the formidable strategy of bacteria against aggressions. *FEMS Microbiol Rev*. 2014;38(6):1126–45.

22. Silpe JE, Bassler BL. A Host-Produced Quorum-Sensing Autoinducer Controls a Phage Lysis-Lysogeny Decision. *Cell*. 2019;176(1–2):268-280.e13.

23. Stecher B, Denzler R, Maier L, Bernet F, Sanders MJ, Pickard DJ, et al. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal *Enterobacteriaceae*. *Proc Natl Acad Sci USA*. 2012 Jan 24;109(4):1269–74.

24. Bertozzi Silva J, Storms Z, Sauvageau D. Host receptors for bacteriophage adsorption. *FEMS Microbiol Lett*. 2016;363(4):fnw002.

25. Bernheim A, Sorek R. The pan-immune system of bacteria: antiviral defence as a community resource. *Nat Rev Microbiol*. 2020;18(2):113–9.

26. Bobay LM, Rocha EPC, Touchon M. The adaptation of temperate bacteriophages to their host genomes. *Mol Biol Evol*. 2013;30(4):737–51.

27. Croucher NJ, Mostowy R, Wymant C, Turner P, Bentley SD, Fraser C. Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict. *PLoS Biol*. 2016 Mar;14(3):e1002394.

28. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, De La Cruz F. Mobility of Plasmids. *Microbiol Mol Biol Rev*. 2010 Sep;74(3):434–52.

29. Guédon G, Libante V, Coluzzi C, Payot S, Leblond-Bourget N. The Obscure World of Integrative and Mobilizable Elements, Highly Widespread Elements that Pirate Bacterial Conjugative Systems. *Genes*. 2017 Nov;8(11):337.

30. McKitterick AC, Seed KD. Anti-phage islands force their target phage to directly mediate island excision and spread. *Nat Commun*. 2018;9(1).

31. Ares-Arroyo M, Nucci A, Rocha EPC. Expanding the diversity of origin of transfer-containing sequences in mobilizable plasmids. *Nat Microbiol*. 2024;12:3240-3253

32. Babakhani S, Oloomi M. Transposons: the agents of antibiotic resistance in bacteria. *J Basic Microbiol*. 2018 Nov;58(11):905–17.

33. Bobay LM, Touchon M, Rocha EPC. Pervasive domestication of defective prophages by bacteria. *Proc Natl Acad Sci USA*. 2014;111(33):12127–32.

34. Cumby N, Davidson AR, Maxwell KL. The moron comes of age. *Bacteriophage*. 2012 Oct;2(4):e23146.

35. Wendling CC, Refardt D, Hall AR. Fitness benefits to bacteria of carrying prophages and prophage-encoded antibiotic-resistance genes peak in different environments. *Evolution*. 2021 Feb;75(2):515–28.

36. Takeuchi N, Hamada-Zhu S, Suzuki H. Prophages and plasmids can display opposite trends in the types of accessory genes they carry. *Proc R Soc B*. 2023 Jun 28;290(2001):20231088.

37. Mazel D. Integrons: agents of bacterial evolution. *Nat Rev Microbiol*. 2006 Aug;4(8):608–20.

38. Che Y, Yang Y, Xu X, Břinda K, Polz MF, Hanage WP, et al. Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. *Proc Natl Acad Sci USA*. 2021 Feb 9;118(6):e2008731118.

39. Bondy-Denomy J, Qian J, Westra ER, Buckling A, Guttman DS, Davidson AR, et al. Prophages mediate defense against phage infection through diverse mechanisms. *ISMEJ*. 2016 Jun 3;10(12):2854–66.

40. Mahata T, Kanarek K, Goren MG, Marimuthu Ragavan R, Bosis E, Qimron U, et al. Gamma-Mobile-Trio systems are mobile elements rich in bacterial defensive and offensive tools. *Nat Microbiol*. 2024 Oct 23

41. Getz LJ, Fairburn SR, Vivian Liu Y, Qian AL, Maxwell KL. Integrons are anti-phage defence libraries in Vibrio parahaemolyticus. *Nat Microbiol*. 2025 Jan 27

42. Mai-Prochnow A, Hui JGK, Kjelleberg S, Rakonjac J, McDougald D, Rice SA. 'Big things in small packages: the genetics of filamentous phage and effects on fitness of their host.' *FEMS Microbiol Rev*. 2015 Jul;39(4):465–87.

43. Dahlberg C, Chao L. Amelioration of the Cost of Conjugative Plasmid Carriage in *Eschericha coli* K12. *Genetics*. 2003 Dec 1;165(4):1641–9.

44. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS. A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation. *Cell Host Microbe*. 2020 Jan;27(1):140-153.e9.

45. Khedkar S, Smyshlyaev G, Letunic I, Maistrenko OM, Coelho LP, Orakov A, et al. Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Res*. 2022 Apr 8;50(6):3155–68.

46. Kieft K, Zhou Z, Anantharaman K. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*. 2020;8(1):1–23.

47. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*. 2021 Dec;9(1):37.

48. Néron B, Littner E, Haudiquet M, Perrin A, Cury J, Rocha E. IntegronFinder 2.0: Identification and Analysis of Integrons across Bacteria, with a Focus on Antibiotic Resistance in *Klebsiella*. *Microorganisms*. 2022 Mar 24;10(4):700.

49. Krupovic M, Makarova KS, Forterre P, Prangishvili D, Koonin EV. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biol*. 2014 Dec;12(1):36.

50. Hackl T, Laurenceau R, Ankenbrand MJ, Bliem C, Cariani Z, Thomas E, et al. Novel integrative elements and genomic plasticity in ocean ecosystems. *Cell*. 2023 Jan;186(1):47-62.e16.

51. Barcia-Cruz R, Goudenège D, Moura De Sousa JA, Piel D, Marbouty M, Rocha EPC, et al. Phage-inducible chromosomal minimalist islands (PICMIs), a novel family of small marine satellites of virulent phages. *Nat Commun*. 2024 Jan 22;15(1):664.

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

53. Roux S, Krupovic M, Daly RA, Borges AL, Nayfach S, Schulz F, et al. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat Microbiol*. 2019 Jul 22;4(11):1895–906.

54. Kalatzis PG, Mauritzen JJ, Winther-Have CS, Michniewski S, Millard A, Tsertou MI, et al. Staying below the Radar: Unraveling a New Family of Ubiquitous "Cryptic" Non-Tailed Temperate Vibriophages and Implications for Their Bacterial Hosts. *IJMS*. 2023 Feb 15;24(4):3937.

55. Zünd M, Ruscheweyh HJ, Field CM, Meyer N, Cuenca M, Hoces D, et al. High throughput sequencing provides exact genomic locations of inducible prophages and accurate phage-to-host ratios in gut microbial strains. *Microbiome*. 2021;9(1):1–18.

56. Xue H, Cordero OX, Camas FM, Trimble W, Meyer F, Guglielmini J, et al. Eco-Evolutionary dynamics of episomes among ecologically cohesive bacterial populations. *mBio*. 2015;6(3):1–10.

57. Kauffman KM, Hussain FA, Yang J, Arevalo P, Brown JM, Chang WK, et al. A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature*. 2018;554(7690):118–22.

58. Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton. *Science*. 2008 May 23;320(5879):1081–5.

59. Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol*. 2014;22(5):235–47.

Chapter I

Tailless and filamentous prophages are predominant in marine Vibrio

Kerrin Steensen[1,2], Joana Séneca[3,4], Nina Bartlau[1], Xiaoqian A. Yu[1], Fatima A. Hussain[5], Martin F. Polz[1*]

[1] Centre for Microbiology and Environmental Systems Science, Division of Microbial Ecology, University of Vienna, Vienna
[2] Doctoral School in Microbiology and Environmental Science, University of Vienna, Vienna, Austria
[3] Centre for Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria
[4] Joint Microbiome Facility of the Medical University of Vienna and the University of Vienna, Vienna, Austria
[5] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[*]Corresponding Author

Running title

Tailless and filamentous prophages in *Vibrio*

Keywords

prophage, tailless phage, filamentous phage, *Vibrio*, induction, integration, phage defense

Author contributions

Conceptualization – KS, MFP
Investigation – KS
Data curation – KS, NB, FAH
Formal analysis – KS, JS, NB
Software – KS, XAY
Visualization – KS
Writing (original draft) – KS
Writing (review and editing) – KS, MFP, JS, NB, XAY
Supervision, Funding acquisition – MFP

The ISME Journal - Open access publication

DOI: 10.1093/ismejo/wrae202

Abstract

Although tailed bacteriophages (phages) of the class *Caudoviricetes* are thought to constitute the most abundant and ecologically relevant group of phages that can integrate their genome into the host chromosome, it is becoming increasingly clear that other prophages are widespread. Here, we show that prophages derived from filamentous and tailless phages with genome sizes below 16 kb make up the majority of prophages in marine bacteria of the genus *Vibrio*. To estimate prophage prevalence unaffected by database biases, we combined comparative genomics and chemical induction of 58 diverse *Vibrio cyclitrophicus* isolates, resulting in 107 well-curated prophages. Complemented with computationally predicted prophages, we obtained 1,158 prophages from 931 naturally co-existing strains of the family *Vibrionaceae*. Prophages resembling tailless and filamentous phages predominated, accounting for 80% of all prophages in *V. cyclitrophicus* and 60% across the *Vibrionaceae*. In our experimental model, prophages of all three viral realms actively replicated upon induction indicating their ability to transfer to new hosts. Indeed, prophages were rapidly gained and lost, as suggested by variable prophage content between closely related *V. cyclitrophicus*. Prophages related to filamentous and tailless phages were integrated into only three genomic locations and restored the function of their integration site. Despite their small size, they contained highly diverse accessory genes that may contribute to host fitness, such as phage defense systems. We propose that, like their well-studied tailed equivalent, tailless and filamentous temperate phages are active and highly abundant drivers of host ecology and evolution in marine *Vibrio*, which have been largely overlooked.

Introduction

When comparing genomes of closely related bacteria, a frequently encountered difference is variation in their integrated bacteriophage genomes (prophages). These viruses, also called temperate phages, can be vertically inherited while genomically inserted in their host's chromosomes or transferred horizontally by prophage activation (induction) and subsequent infection of new hosts. Prophages are widespread bacterial parasites that commonly lead to host lysis upon induction. They also impact their host's phenotype by carrying accessory genes, such as those involved in antibiotic resistance (1), toxin production (2), and bacteriophage (phage) defense (3–7). This diversity of additional functions seems to represent a source of novel genetic material for bacterial hosts in natural environments, as temperate phages are rapidly gained and lost (8–10). However, most of our knowledge on prophage diversity is based on studies using prediction tools for the identification of genomically inserted prophages (11–14). These tools heavily rely on known prophages deposited in public databases, which are biased towards tailed dsDNA phages of the class *Caudoviricetes* (15). Tailed phages are characterized by the classical head and tail structure, using HK97-fold proteins for their icosahedral capsids. Even though most phage biologists are aware of the large bias towards tailed phages, only a few studies have focused on other prophage types. Several computational studies specifically targeting prophages derived from filamentous and tailless phages suggest that these prophages are more widespread than previously thought (16–19).

Filamentous phages of the order *Tubulavirales* differ from the tailed *Caudoviricetes* both morphologically as well as in their mode of infection and reproduction. Unlike most other prophages, they are secreted without killing their host. Instead, their 5-15 kb small, single-stranded DNA genomes are packaged into filamentous virions that are extruded through the bacterial membrane (20). Although specific filamentous phages, such as the phage CTXϕ involved in the virulence of *V. cholerae*, have been intensively studied in the context of disease for decades, the large scope of their natural diversity has only been recently uncovered (2,17).

Prophages belonging to the order *Vinavirales* are even less well studied but have been reported to occur in a diversity of bacterial species and different environments, such as in seawater or the human gut (16,18,19). These prophages that are commonly related to the tailless corticovirus PM2 are inducible spontaneously or after Mitomycin C (MMC) treatment (19,21,22). In contrast to tailed phages, tailless phages possess an icosahedral capsid formed by double jelly-roll-fold major capsid proteins, which contains a lipid layer covering their dsDNA genome. These characteristics, often unaccounted for in standard phage isolation protocols, might introduce an isolation bias further impeding their discovery as shown for other tailless viruses before (21).

Although it is becoming increasingly evident that prophages belonging to the *Vinavirales* and *Tubulavirales* are more widespread than previously thought, their abundance has not been evaluated in the context of other prophages, and how they impact their hosts is largely unexplored. Therefore, we aimed at obtaining a holistic view of the prevalence and diversity of prophages in a coastal ocean isolate collection of *Vibrio*, a bacterial genus that has been reported to carry prophages belonging to the *Vinavirales*, *Tubulavirales*, and *Caudoviricetes*, that we will refer to as "tailless", "filamentous", and "tailed" prophages for better readability in this manuscript (19,21,23,24). We show that filamentous and tailless prophages are more abundant (~80%) than tailed prophages in a well-curated model system of 58 *V. cyclitrophicus* strains using a combination of comparative genomics and chemical induction. This trend was confirmed in over 900 strains of naturally co-occurring marine *Vibrionaceae*. Although most detected prophages had small genomes below 16 kb, they often harbored accessory genes. Prophages were frequently gained and lost, as highlighted by their high diversification among closely related hosts. Our comprehensive prophage collection illustrates that tailless and filamentous prophages have been systematically underestimated in abundance and suggests that they are important drivers of host diversification in environmental bacteria.

Material and Methods

Bacterial isolation and growth conditions

Bacteria were isolated in previous studies from coastal seawater collected in two close-by sampling locations namely Plum Island Sound Estuary, Ipswich, MA in 2007 and Canoe Grove, Nahant, MA in 2010 as described before (53,58–60). In brief, *Vibrio* were isolated by applying a size fractionation approach followed by cultivation on TCBS selective media. If not otherwise stated, *Vibrio* isolates were cultivated using 2216 Marine Broth "2216MB" (Difco) at 25°C.

Vibrionaceae genome collection

For our experimental model system, we sequenced high-quality genomes of 58 *V. cyclitrophicus* strains in several rounds of third generation long-read sequencing (Oxford Nanopore) from bacterial cell pellets (Supplementary Methods). For prophage prediction across different species from the family *Vibrionaceae*, we used publicly available genomes, complemented with newly generated genomes (Table S2). The newly generated *Vibrionaceae* genomes were either sequenced as described previously (21), or sequenced as follows: Genome libraries were prepared using the Nextera DNA Flex library preparation kit (Illumina) utilizing a liquid handler, and then sequenced at 2x150 bp on a Nextseq 550 platform (Illumina). Reads were preprocessed by removing adaptors using hts_SuperDeduper with parameters '*start = 5, min_length = 40*' (v1.1.0, https://github.com/s4hts/HTStream) and phiX sequences were removed using bowtie2 (v2.3.4.2, (Langmead and Salzberg, 2012)). Reads were then quality-trimmed using Trim Galore with the specifications '*length = 30, stringency = 1, error_rate = 0.05, max_n = 1*' (v0.5.0, https://github.com/FelixKrueger/TrimGalore) and downsampled using a digital normalization approach implemented in the normalize-by-median.py script using the parameters '*kmer_size = 22, coverage_cutoff = 20, fp_rate = 0.8*' from the kmher software package (v3.0.0a1, (61)). The remaining reads were assembled using spades (v3.12.0, (62)) with the flag '*--careful*'.

Phylogeny of closely related *Vibrio cyclitrophicus*

The phylogenetic relationship of *V. cyclitrophicus* hosts was calculated as follows: First, we aligned the host core genome with Mugsy (v1.2.3, (63)), and used the alignment to construct a maximum-likelihood tree using PhyML (v3.3, (64)) with a General Time Reversible substitution model, as well as empirical models for the equilibrium nucleotide frequency, nucleotide substitution rates, and the gamma distribution shape parameter (*-m 'GTR' -t 'e' -a 'e' -f 'm' -v 'e'*). We then inferred the vertically inherited and recombined regions in the core genome alignment using ClonalFrameML (v1.12, (65)), with the maximum-likelihood tree and the transversion to transition rate kappa estimated during tree construction as input, and 100 bootstrap replications (*-emsim 100)*. Recombinant regions were masked using maskrc-svg (https://github.com/kwongj/maskrc-svg), and the masked alignment was used to calculate a new maximum-likelihood tree using the same settings in PhyML.

Prophage identification in *Vibrio cyclitrophicus*

Prophage candidate identification using comparative genomics

To get an unbiased prediction of prophages in the 58 *Vibrio cyclitrophicus* strains, we used a graph-based approach and identified 1,286 flexible genome elements using PPanGGOLiN in the panrgp mode (v1.2.105, (66)).

Sequencing of MMC induced cultures

To identify additional active prophages and distinguish prophages from non-phage mobile genetic elements, we treated bacterial cultures with MMC to induce prophages and subsequently sequenced the whole culture. In detail, *V. cyclitrophicus* strains (n=58) were incubated on 1.5% 2216MB streak plates overnight. Single colonies were transferred into 5 ml liquid 2216MB and allowed to grow overnight shaking at 250 rpm. Then, cultures were diluted 1:100 into 4 ml fresh medium in glass tubes, grown to an OD_{600nm} ~ 0.2, and then MMC (Sigma-Aldrich, M0503) was added to a final concentration

of 0.5 µg/ml. After MMC addition, 200 µl culture were transferred to a 96-well plate. OD_{600nm} was measured every 5 min shaking at 240 rpm with the Tecan Spark plate reader. Culture tubes were shaken at 250 rpm in the dark for either 6h or 14 h depending on if the optical density dropped before 6 hours (= sampling after 6 hours) or not (= sampling after 14 hours).

We extracted DNA of 500 µl induced bacterial culture using the Blood and Tissue kit (QIAGEN) according to the manufacturer's protocol with adjustments: The cell lysis step using Proteinase K and lysis buffer (QIAGEN) was increased to 15 minutes to ensure capsid disruption. DNA was eluted in 100 µl AE buffer (QIAGEN) and quantified using the Qubit dsDNA HS kit (Invitrogen). Barcoded sequencing libraries were prepared for sequencing using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) and NEBNext Multiplex Oligos for Illumina (New England Biolabs) and sequenced on a MiSeq using a v3 kit in 2x 300 bp mode (illumina).

Identification of active prophages

To identify the genomic location of the excised prophage elements, we mapped reads of induced cultures against the high-quality reference genomes. Then, we identified regions with increased read coverage compared to the host's background coverage adapted from Zünd et al. (25). We removed sequencing adaptors and phiX residues from the reads, that were then quality filtered (*qtrim=r trimq=14 minlen=45 maq=20 ftl=12)* using BBduk from the BBMap-suite (v. 39.01, (67)). The trimmed reads were mapped against their respective host reference genome using the burrows-wheeler local aligner bwa-mem (v0.7.17, (68)). We filtered the mapping output to obtain only the primary alignment to avoid reads mapping to more than one location, such as repeats, using samtools (v1.19) view with the flags *-h -F 260* (69). The alignments were filtered for an identity >= 99%, alignment length of >= 50 bp, and a read coverage of at least 80% using code adapted from mVIRs (25). The filtered file was sorted by name and a coverage "depth" profile was created using samtools (69). We proceeded with samples >250,000 reads resulting in a mean read coverage ranging from 6 to 95.

To identify high coverage regions, we used a custom python script. We set a tailored median *neighborhood* coverage and corrected standard deviation for each contig longer than 80 kb, instead of using values for the complete genome. To avoid natural skew of the coverage along very long contigs, we sliced contigs >350 kb into pieces of equal length and no more than 300 kb. The corrected standard deviation was calculated using the standard deviation excluding all extreme coverage values >(3 * median) to avoid that actively excising elements influence the corrected standard deviation of the background. The coverage along the host reference genome was then compared to the tailored median coverage value using a sliding window approach. To account for possible prophage plasmids, contigs between 5-80 kb with an elevated coverage of more than one standard deviation above the genome median coverage ($median_{contig} > median_{genome} + 1* std_{contig}$) were completely recorded as high coverage. To identify elements with elevated coverage across the chromosome, we calculated the mean coverage of a sliding window of 250 bp and a step-size of 50 bp. Regions were considered as high coverage if consecutive windows resulted in elements that are 500 – 1500 bp long with coverage >(median + 3*standard deviation) or >1500 bp with a coverage >(median+1*standard deviation). Overlapping elements, or elements being less than 1000 bp apart from each other were combined, including those that spanned contig borders. We manually inspected all genome coverage plots adding slightly elevated elements not found by the algorithm and expanding or decreasing the size of elements based on the base coverage values if necessary.

Prophage candidate curation by gene content

We combined all elements identified by induction and comparative genomics and eliminated non-phage elements as follows: We predicted genes using prodigal in the meta mode (v2.6.3, (70)). and annotated them using hmmsearch (hmmer v3.4, (71)) against VOGdb (v221, (72)) using an e-value of 0.001. Elements were considered as putative prophages if they contained at least 5 genes and a) either contained a clear marker gene, that were terminase (large/endonuclease subunit) or major capsid proteins for tailed phages, a pI-like ATPase for filamentous phages or a double jelly-roll capsid protein for tailless phages or b) contained at least 2 genes indicating a viral origin and contained the words

'coat', 'capsid', 'head', 'tail', 'virion', 'phage', 'holin', 'terminase', 'replication', 'packaging', as well as filamentous phage associated keywords 'Vpf81' and pI-like ATPase hits namely 'BPPHL_Gene_1_protein' and 'BPFD_Gene_1_protein'. The resulting prophage candidates were merged and manually inspected to distinguish between prophages and other microbial genetic elements. Finally, we ensured prophage candidates to be single, distinct elements by manually inspecting gene predictions and alignments of all identified prophage candidates using clinker (v0.0.28, (73)) Adjacent prophages were separated. Additional elements next to the phage that were absent in most other related prophages were discarded. The resulting 107 elements, of which 106 were identified by comparative genomics and 65 by chemical induction, were considered to be prophages.

Prophage border curation

We predicted the prophage borders accurately by combining the results obtained from comparative genomics and the induced prophage coverage data. We used a hierarchical approach: First, we checked for a sufficient number of "clipped" reads close to the predicted borders calculated by mVIRs with the flags -ml 1000 -ML 100000 -m (v1.1.1-3.9.16, (25)). These reads are spanning the part of a circularized phage that is connected to the host genome upon integration. While one part of the read maps well to the host reference, the other part of the read gets clipped at the transition from host to the integrated prophage and therefore provide exact positions of the prophage borders. Second, if clipped reads were unavailable but the phage was actively induced, we used the base coverage of the prophage region to determine the prophage borders. Otherwise, we used the borders predicted by PPanGGOLiN covering the prophage from its first to its last gene discarding the non-coding ends. For phages inserted in a plasmid, the border was set to the first gene not shared with other plasmids having a similar backbone.

Prophage identification across the *Vibrionaceae*

Marker gene collection

We created a marker gene collection from the predicted prophages in *V. cyclitrophicus.* We obtained these genes by creating an all-against-all protein similarity matrix using blastp with an e-value of 0.00001 and subjected it to the markov clustering tool mcl (v22.282) (mcxload with options *--stream-mirror --stream-neg-log10 -stream-tf 'ceil(200)'*) using an inflation factor of 3. Clusters containing marker genes, that were terminase (large/endonuclease subunit) and major capsid proteins for tailed prophages, pI-like ATPase proteins for filamentous prophages and double jelly-roll major capsid proteins for tailless phages were merged into a single marker gene collection per viral realm.

Predicting related phages across the Vibrionaceae and in V. cyclitrophicus draft genomes

We ran ppanngolin (v1.2.105,(66)) species-wise in *panrgp* mode to obtain genomic islands of all strains with clear taxonomic assignment containing at least 5 members. First, we used blastp (v2.15.0,(74)) to compare genes from our custom-made marker gene collection for each viral realm against the genomic islands. Second, we ran the viral gene search described in "Prophage candidate curation by gene content" and sorted elements into the respective viral realms or "other". We applied a viral-realm specific size filtering step (filamentous/other 4 kb, tailless 10 kb, tailed 25 kb). Prophages that were largely disrupted or not showing integrity with related prophages were excluded.

To investigate how the use of fragmented genomes influences prophage identification, we repeated the genomic island prediction and viral gene search in *V. cyclitrophicus* using genomes assembled from next generation sequencing data (Table S2) with an average contig count of 180.6 compared to 3.8 in high quality genomes. We then checked if they were found in both high-quality and draft genomes of *V. cyclitrophicus* using blastn (v2.15.0, (74)).

Prophage annotation and classification

We annotated the curated prophages with accurate borders from *V. cyclitrophicu*s (Table S3) and all prophages across the *Vibrionaceae* with a custom pipeline: Genes were predicted using prodigal in the meta mode (v2.6.3, (70)). Genes were annotated using *hmmsearch* (hmmer v3.4, (71)) against VOGdb (v221) using an e-value of 0.0001, hmmscan (evalue 0.001) against the antidefense database APIS

(2023-09) (47), Defensefinder (v1.2.2,(46,75), blastp (v2.15.0, (74)) against the Swiss-Prot database (last access in 03-2024, (76)), and interproscan (v5.67-99.0-11.0.4, (77)) using the applications TIGRFAM, SUPERFAMILY, ProSiteProfiles, and Pfam.

Viral realms were determined as follows: Prophages containing a double jelly-roll capsid were grouped into the realm *Varidnaviria* ('tailless phages'). Prophages containing a pI-like ATPase, predicted as 'gene 1' or 'zonular occludens toxin', were grouped into the *Monodnaviria* ('filamentous phages') and prophages containing capsid or a terminase (large/endonuclease subunit) were grouped into the realm *Duplodnaviria* ('tailed phages'). The viral groups were further classified using the VICTOR web service (28) by pairwise comparison on nucleotide level using the Genome-BLAST Distance Phylogeny (GBDP) method (78) and subsequent clustering into genera using the recommended clustering thresholds and an F value of 0.5. The resulting phylogenomic GBDP trees were inferred with FASTME (79) using the formula D0 and yielded an average support of 49% (tailless), 58% (filamentous), and 59% (tailed). we considered VICTOR's GBDP method to be better suited for this analysis, because similarity-based approaches as used in VIRIDIC (80) would be highly influenced by the variable locus causing very small clusters.

Comparison to known prophage sequences

We phylogenetically classified the identified prophages per viral realm by running blastp (bit score > 50, evalue <0.00001, (v2.15.0) (74) of their predicted proteins against the INPHARED database (March 2024, (81)). We considered results with genome sizes of 50% - 150% of the query phage genome and sharing >=5% of query phage genes. We calculated a tree of all hits combined with the identified prophages using VipTree (v1.1.2) (82) and extracted sequences clustering together. We then calculated a phylogenetic tree with the identified prophages together with selected representatives of known viruses (Table S4) based on the INPHARED search and current ICTV classification (International Committee on Taxonomy of Viruses (ICTV): https://ictv.global/taxonomy/) using VICTOR (28) as described above. In addition to the whole-genome trees, we also calculated trees of the respective marker genes for each group using mafft (v.7.525, (83)) and fasttree (v.2.1.11, (84)) with standard settings. We checked whether current prophage prediction tools recognized the 107 well-curated prophages as viruses using VirSorter2 (v.2.2.4, (15)) with standard settings and geNomad (v.1.8.0, (29)) in end-to-end mode.

Structural analysis of double jelly-roll major capsid proteins

We extracted the amino acid sequence of the genes predicted as double jelly-roll major capsid proteins of reference phage NO16 and selected tailless prophages. Structural predictions were performed using the the AlphaFold Server powered by the AlphaFold 3 model (85) and visualized using chimeraX (86). The crystal structure of the major capsid protein of PM2 (87,88) was downloaded from the PDB database.

Prophage activity analysis

We used sequenced MMC induced cultures to estimate prophage activity. We first calculated the mean sequencing coverage of each curated prophage element and compared it to their tailored median neighborhood coverage and corrected standard deviation for that specific region described in "Identification of active prophages". Prophages were deemed active if their mean coverage exceeded the median coverage of their genomic neighborhood by >=one corrected standard deviation, or highly active if the median coverage was exceeded by >=three corrected standard deviations. We defined the phage-to-host ratio (PtoH) by dividing the prophage mean coverage by the median neighborhood coverage and further used it as a proxy for phage activity. Plasmid-associated phage activity and PtoH values were refined by first comparing the mean coverage against the whole genome median and multiplied by a plasmid correction factor retrieved from the contig coverage of the original non-induced nanopore reads against the reference genome. We only applied the correction factor if plasmids were present in multiple copies meaning that the corrected PtoH could only decrease.

We checked if tailed and tailless phage activity significantly differed by comparing their PtoH values. Thus, we applied the Shapiro-Wilk test, which showed a non-normal distribution, followed by the Wilcoxon rank sum exact test.

Prophage transfer analysis

Prophage similarity analysis

To identify recently transferred prophages, we calculated the similarity of all identified prophages in *Vibrio* as average nucleotide identity using fastANI (v.1.34, (89)). We then filtered all prophage pairs for an identity of at least 99.9% and complete coverage.

Host Phylogeny

To classify *Vibrionaceae* genomes on species level (Table S2), a phylogenetic tree of the concatenated ribosomal proteins was calculated using RiboTree (https://github.com/philarevalo/RiboTree) using the *Vibrionaceae* RefSeq genomes (downloaded November 2023 from NCBI) as reference. In brief, we removed all contigs <1000 bp from the genomes and ran RiboTree with 100 bootstraps using *Shewanella denitrificans* (GCF_000013765.1) as outgroup. Genomes were classified accordingly if they clustered with the RefSeq genomes. Strains not clustering into species were excluded from the following analysis.

The *V. cyclitrophicus* tree used for inference of the prophage transmission type was calculated as described in "Phylogeny of closely related *Vibrio cyclitrophicus*".

Variable locus analysis

To find accessory gene loci in filamentous and tailless prophages identified in the *Vibrionaceae*, we built fine-grained prophage clusters based (VGCs) using VirClust with standard settings and a clustering distance of 0.6 (90). We dereplicated the prophages at 99.9% using cd-hit resulting in 317 (out of 401) tailless and 239 (out of 302) filamentous prophages (v4.8,1, (91)). Prophage-encoded proteins were assigned to gene families by mcl clustering on a blastp generated protein similarity matrix (*e-value 0.00001*) with the settings *'--stream-mirror --stream-neg-log10 -stream-tf 'ceil(200)''* (74,92) and we calculated their frequency within the prophage clusters. Gene families flanking the variable loci were identified by manual inspection of the synteny plots. We excluded clusters with <5 representatives or not displaying variable genes enclosed by two core phage genes (present in >80% of the phages). We recorded each variable locus from the first to the last variable gene (present in <=30% of the phages) between the respective core genes resulting in 118 filamentous and 243 tailless protein-encoding variable loci and annotated the genes with our custom pipeline (comp. "Prophage annotation and classification"). Defensive genes were reported if found by Defensefinder, or if their annotation was related to abortive infection or toxin/antitoxin systems. Systems were considered complete when reported by Defensefinder, or if neighboring toxin and antitoxin genes were found.

Integration site analysis

Location and functional annotation of integration sites

Prophages integration sites, containing at least one prophage in *V. cyclitrophicus* (Table S1), were inferred using PPanGGOLiN (v1.2.105, (66)). We blasted the flanking regions of selected prophages per hotspot against the genome of 10N.286.51.B1, a *V. cyclitrophicus* strain with no occupied prophage hotspots, to locate the integration sites. To confirm that filamentous phages were integrated at the *dif* sites, we used blastn tailored for short sequences *(-task blastn-short -word_size 12*) (v2.15.0, (74)) to compare the integration site and the prophages themselves against previously published *dif* sequences from *Vibrio cholerae* (41). To check for the functional annotation of hotspot 1, we computationally excised the tailless prophage integrated in hotspot 1 at its predicted borders (comp. Prophage border curation) and inspected the stitched region for their functional annotation assigned by the NCBI Prokaryotic Genome Annotation Pipeline (93).

Integrase Trees

We calculated trees of the integrase genes of both tailless and tailed phages running mafft (v.7.525, (83)) and subsequently fasttree (v.2.1.11, (84)) with standard settings. Filamentous phages were excluded because they do not regularly encode integrase genes.

Data analysis, tools and visualization

All analyses were perfomed using bash and python (v3.12.1). We used the libraries pandas (v2.1.4), matplotlib (v3.8.2), numpy (v1.26.3), and seaborn (v0.12.2). Statistical analyses were conducted in R (v4.3.2). We used ChatGPT (GPT-4) to improve grammar and clarity of the text. Trees were visualized using iToL (94) or ggtree (v3.8.2,(95)). The heatmap was plotted in R (v4.3.2) using the additional packages phangorn (v2.11.1,(96)), ggplot2 (v3.4.4,(97)) with helper functions from reshape2 (v1.4.4), patchwork (v1.1.3), dplyr (v1.1.4), and RColorBrewer (v1.1-3). Figures were edited with Inkscape (v1.1.2).

Results

Vibrio cyclitrophicus harbor prophages from three viral realms

Using a combination of comparative genomics and chemical induction, we identified 107 prophages in 58 marine *V. cyclitrophicus* strains, which were isolated from the same coastal area (Figure 1, Table S1-S2). First, we predicted genomic islands from high-quality *V. cyclitrophicus* genomes, exploiting the fact that prophages belong to the flexible genome. To better differentiate prophages from other mobile genetic elements, we then sequenced DNA from cultures that were exposed to mitomycin C (MMC), which can induce prophages via the host's SOS system. All genomic islands and induced elements were combined, filtered based on their viral gene content (Table S3) and manually curated. This pipeline resulted in a prophage dataset with accurately predicted prophage boundaries (methods), in contrast to prophage prediction tools, which oftentimes imprecisely predict prophage borders (25). Our comprehensive approach revealed that prophages are common with an average prophage count of 1.8 and >80% of the investigated strains carrying at least one prophage (Figure S1), which is in accordance with previously estimated prevalence among diverse bacteria (26).

We found a large diversity of prophages spanning three distinct viral realms: tailless prophages of the order *Vinavirales* possessing a double jelly-roll-fold major capsid protein characteristic for viruses of the realm *Varidnaviria*, filamentous prophages containing the pI-like ATPase marker gene and belonging to the order *Tubulavirales* assigned to the realm *Monodnaviria* and tailed prophages containing a HK97-fold capsid typical for orders within the *Caudoviricetes* from the realm *Duplodnaviria* (17,27). Even though filamentous and tailed prophages have been reported together (12), we show that prophages from all three realms can co-occur in single strains of *V. cyclitrophicus* (Figure 1). Classifying each prophage group using a nucleotide sequence-based approach (28) resulted in 12 genera (I-XII, Figure 1). We compared prophage genomes and their marker genes from all genera to publicly available viral genomes (Figure S2 and S3, Table S4) and, additionally, checked whether the prophages are recognized by two current prophage prediction tools, VirSorter2 (15) and geNomad (29) (methods).

Genomes from the three tailless prophage genera did not cluster with any other tailless phages on genus level (Figure S2a). Despite low amino acid identity, the identified prophages shared blocks of similar open reading frames with the phages PM2 and NO16 (Figure S4a) suggesting that the identified prophages also belong to the *Corticoviridae*. Such PM2-like prophages have previously been found in diverse marine bacteria (16,19). Comparisons of the double jelly-roll major capsid protein, however, show that all three genera cluster together next to capsid proteins originating from *Autolykiviridae* viruses sampled in the same coastal area (Figure S3a). The high nucleotide identity among double jelly-roll major capsid proteins of all different tailless phages observed in our *Vibrio* collection might be a result of their shared environment including similar bacterial hosts. This raises the question whether

genome structure might be more meaningful for the classification of tailless phages, as conserved synteny has been noticed in tailless phages before (30). Most identified tailless prophages from genera I and II were recognized as viral by at least one prophage prediction tool. However, prophages of genus III were never detected (Table S1). These appeared more distantly related to other tailless phages (Figure 2a) and were exclusively encoded on plasmids carrying various mobile genetic elements (Figure S4b). Given that the structural prediction of their double jelly-roll major capsid proteins resembled those of capsid proteins encoded by other tailless phages (Figure S5), genus III might represent a distinct group of tailless phages.

In contrast to the tailless prophages, most filamentous and tailed prophages were related to viruses from known genera (Figure S2b-c) and each identified prophage was also recognized by at least one prophage prediction tool (Table S1). Whole genome comparisons suggested that prophages from the most prevalent filamentous prophages belonging to genus V were related to fibrovirus fs1, which has been reported to release large amounts of phage particles into the supernatant of host cultures (31). Consistently, several filamentous phages from genus V appeared active upon MMC induction as indicated by an increase in the coverage compared to their host's genomes (Figure 1). However, their pI-like ATPase genes seemed more closely related to those of multiple other inoviridae genera (Figure S3b). Filamentous prophages of genus VII did not appear to be closely related to known viruses, but occasionally occurred as tandemly integrated elements as known from other filamentous phages (17,32). Most tailed prophages from our dataset belonged to genus IX and resembled myoviruses of the family *Peduoviridae* (Figure S2c), similar to lytic tailed phages previously isolated from the same coastal area (21).

Filamentous and tailless prophages are highly prevalent across the *Vibrionaceae*

Both tailless and filamentous prophages outnumbered tailed prophages in our model system *V. cyclitrophicus*. Filamentous prophages were the most prevalent group (40.4% of total), followed by tailless prophages (37.6% of total) compared to relatively few detected tailed prophages (22% of total) (Figure 2a). The predominant filamentous and tailless prophages had small genome sizes, ranging between 6.0 – 13.3 kb for filamentous phage genomes and 12.1 – 15.1 kb for tailless phage genomes. Tailed prophages were larger ranging from 32.4 to 42.1 kb and one prophage with 55.6 kb exceeded this size range (Figure 2b). With the predominant prophages having genomes smaller than 16 kb, the size distribution we report here is in contrast to studies reporting the average genome size of prophages to be above 30 kb (26,33).

To investigate whether this prophage distribution was limited to *V. cyclitrophicus*, we expanded our search to elements related to the identified prophages in a broader set of genomes of 11 species belonging to the family *Vibrionaceae* that co-exist with *V. cyclitrophicus* in the same coastal environment (Table S1-S2). We observed that tailed prophages were most abundant (39.6%), followed by tailless (34.3%), and filamentous (26.1%) prophages (Figure 2a). Whereas the entire set of *V. cyclitrophicus* genomes were closed or nearly so, prophage identification across the expanded dataset of *Vibrionaceae* may be influenced by their incomplete assembly. Therefore, we additionally predicted prophages in draft genomes of the 58 *V. cyclitrophicus*, which indeed suggests a systematic underestimation of filamentous prophages (Figure 2a), of which we could identify only 28 in the draft genome set in comparison to 44 in high-quality genomes. Therefore, the observed abundances suggest that all three viral realms are roughly evenly represented in the totality of prophages in the *Vibrionaceae*.

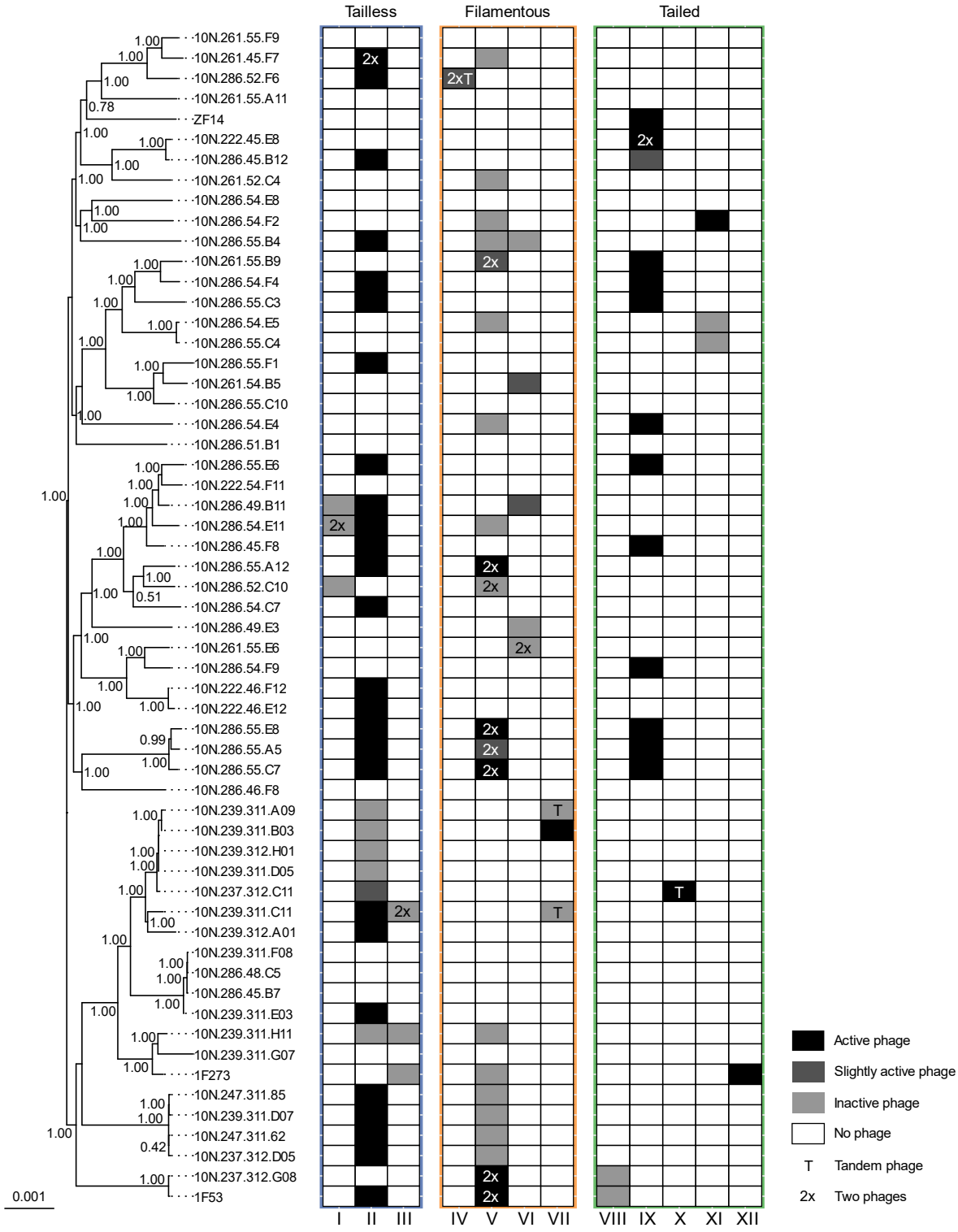

Figure 1 - Prophage presence in a set of 58 *V. cyclitrophicus* isolates.
Clustermap showing the recombination-free phylogenetic tree of 58 *V. cyclitrophicus* along with the presence and activity of induced prophages in the respective bacterial strains. Bootstrap values ranging from 0 to 1 are indicated at the nodes, and the tree scale shows the nucleotide substitutions in the core genome comprising 3,785,366 bp. Columns I-XII reflect VICTOR-predicted genera of prophages belonging to the three different viral realms. Temperate phage activity measured as coverage increase over the host genome in MMC-induced cultures is depicted in greyscale (active: > median + 3*std, slightly active: > median coverage + 1*std, inactive: < median coverage + 1*std). The Labels inside the heatmap denote two similar prophages co-occurring in the same integration site (T) or in different locations in the genome (2x).

Prophages of all viral realms appear to actively replicate

We aimed at measuring prophage activity by sequencing MMC-treated bacterial cultures after 6 or 14 hours of growth (Table S2) and subsequently mapping the reads against a host reference genome. The calculated phage-to-host ratio (PtoH) serves as a proxy for the number of phage particles being produced per cell (methods). Although this assay also includes spontaneously induced prophages, the majority of prophages appeared inducible by MMC, a known activator of the SOS response in bacteria (34).

We found that prophages from all three viral realms amplified their genomes upon MMC treatment (Figure 1). About 48% demonstrated high activity (coverage >median+3*std) 11% seemed to replicate at rather low levels (coverage >median+1*std) and 41% of the prophages identified in *V.cyclitrophicus* appeared to be inactive. To better understand what influences the number of produced phage particles, we checked if the number of produced phage particles might be predicted by their viral realm. Even though the tailless phage average PtoH ratio of 135 was higher than 79 in tailed phages (Figure S6), we did not find evidence for a significant difference in tailed and tailless phage activity (Wilcoxon rank sum test, p-value = 0.6419). We additionally observed that three identical tailless phages had highly variable PtoH ratios under similar experimental conditions in different strains of *V. cyclitrophicus*. Therefore, the number of produced phage particles could also be linked to host features such as growth or its genome architecture including inserted microbial genetic elements as additionally inserted prophages have been shown to change the number of formed phage particles upon induction (35,36).

Filamentous phages generally demonstrated lower phage-to-host ratios than other phages, probably caused by their ssDNA genome not being captured by the sequencing technique. The slightly increased PtoH values likely reflect the double stranded replicative form of filamentous phages occurring during replication (37,38). Therefore, we did not compare filamentous phage activity against the PtoH values of tailless and tailed phages.

Prophages are rapidly gained and lost

Having identified prophages from multiple *Vibrio* species co-existing in the same environment, we aimed at estimating the dynamics of prophage gain and loss in the context of host diversity. To constrain these dynamics to more recent timescales, we only considered closely related prophages that recently descended from the same ancestor. We therefore carried out an all-against-all comparison of the 1,158 prophages found across the *Vibrionaceae* and selected prophages with an average nucleotide identity ≥99.9% and full sequence coverage to allow for single nucleotide substitutions (methods). We observed that the 1474 pairs of highly similar prophages we retrieved were primarily found in hosts of the same species (Figure S7) suggesting that in their recent history, horizontal transfer of temperate phages between different *Vibrio* species was rare. To further explore the dynamics within a species, we focused on our well-curated model system *V. cyclitrophicus*. To infer the evolutionary distance between hosts, we computed a phylogenetic tree unaffected by changes in the flexible genome and homologous recombination (methods). Interestingly, highly similar prophages were still limited to more closely related *V. cyclitrophicus* strains, indicating that temperate phages are specific to hosts on the sub-species level (Figure 3).

We observed prophages from all three viral realms to be rapidly gained and lost within *V. cyclitrophicus*. This was demonstrated by variable prophage content even among the most closely related sisters on the tree (Figure 3), which has been associated with rapid mobile genetic element turnover before (e.g. 9,10). Strikingly, most identified prophages were unique (n=62/107) as they were not highly similar to any other prophages in *V. cyclitrophicus* (Figure 3). This implies that there could be a large pool of diverse temperate phages infecting *V. cyclitrophicus*, which is also demonstrated by the 22 tailed prophages clustering into 5 different genera (Figure 1). As bacterial strains of different *Vibrio* species have been shown to be naturally transformable upon induction with chitin available in the aquatic ecosystem, loss could be mediated through homologous recombination with strains lacking the prophage (9,39,40). Most prophage genotypes appeared not successful enough to spread widely across

V. cyclitrophicus. However, an exception was a tailless prophage genotype that might be currently expanding (Figure 3). We found evidence that this tailless prophage horizontally transferred into three evolutionary separated branches in *V. cyclitrophicus.* In two of these branches, the prophage genotype was also vertically inherited.

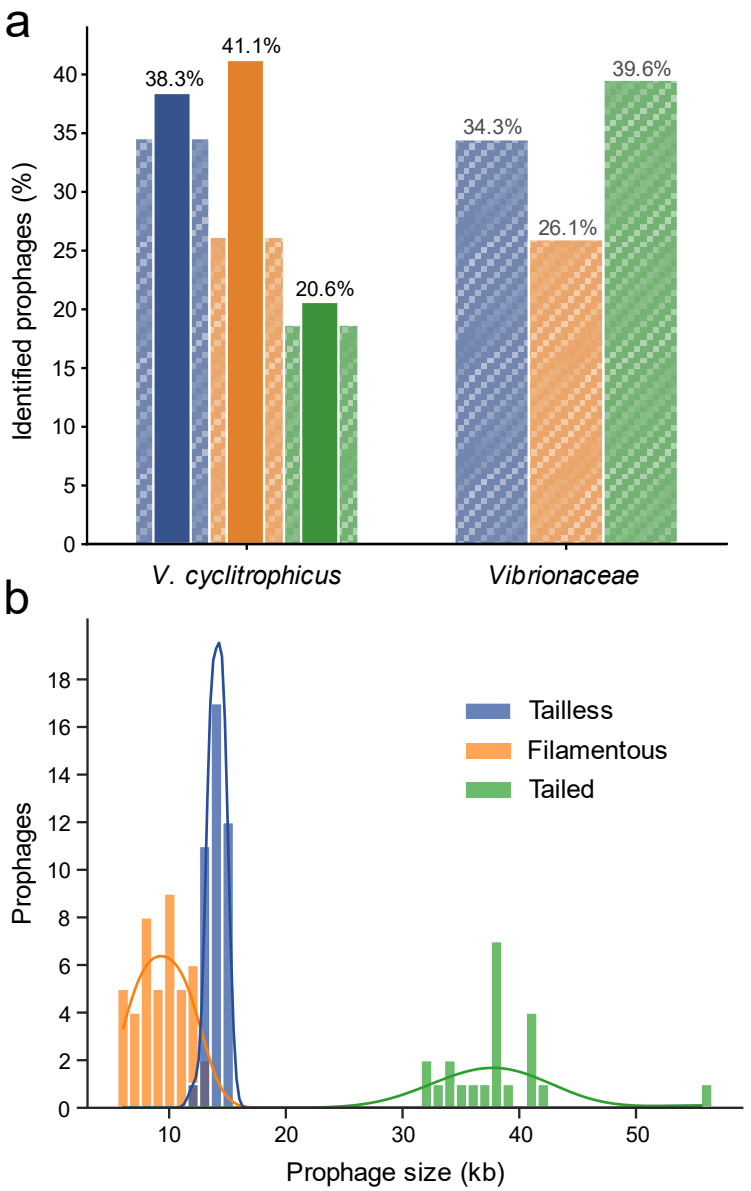

Figure 2 - Prophage distribution and genome size in the *Vibrionaceae*
a) Proportion of each viral realm in the total number of prophages identified in 58 *V. cyclitrophicus* strains and in a broader set of 931 bacterial strains across the *Vibrionaceae* including the 58 *V. cyclitrophicus* strains. Fully colored data indicate highly curated prophages predicted in high-quality genomes. Hashed data indicate computational prediction in draft genomes. b) Histogram showing the genome size of prophages in *V. cyclitrophicus* colored by their respective viral realm.

Prophages integrate by site-specific recombination

Having shown that tailless and filamentous prophages were rapidly turned over in *V. cyclitrophicus*, we were interested in how their integration affects the host's genome architecture. Prophages were specifically recombining into their host's genome demonstrated by the low number (n=9) of integration sites, each of which was mostly containing prophages from a single viral realm. For tailed and most tailless prophages, this specificity was mediated by prophage-encoded integrases that were closely related among prophages integrating into the same site (Figure S8a-b). Furthermore, tailed prophage integrases were diverse, explaining the higher number of integration sites compared to tailless prophages. All filamentous and a few tailless phages recombined site-specifically at the *dif* sites of chromosome 1 and 2. The *dif* sites, conserved sequence motives at each of the two chromosomal termini in *Vibrio* involved in chromosome partitioning (20), have been observed to harbor diverse filamentous prophages in *Vibrio* before (20,41). Filamentous phages use host recombinases for integration (20,42), making the possession of an integrase gene obsolete. In contrast, we observed that tailless prophages targeting the *dif* sites regularly carried integrase genes. Therefore, it remains inconclusive whether tailless prophages recombined into these two hotspots by using host recombinases as suspected before (19), or dependent on their own integrases.

Filamentous and tailless prophages appeared to rescue the function of genomic locations disrupted during integration. Tailless prophages occupying hotspot H1 integrated into the 5' end of the *dusA* gene coding for a tRNA dihydrouridine synthase, a known target of integrases (43). The disrupted *dusA* gene was complemented by a prophage-encoded sequence, likely preserving its function. Furthermore, we observed that several filamentous and tailless prophages integrated at the *dif* sites encoded a small sequence motif showing similarity to their host's *dif* sites possibly ensuring correct chromosome partitioning.

Integration sites targeted by filamentous and tailless phages tended to have higher occupancy than those of tailed phages (Figure 4a). Integration sites containing filamentous or tailless prophages were filled in 33–57%. In contrast, hotspots of tailed prophages were occupied in less than 11% of the 58 strains. One exception was hotspot H6 that was filled in most *V. cyclitrophicus* strains containing a diverse set of elements, such as tailed prophages, transposable elements and defense islands. Our data suggest that tailed temperate phages are more likely to encounter free integration sites, whereas filamentous and tailless temperate phages encounter filled sites more frequently, which might influence the success of prophage integration. Remarkably, filamentous and tailless prophages occasionally co-occurred in a single integration site. Across the *Vibrionaceae*, we found at least 30 such cases. Oftentimes prophages integrated next to each other, however, in *V. cyclitrophicus* we also observed tailless phages integrating into a filamentous prophage separating its transposase from the rest of the filamentous phage genome (Figure 4b). It remains unclear if these two co-occurring prophages can still produce viable virions. Unlike several related filamentous prophages, both prophages remained inactive upon MMC treatment. This co-integration of tailless and filamentous phages suggests possible competition between viruses of different viral realms, for example leading to disruption or inactivation of the resident phage.

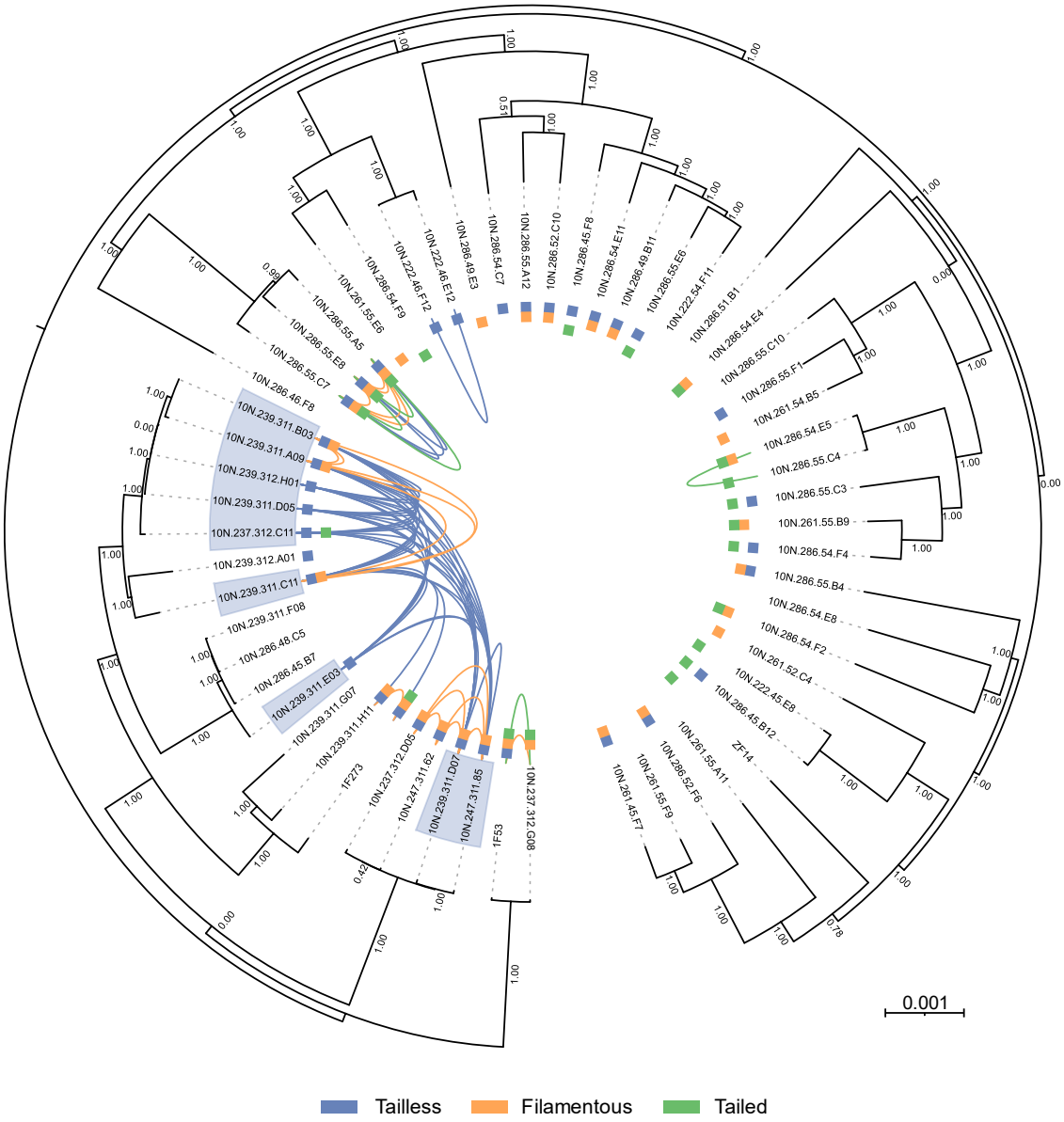

Figure 3 - Prophage transfer and inheritance within 58 co-occurring *V. cyclitrophicus*
Similar prophages (ANI >= 99.9%, full coverage) shared between different strains of *V. cyclitrophicus* are depicted as connections colored by viral realm. The tree shows the relationship of bacterial hosts calculated using a recombination-free alignment of the host core genome. Bootstrap values ranging from 0 to 1 are indicated at the nodes, and the tree scale denotes nucleotide substitutions in the core genome comprising 3,785,366 bp. Hosts carrying prophages are marked with boxes colored by the associated viral realm. Boxes without any connections indicate that the prophages were unique, i.e. not highly similar to other prophages. Hosts colored in light blue carry a similar tailless prophage that seems to currently expand in the population by horizontal gene transfer between, and vertical inheritance within branches.

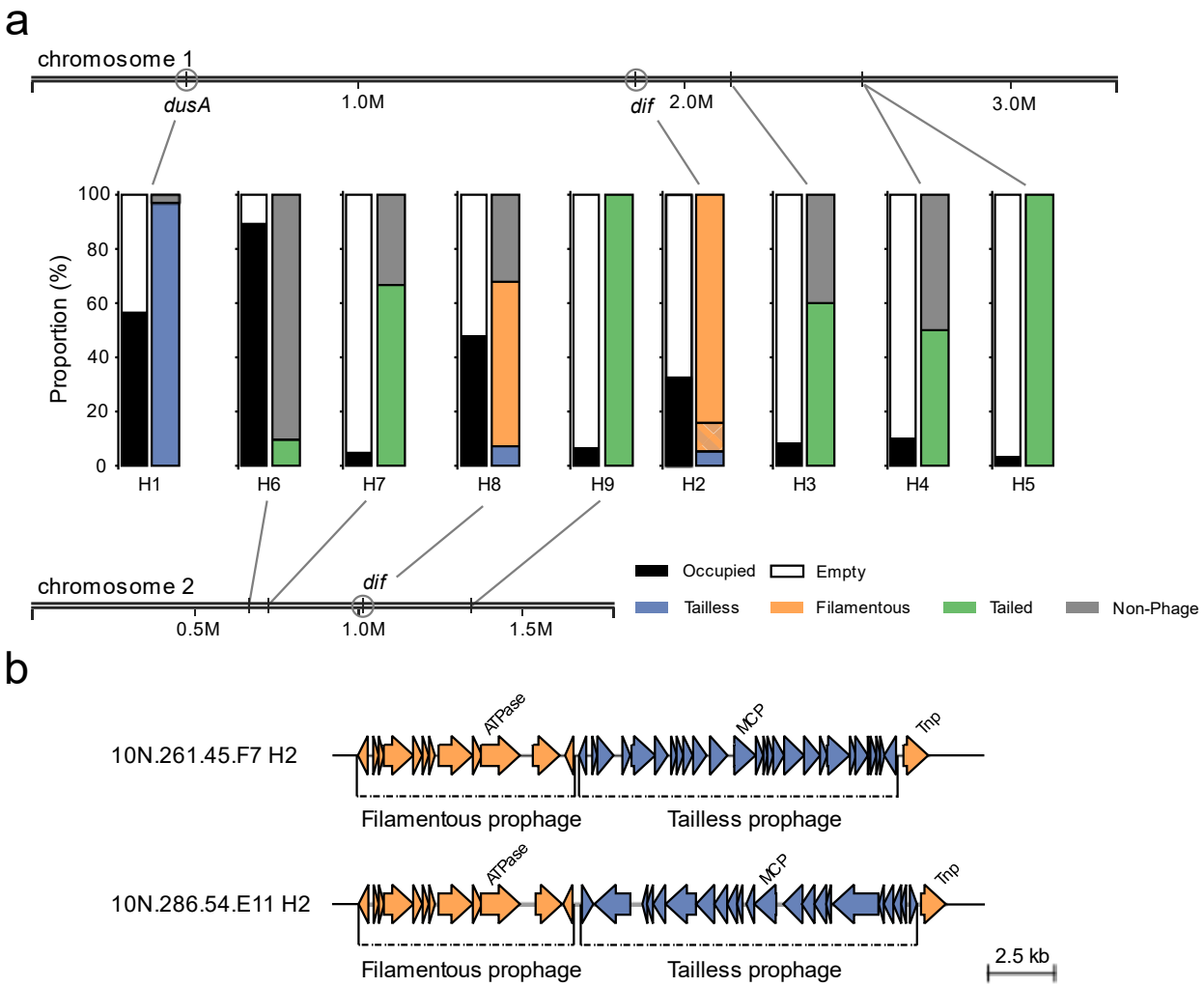

Figure 4 - Prophage integration sites in *V. cyclitrophicus*
a) Depicted are the locations of prophage integration hotspots H1-H9 on both chromosomes of the prophage-free *V. cyc*litrophicus 10N.286.51.B1. Bar plots show the total fill of each integration site in percent (left bar) and which elements the fill is composed of (right bar). Integration sites of tailless and filamentous temperate phages are indicated by circles. The plot shows data of 95 prophages excluding plasmid-encoded prophages and those with inconclusive integration sites. b) The box shows tailless prophages (blue) integrated between the transposase (Tnp) and the rest of a filamentous prophage (orange) in the shared hotspot H2 of V. cyclitrophicus strains 10N.261.45.F7 and 10N.286.54.E11. The signature genes for both phages, the double jelly-roll major capsid protein (MCP), and the pI-like ATPase, are indicated. The scale bar shows the sequence length in kb.

Tailless and filamentous phage genomes possess an accessory gene hot spot.

As many prophages carry genes not needed for the viral life cycle (5,44,45), we investigated the extent to which prophages with small genomes encode for such accessory genes. Assuming that accessory genes are more frequently exchanged than genes required for infection and reproduction, we searched for variable genes in closely related prophages. To this end, we grouped tailless and filamentous prophages predicted across the *Vibrionaceae* into fine-grained clusters using a protein similarity-based approach (methods). Within clusters, we observed highly conserved synteny between prophage genomes. Single proteins varied in their amino acid sequence presumably caused by gene conversion between prophages. The more striking observation was, however, one locus with variable gene content interrupting this conserved gene order (Figure 5a). In fact, the variable genes were rarely shared with any other identified prophages. We grouped similar prophages and identified variable loci in 49.4% of unique filamentous and 76.7% of unique tailless prophages (methods). For the remaining prophages, we were lacking enough representatives or variability between genomes to differentiate core and flexible genes, or flanking core genes were absent. The variable loci were often located in proximity to the prophage borders. Flanking core genes were different between the clusters, although one flanking gene frequently annotated as a transcriptional regulator (Table S5). Despite their small genomes, the variable locus comprised up to 11 genes in a single tailless and up to 6 genes in a single filamentous prophage suggesting selection for these genes.

To understand whether the host might benefit from the variable locus, we investigated the functional gene annotation observing an enrichment for functions related to phage defense. The variable loci we identified in tailless and filamentous prophages across the *Vibrionacaeae* were annotated using databases for both viral and bacterial gene function, complemented with Defensefinder, a tool identifying regions resembling known phage defense systems (46), and a search against the anti-prokaryotic immune systems database (47) (methods). Although most genes in the variable locus remained without functional annotation, 18.6% and 8.5% of the variable loci in tailless and filamentous phage genomes encoded for 27 different putatively complete defense systems (Figure 5b-c). We found a variety of systems associated with abortive infection, such as toxin-antitoxin systems, retrons or Paris, as well as systems degrading nucleic acids, such as restriction modification systems. Most commonly, we observed toxin-antitoxin systems including those where toxin and antitoxin genes were encoded side by side regardless of not being identified as phage defense systems by Defensefinder. Additional to loci encoding for entire defense systems, 20.7% of tailless and 33.1% of filamentous prophages were predicted to contain incomplete defense systems. We found single defense-associated genes that were flanked by genes encoding for putative effector genes that might complete the defense system. For instance, we found a gene predicted as a Thoeris_B sensor next to a hypothetical protein in a tailless prophage, which together could form a defensive sensor-effector pair (Figure 5a). We also observed several accessory genes predicted as counter-defenses targeting CRISPR-Cas and the Thoeris defense system. Our data suggest that prophage-encoded variable loci are under selection and possibly defend both host and its prophage from infecting viruses.

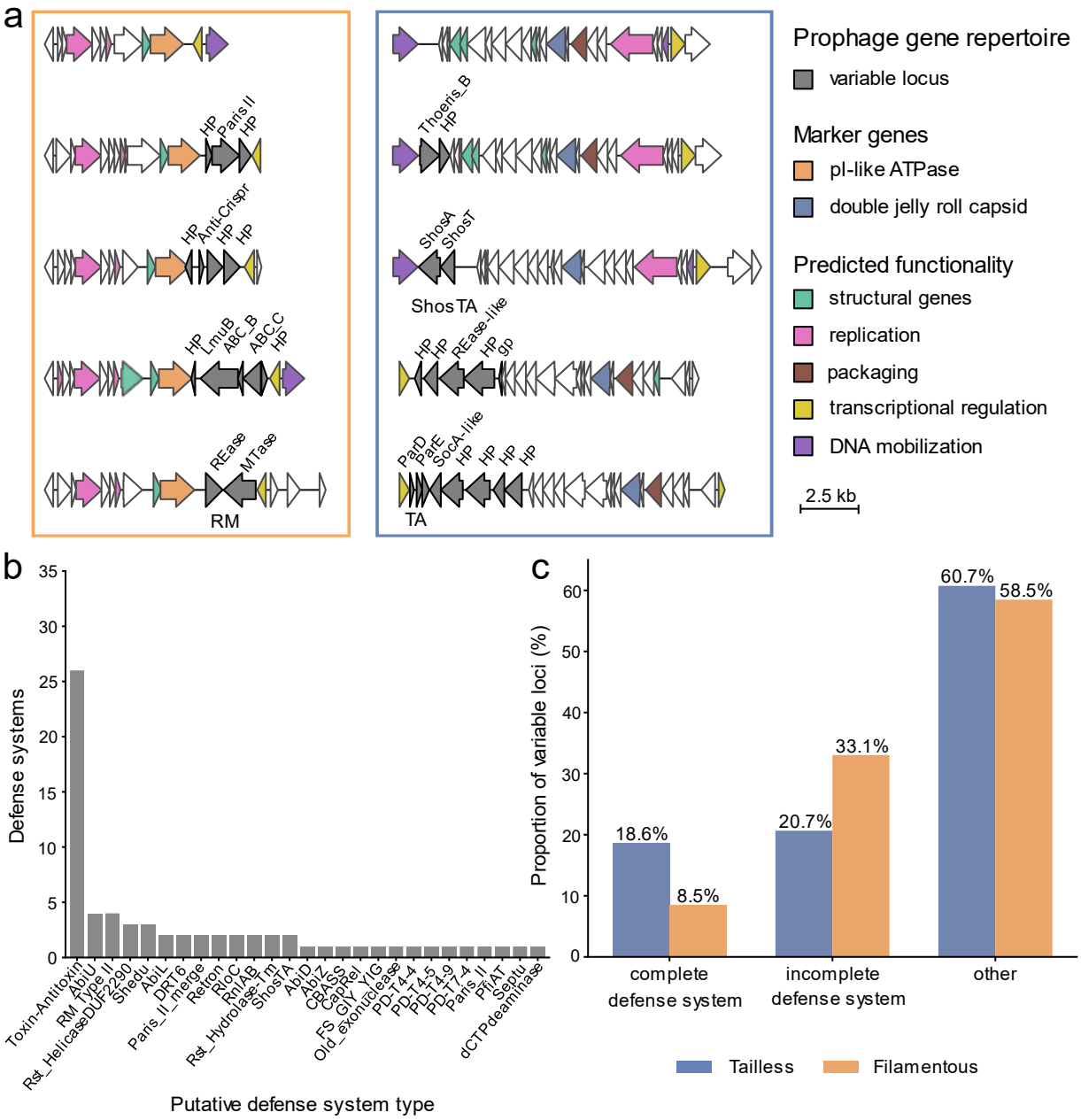

Figure 5 - Accessory gene hotspot in filamentous and tailless prophages in the *Vibrionaceae*
a) Example prophage genomes for filamentous (orange box) and tailless prophages (blue box). Synteny and predicted gene functions are depicted by colored arrows. Proteins within the variable locus are depicted in dark gray with predicted gene functions indicated. DNA mobilization summarizes integrases, excisionases and transposases. White arrows indicate genes, which could not be assigned to a specific function. HP = hypothetical protein, Rease = restriction endonuclease, Mtase = methyltransferase, LmuB = Lamassu_B, ABC = three component system, gp = glycoprotein, RM = restriction modification system, TA = toxin-antitoxin system. Filamentous prophage hosts (top to bottom): 10N.286.55.A12, 10N.239.311.B11, 10N.261.51.C9, 10N.237.312.B10, 10N.239.312.B09. Tailless prophage hosts (top to bottom): 10N.239.312.A01, 10N.286.55.E8, 10N.222.49.B10, 10N.261.45.E12, 1F169. The scale bar shows the sequence length in kb. b) Types of predicted complete defense systems and coupled toxin-antitoxin genes c) Percentage of variable loci with predicted complete defense systems, incomplete defense systems, or other (no defensive genes identified) in filamentous and tailless prophages.

Discussion

Previous studies reported filamentous and tailless prophages across different bacterial taxa suggesting that they are more widespread than previously thought (16–19). In comparison to tailed prophages, their detection remains challenging also due to their underrepresentation in current databases (15,19). By combining chemical induction and comparative genomics for initial prophage candidate prediction instead of using a database-dependent approach, we show that the majority of prophages in naturally co-occurring *Vibrio* are either filamentous or tailless.

Prophage prediction is thought to improve by using high-quality genomes (48). We could confirm higher prophage recovery, however, our comparison of prophage prediction in high-quality and more fragmented genomes revealed that the detection of filamentous prophages is negatively impacted in particular. Possibly caused by several highly similar prophages within the same genome, assembly breaks might lead to filamentous prophage fragments too small to be detected. In addition, bioinformatic workflows introducing prophage detection size thresholds potentially contribute to the underestimation of filamentous and tailless prophages, which are often as small as defective tailed prophages or phage-inducible chromosomal islands (33,49,50).

Filamentous and tailless prophages might balance the negative impact on their host's fitness by recombining site-specifically into a small number of integration sites. The limitation to few integration hot spots could result from counter selection against the disruption of relevant host-encoded genes, as suggested for tailed prophages before (8). In fact, tailless and filamentous prophages seemed to restore the function of genomic regions they integrated into. Although this has been described for filamentous phages targeting the *dif* sites (20) and for mobile genetic elements targeting the *dusA* gene (43), we demonstrate here that tailless phages pursue a similar strategy.

Despite their small genome size, filamentous and tailless prophages frequently carried accessory genes encoded in a defined locus, which were often associated with phage defense. Accessory gene hotspots of tailed prophages have recently been shown to contain phage defense systems (5,6). We observed defense systems and a few counter-defenses encoded on prophages of different viral realms suggesting that phage defense and counter-defense is widespread. Therefore, less-studied tailless and filamentous prophages might harbor many novel defense and counter-defense systems. In addition, accessory genes might confer nucleic acid immunity that provides protection against other mobile genetic elements apart from newly invading phages (51). Prophages might also directly benefit from such accessory genes: We identified numerous toxin-antitoxin systems, that could, similar to the previously investigated toxin-antitoxin system PfiAT (52), improve their replication efficiency. The high variation of accessory genes could result from homologous recombination between closely related co-infecting viruses encoding identical flanking regions of the variable locus (53). Their high variability and possible role in mobile genetic element competition could indicate that accessory genes within the variable locus are under negative-frequency dependent selection. Together, this rapid exchange of the variable loci renders prophages to function as vehicles for phage defense or other virulence genes contributing to the diversification of closely related bacteria.

The variability between prophages of closely related strains indicates the well-known rapid dynamics of prophage gain and loss (9,10). However, whether prophages can successfully lysogenize bacteria phylogenetically distant from their previous host, influences how far their genetic information can travel within bacterial communities. Prophages we identified from naturally co-occurring bacteria were mainly shared within bacterial species boundaries, suggesting that both prophage inheritance and phage-mediated horizontal transfers are often species-specific. This supports previous research stating that most phage-mediated gene transfers take place between closely related bacteria and is in accordance with experimental data demonstrating relatively narrow host ranges of temperate phages (54–56). Recent reports of microbial genetic elements being shared between more distantly related bacteria are challenging this classical view (55). Yet, this genetic barrier might be special to phages because they require specific receptors to bind their hosts (51).

Although this study is limited to prophages in the family *Vibrionaceae*, previous research suggests that the impact of tailless and filamentous prophages also extends to bacterial communities outside of the *Vibrionaceae*. In particular, tailless and filamentous prophages have been reported in various environments and bacterial taxa (16–18). Moreover, the integration sites we identified for tailless and filamentous phages are conserved in a wide range of bacteria (43,57), possibly enabling related prophages to integrate into other bacterial hosts. In conclusion, we here provide a large dataset of 1,158 prophages including tailless and filamentous prophages, that we propose to be systematically underestimated in prophage diversity studies. We further show that filamentous, tailless, and tailed prophages are evenly distributed and might play an equally important role in driving the diversification of closely related bacteria beyond the *Vibrionaceae*.

Acknowledgements

We thank all members of the Joint Microbiome facility for sequencing, especially Petra Pjevac for discussion of sample preparation and Rasmus Kirkegaard for advice on long read nanopore sequencing. We want to acknowledge the former Polz lab members Kathryn Kauffman, Joseph Elsherbini and Phil Arevalo for isolation and sequencing of bacterial strains used in this study. We further would like to thank Rachel Szabo and Carolin Wendling for advice on handling prophages. We want to thank Anna Lopatina, Jacob Bobonis, Ben Roller, and Shaul Pollak for valuable discussions and Michael Wolfram for his assistance in the lab. We thank Thomas Rattei and Lovro Trgovec-Greif for advice on the computational analysis. The computational results of this work have been generated using the Life Science Compute Cluster (LiSC) of the University of Vienna. This work was supported by the Simons Foundation (Life Sciences Project Award-572792) and by the Austrian Science Fund (FWF) [doi.org/10.55776/COE7].

Conflict of Interest

None declared.

Data availability

Reference host reads and genomes were deposited with the accession numbers indicated in Table S2. Genomes of the 107 well-curated prophages are available under the GenBank accession numbers PQ189284 - PQ189390 (Table S2). Reads for MMC sequenced cultures were submitted to the SRA under Bioproject number PRJNA1137887 (Table S2).

References

1. Pfeifer E, Bonnin RA, Rocha EPC. Phage-plasmids spread antibiotic resistance genes through infection and lysogenic conversion. *mBio*. 2022 Oct 26;13(5):e01851-22.

2. Waldor MK, Mekalanos JJ. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science*. 1996 Jun 28;272(5270):1910–4.

3. Bondy-Denomy J, Qian J, Westra ER, Buckling A, Guttman DS, Davidson AR, et al. Prophages mediate defense against phage infection through diverse mechanisms. *ISME J*. 2016 10:12. 2016 Jun 3;10(12):2854–66.

4. Owen SV, Wenner N, Dulberger CL, Rodwell EV, Bowers-Barnard A, Quinones-Olvera N, et al. Prophages encode phage-defense systems with cognate self-immunity. *Cell Host Microbe*. 2021 Nov;29(11):1620-1633.e8.

5. Rousset F, Depardieu F, Miele S, Dowding J, Laval AL, Lieberman E, et al. Phages and their satellites encode hotspots of antiviral systems. *Cell Host Microbe*. 2022 May;30(5):740-753.e5.

6. Vassallo CN, Doering CR, Littlehale ML, Teodoro GIC, Laub MT. A functional selection reveals previously undetected anti-phage defence systems in the *E. coli* pangenome. *Nat Microbiol.* 2022 Sep 19;7(10):1568–79.

7. Patel PH, Taylor VL, Zhang C, Getz LJ, Fitzpatrick AD, Davidson AR, et al. Anti-phage defence through inhibition of virion assembly. *Nat Commun*. 2024 Feb 22;15(1):1644.

8. Bobay LM, Rocha EPC, Touchon M. The adaptation of temperate bacteriophages to their host genomes. *Mol Biol Evol*. 2013;30(4):737–51.

9. Croucher NJ, Mostowy R, Wymant C, Turner P, Bentley SD, Fraser C. Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. *PLoS Biol*. 2016 Mar 2;14(3):e1002394.

10. Hussain FA, Dubert J, Elsherbini J, Murphy M, VanInsberghe D, Arevalo P, et al. Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science*. 2021;374(6566):488–92.

11. Greenrod STE, Stoycheva M, Elphinstone J, Friman VP. Global diversity and distribution of prophages are lineage-specific within the *Ralstonia solanacearum* species complex. *BMC Genomics*. 2022 Oct 6;23(1):689.

12. Johnson G, Banerjee S, Putonti C. Diversity of *Pseudomonas aeruginosa* temperate phages. *mSphere*. 2022 Feb 23;7(1):e01015-21.

13. Jarocki P, Komoń-Janczara E, Młodzińska A, Sadurski J, Kołodzińska K, Łaczmański Ł, et al. Occurrence and genetic diversity of prophage sequences identified in the genomes of *L. casei* group bacteria. *Sci Rep*. 2023 May 26;13(1):8603.

14. Sharma V, Hünnefeld M, Luthe T, Frunzke J. Systematic analysis of prophage elements in actinobacterial genomes reveals a remarkable phylogenetic diversity. *Sci Rep*. 2023 Mar 17;13(1):4410.

15. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*. 2021 Dec;9(1):37.

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

17. Roux S, Krupovic M, Daly RA, Borges AL, Nayfach S, Schulz F, et al. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat Microbiol*. 2019 Jul 22;4(11):1895–906.

18. Yutin N, Rayko M, Antipov D, Mutz P, Wolf YI, Krupovic M, et al. Varidnaviruses in the human gut: A major expansion of the order *Vinavirales*. *Viruses*. 2022 Aug 23;14(9):1842.

19. Kalatzis PG, Mauritzen JJ, Winther-Have CS, Michniewski S, Millard A, Tsertou MI, et al. Staying below the radar: Unraveling a new family of ubiquitous "cryptic" non-tailed temperate vibriophages and implications for their bacterial hosts. *IJMS*. 2023 Feb 15;24(4):3937.

20. Rakonjac J, Bennet NJ, Spagnuolo J, Gagic D, Russell M. Filamentous bacteriophage: Biology, phage display and nanotechnology applications. *Curr Issues Mol Biol*. 2011;13: 51-76.

21. Kauffman KM, Hussain FA, Yang J, Arevalo P, Brown JM, Chang WK, et al. A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature*. 2018;554(7690):118–22.

22. Leigh BA, Breitbart M, Oksanen HM, Bamford DH, Dishaw LJ. Genome sequence of PM2-like phage Cr39582, induced from a *Pseudoalteromonas sp.* isolated from the gut of *Ciona robusta*. *Genome Announc*. 2018 May 24;6(21):e00368-18.

23. Castillo D, Kauffman K, Hussain F, Kalatzis P, Rørbo N, Polz MF, et al. Widespread distribution of prophage-encoded virulence factors in marine *Vibrio* communities. *Scientific Reports*. 2018;8(1):2–10.

24. Nawel Z, Rima O, Amira B. An overview on *Vibrio* temperate phages: Integration mechanisms, pathogenicity, and lysogeny regulation. *Microb Pathog*. 2022 Apr;165:105490.

25. Zünd M, Ruscheweyh HJ, Field CM, Meyer N, Cuenca M, Hoces D, et al. High throughput sequencing provides exact genomic locations of inducible prophages and accurate phage-to-host ratios in gut microbial strains. *Microbiome*. 2021;9(1):1–18.

26. López-Leal G, Camelo-Valera LC, Hurtado-Ramírez JM, Verleyen J, Castillo-Ramírez S, Reyes-Muñoz A. Mining of thousands of prokaryotic genomes reveals high abundance of prophages with a strictly narrow host range. *mSystems*. 2022 Aug 30;7(4):e00326-22.

27. Koonin EV, Kuhn JH, Dolja VV, Krupovic M. Megataxonomy and global ecology of the virosphere. *ISME J*. 2024 Jan 10;wrad042.

28. Meier-Kolthoff JP, Göker M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics*. 2017 Nov 1;33(21):3396–404.

29. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. *Nat Biotechnol*. 2024 Aug;42(8):1303–12.

30. Quinones-Olvera N, Owen SV, McCully LM, Marin MG, Rand EA, Fan AC, et al. Diverse and abundant phages exploit conjugative plasmids. *Nat Commun*. 2024 Apr 12;15(1):3197.

31. Nakasone N, Honma Y, Toma C, Yamashiro T, Iwanaga M. Filamentous phage fs1 of *Vibrio cholerae* O139. *Microbiology and Immunology*. 1998 Mar;42(3):237–9.

32. Davis BM, Waldor MK. CTXϕ contains a hybrid genome derived from tandemly integrated elements. *Proc Natl Acad Sci USA*. 2000 Jul 18;97(15):8572–7.

33. Bobay LM, Touchon M, Rocha EPC. Pervasive domestication of defective prophages by bacteria. *Proc Natl Acad Sci USA*. 2014;111(33):12127–32.

34. Tomasz M. Mitomycin C: small, fast and deadly (but very selective). *Chem Biol*. 1995;2(9):575–9.

35. Refardt D. Within-host competition determines reproductive success of temperate bacteriophages. *ISME J*. 2011 Sep 1;5(9):1451–60.

36. Guo Q, Chen B, Tu Y, Du S, Chen X. Prophage LambdaSo uses replication interference to suppress reproduction of coexisting temperate phage MuSo2 in *Shewanella oneidensis* MR-1. *Environ Microbiol*. 2019 Jun;21(6):2079–94.

37. Geider K, Kornberg A. Conversion of the M13 viral single strand to the double-stranded replicative forms by purified proteins. *J Biol Chem*. 1974;249(13):3999–4005.

38. Hay ID, Lithgow T. Filamentous phages: masters of a microbial sharing economy. *EMBO Rep*. 2019 Jun;20(6):e47427.

39. Meibom KL, Blokesch M, Dolganov NA, Wu CY, Schoolnik GK. Chitin induces natural competence in *Vibrio cholerae*. *Science*. 2005 Dec 16;310(5755):1824–7.

40. Lo Scrudato M, Blokesch M. The regulatory network of natural competence and transformation of *Vibrio cholerae*. *PLoS Genet*. 2012 Jun 21;8(6):e1002778.

41. Das B, Bischerour J, Barre FX. Molecular mechanism of acquisition of the cholera toxin genes. *Indian J Med Res*. 2011;133(2):195–200.

42. Huber KE, Waldor MK. Filamentous phage integration requires the host recombinases *XerC* and *XerD*. *Nature*. 2002 Jun;417(6889):656–9.

43. Farrugia DN, Elbourne LDH, Mabbutt BC, Paulsen IT. A novel family of integrases associated with prophages and genomic islands integrated within the tRNA-dihydrouridine synthase A (*dusA*) gene. *Nucleic Acids Res*. 2015 May 19;43(9):4547–57.

44. Cazares A, Mendoza-Hernández G, Guarneros G. Core and accessory genome architecture in a group of *Pseudomonas aeruginosa* Mu-like phages. *BMC Genomics*. 2014 Dec;15(1):1146.

45. Mai-Prochnow A, Hui JGK, Kjelleberg S, Rakonjac J, McDougald D, Rice SA. 'Big things in small packages: the genetics of filamentous phage and effects on fitness of their host.' *FEMS Microbiol Rev*. 2015 Jul;39(4):465–87.

46. Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun*. 2022 May 10;13(1):2561.

47. Yan Y, Zheng J, Zhang X, Yin Y. dbAPIS: a database of anti-prokaryotic immune system genes. *Nucleic Acids Res*. 2024;52(D1):D419–25.

48. Song W, Sun HX, Zhang C, Cheng L, Peng Y, Deng Z, et al. Prophage Hunter: an integrative hunting tool for active prophages. *Nucleic Acids Res*. 2019;47(W1):W74–80.

49. Touchon M, Bernheim A, Rocha EPC. Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J*. 2016;10(11):2744–54.

50. Barcia-Cruz R, Goudenège D, Moura De Sousa JA, Piel D, Marbouty M, Rocha EPC, et al. Phage-inducible chromosomal minimalist islands (PICMIs), a novel family of small marine satellites of virulent phages. *Nat Commun*. 2024 Jan 22;15(1):664.

51. Haudiquet M, De Sousa JM, Touchon M, Rocha EPC. Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Phil Trans R Soc B*. 2022 Oct 10;377(1861):20210234.

52. Li Y, Liu X, Tang K, Wang W, Guo Y, Wang X. Prophage encoding toxin/antitoxin system *PfiT*/*PfiA* inhibits Pf4 production in *Pseudomonas aeruginosa*. *Microb Biotechnol*. 2020 Jul;13(4):1132–44.

53. Kauffman KM, Chang WK, Brown JM, Hussain FA, Yang J, Polz MF, et al. Resolving the structure of phage–bacteria interactions in the context of natural diversity. *Nat Commun*. 2022 13:1. 2022 Jan 18;13(1):1–20.

54. Popa O, Landan G, Dagan T. Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. *ISME J*. 2017 Feb 1;11(2):543–54.

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

56. Wendling CC, Goehlich H, Roth O. The structure of temperate phage–bacteria infection networks changes with the phylogenetic distance of the host bacteria. *Biol Lett*. 2018;14(11):20180320.

57. Carnoy C, Roten CA. The dif/Xer Recombination systems in *Proteobacteria*. *PLoS ONE*. 2009 Sep 3;4(9):e6531.

58. Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science*. 2008 May 23;320(5879):1081–5.

59. Preheim SP, Boucher Y, Wildschutte H, David LA, Veneziano D, Alm EJ, et al. Metapopulation structure of *Vibrionaceae* among coastal marine invertebrates. *Environ Microbiol*. 2011;13(1):265–75.

60. Martin-Platero AM, Cleary B, Kauffman K, Preheim SP, McGillicuddy DJ, Alm EJ, et al. High resolution time series reveals cohesive but short-lived communities in coastal plankton. *Nat Commun*. 2018 Jan 18;9(1):266.

61. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, et al. The khmer software package: enabling efficient nucleotide sequence analysis. *F1000 Res*. 2015 Sep 25;4:900.

62. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes de novo assembler. *Curr Protoc Bioinformatics*. 2020;70(1):e102.

63. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*. 2011 Feb 1;27(3):334–42.

64. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010;59(3):307–21.

65. Didelot X, Wilson DJ. ClonalFrameML: Efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol*. 2015 Feb 12;11(2):e1004041.

66. Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, et al. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput Biol*. 2020 Mar 19;16(3):e1007732.

67. Bushnell B. BBMap: A Fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States); 2014 Mar. Report No.: LBNL-7065E.

68. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009 Jul 15;25(14):1754–60.

69. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021 Jan 29;10(2):giab008.

70. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010 Dec;11(1):119.

71. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol*. 2011;7(10):e1002195.

72. Trgovec-Greif L, Hellinger HJ, Mainguy J, Pfundner A, Frishman D, Kiening M, et al. VOGDB—Database of virus orthologous groups. *Viruses*. 2024 Jul 25;16(8):1191.

73. Gilchrist CLM, Chooi YH. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics*. 2021 Aug 25;37(16):2473–5.

74. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990 Oct 5;215(3):403–10.

75. Néron B, Denise R, Coluzzi C, Touchon M, Rocha EPC, Abby SS. MacSyFinder v2: Improved modelling and search engine to identify molecular systems in genomes. *Peer Community J*. 2023 Mar ;3:e28.

76. The UniProt Consortium. UniProt: the Universal protein knowledgebase in 2023. *Nucleic Acids Res*. 2023 Jan 6;51(D1):D523–31.

77. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014 May 1;30(9):1236–40.

78. Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*. 2013 Dec;14(1):60.

79. Lefort V, Desper R, Gascuel O. FastME 2.0: A Comprehensive, accurate, and fast distance-based phylogeny inference program: Table 1. *Mol Biol Evol*. 2015 Oct;32(10):2798–800.

80. Moraru C, Varsani A, Kropinski AM. VIRIDIC — A novel tool to calculate the intergenomic similarities of viruses. *Viruses*. 2020;12:1268.

81. Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie M, et al. INfrastructure for a PHAge REference database: identification of large-scale biases in the current collection of cultured phage genomes. *Phage*. 2021;2(4):214–23.

82. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. ViPTree: the viral proteomic tree server. *Bioinformatics*. 2017;33(15):2379–80.

83. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. *Mol Biol Evol*. 2013 Apr 1;30(4):772–80.

84. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS ONE*. 2010 Mar 10;5(3):e9490.

85. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*. 2024 Jun 13;630(8016):493–500.

86. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci*. 2021 Jan 1;30(1):70–82.

87. Abrescia NGA, Grimes JM, Kivela HK, Assenberg R, Sutton GC, Butcher SJ, et al. Crystal structure of the major capsid protein P2 from Bacteriophage PM2. *PDB*. 2008 Sep 16;

88. Abrescia NGA, Grimes JM, Kivelä HM, Assenberg R, Sutton GC, Butcher SJ, et al. Insights into virus evolution and membrane biogenesis from the structure of the marine lipid-containing bacteriophage PM2. *Mol Cell*. 2008 Sep;31(5):749–61.

89. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. 2018;9(1):5114.

90. Moraru C. VirClust—A Tool for hierarchical clustering, core protein detection and annotation of (prokaryotic) viruses. *Viruses*. 2023 Apr 19;15(4):1007.

91. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012 Dec 1;28(23):3150–2.

92. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 2002;30(7):1575–84.

93. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, et al. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res*. 2021 Jan 8;49(D1):D1020–8.

94. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. 2021 Jul 2;49(W1):W293–6.

95. Yu G. Using ggtree to visualize data on tree-like structures. *Curr protoc Bioinformatics*. 2020;69(1):e96.

96. Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics*. 2011 Feb 15;27(4):592–3.

97. Wickham H. ggplot2: elegant graphics for data analysis. Second edition. Switzerland: Springer; 2016. 260 p. (Use R!).

Supplementary Methods

Supplementary Tables S1-S5 are available online.

High quality reference genomes of *Vibrio cyclitrophicus*

Considering that closed genomes are beneficial for prophage discovery, we sequenced 58 selected bacterial isolates using Oxford Nanopore technology (ONT) (Table S2). Bacteria were streaked on 1.5 % agar plates and incubated overnight. Single colonies were transferred into 7 ml liquid 2216MB and grown in a rotator at medium speed ON. Cells were pelleted at 12,000 x g and 4°C until the supernatant cleared (10-40 min).

DNA was extracted from bacterial pellets using the Power Soil Pro Kit (Qiagen) and sequenced in two rounds ("N01" and "N02"). In Brief, the DNA was prepared for sequencing using the rapid barcoding sequencing kit (SQK-RBK110.96, Oxford Nanopore Technologies) following the manufacturer's protocol. Barcoded samples were sequenced in R9.4.1 flowcells on a MinION Mk1b (N01, FLO-MIN106D, Oxford Nanopore Technologies) and on a Promethion P24 (N02, FLO-PRO002, Oxford Nanopore Technologies). The DNA sequencing was carried out using Minknow (v. 22.03.6, Oxford Nanopore Technologies). Raw ONT reads were basecalled using Guppy (v. 6.1.1) in super accuracy mode and assembled using flye (v. 2.9-b1768, (1)) with "–nano-hq". The Illumina reads for polishing that were obtained as described in "*Vibrionaceae* genome collection" were trimmed using cutadapt (v. 3.1, (2)). The assembly was polished three times using short read Illumina data with minimap2 (v. 2.17, (3)) and racon (v. 1.4.3, (4)) and two times using long-reads with medaka (v. 1.6.1, github.com/nanoporetech/medaka). Reads were mapped to the assemblies using minimap2 (v. 2.17, (3)), read mappings were converted using samtools (v. 1.12, (5)) and read coverage was calculated using metabat2 (v. 2.15, (6)). If the polished assemblies contained contigs smaller than 500 bp, they were excluded from this study.

We resequenced genomes not closed after the first runs ("N03"): High molecular weight DNA was extracted with the Monarch HMW DNA extraction kit for tissue (New England Biolabs) with the protocol for Gram negative bacteria. Extracts were left at 4°C for at least one week before quantification and fragment estimation. Samples were diluted and equimolarly barcoded using the SQK-RBK114.96 (Oxford Nanopore Technologies) with the following modifications: the sample input was increased to >80ng/sample, added 0.2ul barcode/sample, and added 0.5ul of rapid adaptor to the barcoded library. About 20 fmol of a >60 kb library was loaded on a R10.4.1 flowcell (FLO-PRO114, Oxford Nanopore Technologies) and sequenced for 48 h on a Promethion P2 solo (Oxford Nanopore Technologies) using Minknow (v. 23.11.4, Oxford Nanopore Technologies). Flowcell light shields were used. Reads were basecalled using either Dorado (v. 7.8.2) or Guppy (v. 6.1.1) using super accuracy mode (Table S2). Raw reads shorter than 500bp and with average Q-scores <15 were removed using chopper (v0.6, (7)) and assembled using flye (v. 2.9.1, (1)) with "–nano-hq". The assembly was polished once with medaka (v. 1.11.1, github.com/nanoporetech/medaka) using model "r1041_e82_400bps_sup_g615". Contigs smaller than 1kb were excluded from the polished assemblies. Genomes were manually inspected and depending on the overall coverage of the main contigs, lower coverage contigs (either <10x or <30x) were removed. The genome quality, completeness and contamination values were checked using QUAST (v. 5.0.2, (8)) and CheckM (v. 1.1.1, (9)).

Supplementary Figures

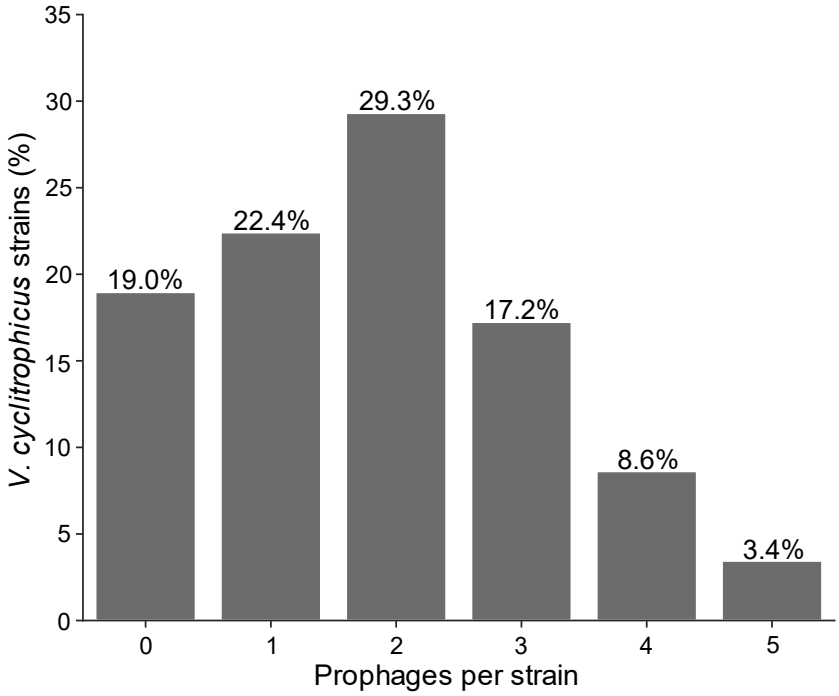

Figure S1 - Prophage distribution in *Vibrio cyclitrophicus*
Histogram showing the distribution of prophages carried per bacterial strain in 58 *Vibrio cyclitrophicus* genomes. Strains carried 1.8 prophages on average.

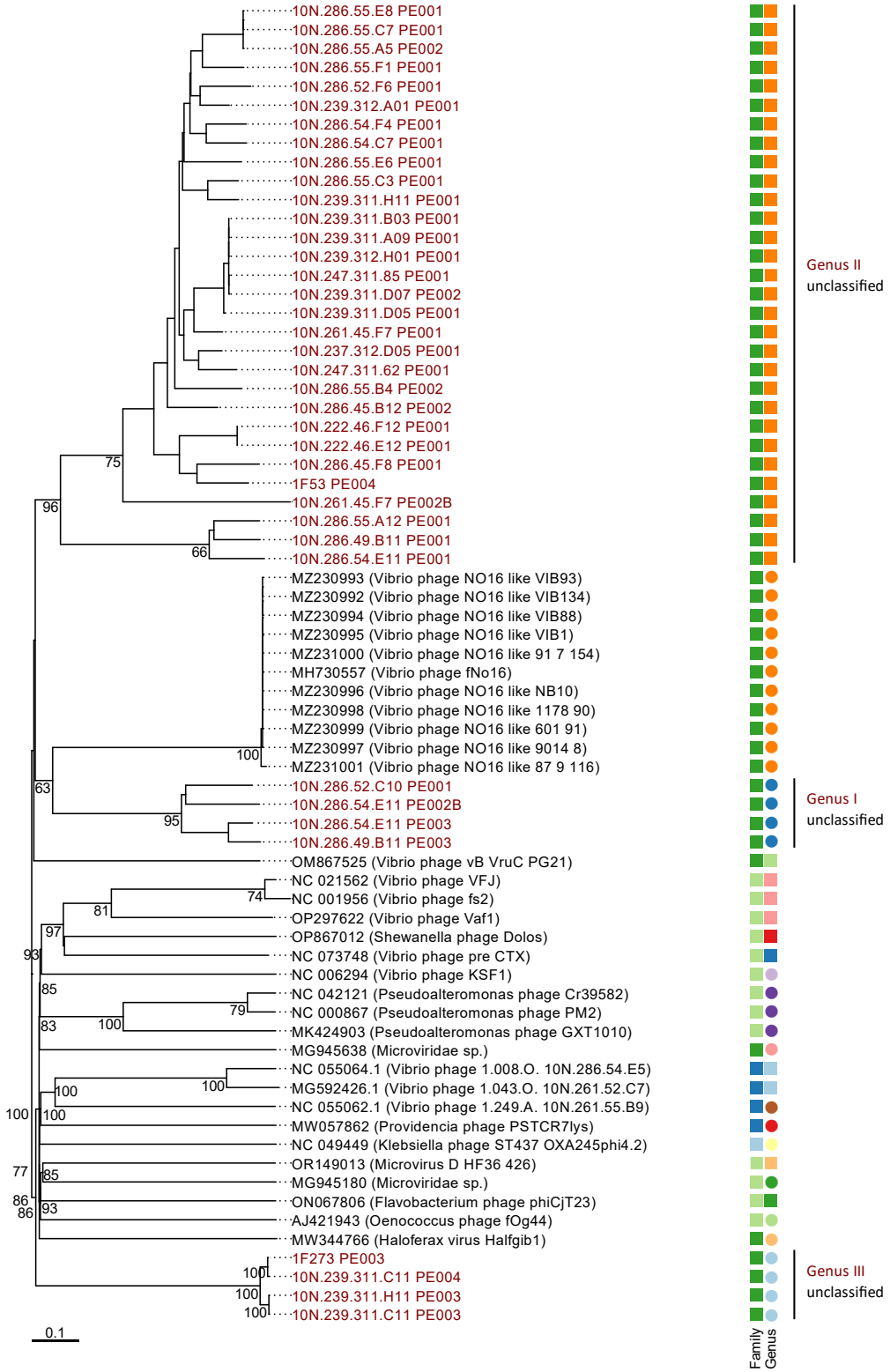

Figure S2a - Classification of tailless prophages.
Balanced minimum evolution tree inferred from intergenomic distances between genome sequences of identified (a) tailless (b) filamentous and (c) tailed prophages (red) and selected known tailless phages. Pairwise distances were calculated using the Genome-BLAST Distance Phylogeny formula D0. Values at the branches denote the support from 100 pseudo-bootstraps. Taxonomic classification was predicted using VICTOR and is represented by colored symbols and the number of the respective genus indicated. The scale bar represents the number of nucleotide substitutions per site.

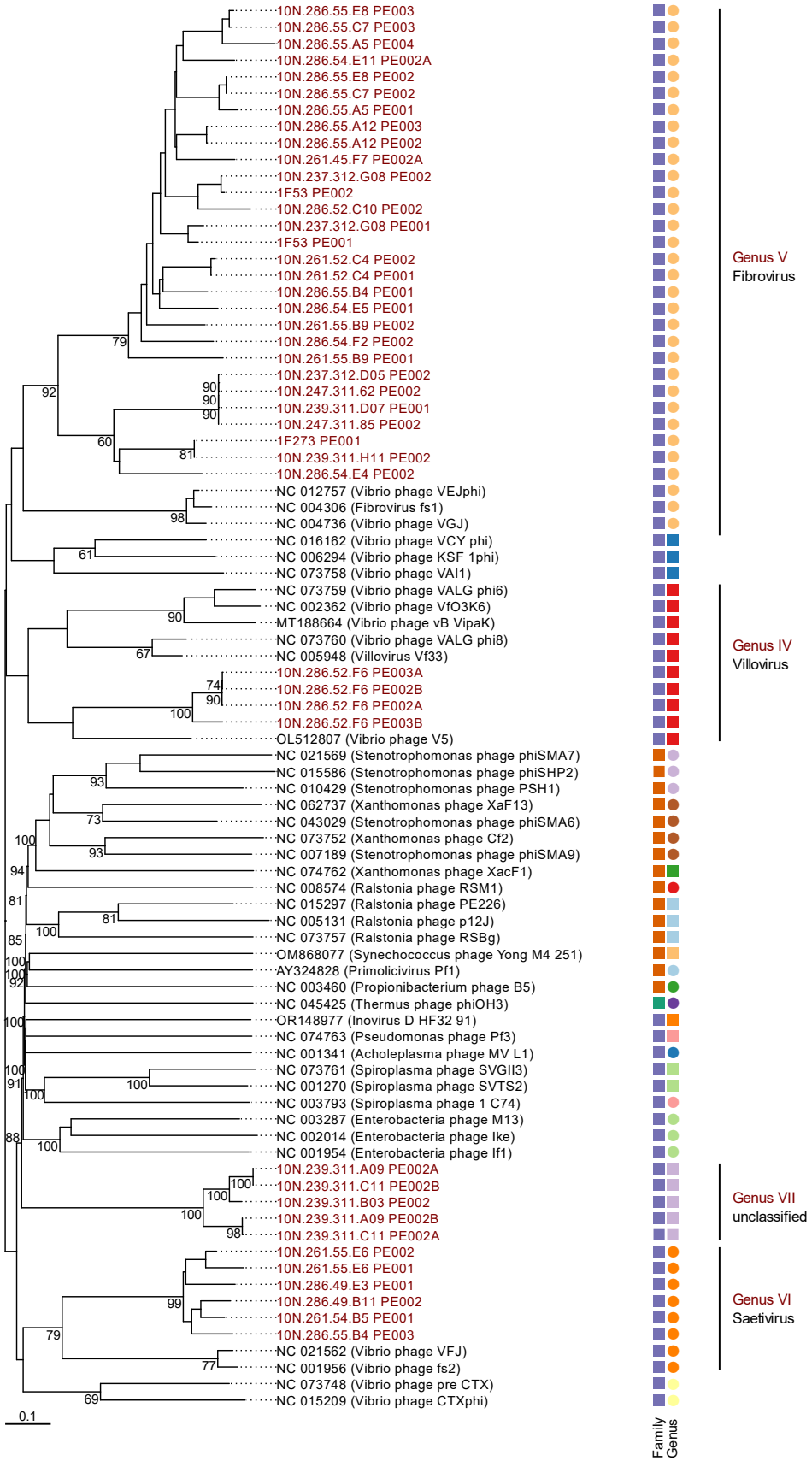

Figure S2b - Classification of filamentous prophages.

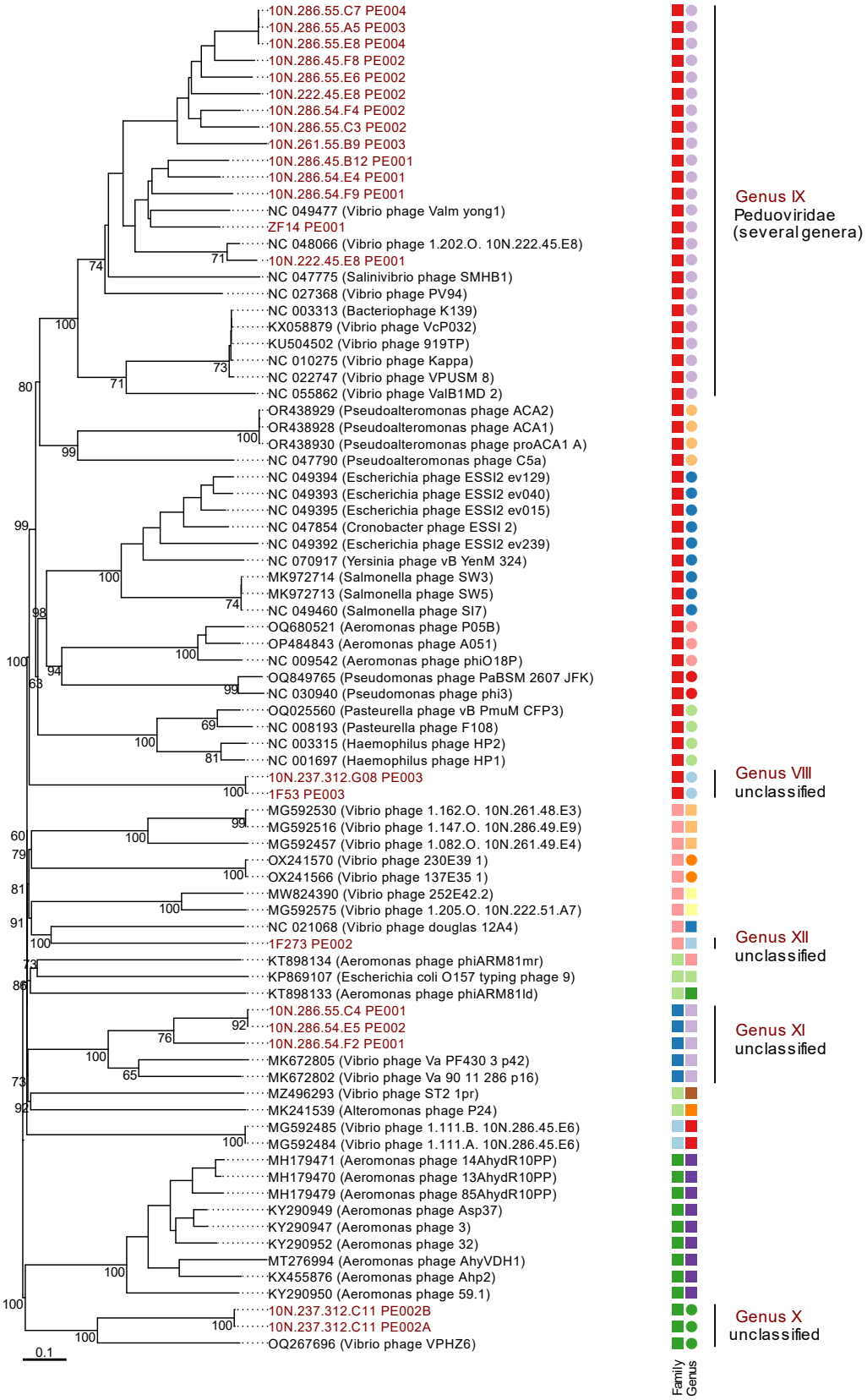

Figure S2c - Classification of tailed prophages.

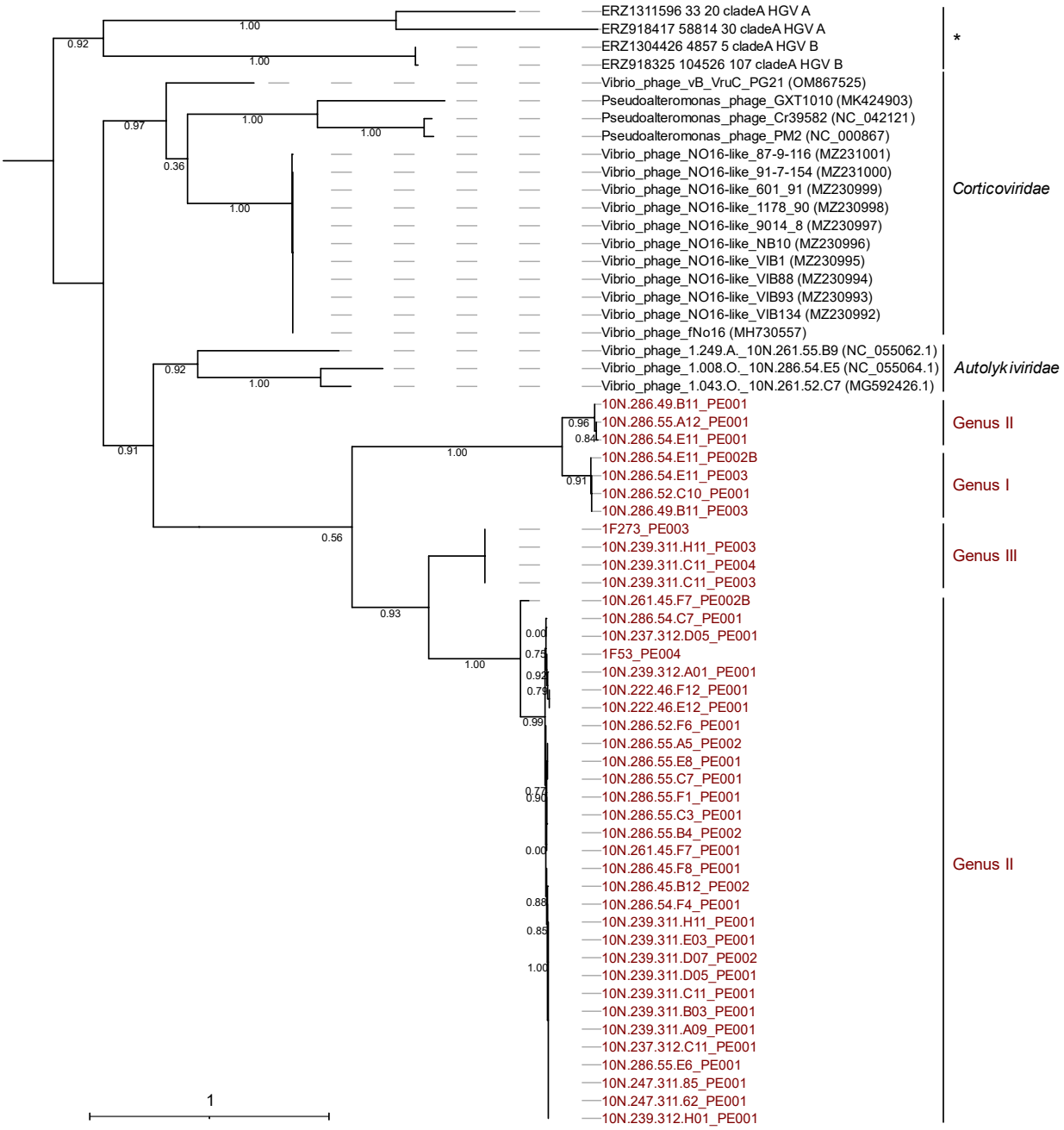

Figure S3a - Marker gene comparison to known tailless phages

Phylogenetic tree based on the alignment of double jelly-roll major capsid proteins of identified prophages and known representatives of the *Vinavirales and Autolykiviridae*. Prophages identified in this study are highlighted in red. Additional double jelly-roll major capsid proteins marked with an asterisk originate from major capsid proteins identified human gut genomes ("clade A") in a previous study (Yutin et al., 2022). The scale bar represents amino acid substitutions per site. Bootstrap values are given in a range from 0 to 1.

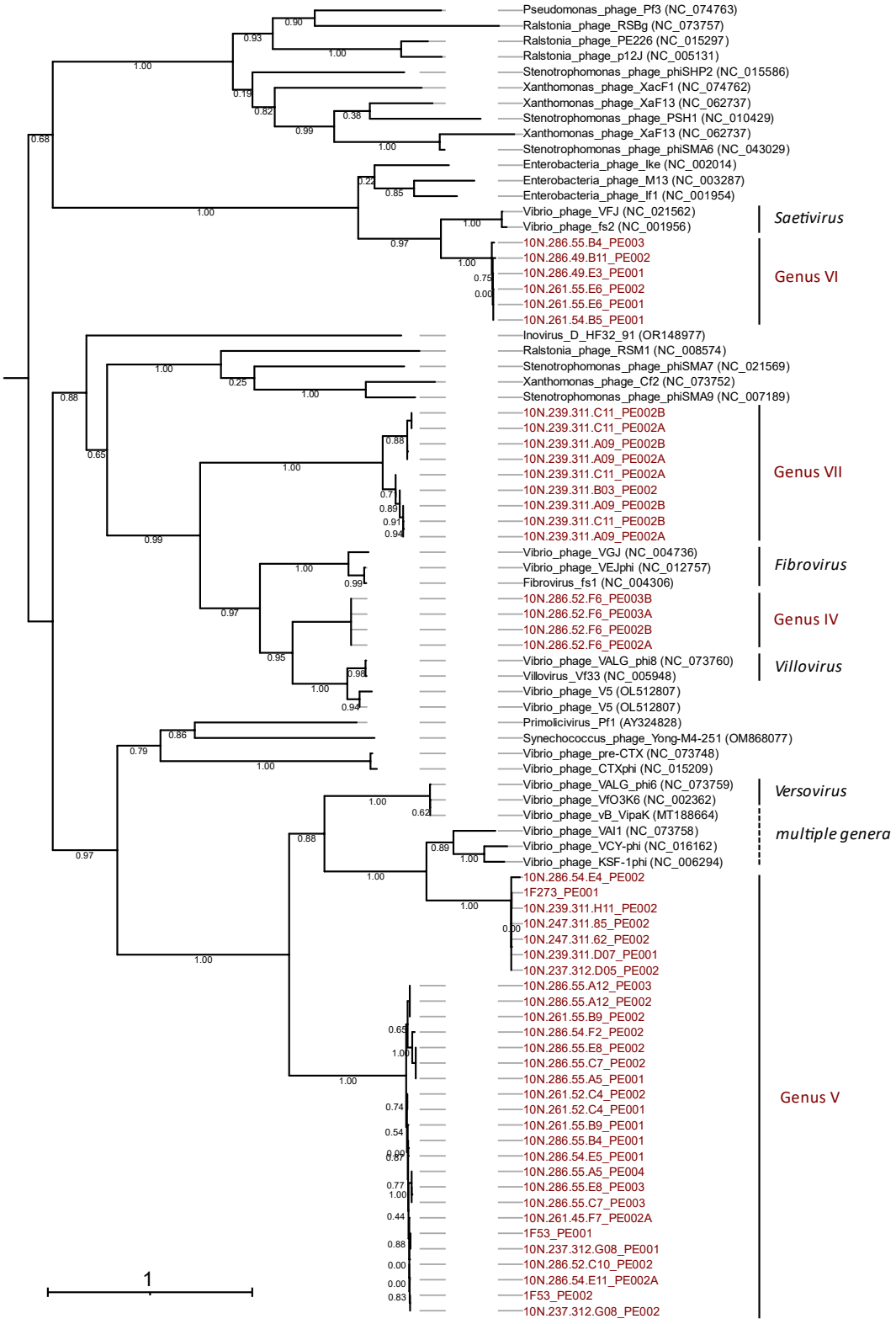

Figure S3b - Marker gene comparison to known filamentous phages

Phylogenetic tree based on the alignment of the pI-like ATPase genes of identified prophages and known representatives of the *Inoviridae*. If two copies of pI-like ATPases are present, both are shown. Prophages identified in this study are highlighted in red. Known filamentous phages represent the majority of genera currently present in the ICTV taxonomy. The scale bar represents amino acid substitutions per site. Bootstrap values are given in a range from 0 to 1.

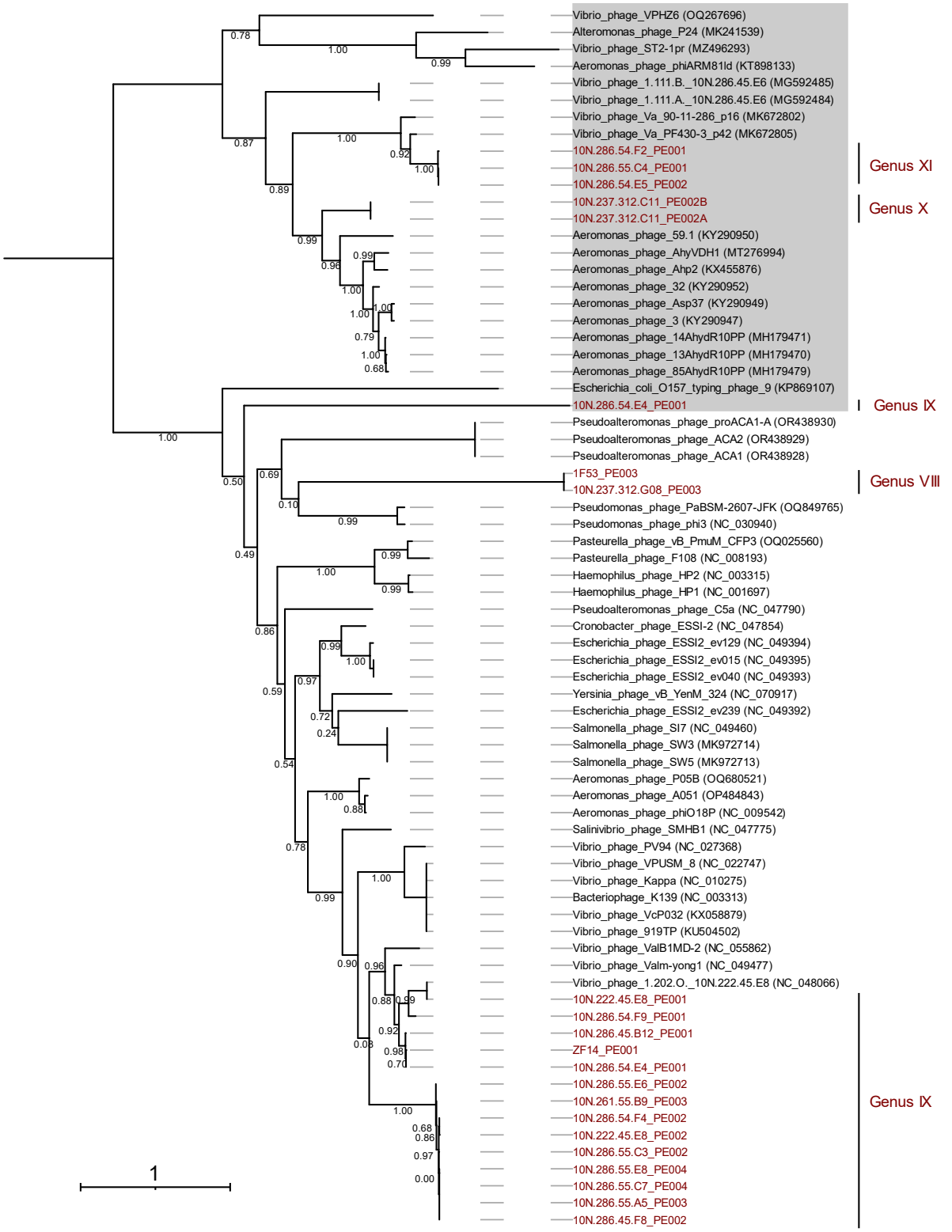

Figure S3c - Marker gene comparison to known tailed phages

Phylogenetic tree based on the alignment of the large subunit (gray box) or endonuclease subunit of terminase genes of identified prophages and known representatives of the *Caudoviricetes*. If two terminase copies are present, both are shown. Prophages identified in this study are highlighted in red. Tailed phages without a clear terminase prediction are excluded. The scale bar represents amino acid substitutions per site. Bootstrap values are given in a range from 0 to 1.

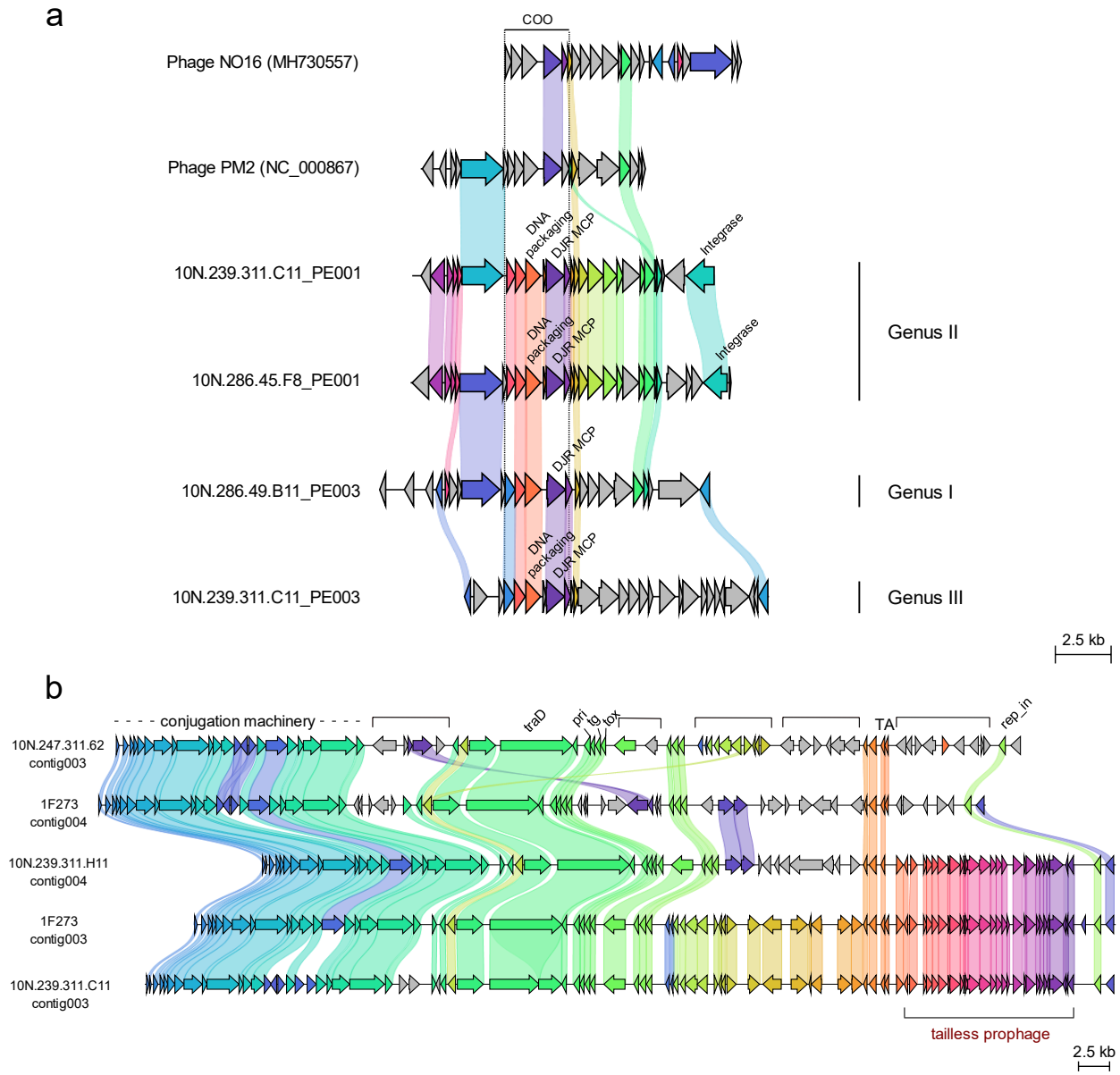

Figure S4 - Genomic maps of tailless phage genomes and plasmids.
a) Representative genomes of tailless phage genera I-III are shown in comparison to known tailless phages PM2 and NO16. Gene annotations for the double jelly-roll major capsid protein (DJR MCP), the DNA packaging ATPase and Integrase genes were added if applicable. The gene order of the identified phages from Genera I-II partially resembled the ones of PM2 and NO16, for example in the marked region with conserved gene order (CGO). b) Plasmids carrying tailless prophages of genus III and plasmids with a similar backbone identified in V. cyclitrophicus. Core gene annotations denote transfer gene D (traD), a primase (pri), transglycosylase (tg), a toxin (tox), a toxin-antitoxin pair (TA) and a replication initiation protein (rep_in). Flexible regions are marked by brackets. In both plots, genes are depicted as arrows, and genes with amino acid identity above 30% are colored and connected. The scale bar shows the sequence length in kb.

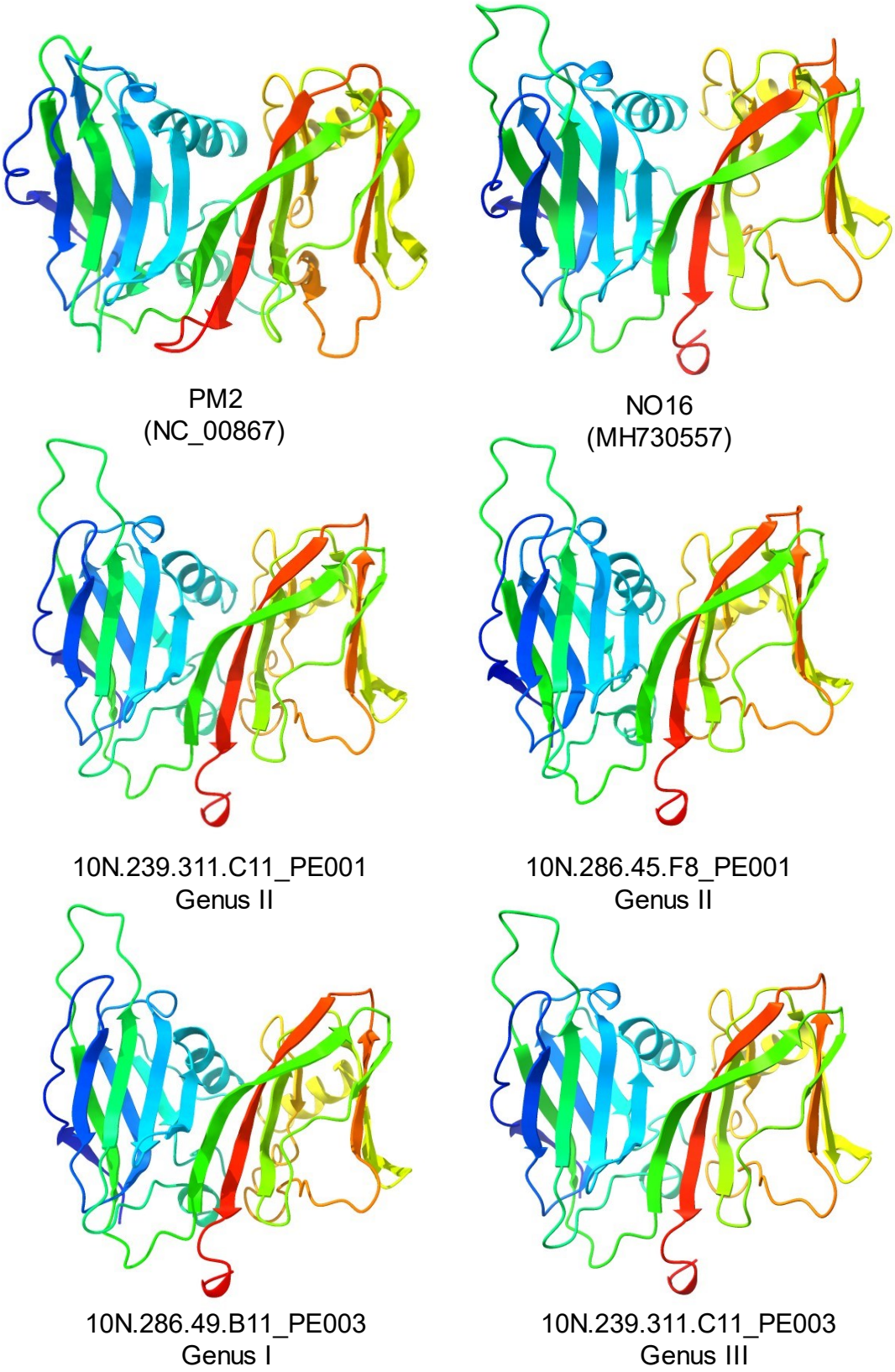

Figure S5 - Structural conservation of double jelly-roll major capsid proteins encoded by tailless phages. Monomers of the crystal structure of the major capsid protein of phage PM2, and structural predictions of the monomers of major capsid proteins obtained by alphafold for phage NO16 and four representative tailless phages from genera I-III.

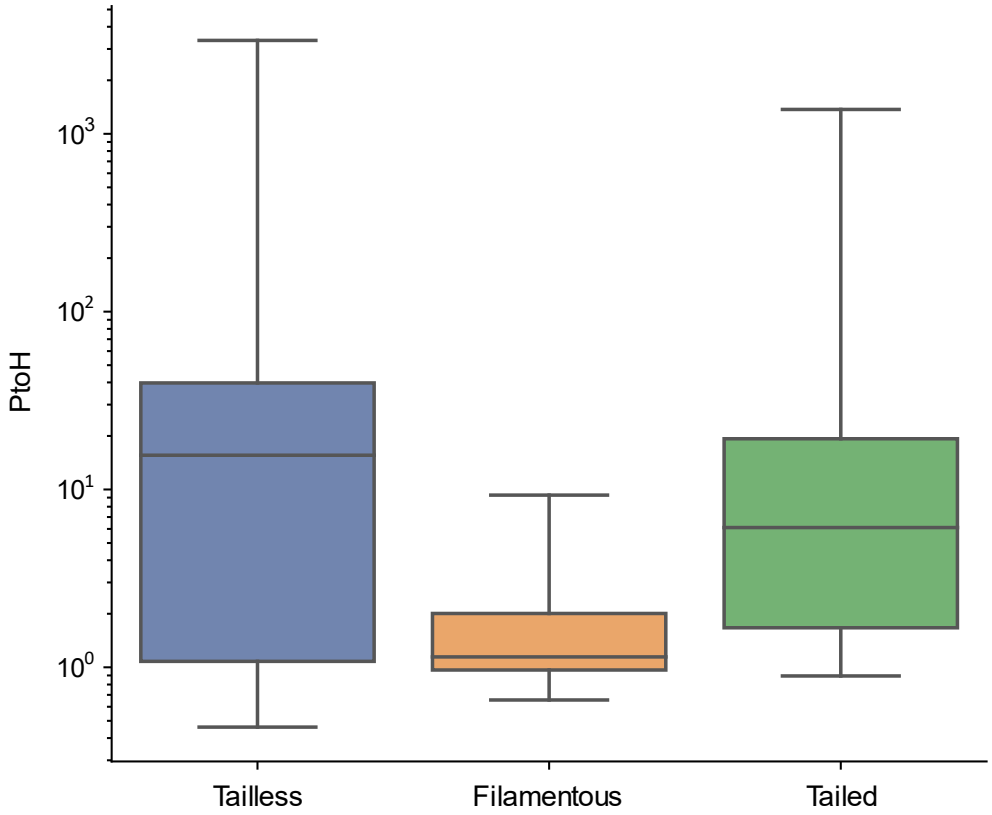

Figure S6 - Prophage activity
Boxplots show the relative coverage of prophages compared to their genomic neighborhood given with the phage-to-host ratio (PtoH) as a measure of the prophage activity. Each boxplot represents the phage-to-host ratios of prophages per viral realm on a logarithmic scale. Whiskers show the range between the minimum and maximum values, and box edges represent the lower and upper quartile of the phage-to-host ratios. The median is given by a horizontal line.

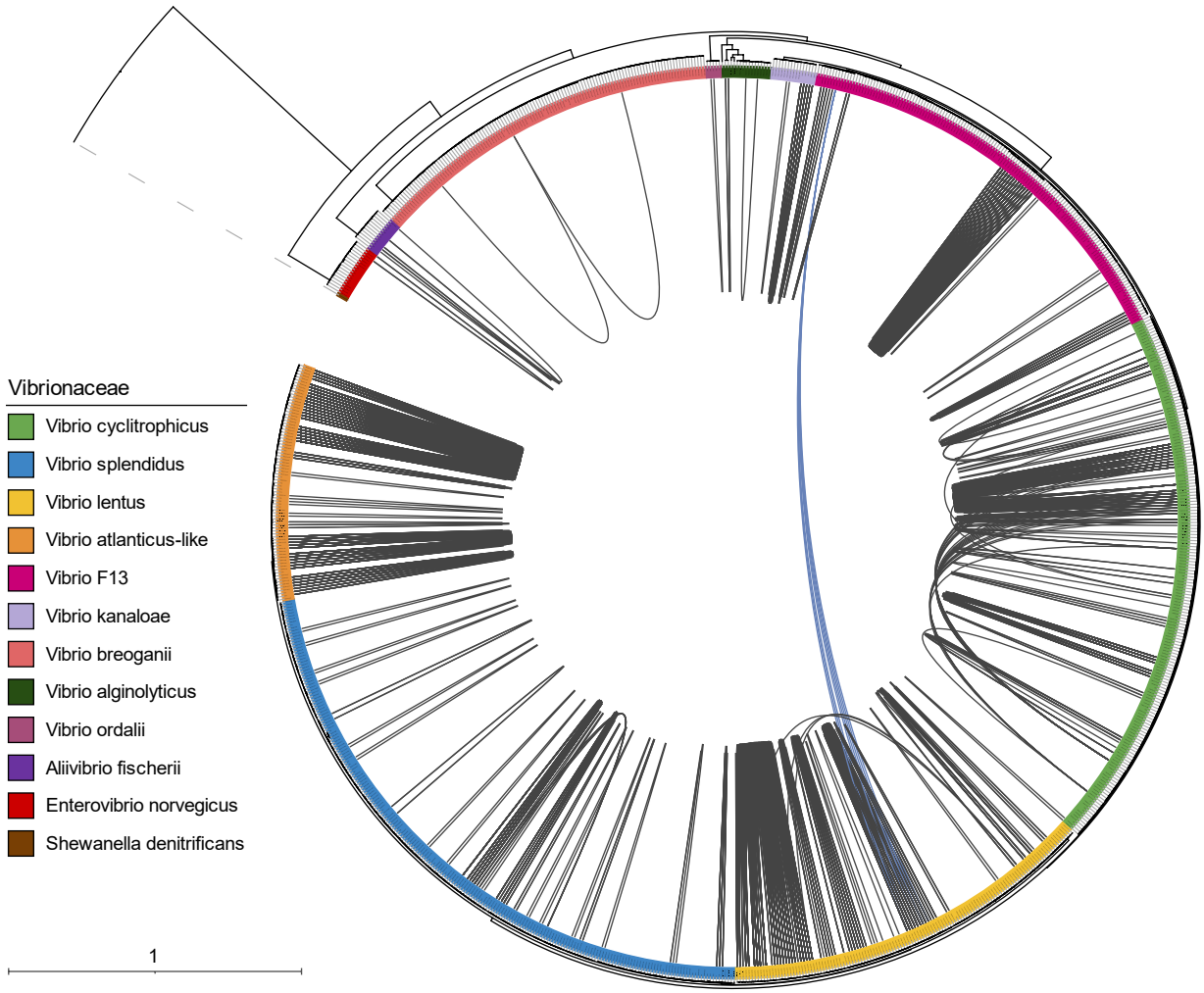

Figure S7 - Shared prophages across the *Vibrionaceae*
Shared prophages (ANI >99.9%, full coverage) across co-occurring populations of different marine *Vibrionaceae* (colored by species). Phylogenetic relationships of the different *Vibrionaceae* strains are based on single copy ribosomal protein. The scale bar denotes substitutions per site. Transfers within population are depicted in black. Transfers between populations are highlighted according to the viral realms (blue = tailless prophages).

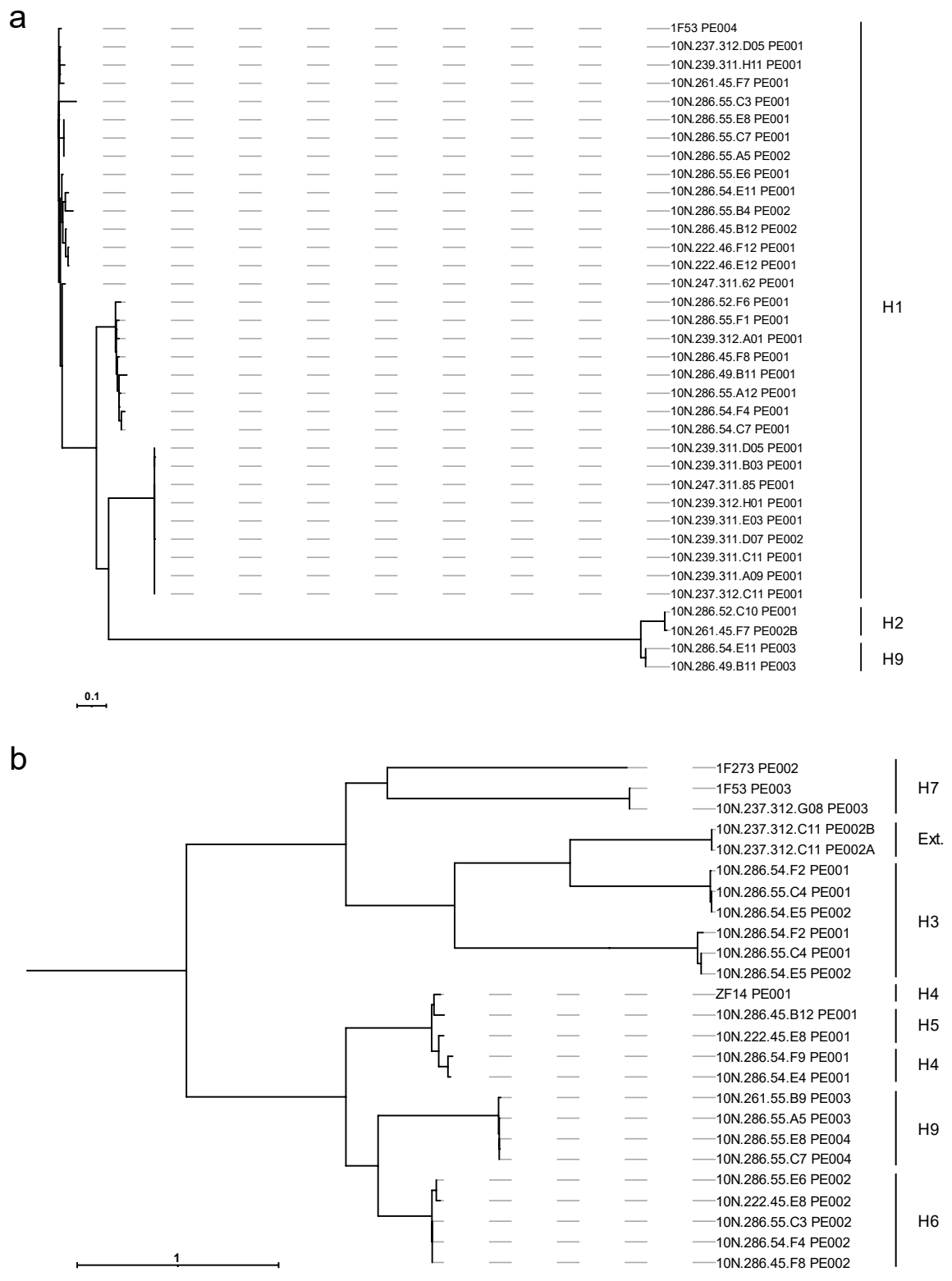

Figure S8 - Prophage integrases and associated integration sites.
Phylogenetic trees based on protein alignments of integrase genes in tailless (a) and tailed (b) prophages. The tree scale represents substitutions per site. Phylogenetically related integrases are mostly carried by prophages integrating into the same genomic location, indicated by hotspots H1-H9.

References

1. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 2019; 37(5):540–6.

2. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011; 17(1):10–2.

3. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018; 34(18):3094–100.

4. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 2017; 27(5):737–46.

5. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021 Jan 29; 10(2):giab008.

6. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019; 7:e7359.

7. De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics*. 2023 May 4; 39(5):btad311.

8. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013; 29(8):1072–5.

9. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015; 25(7):1043–55.

Chapter II

Mobile genetic elements are the main drivers of bacterial pangenome diversity

Kerrin Steensen[1,2], Shaul Pollak[1], Xiaoqian A. Yu[1], Martin F. Polz[1*]

[1] Centre for Microbiology and Environmental Systems Science, Division of Microbial Ecology, University of Vienna, Vienna, Austria
[2] Doctoral School in Microbiology and Environmental Science, University of Vienna, Vienna, Austria
[*]Corresponding Author

Running title

MGEs drive bacterial pangenome diversity

Keywords

mobile genetic elements, pangenome, flexible genome, horizontal gene transfer, *Vibrio*, phage defense

Author contributions

Conceptualization – KS, MFP
Methodology – KS, SP, XAY
Software – KS, SP
Data curation – KS
Formal analysis – KS
Visualization – KS
Writing (original draft) – KS
Writing (review and editing) – KS, MFP
Supervision, Funding acquisition – MFP

Abstract

Bacterial species can demonstrate tremendous genetic variation among individuals. A portion of this variation is driven by the fast gain and loss of mobile genetic elements (MGEs). However, the extent to which MGEs contribute to this genetic diversity in bacterial pangenomes remains poorly constrained. Here, we show that MGEs are highly diverse and represent the main driver of flexible gene pool variation in bacteria of the genus *Vibrio*. To quantify the contribution of MGEs to the flexible gene pool, we first compared sets of closely related genomes using a graph-based comparative genomics approach, yielding precisely defined contiguous flexible genome regions. Then, we defined MGEs based on the presence of mobility signatures. We identified 2,600 putative MGEs among 127 closely related strains of three *Vibrio* species. These MGEs represent the main contributors to the flexible gene pool. Transposons and prophages constituted the predominant MGE types, however most MGEs did not fall into established categories. Among the unclassified MGEs, we describe a prevalent, previously underappreciated type characterized by at least two tyrosine recombinases at its border. Across types, MGEs were genetically diverse. This diversity was driven by a range of MGE subtypes characterized by specific conserved genes, and by variable accessory genes or even entire nested MGEs within MGE types. While large-scale computational studies have already recognized the high global diversity of MGEs, our results demonstrate that even closely related, co-existing microbes harbor large numbers of highly diverse MGEs. Our work highlights that MGEs shape the eco-evolutionary dynamics of bacteria in nature and represent the major driver of bacterial pangenome diversity.

Introduction

Even closely related bacterial genomes exhibit remarkable diversity in gene presence and absence, which has led to the definition of the pangenome (1). The pangenome consists of two parts: genes conserved in all members (the core genome) and the flexible genome (or accessory genome), which is characterized by genes that are not shared among all population members. This flexible genome can rapidly grow with each additional genome analyzed, suggesting a large gene pool with high turnover. There is an ongoing debate about the evolutionary scenarios that lead to the maintenance of such variation (2,3), with neutral and selectionist views existing side by side.

MGEs are an important yet poorly characterized component of the flexible genome. These genetic elements are either self-mobilizable, mobilized by helper elements, or capable of mobilizing other DNA (4), and as such, they can move within or between bacterial genomes (5). MGEs can make up a large proportion of the bacterial genome, as they can comprise several hundred kilobases of DNA that are either inserted in the chromosome or remain in the cell as episomes (5). MGE types are usually characterized by their specific mechanism of mobilization, even though high genetic variation within MGE types is common (4). Many MGE types share the ability to carry accessory genes that are non-essential for the MGE's maintenance (6). These genes can equip the host with various functions, including virulence factors and phage defense (7,8). Several MGE types have been discovered in the last century. In recent years, growing interest in the potential of MGEs to promote or inhibit horizontal gene transfer led to an increase of described MGE types (9–13). Yet many MGE types remain to be described.

Several large-scale computational studies pioneered MGE identification and provided first insights into their diversity (14–16). However, the computational identification of MGEs still faces several challenges. The high variation of MGEs, as well as the lack of a common marker gene among all MGE types makes their identification difficult (15,17). In addition, less studied types often have insufficient numbers of characterized MGE sequences, with core genes described by annotations from individually chosen databases that are further complicating their identification. Even when MGEs are identified, commonly used identification tools that are searching for marker genes often fail to define the element's precise borders (18). This lack of clear boundaries leads to inaccurate estimates of the number of genes that are MGE-derived, thus affecting assessments of the extent to which MGEs contribute to the flexible gene pool. This can be circumvented by using comparative genomics to predict MGEs (15), as MGE borders are inferred based on differential MGE presence in bacteria. However, this approach requires the availability of closely related genomes. Finally, the use of long-read sequenced genomes has improved MGE recovery. MGEs are often incomplete in short-read sequenced genomes, as they can cause assembly breaks, for example when genomes contain MGEs in multiple copies.

Bacteria exhibit considerable variation in their pangenomes. While it is known that a portion of these genomic changes is driven by the gain and loss of mobile genetic elements (MGEs), the extent to which MGEs drive variation in the flexible genome remains unclear. Therefore, we aimed to quantify the contribution of MGEs to the flexible genome and investigate what drives the diversity conferred by MGEs. To this end, we used comparative genomics and closed genomes to precisely identify the number and borders of MGEs encoded in 127 marine bacterial strains from three species within the genus *Vibrio*. We show that MGEs are the main contributors to the flexible gene pool and report a surprisingly large diversity of MGEs present in closely related bacterial strains. We further demonstrate that this diversity is, in part, fueled by the nesting of MGEs within each other and by the spread of accessory genes between MGEs within individual bacterial genomes. Through this detailed analysis of MGE diversity in a larger set of closely related strains, we provide evidence that MGEs are the main contributors to the flexible gene pool of their host species and are, thus, important drivers of the adaptation and diversification of bacteria in the environment.

Results

The *Vibrio* collection harbors numerous MGEs

To investigate the diversity of MGEs and their contribution to the flexible gene pool, we first aimed to comprehensively define flexible genome regions and then identify MGEs among these regions across three *Vibrio* species: *V. cyclitrophicus*, *V. lentus,* and *V. splendidus*. While the flexible genome is often identified based on the frequency of single proteins (19), we used a graph-based comparative genomics approach on closed, or nearly closed, genomes to predict the number and precise borders of flexible genome regions in each species (methods). Regions in the flexible genome were subsequently classified as putative MGEs, if they contained specific mobility signatures typically found in self-mobilizable MGEs and for those unable to autonomously mobilize (methods, Table 1).

We identified 5,782 contiguous flexible genome regions in 127 *Vibrio* genomes, among which we classified 2,600 chromosomally integrated or episomal elements as putative MGEs. The mobility signatures we used to define MGEs built on those established in previous studies (14,15). Thus, we searched for recombinases, such as integrases and transposases, genes encoding conjugation machinery, and terminal inverted repeats at the borders – well-known recognition sites for transposons. We expanded this set of mobility signatures to include *dif* sites, which are known to be involved in the integration of lysogenic phages in *Vibrio* (20). Finally, we also denoted all episomal elements as putative MGEs, since the latest research implies that most, if not all, episomal elements are mobile or mobilizable (21). All regions without any of the above-mentioned mobility signatures were not considered MGEs, but instead as a part of the remaining flexible genome. These regions may either represent MGEs with unknown or degraded mobility signatures, or they represent flexible genomic regions acquired by homologous recombination of the flanking regions. With this approach, we ensured a conservative estimate of the abundance and diversity of MGEs.

MGEs are the main contributors to the flexible gene pool

Leveraging the precisely defined borders of MGEs and the flexible genome, we assessed the total genomic space they occupy in the individual host genomes. MGEs constituted a considerable fraction of their individual host's genome, ranging from 2.1% to 20.7% (Figure 1A). The proportions of both flexible genome regions and putative MGEs were species-dependent (Kruskal-Wallis test, p-value$_{FlexibleGenome}$ = 2.2x10^{-16}, p-value$_{MGEs}$ = 1.8x10^{-15}): We observed that *V. cyclitrophicus* - the species with the smallest average genome size of 5.2 Mb - also had the smallest average proportion of flexible genome regions (8.3%) and putative MGEs (6.3%). Conversely, *V. splendidus* and *V. lentus* had larger genomes averaging 5.7 Mb and 5.8 Mb, respectively, of which 11.4% and 12.7% were putative MGEs. These results confirm previous observations that larger genomes are associated with higher MGE load (22).

We found that putative MGEs were the main contributors to the flexible gene pool, as measured by the number of unique gene families (clustered at 95% identity to combine functionally redundant genes) introduced by MGEs and the remaining flexible genome. We compared the rarefaction curves showing the average increase of unique gene families conferred by MGEs, and the remaining flexible genome, with every additional *Vibrio* genome (Figure 1B). The rarefaction curves showing this increase in MGEs were steep and did not demonstrate any signs of saturation. In contrast, the respective curves of the remaining flexible genome flattened rather quickly, indicating that putative MGEs contributed far more unique gene families to the flexible gene pool than the remaining flexible genome. Interestingly, we observed that the increase in unique gene families conferred by MGEs in *V. splendidus* was higher than in *V. lentus*, as indicated by a steeper rarefaction curve, while the total size of MGEs was comparable in the two species (Wilcoxon rank-sum test, p-value=0.1355). This result indicates a higher diversity of MGEs in *V. splendidus*. However, the difference might be caused by a subset of *V. lentus* isolates, whose core genomes have been shown to be nearly clonal (23) and, thus, likely contain a higher proportion of vertically inherited MGEs.

Although MGEs are responsible for substantial genetic diversity in individual genomes, the large extent of their contribution is evident on the pangenome level. Of the average 4,695 unique gene families encoded on individual genomes, between 7.3% and 13.0% of the gene families were encoded at least once on an MGE (Figure 1C). This proportion increased to at least 39% of the pangenomes when investigating the number of gene families on the species level (Figure 1D). Previous studies reported varying proportions of gene families associated with MGEs in the pangenome, ranging from below 30% to almost 40% in the model organism *Escherichia coli* (2,22). Here, we observed at least 39% of the gene families of the pangenome to be associated with MGEs, whereas the unsaturated rarefaction curve (Figure 1B) suggests that this proportion would further increase with the sampling of more genomes. Hence, MGEs contribute substantially, if not predominantly, to the genetic variation observed in bacterial pangenomes.

Most MGEs carry recombinases

To obtain insights into the common mechanisms of mobilization and transfer of MGEs, we examined the prevalence and proportion of putative MGEs characterized by the different mobility signatures. First, we considered all 194 flexible genome regions for which we could not identify a chromosomal location as “episomal” putative MGEs, since recent studies imply that most, if not all, episomes are mobile or mobilizable. Episomal elements accounted for about 11% of the total size of all putative MGEs. Episomes do not need additional genes mediating genomic insertion, because they can exist inside the host cell as circularized or linear DNA. Diverse MGEs have been shown to exist episomally, including the well-described conjugative plasmids (21), phage plasmids (24) or chromosomally integrated MGEs that are currently replicating or mobilizing via circularized intermediates, such as filamentous phages (25), integrative-conjugative elements (26) or gamma-mobile-trio elements (13). The remaining episomes might represent plasmids that do not encode for their own conjugation machinery. Instead, they are thought to encode an origin of replication, and transfer across bacterial hosts by exploiting the conjugation machinery of other plasmids present in the cell (21).

Most identified putative MGEs (n=1,808) were chromosomally integrated elements containing genes that may mediate the mobility of the MGEs. A small fraction of 85 putative MGEs contained their own conjugative machinery, which we assigned to the “chromosomal-conjugative type”. These could represent integrative-conjugative elements (ICEs), widespread MGEs that can be transferred between bacterial cells by conjugation (9). The largest fraction - 1,723 of the chromosomally integrated putative MGEs - encoded for integrases or transposases, which are recombinases that are known to mediate genomic insertion of diverse MGEs, including transposons, prophages and many, if not most, other currently described MGE types (15). Our results show that 69.5%, the majority of the identified putative MGEs, contain at least one mobility gene. This is more than in a previous study that found 51.4% of the MGEs to contain mobility genes (14).

For the remaining 598 chromosomally integrated putative MGEs, we could not identify genes that may directly mediate their mobility. Instead, they contained non-coding sequences that have been shown to be involved in mobility. Most of these sequences encoded for terminal inverted repeats (TIRs) at their element borders. It is known that MGEs carrying TIRs like the small-sized miniature inverted-repeat transposable elements (MITEs) and transposons with defective transposases can be mobilized *in trans*, meaning that their transposition can be mediated by related, intact transposons in the cell (27). Another 105 putative MGEs contained a sequence with similarity to the *dif* sites near the chromosomal termini (28). Since some MGEs are known to exploit host-encoded recombinases for their integration by encoding such *dif* sites (29), we considered the ”*dif* site” to be a signature of mobility. Finally, 65 putative MGEs were assigned to the “attC-rich” mobility signature group. These attC sites are involved in the mobility of integron cassettes. In contrast to complete integrons, these clusters of *attC* sites are lacking integron-integrases (also called CALINs) and have been proposed to be remnants of integrons (30,31). Although the proportion of MGEs with non-coding mobility signatures could increase as more are identified, most of the putative MGEs in our study contain recombinases that may mediate their mobility.

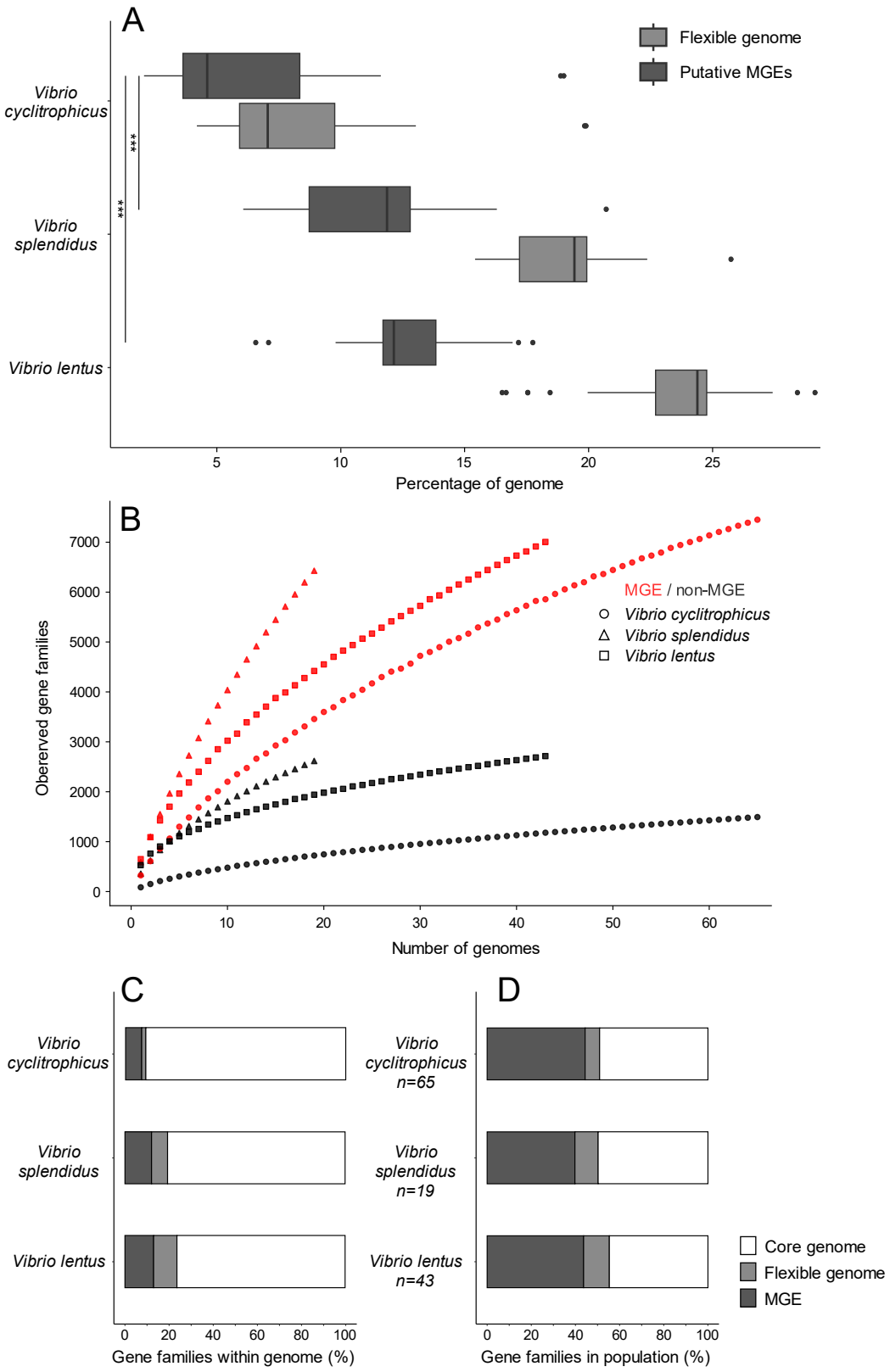

Figure 1 - MGEs confer genetic diversity

A) Boxplot showing the percentage of MGEs and flexible genome regions (including MGEs) compared to the entire genome. Stars show significant difference in MGE load per species (pairwise Wilcoxon test) B) Rarefaction curve of observed gene families (clustered at a threshold of 95%) per species when sampling MGEs or the remaining flexible genome ("non-MGE") of *Vibrio* strains. C) Barplot depicting the mean percentage of gene families present in MGEs and in the remaining flexible genome regions compared to all gene families in individual genomes or D) calculated within all members of each species.

Transposons and prophages predominate MGEs

The mobility signature analysis suggests a broad variety of MGE types to be present in our dataset. Therefore, we aimed to assign the putative MGEs to previously described MGE types. First, we dereplicated the putative MGEs resulting in 2,474 unique MGEs, followed by an average nucleotide identity clustering with a 95% cutoff to group closely related MGEs together. The classification was then based on the functional annotation of the gene content and on the results of common MGE prediction tools within groups of these closely related MGEs (methods). We aimed for the best classifier for each identified putative MGE. For example, plasmids carrying a transposon would be classified as plasmids. If two separate MGEs were clearly co-integrated in the same contiguous flexible genome region, we assigned two classifiers to the region.

While the large majority, in total 1,270 putative MGEs, remained without classification into known MGE types, we found transposons, with 418 putative MGEs, to be the most abundant MGE type in our set (Figure 2B). We observed three commonly described sub-types of transposons: Insertion sequence (IS) elements, composite transposons, and non-composite transposons (Figure S1) (32). IS elements are simple transposons that require only one or two genes for their movement. We identified 209 putative transposons with at least two IS elements that may represent composite transposons, which are transposons composed of two IS elements encoded in proximity to each other and can mobilize the cargo genes in between. Although we did not systematically classify transposon subtypes, we observed more complex, non-composite transposons with transposition genes unrelated to IS transposases.

The second most common MGEs were prophages accounting for 205 MGEs (Figure 2B) and 5 additional prophages that were inserted into plasmids. Of these prophages, 79 derived from tailless, 70 from filamentous and 60 from tailed phages, in proportions that are consistent with previous work on prophages in *Vibrio* (33). Prophages were characterized by many different mobility signature groups, but most of them (n=146) carried recombinases. We observed 4 phage plasmids, and 7 tailless and 41 filamentous phages encoded a *dif* site. Surprisingly, we also identified one tailless and three filamentous phages in the TIR mobility group, consequently lacking an annotated integrase or transposase – all inserted in the same site in *V. lentus*. The remaining three prophages were assigned to the episomal group. Two of these prophages might have been currently replicating filamentous phages and one represented an incomplete tailed phage, which was broken in the assembly.

Other well characterized MGE types were less common, with 97 plasmids, 97 integrons and 85 putative ICEs or integrative mobilizable elements (IMEs). IMEs are elements related to ICEs but do not carry all genes necessary for the conjugation themselves (34). Interestingly, the number of CALINs (n=81), which are thought to be degraded integrons, was almost as high as the number of integrons themselves. We also identified 93 repeats, which are not necessarily MGEs.

In addition, we specifically searched for IS elements across the whole genome. IS elements are very small MGEs that might be missed in our classification approach due to their insertion within larger MGEs or other genomic locations. We found IS elements to be more prevalent (n=5,129) than any other MGE in our dataset (compare Figure 2B), an observation that has been made in a recent large-scale MGE marker-gene search as well (15). In fact, single *Vibrio* strains could contain up to 164 IS elements. We observed significant differences in the IS abundance among different *Vibrio* species (Kruskal-Wallis test, p-value < 2.2×10^{-16}). Furthermore, the number of encoded IS elements varied considerably among strains within species. Generally, strains with larger genome sizes tended to contain more IS elements (Figure 2C). Considering the close relationship of the *Vibrio* strains in this study, we did not perform a statistical test due to the data's lack of independence. However, a previous study showed a positive correlation between IS abundance and genome size across various bacterial strains (35). Interestingly, we observed that IS elements were significantly enriched in the flexible genome with a total of 3,655 IS elements (Wilcoxon signed rank test, p-value < 2.2×10^{-16}). When normalized for sequence length, the median number of IS elements was more than four times higher than in the whole genome (Figure 2D). The enrichment in IS elements further contributes to the number of identified transposons, making them by far the most abundant type of MGE.

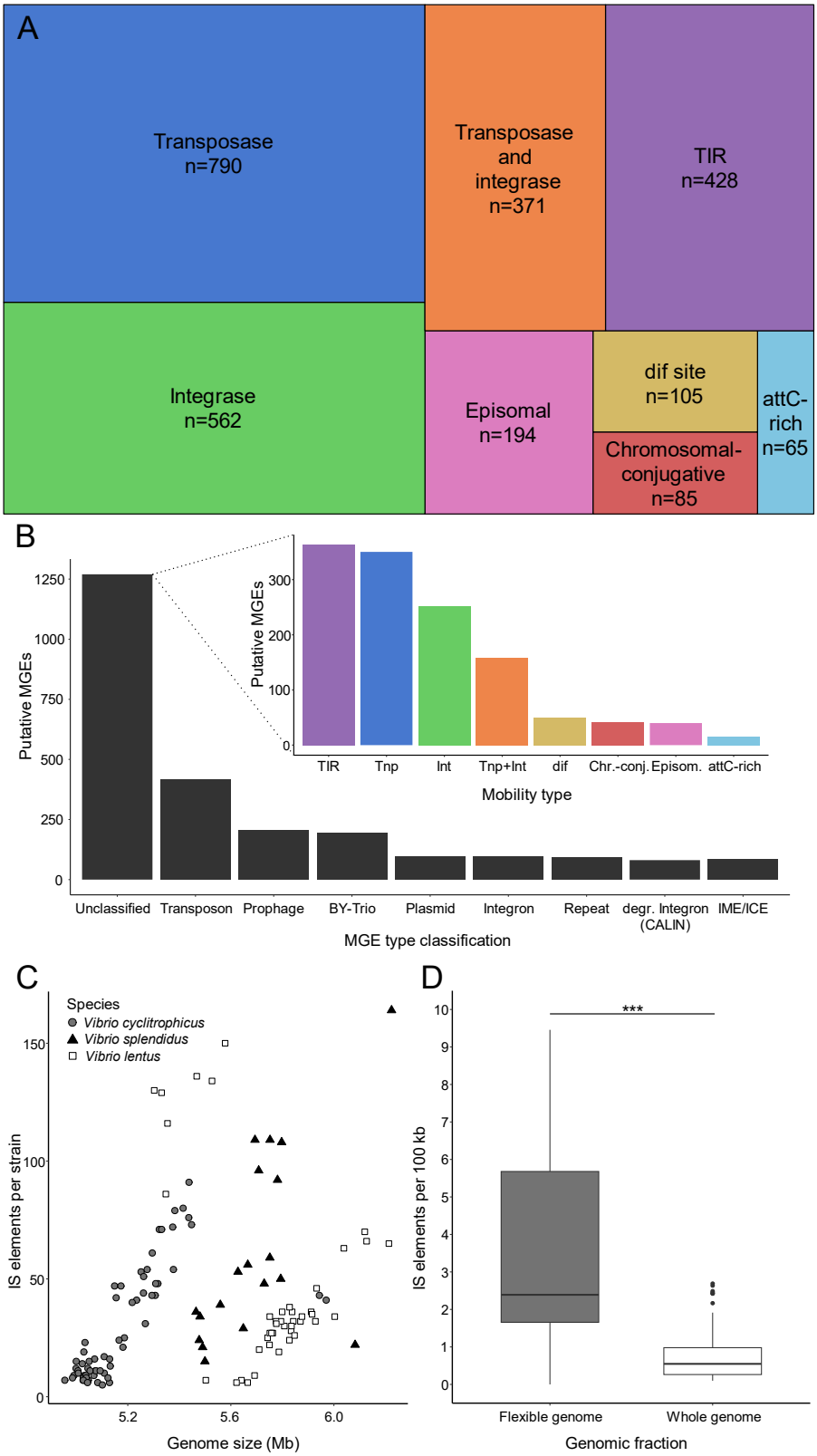

Figure 2 - MGE classification

A) Treemap indicating the mobility signature groups of all putative MGEs. B) Barplot showing the identified elements of known MGE types. The mobility signature group of all unclassified elements is indicated in the top right C) The number of IS elements in whole genomes of different *Vibrio* strains depending on their genome size. Species classification is indicated by colored shapes. D) IS element density (number per 100 kb) in whole genomes and in the flexible genome. Stars indicate significate differences between genomic fractions (paired Wilcoxon test).

Unclassified MGEs in *Vibrio* are a reservoir of novel MGE (sub-)types

As most putative MGEs could not be classified into previously described MGE types (Figure 2B), despite containing typical mobility genes such as transposases or integrases, we asked whether we could identify novel MGE types in our dataset. Known MGE types regularly share at least some conserved proteins, often encoded in a specific gene order, that mediate their movement or are involved in their maintenance. Reasoning that novel MGE types might also demonstrate such core proteins in a specific gene order, we used a synteny-based gene sharing approach to identify putative novel MGE types. To this end, we first clustered the putative MGEs based on their shared syntenic genes (methods). Within the obtained clusters, we then identified conserved core protein combinations by detecting functional gene families (50% identity) that were present in most MGEs within a cluster and showed a high co-occurrence (methods). MGEs carrying all proteins of a specific core protein combination were considered to constitute putative novel MGE types (methods). We manually investigated around 340 putative novel MGE types derived from core protein combinations, that we required to share common syntenic core genes while preserving variation in the remaining MGE genomes to avoid signal from vertically inherited MGEs. Additionally, any subsets or duplicates of MGE types were removed. This analysis resulted in 30 putative MGE (sub-)types.

We first tested whether the synteny-based core protein detection approach successfully recovers known MGE types. In fact, 15 candidates belonged to known MGE types: 5 different subtypes of transposons (Figure S1), one ICE, 4 integron-related elements and 5 prophage subtypes. The prophage subtypes included one that belonged to tailless, two to filamentous and two to tailed genomically integrated phages. Since the prophages in the *Vibrio* collection have been analyzed in detail (33), we used them to assess the prediction quality: The predicted MGE subtypes often included a majority of the MGEs from that subtype in our dataset, with no or few false positives: For instance, one predicted MGE type contained 77 MGEs, of which we confirmed all 77 to belong to genomically integrated tailless phages based on their gene content. Furthermore, the five identified types summed up to 195 prophages, meaning that more than 90% of all prophages were recovered using the synteny-based approach suggesting that it is well-suited to identify novel MGE types.

Using our synteny-based approach, we observed 9 different putative MGE (sub-)types that were characterized by encoding two or three genes annotated as tyrosine integrases at or close to their element border (Figure 3). Usually, the first 3-4 genes of these putative MGEs were conserved, followed by highly variable genes often unrelated among individual MGEs within each type. Accounting for almost 200 elements, these MGEs were nearly as prevalent as prophages in our dataset (Figure 2B). Surprisingly, we observed very little sequence similarity among the conserved tyrosine integrases encoded in the different (sub-)types (Figure 3), indicating different origins of these MGE (sub-)types. However, when aligning the (sub-)types with each other, we observed that they occasionally shared accessory genes with high sequence similarity, an indication for exchange between these putative MGE (sub-)types (Figure 3). Several studies described MGE types that resembled our putative MGE (sub-)types in their descriptions. These were all named by the 3 conserved genes at their element border, namely the recombinase in trio (RIT) elements (36,37), the Genomic Island with three Integrases (GInts) (38) and, most recently, the Gamma-Mobile-Trio elements (13). All these elements, including the putative MGEs we identified encode two or three bordering tyrosine recombinases, which were reported to be variable in their gene lengths (37). Several of these studies reported that the MGEs were able to circularize (13,38). In fact, we observed that 4 of these putative MGEs were episomes. Finally, the previous studies mentioned that the MGEs contain a longer sequence of variable genes, including phage defense, antibiotic resistance, or virulence genes. In accordance with these studies, we also observed phage defense genes encoded by the putative MGE type, as well as other functional annotations such as transporters (Figure 3). We used the MGE borders predicted from the graph-based flexible genome prediction to estimate the length of the identified BY-Trio elements in our study. The median was 16.4 kb, which is in the same size range as the median of the previously described Gamma-Mobile-Trio elements with a median of 13.4 kb (13), although the length of these MGEs can be highly variable (13,38). Based on the similarities, we combined the three previously described subtypes and

our 9 predicted MGE subtypes to form the BY-Trio elements (Bordering tyrosine-Trio) drawing inspiration from prior mentions of similar elements.

We observed 6 additional putatively novel candidate MGE types, all of which were present across species boundaries. Each predicted type contained an integrase gene and was observed to carry accessory genes, such as phage defenses. Three of the putative MGE types shared parts of their predicted core genes (Figure S2). Apart from an integrase, the core gene combinations of these three candidate types contained a RadC-like DNA repair protein, and – dependent on the subtype – proteins with a WYL domain or proteins lacking functional annotation. We observed one more type possibly representing another BY-Trio element subtype, as it has bordering tyrosine recombinases (Figure S3A). However, one of the bordering genes was in the opposite direction, which had not been described for any BY-Trio subtype before. We identified a candidate MGE type that also demonstrated three bordering genes, which were predicted as homing endonuclease, a gene without functional prediction, and a transcriptional regulator followed by trailing variable genes (Figure S3B). Lastly, we observed a rarely occurring MGE with only 7 representatives in 2 species. Its core genes were predicted to be related to a protein from the type IV secretion system, as well as a resolvase and a tyrosine integrase (Figure S3C). While all of these candidate subtypes might represent novel MGE types, additional investigation in a broader set of bacterial strains or experimental verification would be required to confirm their mobility and to characterize their function. Our results show that the diversity of MGEs in only three species of closely related bacteria is sufficient for the discovery of novel MGE types, even though confirmation outside of *Vibrio* is needed.

Most MGEs have only a few close relatives

Motivated by the observation that several MGE types consist of highly diverse MGEs, partly without homology to each other, we investigated how many distinct groups of closely related MGEs could be identified, each potentially having diverged within the host species. To this end, we first dereplicated the putative MGEs and then clustered them at an average nucleotide identity of 95%. We used this threshold as a proxy for within-species differentiation of MGEs, as it corresponds to the sequence divergence often used to distinguish between bacterial species. To account for variation in accessory genes, we set the minimum coverage to 40%. Each resulting divergence cluster harbored closely related MGEs, mainly differing in their accessory gene repertoire.

Using the above criteria, we observed a remarkably high diversity of MGEs divergence clusters. Out of the 2,600 putative MGEs, 126 were removed during the dereplication step as they were identical to other MGEs in the set. This is a small number, especially considering that MGEs from closely related strains are expected to be vertically inherited. The remaining 2,474 putative MGEs grouped into 336 divergence clusters, whereas 518 putative MGEs remained as singletons (Figure 4A). The divergence clusters (excluding singletons) contained an average of 5.8 putative MGEs. While other studies have reported the diversity of MGEs to be high across more distantly related bacteria from various environments (14,15), our data show that even in closely related bacteria sampled in a shared environment we observe a very large diversity of MGEs.

The divergence cluster size (i.e. number of putative MGEs within individual divergence clusters) was often associated with their MGE type (Figure 4B). The largest clusters were carrying transposases or integrases (Figure 4A), often representing transposons and intact or degraded integrons. For example, one of the largest divergence clusters contained rather short transposons that might have been mobilized recently and therefore sustained a relatively high pairwise nucleotide identity. Since integrons are not inherently mobile but much rather exchange gene cassettes (30), clusters predominantly containing integrons might represent vertically inherited, diversifying integrons. Conversely, almost 30% of prophages (n=61) and putative MGEs belonging to the BY-Trio type (n=55) were unique, appearing as singletons in our clustering, thus suggesting that these types are highly diverse. Among bacteriophages, such diversity is well documented (39), and the number of singletons suggests that BY-Trio elements are highly diverse as well. Not all such singletons necessarily represent highly diverse MGEs but could contain several nested or co-integrated MGEs, that are challenging to

separate from each other in the computational analysis. Such nestedness has been observed before (10,15,33) and it might lead to separate divergence clusters or singletons instead of clustering together with each of its contained putative MGEs. With the numerous divergence clusters and singletons, MGEs likely contribute a vast number of gene families to the bacterial pangenome and are thus important drivers of the genetic variation in bacteria.

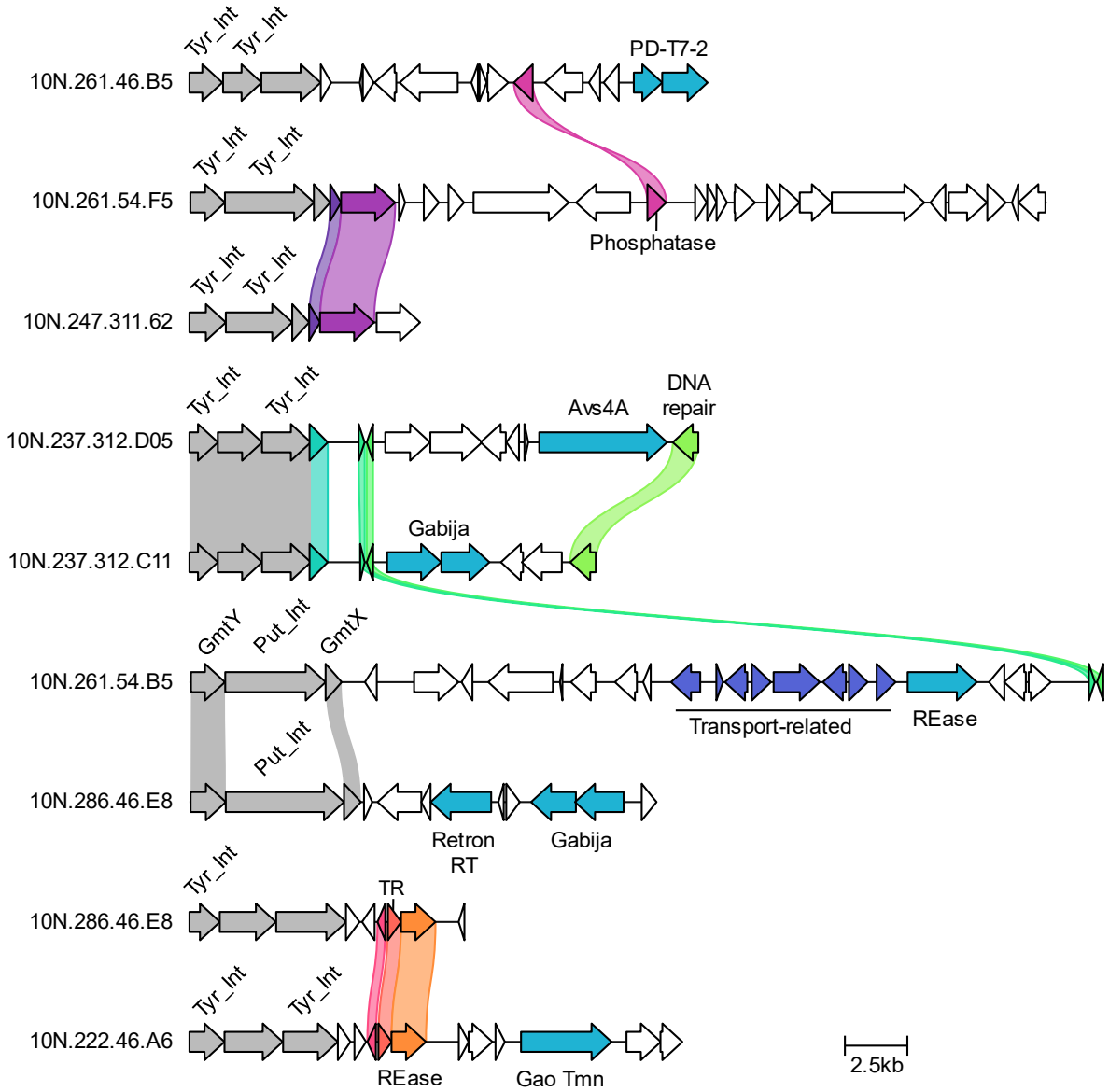

Figure 3 - BY-Trio elements

Genomic maps of different putative MGE types sharing tyrosine recombinases at their element borders. Core proteins are depicted in gray. Genes are connected if their sequence identity was above 30%. Genes in the flexible genome are colored in white with highlighted examples of phage defense genes identified using Defensefinder, or for highlighting other functions. Abbreviations: Tyr_Int = tyrosine integrase, GmtY = tyrosine recombinase in Gamma-Mobile-Trio elements, GmtX = Gamma-Mobile-Trio protein X, Put_Int = putative integrase, REase = restriction endonuclease, TR = transcriptional regulator.

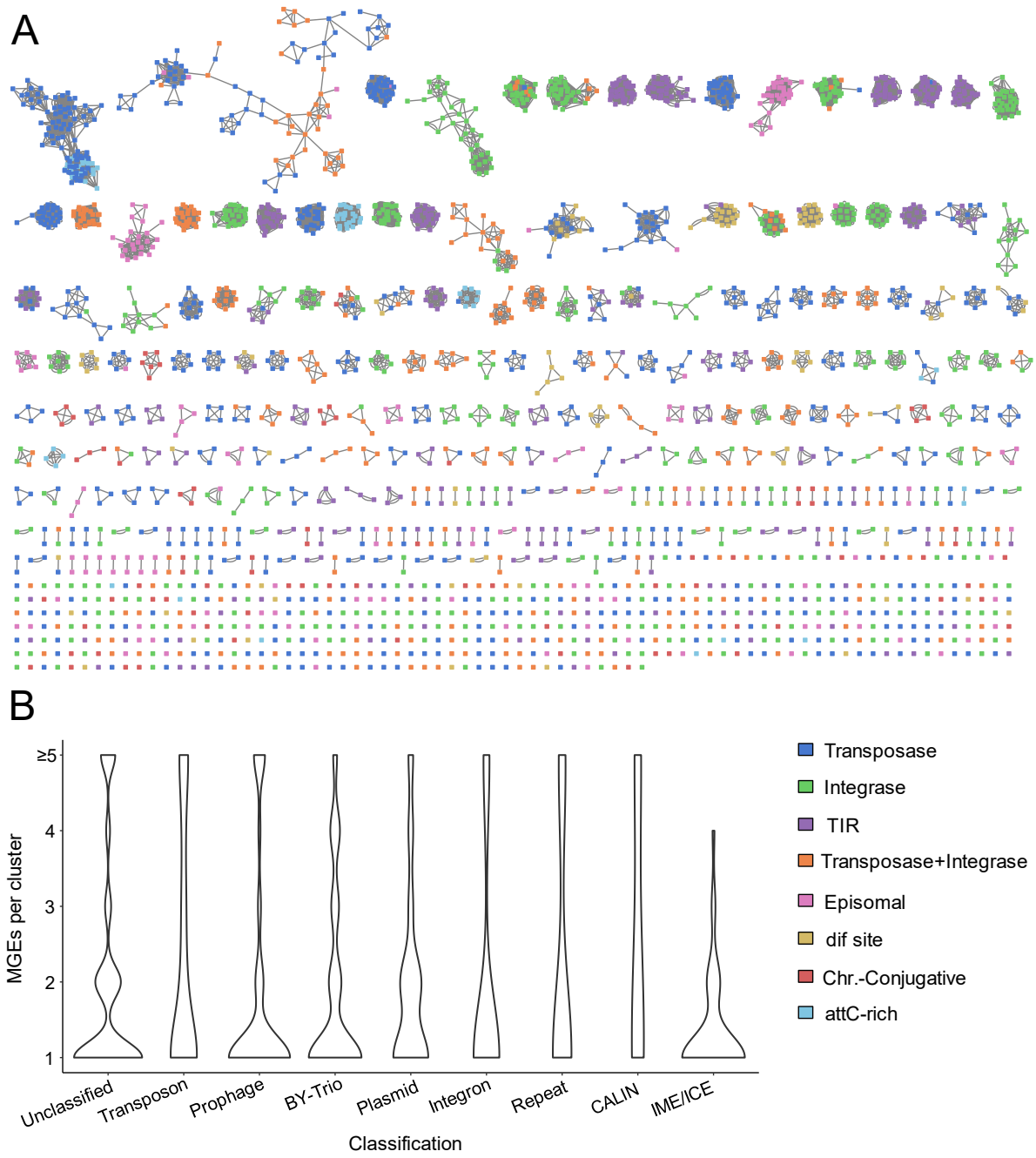

Figure 4 - Sequence diversity of putative MGEs
A) ANI-based network showing divergence clusters of putative MGEs connected if pairwise identity was at least 95% with a coverage of at least 40%. Putative MGEs are colored by mobility signature groups B) Cluster size dependent on their predominant MGE type.

Accessory genes and nested MGEs drive genetic diversification

We observed that MGEs are forming many different divergence clusters in closely related *Vibrio*, and that this diversity is a driver of the expansion of the flexible gene pool of their bacterial hosts. As MGEs are known to exert variation within types (4), we were interested in whether the variation within more closely related MGEs could be an additional driver of the diversity in bacterial pangenomes.

We noticed distinct genomic locations harboring highly variable accessory genes when comparing genomes of MGEs within their respective MGE subtypes (Figure 3, Figure 5A-B). These distinct locations often flank mobility genes such as integrases or are located close to the MGE's borders. For example, prophages regularly encoded such accessory genes directly at their border (Figure 5A). Interestingly, such prophage-flanking genes have previously been described in the context of specialized transduction, where phages mobilize flanking bacterial genes (40). Genes encoded in these loci were often unrelated to any other genes encoded on related MGEs. Therefore, variable loci within MGE genomes are likely providing many novel genes to the flexible gene pool of the host species. Accessory genes encoded in variable loci were often associated with phage defense functions, while annotated transporters or transcriptional regulators were also observed. Yet, many of the accessory genes were lacking any functional annotation. It is well known that MGEs contain accessory genes encoding different functions including phage defense or virulence (13,33,41–43). Furthermore, phage defenses have been reported to be enriched in distinct locations of phage-like MGEs and in gamma-mobile-trio elements (13,33,42). Our data suggest that variable genes encoded in distinct genomic locations might be widespread across different MGEs, and that they are major drivers of the genetic diversity MGEs provide to the flexible gene pool.

Apart from providing various accessory genes to their hosts, we found indications that variable loci of larger MGEs also serve as insertion sites of smaller MGEs, which may hitchhike to transfer between host cells. During the prediction of novel MGE types, we noticed that some putative MGEs were assigned to multiple MGE types. We observed that these regions often represented MGEs that were nested within highly variable loci of another MGE, frequently between blocks of several conserved "core" genes. For instance, we observed a small integrase-encoding MGE, also integrated by itself in another bacterial strain, inside a variable locus flanking the phage's integrase (Figure 5A). We observed similar patterns in conjugative plasmids (Figure 5B), with several smaller MGEs nested between the conjugation machinery, a resolvase and conserved DNA repair proteins. Nestedness of MGEs has been noted before (15), and MGEs such as integrons and transposons are known to move on plasmids (5,30). However, we observed here that MGEs were integrated within variable loci between plasmid or prophage core genes, such as the integrase or the conjugation machinery (Figure 5A-B).

Given the extensive variety of genes in variable loci within MGE types, we wondered whether these accessory genes would be exchanged among MGEs co-infecting the same host. Reasoning that recently transferred genes should be nearly identical, we searched for shared proteins with an identity of at least 99.5% encoded on MGEs within bacterial genomes. We did not observe high gene sharing among MGEs inserted in the same genome. However, we identified numerous identical IS transposases, indicating that they can be mobilized frequently within the genome. In a few cases, these identical IS elements shared highly similar neighboring accessory genes indicating that they were mobilized along with the IS element. For example, we observed a restriction endonuclease, possibly involved in phage defense or genes related to nucleoside metabolism, flanked by two IS elements, which is the hallmark of composite transposons (32). Although our results do not provide a clear explanation of how these variable loci maintain their diversity, the enclosed variable genes and nested MGEs contribute to the MGE-driven diversity in the flexible gene pool of bacteria.

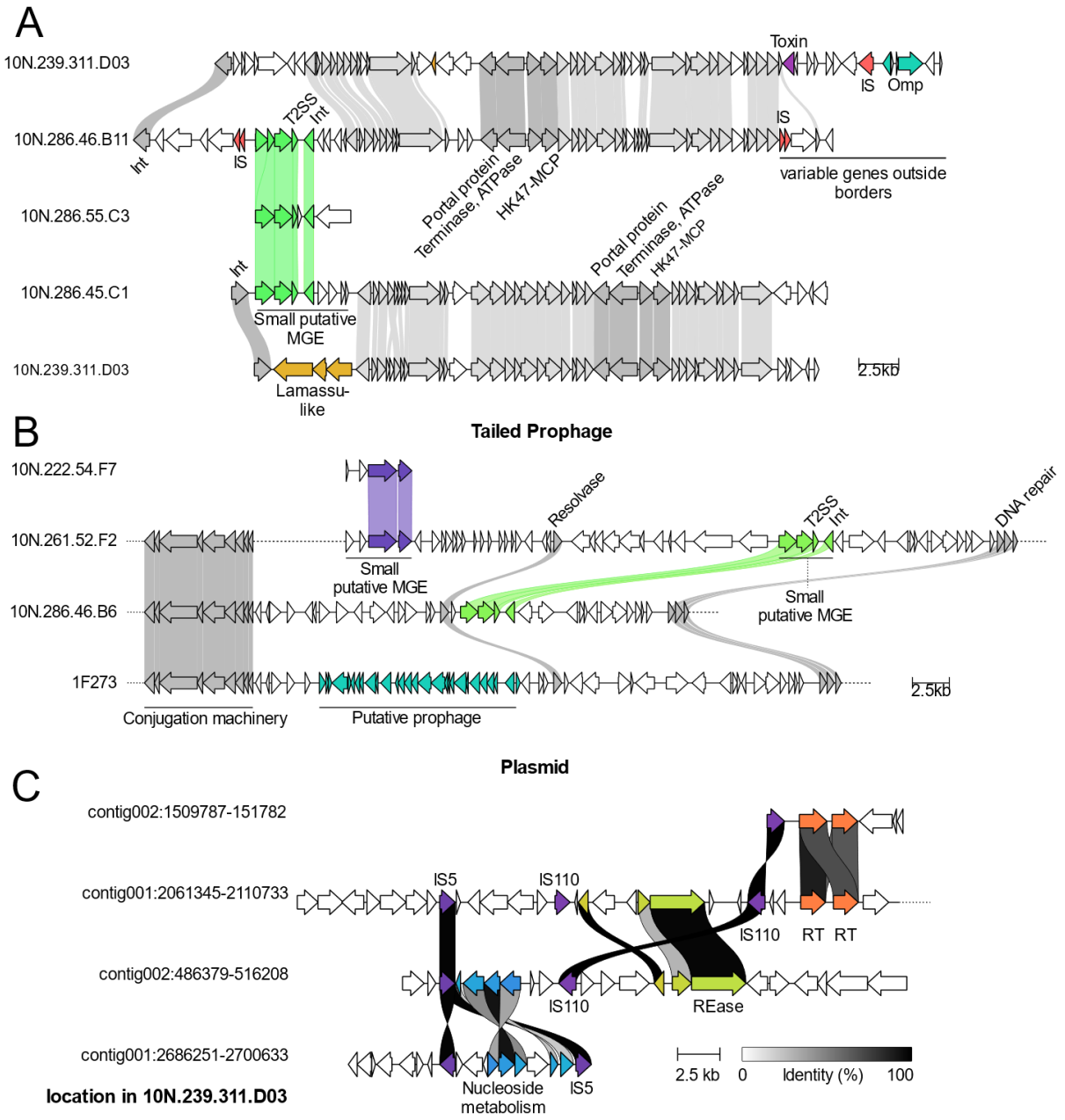

Figure 5 - MGEs accessory genes

Genomic maps of A) tailed prophages and B) plasmids sharing core proteins. Core proteins are depicted in gray, or light gray for core genes of subtypes. Connections indicate a sequence identity of at least 30%. Genes in the variable loci are colored in white with highlighted examples of functionally annotated genes. Nested MGEs are highlighted. Dashed lines indicate larger, not shown, sequence stretches on plasmids for improved visibility. C) Genomic map of elements with composite transposons carrying putative cargo in strain 10N.239.311.D03. Abbreviations: REase=restriction endonuclease, RT=reverse transcriptase, Int=integrase, T2SS=Type II secretion system, IS=insertion sequence, Omp=outer membrane protein.

Discussion

Pangenomes of many bacterial species expand continuously, with every additional sequenced genome adding novel genes to the flexible gene pool. Despite the notion that MGEs contribute to this genetic diversity, the extent of their contribution remains unclear. In this study, we describe a remarkable diversity and number of MGEs, averaging more than 20 MGEs per genome, of closely related bacterial strains of the genus *Vibrio* sampled from the same marine environment. We provide evidence that MGEs constitute the main contributors to the flexible gene pool in *Vibrio*. MGEs appeared in many different divergence clusters, many of which comprise highly variable accessory genes and even entire MGEs located in distinct loci. Both the numerous MGE clusters and the variable locus variation within MGE types are driving bacterial diversification. Our findings demonstrate that MGEs are of major importance for the variation in bacterial pangenomes and drive bacterial adaptation and evolution in the environment.

We created a comprehensive collection of precisely delimited MGEs from closely related *Vibrio* strains using a graph-based comparative genomics approach followed by MGE assignment based on their mobility signatures. We predicted differentially present MGEs and their precise boundaries in the host genome. Unlike current large-scale studies (15,16), we included MGEs that cannot mobilize autonomously, but encode specific motifs known to be involved in mobility. Furthermore, we used long-read generated genomes to improve MGE completeness and assess their genomic insertion sites. The total number of MGEs may still be underestimated due to unknown or unannotated mobility signatures as well as the inability to fully deconvolute all nested MGEs. Since the strains in this study are very closely related, our comparative genomics approach may not detect beneficial MGEs that have become fixed in the host species. Additionally, some identified MGEs may have lost their ability to mobilize, such as domesticated prophages (44).

Of the 2,600 MGEs, most could not be classified into known types. Other studies have addressed this issue by classifying MGEs based on their encoded recombinase type (15,16). However, not all MGEs carry recombinases. For example, some prophages contain only TIR or demonstrate sequence similarity to the *dif* site. We also observed that extensive nestedness further complicates MGE classification. More experimentally characterized elements and future advancements in the deconvolution of the nested MGEs will be needed to improve MGE classification.

IS elements were the most abundant MGEs in our dataset and were enriched in the flexible genome. Individual bacterial strains could contain over a hundred IS elements, which occasionally mobilize neighboring genes. Previous work suggested that while the presence of IS elements is associated with horizontal gene transfer, their abundance is not, as most IS elements in that study were outside horizontally acquired genomic regions (5,35). Both our observations and those from previous studies suggest two possible explanations for the observed enrichment of IS elements in the flexible genome. One possibility is that IS elements enter the host genome via horizontal gene transfer and may expand in the genome upon certain triggers activating transposable elements, such as stress responses (45). Another explanation is that the enrichment of IS elements may be due to selection acting against the disruption of essential genes in the core genome.

We determined two drivers responsible for the large genetic contribution of MGEs to the bacterial pangenome. One driver was the large number of different MGE clusters and singletons we observed in the set of closely related bacteria, also suggesting that previous studies may have underestimated MGE diversity due to the limited availability of high-quality bacterial genomes derived from deeply sampled environments (15). Certain MGE types, such as prophages and the candidate BY-Trio elements, appear to exist in many diverse subtypes, which often lacked closely related MGEs within the set of closely related bacterial hosts. These singletons may substantially contribute to the diversity of MGEs. The other driving force was the presence of accessory genes and even entire MGEs in specific locations on MGEs. While such variable loci have been found in prophages and phage satellites (33,42), we observed that variable loci are widespread among MGEs. Smaller MGEs, unable to horizontally transfer on their own, might thus be efficiently transferred by nesting into plasmids or prophages without affecting their

mobility. It is unclear how these genes are acquired by MGEs. Single genes might get exchanged by homologous recombination of the flanking regions, or MGEs residing in the same genomes insert into these loci. However, we observed accessory genes to be rather shared within similar MGE types than between MGEs infecting the same strain, consistent with previous observations of MGE type-specific accessory gene repertoires (46).

Even though we studied MGE diversity in a set of bacteria from the genus *Vibrio*, similarly high numbers and diversity of MGEs are likely to be found in bacterial communities from other environments. MGE diversity is driven by a range of types and subtypes as well as various accessory genes, which likely contribute to complex interactions between different MGEs. Being the major contributor to bacterial pangenome diversity, MGEs are key players in the ecology and evolution of bacteria.

Acknowledgements

We would like to acknowledge the former or current Polz lab members Kathryn Kauffman, Nina Bartlau, Fatima Hussain, Joseph Elsherbini and Phil Arevalo for isolation and sequencing of the bacterial strains used in this study. The computational results of this work have been generated using the Life Science Compute Cluster (LiSC) of the University of Vienna.

Conflict of interest

None declared.

Material and methods

Genome collection

To obtain intact MGEs not disrupted by assembly breaks, we used closed and nearly closed genomes that were publicly available or generated using Pacbio or Nanopore sequencing as described in previous studies (23,33). The dataset comprises 65 *Vibrio cyclitrophicus*, 43 *Vibrio lentus* and 19 *Vibrio splendidus* strains, together adding up to 127 genomes. Accession numbers of the respective strains are available in supplementary Table S1.

MGE prediction and gene annotation

The flexible genome was predicted using PPanGGOLiN (47) (v. 2.2.1) in panrgp mode per bacterial species and fasta files were created using samtools (48) (v.1.2.1.). We then predicted genes using prodigal (49) (v.2.6.3) with options -c and -meta. Genes were functionally annotated using hmmer (50) (v3.4) against the VOGdb (51) (v225) (e-value < 0.0001) to account for viral annotations. Furthermore, we ran Defensefinder (52) (v1.3.0) including the anti-defense option to predict defense and anti-defense genes. We ran interproscan (53) (v.5.69-101.0-11.0.4) for Pfam (54) and TIGRfam/NCBIfam (55) annotations (e-value <0.001). Conjugative genes were identified with CONJScan (56) (v.2.0.1) in plasmid mode for episomal and chromosome mode for chromosomal elements.

Mobility assignment

We assigned flexible genome regions to “signature of mobility” groups in a hierarchical manner. Thus, they were assigned to the first mobility signature group for which they were fulfilling all conditions, which were based on genomic location, gene content or DNA regions described in table 1. Elements were considered episomal if PPanGGOLiN could not assign a chromosomal spot of integration and its contig was shorter than 300 kb. Conjugative genes were identified using the CONJScan output. Pfam and VOGdb results were used to check for transposases and integrases. We ran integronfinder2 (57) (v.2.0.5) to search for clusters of attC sites lacking integrases (CALINs). Blastn (58) (v.2.16.0) with the options ”*-task blastn-short -word_size 12*“ was used to search for *dif* sites resembling known *dif* sequences from *V. cholerae* (dif1: AGTGCGTATTATGTATGTTATGTTAAAT, dif2:

AATGCGTATTACGTGCGTTATGTTAAAT) and terminal inverted repeats (TIR). Elements showing any of these hallmarks were considered as putative MGEs, the remaining elements were considered as non-MGEs and part of the flexible genome.

Table 1 - Signatures of mobility

#Group	Mobility group	Genomic location	Gene content	DNA signature
1	Episomal	Episomal	-	-
2	Chromosomal-conjugative	Chromosomal	Conjugation machinery	-
3	Transposase + integrase	Chromosomal	Transposase and Integrase	-
4	Transposase	Chromosomal	Transposase	-
5	Integrase	Chromosomal	Integrase	-
6	attC-rich	Chromosomal	-	attC sites
7	TIR	Chromosomal	-	TIR
8	dif	Chromosomal	-	dif site
9	No MGE signature	Chromosomal	-	-

Gene family proportion in the pangenome

To assign genes in the flexible genome to gene families, we clustered them using mmseqs2 (59) (v.17-b804f) in easy-cluster mode with an identity threshold of 0.95 and a coverage threshold of 0.8 in coverage mode 0. We repeated the workflow for genes in whole genomes.

For the rarefaction curve, we listed all gene family clusters encoded in either putative MGEs or the remaining elements in the flexible genome. We used a custom python script to create a rarefaction curve of unique gene families obtained with every additional genome as follows: We first mapped all gene family clusters to their respective host genomes. Then, we sampled genomes iteratively (up to n=1000) to obtain the total number of unique gene families. We calculated and plotted the mean count of unique gene families (y-axis) for each number of sampled genomes (x-axis).

For calculation of unique gene families in the different genomic fractions, we first counted the number of unique gene families encoded on the flexible genome (per strain of a species; and in the whole species). We then checked how many of them were found at least once on an identified putative MGE. To investigate how many gene families were encoded in the respective genomic fractions, we counted the total number of gene families in the whole genome (per strain of a species; and in the whole species) and subtracted the number of gene families encoded at least once on an MGE or in the remaining flexible genome.

Divergence clusters

To investigate the diversity of MGEs on sequence level, we dereplicated the 2,600 putative MGEs using seqkit dedup (60) (v.2.10.0) resulting in 2,474 unique putative MGEs. These were subjected to a pairwise average nucleotide identity (ANI) comparison using fastANI (61) (v.1.34) with an identity threshold of 0.95 or higher, and a coverage threshold of 0.4. We filtered the matches to only include connections where the small element covered more than 40% of the larger elements, to avoid the clustering of very large elements, such as plasmids, with small elements. The resulting connections were imported to Cytoscape (62) (v.3.10.3) to visualize the network with MGEs representing nodes, and MGE pairs above the ANI threshold connected by edges. Final divergence clusters based on the network connectivity (i.e. high ANI between putative MGEs) were obtained by running clusterMaker2 (63) (v.2.3.4) MCL clustering with an inflation value of 5.

MGE classification into known MGE types

To classify the putative MGEs into 6 known types (transposons, plasmids, prophages, ICE/IME, integrons and repeats, which are not necessarily MGEs), and the putative BY-Trio type, we manually inspected the ANI-based divergence clusters and the individual MGE's length, gene annotation or assignment by common prediction tools: When most putative MGEs in the divergence cluster were classified as one MGE type, we classified the whole divergence cluster as such. We allowed exceptions, when two or more MGEs were co-integrated. Plasmids were classified when putative MGEs were episomal and contained plasmid marker genes, such as relaxases identified in the Pfam and VOGdb results, or conjugative genes identified using the CONJScan output generated as described in the annotation section above. We classified putative MGEs as transposons, when they were short in size (<15 kb) and contained transposons as sole mobility signature, or when they contained more than two genes for transposition, such as transposases or genes with annotations associated with specific transposon types (e.g. TnsA, TnsE, Tn7). Prophages were identified using the marker genes for tailless, filamentous and tailed prophages as described previously (33). We identified repeats when genes or their domains were repetitive in a putative MGE. We identified ICEs and IMEs using ICEberg (64) (v.1.0). CALINs and complete integrons were searched by using integronfinder2 (57) (v.2.0.5). CALINs were only assigned if they were all integrated into the same insertion site, and often co-integrated with transposons, as it is a hallmark of certain integrons. The BY-Trio type was identified by the MGE type identification below. All other putative MGEs remained unclassified.

Synteny-based gene sharing clusters

To identify conserved proteins within related MGEs, we searched for putative MGEs sharing gene clusters of at least 3 genes with a maximum gap size of 5 genes using spacedust (65) (v.1-babc6f8). The network formation was performed for putative MGEs of each mobility signature group separately, a) to reduce the very high number of gene connections and b) as we assumed closely related MGEs to share similar mobility signatures. We observed many connections in the transposase+integrase group between otherwise unrelated elements due to IS elements. Therefore, we masked all IS element transposases based on Pfam and VOGdb annotations in this group from the spacedust calculation. Thus, we avoided that IS elements introduced noise in the cluster formation. We visualized the network of putative MGEs represented by nodes and their connections indicating their shared gene clusters in Cytoscape (62) (v.3.10.2). Then, we applied MCL clustering with an inflation factor of 5 using ClusterMaker2 (v.2.3.4) to define MGE clusters for core protein combination predictions based on their connectivity (i.e. the number of shared syntenic gene clusters).

MGE core protein combination prediction

We aimed to search for core proteins with a conserved function shared among several MGEs for the prediction of MGE types. As core proteins might have diversified in sub-types of MGEs while preserving their function, we first assigned every MGE-encoded gene to a protein family. Predicted genes were assigned to the same protein family if they were either found to be related by spacedust or clustered together with mmseqs2 when applying a clustering threshold of 0.5 and a coverage threshold of 0.9 in coverage mode 0.

We assigned a functional consensus prediction to each protein family based on the annotations of the individual genes within the gene family as follows: We considered the individual gene annotations calculcated as described above for the consensus annotation of the gene families: These included the VOGdb and Defensefinder annotations and the interproscan results (Pfam, TIGRfam/NCBIfam). Consensus annotations were assigned if at least half of the genes shared that annotation: If only one tool yielded a result, its annotation was used as consensus. Otherwise, Defensefinder annotations were prioritized. If no Defensefinder annotation was present, VOGdb annotations were prioritized if genes showed viral annotations, while Pfam or TIGRfam/NCBIfam annotations were used for all other genes based on manual investigation and lower e-values.

We predicted putative core protein clusters for each of the synteny-based gene-sharing clusters containing at least four MGEs as follows: We first created a binary matrix showing whether spacedust reported a specific protein family to be present in a shared gene cluster for the individual MGEs. For each cluster, we created a subset of this matrix only containing the cluster-specific elements. Each cluster-specific matrix was visualized as a heatmap, and if large clusters (>30 putative MGEs) contained very different sets of MGEs, we split them using the R stats cuttree function to ensure proper core protein combination prediction. All putative core protein combinations contained all fully present protein families of a cluster (present in at least 95% of all MGEs). To account for subtypes inside larger clusters, several predictions could be made by including additional highly present protein families (present in at least 3 MGEs) if their presence was positively correlated (pearson correlation of at least 0.7).

This way, several putative core protein combinations could be predicted for every synteny-based gene sharing cluster resulting in about 340 different core protein combinations of at least two genes, including subsets of themselves. We manually removed proteins from a small number of these core protein combinations, if they were predicted to belong to defense systems as they commonly form gene clusters of three and might be therefore false positives in the synteny approach.

MGE type prediction

To identify MGE types, we first inspected the functional annotation of the core genes for each of the predicted core protein combinations. All core protein combinations exclusively consisting of IS elements or defense systems were removed. Of the remaining combinations, we extracted MGEs containing all core proteins and manually inspected the aligned genomic maps of putative MGE types containing at least four MGEs. Notably, this approach is only appropriate for MGE types that are not extremely large (< 50-60 kb), as their complexity is often too high for this synteny-based gene sharing approach. All MGE types showing nearly no variation in the encoded genes were excluded from further analysis, as these might be vertically inherited MGEs. For core protein combinations that were subsets of each other, we tried to find a balance between showing genetic variation, while preserving similarity in parts of the putative MGE type. We manually excluded those MGE types with too high variation, for example when the core proteins were not in a similar gene order in the individual elements. Generally, we preferred combinations comprising a small number of proteins. This way, we downsampled the over 340 different core protein combinations yielding 30 putative MGE types.

IS element analysis

We identified IS elements using ISEScan (66) (v.1.7.2.3). IS elements that were recently transferred between MGEs encoded on the same bacterial strain were obtained by blasting genes encoded on all MGEs of single strains in an all-against-all search using bast+ (v.2.16.0) (58) with a similarity cutoff of 99.5%. We then manually inspected genomic maps of all MGEs sharing highly similar genes within individual bacterial strains.

Statistical tests

To test whether there were significant differences in MGE load, or accordingly in the abundance of IS elements, between bacterial species, we first checked the total size of all MGEs per strain in each species for normality using the Shapiro test. We then applied the Kruskal-Wallis test to check whether the three species were different, followed by pairwise Wilcoxon rank-sum tests between the three species.

We checked if there was a significant enrichment in IS elements in the flexible genome compared to the whole genome by applying the paired Wilcoxon test, as counts for IS elements in the flexible genome were dependent on their respective count in the whole genome.

All statistical tests were performed using R (67) (v.4.3.2) and R Studio (v.2024.12.1).

Computational tools and packages

All computational tools were used with standard parameters unless otherwise stated. For genomic maps, genbank files of MGE genomes were created using emboss (68) (v.6.6.0) and maps were visualized using clinker (69) (v. 0.0.31-3.13.2). Unless stated otherwise, downstream data processing and plot visualization was done using base R (67) (v.4.3.2) and R Studio (v.2024.12.1) with the packages ggplot2 (70) (v.3.5.1), pheatmap (71) (v.1.0.12), treemap (72) (v.2.4-4) and helper functions of dplyr (v.1.1.4), tidyr (v.1.3.1), readxl (v.1.4.3), reshape2 (v.1.4.4), grid (v.4.3.2) and purr (1.0.2). Plots, including gene colors and annotation labels, were adjusted manually using inkscape (v.1.1.2).

The use of python (v.3.10.16) scripts was indicated in the respective sections. We used the following packages: pandas (v.2.2.3), numpy (v.2.2.2), matplotlib (v.3.10.0), and seaborn (73) (v.0.13.2).

We used ChatGPT-4 to improve language for grammar and clarity, as well as for troubleshooting and development of python, R and bash code.

References

1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome." *Proc Natl Acad Sci USA*. 2005 Sep;102(39):13950–5.

2. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. *Nat Microbiol*. 2017 Mar;2(4):17040.

3. Shapiro BJ. The population genetics of pangenomes. *Nat Microbiol*. 2017 Nov;2(12):1574–1574.

4. Haudiquet M, De Sousa JM, Touchon M, Rocha EPC. Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Phil Trans R Soc B*. 2022 Oct;377(1861):20210234.

5. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol*. 2005 Sep;3(9):722–32.

6. Owen SV, Canals R, Wenner N, Hammarlöf DL, Kröger C, Hinton JCD. A window into lysogeny: revealing temperate phage biology with transcriptomics. *Microb Genom*. 2020 Feb;6(2).

7. Fullner KJ, Boucher JC, Hanes MA, Haines GK, Meehan BM, Walchle C, et al. The Contribution of Accessory Toxins of *Vibrio cholerae* O1 El Tor to the Proinflammatory Response in a Murine Pulmonary Cholera Model. *J Exp Med*. 2002 Jun;195(11):1455–62.

8. Bondy-Denomy J, Qian J, Westra ER, Buckling A, Guttman DS, Davidson AR, et al. Prophages mediate defense against phage infection through diverse mechanisms. *ISMEJ*. 2016 Jun;10(12):2854–66.

9. Burrus V, Marrero J, Waldor MK. The current ICE age: Biology and evolution of SXT-related integrating conjugative elements. *Plasmid*. 2006 May;55(3):173–83.

10. Krupovic M, Shmakov S, Makarova KS, Forterre P, Koonin EV. Recent Mobility of Casposons, Self-Synthesizing Transposons at the Origin of the CRISPR-Cas Immunity. *Genome Biol Evol*. 2016 Feb;8(2):375–86.

11. McKitterick AC, Seed KD. Anti-phage islands force their target phage to directly mediate island excision and spread. *Nat Commun*. 2018;9(1).

12. Barcia-Cruz R, Goudenège D, Moura De Sousa JA, Piel D, Marbouty M, Rocha EPC, et al. Phage-inducible chromosomal minimalist islands (PICMIs), a novel family of small marine satellites of virulent phages. *Nat Commun*. 2024 Jan;15(1):664.

13. Mahata T, Kanarek K, Goren MG, Marimuthu Ragavan R, Bosis E, Qimron U, et al. Gamma-Mobile-Trio systems are mobile elements rich in bacterial defensive and offensive tools. *Nat Microbiol*. 2024 Oct; 9:3268–3283

14. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS. A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation. *Cell Host Microbe*. 2020 Jan;27(1):140-153.e9.

15. Khedkar S, Smyshlyaev G, Letunic I, Maistrenko OM, Coelho LP, Orakov A, et al. Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Res*. 2022 Apr;50(6):3155–68.

16. Guo J, Aroney S, Dominguez-Huerta G, Smith D, Vik D, Ansah CO, et al. Mobile genetic elements that shape microbial diversity and functions in thawing permafrost soils. Preprint: *bioRxiv.* 2025.

17. Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol*. 2022 Apr;20(4):206–18.

18. Zünd M, Ruscheweyh HJ, Field CM, Meyer N, Cuenca M, Hoces D, et al. High throughput sequencing provides exact genomic locations of inducible prophages and accurate phage-to-host ratios in gut microbial strains. *Microbiome*. 2021;9(1):1–18.

19. Sitto F, Battistuzzi FU. Estimating Pangenomes with Roary. *Mol Biol Evol*. 2020 Mar 1;37(3):933–9.

20. Das B, Bischerour J, Barre FX. Molecular mechanism of acquisition of the cholera toxin genes. *Indian J Med Res*. 2011;133(2):195–200.

21. Ares-Arroyo M, Nucci A, Rocha EPC. Expanding the diversity of origin of transfer-containing sequences in mobilizable plasmids. *Nat Microbiol*. 2024 Nov; 9:3240–3253

22. Touchon M, Perrin A, Sousa JAM de, Vangchhia B, Burn S, O'Brien CL, et al. Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. *PLOS Genet*. 2020 Dec;16(6):e1008866.

23. Hussain FA, Dubert J, Elsherbini J, Murphy M, VanInsberghe D, Arevalo P, et al. Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science*. 2021;374(6566):488–92.

24. Gilcrease EB, Casjens SR. The genome sequence of Escherichia coli tailed phage D6 and the diversity of Enterobacteriales circular plasmid prophages. *Virology*. 2018 Feb;515:203–14.

25. Mai-Prochnow A, Hui JGK, Kjelleberg S, Rakonjac J, McDougald D, Rice SA. 'Big things in small packages: the genetics of filamentous phage and effects on fitness of their host.'. *FEMS Microbiol Rev*. 2015 Jul;39(4):465–87.

26. Scott JR, Kirchman PA, Caparon MG. An intermediate in transposition of the conjugative transposon Tn916. *Proc Natl Acad Sci USA*. 1988 Jul;85(13):4809–13.

27. Bardaji L, Añorga M, Jackson RW, Martínez-Bilbao A, Yanguas-Casás N, Murillo J. Miniature Transposable Sequences Are Frequently Mobilized in the Bacterial Plant Pathogen *Pseudomonas syringae* pv. phaseolicola. *PLoS ONE*. 2011 Oct;6(10):e25773.

28. Carnoy C, Roten CA. The *dif*/*Xer* Recombination Systems in Proteobacteria. *PLoS ONE*. 2009 Sep;4(9):e6531.

29. Huber KE, Waldor MK. Filamentous phage integration requires the host recombinases *XerC* and *XerD*. *Nature*. 2002 Jun;417(6889):656–9.

30. Mazel D. Integrons: agents of bacterial evolution. *Nat Rev Microbiol*. 2006 Aug;4(8):608–20.

31. Cury J, Jové T, Touchon M, Néron B, Rocha EP. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res*. 2016 Jun;44(10):4539–50.

32. Babakhani S, Oloomi M. Transposons: the agents of antibiotic resistance in bacteria. *J Basic Microbiol*. 2018 Nov;58(11):905–17.

33. Steensen K, Séneca J, Bartlau N, Yu XA, Hussain FA, Polz MF. Tailless and filamentous prophages are predominant in marine *Vibrio*. *ISMEJ*. 2024;18(1):wrae202.

34. Bellanger X, Payot S, Leblond-Bourget N, Guédon G. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol Rev*. 2014 Jul;38(4):720–60.

35. Touchon M, Rocha EPC. Causes of Insertion Sequences Abundance in Prokaryotic Genomes. *Mol Biol Evol*. 2007 Apr;24(4):969–81.

36. Van Houdt R, Monchy S, Leys N, Mergeay M. New mobile genetic elements in *Cupriavidus metallidurans* CH34, their possible roles and occurrence in other bacteria. *Antonie van Leeuwenhoek*. 2009 Aug;96(2):205–26.

37. Ricker N, Qian H, Fulthorpe RR. Phylogeny and organization of recombinase in trio (RIT) elements. *Plasmid*. 2013 Sep;70(2):226–39.

38. Bardaji L, Echeverría M, Rodríguez-Palenzuela P, Martínez-García PM, Murillo J. Four genes essential for recombination define GInts, a new type of mobile genomic island widespread in bacteria. *Sci Rep*. 2017 Apr 10;7(1):46254.

39. Koonin EV, Kuhn JH, Dolja VV, Krupovic M. Megataxonomy and global ecology of the virosphere. *ISMEJ*. 2024 Jan 10;wrad042.

40. Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brüssow H. Phage as agents of lateral gene transfer. Curr Opin Microbiol. 2003 Aug;6(4):417–24.

41. Benler S, Faure G, Altae-Tran H, Shmakov S, Zhang F, Koonin E. Cargo Genes of Tn *7* -Like Transposons Comprise an Enormous Diversity of Defense Systems, Mobile Genetic Elements, and Antibiotic Resistance Genes. *mBio*. 2021 Dec;12(6):e02938-21.

42. Rousset F, Depardieu F, Miele S, Dowding J, Laval AL, Lieberman E, et al. Phages and their satellites encode hotspots of antiviral systems. *Cell Host Microbe*. 2022 May;30(5):740-753.e5.

43. Getz LJ, Fairburn SR, Vivian Liu Y, Qian AL, Maxwell KL. Integrons are anti-phage defence libraries in Vibrio parahaemolyticus. *Nat Microbiol*.2025 Jan;10:724–733

44. Bobay LM, Touchon M, Rocha EPC. Pervasive domestication of defective prophages by bacteria. *Proc Natl Acad Sci USA*. 2014;111(33):12127–32.

45. Casacuberta E, González J. The impact of transposable elements in environmental adaptation. *Mol Ecol*. 2013 Mar;22(6):1503–17.

46. Takeuchi N, Hamada-Zhu S, Suzuki H. Prophages and plasmids can display opposite trends in the types of accessory genes they carry. *Proc R Soc B*. 2023 Jun;290(2001):20231088.

47. Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, et al. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput Biol*. 2020 Mar;16(3):e1007732.

48. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021 Jan;10(2):giab008.

49. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010 Dec;11(1):119.

50. Wheeler TJ, Eddy SR. Nhmmer: DNA homology search with profile HMMs. *Bioinformatics*. 2013;29(19):2487–9.

51. Trgovec-Greif L, Hellinger HJ, Mainguy J, Pfundner A, Frishman D, Kiening M, et al. VOGDB—Database of Virus Orthologous Groups. *Viruses*. 2024 Jul 25;16(8):1191.

52. Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun*. 2022 May 10;13(1):2561.

53. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014 May 1;30(9):1236–40.

54. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res*. 2021;49(D1):D412–9.

55. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, et al. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res*. 2021 Jan 8;49(D1):D1020–8.

56. Cury J, Abby SS, Doppelt-Azeroual O, Néron B, Rocha EP. Identifying conjugative plasmids and integrative conjugative elements with CONJscan. *Horizontal Gene Transfer: Methods and Protocols*. 2020;265–83.

57. Néron B, Littner E, Haudiquet M, Perrin A, Cury J, Rocha E. IntegronFinder 2.0: Identification and Analysis of Integrons across Bacteria, with a Focus on Antibiotic Resistance in *Klebsiella*. *Microorganisms*. 2022 Mar;10(4):700.

58. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009 Dec;10(1):421.

59. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. 2017;35(11):1026–8.

60. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS One*. 2016;11(10):e0163962.

61. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat commun*. 2018;9(1):5114.

62. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–504.

63. Utriainen M, Morris JH. clusterMaker2: a major update to clusterMaker, a multi-algorithm clustering app for Cytoscape. *BMC Bioinformatics*. 2023;24(1):134.

64. Liu M, Li X, Xie Y, Bi D, Sun J, Li J, et al. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res*. 2019;47(D1):D660–5.

65. Zhang R, Mirdita M, Söding J. De novo discovery of conserved gene clusters in microbial genomes with Spacedust. Preprint: bioRxiv.2024.

66. Xie Z, Tang H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics*. 2017 Nov 1;33(21):3340–7.

67. R Core Team R. R: A language and environment for statistical computing. 2013.

68. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet*. 2000 Jun 1;16(6):276–7.

69. Gilchrist CLM, Chooi YH. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics*. 2021 Aug 25;37(16):2473–5.

70. Wickham H. ggplot2: elegant graphics for data analysis. Second edition. *Springer*; 2016. 260 p. (Use R!).

71. Kolde R, Kolde MR. Package 'pheatmap.' R package. 2015;1(7):790.

72. Tennekes M, Ellis P. Package 'treemap.' R Package Version. 2017;2–4.

73. Bisong E. Matplotlib and seaborn. In: Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners. *Springer*; 2019. p. 151–65.

Supplementary Materials

Table S1 – Accession number of bacterial strains in this study

Strain	Version	Species	Bioproject	Accession
10N.222.45.E8	N03	Vibrio cyclitrophicus	PRJNA328102	CP170046-CP170047
10N.222.46.E12	N03	Vibrio cyclitrophicus	PRJNA328102	CP170589-CP170597
10N.222.46.F12	N03	Vibrio cyclitrophicus	PRJNA328102	CP170037-CP170045
10N.222.54.F11	N03	Vibrio cyclitrophicus	PRJNA328102	CP170034-CP170036
10N.237.312.C11	N03	Vibrio cyclitrophicus	PRJNA328102	JBFSQG000000000.2
10N.237.312.D05	N03	Vibrio cyclitrophicus	PRJNA328102	CP170054-CP170057
10N.237.312.G08	N03	Vibrio cyclitrophicus	PRJNA328102	CP170583-CP170584
10N.239.311.A09	N03	Vibrio cyclitrophicus	PRJNA328102	CP170058-CP170060
10N.239.311.B03	N02	Vibrio cyclitrophicus	PRJNA328102	JBFSQU000000000.2
10N.239.311.C11	N03	Vibrio cyclitrophicus	PRJNA328102	JBFSRB000000000.2
10N.239.311.D05	N02	Vibrio cyclitrophicus	PRJNA328102	JBFSRD000000000.2
10N.239.311.D07	N03	Vibrio cyclitrophicus	PRJNA328102	CP170064-CP170067
10N.239.311.E03	N02	Vibrio cyclitrophicus	PRJNA328102	JBGFTP000000000.2
10N.239.311.F08	N03	Vibrio cyclitrophicus	PRJNA328102	CP170068-CP170069
10N.239.311.G07	N01	Vibrio cyclitrophicus	PRJNA328102	JBFSRN000000000.2
10N.239.311.H11	N03	Vibrio cyclitrophicus	PRJNA328102	CP170070-CP170074
10N.239.312.A01	N03	Vibrio cyclitrophicus	PRJNA328102	CP170075-CP170076
10N.239.312.H01	N02	Vibrio cyclitrophicus	PRJNA328102	JBFSSI000000000.2
10N.247.311.62	N03	Vibrio cyclitrophicus	PRJNA328102	CP170077-CP170080
10N.247.311.85	N03	Vibrio cyclitrophicus	PRJNA328102	JBFSVD000000000.2
10N.261.45.F7	N03	Vibrio cyclitrophicus	PRJNA328102	CP170598-CP170600
10N.261.52.C4	N03	Vibrio cyclitrophicus	PRJNA328102	CP170031-CP170033
10N.261.54.B5	N03	Vibrio cyclitrophicus	PRJNA328102	MCXH00000000.2
10N.261.54.F1	N03	Vibrio cyclitrophicus	PRJNA328102	to be submitted
10N.261.55.A11	N02	Vibrio cyclitrophicus	PRJNA328102	MCXA00000000.2
10N.261.55.B9	N03	Vibrio cyclitrophicus	PRJNA328102	CP170028-CP170030
10N.261.55.E6	N03	Vibrio cyclitrophicus	PRJNA328102	CP170025-CP170027
10N.261.55.F9	N02	Vibrio cyclitrophicus	PRJNA328102	MCWP00000000.2
10N.286.45.B12	N02	Vibrio cyclitrophicus	PRJNA328102	MCWM00000000.2
10N.286.45.B7	N03	Vibrio cyclitrophicus	PRJNA328102	CP170022-CP170024
10N.286.45.F7	N02	Vibrio cyclitrophicus	PRJNA328102	JBHFAC000000000.2
10N.286.45.F8	N02	Vibrio cyclitrophicus	PRJNA328102	MCVY00000000.2
10N.286.46.B3	N03	Vibrio cyclitrophicus	PRJNA328102	to be submitted
10N.286.46.B6	N03	Vibrio cyclitrophicus	PRJNA328102	CP170050-CP170053
10N.286.46.F8	N03	Vibrio cyclitrophicus	PRJNA328102	MCVL00000000.2

Strain	Version	Species	Bioproject	Accession
10N.286.48.C5	N03	Vibrio cyclitrophicus	PRJNA328102	CP170048-CP170049
10N.286.48.E9	N03	Vibrio cyclitrophicus	PRJNA328102	to be submitted
10N.286.49.B11	N03	Vibrio cyclitrophicus	PRJNA328102	CP170019-CP170021
10N.286.49.E10	N03	Vibrio cyclitrophicus	PRJNA328102	to be submitted
10N.286.49.E3	N03	Vibrio cyclitrophicus	PRJNA328102	MCUQ00000000.2
10N.286.51.B1	N03	Vibrio cyclitrophicus	PRJNA328102	MCUH00000000.2
10N.286.52.C10	N03	Vibrio cyclitrophicus	PRJNA328102	CP170017-CP170018
10N.286.52.F6	N03	Vibrio cyclitrophicus	PRJNA328102	CP170015-CP170016
10N.286.54.C7	N03	Vibrio cyclitrophicus	PRJNA328102	CP170013-CP170014
10N.286.54.E11	N03	Vibrio cyclitrophicus	PRJNA328102	MCTE00000000.2
10N.286.54.E4	N03	Vibrio cyclitrophicus	PRJNA328102	CP170011-CP170012
10N.286.54.E5	N03	Vibrio cyclitrophicus	PRJNA328102	CP170587-CP170588
10N.286.54.E8	N03	Vibrio cyclitrophicus	PRJNA328102	MCSY00000000.2
10N.286.54.F2	N03	Vibrio cyclitrophicus	PRJNA328102	CP170009-CP170010
10N.286.54.F4	N03	Vibrio cyclitrophicus	PRJNA328102	CP170007-CP170008
10N.286.54.F9	N03	Vibrio cyclitrophicus	PRJNA328102	CP170005-CP170006
10N.286.55.A12	N03	Vibrio cyclitrophicus	PRJNA328102	MCSQ00000000.2
10N.286.55.A5	N03	Vibrio cyclitrophicus	PRJNA328102	CP170003-CP170004
10N.286.55.B4	N03	Vibrio cyclitrophicus	PRJNA328102	CP170001-CP170002
10N.286.55.C10	N03	Vibrio cyclitrophicus	PRJNA328102	CP169999-CP170000
10N.286.55.C3	N03	Vibrio cyclitrophicus	PRJNA328102	MCSG00000000.2
10N.286.55.C4	N03	Vibrio cyclitrophicus	PRJNA328102	CP169997-CP169998
10N.286.55.C7	N03	Vibrio cyclitrophicus	PRJNA328102	CP169994-CP169996
10N.286.55.E6	N03	Vibrio cyclitrophicus	PRJNA328102	CP170585-CP170586
10N.286.55.E8	N03	Vibrio cyclitrophicus	PRJNA328102	CP169992-CP169993
10N.286.55.F1	N03	Vibrio cyclitrophicus	PRJNA328102	CP169990-CP169991
1F273	N03	Vibrio cyclitrophicus	PRJNA81867	AIDC00000000.3
1F53	N03	Vibrio cyclitrophicus	PRJNA81861	AICZ00000000.3
ZF14	N03	Vibrio cyclitrophicus	PRJNA81877	CP166824-CP166825
ZF207	N03	Vibrio cyclitrophicus	PRJNA81891	AIDN00000000.2
10N.222.46.E3	P01	Vibrio lentus	PRJNA328102	JAAHTK000000000.1
10N.261.45.C3	N03	Vibrio lentus	PRJNA328102	to be submitted
10N.261.46.B5	P01	Vibrio lentus	PRJNA328102	MCZQ00000000.2
10N.261.49.F8	P01	Vibrio lentus	PRJNA328102	MCYP00000000.2
10N.261.51.A9	P01	Vibrio lentus	PRJNA328102	MCYN00000000.2
10N.261.52.F12	P01	Vibrio lentus	PRJNA328102	MCXR00000000.2
10N.261.52.F1	N03	Vibrio lentus	PRJNA328102	to be submitted
10N.261.52.F7	N02	Vibrio lentus	PRJNA328102	MCXM00000000.2

Strain	Version	Species	Bioproject	Accession
10N.261.54.A2	N03	Vibrio lentus	PRJNA328102	to be submitted
10N.261.55.C8	N04	Vibrio lentus	PRJNA328102	to be submitted
10N.261.55.E5	N02	Vibrio lentus	PRJNA328102	MCWT00000000.2
10N.261.55.E8	N02	Vibrio lentus	PRJNA328102	MCWR00000000.2
10N.286.45.B8	N02	Vibrio lentus	PRJNA328102	MCWI00000000.2
10N.286.45.B9	N02	Vibrio lentus	PRJNA328102	MCWH00000000.2
10N.286.45.C1	N02	Vibrio lentus	PRJNA328102	MCWG00000000.2
10N.286.45.C3	N02	Vibrio lentus	PRJNA328102	MCWF00000000.2
10N.286.45.C8	N02	Vibrio lentus	PRJNA328102	MCWE00000000.2
10N.286.45.E6	N02	Vibrio lentus	PRJNA328102	MCWC00000000.2
10N.286.45.E9	N02	Vibrio lentus	PRJNA328102	MCWB00000000.2
10N.286.46.A10	N02	Vibrio lentus	PRJNA328102	MCVX00000000.2
10N.286.46.A12	N02	Vibrio lentus	PRJNA328102	JBHOMU000000000.1
10N.286.46.B11	N02	Vibrio lentus	PRJNA328102	MCVU00000000.2
10N.286.46.E8	N02	Vibrio lentus	PRJNA328102	MCVM00000000.2
10N.286.48.A12	N02	Vibrio lentus	PRJNA328102	JAJGZN000000000.2
10N.286.48.A2	N02	Vibrio lentus	PRJNA328102	MCVK00000000.2
10N.286.48.A4	N02	Vibrio lentus	PRJNA328102	MCVJ00000000.2
10N.286.52.E12	P01	Vibrio lentus	PRJNA328102	JAJGZL000000000.1
10N.286.54.A11	P01	Vibrio lentus	PRJNA328102	MCTN00000000.2
10N.286.54.A1	P01	Vibrio lentus	PRJNA328102	MCTM00000000.2
10N.286.54.A6	P01	Vibrio lentus	PRJNA328102	MCTL00000000.2
10N.286.54.A8	P01	Vibrio lentus	PRJNA328102	MCTK00000000.2
10N.286.54.B1	P01	Vibrio lentus	PRJNA328102	MCTJ00000000.2
10N.286.54.B4	P01	Vibrio lentus	PRJNA328102	JAJGZK000000000.1
10N.286.54.B8	P01	Vibrio lentus	PRJNA328102	MCTI00000000.2
10N.286.54.E12	P01	Vibrio lentus	PRJNA328102	MCTD00000000.2
10N.286.54.E1	P01	Vibrio lentus	PRJNA328102	MCTC00000000.2
10N.286.54.E7	P01	Vibrio lentus	PRJNA328102	MCSZ00000000.2
10N.286.54.F6	P01	Vibrio lentus	PRJNA328102	MCSU00000000.2
10N.286.54.F7	N04	Vibrio lentus	PRJNA328102	to be submitted
10N.286.54.F8	P01	Vibrio lentus	PRJNA328102	to be submitted
10N.286.55.A1	N03	Vibrio lentus	PRJNA328102	to be submitted
10N.286.55.C12	N03	Vibrio lentus	PRJNA328102	to be submitted
10N.286.55.E1	P01	Vibrio lentus	PRJNA328102	MCSB00000000.2
10N.222.46.A6	N03	Vibrio splendidus	PRJNA328102	to be submitted
10N.222.46.B1	N03	Vibrio splendidus	PRJNA328102	to be submitted
10N.222.46.C2	N03	Vibrio splendidus	PRJNA328102	to be submitted

Strain	Version	Species	Bioproject	Accession
10N.222.51.B4	N03	Vibrio splendidus	PRJNA328102	to be submitted
10N.222.51.F8	N03	Vibrio splendidus	PRJNA328102	to be submitted
10N.222.52.A7	N03	Vibrio splendidus	PRJNA328102	to be submitted
10N.222.54.F7	N03	Vibrio splendidus	PRJNA328102	to be submitted
10N.239.311.D03	N03	Vibrio splendidus	PRJNA328102	CP170061-CP170063
10N.261.45.A2	N03	Vibrio splendidus	PRJNA328102	to be submitted
10N.261.48.B7	N03	Vibrio splendidus	PRJNA328102	to be submitted
10N.261.49.C11	P01	Vibrio splendidus	PRJNA328102	MCYV00000000.2
10N.261.52.F2	N02	Vibrio splendidus	PRJNA328102	MCXP00000000.2
10N.261.54.A1	N02	Vibrio splendidus	PRJNA328102	MCXK00000000.2
10N.261.54.F5	N02	Vibrio splendidus	PRJNA328102	MCXB00000000.2
10N.286.45.F12	N02	Vibrio splendidus	PRJNA328102	MCWA00000000.2
10N.286.46.A4	N02	Vibrio splendidus	PRJNA328102	MCVV00000000.2
10N.286.46.C9	N02	Vibrio splendidus	PRJNA328102	MCVP00000000.2
10N.286.52.C6	N03	Vibrio splendidus	PRJNA328102	to be submitted
10N.286.52.F10	N03	Vibrio splendidus	PRJNA328102	to be submitted

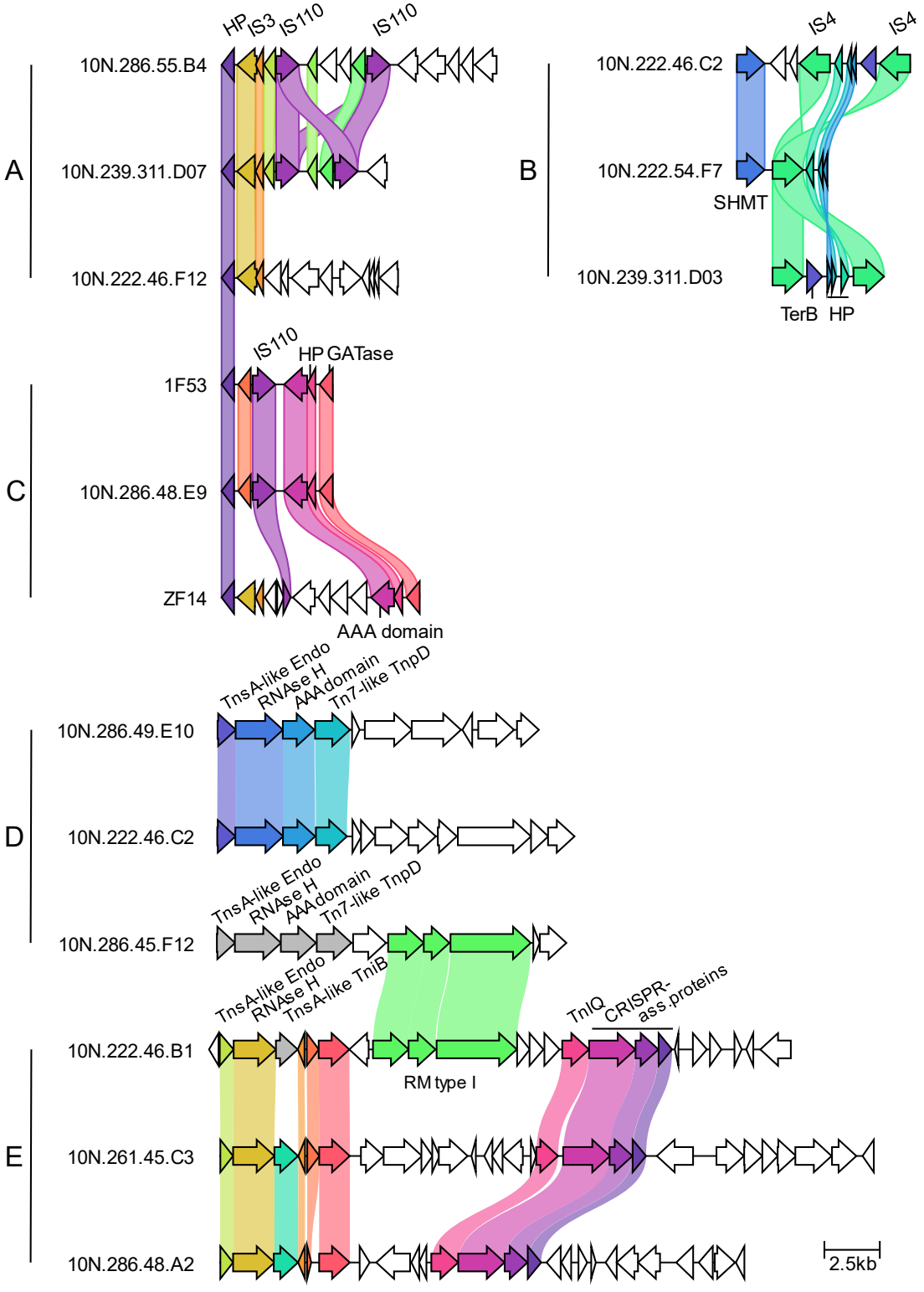

Figure S1 – Overview of transposons

Genomic maps of different candidate MGE types encoding core protein combinations. Hosts are denoted on the left side. Core protein annotations are displayed in diagonal orientation and selected accessory gene annotations in horizontal orientation. Proteins with an identity above 30% are connected. A) Putative nested transposons B) Composite transposons C) Transposons (unknown type due to a missing core gene annotation). D-E) Non-composite transposons. Abbreviations: IS = Insertion Sequence, GATase = glutamine amidotransferase, HP = hypothetical protein (no functional annotation), TerB = tellurite resistance protein, Endo = Endonuclease, RNase H = Ribonuclease H, SHMT = serine-hydroxymethyltransferase, CRISPR-ass.proteins= CRISPR-associated proteins.

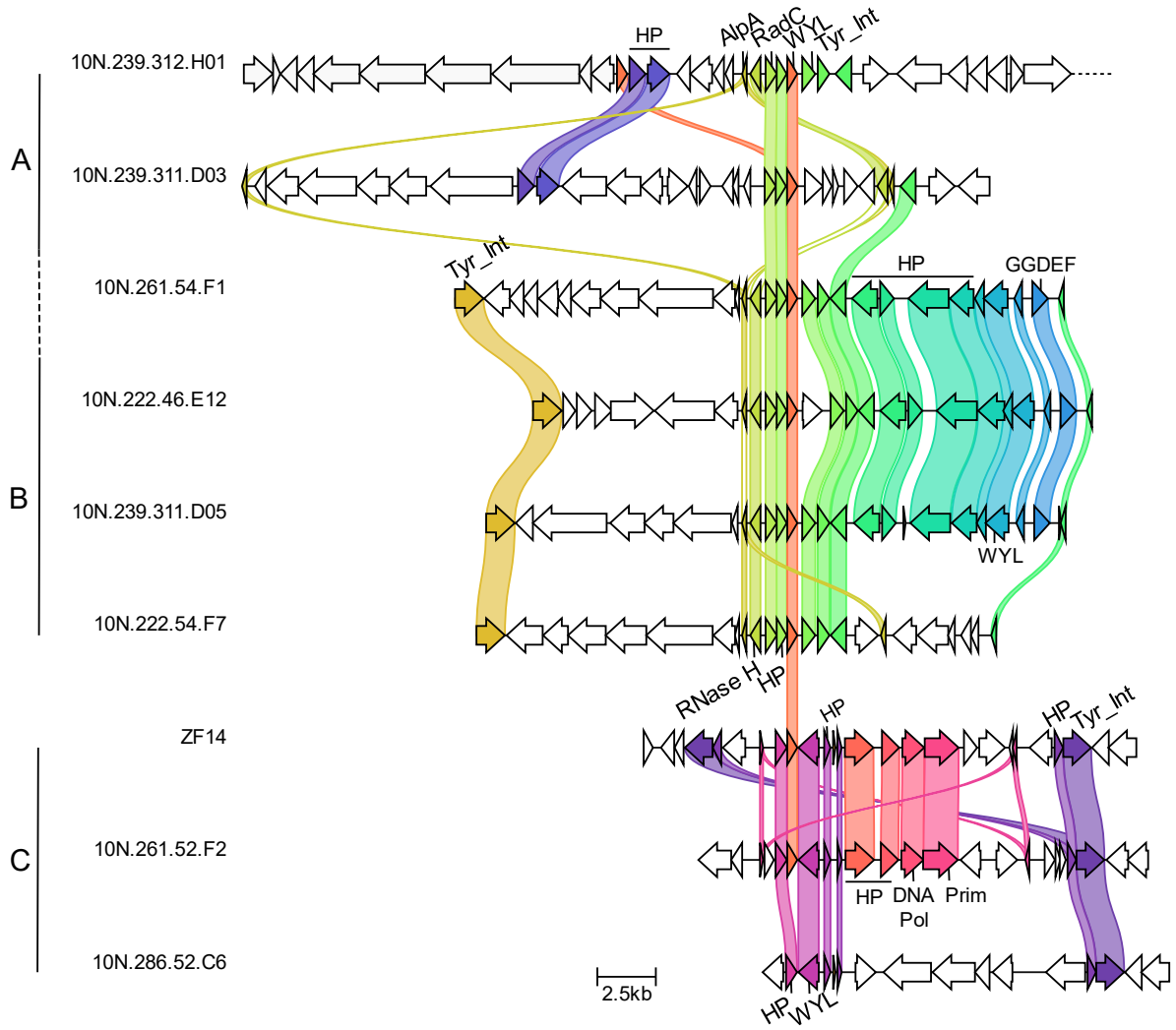

Figure S2 - Putative novel MGE types with shared core proteins
A-C) Genomic maps of different candidate MGE types encoding core protein combinations. Hosts are denoted on the left side. Core protein annotations are added in diagonal orientation. Selected accessory gene functional annotations in horizontal orientation. Dashed vertical lines indicate nested MGEs between two types, while dashed horizontal lines show that the flexible genome region is not fully displayed for better visualization. Proteins with an identity above 30% are connected. Abbreviations: AlpA = transcriptional activator AlpA, WYL = protein with WYL domain, RadC = DNA repair protein RadC, GGDEF = diguanylate cyclase, HP = hypothetical protein (no functional annotation), Tyr_Int = tyrosine integrase, DNA pol = DNA polymerase, Prim = primase.

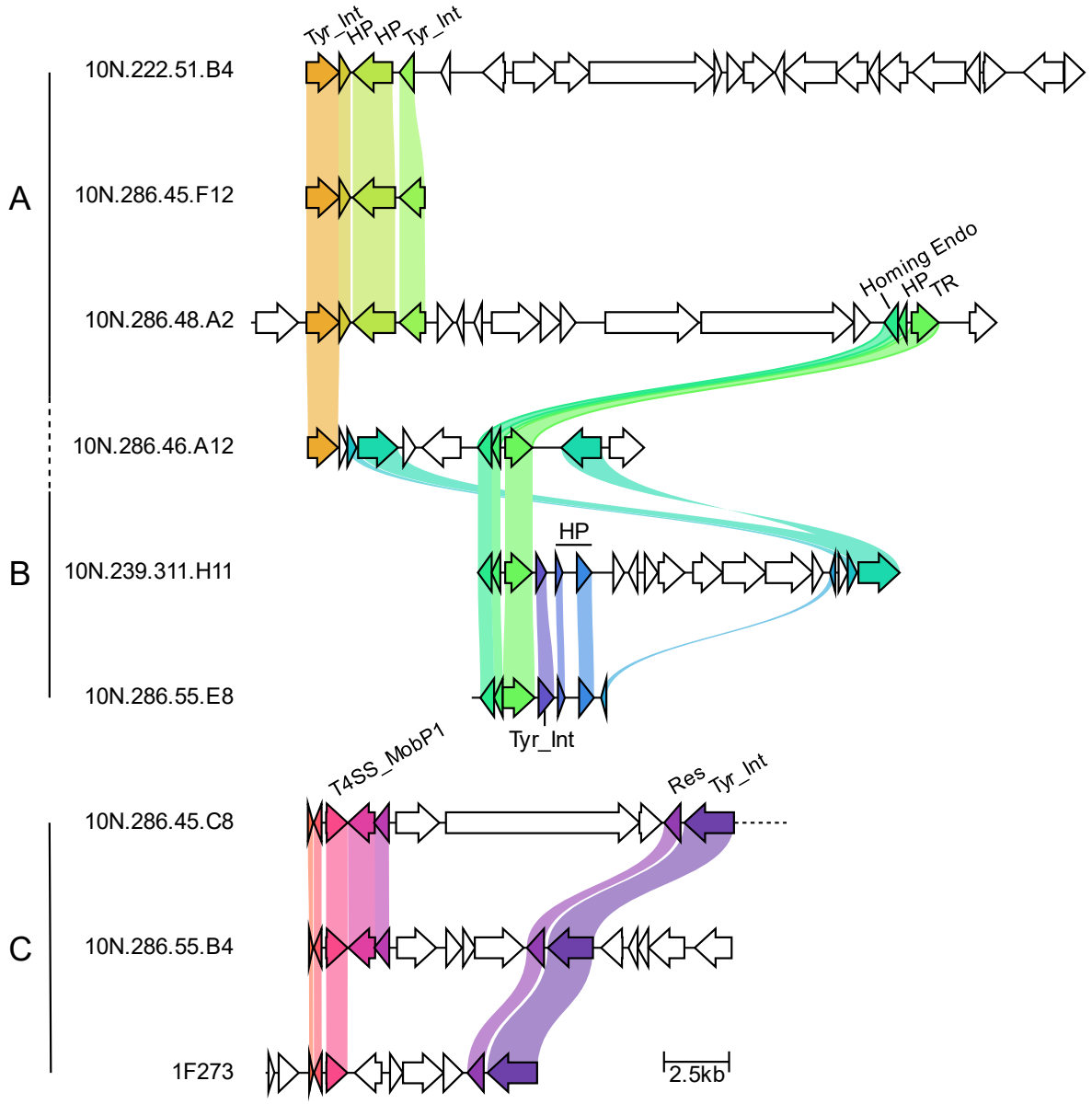

Figure S3 - Putative novel integrase-based MGE types
A-C) Genomic maps of different candidate MGE types encoding core protein combinations. Hosts are denoted on the left side. Core protein annotations are displayed in diagonal orientation and selected accessory gene functional annotations in horizontal orientation. Dashed vertical lines indicate nested MGEs between two types, while dashed horizontal lines show that the flexible genome region is not fully displayed for better visualization. Proteins with an identity above 30% are connected. Abbreviations: Tyr_Int = tyrosine integrase, HP = hypothetical protein, Endo = Endonuclease, TR = transcriptional regulator, T4SS_MobP1 = Type IV secretion system mobility protein 1, Res = Resolvase.

Conclusion and outlook

The large variation in bacterial pangenomes is thought to be caused by the horizontal acquisition of genetic material, part of which is driven by the rapid gain and loss of diverse MGEs. These mobile segments of DNA influence both their host and coexisting MGEs. For example, MGEs can affect the rate of further horizontal gene transfers (1) or provide accessory genes that alter their host's phenotype (2,3). Due to their impact, the computational prediction of MGEs has become increasingly important but remains biased toward the few known and well-described MGEs. Several studies have explored the diversity of known MGEs on a global scale. However, our understanding of their eco-evolutionary dynamics, specifically the abundance and diversity of MGEs, as well as their impact on hosts in the environment, remains limited. In this thesis, I investigated the abundance and diversity of MGEs and their contribution to bacterial diversification using a model system consisting of bacteria from the family *Vibrionaceae*, sampled from the same coastal ocean environment.

In chapter I, I investigated the abundance and diversity of prophages. I also aimed to include previously underrepresented prophages and to precisely locate their borders in the bacterial genome. Therefore, I applied a combination of comparative genomics and chemical prophage induction in the lab to predict prophages *de novo*. I identified and carefully curated 107 prophages found in 58 *V. cyclitrophicus* isolates. I computationally extended the search resulting in 1,158 prophages present in a collection of 931 bacterial strains from the *Vibrionaceae*.

I demonstrated that prophages resembling filamentous and tailless phages predominate in *Vibrionaceae* genomes, which is surprising given that tailed phages constitute the majority of phage genomes represented in public databases. The previous underestimation of tailless and filamentous prophages was also reflected in several prophages remaining undetected by commonly used prophage prediction tools. This work contributed to future prophage prediction improvements by increasing the available tailless and filamentous phage genomes in current databases. However, these results underscore the necessity for experimental verification of novel prophages, as the quality of computational predictions depends on the availability of known phage sequences.

While I used mitomycin C for prophage induction, an antibiotic known to trigger the SOS response in sublethal doses by causing DNA stress (4–6), not all prophages may be inducible via the SOS response. Recent studies have shown that prophages can also be induced independently of the SOS response, for instance, by natural compounds produced by other bacteria (7–9). Exploring alternative, environmentally relevant inducing agents may lead to the discovery of novel prophages and enhance our understanding of prophage induction in natural environments.

The identified prophages were highly diverse, as even very closely related strains did not necessarily share the same integrated prophages and identical prophages were rare. These observations raise the question from where this immense diversity originates. One possibility is the existence of a vast pool of lysogenic bacteriophages present in the environment, including those phages that are packaged into infectious viral particles. Another possibility is that these prophages might rapidly diversify while integrated in the host genome, for example by homologous recombination with other co-infecting MGEs. Which molecular mechanisms lead to this high diversity in prophages will require further investigation. The maintenance of this diversity, however, suggests that accessory genes in prophages provide a fitness advantage when they are rare, implying they might be under negative frequency-dependent selection. This might apply to different types of genes, such as toxins or phage defense genes. For instance, if a particular phage defense gene becomes highly abundant in a bacterial population, co-occurring bacteriophages may rapidly evolve counter-defenses, creating a dynamic referred to as the phage-host arms race (10).

The use of comparative genomics provided insights into the genomic insertion sites of prophages. Remarkably, tailed phages integrated into diverse and relatively infrequently occupied sites, while the

opposite was true for tailless and filamentous phages. For a subset of these phages, their tropism can be explained by their reliance on host-encoded recombinases *XerC* and *XerD*, which are known to mediate integration near the *dif* sites (11). However, this dependency alone is unlikely to fully explain this phenomenon, as both tailless and filamentous phage genomes often contain mobility genes. Furthermore, most tailless phages integrate into a different genomic site. The higher occupation of these few integration sites of tailless and filamentous phages might, however, lead to competition between phages targeting the same integration site, which was indeed observed as co-inserted, possibly disrupted, prophages. Future research could provide further insights into the interactions of MGEs targeting similar integration sites, such as whether co-integration affects the prophage's ability to induce.

Chapter II explored the diversity of MGEs in a model system of marine *Vibrio*, and I asked to what extent MGEs contribute to the flexible gene pool of their bacterial hosts. As in chapter I, I leveraged the close relatedness of the bacterial hosts to infer the precise boundaries of the flexible genome by comparing their - mostly closed - genomes to each other in a graph-based approach. Within the predicted contiguous regions of the flexible genome, I identified MGEs based on the presence of various signatures known to be involved in MGE mobility, such as recombinases and their recognition sites. This approach resulted in a collection of numerous precisely delimited and highly diverse MGEs, which were distinguished from other contiguous regions in the remaining flexible genome. These predictions allowed for a direct comparison of the genomic fractions accounting for MGEs, the remaining flexible genome, and the core genome of the host species.

MGEs were the main contributors to the flexible gene pool of the bacterial host species. I could identify two drivers for this diversity: the large number of different MGE types and subtypes, and highly variable accessory genes within MGE types. First, the large diversity within MGEs from a shared environment indicates their major role in bacterial adaptation and evolution in the environment. It also points to a large pool of MGEs present in the environment, which were postulated for prophages in chapter I as well. However, it is unclear how this large pool of MGEs could be maintained if not inside bacterial hosts. Even if MGEs might be residing outside the host either as free DNA or protected within capsids or vesicles (12), these are not necessarily stable over longer periods of time in the environment (13). Second, I observed a variety of accessory genes, and even entire MGEs, within various MGEs, mostly in distinct loci between conserved genes or at the border of MGEs. Related MGEs often exhibited completely different variable loci, posing the question of how these genes are exchanged at such a fast pace. In cases where entire MGEs are incorporated, future research could investigate whether they specifically target sites within the variable loci of larger MGEs as a strategy to travel along with them. Another possibility is the homologous recombination of the conserved regions flanking the variable loci, with other MGEs co-infecting the cell. Similar to variable regions in prophages, the variable loci that were widespread in MGEs are likely to be under negative-frequency dependent selection.

The classification of MGEs in chapter II covered only the best-studied groups of MGEs, such as plasmids, prophages, or integrons. Other MGE diversity studies generally covered similar categories, supplemented with other, individually defined categories such as phage-like elements (14–16). However, there is currently no widely used framework for MGE classification. This is problematic, as the results might not be easily reproducible nor comparable between studies. Furthermore, the field is challenged by individual descriptions of newly discovered MGE types (15). Although these MGE types are described, defining rules for computational prediction is difficult due to relatively few characterized sequences, or sequences limited to specific bacterial clades. Therefore, these less well studied MGEs are often not taken into consideration. Early attempts also included the formation of an MGE-specific database (17), which has not been updated in recent years. To date, targeted prediction tools for individual MGE types dominate the field. Generally, phage prediction tools are most common (18,19), and other tools targeting specific MGEs like ICEs (20) or integrons (21) exist. The growing interest in

MGE research may eventually lead to combined prediction tools, or specific, well-curated MGE databases, which could improve MGE classification and make it more reproducible.

Improvements in the classification of MGEs may also lead to the identification of further commonalities and differences among MGE types. For example, we identified variable loci in a range of different MGE types, suggesting the potentially universal fitness advantage of these variable loci. However, there are indications that the variable gene repertoire differs among MGE types, such as accessory gene sharing between BY-Trio elements, which were unrelated in their integrases, or as demonstrated in another study (22). Furthermore, there could be differences in the diversity of different MGE types. For example, tailed phages and BY-Trio elements were highly diverse, since they were more commonly singletons in our clustering analysis. Identifying such similarities and differences of MGEs can provide important knowledge on the underlying biology of MGEs and their interactions in nature.

Recent work has improved our understanding of the general diversity of MGEs. However, most studies focused on prediction of MGEs on a global scale (15,23) or on MGEs derived from microbes confined by their taxonomy, such as pathogenic or model species (14,24). In this thesis, I approach MGE diversity from an eco-evolutionary angle by using genomes of closely related bacteria originating from the same environment. This approach improves the quality and completeness of the predicted MGEs, the depth of genomic investigations, and provides insights into the ecological impact of MGEs. By precise prediction of the flexible genome including known and, so far, unclassified MGEs, I am expanding the previous knowledge by several means: I show that the previously underappreciated filamentous and tailless prophages are predominant in *Vibrio*. Furthermore, I demonstrate that MGEs represent the main contributors to the flexible gene pool in *Vibrio*. I reveal two drivers of bacterial diversification conferred by MGEs: 1) the variety of different MGE types and subtypes, and 2) the highly variable loci encoded in distinct locations of many MGEs. Most importantly, I demonstrate the remarkable diversity of MGEs in closely related bacteria isolated from the same environment. This high diversity indicates the large impact of MGEs on the eco-evolutionary dynamics of bacteria in natural environments. Although these studies were performed in marine *Vibrionaceae*, the findings are likely applicable to many other environments, including freshwater systems, soil, or the human microbiome.

There are still many open questions beyond the scope of this thesis, particularly those concerning the highly diverse accessory genes encoded on MGEs: What is the function of these genes? And where do these genes originate from? The question about their function is not new, as the lack of annotation of the numerous accessory genes on MGEs has intrigued many researchers (3,25). In fact, some studies have discussed the existence of so-called ORFans on MGEs (26,27), which are genes that lack homology to any other genes in more distantly related bacteria. In recent years, the discovery of phage defenses and anti-defenses has vastly contributed to the functional annotations of these accessory genes (28,29). However, there are still numerous genes without functional annotation. Given that MGEs carry many phage defense genes, it is possible that they also defend against other MGEs. This hypothesis needs to be tested with future laboratory experiments similar to those in phage defense studies. These studies will require the development of other high-throughput screening methods, as other MGEs do not show easily detectable inhibition zones (so-called plaques) on bacterial lawns. The question of how these accessory genes are acquired is even more challenging. As demonstrated in this thesis, the diversity of the variable genes in MGEs is very high, indicating a fast diversification of MGEs. Given that the *Vibrio* collection is deeply sampled, this question might not be computationally solvable with the current techniques. Advances in long-read metagenomic sequencing, or experimental investigation of the physical interactions between MGEs might help gain better insights in the future.

For many decades, transposable elements have been regarded as being selfish and recent literature has expanded this concept to other MGEs (1,30). This concept might help to explain many of the observations of MGE abundance and diversity, whereas the question of whether MGEs are selfish or not is almost philosophical. However, the field might benefit from viewing MGEs as individual biological

entities, much like many other organisms in nature. This perspective would prevent overlooking possible mutualisms that have formed during co-evolution with their hosts. Bacteria could be perceived as individual ecosystems, inhabited by different MGEs. Like other organisms, MGEs interact both with their environment (i.e., the host), and with each other. When conditions allow, MGEs might remain within an environment. Alternatively, MGEs could go extinct or transfer to other environments in search of better conditions to ensure their survival. Whether selfish or not, future work will need to delineate the complex interactions MGEs undergo to fully understand how MGEs drive the ecology and evolution of bacteria.

References

1. Haudiquet M, De Sousa JM, Touchon M, Rocha EPC. Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Phil Trans R Soc B*. 2022 Oct 10;377(1861):20210234.

2. Fullner KJ, Boucher JC, Hanes MA, Haines GK, Meehan BM, Walchle C, et al. The Contribution of Accessory Toxins of *Vibrio cholerae* O1 El Tor to the Proinflammatory Response in a Murine Pulmonary Cholera Model. *J Exp Med*. 2002 Jun 3;195(11):1455–62.

3. Cumby N, Davidson AR, Maxwell KL. The moron comes of age. *Bacteriophage*. 2012 Oct;2(4):e23146.

4. Levine M. Effect of mitomycin C on interactions between temperate phages and bacteria. *Virology*. 1961 Apr 1;13(4):493–9.

5. Jiang SC, Paul JH. Seasonal and diel abundance of viruses and occurrence of lysogeny/bacteriocinogeny in the marine environment. *Mar Ecol Prog Ser*. 1994;163–72.

6. Tomasz M. Mitomycin C: small, fast and deadly (but very selective). *Chem Biol*. 1995;2(9):575–9.

7. Selva L, Viana D, Regev-Yochay G, Trzcinski K, Corpa JM, Lasa Í, et al. Killing niche competitors by remote-control bacteriophage induction. *Proc Natl Acad Sci USA*. 2009;106(4):1234–8.

8. Silpe JE, Bassler BL. A Host-Produced Quorum-Sensing Autoinducer Controls a Phage Lysis-Lysogeny Decision. *Cell*. 2019;176(1–2):268-280.e13.

9. Jancheva M, Böttcher T. A Metabolite of Pseudomonas Triggers Prophage-Selective Lysogenic to Lytic Conversion in *Staphylococcus aureus*. *J Am Chem Soc*. 2021;143(22):8344–51.

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

11. Huber KE, Waldor MK. Filamentous phage integration requires the host recombinases XerC and XerD. *Nature*. 2002 Jun;417(6889):656–9.

12. Biller SJ, Schubotz F, Roggensack SE, Thompson AW, Summons RE, Chisholm SW. Bacterial Vesicles in Marine Ecosystems. *Science*. 2014 Jan 10;343(6167):183–6.

13. Noble RT, Fuhrman JA. Virus decay and its causes in coastal waters. *Appl Environ Microbiol*. 1997 Jan;63(1):77–83.

14. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS. A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation. *Cell Host Microbe*. 2020 Jan;27(1):140-153.e9.

15. Khedkar S, Smyshlyaev G, Letunic I, Maistrenko OM, Coelho LP, Orakov A, et al. Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Res*. 2022 Apr 8;50(6):3155–68.

16. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. *Nat Biotechnol*. 2024 Aug;42(8):1303–12.

17. Leplae R. ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic Acids Res*. 2004 Jan;32(90001):45D – 49.

18. Kieft K, Zhou Z, Anantharaman K. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*. 2020;8(1):1–23.

19. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*. 2021 Dec;9(1):37.

20. Liu M, Li X, Xie Y, Bi D, Sun J, Li J, et al. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res*. 2019;47(D1):D660–5.

21. Néron B, Littner E, Haudiquet M, Perrin A, Cury J, Rocha E. IntegronFinder 2.0: Identification and Analysis of Integrons across Bacteria, with a Focus on Antibiotic Resistance in *Klebsiella*. *Microorganisms*. 2022 Mar;10(4):700.

22. Takeuchi N, Hamada-Zhu S, Suzuki H. Prophages and plasmids can display opposite trends in the types of accessory genes they carry. *Proc R Soc B*. 2023 Jun;290(2001):20231088.

23. López-Leal G, Camelo-Valera LC, Hurtado-Ramírez JM, Verleyen J, Castillo-Ramírez S, Reyes-Muñoz A. Mining of Thousands of Prokaryotic Genomes Reveals High Abundance of Prophages with a Strictly Narrow Host Range. *mSystems*. 2022 Aug 30;7(4):e00326-22.

24. Hochhauser D, Millman A, Sorek R. The defence island repertoire of the *Escherichia coli* pan-genome. *PLoS Genet*. 2023;19(4):e1010694

25. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol*. 2005 Sep;3(9):722–32.

26. Daubin V, Ochman H. Bacterial Genomes as New Gene Homes: The Genealogy of ORFans in *E. coli*. *Genome Res*. 2004 Jun;14(6):1036–42.

27. Yin Y, Fischer D. Identification and investigation of ORFans in the viral world. *BMC Genomics*. 2008;9(1):24.

28. Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun*. 2022 May 10;13(1):2561.

29. Tesson F, Huiting E, Wei L, Ren J, Johnson MC, Planel R, et al. Exploring the diversity of anti-defense systems across prokaryotes, phages, and mobile genetic elements. *Nucleic Acids Res*. 2025; 53(1):gkae1171

30. Feschotte C. Transposable elements: McClintock's legacy revisited. *Nat Rev Genet*. 2023 Nov;24(11):797–800.