Identifying functional groups in microbial communities based on ecological patterns

By

Xiaoyu Shan

B.E. Zhejiang University (2016)

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

Doctor of Philosophy in Civil and Environmental Engineering

at the

Massachusetts Institute of Technology

September 2023

Authored by: Xiaoyu Shan
Department of Civil and Environmental Engineering
August 11, 2023

Certified by: Otto X. Cordero
Associate Professor of Civil and Environmental Engineering
Thesis supervisor

Accepted by: Heidi Nepf
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Identifying functional groups in microbial communities based on ecological patterns

By

Xiaoyu Shan

Submitted to the Department of Civil and Environmental Engineering
on August 11, 2023, in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy in Civil and Environmental Engineering

Abstract

Recent development in sequencing technologies has greatly advanced our understandings of structure and function of microbial communities in various ecosystems. In microbial communities, a metabolic function is often performed by a group of multiple species (i.e., a functional group) at the same time. However, identifying these functional groups remains to be a major challenge for structure-function mapping in microbiome studies. Instead of relying on annotation-based methods that are highly biased for a few model microorganisms, here I tackle this challenge by developing a novel annotation-free approach. In chapter two, I develop the mathematical framework behind the new approach – which we call EQO – and show its power by applying it to a few existing microbiome datasets. I show that, based solely on the patterns of statistical variation in species abundances, EQO identifies functional groups in soil, ocean and animal gut microbiome. The following two chapters discuss an application of this method, which has led to the discovery a potential new form of interaction between bacteria in animal guts, and an unexpected finding in the lab regarding the ecological dynamics of phage-plasmids in marine bacterial populations. In chapter three, I show how applying EQO to an aquaculture dataset leads us to identify potential pathogen-inhibiting groups of bacteria in an animal-associated microbiome. Guided by the computational prediction, I successfully isolate a member of this group that is a novel species with a broad spectrum of interaction against various *Vibrio* pathogens. By synthesizing and secreting polysaccharides, the novel species causes limited dispersion and reduced virulence of *Vibrio*. My efforts to understand the ecology of marine bacteria also lead me to study the role of widely distributed phage-plasmids. Combining mathematical models and experimental evidence, I show that loss-of-function mutations and segregational drift recurrently drive productive infections of phage-plasmids within marine bacterial populations. Together, this thesis provides a simple yet powerful approach to abstract functional groups from taxonomic composition in complex microbiome. As a useful hypothesis generating tool, this approach will pave the way for more mechanistic studies of microbiome in the future.

Thesis supervisor: Otto X. Cordero
Title: Associate Professor

Acknowledgement

The whole story started from the moment I knocked at the door of 48-429. Since then, I have been immensely grateful to Professor Otto Cordero for giving me an opportunity to pursue scientific research with him. His mentorship extends far beyond his vision of evolutionary ecology; he also devotes countless hours to helping me improve my oral and written presentations. Moreover, he offered warm care and strong support during the dark and chaotic times in the early period of the pandemic. I am also very appreciative of all the current and former members of the Cordero lab - Tim Enke, Manoshi Datta, Elise Ledieu, Rachel Szabo, Annika Gomez, Suryateja Jammalamadaka, Gabriel Vercelli; Gabriel Leventhal, Julia Schwartzman, Ali Ebrahimi, Leonora Bittleston, Shaul Pollak, Matti Gralka, Rachel Gregor, Andreas Sichert, Stefany Moreno Gamez; Patrizia Stadler, Denise Stewart, Lu Lu, Philip Wasson and Samantha Pennino. Special thanks go to Chuliang Song for introducing me to the Cordero lab.

I am more than fortunate to have Professors Tami Lieberman, Jeff Gore, and Anthony Sinskey in my thesis committee, who generously share with me their insights. Professor Tami Lieberman deserves a special mention for an unforgettable semester when I was a teaching assistant for her class, *Genomics and Evolution of Infectious Disease*. It is my great pleasure to work closely with all members in the Principles of Microbial Ecosystems (PriME) collaboration funded by the Simons Foundation. In particular, I would like to express my gratitude to Professor Terence Hwa at UCSD for regular meetings, which enhanced my understanding of microbial physiology. I am indebted to Professors Hung Cheng, Gilbert Strang and Manolis Kellis at MIT, whose classes laid a solid foundation for my quantitative skills in research. I would like to thank Professors Salvador Almagro-Moreno, Lone Gram, Alvaro San Millan, Itzik Mizrahi, Mikhail Tikhonov, Sallie Chisholm, Laura Kiessling, Alan Grossman, Martin Polz, Michael Follows, Daniel Rothman, Serguei Saavedra, Darcy McRose, Jonathan Friedman, Stilianos Louca, as well as Fatima Aysha Hussain, Yuanqiao Rao, Wei Gao, Akshit Goyal, Joshua Goldford, Victoria Marando, Jiliang Hu, Ben Liu, Qingyi Wang, Jie Yun, Shiwei Wang, Qi Xie, Jane Maureen Jayakumar, Vivian Cruz-Lopez, Omar Tantawi, Marc Foster, Xiao Huan, Ruiqi Xiao, Jie Deng, Ruijiao Sun, Wentao Huang, Xiaoqian Wan, Dongdong Xu, Nichakan Khuichad, Hongyu Chen and many others for warm help and illuminating discussions in scientific research. I also want to thank Professors Yunfeng Yang, Daliang Ning and Jizhong Zhou for their invaluable guidance into microbial ecology during my earlier years at Tsinghua.

Last but not least, I want to thank Dr. Shengchun Zhou, Professor of History at Zhejiang University, for his life-changing education. It was he who introduced me to MIT. I received his phone call at the end of 2021, when he asked me how my research went at MIT. At that time, I told him that I was working on demystifying a phage in the lab. He sounded pleased, saying that I should let him know when I solve the puzzle. I promised him, not realizing that he was already in a hospital bed then. After dedicating his whole life to the liberal education of more than 400 Morningside Scholars, he passed away on July 20, 2022, at the age of 75. When the phage study finally got published in early 2023, I no longer had a chance to let him know. Throughout his life, he taught students not only intellectual excellence but also humane decency. Students are what he loves and cares for; and he will be respected and remembered by the students.

Table of Contents

List of figures

List of tables

Chapter 1 Introduction

This chapter is part of the following perspective paper in preparation.

Finding the right variables in microbial communities. Xiaoyu Shan*, Akshit Goyal*, Mikhail Tikhonov, Otto X. Cordero. In preparation. (* equal contribution)

1.1 Background

One of the main goals in microbial ecology is to gain a predictive understanding of how changes in community composition, or structure, relate to changes in metabolic processes (e.g., rates of organic matter degradation, concentrations of fermentation byproducts, etc.)[1,2]. Doing so requires mapping community structure to function, which would convert DNA sequencing into an environmental sensor and enable rational design of synthetic communities[3,4]. Without predictive structure-function maps, microbiome engineering remains a slow grind due to the high dimensionality of microbial communities.

But we are still far from having well-established, general strategies to predict a microbiome's function from data about its (metagenomic) structure. For instance, we cannot currently predict whether a soil community will release or capture net CO_2 solely from its metagenome, let alone estimating the flux[5]. Such predictions may never be perfect, as there are many features, such as phenotypic plasticity dependent on environmental context, that determine community function besides species composition and genetic makeup. However, there is reason to believe, as explained below, that the

genomic structure of communities, and even just its species composition, holds a significant amount of information that could be used to predict metabolic fluxes.

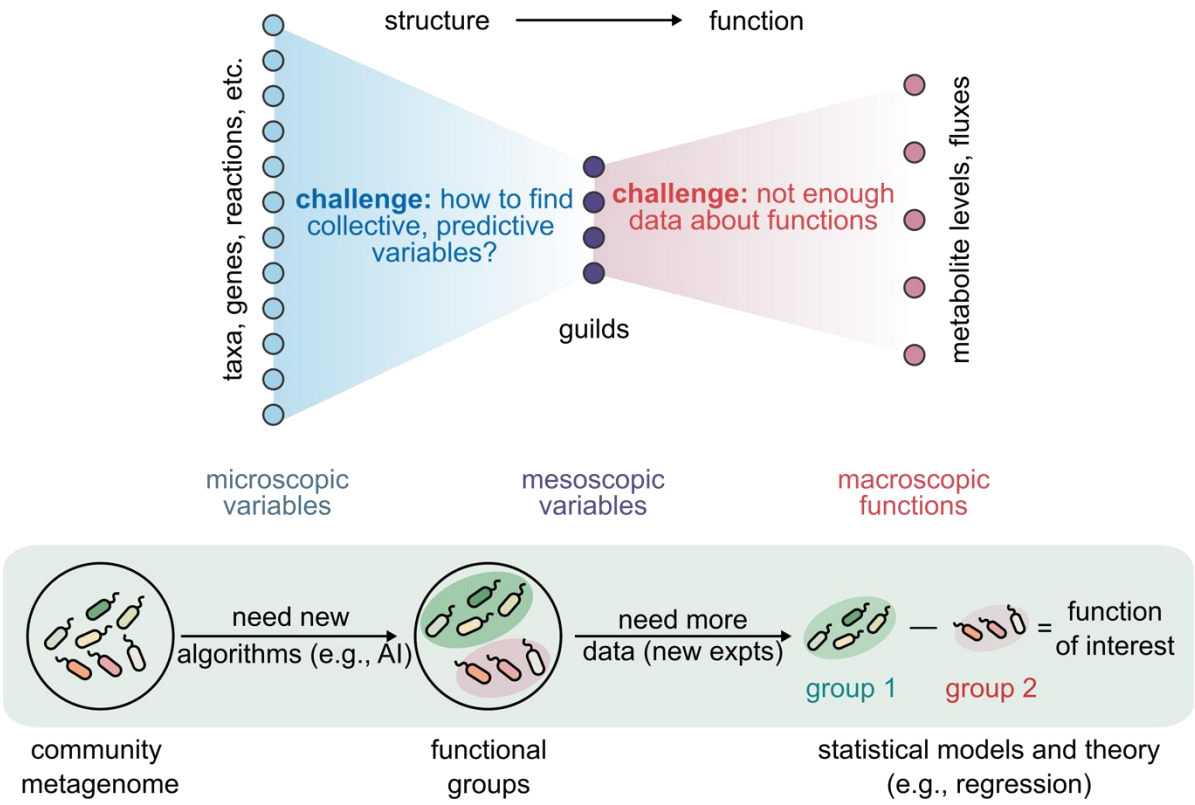


Figure 1.1 Solving the microbiome structure-function problem using mesoscopic variables. Solving the structure-function problem in microbial communities requires mapping community structure (i.e., composition) with some aspect of its function (e.g., metabolic fluxes). However, we can describe community structure either in microscopic detail (strains, genes, etc.) or in coarser, mesoscopic detail (guilds or functional groups). To predict function from structure, we argue that mesoscopic approaches in microbial ecology have lost traction in the age of 'omics, and ought to be reinvigorated. We identify two big challenges facing such approaches: the lack of data on community function, and the paucity of algorithms, methods and theory, to coarse-grain communities and detect functional groups.

Given the obvious importance of being able to predict ecosystem function from microbiome structure, why haven't we been able to solve this problem? We argue in this perspective that there are at least two barriers, stemming from historical contingencies in the field: the focus on microscopic variables and the lack of data on community

function (Fig 1.1). To overcome the first barrier, we advocate a change in approach — to coarsen the level of description of both structure and function, from microscopic to mesoscopic. To overcome the second one, we provide a few concrete suggestions on how to enrich microbiome surveys with functional data.

1.2 Microscopic variables and their limitation in microbial ecology

The structure of microbial communities can be quantified using a variety of descriptors, implicitly defined at specific levels of biological organization. For instance, genomic descriptors range from genes and strains[6,7] at the lowest level of biological organization, to metabolic pathways, whole genomes or phylogenetic groups at a level higher. Likewise, the activity of communities can be studied at the level of single cells[8], or at the level of the whole biome[9]. Although there is no single level of organization that is an intrinsically 'better' than the others for quantifying structure, the focus on molecular processes has driven the development of new technologies primarily towards the lowest levels of biological organization (molecules, genes, strain variants), or what we refer here to as *microscopic variables*. Accordingly, quantitative approaches to map structure to function also tend to be quite microscopic. For instance, genome-scale metabolic models require a detailed description of an organism's genomic structure to predict fluxes[10–12], while ecological models, such as those based on consumer-resource interactions require a metabolite consumption and excretion network to infer community dynamics[13–15]. These approaches have been successful in predicting growth rates in organisms like *E. coli* under different environments[16] or explaining how multiple species can coexist[17–19], to name a few. However, when it comes to the highly diverse

communities found in nature, we claim that it would be challenging to only use microscopic approaches to build a predictive understanding.

To a reader who is familiar with the enormous power that molecular tools offer to learn about the life of microorganisms in-situ and in-vitro, this criticism of pure microscopic approaches may seem misplaced. If anything, one may think, what we need is to 'see' biological processes at even lower levels of biological organization (e.g., single molecules) to understand the inner workings of complex microbial communities. To be clear, high-resolution measurements can indeed lead to exciting and fundamental discoveries; we share the enthusiasm – and even partake in the effort. However, when it comes to the specific and key challenge of learning how to map structure to community function, the bottom-up strategy is likely to fail because it requires tracking an enormous number of variables and an even larger number of parameters.

To illustrate this point more precisely, let us focus on the case of genomic data. Microscopic approaches that attempt to use genomic data to predict metabolic function rely on gene annotations to build models of the chemical reactions a cell performs. In a few key cases, those models focus on specialized forms of metabolism for which qualitative predictions are clear. For example, the presence of methanogenic archaea in a community, or *amoA* genes in a genome, suggests that methanogenesis and ammonia oxidation take place, respectively, and so on and so forth. In many other cases, however, qualitative predictions are less obvious (e.g., whether an organism would ferment, and if so, what are the fermentation end-products) and quantitative predictions demand well-parameterized metabolic reaction networks, which are bound to be incomplete for most taxa across the tree of life.

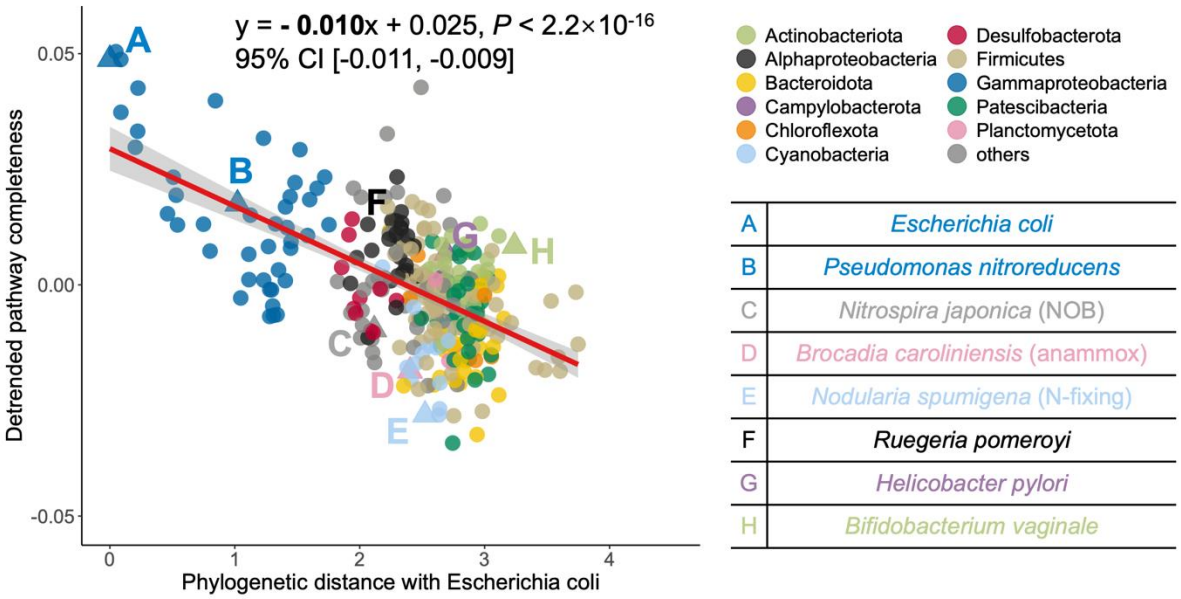


Figure 1.2 Annotation quality deteriorates with increasing phylogenetic distance from *Escherichia coli*. We annotated 3943 genomes from NCBI RefSeq (one genome per genus) using eggNOG-mapper. Pathway completeness decreases as phylogenetic distance with *Escherichia coli* increases, after controlling the covariation introduced by genome size. For each genome, pathway completeness is calculated as the number of genes annotated as part of a KEGG pathway over the total number of genes in that pathway, which is further averaged across all pathways identified in that genome. As pathway completeness covaries with genome size, we calculate "de-trended pathway completeness" as the residual of pathway completeness after a polynomial fitting against genome size. Shaded grey area indicates 95% confidence intervals for the slope of linear fitting are shown in both panels. Linear regression is performed with all 3943 genomes, whereas only 300 genomes randomly chosen are visualized in the figure panel to avoid an overcrowded dot cloud. Eight genomes with clinical or environmental relevance are separately annotated in the figure as examples.

For what is probably the best studied organism, *Escherichia coli*, the fraction of genes with unknown function is around 17%, and likely to contain mostly phage defense genes rather than metabolic enzymes[20]. However, for the typical bacterium, which is not necessarily closely related to *E. coli*, the situation is different, with an average of 45% ± 9% of genes with no known function. The bias towards model organisms is systemic and clear: across ~ 4000 publicly available bacterial genomes from across the tree of

bacterial life[21], the level of pathway completeness increases with phylogenetic distance to *E. coli*, the canonical model organism. (Fig. 1.2). This means that either enzymes have diverged beyond recognition from their common ancestor with an *E. coli* or, perhaps more likely, that the farther an organism is from our laboratory model systems the less we know about its biology. This is a testament to the incredible genomic diversity manifested in natural environments. This issue may be solved, in theory, with more and better efforts to cultivate and biochemically characterize poorly studied organisms. However, in reality, this is likely to remain a slow process compared to the rate of genome sequencing, especially for the many slow growing bacteria that are relevant for ecosystem processes. As a result, methods based on current versions of functional annotations are bound to remain biased towards a very small fraction of cultivable strains, especially regarding functions that expand beyond energy generation or anabolism.

But beyond these technical limitations discussed for genomic studies, there is a broader point to make: bottom-up approaches, whereby one tries to explain complex phenomena from microscopic measurements, are typically inefficient in generating a predictive understanding of the system. While this limitation is well-recognized in other disciplines, like physics or ecosystem ecology, it is a hard sell for biology, given the enormous world of discovery that still lies within the (microscopic) molecular world. However, we argue that to gain a predictive understanding of microbial communities, and more broadly, to develop a theory of microbiomes, it is necessary to zoom out and identified coarser, mesoscopic variables (like groups of functionally similar species) that are more directly linked to community dynamics and functional outputs.

1.3 Towards mesoscopic variables: the promise and challenge of functional groups

In our context, a mesoscopic variable is one that describes a higher level of biological organization, averaging out microscopic variation and providing additional predictive power. In classical ecology, a prime example of a mesoscopic variable is the concept of guild, or functional group, which has long been recognized as a way to construct more compact and interpretable descriptions of communities or ecosystems. Indeed, encapsulating several species into functional groups can reveal "convergent emergent properties" at the ecosystem level[22,23], e.g., reproducible predator-prey ratios and trophic structures[24].

The problem, both for classical ecology and for microbiology, is that we do not have a general methodology for discovering such functional groups in a systematic manner. Instead, we rely on hypotheses that impute function on organisms based on expert knowledge (by definition subjective). This approach may fail and lead to incongruencies between studies. As said earlier, in the case of microbes, the imputation of function based on gene annotation and taxonomy may work well for specialized forms of metabolism like methanogenesis, but it is not a generally reliable strategy. More importantly, we need a way to group taxa that improves our ability to make quantitative predictions about microbiome functions.

How to solve this problem? Encouraged by recent progress, we advocate here for a data-driven, as opposed to an expert driven approach, whereby functional groups can be discovered from controlled, high-throughput community assembly experiments that

pair functional readouts with community composition. We argue that data-driven approaches may provide more principled, intuition-free ways of inferring guilds (Fig. 1.1).

The potential for such approaches is illustrated in a recent study, where we reported an algorithm called EQO, that aims to identify functional groups solely based on patterns of covariation in taxa abundances and environmental measurements, without the need for gene annotations or phylogenetic information[25]. The algorithm exploits ecological patterns in species abundance to identify minimal assemblages of taxa that, when grouped together, maximize the correlation with a chosen functional readout of the whole community. This technique was used to show that the abundance of aerobic and anaerobic ammonia oxidizers, when combined, is the best biological predictor of nitrate concentration in the ocean's water column (as measured in the TARA oceans dataset), and to identify butyrate producing groups of microbes in animal guts as well as groups of cross-feeding microbes in laboratory microcosms. In many contexts, this type of analysis can also serve as a powerful hypothesis generation tool regarding the function of poorly studied taxa.

Despite these encouraging results, the problem of finding functional groups in a data-driven manner is still wide-open. Approaches such as EQO are only a scratch on the surface. For example, EQO identifies a single group per measured function, leaving untouched the main question of how to partition the complete system into functional groups. There is also space for data-driven methods to include some biological guidance, e.g., in the form of statistical priors (e.g., phylogenetically-guided regression[26]). Further, pursuing theory may help us understand how predictive

functional groups emerge, and help suggest new methods to directly learn them from data[27]. It should be also noted that coarse-graining species into guilds does not mean species does not matter. In contrast, it remains to be better understood how is within-guild variation is shaped by species-level or strain level interactions.

However, the main limitation impeding the discovery of better structure-function relations is not computation but rather the lack of appropriate datasets. This may seem *a priori* surprising, because there is an enormous wealth of rapidly expanding genomic datasets easily available. However, for any data-driven guild discovery method to work, genomic or community data needs to be paired with parallel measurements of the shared community function. Unfortunately, most microbiome surveys (16S rRNA or metagenomic) lack any meaningful metadata, and when available, they focus primarily on abiotic parameters that describe the environmental context, more than microbiome function. The fact that most microbial community surveys lack a systematic, quantitative tracking of the metabolic processes mediated by microbial communities is the second major barrier to solving the structure-function problem.

1.4 New datasets to discover functional groups

We argue that to partition microbiome composition into ecological guilds in a data-driven manner, we need datasets that pair microbiome data with measurements of the inputs and outputs of metabolism, such as electron donors and acceptors, nutrients, end-products like CH_4, CO_2, SO_4, fermentation products, etc. To illustrate this idea, in the case of marine and soil microbiomes, where there is a large diversity of taxa involved in the processing of high-molecular weight organic matter, there is a strong need to

identify functional guilds that help us predict carbon turnover and fluxes. However, ocean or soil microbiome datasets are rarely, if ever, paired with a description of the type of organic matter digested by the system, at least in terms of broad classes of compounds like proteins, sugars or lipids, and the metabolic responses of the community (type and rate of respiration, fermentation products, biomass production). Instead, microbiome surveys are paired with chemical and physical descriptions of the environment, like pH, salinity or water depth. Although this context is obviously important, it is knowledge of substrates and metabolic responses what is needed to infer structure-function relations.

It is not for lack of awareness that such input / output measurements are absent from microbiome surveys. Abiotic measurements are low-throughput or require specialized equipment that increases costs and labor, making it difficult to justify their deployment across ambitious sampling projects. Moreover, it is unclear how to measure what a microbiome consumes and produces *in situ*. Instead, it is much easier to collect a sample, extract DNA, sequence and run through the analysis pipelines that output community diversity or taxonomy. However, we argue that is time to develop new strategies to study microbiome function in a systematic, high-throughput manner that would allow us to tackle the structure-function problem.

Below we suggest three elements that we think are key to redefine microbiome surveys:

1. Ecological replicates. Traditionally, microbiome surveys have been structured to compare different but related environments, such as body sites in the human body[28], water depths in the ocean[29], or different soil types[30], to name a few. However, data-

driven guild discovery works best when comparing many samples from the same environment that differ only in a small number of dimensions due to small biological or chemical fluctuations around some mean values. We refer to these as *ecological replicates*. For example, there can be samples from the same body site under similar health conditions and physiological states. With sufficient replication of samples whose composition and function vary around a mean, one should be able to learn how to predict function from structure in a systematic manner.

An important and frequently asked question is, *how many ecological replicates are enough to be able to learn structure from function?* The answer can unfortunately be complicated, depending on species richness, covariance between species, sparsity, etc. However, in our experience with EQO, the order of 100 samples has allowed us to obtain results in a variety of environments, although this should be considered a lower bound[25].

2. We propose a perturbation approach to probe microbiome responses to variation in the chemical environment and substrate supply. The basic strategy consists in creating replicate enrichments with a small variation in chemical composition (e.g., a shift in carbon or nitrogen source). Microbiome responses can be measured for each enrichment by sequencing the community before and after the perturbation, and by measuring metabolite production or consumption. All these measurements need to be compared against a baseline control that accounts for the changes that occur in the unperturbed enrichment.

An example of this strategy, although in relatively low throughput, was recently reported for human microbiome samples[31]. By growing natural microbial inocula from

the human gut in media with various carbohydrates and fermentation intermediates, functional niche of different taxa abundant was identified in the human microbiome. For instance, an organism could be identified as a consumer of a given dietary fiber, like pectin, and producer of a particular fermentation product, like succinate, and so on and so forth along the fermentation cascades of the human gut microbiome.

3. Development of standards. A more widespread consensus and measurement of a certain set of core functions would significantly benefit the scientific community and facilitate the development of data-driven methods. This is because such data could then be aggregated across many microbial studies, greatly improving statistical power.

An example in this regard is probably microbial studies in wastewater treatment plants[32]. Most studies of these systems have chemical measurements of carbon, nitrogen and phosphorous concentrations in the influent and effluent, as well as operational estimates for microbial growth traits such as mixed liquid volatile suspended solids (MLVSS) and sludge retention time (SRT). These measurements are often carried out under standard protocols to evaluate and monitor process performance. When combined together, data from different plants or different time can form an extensive dataset to address important ecological questions.

1.5 Summary

In summary, we advocate the use of mesoscopic approaches to quantitatively study microbial community function. Such approaches have proven useful and predictive in classical ecosystems[33–35], but also in complex systems in several other disciplines such as sociology (e.g., “status groups” by Marx Weber[36]). In the microbial context, however,

they have been almost entirely displaced by the advent of omics technologies. Yet for microbiology, the need for mesoscopic descriptions to tame the microscopic complexity is even greater, especially since the microscopic approaches like consumer-resource models and FBA are data-limited and struggle to scale to the complexity of natural communities.

To find the right mesoscopic variables, we argue that the field should prioritize data-driven approaches over subjective, experience-based ones. The ongoing efforts in this regard, based on statistical approaches, feature selection and machine learning, would be aided by a better theoretical understanding of when and why certain microscopic details can be ignored, but also and most importantly, by a more systematic effort towards collecting data on community function. To this end, we urge the field to build consensus on a set of relevant “model community functions” and suggest experimental design principles for choosing and measuring such functions, including their dynamics in response to perturbations.

The successful deployment of methods that can learn the “right” mesoscopic community variables from data would revolutionize the microbial sciences. It would enable mapping community structure with collective metabolic processes, facilitating rational community design and environmental inference. Additionally, it would have a broader impact on ecology and microbiology, providing objective, data-driven methods for deriving guilds and functional groups without relying on intuition. This could lead to new biological interpretations of data-derived guilds, and hypotheses about possible interactions between them. Ultimately, we may arrive at a unified understanding of

microbial communities as modular interacting systems consisting of functionally cohesive groups of taxa.

Chapter 2 Annotation-free discovery of functional groups in microbial communities

This chapter illustrates the mathematical framework of EQO and demonstrates its application in several existing microbiome datasets from soil, ocean and animal gut. This chapter has been published as the following research article.

Xiaoyu Shan, Akshit Goyal, Rachel Gregor, Otto X. Cordero. (2023). Annotation-free discovery of functional groups in microbial communities. *Nature Ecology & Evolution*, 7,716–724.

Abstract

Recent studies have shown that microbial communities are composed of groups of functionally cohesive taxa, whose abundance is more stable and better associated with metabolic fluxes than that of any individual taxon. However, identifying these functional groups in a manner that is independent from error-prone functional gene annotations remains a major open problem. Here, we tackle this structure-function problem by developing a novel unsupervised approach that coarse-grains taxa into functional groups, solely based on the patterns of statistical variation in species abundances and functional read-outs. We demonstrate the power of this approach on three distinct data sets. On data of replicate microcosm with heterotrophic soil bacteria, our unsupervised algorithm recovered experimentally validated functional groups that divide metabolic labor and remain stable despite large variation in species composition. When leveraged

against the ocean microbiome data, our approach discovered a functional group that combines aerobic and anaerobic ammonia oxidizers, whose summed abundance tracks closely with nitrate concentrations in the water column. Finally, we show that our framework can enable the detection of species groups that are likely responsible for the production or consumption of metabolites abundant in animal gut microbiomes, serving as a hypothesis generating tool for mechanistic studies. Overall, this work advances our understanding of structure-function relationships in complex microbiomes and provides a powerful approach to discover functional groups in an objective and systematic manner.

2.1 Introduction

Microbial communities often involve thousands of different taxa (e.g., 16S rRNA ribotypes, strains, etc.) despite their sharing of many metabolic functions[37,38]. This observation has led to the notion that the major metabolic fluxes in a community of microbes are better captured not by the overall taxa composition, but by the abundance of a much smaller set of so-called 'functional groups' or guilds[18,39,40], akin to those defined for plant and animal communities[41–43]. Members of the same group can fluctuate widely in abundance, replacing each other across space and time due to (often stochastic) ecological interactions, like predation, which can also be highly strain specific[7,20]. In contrast, the abundance of the group as a whole and its combined metabolic output – i.e. the ecosystem service provided by the group, remains stable and is better-associated with the environmental parameters (pH, nutrient concentration, etc.) that control metabolism[39,40]. Thus, as in the case of financial portfolios in which

diversification stabilizes returns[44], functional groups buffer microbial ecosystems from the intrinsic volatility of species dynamics. Moreover, functional groups also provide a more interpretable and simpler description of microbial ecosystems – in terms of key metabolic functions and the groups performing them – than the more readily accessible taxa and gene catalogs.

Despite consensus about the importance of functional groups in microbial ecology, identifying and discovering them is still a major challenge. This is a problem that is not specific to microbial ecosystems. For animals and plants, functional groups are defined based on shared traits, i.e. morphological or physiological differences that impact an organism's role in the ecosystem, but more often than not it is unclear what the relevant traits are and how to measure them in a systematic manner[45]. As a result, there can be a large degree of subjectivity in how functional groups are defined, compromising the value of the approach. In the case of microbes, morphological or physiological traits are much harder to define, and the best way we have to infer phenotypic differences is based on functional gene annotations and taxonomy. However, these functional annotations can carry a high degree of uncertainty[46–48], especially when considered in the context of a community. At the core of the problem is that most microbial taxa have not been cultured, let alone phenotypically characterized. Therefore, functional gene annotations rely on what is known for only a handful of model organisms, like *E. coli* and *B. subtilis*. The less related the organism of interest is to these well-studied model systems, the lower the quality of annotations[49]. To complicate matters further, the presence of a pathway in a genome, even if uncontroversial, does not mean that it is expressed by the organism under the conditions in which the community exists[50,51]. This

implies that the power of annotation-based approaches to identify groups of functionally redundant taxa is limited to special cases in which there is a clear, unambiguous relation between taxa identity and function, e.g., oxygenic photosynthesis and cyanobacteria[39].

To address this challenge, here we present an unsupervised, annotation-free approach that identifies functional groups of taxa based on the patterns of statistical variation in microbiome composition and environmental variables. The approach shares similarities with microbiome-wide association studies (MWAS)[52], which use statistical correlations in microbiome datasets to identify biomarkers of phenomena like disease. Here too, we leverage the correlations between microbiome composition and functional read-outs, but in contrast to MWAS, our main interest is to find groups of functionally redundant taxa, which when combined improve our ability to predict functional read-outs, like the concentration of a relevant metabolic byproduct. To solve this problem, we developed a new algorithm that builds on two assumptions: i) groups of functionally redundant taxa are better correlated with functional read-outs than individual species, and ii) functional redundancy leaves a statistical footprint in microbiome data that can be leveraged to reconstruct the structure-function mapping of the system. Below, we present ample evidence showing that this approach succeeds in identifying groups of functionally redundant species. We start by providing a detailed explanation of the methodology before showing how it can be applied across various data sets, ranging from replicate microcosms[18] to the Tara Oceans microbiome[29] to metabolomic profiles of animal gut microbiomes[53].

2.2 Results

2.2.1 Theory of extracting functional groups from microbiome data

We interpret the challenge of finding functional groups as an optimization problem, i.e. finding the smallest group of taxa that maximize the correlation between their combined abundance and a given environmental variable. To gain intuition into how to solve this problem, consider a group of two taxa and one variable, *y*, representing, for instance, the concentration of a metabolite. From a statistical standpoint, what would make the group a better predictor of *y* than individual taxa (Figure 2.1A-B)? The answer depends on how the two species covary with *y*, as well as with each other. It is straightforward to show that, for the coefficient of determination to increase when the two taxa are counted as one, their individual abundances should not be strongly positively correlated (Figure 2.1C-D). By contrast, if the two species are anticorrelated or uncorrelated in abundance, they may complement each other – the species may be good predictors of *y* in non-overlapping subsets of samples, thus increasing the overall value of the correlation between the group and *y*. At the same time, if one species has a positive covariance with y and the other a negative covariance, their individual effects may cancel out. As shown below, these intuitions can be generalized: the best functional group is one in which members' abundances tend to be correlated with the external variable in a consistent manner (same sign), while at the same time complementing each other

(uncorrelated or anticorrelated). Solving this tension between consistency and complementarity is the optimization challenge.

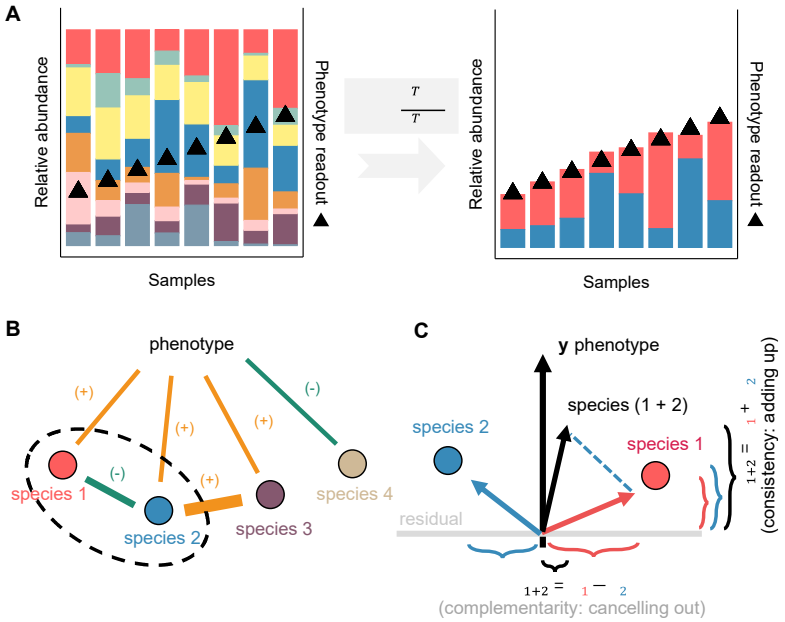


Figure 2.1 Schematic illustration of Ensemble Quotient Optimization (EQO). (A) A typical microbiome is composed of a diversity of taxa, of which individual taxa are only poorly coupled with the measured phenotypic readout. (B) With EQO, species 1 (red) species 2 (blue) are selected to be grouped into an assemblage, whose relative abundance is strongly correlated with the phenotypic readout. Species 1 and species 2 are both positively correlated, though not strongly, with the phenotype ($r_{1,y} = 0.41$, $r_{2,y} = 0.53$), while they are anti-correlated with each other ($r_{1,2} = -0.54$). They are both *consistent* and *complementary*. (C) Consistency implies that two species "add up" in the direction of the phenotype axis while complementarity implies that two species "cancel out" each other with their residuals orthogonal to the phenotype axis. Black vertical axis marks the phenotypic variable. Gray horizontal axis marks the residual of species after projecting onto the phenotypic variable axis.

We derived a simple mathematical expression that captures this tension and that can be used to find functional groups of taxa, as here defined. Consider a community matrix with the abundances of *n* species abundances over *m* samples, and an environmental variable *y*. A group of taxa is defined by a vector of $\boldsymbol{x} \in (0,1)^n$, where the i^{th} position is 1 if the corresponding species belongs to group, or 0 otherwise. The correlation between the group abundance and *y* can be expressed in a compact form,

which we term the Ensemble Quotient, $EQ = \frac{\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x}}$, where P and Q are algebraic transformation of the community matrix that capture the covariance between species (*complementarity*), and the covariance between species and *y* (*consistency*), respectively (Methods and Supplementary Notes). The exact form of these algebraic transformations depends on whether *y* is a continuous variable (e.g., a metabolite concentration), a categorical variable (e.g. healthy or disease states) or a constant (in cases when we want to identify stable groups from biological replicates) (Figure A.1). Independently of the type of variable *y* represents, the EQ is the objective function we want to maximize over $\boldsymbol{x}$.

Two problems that arise when searching for a group that maximizes EQ are the large number of possible solutions and the risk of overfitting. With only 100 species, there are over 75 million groups of size 5, and this number increases exponentially with species richness. As described in Methods, we circumvent this problem using a genetic algorithm that allows us to efficiently search functional groups in communities with hundreds of taxa. The risk of overfitting appears because, the larger the group, the easier it should be to find a species combination that produces a good fit. We solve this problem by putting a penalty on group size (regularization) when appropriate and by assessing the statistical significance of out-of-bag predictions.

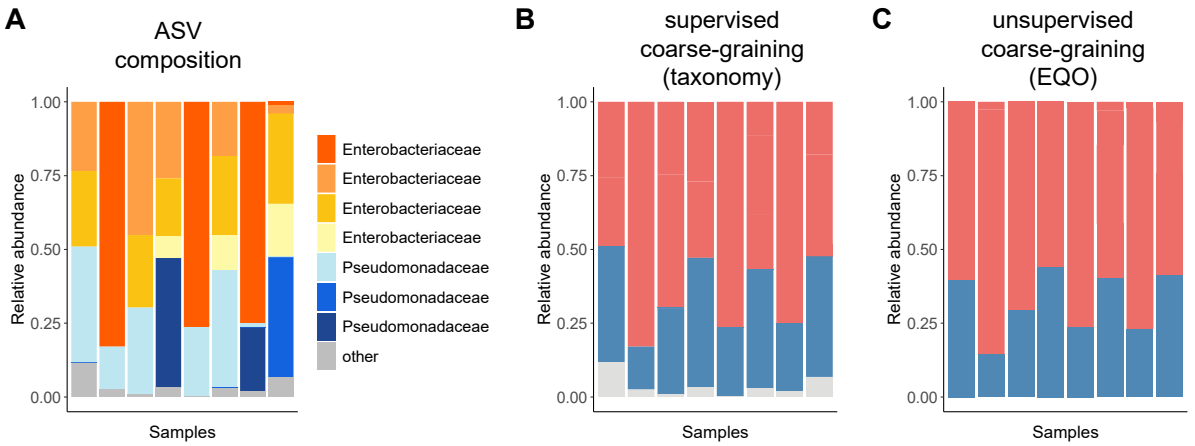
We named the functional group discovery approach EQO, for Ensemble Quotient Optimization. Below, we illustrate its power on three distinct datasets: a set of communities assembled through controlled laboratory microcosms[18] in which amplicon sequence variants (ASVs) fall into two, phylogenetically constrained, functional groups; the Tara oceans microbiome data[29] coupled with environmental parameters such as the

nitrate or oxygen concentration in the water column, and an animal microbiome dataset[53] in which taxa composition data is paired with detailed metabolomic profiling.

2.2.2 Stable functional groups in replicate microcosms

We start by applying EQO to a data set developed from laboratory scale enrichments of soil communities under controlled conditions. The communities were assembled by serial passaging in minimal media with glucose as the limiting resource until they reached a stable composition[18]. Despite identical environmental conditions among replicates, at the fine-grained level of genetic resolution (ASVs), replicate microcosms stabilized into communities with very different compositions. However, the communities were much more similar to each other at the level of major taxonomic families, with two families (*Enterobacterales* and *Pseudomonadales*) dominating the assemblage. It was later confirmed that these two families constituted bona fide functional groups, with one group (*Enterobacterales*) performing partial glycolysis, and the other (*Pseudomonadales*) performing organic acid utilization to complete the full respiration of carbon. Because this is a small scale, controlled experiment, with experimentally validated functional groups, it serves an ideal case to test whether out approach can reproduce results based on taxonomic annotations.

Importantly, this is a case in which we do not have an external signal to guide the functional group search, since all replicate communities were assembled from the same inoculum and under the same environmental conditions. Instead, we want to find the partitioning of the community that is most stable across replicates. To this end, we define y, the environmental variable, as a constant across all replicate microcosms. By finding the group of taxa that best correlates with this constant value, we identify the most stable bi-partition of the community. The EQ formulation for a constant *y* is



described in Methods.

Figure 2.2 Coarse-graining functional groups in replicate microcosms. (A) original ASV-level composition of the replicate microcosms assembled with glucose as carbon source. (B) Supervised coarse-graining into Family-level groups based on taxonomic annotation. (C) Unsupervised coarse-graining by EQO for groups with stable relative abundances across samples, 94.0% ± 2.2% of the ASV reads were co-classified in both approaches, highlighting the power of the unsupervised method.

Remarkably, our unsupervised approach was able to reproduce the bipartition of the community into sugar specialists (represented by *Enterobacterales* ASVs) and acid specialists (represented by *Pseudomonadales* ASVs), but without any prior information about taxonomy, purely from the statistical patterns of ASV variation in the data (Figure 2.2, 94.0% ± 2.2% of the reads correctly mapped; Pearson's r = 0.97 between the

supervised and unsupervised groups). Interestingly, in this specific case we did not need to penalize group size. The optimal group, without constraining for size, is the one that partitions the community into the two, experimentally validated functional groups. Overall, these results show that the approach here developed is capable of identifying functional groups. Our next step is to test its value on real world, environmental data.

2.2.3 Nitrogen cycling in the ocean microbiome

To study EQO's performance in a complex, environmental datasets, we applied it on the Tara oceans dataset[29,39]. Samples in this data set originate from depths ranging from 5.3 to 792 meters and contain a total of 2451 taxa classified at the genus level. Of those, 97 are over 1% abundance in at least one sample. Besides depth, other environmental variables are pH, temperature, nitrate, phosphate, silicate and oxygen, with most of them systematically changing as a function of depth. Of these, we chose to focus on nitrate because of its importance in nutrient cycling, and because it is a direct intermediate of microbial metabolism – a product of nitrification and a substrate for denitrification.

EQO was able to discover a functional group of genera involved in nitrogen cycling in the ocean. Using the Akaike Information Criterion (AIC), calculated as $-2k - \ln(L_k)$, where L_k, is the regression likelihood and *k* the group size, we established an optimal group size of 11 members (Figure A.2). To estimate the relative contribution of each individual member, or member-member pairs to the group's performance, we took advantage of the large number of samples in Tara to apply cross-validation (Figure 2.3A). We perform cross-validation by dividing the data in training and test (50-50) sets

and iterating 1000 times, resulting in an equal number of potentially different groups. We estimated the relative importance of a group member, *i*, by summing the R^2 of all groups in which *i* was present ($\sum x_i R^2$), and similarly for pairs of members ($\sum x_i x_j R^2$).

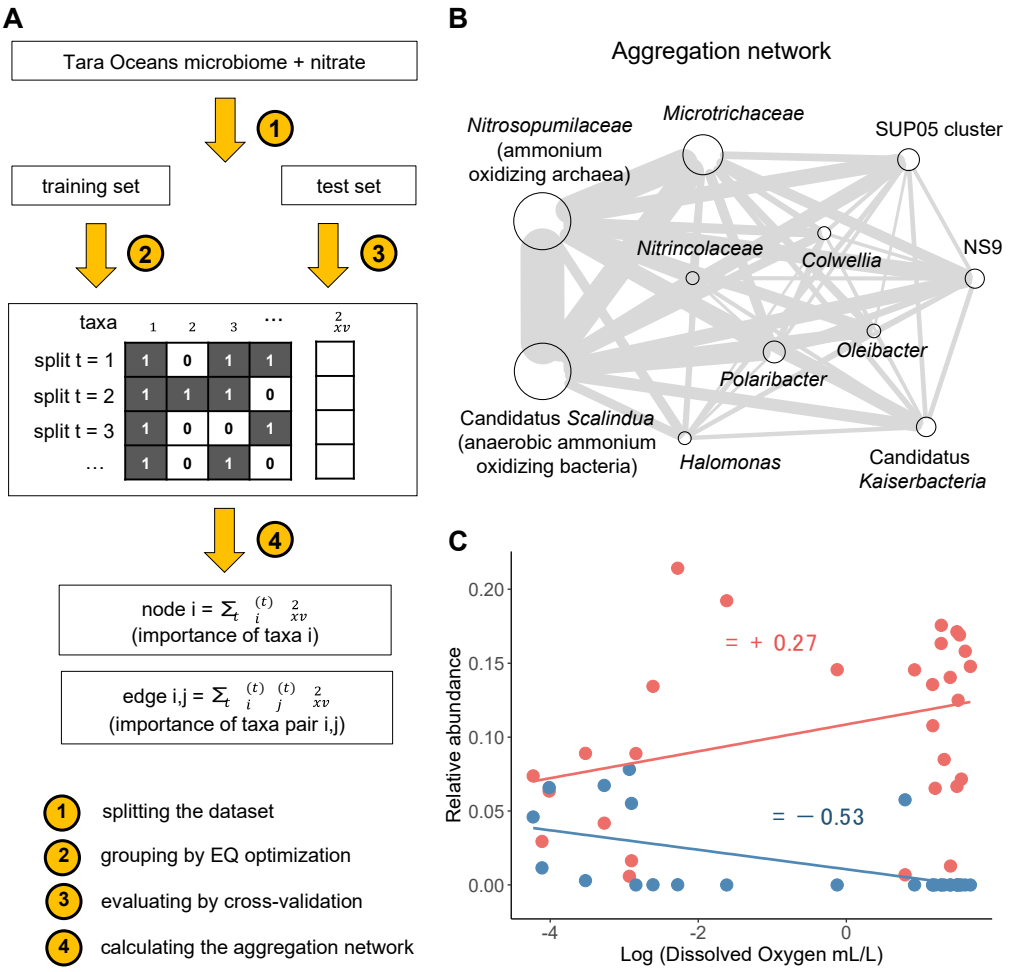


Figure 2.3 Functional guilds of nitrogen cycling in the ocean microbiome. (A) A cross-validation--based algorithm to construct the aggregation network for functional grouping (Methods). The microbiome dataset with the accompanying metadata is randomly split into a training set and a test set. The EQO was applied to the training set to generate a best assemblage, which was then validated with the test set for cross-validated R^2. After iterating 100 times with the cross-validation process, the cumulative cross-validated R^2 for assemblages where a single taxon was present or a pair of taxa was co-present were calculated to indicate the node size and edge width in the aggregation network. (B) Aggregation network for nitrate concentration of the Tara Oceans microbiome. *Nitrosopumilaceae*, the ammonium oxidizing archaea (highlighted in red) and Candidatus *Scalindua*, the anaerobic ammonium oxidizing bacteria (highlighted in blue) tended to be always co-selected by the algorithm, which led to strong cross-validated R^2 with nitrate concentration, suggesting these two taxa together

formed a functional guild. (C) *Nitrosopumilaceae* and Candidatus *Scalindua* showed opposite trend of variation in response to dissolved oxygen in the water column, indicating that they are alternating based on the level of oxygen availability. Shade area indicates 95% confidence interval based on linear regression.

Out of the 11 members, two taxa, *Nitrosopumilaceae* and Candidatus *Scalindua*, had the highest relative contribution to group's ability to predict nitrate concentrations. *Nitrosopumilaceae* is a well-known clade of ammonia oxidizing archaea[54], while Candidatus *Scalindua* is the most abundant anaerobic ammonia oxidizing (annamox) bacteria in marine environment[55]. These two taxa are always co-selected by EQO during cross-validations, as shown in the network of relative pair importance (Figure 2.3B). This is because these two taxa alternate across sampling stations as a function of oxygen concentration (across samples where both taxa are >1% relative abundance, their correlation is Pearson's r = - 0.64), in complete agreement with their predicted roles as aerobic and anaerobic ammonia oxidizers (Figure 2.3C). Because conditions with low oxygen concentration are rare in the Tara samples, the correlation between Candidatus *Scalindua* alone and nitrate is weak (r = 0.31). However, when combined with *Nitrosopumilaceae*, these two taxa complement each other and the correlation was enhanced to r = 0.82 (r = 0.89 for the whole group of 11 taxa), illustrating the power of grouping functionally relevant taxa in association studies. These results show that nitrate concentration in the water column is tightly controlled by the abundance of ammonia oxidizing *archaea* and bacteria, to the extent that nitrate measurements are a good proxy for the overall abundance of this functional group. However, in addition to the ammonia oxidizing *Nitrosopumilaceae* and Candidatus *Scalindua*, there are other nine taxa selected by the algorithm but with lower relative importance as shown in the

aggregation network. Among them, *Nitrincolaceae* is known to oxidize nitrite to nitrate[56] and *Microtrichaceae* found to be abundant in partial nitrification bioreactors, suggesting its close relation with nitrogen cycling[57], Overall, these results illustrates the power of EQO as a hypothesis generating tool.

2.2.4 Mapping metabolites to functional groups in gut microbiome

One of the most compelling potential applications of EQO is the identification of functional groups responsible for the production or consumption of metabolites in gut microbiome data. We leveraged an animal gut microbiome dataset with 101 fecal samples from a wide range of 25 mammalian species, accompanied by 74 peak features that are detected by gas chromatography-mass spectrometry (GC-MS) in >80% samples[53]. Unlike the case of nitrate discussed above, which has been a major focus of research in since the early days of environmental microbiology, the metabolic processes that drive the production or consumption of metabolites in animal guts are much less understood. This makes interpreting the results of the algorithm much harder and demands that we focus only on those predictions that are well-above any significance threshold.

We took a two-fold approach to make sure we focus on statistically significant functional group predictions (Fig. A.3). As before, we use cross-validation to assess out-of-bag predictions. Of all the AIC-based optimal groups, 77% of groups with cross-validated R^2 (xvR^2) below 20% were filtered from further analysis (e.g., most small groups (1-3 taxa) have xvR^2 ~ 0). Groups with xvR^2 > 20% were composed of 7 ± 1 taxa. For these groups, we asked whether their xvR^2 was statistically significant relative

to random groups of the same size. We used an p-value adjustment, multiplying the p-values by a factor $\binom{n}{k}$, where n is the total number of taxa in the microbiome and k is the number of taxa in the group, to account for the fact that the number of hypotheses to test increases rapidly with group size. After accepting groups with adjusted p-values below 0.01, we ended with 12 (16.2%) metabolites with a significant functional group prediction (Figure 2.4A).

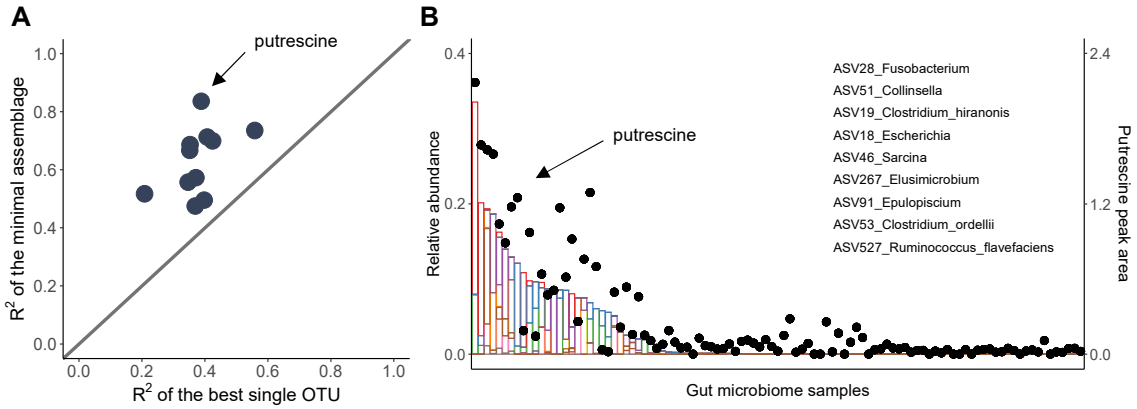


Figure 2.4 Predicting the level of metabolites with minimal assemblages in gut microbiome. (A) Linear regression R^2 of the best single ASV compared to the linear regression R^2 of the best minimal assemblages for 74 metabolites prevalent in > 80% samples. Metabolites that passed significance tests were highlighted as dark black dots (Methods). (B) Putrescine is among the metabolites that passed the stringent significance test, whose level can be strongly predicted by a minimal assemblage of 9 ASVs, which are alternating across different host animals. Host animals are ranked based on descending order of putrescine detected.

The best predicted metabolite was putrescine (1,4-di-amino-butane), a common polyamine molecule derived from amino acids such as arginine[58]. Polyamines such as spermidine and putrescine have been found with remarkable importance in aging[59], cognitive function[60], inflammation suppression[61] and cancer development[62]. Putrescine is mainly enriched in carnivores such as leopards, lions and tigers, as well as the omnivorous bears and coatis, compared to herbivores such as elephants, rhinos and

zebras, consistent with the idea that it is a derivative of protein metabolism (Figure 2.4B). Of the different group members selected by EQO, an ASV of the genus *Fusobacteria* had the highest relative importance (Figure 2.4B). *Fusobacteria* have been experimentally found to be enriched in animal systems with high production of putrescine and is known to be able to synthesize it both in vitro and in vivo[63,64]. However, this ASV alone only explains 29% total variance in putrescine, as compared to 85% total variance explained by the group identified by EQO, suggesting roles of other taxa in the group in putrescine metabolism. Other taxa in the group such as *Clostridium* might also play a role in putrescine production through anaerobic fermentation of proteins. Interestingly, a previous study has found that *Fusobacteria* and *Clostridia* species together dominate the gut microbiome of New World vultures that scavenge dead animals, suggesting their potential functional coupling in protein metabolism[65]. Given the diversity of polyamine biosynthesis pathways and microbes that carry them, the production of putrescine is a good example of a function that is better understood as the result of the collective action of multiple species[66,67]. Our framework allows us to deal with this scenario and generates hypotheses regarding the processes and players involved in polyamine production.

To further validate our methodology in the context of gut-associated microbiomes, we asked whether EQO could identify well-known fermenters based on the abundance of fermentation products in the metabolomics data. To this end, we examined levels of butyrate and lactate quantified across the animal gut microbiome samples. Remarkably, we found that the functional groups EQO was able to identify contained taxa well-known to perform the corresponding fermentation reactions (Figure A.4). For instance, in the

butyrate producing group, EQO found *Ruminococcaceae*, *Faecalibacterium prausnitzii*[68], *Blautia producta*[69] as well as other *Lachnospiraceae* species[68], all of which have been reported to be butyrate producers in the literature. In contrast, in the lactate producing group, EQO identified lactic acid bacteria, and in particular two variants of *Streptococcus luteciae* which when combined are strongly correlated with lactate concentration[70]. Together with the previous cases, the congruence between the functional groups found by EQO and findings in previous experimental studies illustrate the power of this approach.

2.3 Discussion

Over the last couple of decades, DNA sequencing technologies revolutionized the study of microbial communities, by allowing us to construct catalogs of genes and taxonomic markers in a rapid, high-throughput and inexpensive manner. However, rarely are species and gene catalogs themselves the direct subject of a research question, but rather the variables at our disposal to address questions that relate to what functions microbes mediate: anaerobic respiration, digesting organic matter, producing a toxin, etc., are examples of the functions that microbes mediate. Finding a way to map from the variables we can measure (taxa and genes) to the ones we care about (function), or in other words, solving the structure-function mapping, is a primary intellectual challenge in the field of microbial ecology. Current approaches are bottom-up in nature, building a picture of the community based on functional gene annotations or taxonomic descriptions. We argued that this approach is limited. Instead, here we proposed a framework to coarse-grain communities into functional groups in an unsupervised

manner by exploiting the statistical patterns of taxonomic abundance microbial communities and the metadata associated with it. We showed that this approach is surprisingly effective at identifying functional groups without the need for functional or taxonomic annotation in synthetic and natural datasets. The apparent success of EQO suggests that high-throughput surveys of microbial environments in which microbial composition data is paired with metadata (e.g. nutrient levels) can be systematically leveraged to study the structure-function mapping of microbial ecosystems.

EQO builds on the assumption that members of a functional group covary weakly or negatively with one another – a basic tenet of portfolio theory[44]. However, this is not always the case. During early stages of community assembly, e.g. during the colonization of a new environment, environmental filters couple the dynamics of functionally equivalent species[71]. In contrast, as community composition approaches a steady state, functional group members can be decoupled by various factors, including historical contingencies, different sensitivities to environmental parameters (O_2 concentration), and biotic interactions, to name a few[72]. It is in this scenario, where the community composition is not undergoing early successional changes, that approaches such as ours can be applied.

Microbial communities can be composed of many thousands of taxa, especially if defined at high levels of genetic resolution, like strains. If diversity is too high (e.g., 1000 taxa or more) the unsupervised identification of functional groups may become too impractical due the explosion of the search space size. However, as shown in this paper, even in systems with high natural diversity, like oceans or gut microbiomes, we have been able to constrain the search space to less than 100 taxa. This reduction in

complexity relied on two inherent aspects of microbiome data. The first one is that rank abundance curves have very long tails, meaning that although total species counts are large (tens of thousands), the vast majority of taxa appear in very few (e.g., only one) samples. Since those extremely rare taxa do not carry any useful information in the statistical sense, we discard them from the analysis, drastically reducing search space dimensions. Second, taxa can be collapsed at the different levels of phylogenetic resolution (strains, species, genera, families, etc.). For instance, with the Tara Oceans dataset we chose to collapse taxa at the level of genera (e.g., 97% similarity cutoff in 16S rRNA), a common practice in the field, before applying EQO. Although this choice proved effective, it is partly ad-hoc and more systematic approaches to reduce phylogenetic dimensions should be explored.

Naturally, the approach has limitations that are important to discuss. First the statistical association between a functional group and the relevant functional variable might be complicated when responses are decoupled from microbial abundances – e.g. the production of a metabolite does not correlate with an increase in biomass[41]. In such a case, only transcriptomics or proteomics data could reveal the association. Second, the current version of EQO identifies groups by finding associations with single environmental or functional variables, while in some cases a functional group might be strongly coupled with a combination of read-outs. While it is technically possible to deal with these problems, the explosion in the number of possible combinations would demand much larger datasets than currently available. Third, handling large-scale problem efficiently is a major computational challenge for not only EQO but for the field of combinatorial optimization in general. When we attempted to reduce the scale of the

datasets by filtering rare species, we are making a compromise which may not always be acceptable. Finally, and perhaps most importantly, EQO requires relevant functional variables (e.g., metabolic inputs and outputs) to be available across samples. This information is absent from most datasets and requires measurements that are much harder and expensive to generate than sequencing data. Moreover, even the measurements were available, metabolites that are rapidly produced and consumed by microbes may not be measurable unless the system is perturbed. Notwithstanding these limitations, we expect EQO to be a useful hypothesis generating tool to solve structure-function problems in the context of microbial communities.

We have presented the first systematic and unsupervised approach to resolve the structure function mapping of microbial communities. The approach relies on study designs that measure microbiome composition and environmental measurements across multiple sampling stations (ideally hundred or more). Yet, most metagenomic datasets typically lack this structure, either because of the low number of samples or the lack of functional measurements. As we continue to develop the approach, we hope our work clarifies the critical importance of designing microbiome surveys in suitable for the discovery of structure-function relations.

2.4 Methods

Microbiome datasets

Analyses of replicate microbial microcosms[4] and gut microbial communities[53] were based on Amplicon Sequence Variants (ASVs) of 16S rRNA gene amplicon sequencing generated by the original authors. Analysis on the Tara Oceans microbiome was based

on closed-reference taxonomic clustering of metagenome-extracted 16S miTAGs[29,39]. Briefly, 16S rRNA short reads extracted from metagenomes by the original authors were mapped to SILVA 138 reference database[75,76] at 99% similarity with vsearch v2.21[77]. Considering the large number of sequence clusters (~ 40,000) generated from short reads (~100 bp), we used taxonomic composition resolved at the genus level as the input for our algorithm. Reads that cannot be classified at the genus level are binned to the closest taxonomic level possible. Environmental metadata for the Tara Oceans microbiome including nitrate concentration and dissolved oxygen level as well as targeted and untargeted measurements of metabolites by gas chromatography for the animal gut microbiome were requested from the original authors of the previous studies[39,53]. The dimensions (sample × taxa) of the soil, ocean and gut microbiome datasets on which we applied EQO were 8 × 23, 136 × 97 and 101 × 232, respectively.

Formulation of the Ensemble Quotient

Microbiome coarse-graining using uniform, continuous or categorical phenotypic variables can be generalized into an optimization problem as follows

$$max\ \frac{\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x}}$$

where $\boldsymbol{x}$ is a Boolean vector of length n to be solved from the optimization, for which 1 or 0 represent presence or absence of a species in the ensemble. The number n indicates the dimension of the microbiome, e.g., the number of species in an OTU table. For instance, $\boldsymbol{x} = [1,1,0,1,0]$ indicates that there are in total 5 species in the microbiome, in which the 1st, 2nd and 4th species should be coarse-grained into the ensemble. Matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ are given by the algebraic transformation of the taxa and phenotype, with slightly different forms for uniform, continuous or categorical phenotypic

variables detailed as follows. Derivation of the whole mathematical framework is detailed in Supplementary Notes.

(a) uniform phenotypic variable (i.e., composition of the assemblage is stable)

$$\boldsymbol{Q} = \boldsymbol{M}^T \mathbf{1}\mathbf{1}^T \boldsymbol{M}$$

$$\boldsymbol{P} = \boldsymbol{M}^T \boldsymbol{M} - \frac{2}{n} \boldsymbol{M}^T \mathbf{1}\mathbf{1}^T \boldsymbol{M} + \frac{1}{n^2} \boldsymbol{M}^T \mathbf{1}\mathbf{1}^T \mathbf{1}\mathbf{1}^T \boldsymbol{M}$$

$\boldsymbol{M}$ denotes the community composition matrix with m rows of samples and n columns of species, where the element in i-th row and j-th column M_{ij} is the relative abundance of species j in sample i. $\mathbf{1}$ is a unit vector with length m whose elements are all ones. In addition, two simple linear constrains are required to avoid ending up with an empty group or a group with all taxa that is numerically stable but ecologically trivial (Supplementary Notes).

(b) continuous phenotypic variable

$$\boldsymbol{Q} = \boldsymbol{M_0}^T \boldsymbol{y_0} \boldsymbol{y_0}^T \boldsymbol{M_0}$$

$$\boldsymbol{P} = \boldsymbol{M_0}^T \boldsymbol{M_0}$$

$\boldsymbol{M_0}$ is the centered community matrix $\boldsymbol{M}$ whose column means are zero (i.e., relative abundance of each taxon is scaled by subtracting the mean abundance of that taxon across samples). The phenotypic vector $\boldsymbol{y}$ denotes that the level of phenotype at each sample is also centered to $\boldsymbol{y_0}$ but subtracting the mean.

(c) categorical phenotypic variable

$$\boldsymbol{Q} = \boldsymbol{M_0}^T \boldsymbol{Y}\boldsymbol{L}\boldsymbol{L}^T \boldsymbol{Y}^T \boldsymbol{M_0}$$

$$\boldsymbol{P} = \boldsymbol{M_0}^T \boldsymbol{M_0}$$

$\boldsymbol{Y}$ is an augmented categorical matrix with m rows and c columns. The row number m is the same as the number of total samples in the microbiome and the column

number c is the same as the number of categories. $\boldsymbol{L}$ is a diagonal matrix whose diagonal elements are given by the inverse square root of the number of samples in each category.

Optimization of the Ensemble Quotient can be achieved either by reformulating into an equivalent mixed integer linear programming problem (for small-scale problems) or by Markov-chain based heuristics like genetic algorithms (for large-scale problems). Full details of algorithms are in Supplementary Notes.

Cross-validation and aggregation network

Cross-validation consists of the following four steps. (a) Randomly splitting the dataset into a training subset and a test subset (e.g. 50-50 with the Tara oceans data). (b) Finding the best group with EQO on the training subset generated from each time of splitting. We regularized the model by finding the group that minimizes the Akaike Information Criterion (AIC) (Fig. S2A). (c) Evaluating the performance of the best assemblage generated from the training subset by calculating the cross-validation R^2 with the corresponding test subset in each random splitting. (d) Computing the importance each single taxon as cumulative cross-validation R^2 for assemblages where that taxon is present

$$I_i = \sum_t x_i(t)\, R^2_{xv}(t)$$

and importance of taxa pairs as cumulative cross-validation R^2 for assemblages where the two taxa are present

$$I_{i,j} = \sum_t x_i(t) x_j(t)\, R^2_{xv}(t)$$

In the above expressions, t is the split number (1~100). $R^2_{xv}(t)$ is the cross-validated R^2 for the test subset in the t-th splitting. $x_i(t)$ and $x_j(t)$ are 0/1 numbers denoting whether species i and *j*, respectively, are present (1) or absent (0) in the group generated from the training subset in the t-th split. The importance calculated above is normalized to the maximal importance, in order to obtain a value between 0 and 1. A taxon with high relative importance means that if this taxon is included in a group the cross-validated R^2 is likely to be high, and likewise for pairs of taxa. The relative importance of single taxon and relative importance of taxa pairs are indicated by the size of nodes and the width of edges in an aggregation network. The aggregation network in Fig. 3B showed 11 taxa with relative importance larger than 0.5.

Predicting gut metabolites with minimal microbiome assemblages

We ran EQO for each of the 74 metabolites that are detected in > 80% samples, for which the best group size was determined by minimizing AIC value as previously detailed. To assess the significance of metabolite prediction with minimal assemblages, we generated a null model to calculate the expected linear regression R^2, by creating 999 random groups of he same size for each metabolite. A significance value (*P*-value) was evaluated from a Gaussian probability density function with the mean and standard deviation inferred from the linear regression R^2 of random groups. Then we implemented stringent multiple hypotheses testing correction by multiplying the calculated *P*-value by a factor $\binom{n}{k}$ to account for the fact that the number of hypotheses to test increases rapidly with group size, where n is the total number of taxa in the microbiome and k is the number of taxa in the minimal group. For metabolites whose minimal group prediction has an adjusted *P*-value smaller than 0.01,

we applied a second filter by only keeping metabolites whose minimal group predictions have a cross-validated R^2 higher than 0.2.

Other statistical analysis

All the other statistical analyses were performed in R version 4.1.3[78]. Aggregation network visualization was performed with Cytoscape version 3.9.1. Solution of the reformulated mixed integer programming problems were performed by Gurobi optimizer (https://www.gurobi.com) version 8.1.1 (MIT licensed) with an R 3.5.3 interface on a high-performance computing cluster at MIT. Genetic algorithm was implemented by R package GA[79,80] version 3.2.2, with parallel computing enabled by R package doParallel 1.0.17. Fast numerical multiplication of matrices was executed by a customized C++ script, which was integrated into R script by R packages Rcpp 1.0.8 and RcppEigen 0.3.3. Visualization of other plots are performed by R package ggplot2 version 3.3.5.

Chapter 3 Inhibiting aquaculture pathogens: from statistical patterns to biological mechanism

In this chapter we leveraged EQO to identify a group of species inhibitory of a pathogen in marine aquaculture microbiome. This effort led to the isolation of novel bacterial species secreting putative polysaccharides with promising biological effects. We are performing more experiments with our collaborators to further understand this novel species (see future perspective in chapter 5). This chapter is part of the following research paper in preparation.

Xiaoyu Shan, Patrizia Stadler, Rachel Gregor, Gabriel Vercelli, Andreas Sichert, Otto Cordero. Inhibiting aquaculture pathogens: from statistical patterns to biological mechanism. In preparation.

Abstract

Pathogen infections in aquaculture pose significant threats to food security and ecological health. Recent advancements in microbiome analysis offer promising avenues for understanding pathogen control from a microbial ecology perspective. However, moving beyond statistical descriptions to uncovering biological mechanisms remains a major challenge in microbiome studies. In this study, we tackle this challenge by investigating the potential of microbiomes associated with shrimp larvae to inhibit *Vibrio parahaemolyticus* (Vp), a pathogen responsible for massive production losses around the globe. To this end, we analyze the relationship between animal microbiome

composition and survival rates of shrimp larvae across 130 production tanks, revealing a core assemblage of key species whose abundance is inversely correlated with that of Vp. Guided by our computational predictions, we successfully isolated a new species within the marine *Roseobacter* clade and experimentally validated its inhibitory effects on Vp. Intriguingly, we find that the novel species leads to hindered dispersion and reduced virulence of Vp by secreting putative polysaccharides. Overall, these findings advance our understandings of pathogen inhibition in the shrimp larvae microbiome. More broadly, our work also highlights the power of a pipeline involving computational and experimental approaches in addressing real-world environmental challenges with microbiome studies.

3.1 Introduction

Aquaculture farming has experienced remarkable growth over the past a few decades, serving as an important solution to the increasing demand for high-quality food proteins in the modern society[81,82]. However, the expansion of aquaculture systems has also brought about various challenges, in particular the emergence and spread of bacterial pathogens that pose significant threat to food security and ecological health[81,83]. One of the most notable bacterial pathogens in marine aquaculture is *Vibrio parahaemolyticus* (Vp), a microorganism responsible for acute hepatopancreatic necrosis disease (AHPND). AHPND, formerly known as early mortality syndrome, can lead to massive mortalities up to 100% of shrimps and has caused huge economic losses of farmers across the globe[84,85].

Traditional approaches to pathogen control such as antibiotic treatment are facing growing concerns due to the rise of antibiotic resistance and its implications for food security[86,87]. In recent years, the development of high-throughput sequencing has inspired a growing trend to link the health of hosts with the composition of their associated microbiomes[88,89]. Despite the promising vision of advancing pathogen control from a microbial ecology perspective, efforts in this direction still faces major obstacles. Firstly, host-associated microbiomes often comprise of hundreds or thousands of species with highly variable dynamics, few of which is strongly correlated with host survival or pathogen abundance[88]. The statistical power for individual species is questionable, particularly considering that multiple species can fulfill similar ecological roles or perform the same metabolic functions at the same time (i.e., functional redundancy)[37,39,90]. Therefore, it is crucial to employ appropriate algorithms capable of extracting robust functional groups from noisy species distributions[25]. Secondly, most studies on host-associated microbiomes, especially in the context of aquaculture animals, have mainly focused on statistical patterns of microbiome composition[88]. However, more important than calculated coefficients or fitted parameters are fundamental questions such as confirming responsible species and understanding relevant mechanisms. Addressing these questions would require experimental validation of the hypotheses generated from microbiome-based computations.

Here, we address these challenges by studying the shrimp larvae-associated microbiome in a marine aquaculture farm located in Ecuador. We extensively sampled 130 tanks for shrimp larvae microbiome composition profiling via high-throughput amplicon sequencing and records of shrimp larvae survival rate. Moreover, we sampled

4 tanks throughout the developmental stages of shrimp larvae from the fifth day of nauplius (N5) to the eighth day of postlarvae (PL8) to track the introduction of potential pathogens. By combining computational and experimental efforts, we successfully identified and isolated a novel species in the marine *Roseobacter* clade inhibiting Vp. We find that the inhibition is likely to be mediated by secretion of a soluble, high-molecular-weight compound which limits dispersal and reduces virulence of Vp. Altogether, this work provides new insights into the potential of biological control and treatment of aquaculture bacterial pathogens.

3.2 Results

3.2.1 Microbiome composition predicts survival rate of shrimp larvae

To characterize the composition of the larvae-associated microbiome, we sampled 130 tanks of 11 m^3 volume with white-leg shrimp larvae (*Litopenaeus vannamei*) from a hatchery in San Pablo, Ecuador. A total of 3907 amplicon sequence variants (ASVs) were generated from high-throughput amplicon sequencing, where the most abundant taxa include Vibrionales (25.2% of total relative abundance), Rhodobacterales (21.4%) and Flavobacteriales (14.6%, Figure 3.1A). Interestingly, these taxa were also among the most abundant taxa in microbial communities assembled on marine particles made of polysaccharides such as chitin[71,91]. In those microbial communities assembled on nutrient patches in marine ecosystems, Flavobacteriales and Vibrionales species often act as primary degraders of high molecular weight compounds such as polysaccharides or proteins, while Rhodobacterales and Pseudomonadales species often act as scavengers of intermediate metabolites such as amino acids or organic acids[92,93].

We further sought to understand the relationship between shrimp larvae health and shrimp larvae-associated microbiome composition. We find that the microbiome composition is significantly correlated with the survival rate of shrimp larvae (Mantel test $P = 0.001$). The significant correlation raises an intriguing question regarding the potential of using machine learning to predict shrimp larvae survival based on microbiome composition. Indeed, leveraging a gradient boosting tree regressor, we find that ~ 60% of the total variance in shrimp larvae survival rate can be correctly predicted solely with the microbiome composition (Figure 3.1C). This indicates a potential to leverage rapid microbiome sequencing as a tool to forecast shrimp larvae health in aquaculture farms in practice.

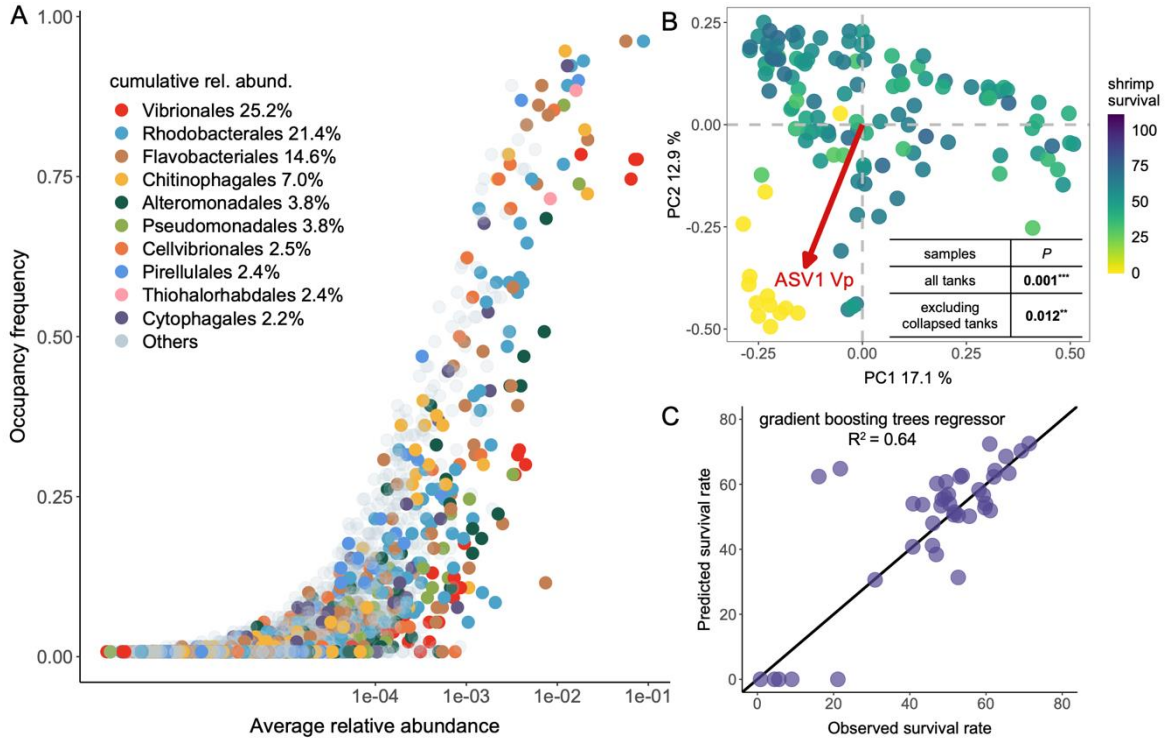


Figure 3.1 Shrimp larvae microbiome composition is coupled with shrimp survival rate. (A) Overview of shrimp larvae microbiome composition sampled from 130 tanks - the horizontal axis indicates the average relative abundance of each ASV across all tanks and the vertical axis indicates the fraction of tanks where each ASV is present. Vibrionales, Rhodobacterales and Flavobacteriales are among the most abundant and common taxa in the shrimp larvae-associated microbiome. (B) Principal coordinate analysis for shrimp larvae microbiome composition, where each tank is also colored

according to its shrimp larvae survival rate. A cluster of tanks colored in yellow are those suffered from an outbreak of AHPND. Microbiome composition of those tanks diverged from the rest of healthy tanks because of high relative abundance of ASV1 Vp. Inset table indicates that microbiome composition and shrimp larvae survival rate are significantly correlated with each other, even when the collapsed tanks were excluded from the test. (C) With machine learning techniques, microbiome composition can predict ~ 60% of the total variance of the shrimp larvae survival rate.

We found among all ASVs that ASV1, classified as Vp, shows the strongest negative correlation with the survival rate (Pearson's r = - 0.70, Figure B.1). The relative abundance of this ASV was significantly higher in the 12 tanks where the whole shrimp larvae population collapsed as a result of a severe outbreak of AHPND (50.3% ± 6.6% in these tanks vs. 2.8% ± 0.5% in the rest of tanks, Mann Whitney U test $P = 2.0 \times 10^{-7}$, Figure 3.1B). We managed to isolate a Vp strain from the larval samples with 100% identity 16S sequence with this ASV. Further whole-genome sequencing confirms the presence of plasmid-borne *pirA* and *pirB* in its genome, which are the virulence factors responsible for AHPND[94]. In addition to this Vp ASV, there is also a Flavobacteriaceae ASV showing strong negative correlation with shrimp larvae survival (Pearson's r = - 0.54, Figure B.1). However, this ASV is only classifiable at the family level, whose ecophysiology and potential pathogenicity remains to be better studied in future studies.

In this work, we focus primarily on Vp as it shows the strongest statistical signal as a threat to shrimp larvae survival.

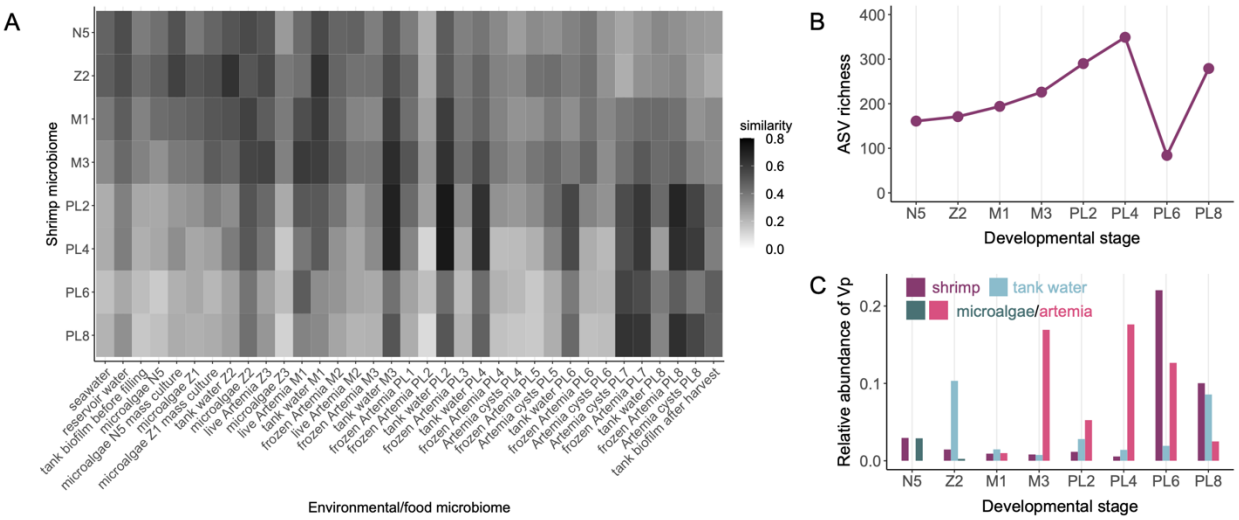


Figure 3.2 Interplay between shrimp larvae microbiome and environmental/food microbiome. (A) Similarity between shrimp larvae microbiome (vertical axis) sampled across developmental stages from Nauplius (N), zoea (Z), mysis (M) to postlarvae (PL), and microbiome from environmental or food samples (horizontal axis) at different timepoints. A strong diagonal indicates an interplay between shrimp larvae microbiome and environmental/food microbiome as the shrimp larvae develops. (B) ASV richness of shrimp larvae microbiome across developmental stages, where a steady increase is interrupted by a sharp reduction at PL6 stage. (C) The sharp reduction in richness coincides with a burst of Vp in the shrimp larvae microbiome. Vp was found to be persistently abundant in the feeding *Artemia*, suggesting that feeding *Artemia* might be a potential source for pathogen infections of the shrimp larvae.

3.2.2 *Artemia* as a potential source for Vp infections

To better understand the potential source of the shrimp larvae microbiome, we focused on one larval tank where we temporally tracked microbiome composition of environmental samples (e.g., seawater, tank water, tank biofilm) and food samples (e.g., microalgae, *Artemia*) at different developmental stages, from nauplii (N), zoea (Z), mysis (M) to postlarvae (PL). In general, we found a close and dynamic interplay between the shrimp larvae microbiome and the microbiome from environmental or food

samples along developmental stages (Figure 3.2A). For instance, the seawater microbiome is most similar to the shrimp larvae microbiome at the early Nauplius stage (46.5% shared ASVs at N5, Figure 3.2A top left cell), while the dissimilarity increases as the shrimp larvae continues to develop into the postlarvae stage (only 21.6% shared ASVs at PL8, Figure 3.2A bottom left column). In contrast, as the shrimps approach later postlarvae stage, their microbiome composition becomes more similar to the tank biofilm sampled after harvest (from 28.6% shared ASVs at N5 to 46.7% shared ASVs at PL8, Figure 3.2A rightmost column). The shift from a seawater-related microbes to a tank-related microbes suggests potential environmental influences on the temporal succession of shrimp larvae microbiome.

We then focus on Vp and sought to identify potential source of Vp infections. Tracking shrimp larvae-associated microbiome along different developmental stages, we found a surge of Vp from 0.5% at PL4 to 22.0% at PL6 in the shrimp microbiome (Figure 3.2C). The surge of Vp corresponds to an abrupt decline of species richness of shrimp larvae microbiome at PL6 (Figure 3.2B). Examining the tank water as well as the microbiome of the microalgae or *Artemia* that is fed to larvae, we found that *Artemia* persistently carries a high abundance of Vp from M3 to PL6 (13.1% ± 2.8% relative abundance, Figure 3.2C). Analysis of the other three tanks we temporally tracked also illustrated the similar pattern (Figure B.2), suggesting that feeding *Artemia* might be a potentially important source for Vp. Interestingly, in some cases the abundance of Vp decreases rapidly after its surge (Figure 3.2C, Figure B.2), begging the question of whether certain variable host factors, such as its internal microbiome, may protect it

from Vp infections. Therefore, we next try to identify species in the shrimp larvae microbiome possibly inhibiting Vp.

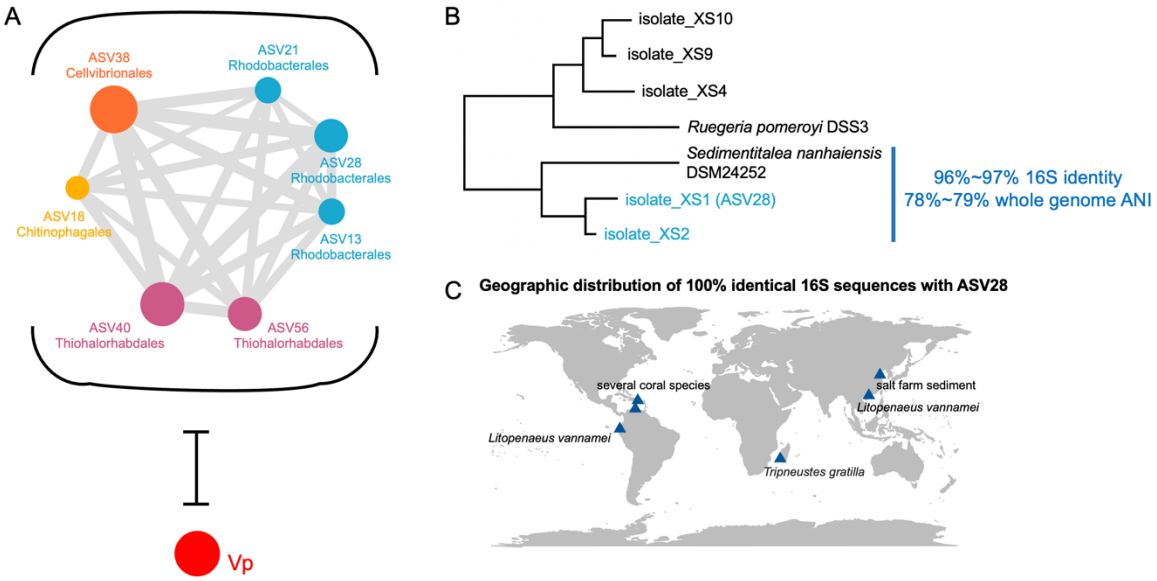


Figure 3.3 Computational prediction of a core inhibitory assemblage for Vp leads to the isolation of novel, cosmopolitan *Sedimentitalea* species. (A) A core assemblage of seven ASVs were predicted by the Boolean least square coarse-graining (methods). The size of nodes in the aggregation network indicates the relative importance of each ASV in inhibiting Vp while the size of edges indicates the relative importance of coarse-graining the connected nodes. (B) Rhodobacterales ASV28 predicted in the core assemblage was isolated in the lab, which is a novel species in the genus of *Sedimentitalea* within the marine Roseobacter clade. (C) Although this species has not been isolated before, it has been detected in amplicon sequencing efforts throughout the globe in previous surveys from various marine host-associated ecosystems.

3.2.3 Identifying a minimal assemblage Vp-inhibiting species

We employ a data-driven approach to identify microbiome features that can potentially protect the host from Vp. Correlating individual ASVs with Vp does not yield any strong statistical signal, with the distribution of Pearson's correlation coefficients densely centered around zero (Figure B.3). The weak correlations suggest that there might be multiple species capable of inhibiting Vp via different mechanisms, such as antibiotic production, contact-dependent killing, resource competition or anti-colonization.

Therefore, we shifted our efforts to look for a minimal assemblage of species that together showing the strongest anti-correlation with Vp. To that end, we leveraged a Boolean least square algorithm to coarse-grain individual species to functional groups, guided by the patterns of statistical variation of species across samples to optimize the performance of the group as a whole[25]. The outcome of the algorithm is visualized as an aggregation network, where the size of nodes indicates the relative importance of the individual species and the size of edges indicate the relative importance of coarse-graining the pair of species (Figure 3.3A). We find that seven species are finally selected by the algorithm, including three Rhodobacterales species. Rhodobacterales is a clade widely distributed in aquaculture ecosystems with a well-understood ecological role as scavengers of small-molecule metabolites[95,96]. Previous studies have found several members in this clade with positive effects on pathogen inhibition and aquatic animal health, such as *Phaeobacter inhibens*[97] and *Tritonibacter mobilis*[98]. However, the three Rhodobacterales species predicted here are neither *Phaeobacter* nor *Tritonibacter*.

We isolated one of the predicted Rhodobacterales species (ASV28) by designing a customized growth media with organic acids as the only carbon source. Phylogenetic placement indicates that the isolates whose 16S sequences are 100% identical to ASV28 are in the genus of *Sedimentitalea*[99]. However, the whole genome average nucleotide identity (ANI) between these isolates and their closest relative *Sedimentitalea nanhaiensis* DSM24252 is only 78%~79%, despite a 96%~97% similarity in their 16S sequences. This indicates that the ASV28 isolate represents a novel species within the genus of *Sedimentitalea* (Figure 3.3B). Although this species

has not been isolated before, we found in publicly available amplicon sequencing datasets that 16S rRNA gene sequences 100% identical to ASV28 were wide-spread throughout the globe from previous environmental surveys, especially in marine aquatic creature-associated microbiome. In addition to the South American *Litopenaeus vannamei* microbiome in our study, 100% identical rRNA gene sequences are also present in *Litopenaeus vannamei* microbiome in China, *Tripneustes gratilla* microbiome in Africa, as well as coral reef microbiome in the Caribbean region (Figure 3.3C).

3.2.4 *Sedimentitalea* sp. inhibits Vp by secreting putative polysaccharides

We leveraged several experimental assays to test the hypothesis that the isolates of the novel *Sedimentitalea* sp. inhibits Vp. With a halo assay, we found that *Sedimentitalea* sp. produced a halo zone on a lawn of Vp in an agar plate. The halo zone can be either resulted from a contact-dependent mechanism or a compound secreted by *Sedimentitalea* sp. To further distinguish between these alternative possibilities, we repeated the halo assay with cell-free supernatant of *Sedimentitalea* sp. We found that the halo zone was still produced by the supernatant (Figure 3.4B), indicating that the halo zone was resulted from a secretant of *Sedimentitalea* sp. Moreover, we found that the halo zone emerged from morphological differences of Vp colonies when they were affected by the *Sedimentitalea* sp. secretant (Figure B.4). The Vp colonies unaffected by *Sedimentitalea* sp. supernatant are opaque, with Vp cells dispersing beyond the boundary of the colonies. In contrast, Vp colonies affected by *Sedimentitalea* sp. supernatant are translucent with smooth boundaries, without beyond-boundary dispersion of cells. To better understand if there is any other phenotypic difference

associated with the two distinct colony morphologies, we extracted mRNA from the affected and unaffected Vp colonies and performed RNA sequencing. Further analysis with transcriptomics showed that the gene encoding quorum sensing signal synthase (*luxM*) and the gene encoding virulence factor (*pirA*) were significantly down-regulated in Vp colonies affected by *Sedimentitalea* sp. supernatant (Figure 3.4D).

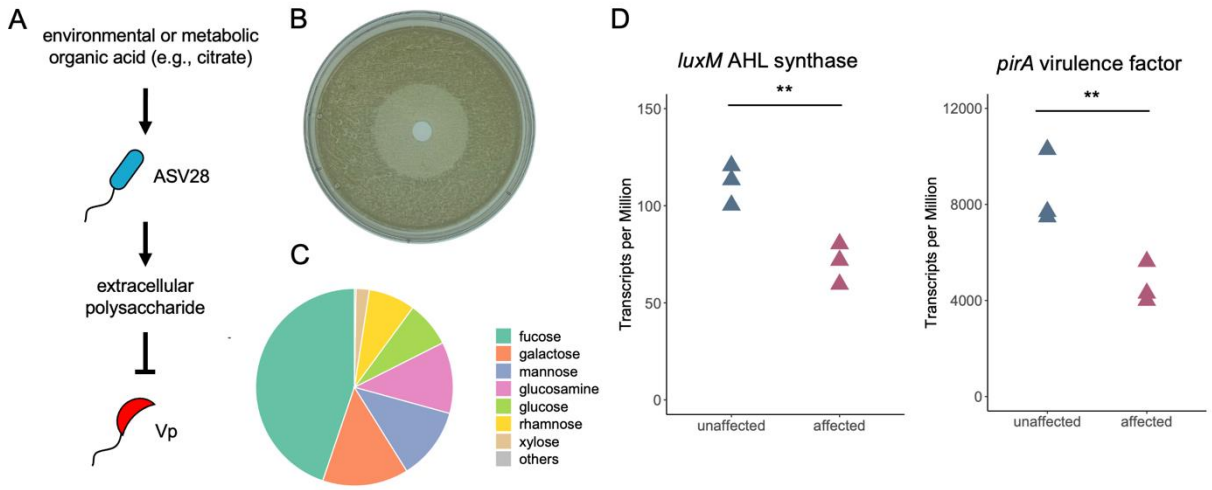


Figure 3.4 The *Sedimentitalea* species caused limited dispersion and reduced virulence of Vp by secreting extracellular polysaccharides. (A) Schematic illustration of the inhibition of *Sedimentitalea* species on Vp. Induced by citrate, *Sedimentitalea* species highly produces the extracellular polysaccharides, which has inhibitory effects on Vp. (B) The excreted extracellular polysaccharides form a halo zone on the lawn of Vp on an agar plate. (C) Monosaccharide composition of secreted extracellular polysaccharides. (D) The secreted extracellular polysaccharides lead to significant downregulation of quorum sensing synthase *luxM* and virulence factor *pirA* in Vp.

Interestingly, the active part of the *Sedimentitalea* sp. supernatant triggering the aforementioned phenotypic changes of Vp is a high-molecular-weight compound. To identify the size range of the compound, we fractionated the cell-free supernatant of *Sedimentitalea* sp. with centrifugal filtering and found that the active compound was highly enriched in the >10K Dalton fraction. A biomolecule with >10K Dalton molecular weight is mostly likely to be a high molecular weight polymer such as nucleic acid,

protein, or polysaccharide. To determine the identity of the active compound, we treated the >10K Dalton fraction with DNase, RNase, proteinase K, or 95℃ incubation, as nucleic acids and proteins should be deactivated after these treatments. Surprisingly, the activity of the post-treatment fraction remained unaffected, suggesting the active compound was neither a protein nor a nucleic acid. To test if polysaccharides were present in the fraction, we performed a phenol-sulfuric acid assay for carbohydrate detection as well as SDS-PAGE staining for carbohydrate-characteristic *cis*-diol groups. Both assays gave positive signals, confirming the presence of polysaccharides in the >10K Dalton fraction. Indeed, we found with acid-hydrolysis-based monosaccharide analysis that the secreted polysaccharides of *Sedimentitalea* sp. were composed of fucose (44.8%), galactose (14.1%), glucosamine (11.7%), mannose (11.8%), rhamnose (7.5%), glucose (7.5%) and xylose (2.2%, Figure 3.4C). Interestingly, we further found that the extracellular polysaccharides were only highly produced and secreted when citrate was present, indicating that citrate acted as an environmental cue to trigger the polysaccharide-mediated interspecies interactions (Figure 3.4A).

We are continuing to better understand the chemical structure and ecological relevance of the secreted compound. Ideas and plans are detailed in chapter 5 as future perspectives.

3.3 Discussion

In this work, we identified and isolated a novel *Sedimentitalea* species from the shrimp-larvae associated microbiome capable of inhibiting aquaculture pathogen Vp. We started from the microbiome data by computationally mapping taxonomic composition to

functional roles, i.e., pathogen exclusion, with an unsupervised, ecology-guided coarse-graining algorithm. Then we moved from statistical patterns to biological mechanism by experimentally confirming and characterizing the inhibitory effects of the predicted *Sedimentitalea* species on Vp. Overall, our work illustrated a powerful workflow in microbiome studies, where structure-function mapping algorithms serve as hypothesis-generating tools to guide experimental follow-ups to gain mechanistic insights. While our study here mainly focuses on analyzing a microbiome dataset from aquaculture shrimp larvae, the workflow itself is highly applicable to various ecosystems. This approach has the potential to revolutionize microbiome research, enabling us to transcend descriptive patterns in microbiome surveys and to advance deeper understandings into ecology and biology.

Our results suggest that the novel *Sedimentitalea* species might inhibit Vp by synthesizing and secreting extracellular polysaccharides. The secreted extracellular polysaccharides trigger phenotypic changes of Vp including limited dispersion, suppressed quorum sensing and reduced expression of virulence factor. We are still working to better elucidate the chemical nature of the active compound, as well as its ecological relevance (chapter 5). Some recent studies have underscored the important role of human mucin glycans in controlling the quorum sensing, toxigenicity or pathogenicity for various human pathogens such as *Candida albicans*[100], *Vibrio cholerae*[101] and *Streptococcus mutans*[102]*.* If our hypothesis of polysaccharides-mediated inhibition is true, then our results would suggest that, in addition to host-derived glycans, glycans derived from bacteria also play a role in modulating phenotypes of other bacteria. Our hypothesis can be partially supported by some

previous works illustrating anti-biofilm effects of some bacterial extracellular polysaccharides[103], especially considering that regulations of biofilm formation, quorum sensing and virulence are often tightly coupled[104,105]. Interestingly, the *Sedimentitalea* species only highly produces and secretes the inhibitory polysaccharides in the presence of citrate, which is likely an intermediate metabolite from the host shrimp larvae or some other bacteria. Therefore, the interaction we suggested here between *Sedimentitalea* species and Vp might be part of a larger interaction network in the complex microbiome. Furthermore, citrate has been reported to affect bacterial physiology as a strong chelator of divalent ions such as iron[106,107], suggesting the importance of chemical context in shaping interspecies interactions.

The rapid advancement in microbiome studies provide us with enormous opportunities to tackle the challenges of pathogen control from an ecological perspective, which is especially invaluable given the growing concern of antibiotic resistance. Here we present in this aquaculture shrimp larvae system how we can gain interesting biological insights from data-mining of the host-associated microbiome composition. As the volume of accessible and available microbiome data continues to increase each year, the crucial mission lies in effectively harnessing and extracting biological underpinnings from it, so as to pave the way for microbiome engineering in the future to address more real-world challenges.

3.4 Methods

Sampling

Microbiome sampling was performed within Biogemar hatchery at San Pablo, Santa Elena province, Ecuador (2°15'S, 80°56'W). A total of 144 tanks distributed in 11 rooms with *Penaeus vannamei* was sampled at the end of the production cycle, when the survival rate of shrimp larvae was also recorded. All tanks of room 12 presented an outbreak disease and samples of larvae of these tanks were collected before population has been discarded. Furthermore, *Penaeus vannamei* larvae and water from four tanks (two from room 9 and two from room 11) were sampled every two days during the entire production cycle from January 27 to February 19, 2021. Possible sources of introduction of bacteria to the larva digestive tract were also sampled, such as seawater, tank biofilm, microalgae culture and *Artemia* as the shrimp food. During the production cycle, shrimp developed by passing through four stages: Nauplius (N), zoea (Z), mysis (M) and postlarvae (PL).

DNA extraction and amplicon sequencing

DNA extraction was performed with Gentra Puregene Tissue Kit (QIAGEN) following the protocol provided by the manufacturer, except that bead beating was used to increase the extraction efficiency for 60 seconds at 5000 rpm. In brief, the extraction procedure includes cell lysis (1:1 cell lysis solution), RNA removal (4μL of RNAse A), protein precipitation (250 μL Protein Precipitation Solution), DNA precipitation (100% isopropanol) and purification (70% ethanol). Purified DNA was shipped to Argonne National Laboratory (Lemont, IL) for amplicon sequencing on a MiSeq targeting the V4 region of 16S rRNA using the 515F and 806R primers. Sequence-specific Peptide nucleic acid (PNA) clamps were used to block the amplification of host-derived mitochondrial 16S sequences at V4 region. Amplicon sequences analyses were

performed on the MIT Engaging computing cluster, where we used QIIME2 v2019.2[108] to demultiplex the raw reads, and DADA2 plugin[109] to generate Amplicon Sequence Variants (ASVs) of ~250 base pairs. Taxonomy of representative ASVs were assigned with the classify-sklearn method by QIIME2 v2019.2.

Boolean least square coarse-graining of microbiome composition

We leveraged a Boolean least square algorithm to coarse-grain individual species into minimal core assemblages[25]. Briefly, the algorithm looks for a linear combination of species with strictly binary coefficients (either zero or one), to maximize the least square error with the targeted variable.

$$\boldsymbol{e^2} = \|\boldsymbol{Az} - \boldsymbol{y}\|_2 = \boldsymbol{z}^T\boldsymbol{A}^T\boldsymbol{Az} - 2\boldsymbol{y}^T\boldsymbol{Az} + \boldsymbol{y}^T\boldsymbol{y}$$

where $\boldsymbol{A}_{m\times(n+1)}$ is the augmented microbiome matrix $[\mathbf{1}_m : \boldsymbol{M}_{m\times n}]$ since we need an additional column of 1 to map the linear intercept b. $\boldsymbol{z}$ is the augmented vector for unknown variable $[b : k\boldsymbol{x}]$, where x_j is the binary variable denoting whether species j should be included in the assemblage. This mixed integer quadratic programming problem is then solved by a commercial optimizer Gurobi v8.0 for the optimal core assemblage of species mostly anticorrelated with *Vp*. Cross-validation with 50-50 random split of the 130 tanks was repeated 100 times to construct the aggregation network. The size of nodes shows the cumulative cross-validation R^2 when the node is selected into the optimal assemblage, while the size of edges shows the cumulative cross-validation R^2 with both of the connecting nodes selected into the optimal assemblages.

Strain isolation and genotyping

Vp was isolated by the CHROMagar *Vibrio* (CaV) selective media, where the mauve colonies are picked for genotyping. A total of 190 Rhodobacterales strains were isolated with a customized media recipe favoring growth of Rhodobacterales species among other marine copiotrophic heterotrophs such as Flavobacteriales, Vibrionales and Alteromonadales. The media is adapted from a MBL minimal media, containing 10 mM NH_4Cl, 10 mM Na_2HPO4, 1 mM Na_2SO_4, 50 mM HEPES buffer (pH 8.2), NaCl (20 g/liter), $MgCl_2*6H_2O$ (3 g/liter), $CaCl_2*2H_2O$ (0.15 g/liter), and KCl (0.5 g/liter). Trace metals and vitamins were added by 1:1000 of the following stock solution. Trace metals stock solution included $FeSO_4*7H_2O$ (2100 mg/liter), H_3BO_3 (30 mg/liter), $MnCl_2*4H_2O$ (100 mg/liter), $CoCl_2*6H_2O$ (190 mg/liter), $NiCl_2*6H_2O$ (24 mg/liter), $CuCl_2*2H_2O$ (2 mg/liter), $ZnSO_4*7H_2O$ (144 mg/liter), $Na_2MoO_4*2H_2O$ (36 mg/liter), $NaVO_3$ (25 mg/liter), $NaWO_4*2H_2O$ (25 mg/liter), and $Na_2SeO_3*5H_2O$ (6 mg/liter). Vitamins, which were dissolved in 10 mM MOPS (pH 7.2), contained riboflavin (100 mg/liter), D-biotin (30 mg/liter), thiamine hydrochloride (100 mg/liter), L-ascorbic acid (100 mg/liter), Ca-D-pantothenate (100 mg/liter), folate (100 mg/liter), nicotinate (100 mg/liter), 4-aminobenzoic acid (100 mg/liter), pyridoxine HCl (100 mg/liter), lipoic acid (100 mg/liter), NAD (100 mg/liter), thiamine pyrophosphate (100 mg/liter), and cyanocobalamin (10 mg/liter). Sodium succinate of 40mM is added as the carbon source. Marine Broth 2216 was spiked into the media with a dilution factor of 1:80 to facilitate bacterial growth. Genotyping of all isolates were performed by full-length 16S Sanger sequencing (GENEWIZ at Azenta Life Sciences). 16S sequences of isolates were aligned to that of ASV1 and ASV28 with blastn in the blast suite v2.12.0. A

maximal-likelihood phylogenetic tree of representative 16S sequences was constructed with MEGA v11.

Whole genome sequencing and genomic data processing

Shotgun whole genome sequencing for Vp and five representative Rhodobacterales isolates were performed at the SeqCenter (Pittsburgh, PA), on an Illumina NextSeq 2000 platform (2x151bp pair-ended). Raw reads were trimmed to remove adaptors and low-quality bases (-m pe -q 20) with Skewer v0.2.2[110]. The remaining paired reads were checked for quality with FastQC v0.11.9. Quality-filtered reads were assembled into contigs with MEGAHIT v1.2.9[111].

Nanopore long-read sequencing was also performed at the SeqCenter (Pittsburgh, PA) to close the genome of Vp. Closed genome was assembled using Unicycler v0.4.9[112] by combining Illumina short reads and Nanopore long reads, resulting in two chromosomes and three circular plasmids. The assembly graph in gfa format was visualized by Bandage v0.8.1[113]. Coding sequences are predicted by prodigal v2.6.3, followed by functional annotation by eggnog-mapper v2[114] (--go_evidence non-electronic --target_orthologs all --seed_ortholog_evalue 0.001 --seed_ortholog_score 60).

Agar plate-based screening of Vp inhibition

We screened for inhibitory effects on pathogens with agar plate-based halo assays. A lawn of the target species (e.g., Vp) was prepared by spreading 50 µL overnight Vp culture on top of Marine Broth 2216 agar with 5~10 glass beads. Rhodobacterales species are allowed to grow in Marine Broth rich media or MBL minimal media with citrate as carbon source for 48 hours, before transferring to the lawn of Vp. Rhodobacterales culture is centrifuged at 5,000g for 5 minutes before being filtered with

0.22 μm filter units for cell-free supernatant. A hole is punched on the center of the agar plate with the lawn of Vp with the opposite end of a 1mL pipette tip, in which 100 μL of cell-free supernatant of Rhodobacterales culture was spiked. The agar plates were left at room temperature for 24 hours before a halo zone becomes visible.

Extraction of *Sedimentitalea* species extracellular polysaccharides

Sedimentitalea species was growing in 100mL of MBL minimal media with 27 mM citrate as carbon source for 48 hours in Innova 42R incubator shaking at 220 RPM 25°C. Cells were removed by first being spined down at 5000 RPM at 4°C for 5 minutes and then being filtered through 0.22 μm filter units. Cell-free flow-through was concentrated by centrifugal filtering units (MilliporeSigma Amicon) with 10K Dalton molecular weight cutoff membrane. A 2.5 volume of 100% ethanol was added into the > 10K Da fraction for ethanol precipitation at 4°C overnight. The precipitated material was spined down at 4500 RPM for 10 minutes and dried in a speed vacuum for 4 hours at room temperature to remove the remaining ethanol. Phenol-sulfuric acid assay and SDS-PAGE-based polysaccharides staining were both performed to confirm the presence of carbohydrates. The biological activity of the purified material was also tested after treatment with DNase, RNase, proteinase K or high temperature (95°C for 20 minutes) to rule out the possibility of being nucleic acids or peptides/proteins.

Determination of extracellular polysaccharide composition

Monosaccharide composition of extracellular polysaccharides were determined using acid hydrolysis and mass spectrometry analysis[115,116]. Dry pellets of extracted extracellular polysaccharides from *Sedimentitalea* species was resuspended into Milli-Q water, where 5 μL of sample were diluted with 45 μL of ddH2O and mixed with 50 uL of

2 M HCl. Samples were hydrolyzed in a thermocycler for 24 hours at 100°C and afterwards neutralized by addition of 4 M NaOH. Samples containing released monosaccharides after acid hydrolysis (25 µL) were derivatized with 0.1M PMP in 2:1 Methanol:ddH2O with 0.4 Ammonium hydroxide (75 µL) for 100 minutes at 70°C following a previously published protocol . For quantification, we derivatized a serial dilution of a standard mix containing Galacturonic acid, D-Glucuronic acid, Mannuronic Acid, Guluronic Acid, Xylose, Arabinose, D-Glucosamine, Fucose, Glucose, Galactose, Mannose, N-Acetyl-D-glucosamine, Ribose, Rhamnose and D-galactosamine. PMP-derivatives were measured on a SCIEX qTRAP5500 and an Agilent 1290 Infinity II LC system equipped with an Agilent Poroshell 120 EC-C18, 2.1x50mm,1.9um reversed phase column with guard column. The mobile phase consisted of Buffer A (25 mM NH4Acetate in ddH2O, 5% acetonitrile, pH=5.6 adjusted with formic acid) and Buffer B (5% ddH2O and 95% acetonitrile). PMP-derivatives were separated in a gradient from 9 % to 23% Buffer B in 2 minutes with a flow of 1 mL/min. The ESI source settings were 625°C, with curtain gas set to 30 (arbitrary units), collision gas to medium, ion spray voltage 5500, temperature to 625, Ion source Gas 1 to 90 and Ion Source Gas 2 to 90. PMP-derivatives were measured by multiple reaction monitoring (MRM) with previously optimized transitions and collision energies. For example, a glucose derivative has a Q1 mass of 511 and was fragmented with a collision energy of 35V to yield the quantifier ion of 175Da and the diagnostic fragment of 217Da. Different PMP-derivatives were identified by their mass and retention in comparison to known standards and their peak areas (175Da fragment) was used for normalized by the amount of internal standard.

RNA isolation and transcriptomics

To isolate RNA from Vp colonies affected or unaffected by the Rhodobacterales supernatant, we sampled colonies within or beyond the halo zone in triplicates. To ensure sufficient amount of RNA for sequencing, we picked 10 colonies per sample. The picked colonies are vortexed in RNA Protect Bacterial Reagent (Qiagen, Hilden, Germany) for mixing. RNA isolation was performed with a Qiagen RNeasy kit following the manufacturer's protocol except that cells were resuspended in 15 mg/mL lysozyme in TE buffer and incubated for 30 min at room temperature before adding buffer RLT. RNA library preparation, rRNA depletion, and pair-ended Illumina sequencing were all performed at the SeqCenter (Pittsburgh, PA). We trimmed paired-end RNA reads by Skewer v0.2.2[110] to remove sequencing adapters and low-quality reads (-m pe -q 20). The remaining paired reads were mapped to Vp closed genome using Bowtie2 v2.2.6[117]. We leverage HTSeq v0.11.3[118] to obtain count table of transcripts and use DeSeq2 R package[119] for differential gene expression analysis. Normalized transcript abundance was generated from count tables by transcripts per kilobase million (TPM) calculations.

Other statistical analysis

Gradient boosting-based prediction of shrimp larvae survival rate was performed by R package xgboost v1.7.5.1, for which 90 tanks are randomly selected as the training set and the remaining 40 tanks are used as the test set. All other statistical analyses are implemented with R v4.1.3. World map is visualized using the default world map in R package ggmap v3.0.0.

Chapter 4 Mutation-induced infections of phage-plasmids

In this chapter I report a serendipitous finding regarding phage-plasmids. When I studied the metabolic functions of a marine strain, I noticed that its liquid culture became clumpy every week. At the beginning I thought it was because I did something wrong in my experiments. Then I repeated the experiments, but always ended up with the same observation. Later, supported by Professor Otto Cordero and Professor Tami Lieberman, I was able to demystify the “doomed fate” of the that bacteria strain. It gradually became clear that behind the scene were mutation-induced infections of phage-plasmids.

This work has been published as the following research article.

Xiaoyu Shan, Rachel Szabo, Otto Cordero. (2023). Mutation-induced infections of phage-plasmids. *Nature Communications*, 14(1), 2049.

Abstract

Phage-plasmids are extra-chromosomal elements that act both as plasmids and as phages, whose eco-evolutionary dynamics remain poorly constrained. Here, we show that segregational drift and loss-of-function mutations play key roles in the infection dynamics of a cosmopolitan phage-plasmid, allowing it to create continuous productive infections in a population of marine *Roseobacter*. Recurrent loss-of-function mutations in the phage repressor that controls prophage induction leads to constitutively lytic

phage-plasmids that spread rapidly throughout the population. The entire phage-plasmid genome is packaged into virions, which were horizontally transferred by re-infecting lysogenized cells, leading to an increase in phage-plasmid copy number and to heterozygosity in a phage repressor locus in re-infected cells. However, the uneven distribution of phage-plasmids after cell division (i.e., segregational drift) leads to the production of offspring carrying only the constitutively lytic phage-plasmid, thus restarting the lysis-reinfection-segregation life-cycle. Mathematical models and experiments show that these dynamics lead to a continuous productive infection of the bacterial population, in which lytic and lysogenic phage-plasmids coexist. Furthermore, analyses of marine bacterial genome sequences indicate that the plasmid backbone here can carry different phages and disseminates trans-continentally. Our study highlights how the interplay between phage infection and plasmid genetics provides a unique eco-evolutionary strategy for phage-plasmids.

4.1 Introduction

A key distinction among temperate phages is whether they integrate into the host chromosome (e.g., the well-known *Escherichia coli*'s phage lambda) or replicate as an extrachromosomal element. In this latter group are phage-plasmids, circular elements that appear to have evolved by the fusion of a plasmid and phage. Although a few examples such as phage P1[120] infecting *Escherichia coli* and phage VP882[121] infecting *Vibrio cholerae* have been extensively-studied, it is only very recently that we have become aware of the prevalence and relevance of these hybrid elements[122]. Recent surveys have found that phage-plasmids are abundant[122,123] and carry a large diversity

of clinically relevant antibiotic resistant genes across bacteria[124]. Despite their significance, however, most of these elements have not been experimentally characterized and their ecology and evolution as not only phages but also plasmids remain poorly understood.

As most other plasmids, phage-plasmids can also be found in multiple copies per cell (polyploidy). This fact has surprising implications for the population genetics of these elements and their dynamics of infection. Polyploidy makes it possible for cells to be heterozygous at any phage-plasmid encoded locus[125–127], including key genes such as the transcriptional repressor that maintains the phage in its lysogenic state. This intra-cell genetic variation can have a significant impact on phage-plasmid dynamics. If the prophage was chromosomally integrated, loss of function mutations in the phage repressor would be effectively suicide mutations, committing the phage to a lytic phage. However, in theory, such defective allele variants could be recessive in a polyploid phage-plasmid. Polyploidy also implies that the intergenerational dynamics of phage-plasmids are affected by segregational drift – i.e., the random assortment of plasmid copies among daughter cells after cell division. Segregational drift can lead to large fluctuations in the degree of heterozygosity (including the production of homozygous offspring) in subsequent generations[128–130]. As shown below, the interplay between heterozygosity and segregational drift in phage-plasmids can lead to a type of eco-evolutionary dynamics unique for phage-plasmids.

We explore these dynamics focusing on a cosmopolitan type of phage-plasmid widespread among marine *Roseobacter* – an abundant copiotroph in the ocean[98]. Using a combination of experiments and mathematical models we show that the hybrid nature

of phage-plasmids allows loss-of-function mutations in the phage repressor gene to be maintained in the population, leading to continuous productive infections. We show that phage-plasmid variants transmit rapidly throughout the population via horizontal transfer, increasing ploidy and producing heterozygous cells. This force is counterbalanced by segregational drift, which restores homozygosity. The combination of these forces leads to the continuous production of phages and the stable coexistence of infected and resistant cells. We continue to show that the phage-plasmids such as this are formed frequently in the environment via fusion of plasmid backbones and phages and widespread across disparate geographic regions, suggesting a successful life-style strategy for these parasitic elements.

4.2 Results

4.2.1 Recurrent productive infection of a phage-plasmid in *T. mobilis* after ~40 generations

Tritonibacter mobilis (previously known as *Ruegeria mobilis*) is a member of the *Roseobacter* clade[98], which collectively represents one of the most ubiquitous groups of marine heterotrophic bacteria[131]. *Tritonibacter mobilis* A3R06, carrying a temperate phage-plasmid, was isolated from an agarose particle inoculated with coastal seawater bacterial communities. Its genome has 4.65 million base pairs (Mbp), with a chromosome of 3.2 Mbp plus four (mega)plasmids of 1.2 Mbp, 0.1 Mbp, 78 thousand base pairs (Kbp) and 42 (Kbp), respectively (Figure C.1). The 42 Kbp plasmid is also a circular phage with 51 predicted genes. These include genes encoding a phage head, tail, capsid and portal proteins, lysozyme, cell wall hydrolases as well as a C1-type

phage repressor. The rest of the genes are involved in plasmid stability, replication and segregation, such as the y*oeB-yefM* toxin-antitoxin system, the *parAB* plasmid segregation system and P4-family plasmid primase (Figure C.2).

When growing *Tritonibacter mobilis* A3R06 under serial dilution cycles of approximately 6 generations per transfer in minimal media, we observed a reproducible, sharp decline in optical density (OD600) after approximately 40 generations (Figure 4.1, Figure C.3). The decline in OD600 was due to the formation of cell clumps containing extracellular DNA (eDNA) (Figure S5), consistent with the idea that cell lysis promoted clump formation[132]. By sequencing the time courses and analyzing differences in genome coverage, we confirmed the induction of the phage-plasmid in all of the 15 independent biological replicates after 30-50 generations (Figure 4.1A, Table C.1 and Figure C.4). To further validate the lysogeny-lysis switch of the phage-plasmid, we did transmission electron microscopic imaging of the 0.22 μm filtered supernatant from the clumpy bacterial culture, confirming the production of virion particles. The phage particle has *Siphoviride*-type morphology, with an isometric head of ~50 nm diameter and a long tail of ~ 180 nm length (Figure 4.1B).

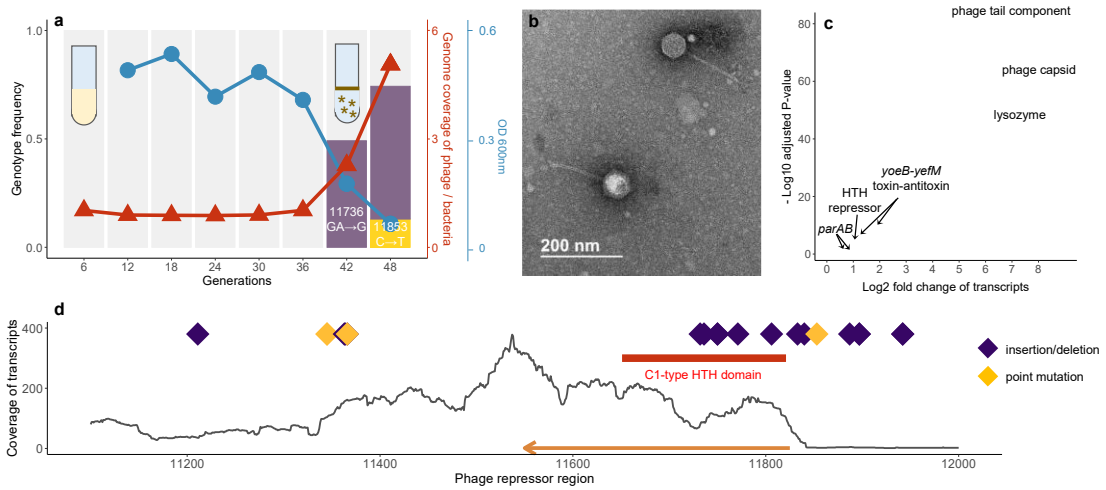


Figure 4.1 Mutations in a phage repressor region recurrently drives productive infection of the phage-plasmid. (a) Productive switch of a phage-plasmid in *Tritonibacter mobilis* A3R06 was observed after ~ 40 generations of serial-dilution growth (red). A deletion mutation (11736: GA→G, purple bar) rapidly increased to ~50% relative genotypic frequency within one dilution cycle, before the increase slowed down in the next dilution cycle. A second mutation (11853: C→T, yellow bar) was observed in the last dilution cycle. Planktonic bacterial culture became highly clumpy after the productive switch, as indicated by the sharp decrease in OD600 (blue). For eco-evolutionary trajectories for the other 8 populations temporally-tracked with genomic sequencing, see Figure S4 and Table S1. **(b)** Transmission electron microscope image of the phage-plasmid particle. Imaging was performed for seven times with biological triplicates, all yielding similar results. **(c)** Differential expression of phage-plasmid genes before and after observing the mutation. Genes related to phage production were significantly upregulated after the productive switch, in particular the phage structural genes and the phage lysozyme gene. Expression of genes that are housekeeping for plasmid replication and stability were only increased because of copy-number increase of genes. P values are calculated based on Wald test and are adjusted by the Benjamin-Hochberg (BH) procedure. **(d)** All 21 mutations identified in 15 independent lines of populations were within a short ~1,000 bp region encoding a C1-type phage repressor (orange arrow). Most of mutations are insertions or deletions (purple diamond).

4.2.2 Productive infection of the phage-plasmid is driven by mutations in a phage repressor region

The observed 30-50 generation lag before the productive infection suggested that the lysogenic-lytic switch was less likely to be driven by metabolite accumulation or physiological signaling. Alternatively, we hypothesized that the prophage induction was

driven by genotypic changes. To test this hypothesis, we did genomic sequencing for 9 independent bacterial populations at the end of each dilution cycle. We found that only a short, ~1000 bp region in the phage genome encoding a C1-type phage repressor, consistently contained mutations across all the independent lines (Figure 4.1D, Table C.1). A large fraction (15/19, 78.9%) of the mutations were insertion/deletion mutations leading to frame-shift within either the helix-turn-helix DNA binding domain of the C1-type repressor or the putative upstream promoter region as inferred from the level of transcripts (Figure 4.1D), suggesting that the mutations resulted in a loss of repressor function. Interestingly, a majority of these insertion/deletion mutations (9/15, 60.0%) were related to tandem repeats sequences in the genome (e.g., nucleotide position 11897: from GAAAAA to GAAAA, Table C.1), which act as mutational hotspots due to replication slippage[133–135], suggesting that these mutations occurred faster than the background rate. Interestingly, several previous studies have found that mutations of tandem repeats can be often reversed[136–138], which might enable an evolutionary switch between lysis and lysogeny.

In order to learn more about the consequences of the repressor mutations, we performed RNA-seq experiments for 3 independent populations, which allowed us to quantify the transcription of phage coding sequences before and after the mutation was observed. We found that the expression of genes related to a lytic phage lifestyle in the phage-plasmid were highly up-regulated after the observation of mutations in the repressor sequence, such as the phage capsid, phage tail and lysozyme (128~256 folds, Figure 4.1C), showing these genes related to phage production were indeed de-repressed after the loss-of-function mutations. In contrast, those genes related to

plasmid replication and stability such as the y*oeB-yefM* toxin-antitoxin system and the *parAB* plasmid segregation system were only increased similarly with the copy-number increase of the phage-plasmid (2~4 folds, Figure 4.1C). These genes were actively expressed even before the mutation was observed (Figure C.6), suggesting that they were functioning for a lysogenic lifestyle.

DNA sequencing across different timepoints during dilution cycles showed that, after repressor mutations appeared, their frequency in the population increased at an extremely rapid rate. e.g., jumping to ~50% within one dilution cycle of 6 generations in a population of ~10^8 cells. Despite this rapid increase, the mutant genotype never reached fixation, stabilizing at around 60% (Figure 4.1A and Figure C.4). This pattern of evolutionary dynamics was intriguing in two respects. First, if transmission was only vertical, the drastic increase in the mutant genotype frequency would imply unrealistically high relative fitness coefficients (s ~ 100). Therefore, the evolutionary dynamics can only be explained by the infection spreading horizontally throughout the population. This presents an apparent conundrum, as we expected a host population lysogenized with the wild-type phage-plasmid to be immune to the same type of phage[139,140]. Second, if the mutated phage genotype was able to spread so rapidly throughout the population, why did not it reach fixation?

4.2.3 Reinfection and segregational drift together explain the observed evolutionary dynamics

We hypothesized that the observed evolutionary dynamics could be explained by the unique population genetic features of a phage-plasmid hybrid. If the mutated phage was

able to spread through the population via virion production and reinfection, then the infected cells would likely contain multiple plasmid copies and be heterozygous at the repressor locus (e.g., one copy of lysogenic wild-type phage plasmid and one copy of mutated phage plasmid). In that case, segregational drift during the stochastic partitioning of plasmids between daughter cells should impact the subsequent bacterial and phage population dynamics. Indeed, recent studies have shown that the evolution of multiple-copy plasmids is affected by segregational drift[125,128,129], akin to the case of mitochondria in eukaryotes[130], resulting in variation in intracellular frequencies of plasmid-encoded alleles between mother cells and daughter cells. In the simplest scenario where plasmids were randomly distributed into daughter cells with equal opportunity, while the copy number of plasmids per cell remained constant, cell division could result in the maintenance of repressor heterozygosity at the single cell level, or the production of two homozygote cells, one carrying only wild-type and one carrying only mutated phage (Figure C.7). In the latter case, the daughter cell with only mutated phages would be lysed, releasing more mutated phage particles and continuing the spread of the phage.

Further experiments confirmed that the lytic phage-plasmid re-infected cells lysogenized with wild-type (Figure 4.2A-B). To show this, we spiked cell-free supernatant containing the mutant phage-plasmid into a culture of the host carrying only the wild-type variant. After an overnight incubation we observed the appearance of clumps, identical to those that appear spontaneously after 30-50 generations (Methods). Genome sequencing of clones streaked out of this culture showed that they carried the full mutant phage-plasmid, whose genotype were identical to the one of present in the

cell-free supernatant used in the re-infection experiment, and that they were heterozygous at the repressor locus (Figure 4.2B). In contrast, when we repeated the same experiment but with the cell-free supernatant 0.02 μm filtered to remove the phage particles, the victim host population remained planktonic, indicating no re-infection.

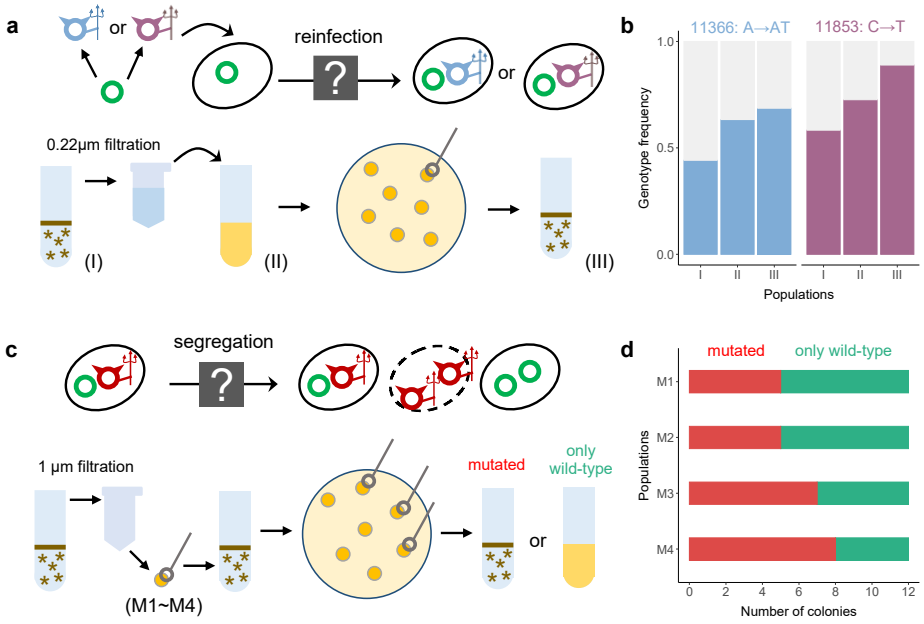


Figure 4.2 Experimental confirmation of reinfection and segregational drift. (**a**) Schematic illustration of reinfection, for which we hypothesize that mutated phages are able to infect hosts lysogenized by wild-type phages. Wild-type and mutated phage-plasmids were showed as green circles and blue color circles. See Methods for full experimental details. (**b**) Mutated phages with the same genotype were observed across (I) the initial source host population carrying the mutated phage, (II) the victim host population re-infected by the mutated phage and (III) the descendant of (II), supporting our hypothesis of reinfection. Experiments were performed in biological duplicates with two source host populations carrying different genotypes of mutated phages (blue and purple). (**c**) Schematic illustration of segregational drift, for which we hypothesize that a host infected by mutated phages is able to generate offspring with only wild-type phages. See Methods for full experimental details. (**d**) Each of the 4 single-cell mother colonies (M1~M4) carrying a mixture of mutated phages and wild-type phages was able to generate descendants only carrying the wild-type phages, supporting our hypothesis of segregational drift.

As a result of re-infection, we observed an increase in phage-plasmid copy number, opening the possibility for segregational drift to impact its evolutionary dynamics (Figure 4.2C-D). Starting with a single colony carrying both the wild-type and mutant phage-plasmids, we questioned whether it was able to generate homozygous descendants, carrying only wild-type phages (Methods). As expected, we found that a heterozygous mother host cell was able to produce offspring that are free of the mutant phage. Genome sequencing of four post-segregation descendant populations confirmed that they only contained the wild-type phage-plasmid and its average copy number was significantly increased (Figure S8, Kruskal-Wallis test $P = 0.02$) as a consequence of segregational drift (Figure 4.2C).

A higher dosage of wild-type repressor gene copies should in principle provide a stronger buffer against phage-plasmid induction, at least in part because the probability of generating zero wild-type repressor after segregational drift would be lower. To test this, we restarted the serial dilution cycles with the post-segregation populations carrying a higher copy number of wild-type phages. Indeed, we found that it took at least 66 generations to observe phage-plasmid induction (Figure C.9), which was significantly longer than the 30~50 generations observed for wild-type populations with single-copy repressor gene.

With reinfection and segregational drift as the only two basic components, we found that a minimal probabilistic model was sufficient to reproduce the observed evolutionary dynamics of the phage-plasmid mutations (Figure 4.3 and Methods for full details of simulation). We started with a population of one million host cells each carrying a single copy of wild-type phage-plasmid and doubling 6 times per serial passage exactly the

same as in the experimental condition (Method). A loss-of-mutation happened in a random phage-plasmid, leading to the lysis of host cell and release of mutated phage-plasmid particles. Some of the released phage-plasmids managed to reinfect another randomly-encountered host based on an efficiency of re-infection (R), which is reminiscent of the production efficiency (R_0) in epidemiology (Methods). The re-infected hosts carrying heterozygous repressor loci underwent segregational drift, after which descendants carrying only mutated phage-plasmids were lysed and re-entered the infection cycle. With these ingredients, our simulation displayed a rapid spread of the mutated phage infection before quickly saturating, which is consistent with experimental observations (Figure 4.3).

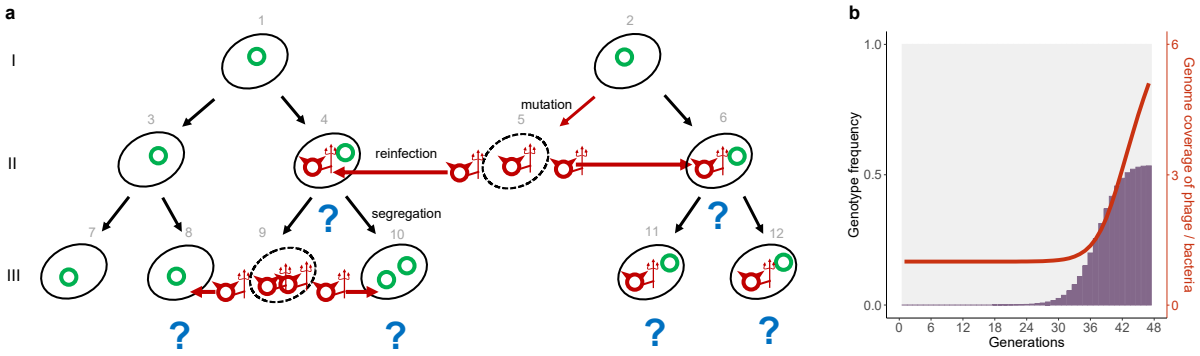


Figure 4.3 A minimal model for phage-plasmid hybrid reproduces the observed eco-evolutionary dynamics. (**a**) Schematic illustration of the model simulation. A dice is drawn under a host cell when it carries more than one copies of phage-plasmids with different genotypes (a heterozygote). In those cases, the phage-plasmid genotype in descendants becomes stochastic due to segregational drift (e.g., cells 8, 10, 11 and 12). (**b**) With reinfection and segregational drift as the only two components, the simulated eco-evolutionary dynamics well matches the observed patterns in the experiment. For the experimentally observed eco-evolutionary dynamics, see Figure 4.1A and Figure C.4. See Methods for full details of the model simulation.

4.2.4 Dissemination of phage-plasmids across geographic and phylogenetic distances

Using comparative genomics, we found that the plasmid backbone of the *Tritonibacter mobilis* A3R06 phage-plasmid carries different phage head and tail components across

different continents (Figure 4.4). To better understand the ecological distribution of phage-plasmids, we leveraged 1,849 genomes available in the RefSeq database from the order of *Rhodobacterales*, to which the *Tritonibacter mobilis* was affiliated. With these genomes, we searched for homologs of the phage-related genes or plasmid-related genes in our *Tritonibacter mobilis* phage-plasmid (isolated in Massachusetts, USA). Strikingly, we found clusters of nearly identical (~ 100%) homologs of plasmid-related genes, such as *parAB* segregation system and P4-family plasmid primase, in another phage-plasmid of another *Tritonibacter mobilis* strain isolated in marine aquaculture in Denmark[98] (Figure 4.4). These gene clusters were also found in perfect synteny, which strongly indicated a recombination event. However, the phage structural genes (e.g., phage head and tail) of these two phages were very different, both in terms of homology and synteny. The structural genes of our *Tritonibacter mobilis* phage-plasmid was both homologous and syntenic to those found in another phage integrated in the genome of a *Roseobacter* strain, which was isolated from 2,500 m deep water in the Arabian Sea[141] (Figure 4.4). Taking together, our results showed that the evolution of the plasmid-related genes and the phage-related genes for phage-plasmids could be decoupled. Different phages could become the genetic cargo of the same plasmid, which was able to transmit across continents carrying their phage components. This was consistent with recent findings showing that the core plasmid backbone could be recombined with different cargo genes in marine or human gut microbiome[142,143], with our findings suggesting that this could be exploited by phages to disseminate across large geographic distances. Additionally, we identified a second example of nearly identical plasmid-related genes but very different phage-related genes between two

other phage-plasmids[144,145] (Figure C.11). These two phages were found in two strains of different though related bacterial species (average nucleotide identity ~92%), suggesting that plasmid backbones were also able to transmit across phylogenetic distances. All those phage-plasmids resemble *Tritonibacter mobilis* A3R06's phage-plasmid: a 40K-50K genome size with the presence of independent replication systems (such as *ParABS* and *RepABC*) but absence of genes known for effective reinfection blockage (such as *SieA* of *E.coli* phage P1[140]), suggesting those phages can be also subject to mutation-driven induction in natural environment.

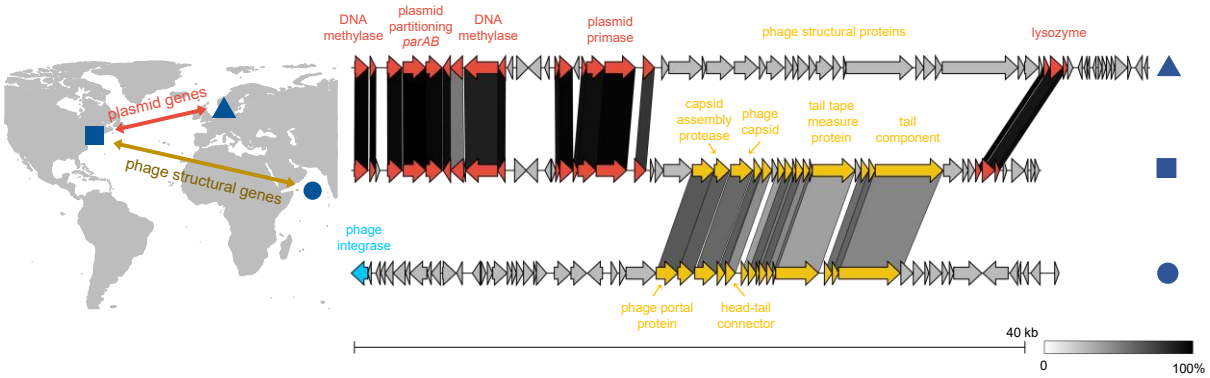


Figure 4.4 The same plasmid backbone carrying different phages genes disseminate vast geographic distance. *Tritonibacter mobilis* M41-2.2, isolated in Denmark, contains a phage-plasmid (triangle) whose plasmid-related genes are homologous and syntenic to those of *Tritonibacter mobilis* A3R06 phage-plasmid (square). However, their phage structural genes are very different from each other. Phage structural genes of *Tritonibacter mobilis* A3R06 phage-plasmid is both homologous and syntenic to that of a chromosome-integrated phage found in *Roseobacter* sp. SK209-2-6 (circle), which was isolated from deep water column in the Arabian Sea. Background map is from R package ggmap.

A bioinformatic search across publicly available metagenomes identified significant hits ($e < 10^{-15}$) of the *Tritonibacter mobilis* A3R06 phage-plasmid repressor in several natural microbial communities across the globe (Figure C.12). Hits were found mainly in nutrient rich marine environment, such as an shrimp aquaculture in the eastern coast of China[146], particle-attached Mediterranean water column[147], as well as an intense algal

bloom in California[148]. Further studies are required to get more complete sequence information of phage-plasmids and to capture and characterize their eco-evolutionary dynamics in natural environments.

4.3 Discussion

In this study, we found that mutation and segregational drift controlled the dynamics of transmission of a cosmopolitan phage-plasmid. First, we showed that a spontaneously mutated phage-plasmid was able to re-infect a host lysogenized with a wild-type phage, which prevented the mutant phage-plasmid from turning lytic. The occurrence of phage re-infection indicates that superinfection exclusion is not always effective, which is consistent with experimental observations for some other phages[149,150] and models suggesting a short-term evolutionary benefit of allowing super-infection[151,152]. Second, we showed that segregational drift diversified the phenotypic outcomes of daughter cells, with cells carrying only mutant phage-plasmids lysing and releasing virions. This highlights the impact that multi-copy elements like plasmids can have on the evolutionary dynamics of bacteria. Taken together, these observations reflect a mixture of both phage and plasmid properties: the phage facet enables rapid horizontal proliferation through virion production while the plasmid facet enables heterozygosity and segregational drift. Consequently, the phage-plasmids proliferated rapidly through iterative reinfection and lysis of a stochastically selected proportion of host descendants, leading to the co-existence of mutant phages and wild-type phage-plasmids. This strategy also reflects how genes and alleles can use viruses to rapidly propagate through a population without driving it to a sudden collapse.

The mutation-driven switch from lysogeny to lysis we observed in this study is distinct from the traditional model of prophage induction, in which lysis is triggered by a regulatory response to stress, chemical signals, etc. One important consequence of this difference is that the rate of mutation-driven induction becomes proportional to the rate of mutation accumulation in a population. Therefore, a rapidly growing host population with large population sizes can develop continuous productive infections, as observed in this study. This feature can be relevant for “opportunitrophs” like members of the *Roseobacter* clade[153]. Members of this clade are known to frequently switch between two types of ecological life-styles, i.e., a survival mode in low-nutrient regions, and rapid growth mode on transient nutrient hotspots such as the phycosphere of marine algae[154]. Thus, we hypothesize that the mutational switch is a phage-plasmid adaptation to rapidly propagate through those fast-growing population, without driving them to a collapse.

We found that phage plasmids such as the one here described evolve by the rapid mixing and matching of plasmid backbones and prophages. This is evident in the fact that the phage region of the phage-plasmid was homologous to a phage found in the Arabian Sea, while its plasmid backbone was homologous to another phage-plasmid found in Denmark. Considering that the element that is the focus of this paper was isolated from the coast of Massachusetts, our findings suggest that phage plasmids might be evolutionary chimeras that combine elements with disparate evolutionary histories and disseminate across vast geographic distances. The wide distribution of these elements in natural environment and their ability to maintain continuous productive infections with rapid transmission of new genetic variants, suggest that these

elements may be major vectors of horizontal transfer. Further work is needed to better-understand the ecological relevance of this hybrid elements and their potential of mutation-driven induction to trigger continuous productive infections in natural and synthetic systems.

4.4 Methods

Media

The minimal marine media, MBL media, was used for serial dilution growth of Tritonibacter mobilis A3R06. It contained 10 mM NH_4Cl, 10 mM Na_2HPO4, 1 mM Na_2SO_4, 50 mM HEPES buffer (pH 8.2), NaCl (20 g/liter), $MgCl_2*6H_2O$ (3 g/liter), $CaCl_2*2H_2O$ (0.15 g/liter), and KCl (0.5 g/liter). Glucose was added as the only carbon source at a concentration of 27 mM. Trace metals and vitamins were added by 1:1000 of the following stock solution. Trace metals stock solution included $FeSO_4*7H_2O$ (2100 mg/liter), H_3BO_3 (30 mg/liter), $MnCl_2*4H_2O$ (100 mg/liter), $CoCl_2*6H_2O$ (190 mg/liter), $NiCl_2*6H_2O$ (24 mg/liter), $CuCl_2*2H_2O$ (2 mg/liter), $ZnSO_4*7H_2O$ (144 mg/liter), $Na_2MoO_4*2H_2O$ (36 mg/liter), $NaVO_3$ (25 mg/liter), $NaWO_4*2H_2O$ (25 mg/liter), and $Na_2SeO_3*5H_2O$ (6 mg/liter). Vitamins, which were dissolved in 10 mM MOPS (pH 7.2), contained riboflavin (100 mg/liter), D-biotin (30 mg/liter), thiamine hydrochloride (100 mg/liter), L-ascorbic acid (100 mg/liter), Ca-D-pantothenate (100 mg/liter), folate (100 mg/liter), nicotinate (100 mg/liter), 4-aminobenzoic acid (100 mg/liter), pyridoxine HCl (100 mg/liter), lipoic acid (100 mg/liter), NAD (100 mg/liter), thiamine pyrophosphate (100 mg/liter), and cyanocobalamin (10 mg/liter). Marine Broth 2216, a rich media

commonly used for growing marine bacterial strains, was purchased from the Fisher Scientific (BD 279110).

Culture growth

Tritonibacter mobilis A3R06 was isolated from an agarose particle in coastal seawater from the Nahant Beach, Massachusetts, USA[71]. Single colonies of *Tritonibacter mobilis* A3R06 on Marine Broth agar plates were picked for enrichment in 2 mL liquid Marine Broth 2216 media for 6 hours. After that, 50 µL of enriched culture was transferred into 4 mL MBL minimal media for pre-culture growth. Cells in the pre-culture was grown to mid-exponential phase before being diluted into 4 mL fresh MBL minimal media to an OD600 of roughly 0.01 to initiate the serial dilution cycles. Each cycle lasted for 24 hours, corresponding to roughly 6 generations per cycle considering the doubling time of *Tritonibacter mobilis* A3R06 being ~ 4 hours in MBL minimal media with glucose (Figure S3). At the end of each cycle, cells were still within the exponential phase of growth (Figure S3), except for those very late cycles where cell clump formed following the productive switch of the phage-plasmid. All liquid culture growth was performed in Innova 42R incubator shaking at 220 rpm at 25℃.

DNA extraction, Illumina sequencing and reads processing

Prior to DNA extraction, the cell culture samples were centrifuged at 8000 *g* for 60 seconds to remove the liquid. The cells were then resuspended into fresh MBL media by thoroughly pipetting for at least fifteen times. For each sample, the resuspension-centrifuge procedure was repeated for three times so as to wash away free virion particles outside of the cells. For Illumina sequencing, DNA was extracted with the Agencourt DNAdvance Genomic DNA Isolation Kit (Beckman Coulter). DNA

concentration was quantified with Quant-iT PicoGreen dsDNA Assay kit (Invitrogen) on a Tecan plate reader. Short-read sequencing was performed on an Illumina NextSeq 2000 platform (2x151bp pair-ended). Library preparations and sequencing were performed at the Microbial Genome Sequencing Center (Pittsburgh, PA). Sequencing reads were trimmed to remove adaptors and low-quality bases (-m pe -q 20) with Skewer v0.2.2[110]. The remaining paired reads were checked for quality with FastQC v0.11.9.

Closing the genome of *Tritonibacter mobilis* A3R06

Nanopore long-read sequencing was used to close the genome of Tritonibacter mobilis A3R06. DNA was extracted with a Qiagen DNeasy kit for higher DNA yield following the manufacturer's protocol. Long-read sequencing was performed on the Oxford Nanopore platform with a PCR-free ligation library preparation at the Microbial Genome Sequencing Center (Pittsburgh, PA). Closed genome of Tritonibacter mobilis A3R06 was assembled using Unicycler v0.4.9[112] by combining Illumina short reads and Nanopore long reads, resulting in one chromosome plus four circular plasmids. The assembly graph in gfa format was visualized by Bandage v0.8.1[113]. Coding sequences were predicted using prodigal v2.6.3[155] and functionally annotated with eggnog-mapper v2[114] (--go_evidence non-electronic --target_orthologs all --seed_ortholog_evalue 0.001 --seed_ortholog_score 60). The phage genome map was visualized by SnapGene v6.0 (Insightful Science; available at snapgene.com).

Read mapping and variant calling

The complete genome of Tritonibacter mobilis A3R06 was used as the reference genome. Quality-filtered pair-ended Illumina sequencing reads were mapped the

reference genome using Minimap2 v2.17 with stringent settings (-ax sr)[156]. Genetic variants were identified from the aligned reads using BCFtools v1.13 with only variants with quality score ≥ 20 and a local read depth ≥ 20 were remained[157].

RNA isolation and sequencing

RNA Protect Bacterial Reagent (Qiagen, Hilden, Germany) was added to the cell culture samples at a 2:1 volume ratio. RNA was isolated with a Qiagen RNeasy kit following the manufacturer's protocol except for the following changes[158]: cells were resuspended in 15 mg/mL lysozyme in TE buffer and incubated for 30 min at room temperature before adding buffer RLT. Mechanical disruption of samples using lysing matrix B (MPBio, Santa Ana, CA) were performed by shaking in a homogenizer (MPBio) for 10X 30 seconds intervals. Dry ice was added in the homogenizer to prevent overheating. Illumina Stranded RNA library preparation with RiboZero Plus rRNA depletion and pair-ended Illumina sequencing (2x51bp) were performed at the Microbial Genome Sequencing Center (Pittsburgh, PA).

Transcriptomic analysis

Paired-end RNA reads were trimmed using Skewer v0.2.2 to remove sequencing adapters and low-quality reads (-m pe -q 20)[110]. The remaining paired reads were checked for quality with FastQC v0.11.9 and mapped to *Tritonibacter mobilis* A3R06 genome using Bowtie2 v2.2.6[117]. The generated SAM files were sorted by position using SAMTools v1.3.1[157]. Count table of transcripts were obtained by HTSeq v0.11.3[118] and differential gene expression was evaluated with DeSeq2 R package[119]. Normalized transcript abundance was generated from count tables by transcripts per kilobase million (TPM) calculations.

Fluorescence staining and light microscope

Live cells were stained by 5 μM SYTO9 which emits green fluorescence when it is bound to DNA. Dead cells were stained by 20 μM propidium iodide which emits red fluorescence when it is bound to DNA but was unable to permeate the cell membrane. Fluorescence was visualized using an ImageXpress high content microscope equipped with Metamorph Software (Molecular devices, San Jose, CA), operating in widefield mode. Images were acquired in widefield mode at 40x with a Ph2 ELWD objective (0.6 NA, Nikon) and filter sets: Ex 482/35 nm, Em: 536/40 nm, dichroic 506 nm to detect SYTO9 and Ex 562/40 nm, Em 624/40 nm, dichroic 593 nm to detect propidium iodide. Images were collected with exposure times of 100 ms and processed with ImageJ v1.53[159].

Transmission electron microscopic imaging

Transmission electron microscopic imaging was performed at Koch Institute's Robert A. Swanson (1969) Biotechnology Center Nanotechnology Materials Core (Cambridge, MA). Samples were negatively stained with 2% uranyl acetate and were imaged on an JEOL 2100 FEG microscope. The microscope was operated at 200 kV and with a magnification in the ranges of 10,000~60,000 for assessing particle size and distribution. All images were recorded on a Gatan 2kx2k UltraScan CCD camera.

Comparative genomics of Rhodobacterales phages

A total of 1,849 genomes in the Family of *Rhodobacterales* were downloaded from NCBI RefSeq database on Jan 1st 2022. Coding sequences were annotated by eggnog-mapper v2 (--go_evidence non-electronic --target_orthologs all --seed_ortholog_evalue 0.001 --seed_ortholog_score 60)[114]. Phage sequences were

predicted with VIBRANT v.1.2.0[160]. MMseqs2[161] was used to search for homologs of *Tritonibacter mobilis* A3R06 phage genes, with high sensitivity parameters (-s 7.5 -c 0.8). The search output of MMseqs2 were sorted for the most significant hits as well as the highest number of hits, leading to the finding of *Tritonibacter mobilis* M41-2.2 phage and *Roseobacter* sp. SK209-2-6 phage. Alignment map was visualized with clinker clinker v0.0.23[162]. Map is visualized using the default world map in R package ggmap v3.0.0[163].

Model simulation of eco-evolutionary dynamics

In order to simulate the evolutionary dynamics of the phage-plasmid, we developed a minimal model combining the population genetics of a plasmid as well as infection dynamics of a phage.

Our model in part resembles a classical Wright-Fisher model, which assumes non-overlapping generations in a discrete Markov process. However, we considered dynamic population size in our model, which incorporates cell doubling within a serial dilution cycle as well as cell lysis due to phage production. To start with, we have a population of N_0 host cells. When there is no phage-plasmid productive infection, all cells divide into two daughter cells thus the population size doubled every generation following $N_t = N_0 2^t$, which reaches $2^6 \times N$ at the end of each serial dilution cycle. Then a bottleneck indicated by the dilution factor d was applied to the population so that $1/d$ cells were randomly sampled from the current population to enter the next serial dilution cycle. In our simulation, we use $N = 10^6$ at the beginning of each dilution cycle and dilution factor $d = 64$ as we did in experiment. Each host cell carries one

copy of wild-type lysogenic phage-plasmid, replicating and segregating into two daughter cells as the host cell divides.

Loss of function mutations in the phage repressor gene result in this lysogenic phage becoming constitutively lytic. The host cell carrying the mutated phage-plasmid is then killed, releasing virion particles with the mutated phage genome to randomly infect other host cells. Previous studies have reported burst size of marine prokaryotic phages ranging from 4 to more than 100[164]. In our model, what matters in population genetics is the average number of released mutated phage-plasmids that successfully re-infect a host cell per host cell lysed. This parameter, termed as re-infection efficiency R, is similar to the parameter R_0 in epidemiology and should be lower than the empirical burst sizes[165], especially considering that the other host cells have been already lysogenized with a wild-type phage-plasmid larger than 40KB. We found that the saturating frequency of the mutated genotype was affected by R, for which $R = 5$ best fitted the experimentally observation (Figure C.10).

The host cells re-infected by the mutated phage-plasmids become heterozygote with more than one copies of phage-plasmids. Segregation of multiple copies of phage-plasmids with different genotypes can lead to genetic heterogeneity among daughter cells. In our minimal model, we assume a simplest scenario where phage-plasmids were randomly distributed into daughter cells with equal opportunity, while the copy number of plasmids per cell remained constant. For a cell host with a copies of wild-type phage-plasmids and b copies of mutated phage-plasmids, the segregation can be described using a Binomial distribution $B(2a + 2b, a + b)$. For instance, the probability of having a_1 copies of wild-type phage-plasmids in the first daughter cell follows

$$P(a_1) = C_{2a}^{a_1} C_{2b}^{a+b-a_1} / C_{2a+2b}^{a+b} \quad (1)$$

To simulate segregational drift at cell division, we performed Binomial sampling for all heterozygotic cell hosts containing more than one copies of phage-plasmids at each generation, generating daughter cells with stochastically different genotypes. Cells carrying at least one copy of wild-type phage-plasmid are prevented from lysis since the repressor gene is normally functioning. Host cells carrying only mutated phage-plasmids after segregational drift will be killed since the lytic genes on the phage-plasmids are no longer repressed. These lysed cells will be used to produce more virion particles to re-infect more host cells in the next cycle.

We implemented the simulation in R 4.1.0. The script simulates the life cycle of the phage-plasmid as detailed above, with the following parameters: initial population size ($N = 10^6$), re-infection efficiency ($R = 5$) and dilution factor ($D = 64$).

Experimental confirmation of reinfection

To verify re-infection, mutated phage-plasmids were used to infect the host population carrying the wild-type phage-plasmid. Mutated phage-plasmids were separated from the source host population cells by filtering through a 0.22 μm pore size membrane. The supernatant was then spiked into a victim host population carrying only wild-type phage-plasmids growing in fresh MBL minimal media. The planktonic culture became highly clumpy after overnight growth, indicative of phage-plasmid induction. The infected population was then used to streak an agar plate for descendant single colonies. Colonies considered to harbor the mutated phage through re-infection, as indicated by the clump formation after re-growing in liquid media, were sequenced for phage-plasmid

genotyping. The experiment was performed in biological duplicates with two source host populations carrying different mutated genotypes (11366: A→AT and 11853: C→T).

Experimental confirmation of segregational drift

To verify segregational drift, host populations carrying both wild-type phages and mutated phages were tested for whether they were able to generate offspring with only wild-type phages. To ensure the host population really came from a heterozygote single cell rather than a clump of cells with mixed genotypes, we filtered the host population carrying mutated phages with 1 μm cell strainer (Pluriselect 437000103) to remove the multicellular clumpy aggregates. The filtered planktonic subpopulation was carefully examined under the microscope to ensure it contained planktonic cells clearly separated from each other. We then streaked this planktonic subpopulation on an agar plate for single colonies, of which 4 single colonies carrying both wild-type and mutated phages were picked as mother colonies. For each mother colony, liquid culture after overnight growth was then used to streak agar plates for daughter colonies, of which 12 daughter colonies were picked per mother colony. All the 48 daughter colonies were screened in liquid culture for whether they were planktonic, which is indicative of carrying only wild-type phage-plasmids, or clumpy, which is indicative of induction of mutated phage-plasmids. Further, we performed whole-genome sequencing of four daughter colonies that are planktonic in liquid culture, confirming that 1) they were indeed free of any mutated phage-plasmids and only contained wild-type phage-plasmids and 2) the copy number of phage-plasmid in their genomes are significantly increased.

Phage susceptibility of other Rhodobacterales strains

The *yoeB-yefM* toxin-antitoxin system encoded by the phage-plasmid makes it difficult to cure the plasmid for the *Tritonibacter mobilis* A3R06 host. We therefore tried to test whether this phage-plasmid is able to infect any other bacterial host with a plaque assay, including *Tritonibacter mobilis* F1926 which is the model strain for the *Tritonibacter* genus[98] and other 28 *Rhodobacterales* isolates in the Cordero lab strain collection. However, none of those isolates were subject to infection. Recent studies suggested that specificity of phage infection may be related to the structure of bacterial capsule[150,166,167]. This may be the case for our phage-plasmid since *Tritonibacter mobilis* A3R06 harbors another 78Kb plasmid encoding a capsule, which was not found in the genome of other strains we tested for susceptibility.

Chapter 5 Conclusion and Perspective

How to map taxonomic composition to metabolic function is one of most fundamental questions in microbial ecology. In this thesis, I developed Ensemble Quotient Optimization (EQO) to infer functional groups based on ecological patterns and demonstrated the biological insights we gained by applying EQO. In chapter 2, I proposed a generalized mathematical framework for EQO and illustrated its power with a few existing microbiome datasets. In chapter 3, I isolated a novel marine bacterial species guided by the computational predictions of EQO, whose secreted polysaccharides seemed to be inhibitory of *Vibrio* pathogens. In chapter 4, I deciphered unique eco-evolutionary dynamics of a phage-plasmid driven by mutation-induced infections.

However, EQO is just a baby step towards structure-function mapping in microbial communities. Both computational and experimental efforts are needed to tackle the challenges of depicting a complete structure-function landscape. On the experimental side, as argued earlier in chapter 1, experimental datasets with sufficient ecological replicates, perturbed environmental conditions and comprehensive functional measurements will be particularly needed. On the computational side, further efforts should be made to enable identification of multiple functional groups within the same community, or functional groups based on partial *a priori* knowledge. Furthermore, computational speed of EQO as a combinatorial optimization problem remains to be improved. For instance, *semidefinite relaxation* for integer programming might play a promising role in this direction, especially for microbiome datasets that are too large for EQO to solve efficiently. Finally, EQO infers functional groups based on a statistical

interpretation of functional redundancy. However, there can be cases where the metabolic function a species is highly dependent on specific interactions with other species, or even phages. In these cases, additional computational tools need to be developed to map species to function.

Experimental results of the novel *Sedimentitalea* species raise an interesting hypothesis that its interaction with *Vibrio* pathogens might be mediated by extracellular polysaccharides. To me, this is reminiscent of personal communications with Professor Jan-Hendrik Hehemann during the PriME annual conference in New York in 2022. Professor Hehemann has an interesting theory that polysaccharides are antimicrobial in nature, which impressed me a lot (personal conversation, 2022). However, in our story of chapter 3, ecological relevance and chemical structure of the putative secreted polysaccharides still remain to be better understood. Regarding its ecological relevance, we have a hypothesis that the translucent Vp colonies affected by the putative polysaccharides might be more susceptible to macrophage ingestion, probably because of changes in cell wall composition. To test this hypothesis, we are collaborating with Professor Salvador Almagro-Moreno to perform macrophage assay. Furthermore, changes in cell wall might also lead to differences in susceptibility to phage infections. To test this hypothesis, we are isolating lytic phages of Vp for plague assay-based studies. Finally, solving the structure of the putative polysaccharide requires separation and purification for a pure compound. To that end, we are also working closely with our colleagues in chemistry.

Mutation-induced infections of phage-plasmid provides novel insights into productive switch of prophages. Most previous studies have focused on physiological

state or ecological signals in productive switch, while here evolutionary forces such as mutations and drift play a critical role. However, what still remains elusive is how common such mutation-driven productive infections are in natural environment across different ecosystems. Addressing this question would require deeply sequenced environmental metagenome or virome, where one can be provided with sufficient coverage to examine genetic variation in phage repressor regions. With this type of analysis, one can investigate when mutation-driven productive infections tend to be favored (e.g., high density of hosts, rich nutrient, strong environmental fluctuations, etc.). A systematic investigation into this question will greatly advance our understandings of ecology and evolution of phage-host interactions.

Bibliography

1. Gopalakrishnappa, C., Gowda, K., Prabhakara, K. H. & Kuehn, S. An ensemble approach to the structure-function problem in microbial communities. *ISCIENCE* **25**, 103761 (2022).

2. Sanchez, A. *et al.* The community-function landscape of microbial consortia. *Cell Syst.* **14**, 122–134 (2023).
3. Bond-Lamberty, B. & Thomson, A. Temperature-associated increases in the global soil respiration record. *Nat. 2010 4647288* **464**, 579–582 (2010).
4. Jian, J., Steele, M. K., Thomas, R. Q., Day, S. D. & Hodges, S. C. Constraining estimates of global soil respiration by quantifying sources of variability. *Glob. Chang. Biol.* **24**, 4143–4159 (2018).
5. Shi, Z., Crowell, S., Luo, Y. & Moore, B. Model structures amplify uncertainty in predicted soil carbon responses to climate change. *Nat. Commun. 2018 91* **9**, 1–11 (2018).
6. Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLOS Biol.* **17**, e3000102 (2019).
7. Goyal, A., Bittleston, L. S., Leventhal, G. E., Lu, L. & Cordero, O. X. Interactions between strains govern the eco-evolutionary dynamics of microbial communities. doi:10.7554/eLife.74987
8. Dar, D., Dar, N., Cai, L. & Newman, D. K. Spatial transcriptomics of planktonic and sessile bacterial populations at single-cell resolution. *Science (80-.).* **373**, (2021).
9. Fan, K. *et al.* Soil biodiversity supports the delivery of multiple ecosystem functions in urban greenspaces. *Nat. Ecol. Evol. 2023 71* **7**, 113–126 (2023).
10. Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? *Nat. Biotechnol. 2010 283* **28**, 245–248 (2010).
11. Khandelwal, R. A., Olivier, B. G., Röling, W. F. M., Teusink, B. & Bruggeman, F. J. Community Flux Balance Analysis for Microbial Consortia at Balanced Growth. *PLoS One* **8**, e64567 (2013).
12. Henson, M. A. & Hanly, T. J. Dynamic flux balance analysis for synthetic microbial communities. *IET Syst. Biol.* **8**, 214–229 (2014).
13. Goyal, A., Wang, T., Dubinkina, V. & Maslov, S. Ecology-guided prediction of cross-feeding interactions in the human gut microbiome. *Nat. Commun. 2021 121* **12**, 1–10 (2021).
14. Wang, T., Goyal, A., Dubinkina, V. & Maslov, S. Evidence for a multi-level trophic organization of the human gut microbiome. *PLOS Comput. Biol.* **15**, e1007524 (2019).
15. Ho, P.-Y. *et al.* Resource competition predicts assembly of in vitro gut bacterial communities. *bioRxiv* 2022.05.30.494065 (2022). doi:10.1101/2022.05.30.494065
16. Segrè, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 15112–15117 (2002).
17. Goyal, A. & Maslov, S. Diversity, Stability, and Reproducibility in Stochastically Assembled Microbial Ecosystems. *Phys. Rev. Lett.* **120**, 158102 (2018).
18. Goldford, J. E. *et al.* Emergent simplicity in microbial community assembly. *Science (80-.).* **361**, 469–474 (2018).
19. Marsland, R. *et al.* Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLOS Comput. Biol.* **15**, e1006793 (2019).

20. Hussain, F. A. *et al.* Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science (80-.).* **374**, 488–492 (2021).
21. Pollak, S. *et al.* Public good exploitation in natural bacterioplankton communities. *Sci. Adv.* **7**, 4717–4745 (2021).
22. Simberloff, D. & Dayan, T. THE GUILD CONCEPT AND THE STRUCTURE OF ECOLOGICAL COMMUNITIES. (1991).
23. Salt, G. W. A Comment on the Use of the Term Emergent Properties. *https://doi.org/10.1086/283370* **113**, 145–148 (1979).
24. Heatwole, H. & Levins, R. Trophic Structure Stability and Faunal Change during Recolonization. *Ecology* **53**, 531–534 (1972).
25. Shan, X., Goyal, A., Gregor, R. & Cordero, O. X. Annotation-free discovery of functional groups in microbial communities. *Nat. Ecol. Evol. 2023 75* **7**, 716–724 (2023).
26. Huang, C. *et al.* Phylogeny-guided microbiome OTU-specific association test (POST). *Microbiome* **10**, 1–15 (2022).
27. Moran, J. & Tikhonov, M. Defining Coarse-Grainability in a Model of Structured Microbial Ecosystems. *Phys. Rev. X* **12**, 021038 (2022).
28. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* (2012). doi:10.1038/nature11234
29. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science (80-.).* (2015). doi:10.1126/science.1261359
30. Bahram, M. *et al.* Structure and function of the global topsoil microbiome. *Nature* (2018). doi:10.1038/s41586-018-0386-6
31. Anthamatten, L. *et al.* Mapping gut bacteria into functional niches reveals the ecological structure of human gut microbiomes. *bioRxiv* 2023.07.04.547750 (2023). doi:10.1101/2023.07.04.547750
32. Wu, L. *et al.* Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat. Microbiol. 2019 47* **4**, 1183–1195 (2019).
33. Franklin, O. *et al.* Organizing principles for vegetation dynamics. *Nat. Plants 2020 65* **6**, 444–453 (2020).
34. Kunstler, G. *et al.* Plant functional traits have globally consistent effects on competition. *Nat. 2015 5297585* **529**, 204–207 (2015).
35. Shipley, B., Vile, D. & Garnier, É. From plant traits to plant communities: A statistical mechanistic approach to biodiversity. *Science (80-.).* **314**, 812–814 (2006).
36. Weber, M., Fischoff, E., Roth, G. D. & Wittich, C. Economy and society. An outline of interpretive sociology. Edited by Guenther Roth and Claus Wittich. (Translators: Ephraim Fischoff [and others].). 3 vol.: pp. cviii, xxviii, xxviii, 1469, lxiv. 24 (1968).
37. Louca, S. *et al.* Function and functional redundancy in microbial systems. *Nature Ecology and Evolution* (2018). doi:10.1038/s41559-018-0519-1
38. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* (2016). doi:10.1038/ncomms13219
39. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science (80-.).* (2016). doi:10.1126/science.aaf4507

40. Nelson, M. B., Martiny, A. C. & Martiny, J. B. H. Global biogeography of microbial nitrogen-cycling traits in soil. *Proc. Natl. Acad. Sci. U. S. A.* (2016). doi:10.1073/pnas.1601070113
41. Díaz, S. & Cabido, M. Vive la différence: plant functional diversity matters to ecosystem processes. *Trends Ecol. Evol.* **16**, 646–655 (2001).
42. Benoit, D. M., Jackson, D. A. & Chu, C. Partitioning fish communities into guilds for ecological analyses: an overview of current approaches and future directions. *https://doi.org/10.1139/cjfas-2020-0455* **78**, 984–993 (2021).
43. Blaum, N., Mosner, E., Schwager, M. & Jeltsch, F. How functional is functional? Ecological groupings in terrestrial animal ecology: Towards an animal functional type approach. *Biodivers. Conserv.* **20**, 2333–2345 (2011).
44. Schindler, D. E., Armstrong, J. B. & Reed, T. E. The portfolio concept in ecology and evolution. *Front. Ecol. Environ.* **13**, 257–263 (2015).
45. Levin, S. A. Encyclopedia of Biodiversity. (2nd ed.). (2013).
46. Shi, L. D. *et al.* Methane-dependent selenate reduction by a bacterial consortium. *ISME J. 2021 1512* **15**, 3683–3692 (2021).
47. Lobb, B., Tremblay, B. J. M., Moreno-Hagelsieb, G. & Doxey, A. C. An assessment of genome annotation coverage across the bacterial tree of life. *Microb. Genomics* **6**, (2020).
48. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLOS Comput. Biol.* **5**, e1000605 (2009).
49. Griesemer, M., Kimbrel, J. A., Zhou, C. E., Navid, A. & D'Haeseleer, P. Combining multiple functional annotation tools increases coverage of metabolic annotation. *BMC Genomics* **19**, 1–11 (2018).
50. Lee, D. J., Minchin, S. D. & Busby, S. J. W. Activating Transcription in Bacteria. *Annu. Rev. Microbiol* **66**, 125–152 (2012).
51. Aidelberg, G. *et al.* Hierarchy of non-glucose sugars in Escherichia coli. *BMC Syst. Biol.* **8**, 1–12 (2014).
52. Gilbert, J. A. *et al.* Microbiome-wide association studies link dynamic microbial consortia to disease. *Nat. 2016 5357610* **535**, 94–103 (2016).
53. Gregor, R. *et al.* Mammalian gut metabolomes mirror microbiome composition and host phylogeny. *ISME J. 2021 165* **16**, 1262–1274 (2021).
54. Qin, W. *et al.* Nitrosopumilus maritimus gen. nov., sp. nov., Nitrosopumilus cobalaminigenes sp. nov., Nitrosopumilus oxyclinae sp. nov., and Nitrosopumilus ureiphilus sp. nov., four marine ammoniaoxidizing archaea of the phylum thaumarchaeo. *Int. J. Syst. Evol. Microbiol.* **67**, 5067–5079 (2017).
55. Schmid, M. C. *et al.* Anaerobic ammonium-oxidizing bacteria in marine environments: widespread occurrence but low diversity. *Environ. Microbiol.* **9**, 1476–1484 (2007).
56. Mori, J. F. *et al.* Putative mixotrophic nitrifying-denitrifying gammaproteobacteria implicated in nitrogen cycling within the ammonia/oxygen transition zone of an oil sands pit lake. *Front. Microbiol.* **10**, 2435 (2019).
57. Wang, B. *et al.* Recovering partial nitritation in a PN/A system during mainstream wastewater treatment by reviving AOB activity after thoroughly inhibiting AOB and NOB with free nitrous acid. *Environ. Int.* **139**, 105684 (2020).

58. Kibe, R. *et al.* Upregulation of colonic luminal polyamines produced by intestinal microbiota delays senescence in mice. *Sci. Reports 2014 41* **4**, 1–11 (2014).
59. Matsumoto, M., Kurihara, S., Kibe, R., Ashida, H. & Benno, Y. Longevity in mice is promoted by probiotic-induced suppression of colonic senescence dependent on upregulation of gut bacterial polyamine production. *PLoS One* **6**, (2011).
60. Gupta, V. K. *et al.* Restoring polyamines protects from age-induced memory impairment in an autophagy-dependent manner. *Nat. Neurosci. 2013 1610* **16**, 1453–1460 (2013).
61. Zhang, M. *et al.* Spermine Inhibits Proinflammatory Cytokine Synthesis in Human Mononuclear Cells: A Counterregulatory Mechanism that Restrains the Immune Response. *J. Exp. Med.* **185**, 1759–1768 (1997).
62. Gerner, E. W. & Meyskens, F. L. Polyamines and cancer: old molecules, new understanding. *Nat. Rev. Cancer 2004 410* **4**, 781–792 (2004).
63. Noack, J., Kleessen, B., Proll, J., Dongowski, G. & Blaut, M. Dietary Guar Gum and Pectin Stimulate Intestinal Microbial Polyamine Synthesis in Rats. *J. Nutr.* **128**, 1385–1391 (1998).
64. Noack, J., Dongowski, G., Hartmann, L. & Blaut, M. The Human Gut Bacteria Bacteroides thetaiotaomicron and Fusobacterium varium Produce Putrescine and Spermidine in Cecum of Pectin-Fed Gnotobiotic Rats. *J. Nutr.* **130**, 1225–1231 (2000).
65. Roggenbuck, M. *et al.* The microbiome of New World vultures. *Nat. Commun. 2014 51* **5**, 1–8 (2014).
66. Nakamura, A., Ooga, T. & Matsumoto, M. Intestinal luminal putrescine is produced by collective biosynthetic pathways of the commensal microbiome. *Gut Microbes* **10**, 159–171 (2019).
67. Kitada, Y. *et al.* Bioactive polyamine production by a novel hybrid system comprising multiple indigenous gut bacterial strategies. *Sci. Adv.* **4**, 62–89 (2018).
68. Kettle, H., Louis, P., Holtrop, G., Duncan, S. H. & Flint, H. J. Modelling the emergent dynamics and major metabolites of the human colonic microbiota. *Environ. Microbiol.* **17**, 1615–1630 (2015).
69. Benítez-Páez, A. *et al.* Depletion of Blautia Species in the Microbiota of Obese Children Relates to Intestinal Inflammation and Metabolic Phenotype Worsening . *mSystems* **5**, (2020).
70. McKenzie, V. J. *et al.* The Effects of Captivity on the Mammalian Gut Microbiome. *Integr. Comp. Biol.* **57**, 690–704 (2017).
71. Datta, M. S., Sliwerska, E., Gore, J., Polz, M. F. & Cordero, O. X. Microbial interactions lead to rapid micro-scale successions on model marine particles. *Nat. Commun.* **7**, (2016).
72. Szabo, R. E. *et al.* Historical contingencies and phage induction diversify bacterioplankton communities at the microscale. *Proc. Natl. Acad. Sci.* **119**, e2117748119 (2022).
73. Khanijou, J. K. *et al.* Metabolomics and modelling approaches for systems metabolic engineering. *Metab. Eng. Commun.* **15**, e00209 (2022).
74. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8**, 2224 (2017).

75. Yilmaz, P. *et al.* The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Res.* **42**, D643–D648 (2014).
76. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
77. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* (2016). doi:10.7717/peerj.2584
78. R Development Core Team, R. *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing* (2011). doi:10.1007/978-3-540-74686-7
79. Scrucca, L. GA: A Package for Genetic Algorithms in R. *J. Stat. Softw.* **53**, 1–37 (2013).
80. Scrucca, L. On some extensions to GA package: hybrid optimisation, parallelisation and islands evolution. *R J.* **9**, 187–206 (2016).
81. Naylor, R. L. *et al.* A 20-year retrospective review of global aquaculture. *Nat. 2021 5917851* **591**, 551–563 (2021).
82. Golden, C. D. *et al.* Aquatic foods to nourish nations. *Nat. 2021 5987880* **598**, 315–320 (2021).
83. Stentiford, G. D. *et al.* Sustainable aquaculture through the One Health lens. *Nat. Food 2020 18* **1**, 468–474 (2020).
84. Kumar, V., Roy, S., Behera, B. K., Bossier, P. & Das, B. K. Acute Hepatopancreatic Necrosis Disease (AHPND): Virulence, Pathogenesis and Mitigation Strategies in Shrimp Aquaculture. *Toxins 2021, Vol. 13, Page 524* **13**, 524 (2021).
85. Kumar, R., Ng, T. H. & Wang, H. C. Acute hepatopancreatic necrosis disease in penaeid shrimp. *Rev. Aquac.* **12**, 1867–1880 (2020).
86. Reverter, M. *et al.* Aquaculture at the crossroads of global warming and antimicrobial resistance. *Nat. Commun. 2020 111* **11**, 1–8 (2020).
87. Van Boeckel, T. P. *et al.* Global trends in antimicrobial use in food animals. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 5649–5654 (2015).
88. Holt, C. C., Bass, D., Stentiford, G. D. & van der Giezen, M. Understanding the role of the shrimp gut microbiome in health and disease. *J. Invertebr. Pathol.* **186**, 107387 (2021).
89. Reyes, G. *et al.* Microbiome of Penaeus vannamei Larvae and Potential Biomarkers Associated With High and Low Survival in Shrimp Hatchery Tanks Affected by Acute Hepatopancreatic Necrosis Disease. *Front. Microbiol.* **13**, 838640 (2022).
90. Louca, S. *et al.* High taxonomic variability despite stable functional structure across microbial communities. *Nat. Ecol. Evol.* (2017). doi:10.1038/s41559-016-0015
91. Enke, T. N. *et al.* Modular Assembly of Polysaccharide-Degrading Marine Microbial Communities. *Curr. Biol.* **29**, 1528-1535.e6 (2019).
92. Gralka, M., Pollak, S. & Cordero, O. X. Fundamental metabolic strategies of heterotrophic bacteria. *bioRxiv* 2022.08.04.502823 (2022). doi:10.1101/2022.08.04.502823

93. Buchan, A., LeCleir, G. R., Gulvik, C. A. & González, J. M. Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat. Rev. Microbiol. 2014 1210* **12**, 686–698 (2014).
94. Lee, C. Te *et al.* The opportunistic marine pathogen Vibrio parahaemolyticus becomes virulent by acquiring a plasmid that expresses a deadly toxin. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10798–10803 (2015).
95. Gralka, M., Pollak, S. & Cordero, O. X. Fundamental metabolic strategies of heterotrophic bacteria. *bioRxiv* 2022.08.04.502823 (2022). doi:10.1101/2022.08.04.502823
96. Buchan, A., LeCleir, G. R., Gulvik, C. A. & González, J. M. Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat. Rev. Microbiol.* **12**, 686–698 (2014).
97. Sonnenschein, E. C., Jimenez, G., Castex, M. & Gram, L. The Roseobacter-Group Bacterium Phaeobacter as a Safe Probiotic Solution for Aquaculture. *Appl. Environ. Microbiol.* **87**, 1–15 (2021).
98. Sonnenschein, E. C. *et al.* Global occurrence and heterogeneity of the Roseobacter-clade species Ruegeria mobilis. *ISME J. 2017 112* **11**, 569–583 (2016).
99. Breider, S. *et al.* Genome-scale data suggest reclassifications in the Leisingera-Phaeobacter cluster including proposals for Sedimentitalea gen. nov. and Pseudophaeobacter gen. nov. *Front. Microbiol.* **5**, 102427 (2014).
100. Takagi, J. *et al.* Mucin O-glycans are natural inhibitors of Candida albicans pathogenicity. *Nat. Chem. Biol. 2022 187* **18**, 762–773 (2022).
101. Wang, B. X. *et al.* Host-derived O-glycans inhibit toxigenic conversion by a virulence-encoding phage in Vibrio cholerae. *EMBO J.* **42**, e111562 (2023).
102. Werlang, C. A. *et al.* Mucin O-glycans suppress quorum-sensing pathways and genetic transformation in Streptococcus mutans. *Nat. Microbiol. 2021 65* **6**, 574–583 (2021).
103. Valle, J. *et al.* Broad-spectrum biofilm inhibition by a secreted bacterial polysaccharide. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12558–12563 (2006).
104. Zhu, J. *et al.* Quorum-sensing regulators control virulence gene expression in Vibrio cholerae. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3129–3134 (2002).
105. O'Loughlin, C. T. *et al.* A quorum-sensing inhibitor blocks Pseudomonas aeruginosa virulence and biofilm formation. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 17981–17986 (2013).
106. Imai, K., Banno, I. & Iijima, T. INHIBITION OF BACTERIAL GROWTH BY CITRATE. *J. Gen. Appl. Microbiol.* **16**, 479–487 (1970).
107. Balado, M. *et al.* Secreted citrate serves as iron carrier for the marine pathogen Photobacterium damselae subsp damselae. *Front. Cell. Infect. Microbiol.* **7**, 283346 (2017).
108. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol. 2019 378* **37**, 852–857 (2019).
109. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods 2016 137* **13**, 581–583 (2016).

110. Jiang, H., Lei, R., Ding, S. W. & Zhu, S. Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**, 1–12 (2014).
111. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
112. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput. Biol.* **13**, e1005595 (2017).
113. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
114. Cantalapiedra, C. P., Hern̨andez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
115. Rühmann, B., Schmid, J. & Sieber, V. Fast carbohydrate analysis via liquid chromatography coupled with ultra violet and electrospray ionization ion trap detection in 96-well format. *J. Chromatogr. A* **1350**, 44–50 (2014).
116. Xu, G., Amicucci, M. J., Cheng, Z., Galermo, A. G. & Lebrilla, C. B. Revisiting monosaccharide analysis – quantitation of a comprehensive set of monosaccharides using dynamic multiple reaction monitoring. *Analyst* **143**, 200–207 (2017).
117. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods 2012 94* **9**, 357–359 (2012).
118. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
119. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
120. Hoess, R. H., Ziese, M. & Sternberg, N. P1 site-specific recombination: nucleotide sequence of the recombining sites. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 3398 (1982).
121. Silpe, J. E. & Bassler, B. L. A Host-Produced Quorum-Sensing Autoinducer Controls a Phage Lysis-Lysogeny Decision. *Cell* **176**, 268-280.e13 (2019).
122. Gilcrease, E. B. & Casjens, S. R. The genome sequence of Escherichia coli tailed phage D6 and the diversity of Enterobacteriales circular plasmid prophages. *Virology* **515**, 203–214 (2018).
123. Pfeifer, E., Moura De Sousa, J. A., Touchon, M. & Rocha, E. P. C. Bacteria have numerous distinctive groups of phage–plasmids with conserved phage and variable plasmid gene repertoires. *Nucleic Acids Res.* **49**, 2655–2673 (2021).
124. Pfeifer, E., Bonnin, R. A. & Rocha, E. P. C. Phage-Plasmids Spread Antibiotic Resistance Genes through Infection and Lysogenic Conversion. *MBio* **13**, e0185122 (2022).
125. Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R. C. & San Millán, Á. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat. Rev. Microbiol. 2021 196* **19**, 347–359 (2021).

126. Millan, A. S., Escudero, J. A., Gifford, D. R., Mazel, Di. & MacLean, R. C. Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat. Ecol. Evol. 2016 11* **1**, 1–8 (2016).
127. Rodriguez-Beltran, J. *et al.* Multicopy plasmids allow bacteria to escape from fitness trade-offs during evolutionary innovation. *Nat. Ecol. Evol. 2018 25* **2**, 873–881 (2018).
128. Ilhan, J. *et al.* Segregational Drift and the Interplay between Plasmid Copy Number and Evolvability. *Mol. Biol. Evol.* **36**, 472–486 (2019).
129. Garoña, A., Hülter, N. F., Romero Picazo, D. & Dagan, T. Segregational Drift Constrains the Evolutionary Rate of Prokaryotic Plasmids. *Mol. Biol. Evol.* **38**, 5610–5624 (2021).
130. Birky, J. The Inheritance of Genes in Mitochondria and Chloroplasts: Laws, Mechanisms, and Models. *http://dx.doi.org/10.1146/annurev.genet.35.102401.090231* **35**, 125–148 (2003).
131. Luo, H. & Moran, M. A. Evolutionary Ecology of the Marine Roseobacter Clade. *Microbiol. Mol. Biol. Rev.* **78**, 573–587 (2014).
132. Gödeke, J., Paul, K., Lassak, J. & Thormann, K. M. Phage-induced lysis enhances biofilm formation in Shewanella oneidensis MR-1. *ISME J. 2011 54* **5**, 613–626 (2010).
133. Fan, H. & Chu, J. Y. A Brief Review of Short Tandem Repeat Mutation. *Genomics. Proteomics Bioinformatics* **5**, 7–14 (2007).
134. Flynn, J. M., Lower, S. E., Barbash, D. A. & Clark, A. G. Rates and Patterns of Mutation in Tandem Repetitive DNA in Six Independent Lineages of Chlamydomonas reinhardtii. *Genome Biol. Evol.* **10**, 1673–1686 (2018).
135. Mitra, I. *et al.* Patterns of de novo tandem repeat mutations and their role in autism. *Nat. 2020 5897841* **589**, 246–250 (2021).
136. Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4**, 24–33 (1994).
137. Zhou, K., Aertsen, A. & Michiels, C. W. The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiol. Rev.* **38**, 119–141 (2014).
138. Yi, H. *et al.* The Tandem Repeats Enabling Reversible Switching between the Two Phases of β-Lactamase Substrate Spectrum. *PLOS Genet.* **10**, e1004640 (2014).
139. Lwoff, A. LYSOGENY. *Bacteriol. Rev.* **17**, 269–337 (1953).
140. Labrie, S. J., Samson, J. E. & Moineau, S. Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol. 2010 85* **8**, 317–327 (2010).
141. Yooseph, S. *et al.* Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468**, 60–66 (2010).
142. Petersen, J. *et al.* A marine plasmid hitchhiking vast phylogenetic and geographic distances. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 20568–20573 (2019).
143. Yu, M. K., Fogarty, E. C. & Eren, A. M. The genetic and ecological landscape of plasmids in the human gut. *bioRxiv* 2020.11.01.361691 (2022). doi:10.1101/2020.11.01.361691
144. Freese, H. M., Methner, A. & Overmann, J. Adaptation of Surface-Associated Bacteria to the Open Ocean: A Genomically Distinct Subpopulation of

Phaeobacter gallaeciensis Colonizes Pacific Mesozooplankton. *Front. Microbiol.* **8**, 1659 (2017).
145. Freese, H. M. *et al.* Trajectories and Drivers of Genome Evolution in Surface-Associated Marine Phaeobacter. *Genome Biol. Evol.* **9**, 3297–3311 (2017).
146. Wang, J. H. *et al.* Metagenomic analysis of antibiotic resistance genes in coastal industrial mariculture systems. *Bioresour. Technol.* **253**, 235–243 (2018).
147. López-Pérez, M., Kimes, N. E., Haro-Moreno, J. M. & Rodriguez-Valera, F. Not all particles are equal: The selective enrichment of particle-associated bacteria from the mediterranean sea. *Front. Microbiol.* **7**, 996 (2016).
148. Nowinski, B. *et al.* Microbial metagenomes and metatranscriptomes during a coastal phytoplankton bloom. *Sci. Data 2019 61* **6**, 1–7 (2019).
149. Bondy-Denomy, J. *et al.* Prophages mediate defense against phage infection through diverse mechanisms. *ISME J. 2016 1012* **10**, 2854–2866 (2016).
150. de Sousa, J. A. M., Buffet, A., Haudiquet, M., Rocha, E. P. C. & Rendueles, O. Modular prophage interactions driven by capsule serotype select for capsule loss under phage predation. *ISME J. 2020 1412* **14**, 2980–2996 (2020).
151. Hunter, M. & Fusco, D. Superinfection exclusion: A viral strategy with short-term benefits and long-term drawbacks. *PLOS Comput. Biol.* **18**, e1010125 (2022).
152. Wodarz, D., Levy, D. N. & Komarova, N. L. Multiple infection of cells changes the dynamics of basic viral evolutionary processes. *Evol. Lett.* **3**, 104–115 (2019).
153. Moran, M. A. *et al.* Genome sequence of Silicibacter pomeroyi reveals adaptations to the marine environment. *Nat. 2005 4327019* **432**, 910–913 (2004).
154. Fu, H., Uchimiya, M., Gore, J. & Moran, M. A. Ecological drivers of bacterial community assembly in synthetic phycospheres. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 3656–3662 (2020).
155. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 1–11 (2010).
156. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
157. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
158. Schwartzman, J. A. *et al.* Bacterial growth in multicellular aggregates leads to the emergence of complex lifecycles. *bioRxiv* 2021.11.01.466752 (2021). doi:10.1101/2021.11.01.466752
159. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods 2012 97* **9**, 671–675 (2012).
160. 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* **8**, 1–23 (2020).
161. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol. 2017 3511* **35**, 1026–1028 (2017).
162. Gilchrist, C. L. M. & Chooi, Y. H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).
163. Kahle, D. & Wickham, H. ggmap: Spatial Visualization with ggplot2.

164. Parada, V., Herndl, G. J. & Weinbauer, M. G. Viral burst size of heterotrophic prokaryotes in aquatic systems. *J. Mar. Biol. Assoc. United Kingdom* **86**, 613–621 (2006).
165. Weitz, J. S., Li, G., Gulbudak, H., Cortez, M. H. & Whitaker, R. J. Viral invasion fitness across a continuum from lysis to latency. *Virus Evol.* **5**, (2019).
166. Haudiquet, M., Buffet, A., Rendueles, O. & Rocha, E. P. C. Interplay between the cell envelope and mobile genetic elements shapes gene flow in populations of the nosocomial pathogen Klebsiella pneumoniae. *PLOS Biol.* **19**, e3001276 (2021).
167. Rendueles, O., de Sousa, J. A. M., Bernheim, A., Touchon, M. & Rocha, E. P. C. Genetic exchanges are more frequent in bacteria encoding capsules. *PLOS Genet.* **14**, e1007862 (2018).

Appendix

Appendix A Supplementary material for chapter 2

A.1 Formulation of Ensemble Quotient

In this section, we will prove that microbiome coarse-graining guided by a uniform phenotypic variable, a continuous phenotypic variable or a categorical phenotypic variable (Figure S1) can be generalized into a unified and simple mathematical framework featuring Ensembled Quotient

$$Ensemble\ Quotient \coloneqq \frac{\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x}}, \boldsymbol{x} \in (0,1)^n$$

where n is the dimension of the microbiome, e.g., the number of OTUs in a microbiome. The $\boldsymbol{x}$ is a Boolean vector of length n, for which 1 or 0 represent presence or absence of a species in the ensemble. For instance, $\boldsymbol{x} = [1,1,0,1,0]$ indicates that there are in total 5 species in the microbiome, in which the 1st, 2nd and 4th species are coarse-grained into the ensemble. Clearly, Ensemble Quotient has a quadratic fractional form.

A.1.1 Continuous phenotypic variable

For a continuous phenotypic variable, we need to determine which individuals should be taken into the ensemble and which ones not, so that the ensemble is strongly correlated with the external phenotypic variable of interest.

Let us start by considering a microbiome matrix $\boldsymbol{M}$ with m rows of samples and n columns of species, where the element in i-th row and j-th column M_{ij} is the relative abundance of species j in sample i. Then the product of $\boldsymbol{M}$ and $\boldsymbol{x}$

$$\boldsymbol{s} = \boldsymbol{M}\boldsymbol{x}$$

gives us a new vector $\boldsymbol{s}$ of length m whose k-th element is the relative abundance of the assemblage in the k-th sample. As for the external phenotypic variable (e.g., concentration of a metabolite), let us denote it as a vector $\boldsymbol{y}$ of length m, so that the l-th element of $\boldsymbol{y}$ becomes the concentration readout in sample #l. We then need to examine the correlation between the assemblage vector $\boldsymbol{s}$ and the external variable $\boldsymbol{y}$. Note that vectors $\boldsymbol{s}$, $\boldsymbol{y}$, and all the column vectors in matrix $\boldsymbol{M}$ representing each species can be scaled by subtracting their means without affecting correlations. Let us denote the scaled microbiome matrix, assemblage vector and phenotypic vector as $\boldsymbol{M_0}$, $\boldsymbol{s_0}$ and $\boldsymbol{y_0}$, respectively, since the scaling can greatly simplify the algebra below. However, note that for a uniform phenotypic variable to be discussed in section 1.3, vector scaling will not be allowed as a zero in the denominator will ruin the analytics. Still, we have

$$\boldsymbol{s_0} = \boldsymbol{M_0}\boldsymbol{x}\ (Eq.S1)$$

For linear regression with continuous variables, the goodness of prediction is often quantified as coefficient of determination, which is statistically defined as the ratio of explained sum of squares (ESS) over the total sum of squares (TSS)

$$R^2 = \frac{ESS}{TSS} = \frac{(\widehat{\boldsymbol{y_0}} - \overline{\boldsymbol{y_0}})^T(\widehat{\boldsymbol{y_0}} - \overline{\boldsymbol{y_0}})}{(\boldsymbol{y_0} - \overline{\boldsymbol{y_0}})^T(\boldsymbol{y_0} - \overline{\boldsymbol{y_0}})}$$

For scaled $\boldsymbol{y_0}$ we have the mean $\overline{\boldsymbol{y_0}} = \boldsymbol{0}$. Combining the projection formulation in linear algebra

$$\widehat{\boldsymbol{y_0}} = \frac{\boldsymbol{s_0}^T\boldsymbol{y_0}}{\boldsymbol{s_0}^T\boldsymbol{s_0}}\boldsymbol{s_0}$$

we have

$$R^2 = \frac{\boldsymbol{s_0}^T \boldsymbol{y_0} \boldsymbol{y_0}^T \boldsymbol{s_0}}{\boldsymbol{s_0}^T \boldsymbol{s_0} \cdot \boldsymbol{y_0}^T \boldsymbol{y_0}}$$

in which $\boldsymbol{y_0}^T \boldsymbol{y_0}$ can be neglected in the optimization since it is just comprised of known constants.

Plugging in (Eq. S1), the problem of maximizing R^2 can be reformulated as the following integer programming with an objective function in the form of quadratic quotient

$$max\ \frac{\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x}}$$

$$s.t. \boldsymbol{x} \in (0,1)^n$$

where

$$\boldsymbol{Q} = \boldsymbol{M_0}^T \boldsymbol{y_0} \boldsymbol{y_0}^T \boldsymbol{M_0}$$

$$\boldsymbol{P} = \boldsymbol{M_0}^T \boldsymbol{M_0}$$

Note that the coefficient of determination (R^2) for linear regression is numerically equivalent to Pearson's correlation coefficient in its square form (r^2), although they are based on slightly different statistical contexts. In theory, it is possible that maximizing R^2 might lead to an assemblage with very strong negative correlation with the phenotype. This is in fact not an issue for microbiome datasets due to its compositionality, since the remaining set of species will automatically have strong positive correlation with the phenotype (same correlation strength but just with the opposite sign). In a few cases where the sign of the optimized assemblage indeed matters, we can adopt a modified formulation to maximize Pearson's correlation coefficient, but not in its square form

$$r^2 = \frac{Cov(\boldsymbol{s_0}, \boldsymbol{y_0})}{\sigma(\boldsymbol{s_0})\sigma(\boldsymbol{y_0})} = \frac{\boldsymbol{s_0}^T \boldsymbol{y_0}}{\sqrt{\boldsymbol{s_0}^T \boldsymbol{s_0} \cdot \boldsymbol{y_0}^T \boldsymbol{y_0}}}$$

which can be also rewritten with $\boldsymbol{x}$ as the unknown variable as

$$max \frac{\boldsymbol{x}^T \boldsymbol{M_0}^T \boldsymbol{y_0}}{\sqrt{\boldsymbol{x}^T \boldsymbol{M_0}^T \boldsymbol{M_0 x}}}$$

A.1.2 Categorical phenotypic variable

The case with categorical phenotypic variable (healthy vs. disease, or several different sub-types of diseases) is conceptually the same as the case with continuous phenotypic variable. For a categorical phenotypic variable, an equivalent statistical metric to the coefficient of determination can be also constructed by the sum of square between/among treatments against the total of sum of square

$$R^2 = \frac{SS_{bewteen\ treatment}}{SS_{total}} = \frac{(\boldsymbol{s_0}^T \boldsymbol{YL})^2}{\boldsymbol{s_0}^T \boldsymbol{s_0}}$$

where $\boldsymbol{Y}$ is an augmented categorical matrix with m rows and c columns. The row number m is the same as the number of total samples and the column number c is the same as the number of categories. For instance, if there are altogether 5 samples in the microbiome dataset where samples 1, 2, 4 are from healthy hosts and samples 3,5 from hosts with a disease, then the matrix $\boldsymbol{Y}$ will be

$$\boldsymbol{Y} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$\boldsymbol{L}$ is a diagonal matrix whose diagonal elements are given by the inverse square root of the number of samples in each category.

Combining (Eq. S1) again, we get

$$R^2 = \frac{\boldsymbol{x}^T \boldsymbol{M_0}^T \boldsymbol{Y} \boldsymbol{L} \boldsymbol{L}^T \boldsymbol{Y}^T \boldsymbol{M_0} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{M_0}^T \boldsymbol{M_0} \boldsymbol{x}}$$

Clearly, we again reach the same quadratic quotient form of

$$min\ \frac{\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x}}$$

$$s.t. \boldsymbol{x} \in (0,1)^n$$

where

$$\boldsymbol{Q} = \boldsymbol{M_0}^T \boldsymbol{Y} \boldsymbol{L} \boldsymbol{L}^T \boldsymbol{Y}^T \boldsymbol{M_0}$$

$$\boldsymbol{P} = \boldsymbol{M_0}^T \boldsymbol{M_0}$$

A.1.3 Uniform phenotypic variable

With a uniform phenotypic variable, we are not allowed to directly construct a correlation/regression-type of R^2 although in spirit it can be also regarded as special correlation/regression problem with a constant number across all samples. Doing so implies that the variance of the phenotypic variable is zero, which becomes illegal as a denominator.

Here, we examine the coefficient of variation of the ensemble, which is commonly adopted to quantify the level of variation across samples. In statistics, the coefficient of variation is defined as standard deviation divided by mean,

$$CV = \frac{\sigma}{\mu}\ (Eq. S2)$$

As aforementioned, here we are no longer able to simply the expression of variance by scaling the vectors as $\mu(\boldsymbol{s}) = 0$ is not acceptable as a denominator. To derive the full

expression of the variance of $\boldsymbol{s}$, we need an auxiliary unit vector $\mathbf{1}$ with length m whose elements are all ones. Then we have

$$nVar(\boldsymbol{s}) = (\boldsymbol{s} - \mu\mathbf{1})^T(\boldsymbol{s} - \mu\mathbf{1}) \ (Eq.S3)$$

Where μ, a scalar, denotes the mean of the ensemble $\boldsymbol{s}$, which can be at the same time expressed as

$$n \cdot \mu = \mathbf{1}^T\boldsymbol{s} = \boldsymbol{s}^T\mathbf{1} \ (Eq.S4)$$

Combining (Eq. S1~S4), we can derive that

$$CV^2 \sim \frac{\boldsymbol{x}^T\left(\boldsymbol{M}^T\boldsymbol{M} - \frac{2}{n}\boldsymbol{M}^T\mathbf{1}\mathbf{1}^T\boldsymbol{M} + \frac{1}{n^2}\boldsymbol{M}^T\mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T\boldsymbol{M}\right)\boldsymbol{x}}{\boldsymbol{x}^T(\boldsymbol{M}^T\mathbf{1}\mathbf{1}^T\boldsymbol{M})\boldsymbol{x}}$$

Again, we arrive at the form of the microbiome ensemble quotient. Similarly, we are able to reformulate coarse-graining problem into the following integer programming format

$$max \ \frac{\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{P}\boldsymbol{x}}$$

$$s.t.\boldsymbol{x} \in (0,1)^n$$

where

$$\boldsymbol{Q} = \boldsymbol{M}^T\mathbf{1}\mathbf{1}^T\boldsymbol{M}$$

$$\boldsymbol{P} = \boldsymbol{M}^T\boldsymbol{M} - \frac{2}{n}\boldsymbol{M}^T\mathbf{1}\mathbf{1}^T\boldsymbol{M} + \frac{1}{n^2}\boldsymbol{M}^T\mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T\boldsymbol{M}$$

Conceptually, the efforts to minimize the coefficient of variation is analogous to maximize the “correlation” with a uniform vector (e.g., $\boldsymbol{y} = [1,1,1,1,1]$) although it is illegal in practice since such uniform vectors have zero variance. Nevertheless, this

analogy can be helpful for understanding the generalization of stabilizing grouping and associating grouping into the same formulation.

In addition, two simple linear constrains are required to avoid generating an empty group or a group with all taxa that are numerically stable but ecologically trivial

$$(\boldsymbol{M}^T \mathbf{1}_m)^T \boldsymbol{x} \geq me$$

$$(\boldsymbol{M}^T \mathbf{1}_m)^T \boldsymbol{x} \leq m(1-e)$$

In those constraints, $\mathbf{1}_m$ is a unit vector with length m whose elements are all ones. These constraints imply that the average abundance of the aggregated group across all the samples should be no smaller than a threshold given by e (for instance 5%) and no larger than a threshold given by $1-e$ (for instance 95%).

A.2 Optimization of Ensemble Quotient

A.2.1 Reformulation into a mixed integer linear programming (MILP) problem

The binary nature of coarse-graining provides the ensemble quotient with inherent mathematical simplicity. It has been shown in integer programming studies that quadratic fractional optimization, when the variables are integers, can be reduced to an equivalent mixed integer linear programming problem (Gaur and Arora, 2008). The method includes Charnes-Cooper transformation incorporated with Glover's linearization, which has been widely used in solving fractional programming problems (Yue et al., 2013). Briefly, the reformulation-linearization algorithm works by introducing

new variables with linear constrainwts. In this sense, the above quadratic fractional programming problem can be reformulated into

$$min\ \mathbf{1}^{\boldsymbol{T}}\boldsymbol{\alpha} - m_p\boldsymbol{k}$$

$$s.t.\ \boldsymbol{\alpha} + \boldsymbol{\beta} - \boldsymbol{P}\boldsymbol{k} - m_p u\mathbf{1} = \mathbf{0}\ (C1)$$

$$\boldsymbol{\alpha} - 2m_p u\mathbf{1} + 2m_p\boldsymbol{k} \leq \mathbf{0}\ (C2)$$

$$\boldsymbol{\beta} - 2m_p\boldsymbol{k} \leq \mathbf{0}\ (C3)$$

$$\boldsymbol{\gamma} + \boldsymbol{\delta} - \boldsymbol{Q}\boldsymbol{k} - m_q u\mathbf{1} = \mathbf{0}\ (C4)$$

$$\boldsymbol{\gamma} - 2m_q u\mathbf{1} + 2m_q\boldsymbol{k} \leq \mathbf{0}\ (C5)$$

$$\boldsymbol{\delta} - 2m_q\boldsymbol{k} \leq \mathbf{0}\ (C6)$$

$$\mathbf{1}^T\boldsymbol{\delta} - m_q\boldsymbol{k} = \mathbf{1}\ (C7)$$

$$\boldsymbol{k} - u\mathbf{1} - R\boldsymbol{x} \geq -R\mathbf{1}\ (C8)$$

$$\boldsymbol{k} - u\mathbf{1} \leq \mathbf{0}\ (C9)$$

$$\boldsymbol{k} - R\boldsymbol{x} \leq \mathbf{0}\ (C10)$$

$$u - R \leq 0\ (C11)$$

Where $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{k} \geq \mathbf{0}$ are vectors of length n, $\boldsymbol{x} \in (0,1)^n$ is the binary variable to be solved through optimization, R is a sufficiently large number, m_p and m_q are defined as an upper bound of row sums of $\boldsymbol{P}$ and $\boldsymbol{Q}$, respectively.

$$m_p = \max_i(\sum_j |P_{ij}|)$$

$$m_q = \max_i(\sum_j |Q_{ij}|)$$

In this way, the binary quadratic fraction programming problem with n unknown variables is reformulated into an MILP with $6n + 1$ unknown variables embedded with $9n + 2$ linear constraints. The reformulated MILP is directly accessible by commercially available integer optimizers such as Gurobi and Cplex. However, MILP itself is still NP-hard in nature, making this only applicable to small-scale problems.

A.2.2 Genetic algorithm and Aggregation Network

An alternative approach to optimize the Ensemble Quotient to leverage heuristic algorithms. In particular, the binary nature of the unknown variable x in the formulation makes Genetic Algorithm an ideal candidate. Briefly, the presence or absence of a species in the assemblage is to be regarded as dominance or recessive of a loci in a genetic algorithm. Then a Markov-chain based stochastic search is implemented to simulate the mutation, recombination and selection of the “genotype” based on its fitness given by the object function for optimization. In our case, the fitness function is the Ensemble Quotient. In this way, the optimal assemblage given by the genetic algorithm is analogous to a genotype “evolved” towards the peak in the fitness landscape.

A potential shortcoming of genetic algorithm lies in the fact that it only converges in probability as heuristics. However, for microbiome studies we are more interested in gaining biological insights from the most robust statistical patterns than recovering the very exact optimal solutions in complex high-dimensional datasets. In light of that, we have developed a cross-validation-based algorithm to capture and characterize the

most important cohesive guilds. Briefly, the relative importance of single species and species pairs are evaluated as cumulative cross-validation R^2 for assemblages where that single species or the species pair is present, which can be visualized as the node size and edge width in an Aggregation Network (Methods). One can then infer the most important species that should be grouped together by examining strongly connected big nodes in an Aggregation network, as in the Tara Oceans case (Figure 3B).

A.2.3 Boolean least square regression for a continuous phenotypic variable

Specifically, when the phenotypic variable is continuous, the Ensemble Quotient can be reformulated into the most ordinary least square format. In addition to the slope k and intercept b, we also need Boolean variables to indicate the presence or absence of each species in the assemblage.

$$\boldsymbol{y} \sim k(\sum_{\boldsymbol{j}} x_j \boldsymbol{m_j}) + b$$

Where $\boldsymbol{m_j}$ is column j in microbiome matrix $\boldsymbol{M}$ indicating the relative abundance of species j in each sample. Again, x_j, the binary variable, denotes whether species j should be included in the assemblage. Then the least square error is

$$\boldsymbol{e^2} = \|\boldsymbol{Az} - \boldsymbol{y}\|_2 = \boldsymbol{z}^T\boldsymbol{A}^T\boldsymbol{Az} - \boldsymbol{2y}^T\boldsymbol{Az} + \boldsymbol{y}^T\boldsymbol{y}$$

Where $\boldsymbol{A}_{m\times(n+1)}$ is the augmented microbiome matrix $[\mathbf{1}_m \vdots \boldsymbol{M}_{m\times n}]$ since we need an additional column of 1 to map the linear intercept. $\boldsymbol{z}$ is the augmented vector for unknown variable $[b \vdots k\boldsymbol{x}]$.

Also noting that $\boldsymbol{y}^T\boldsymbol{y}$ can be neglected as a known term, we have

$$min\ \boldsymbol{z}^T\boldsymbol{Q}\boldsymbol{z} + \boldsymbol{L}\boldsymbol{z}$$

where $\boldsymbol{Q} = \boldsymbol{A}^T\boldsymbol{A}$, $\boldsymbol{L} = -\mathbf{2}\boldsymbol{y}^T\boldsymbol{A}$.

In this way, we are going to solve a mixed integer quadratic programming problem, which is also solvable by commercial optimizer such as Gurobi (later than v8.0).

A.3 Interpretation of Ensemble Quotient

Here we discuss the simplest case with a continuous phenotypic variable. It can be easily generalized since categorical/uniform phenotypic variables can be regarded as special forms of continuous phenotype cases as detailed in section 1. Recall the expression of R^2 for a continuous phenotypic variable as

$$R^2 \sim EQ = \frac{(\boldsymbol{y_0}^T\boldsymbol{s_0})^2}{\boldsymbol{s_0}^T\boldsymbol{s_0}}$$

Which can be rewritten as

$$EQ = \frac{[\boldsymbol{y}^T(\boldsymbol{S}\cdot\mathbf{1})]^2}{\mathbf{1}^T\boldsymbol{S}^T\boldsymbol{S}\mathbf{1}}$$

Where $\boldsymbol{S}$ is the assemblage matrix (the subset of the whole microbiome matrix $\boldsymbol{M}$ but only with species included in the assemblage). Again, neglecting the constant term $\boldsymbol{y}^T\boldsymbol{y}$, the expression above can be transformed into a form that is more statistically interpretable,

$$EQ = \frac{(\overline{cov_{(x_i,y)}})^2}{\overline{cov_{(x_i,x_j)}}}$$

Where $\overline{cov_{(x_i,y)}}$ means the average covariance between each species and the phenotypic variable, $\overline{cov_{(x_i,x_j)}}$ means the average covariance between each species and each species. In the simplest scenario where all vectors have unit standard deviation, it can be further simplified as

$$EQ = \frac{(\overline{r_{(x_i,y)}})^2}{\overline{r_{(x_i,x_j)}}}$$

Which basically reflects the ratio of average species-phenotype correlation and average species-species correlation. On one hand, if species within an assemblage have opposite signs of correlation with a phenotype, then their average would be nearly zero, resulting in small value of the nominator (i.e., poor consistency). On the other hand, if species within an assemblage tend to be strongly correlated with each other, then their average would lead to a large value of the denominator (i.e., poor complementarity). An optimal assemblage should be formed by species with strong and consistent correlations with the phenotype but weak or even negative correlations with each other.

A.4 Supplementary figures

Figure A.1 Schematic illustration of three scenarios for EQO. Bar plots show relative abundance of different taxa. Appropriate grouping of taxa leads to strong coupling with the phenotypic variable. For a continuous phenotypic variable such as measured concentration of a metabolite across samples, the coupling can be captured by a strong correlation. For a uniform phenotypic variable, the coupling can be marked as high stability or low variability. For a categorical phenotypic variable, the coupling can be interpreted as significant discrimination between treatments.

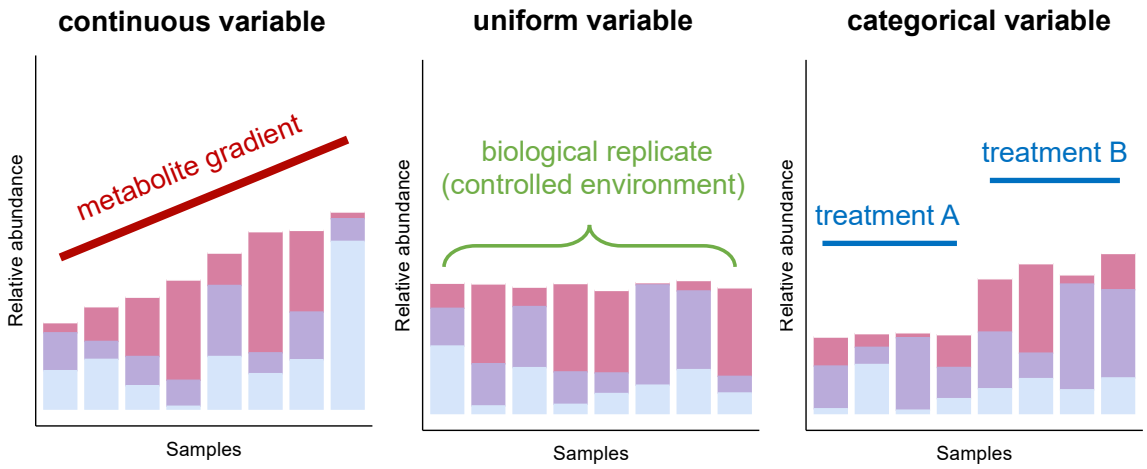


Figure A.2 Details of EQO results on Tara Oceans microbiome. (A) The best group size for EQO was determined to be 11 based on an AIC minimization criterion. (B) Relative abundance distribution of the 11 taxa selected by the algorithm across all sampling sites (left y-axis). Nitrate concentration measured at each sampling site was shown as black dots (right y-axis).

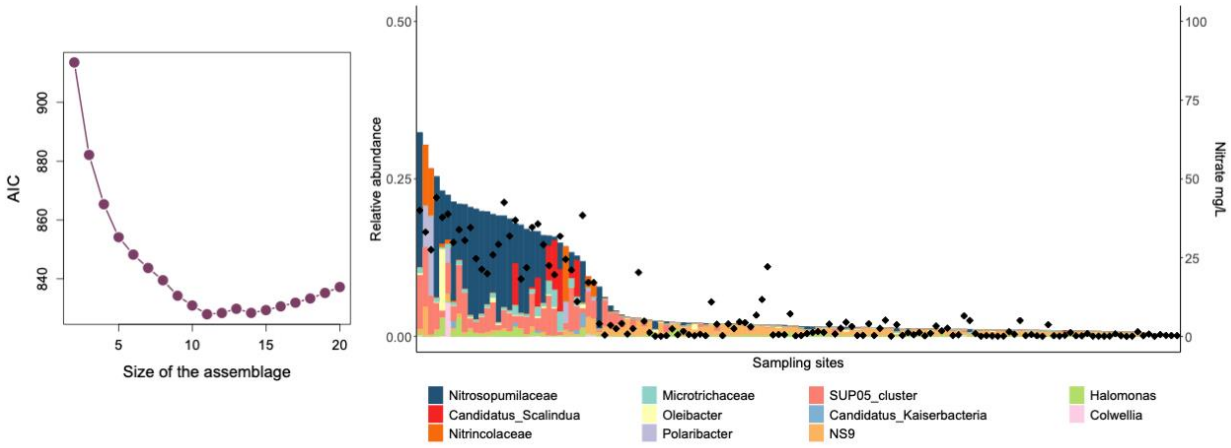


Figure A.3 Metabolite selection in the animal gut microbiome based on statistical tests. Metabolites with cross-validated R2 lower than 0.2 as well as adjusted P-value higher than 0.01 were filtered out from further analysis (Methods).

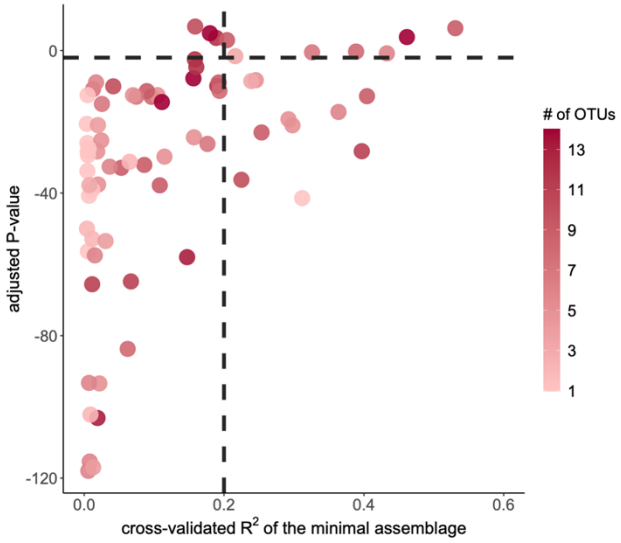
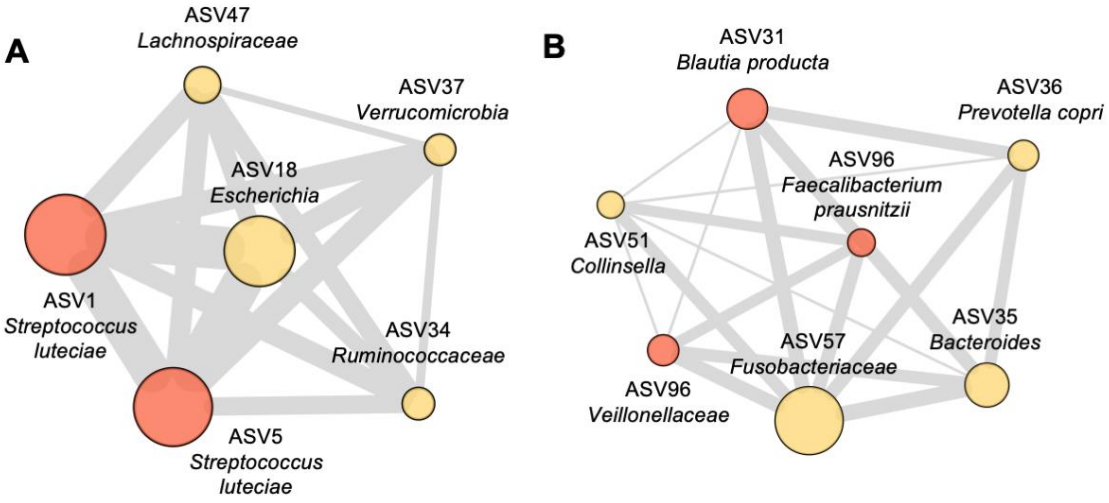


Figure A.4 Predicting lactate or butyrate producers in animal gut microbiome. Species that are confirmed to be lactate producers or butyrate producers in previous experimental studies are highlighted in red. The two *Streptococcus luteciae* species in the left panel are well-known lactate producers in the Lactobacillales clade. *Faecalibacterium prausnitzii* and *Blautia producta* as well as Lachnospiraceae species have been reported to be butyrate producers.



Appendix B Supplementary material for Chapter 3

B.1 Supplementary figures

Figure B.1 Pearson's correlation coefficient between all ASVs and survival rate of shrimp larvae. Each dot represents an ASV colored according family-level taxonomy. Size of dots represents the average relative abundance across all tanks.

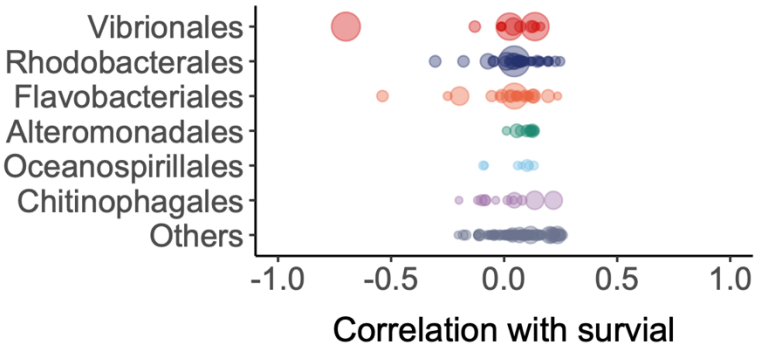


Figure B.2 Relative abundance of Vp in the other three temporally-tracked tanks.

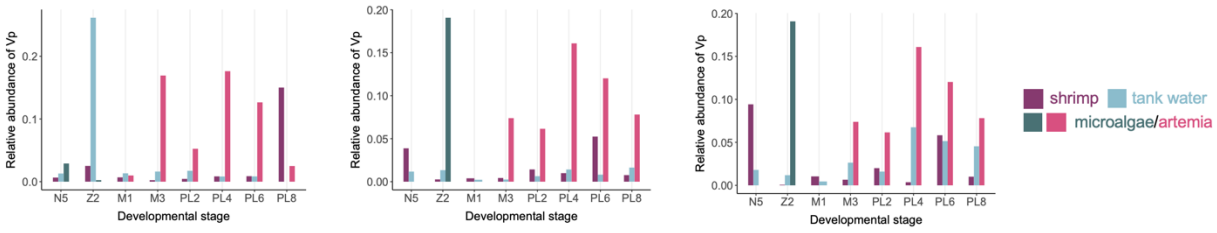


Figure B.3 Distribution of Pearson's correlation coefficient of Vp and all the other ASVs

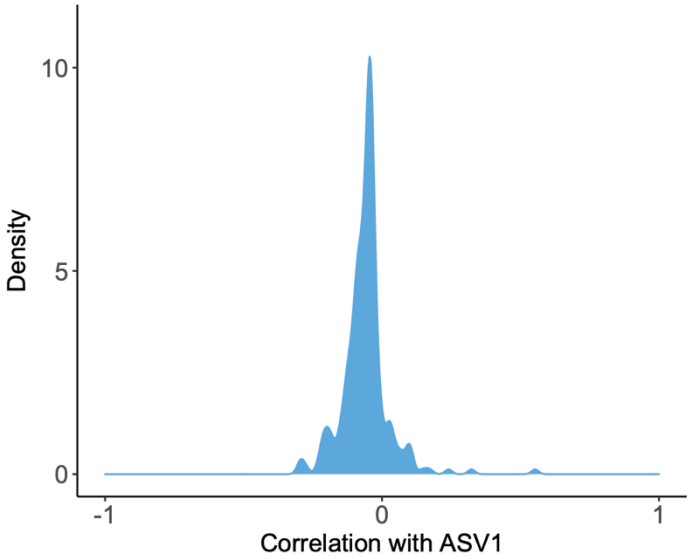
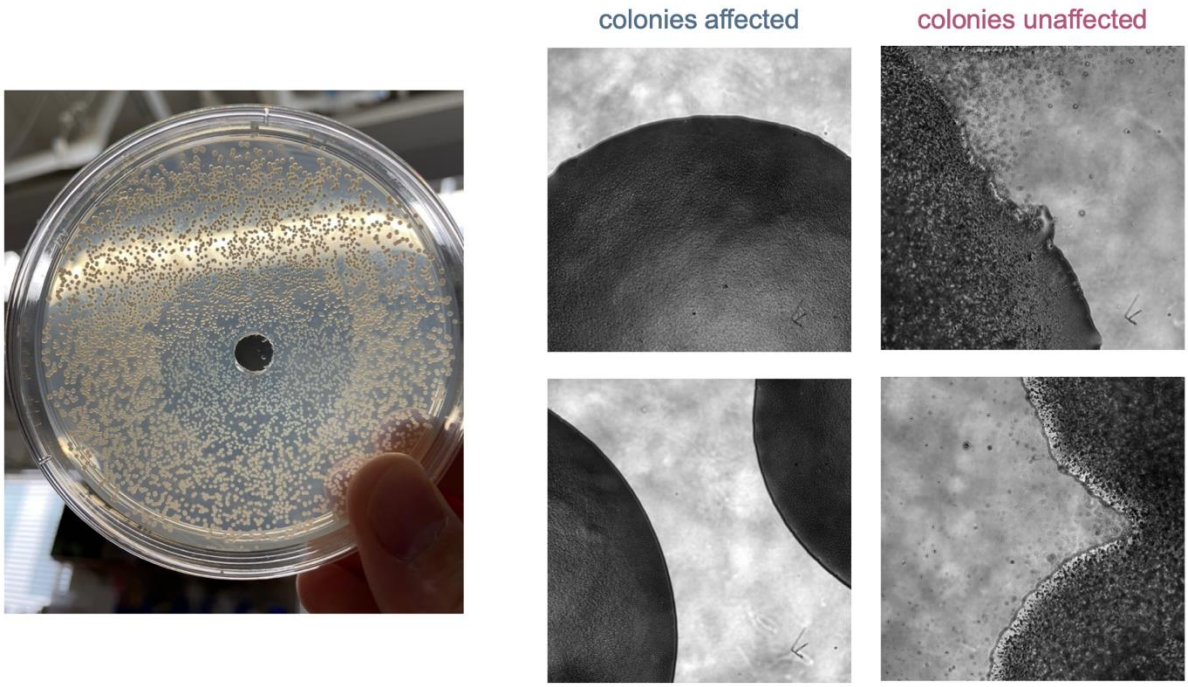


Figure B.4 Morphological changes of Vp colonies affected or unaffected by the cell-free supernatant of *Sedimentitalea sp*. Colonies affected by the *Sedimentitalea sp.* are smooth with clear boundaries, while colonies unaffected have a lot of cells dispersed beyond the colony boundary.



Appendix C Supplementary material for Chapter 4

C.1 Supplementary table

Table C.1 Details of all 21 mutations identified in 15 independent lines of populations. These include 9 lines (populations a-i) of temporally-tracked populations with genomic sequencing, 3 lines with transcriptomic sequencing (populations j-l) and 3 lines of populations that we only sequenced colonies at the end timepoint (populations m-o). Notably, some mutations happened in parallel in two independent lines, which are marked in red.

	position	population	source	reference	alteration
1	11211	j	transcriptomics	GTTTTTT	GTTTTTTT
2	11345	m	genomics	T	G
3	11345	e	genomics	T	G
4	11363	g	genomics	AG	AGG
5	11364	h	genomics	GT	GTT
6	11366	n	genomics	A	AT
7	11366	c	genomics	A	T
8	11732	b	genomics	TGGTGAGGT	TGGTGAGGTGAGGT
9	11736	i	genomics	GA	G
10	11750	b	genomics	CAA	CAAA
11	11750	d	genomics	CAA	CAAAA
12	11771	d	genomics	CG	CGG
13	11806	l	transcriptomics	TC	T
14	11833	k	transcriptomics	G	GC
15	11840	a	genomics	ATTT	ATT
16	11853	i	genomics	C	T
17	11887	a	genomics	CG	C
18	11897	o	genomics	GAAAAA	GAAAA
19	11897	b	genomics	GAAAAA	GAAAA
20	11942	a	genomics	CTT	CT
21	11942	f	genomics	CTT	CT

C.2 Supplementary figures

Figure C.1 Assembly graph of *Tritonibacter mobilis* A3R06 complete genome. The whole genome has 4.65 million base pairs (Mbp), with a chromosome of 3.2 Mbp plus four (mega)plasmids of 1.2 Mbp, 0.1 Mbp, 78 thousand base pairs (Kbp) and 42 (Kbp).

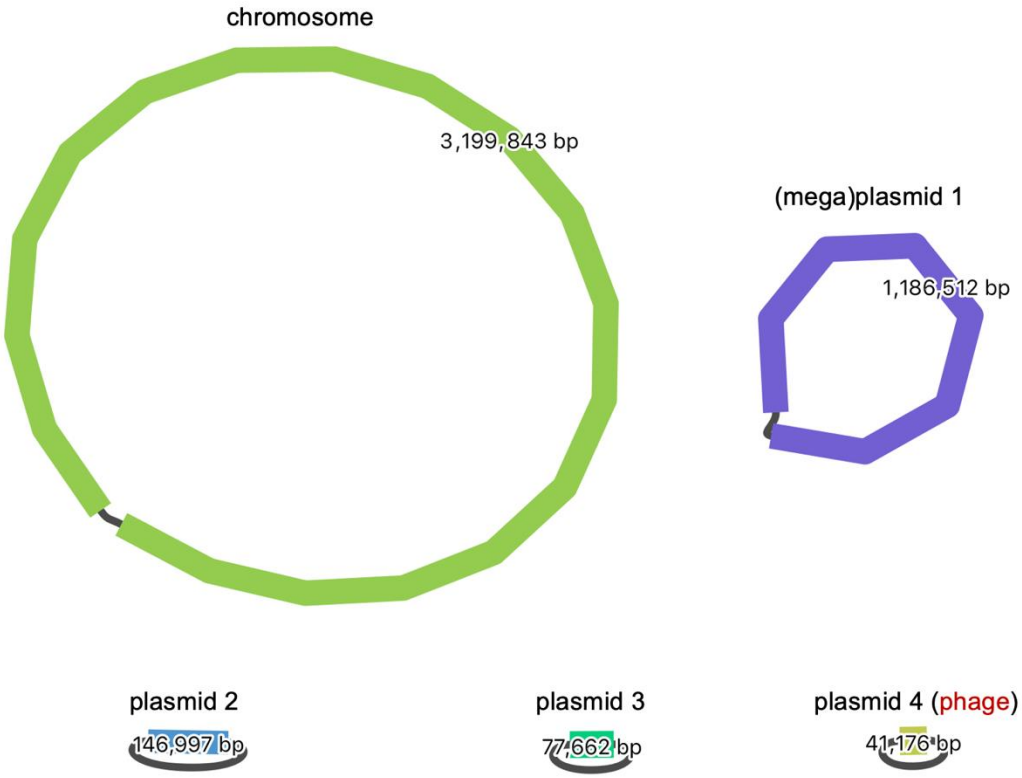


Figure C.2 Genome map of *Tritonibacter mobilis* A3R06 phage-plasmid. The phage-plasmid has 51 predicted genes, including phage structural genes (e.g., a phage head, tail, capsid and portal proteins, purple), lysozyme (orange), cell wall hydrolases (orange) as well as a C1-type phage repressor (red). The rest of the genes are involved in plasmid stability, replication and segregation, such as the y*oeB-yefM* toxin-antitoxin system (pink), the *parAB* plasmid segregation system (blue), P4-family plasmid primase (yellow) as well as various methylases (green) for epigenetic modifications.

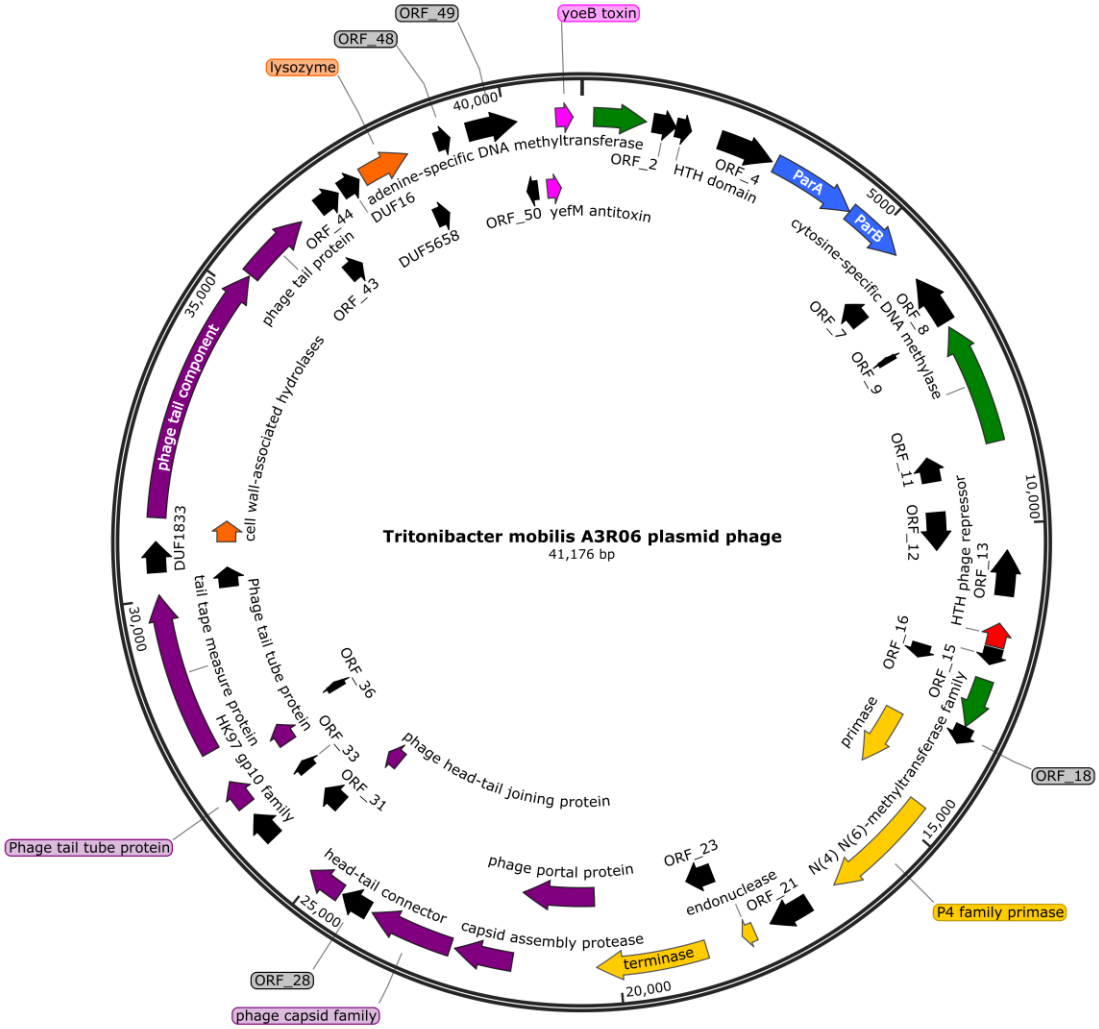


Figure C.3 Manually measured growth curve of *Tritonibacter mobilis* A3R06. Culture growth followed an experimental physiology protocol including appropriate pre-culturing and temperature-controlled shaking conditions, as detailed in Methods. Guided by the measured doubling time of ~ 4h, we set the period of each serial dilution to be 24 hours and dilution factor to be 1/64. This translated to ~6 generations per cycle and the cells are always in the exponential phase throughout the cycles. Source data are provided in the Source Data file.

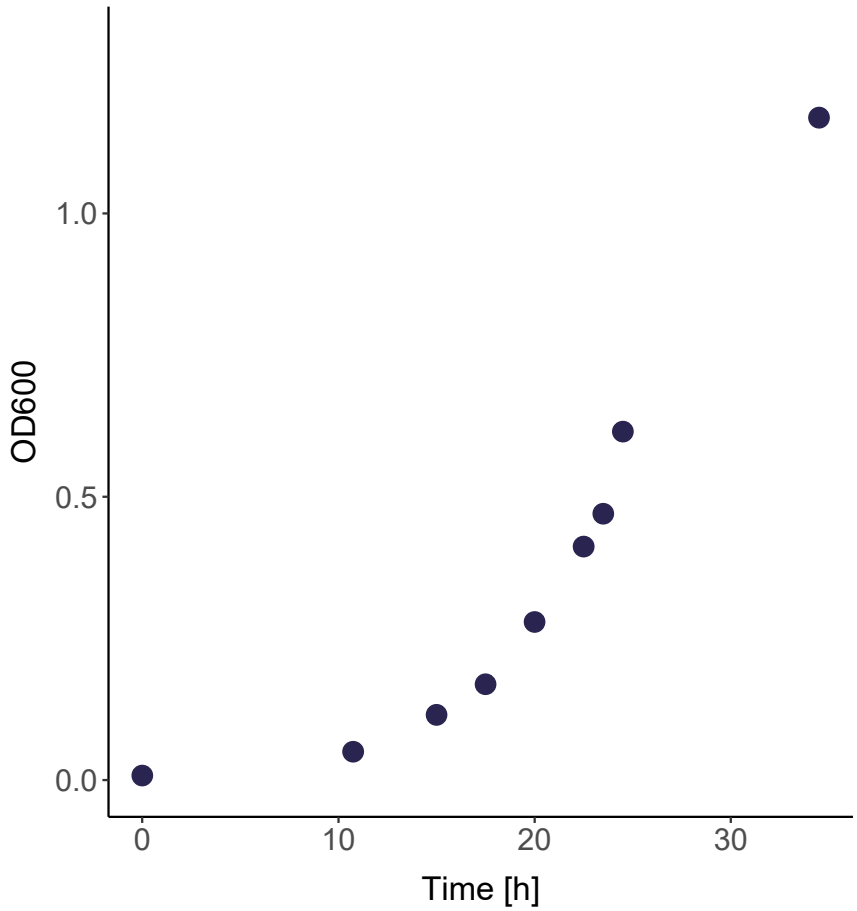


Figure C.4 Eco-evolutionary dynamics of the other 8 independent lines of populations that were temporally-tracked with genomic sequencing. Each line of population (populations a-i) started from a different single colony. Bacterial culture was collected at the end of each dilution cycle, for which we did genome sequencing to determine the genotype frequencies of mutations (A) as well as the relative copy number of the phage-plasmid compared to the host chromosome (C). OD600 was also measured at the end of each dilution cycle (B). Source data are provided in the Source Data file.

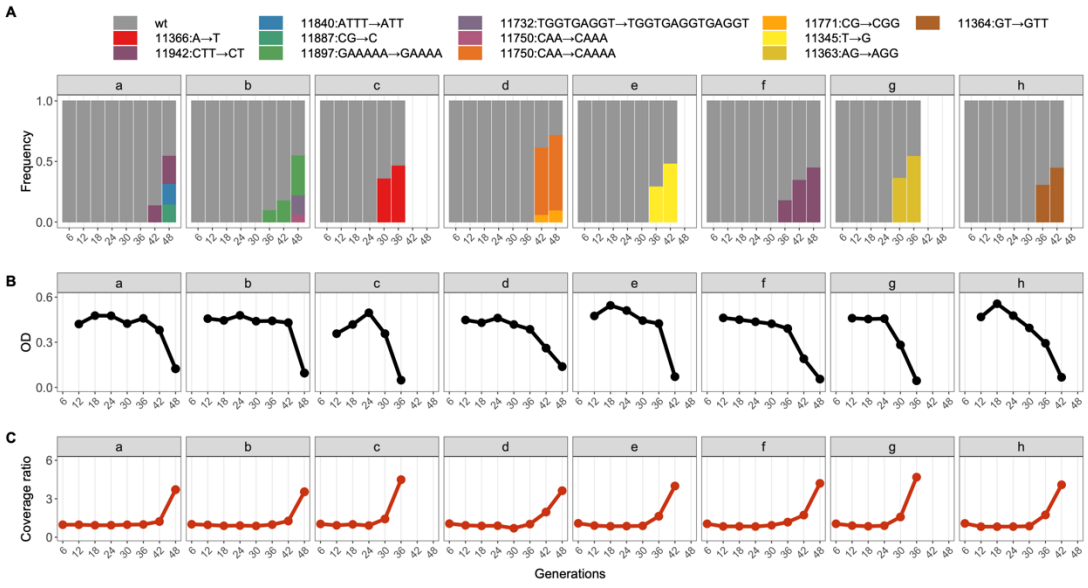


Figure C.5 Images of *Tritonibacter mobilis* A3R06 culture. (a) Liquid culture was planktonic without phage-plasmid induction. (b) Culture became clumpy with cell aggregates when the mutated phage-plasmid was induced. (c) Fluorescent microscopic image of the cell aggregates. Green light by SYTO9 visualize live cells by binding to intracellular DNA while red light by propidium iodide visualize dead cells by binding to extracellular DNA. The same experiments were repeated for 4 times, all yielding similar results.

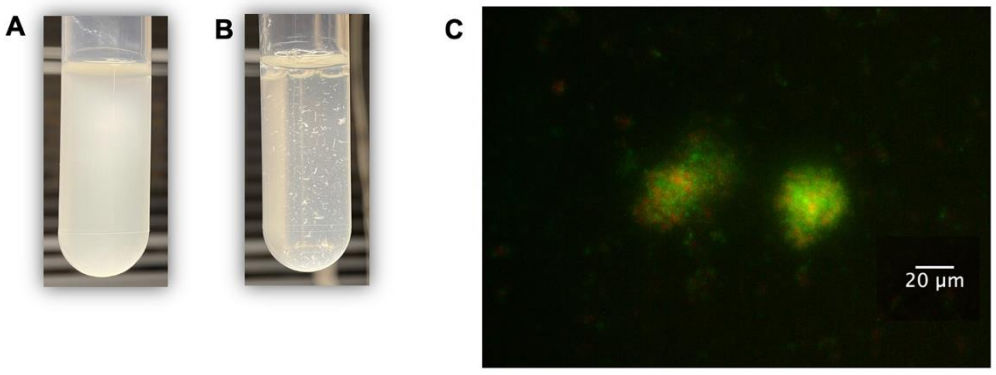


Figure C.6 Expression levels of phage-plasmid genes before and after the mutation was observed. Genes related to plasmid replication and stability were constituently expressed even when the phage is lysogenic. Genes that are related to phage production became highly up-regulated after productive switch. Data are presented as mean ± SEM based on N=3 biologically independent replicates. Source data is provided in the Supplementary Data 1.

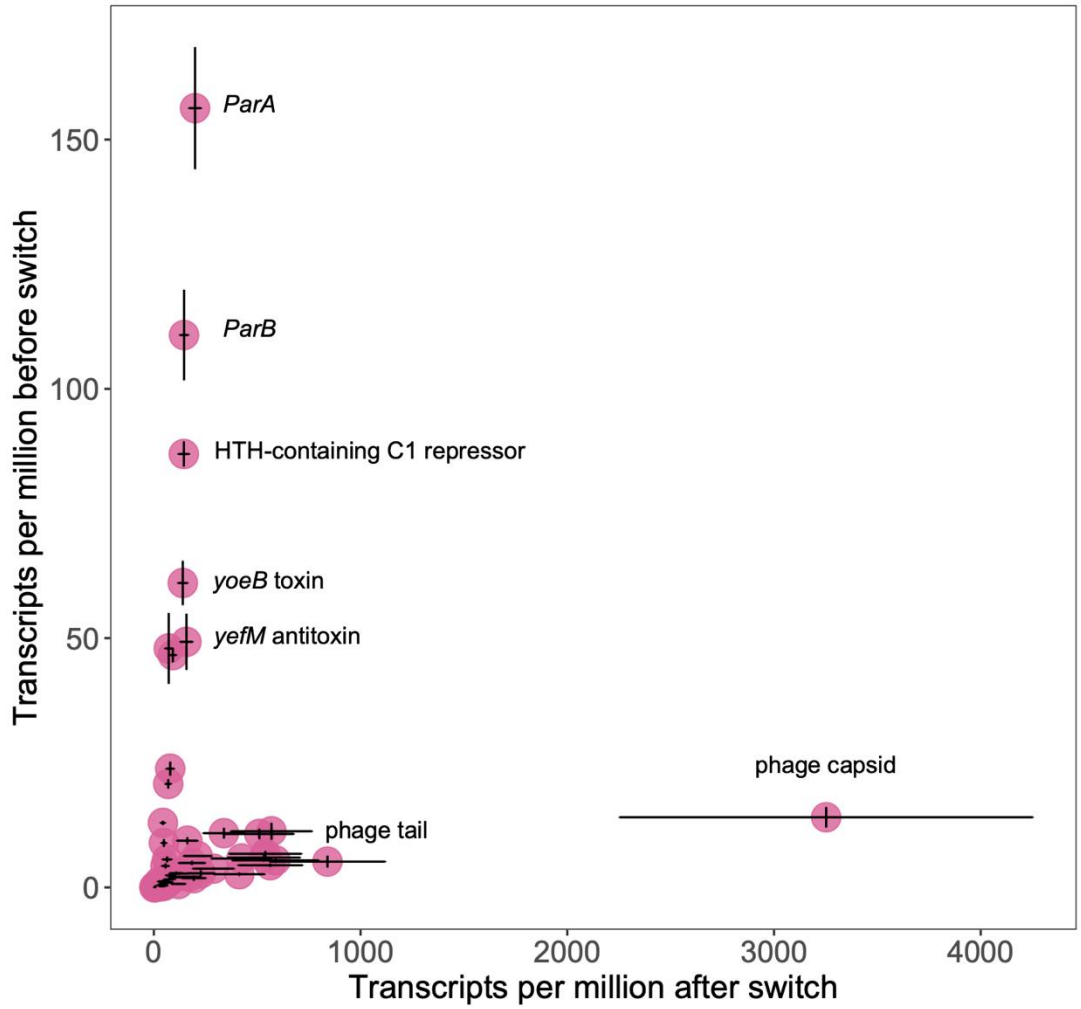


Figure C.7 Schematic illustration of segregational drift. Here we showed a simplest scenario where two copies of phage-plasmids were randomly distributed into daughter cells with equal opportunity, while the copy number of phage-plasmids per cell remained constant. Following a Binomial distribution, we have 1/3 probability of generating two homozygotic descendants and 2/3 probability of generating two heterozygotic descendants.

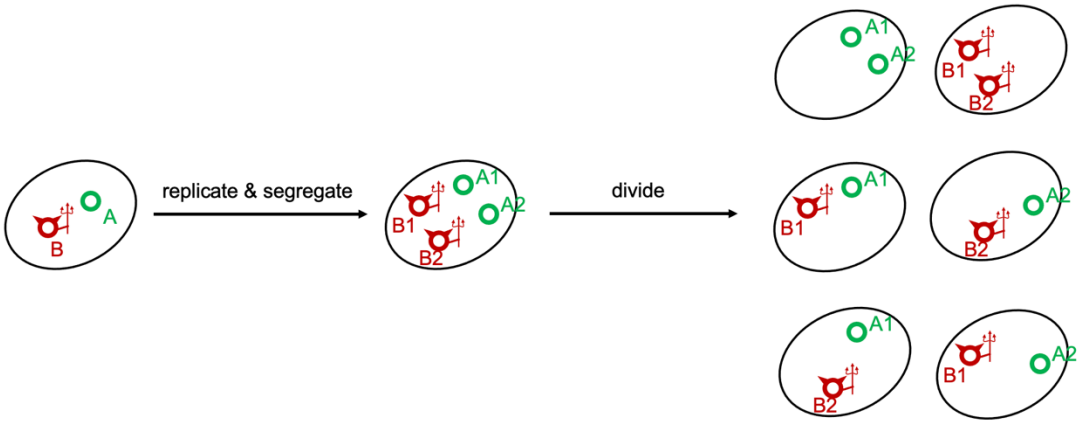


Figure C.8 Copy number of phage-plasmid per host cell. Post segregational drift descendant populations carrying only wild-type phage-plasmid harbored a significantly elevated copy number of the phage-plasmid compared to the original host population before mutation-driven phage-plasmid induction. Notably, the average copy number in post segregational drift descendant populations was not an integer but between one and two, suggesting that some individuals may lose some copies of phage-plasmids during cell division. Source data are provided in the Source Data file.

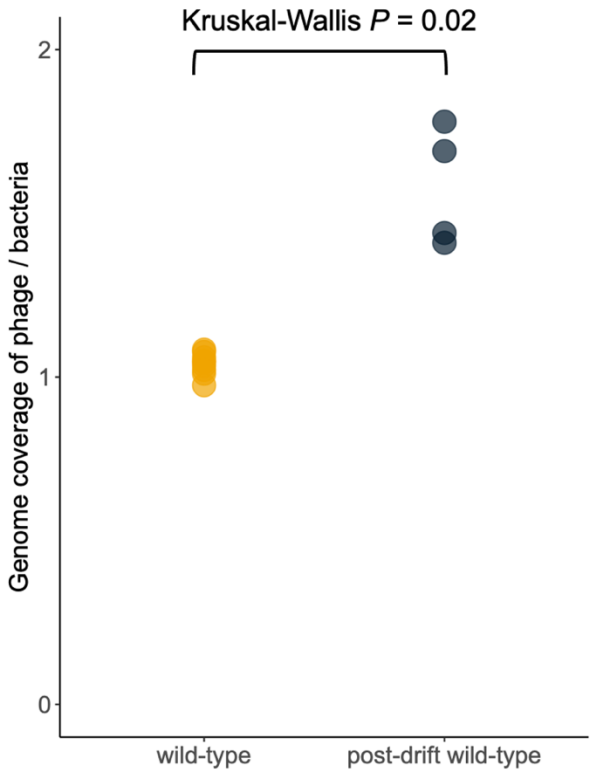


Figure C.9 Post segregational drift descendant populations carrying only wild-type phage-plasmid but with a higher copy number became more resistant to mutation-driven phage-plasmid induction. Compared to the timing of 30~50 generations for populations carrying single copy of phage-plasmid, these post segregational drift populations were resistant to phage induction for 66 generations or longer under the same experimental serial dilution regime. Each color represents an independent line of post-segregational drift population under serial dilution scheme. Source data are provided in the Source Data file.

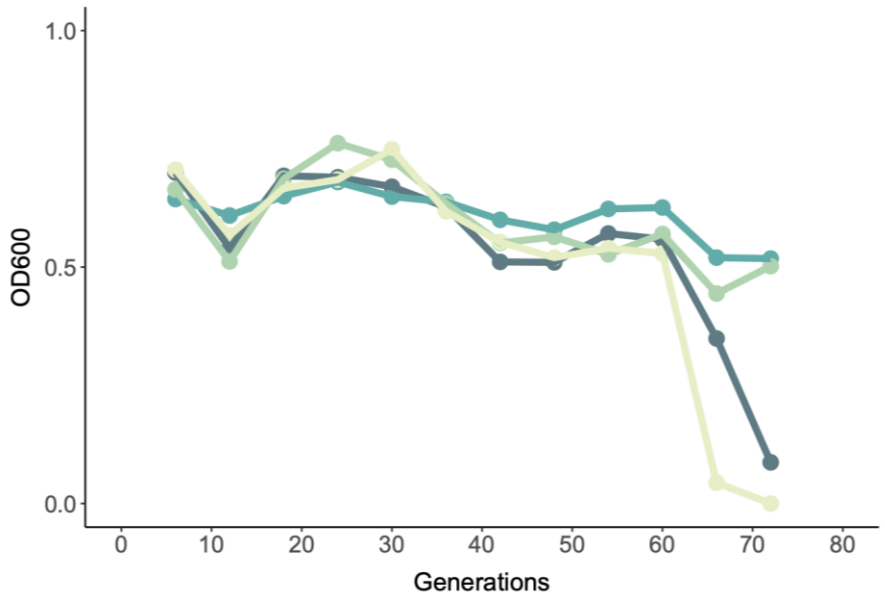


Figure C.10 The relationship between reinfection efficiency and saturating genotypic frequency in model simulation. In our model, reinfection efficiency (R) indicates the average number of released mutated phages that successfully re-infect a host cell per host cell lysed. The resulting saturating frequency of the mutated genotype increases when reinfection efficiency increases, but the increase decelerates indicating the mutated genotype never reaches fixation. We found that $R = 5$ best fitted the experimentally observation of the saturating genotypic frequency of roughly 0.5. Source data are provided in the Source Data file.

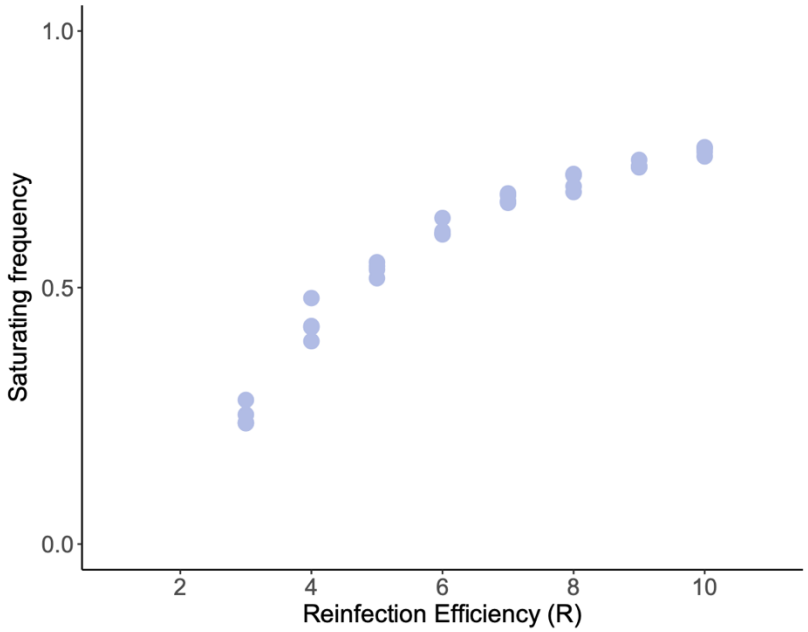


Figure C.11 A phage-plasmid in *Phaeobacter gallaeciensis* DSM26640 and *Phaeobacter piscinae* P18 also shares a highly homologous plasmid backbone but with very different phage structural genes. These two strains are isolated in geographically close locations but are phylogenetically distinct (different species, average nucleotide identity = 92%). Source data are provided in the Source Data file.

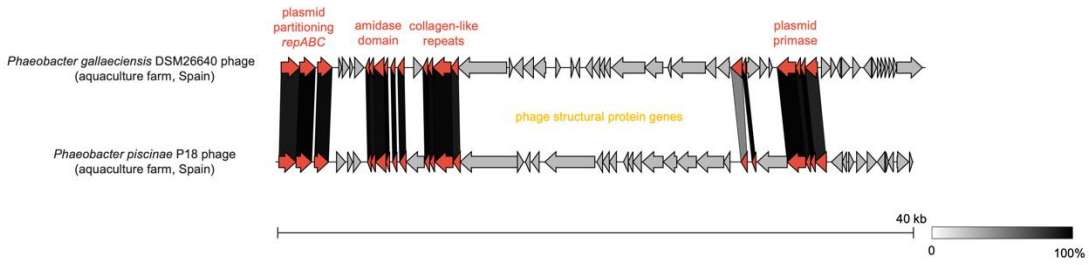


Figure C.12 Geographic distribution of phage repressors that are highly significant hits ($P < 10^{-15}$) of *Tritonibacter mobilis* A3R06 phage-plasmid C1-type repressor gene, identified from currently available environmental metagenomic datasets. Background world map is from R package ggmap.