INFERRING ECOLOGICAL INTERACTIONS FROM DYNAMICS IN PHAGE-BACTERIA COMMUNITIES

A Dissertation
Presented to
The Academic Faculty

By

Ashley Coenen

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Physics

Georgia Institute of Technology

December 2021

INFERRING ECOLOGICAL INTERACTIONS FROM DYNAMICS IN PHAGE-BACTERIA COMMUNITIES

Thesis committee:

Dr. Joshua Weitz, Advisor
School of Biological Sciences and School of Physics
Georgia Institute of Technology

Dr. Justin Romberg
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Simon Sponberg
School of Physics
Georgia Institute of Technology

Dr. Flavio Fenton
School of Physics
Georgia Institute of Technology

Dr. Matthew Sullivan
Department of Microbiology
Ohio State University

Date approved: November 18, 2021

ACKNOWLEDGMENTS

I am immensely grateful to friends, roommates, partners, colleagues, and mentors for their companionship throughout these strange times. To Devika Singh and Amen Ben Hamida: my anchors. To Daniel Holmes: for impeccable music, food, and game curation. To Hallie Andrews: for being the best roommate I could have asked for and least of all for saving my life. To Karl Lundquist: for keeping in touch and for growing with me. Thank you for your endless support during, as well as before and beyond, the pandemic. And thank you to friends both new and old – you know who you are. I would not have finished this endeavor without you all.

To my dear friends: Luke Rivard, Annica Mandeltort, Tom Sasani, and Micah Price. To the Lawrence University physics crew. To my family: Terry and Chris Coenen, Zach and Kelsey and Nola Coenen, and Susan Massey Nickel. To my grandma, Marlene Coenen, for chatting about the weather and for lighting a candle for me. Thank you for always being there for me, without question or hesitation.

Throughout my time in graduate school, I have had wonderful connections with too many fellow graduate students to list: Rozenn Pineau, Roth Conrad, Juan Pablo Barraza, Yael Toporek, Keith Paarporn, Logan Kageorge, Marcus Daum, Matthew Johnson, Félix Thouin, Shane Jacobeen, Anna Miettinen, and John Hendry. To my roommate and partner in crime Keshav Joshi, who is dearly, dearly missed. These years have been brighter and happier for having had you all in my life.

To Stephen Beckett of the Weitz group, for being my mentor these past six years. To current Weitz group members, who have taught, debated with, and learned with me: David Demory, Daniel Muratore, Rogelio Rodriguez-Gonzalez, Andreea Magalie, Marian Dominguez-Mirazo, Conan Zhao, Shashwat Shivam, Jeremy Harris, Adriana Lucia-Sanz, Jacopo Marchi, Rachel Calder, Audra Davidson, and Gabi Steinbach. To my mentees, Ellen and Anup, who taught and learned with me too. To past Weitz group members who

have provided me with guidance, both scientific and professional: Jessica Irons, Ceyhun Eksin, Joey Leung, Chad Wigington, Yu-Hui Lin, and César Flores.

Throughout this thesis work, I have had incredible support from many colleagues and collaborators. To my co-authors: Natalie Solonenko, Christine Sun, Daniel Muratore, Sarah Hu, Elaine Luo, and many others who supported theory, coding, and experimental work, particularly on Chapter 3. Thank you also to those who provided code review and feedback throughout the manuscript writing process. On Chapter 2: Ben Buldoc, Stephen Beckett, Yu-Hui Lin, David Demory, and five anonymous reviewers. On Chapter 4: the entire BioOceans team, including Steven Wilhelm, Gary LeCleir, Naomi Gilbert, and Debbie Lindell, as well as members from the Weitz Group, and Jessica Irons for administrative support. Many thanks to the Simons Foundation and the National Science Foundation for funding.

To my proposal committee: Justin Romberg and Simon Sponberg, for being there from the very beginning. To Matt Sullivan, for welcoming me to the Ohio State, and to Flavio Fenton for joining the thesis committee later and with just as much impact. To faculty members who have guided me both in academia and beyond: Jennifer Curtis and David Ballantyne.

And, finally, to my advisor: Joshua Weitz. Thank you for your mentorship, your encouragement, and your understanding – it was not always an easy journey for me, and I can point to many times where I would have been lost without your support. Thank you, also, for connecting me with collaborations beyond the standard academic purview, and for giving me room to grow as both a researcher and as a citizen. I look forward keeping in touch and meeting your future graduate students, wherever our paths align.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Bacteria and the viruses that infect them are ubiquitous, abundant, and highly diverse. Characterizing how viruses interact with their microbial hosts is critical to understanding microbial community structure and function, as well as downstream effects on the surrounding environment. However, existing methods for quantifying bacteria-phage interactions are not widely applicable to natural communities. First, many bacteria are not culturable, preventing direct experimental testing. Second, "-omics" based methods, while high in accuracy and specificity, have been shown to be extremely low in power. Third, inference methods based on time-series or co-occurrence data, while promising, have for the most part not been rigorously tested. This thesis work focuses on this final category of quantification strategies: inference methods.

In this thesis, we further our understanding of both the potential and limitations of several inference methods, focusing primarily on time-series data with high time resolution. We emphasize the quantification of efficacy by using time-series data from multi-strain bacteria-phage communities with known infection networks. We employ both *in silico* simulated bacteria-phage communities as well as an *in vitro* community experiment. We review existing correlation-based inference methods, extend theory and characterize trade-offs for model-based inference which uses convex optimization, characterize pairwise interactions in a 5x5 virus-microbe community experiment using Markov chain Monte Carlo, and present analytic tools for microbiome time-series analysis when a dynamical model is unknown. In doing so, we provide evidence in favor of model-based inference in recovering phage-bacteria infection networks with high accuracy and specificity. Together, these chapters bridge gaps in existing literature, as well as identify future research directions, in inference of ecological interactions from time-series data.

CHAPTER 1
INTRODUCTION

Viruses are ubiquitous, abundant, and highly diverse in marine, soil, and human-associated environments. Viruses can infect all domains of life, although most viruses infect bacteria and archaea [1]. Viruses of microbes (bacteriophages, archaeal viruses, and some eukaryotic viruses) are known to play important roles in microbial communities. In ocean ecosystems for example, viruses control microbial population sizes through infection and cell lysis. As a result, the flow of organic matter may be redirected from higher trophic levels back to the microbial loop [2, 3]. Viruses can also alter host physiology and metabolism. For example, viruses infecting the marine bacteria *Prochlorococcus* and *Synechococcus* carry genes that augment host photosynthetic capacity [4, 5]. Both of these examples point to how viruses can ultimately affect ecosystem-level processes. Identifying and characterizing how viruses interact with their microbial hosts *microscopically* (i.e. individually and locally) is critical to understanding microbial community structure and function within an ecosystem [2, 6, 7, 8].

Viruses interact with microbial hosts primarily through two modes of infection: lysis and lysogeny. In a lytic infection, viral genetic material is injected into a microbial host, hijacking the host cell machinery to produce new viral particles inside of the host cell. After some delay, the host cell is lysed (and killed) and the new viral particles are released into the environment. In a lysogenic infection, viral genetic material is injected into a microbial host and integrated into the host genome. The virus remains dormant and is transmitted vertically as the microbial host replicates. Typically an environmental cue, such as UV radiation, induces the virus into the lytic infection mode. Some viruses are exclusively lytic. Other viruses can "choose" either the lytic or lysogenic mode upon infection, although the mechanisms underlying this "decision" are understood for only a few model systems [9].

In addition to immediate infection outcomes, virus-microbe interactions affect host cells in other ways. Viral infection can result in e.g. altered host metabolism, host pathogenicity or toxin production, and horizontal gene transfer among hosts [4, 5, 10, 11]. In this thesis, we are primarily concerned with lytic infections.

A detailed understanding of infection mechanism is lacking for most viruses and microbes in nature [9]. The majority of viruses and microbes on the planet are not culturable and therefore cannot be studied directly in a laboratory setting [10]. For natural communities, virus-microbe interactions must be probed in other ways. There are several partially culture-independent methods (e.g. viral tagging [12, 13] and digital PCR [14]) and single-cell methods (e.g. single-cell amplified genome analysis [15, 16, 17]) that are useful but ultimately limited in scope and scalability. On the other hand, recent advances in metagenomic sequencing and analyses allow for identification and (partial) quantification of viruses and microbes *in situ* in a direct and high-throughput manner [18, 19, 20]. Viral and bacterial sequences from assembled metagenomes can be analyzed directly and putatively linked on a genetic basis [21, 22]. Alternatively, metagenomic sampling of a community *over time* provides estimates of changing abundances of viral and microbial populations. From these time-series, a variety of statistical and mathematical methods can be used to infer virus-microbe interactions [23, 24, 25, 26, 27].

This thesis focuses on the inference of ecological interactions from observed community dynamics. The systems of interest are virus-microbe communities, in which *in situ* population dynamics can be observed via metagenomic sampling, although much of the theoretical work here is generalizable to ecological systems broadly. In microbial and viral ecology, identifying virus-microbe interactions and characterizing their ecological mode (e.g. lysis or lysogeny) is an active area of research [28, 29, 30]. Currently, there does not exist a "gold standard" virus-microbe community in which interactions are known, which makes truth-testing of any given method difficult. In this thesis, we assess the effectiveness of several inference methods using a combination of *in silico* simulated communities and

in vitro experiments of relatively small communities with well-characterized life history traits.

In Chapter 2, we review several correlation-based inference methods for predicting virus-microbe interactions, which are widely used in existing literature [31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 24, 42]. Using *in silico* virus-microbe communities, we show that correlation and correlation-based methods are poor predictors of true interactions. The work in Chapter 2 is published [43].

In Chapter 3, we introduce model-based inference methods as an alternative to correlation-based methods. We extend an existing technique [44], based on linearized differential equations of virus-microbe ecological dynamics, and show that both virus-microbe and microbe-microbe interactions can be accurately inferred in more complex settings than originally proposed. Chapter 3 also uses *in silico* simulations, allowing us to rapidly characterize tradeoffs in inference efficacy with experimental design. The work in Chapter 3 is currently in prep for submission.

In Chapter 4, we integrate models and theory from previous chapters with an *in vitro* 5x5 virus-bacteria community experiment. The life history traits of all 5 bacteria and 5 phage strains, as well as 25 potential virus-bacteria interactions, were characterized experimentally, providing a "gold standard" against which to compare inferred interactions. The work in Chapter 4 is in collaboration with an experimental lab at the Ohio State University and is currently in prep for submission.

In Chapter 5, we present a suite of methods for analyzing microbiome time-series when a dynamical model is not explicitly known, with example demonstrations on published datasets. We use regularized regression to predict interactions, emphasizing that the underlying assumptions of normality and independence among samples hold only in limited contexts. We also show how clustering and periodicity analyses can be used to investigate patterns in dynamics. The content in Chapter 5 is published as a multi co-first author paper, with publicly available interactive tutorials in R and MATLAB [45].

Together, these chapters bridge gaps in existing literature concerning inference methods for ecological systems using time-series data, in particular for virus-microbe communities. Inference methods which leverage high-throughput time-series data, like metagenomes and viromes, will be integral for characterizing ecological interactions in natural environments across large spatial and temporal scales [46]. Yet, the efficacy of such methods has not been well characterized. This thesis characterizes efficacy for several inference methods in idealized *in silico* communities as well as in a relatively small *in vitro* community experiment. In using idealized *in silico* and simple *in vitro* communities, we identify contexts in which inference methods are likely to succeed, fail, or be inconclusive or unreliable in more complex, natural environments. Of course, many open questions remain on the viability of applying these inference methods to communities in natural environments, such as the effects of compositional "-omics" data and complex environmental effects like seasonality or micro-scale physics of particles. These open questions are discussed in more detail in Chapter 6.

CHAPTER 2

THE LIMITATIONS OF CORRELATION-BASED INFERENCE IN COMPLEX VIRUS-MICROBE COMMUNITIES

Adapted from A R Coenen and J S Weitz, The limitations of correlation-based inference in complex virus-microbe communities, mSystems, (2018) [43].

2.1 Abstract

Microbes are present in high abundances in the environment and in human-associated microbiomes, often exceeding one million per milliliter. Viruses of microbes are present in even higher abundances and are important in shaping microbial populations, communities, and ecosystems. Given the relative specificity of viral infection, it is essential to identify the functional linkages between viruses and their microbial hosts, particularly given dynamic changes in virus and host abundances. Multiple approaches have been proposed to infer infection networks from time-series of *in situ* communities, among which correlation-based approaches have emerged as the *de facto* standard. In this work, we evaluate the accuracy of correlation-based inference methods using an *in silico* approach. In doing so, we compare predicted networks to actual networks to assess the self-consistency of correlation-based inference. At odds with assumptions underlying its widespread use, we find that correlation is a poor predictor of interactions in the context of viral infection and lysis of microbial hosts. The failure to predict interactions holds for methods which leverage product-moment, time-lagged, and relative-abundance based correlations. In closing, we discuss alternative inference methods, particularly model-based methods, as a means to infer interactions in complex microbial communities with viruses.

2.2 Importance

Inferring interactions from population time-series is an active and ongoing area of research. It is relevant across many biological systems – in particular in virus-microbe communities, but also in gene regulatory networks, neural networks, and ecological communities broadly. Correlation-based inference – using correlations to predict interactions – is widespread. However, it is well known that "correlation does not imply causation". Despite this, many studies apply correlation-based inference methods to experimental time-series without first assessing the potential scope for accurate inference. Here, we find that several correlation-based inference methods fail to recover interactions within *in silico* virus-microbe communities, raising questions on their relevance when applied *in situ*.

2.3 Introduction

Viruses of microbes are ubiquitous and highly diverse in marine, soil, and human-associated environments. Viruses interact with their microbial hosts in many ways. For example, they can transfer genes between microbial hosts [10, 11], alter host physiology and metabolism [4, 5], and redirect the flow of organic matter in food webs through cell lysis [2, 3]. Viruses are important parts of microbial communities, and characterizing the interactions between viruses and their microbial hosts is critical for understanding microbial community structure and ecosystem function [2, 6, 7, 8].

A key step in characterizing virus-microbe interactions is determining which viruses can infect which microbes. Viruses are known to be relatively specific but not exclusive in their microbial host range. Individual viruses may infect multiple strains of an isolated microbe or they may infect across genera as part of complex virus-microbe interaction networks [47, 48]. For example, cyanophage can infect both *Prochlorococcus* and *Synechococcus* which are two distinct genera of marine cyanobacteria [49]. However, knowledge of viral host range remains limited because existing experimental methods for di-

rectly testing for viral infection are generally not applicable to an entire *in situ* community. Culture-based methods such as plaque assays are useful for checking for viral infection at the strain level and permit high confidence in their results, but they are not broadly applicable as many viruses and microbes are difficult or currently impossible to isolate and culture [10]. Partially culture-independent methods, such as viral tagging [12, 13] and digital PCR [14], overcome some of these hurdles but only for particular targetable viruses and microbes. Similarly, single-cell genome analysis is able to link individual viruses to microbial hosts [15, 16, 17] but for a relatively small number of cells.

Viral metagenomics offers an alternate route for probing virus-microbe interactions for entire *in situ* communities, bypassing culturing altogether [18, 19, 20]. The viral sequences obtained from metagenomes can be analyzed directly using bioinformatics-based methods to predict microbial hosts [21, 22] although such methods may only be appropriate for a subset of viruses (phages and archaeal viruses but not eukaryotic viruses) and putative hosts (prokaryotes but not eukaryotes). Alternatively, metagenomic sampling of a community *over time* can provide estimates of the changing abundances of viral and microbial populations at high time- and taxonomic- resolution. Once these high-resolution time-series are obtained, they can be used to predict virus-microbe interactions using a variety of statistical and mathematical inference methods (see reviews [23, 24, 25, 26, 27]).

Correlation and correlation-based methods are among the most widely used network inference methods for microbial communities [24]. For example, Extended Local Similarity Analysis (eLSA) is a correlation-based method which allows for both local and time-lagged correlations [31, 32, 33] and has been used to infer interaction networks in communities of marine bacteria [34, 35]; bacteria and phytoplankton [36, 37]; bacteria and viruses [38]; and bacteria, viruses, and protists [39, 40]. In addition, several correlation-based methods have been developed to address challenges associated with the compositional nature of "-omics" datasets [41, 24], including Sparse Correlations for Compositional data (SparCC) [42].

Regardless of the particular details of these methods, all correlation-based inference operates on the same core assumption: that interacting populations trend together (are correlated) and that non-interacting populations do not trend together (are not correlated). Particular correlation-based methods may relax or augment this assumption. For example, with eLSA the trends may be time-lagged [31, 32, 33]; with simple rank correlations the trends may be non-parametric; and with compositional methods like SparCC the trends may occur between ratios of relative abundances [42]. In communities with only a few populations and simple interactions, population trends may indeed be indicative of ecological mechanism. In these contexts, some correlation-based methods have been shown to recapitulate microbe-microbe interactions with limited success [24]. Typically however the challenge of inferring interaction networks applies to diverse communities and complex ecological interactions. Microbial communities often have dozens, hundreds, or more distinct populations, each of which may interact with many other populations through non-linear mechanisms such as viral lysis, as well as be influenced by fluctuating abiotic drivers. In these contexts, the relationship between correlation and ecological mechanism is poorly understood. Often correlations do not have a simple mechanistic interpretation, a well-known adage ("correlation does not imply causation") that is often disregarded.

Despite the challenge of interpretation, correlation-based inference methods are widely used with *in situ* datasets [31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 24, 42]. Benchmarking inferred networks – connecting correlations to specific ecological mechanisms – is difficult. In the context of lytic infections of environmental microbes by viruses, there is (usually) no existing "gold standard" interaction network for which to validate inferred interactions. Therefore, in this work, we take an *in silico* approach to assess the accuracy of correlation-based inference. To do this, we simulate virus-microbe community dynamics with an interaction network which is prescribed *a priori* and use it to benchmark inferred networks. Several existing studies have applied similar *in silico* approaches in the case of both microbe-microbe and microbe-virus interactions and found that simple Pearson cor-

relation [50, 41] and several correlation-based methods [24] either fail or are inconsistent in recapitulating interaction networks. Here, we provide an in-depth assessment of the potential for correlation-based inference in diverse communities of microbes and viruses. As we show, correlation-based inference fails to recapitulate virus-microbe interactions and performs worse in more diverse communities. The failure of correlation-based inference in this context raises concerns over its use in inferring microbe-parasite interactions as well as microbe-predator and microbe-microbe interactions more broadly.

2.4 Methods

2.4.1 Dynamical model of a virus-microbe community

We model the ecological dynamics of a virus-microbe community with a system of nonlinear differential equations:

$$\dot{H}_i = \overbrace{r_i H_i \left(1 - \frac{\sum_{i'}^{N_H} a_{ii'} H_{i'}}{K}\right)}^{\text{microbial growth and competition}} - \overbrace{H_i \sum_{j}^{N_V} M_{ij}\phi_{ij} V_j}^{\text{death by viral lysis}} \tag{2.1}$$

$$\dot{V}_j = \underbrace{V_j \sum_{i}^{N_H} M_{ij}\phi_{ij}\beta_{ij} H_i}_{\text{release of virions}} - \underbrace{m_j V_j}_{\text{viral decay}} \tag{2.2}$$

where H_i and V_j refer to the population density of microbial host i and virus j respectively. There are N_H different microbial host populations and N_V different virus populations. For our purposes, a "population" is a group of microbes or viruses with identical life history traits, that is microbes or viruses which occupy the same functional niche.

In the absence of viruses, the microbial hosts undergo logistic growth with growth rates r_i. The microbial hosts have a community-wide carrying capacity K, and they compete with each other for resources both inter- and intra-specifically with competition strength $a_{ii'}$. Each microbial host can be infected and lysed by a subset of viruses determined by the

interaction terms M_{ij}. If microbial host i can be infected by virus j, $M_{ij} = 1$; otherwise $M_{ij} = 0$. The collection of all the interaction terms is the interaction network represented by matrix $\mathbf{M}$ of size N_H by N_V. The adsorption rates ϕ_{ij} denote how frequently microbial host i is infected by virus j.

Each virus j's population grows from infecting and lysing their hosts. The rate of virus j's growth is determined by its host-specific adsorption rate ϕ_{ij} and host-specific burst size β_{ij}, which is the number of new virions per infected host cell. The quantity $\tilde{M}_{ij} = M_{ij}\phi_{ij}\beta_{ij}$ is the effective interaction strength between virus j and host i, and the collection of all the interaction strengths is the weighted interaction network $\tilde{\mathbf{M}}$. Finally, the viruses decay at rates m_j.

2.4.2 Generating interaction networks and characterizing network structure

Virus-microbe interaction networks, denoted $\mathbf{M}$, are represented as bipartite networks or matrices of size N_H by N_V where N_H is the number of microbial host populations and N_V is the number of virus populations. The element M_{ij} is 1 if microbe population i and virus population j interact and 0 otherwise. In this chapter, we consider only square networks ($N = N_H = N_V$) although the analysis is easily extended to rectangular networks. We consider three network sizes $N = 10, 25, 50$.

For each network size N, we generate an ensemble of networks varying in nestedness and modularity (Figure 2.1). We first generate the maximally nested (Figure 2.1A) and maximally modular (Figure 2.1B) networks of size N using the BiMat Matlab package [51]. In order to achieve maximal nestedness and modularity, the network fill F (fraction of interacting pairs) is fixed at $F = 0.55$ for the nested networks and $F = 0.5$ for the modular networks. For the modular networks, the number of modules is set to 2, 5, and 10 for the three network sizes respectively.

To generate networks that vary in nestedness and modularity, we perform the following "rewiring" procedure. Beginning with the maximally nested or maximally modular

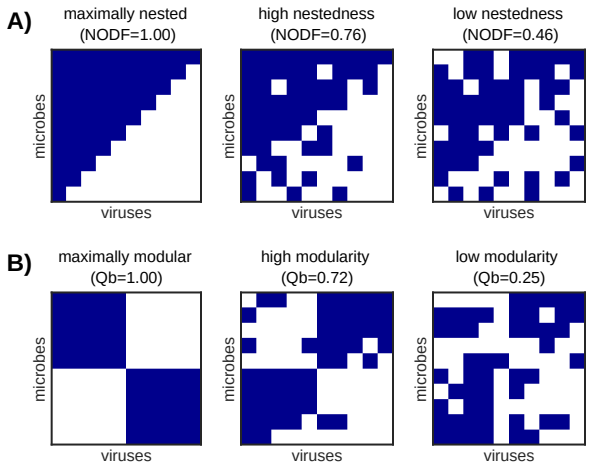

Figure 2.1: Example interaction networks characterized by A) nestedness and B) modularity. The networks shown here have size $N = 10$ and fill A) $F = 0.55$ and B) $F = 0.5$. Within each network, rows represent microbe populations and columns represent virus populations, while navy squares indicate interaction ($M_{ij} = 1$). Networks were generated according to subsection 2.4.2. Nestedness (NODF) and modularity (Q_b) were measured with the BiMat package and are arranged in their most nested or most modular forms [51].

network, we randomly select an interacting virus-microbe pair ($M_{ij} = 1$) and a non-interacting virus-microbe pair ($M_{i'j'} = 0$) and exchange their values. We do not allow exchanges that would result in an all-zero row or column, as that would isolate the microbe or virus population from the rest of the community. We continue the random selection of pairs without replacement until the desired nestedness or modularity has been achieved. To calculate nestedness and modularity, we use the default algorithms in the BiMat Matlab package. The nestedness metric used is NODF [52], and the algorithm used to calculate modularity is AdaptiveBRIM [53]. The modularity is additionally normalized according to a maximum theoretical modularity as detailed in [54].

2.4.3 Choosing life history traits for coexistence

The life history traits for a given interaction network are chosen to ensure that all microbial host and virus populations can coexist, adapted from [55].

First we sample target fixed point densities H_i^* and V_j^* for each microbial host and

virus population. In addition we sample adsorption rates ϕ_{ij} and burst sizes β_{ij}. All of these parameters are independently and randomly sampled from uniform distributions with biologically feasible ranges specified in Table 2.1. We use a fixed carrying capacity density $K = 10^6$ cells/mL for all parameter sets.

Table 2.1: Sampling ranges for parameters in the virus-microbe dynamical model (Equation 2.1 and Equation 2.2).

	parameter	sampling range	units
H_i^*	host i target steady-state density	$10^3 - 10^4$	cells/mL
V_j^*	virus j target steady-state density	$10^6 - 10^7$	virions/mL
K	community-wide host carrying capacity	10^6	cells/mL
ϕ_{ij}	adsorption rate of virus j into host i	$10^{-7} - 10^{-6}$	mL/ (virion · day)
β_{ij}	burst size of virus j per host i	$10 - 100$	virions/cell
H_i^{0*}	host i target steady-state density in the absence of viruses	$10^3 - 10^6$	cells/mL
$a_{ii'}$	competitive effect of host i' on host i	$0 - 1$	

Next we sample microbe-microbe competition terms $a_{ii'}$. We introduce an additional constraint that microbial populations should coexist in the absence of all viruses. To this end, we sample target virus-free fixed point densities H_i^{0*} from a uniform distribution with a range specified in Table 2.1. After sampling, the $H_i^{0^*}$ remain fixed. According to Equation 2.1, coexistence in the virus-free setting is satisfied when

$$K = \sum_{i'}^{N_H} a_{ii'} H_{i'}^{0*} \tag{2.3}$$

for each microbial host i. To start, we set all intraspecific competition to one ($a_{ii} = 1$) and all interspecific competition to zero ($a_{ii'} = 0$ for $i' \neq i$). Then for each microbial host i we randomly choose an index $k \neq i$ and sample a_{ik} uniformly between zero and one. If the updated sum in Equation 2.3 does not exceed the carrying capacity K, we repeat for a new index k. Once the carrying capacity is exceeded, we adjust the most recent a_{ik} so that Equation 2.3 is satisfied exactly.

Finally, the viral decay rates m_j and host growth rates r_i are computed from the fixed

point versions of Equation 2.1 and Equation 2.2:

$$m_j = \sum_{i}^{N_H} M_{ij}\phi_{ij}\beta_{ij}H_i^* \tag{2.4}$$

$$r_i = \left(\sum_{j}^{N_V} M_{ij}\phi_{ij}V_j^*\right) / \left(1 - \frac{\sum_{i'}^{N_H} a_{ii'}H_{i'}^*}{K}\right) \tag{2.5}$$

2.4.4 Simulating and sampling time-series

We use Matlab's native `ODE45` function to numerically simulate the virus-microbe dynamical model specified in subsection 2.4.1 with interaction network and life history traits generated as described subsection 2.4.2 and subsection 2.4.3. We use a relative error tolerance of 10^{-8}. Initial conditions are chosen by perturbing the fixed point densities H_i^* and V_j^* by a multiplicative factor δ where the sign of δ is chosen randomly for each microbial host and virus population. We note that δ can be used to tune the amount of variability in the simulated time-series (see Figure A.1).

After simulating virus and microbe time-series, we sample the time-series at regularly spaced sample times (every 2 hours) for a fixed duration (200 hours, or 100 samples). Therefore, for each virus and each microbe in the community we take S samples at times $t_1, \ldots, t_S$. We use the same sampling frequency and the same S for each inference method, except for time-delayed correlation (see subsection 2.4.5).

2.4.5 Standard and time-delayed Pearson correlation networks

We assume S regularly spaced sample times $t_1, \ldots, t_S$ for each host type H_i and each virus type V_j. The samples are log-transformed, that is $h_i(t_k) = \log_{10} H_i(t_k)$ and $v_j(t_k) = \log_{10} V_j(t_k)$ for each sampled time-point t_k. The standard Pearson correlation coefficient

between host i and virus j is then

$$r_{ij} = \frac{\sum_{k=1}^{S} \left(h_i(t_k) - \bar{h}_i\right)\left(v_j(t_k) - \bar{v}_j\right)}{\sqrt{\sum_{k=1}^{S} \left(h_i(t_k) - \bar{h}_i\right)^2}\sqrt{\sum_{k=1}^{S} \left(v_j(t_k) - \bar{v}_j\right)^2}} \tag{2.6}$$

where $\bar{h}_i = \frac{1}{S}\sum_{k=1}^{S} h_i(t_k)$ and $\bar{v}_j = \frac{1}{S}\sum_{k=1}^{S} v_j(t_k)$ are the sample means. The correlation coefficients for all virus-host pairs are represented as a bipartite matrix $\mathbf{R}$ of size $N_H \times N_V$ analogous to the interaction network (see subsection 2.4.2).

Time-delayed correlations are computed by sampling the virus time-series later in time. Each virus-host pair may have a unique time-delay τ_{ij}. For example, if host i is sampled at times $t_1, \ldots, t_S$ then virus j is sampled at times $t_1 + \tau_{ij}, \ldots, t_S + \tau_{ij}$. We keep the number of samples S fixed, and consequently allow virus j to be sampled beyond the final sample time t_S of the hosts. The time-delayed Pearson correlation coefficient is

$$r_{ij}^{\tau} = \frac{\sum_{k=1}^{S} \left(h_i(t_k) - \bar{h}_i\right)\left(v_j(t_k + \tau_{ij}) - \bar{v}_j^{\tau_{ij}}\right)}{\sqrt{\sum_{k=1}^{S} \left(h_i(t_k) - \bar{h}_i\right)^2}\sqrt{\sum_{k=1}^{S} \left(v_j(t_k + \tau_{ij}) - \bar{v}_j^{\tau_{ij}}\right)^2}} \tag{2.7}$$

where $\bar{v}_j^{\tau_{ij}} = \frac{1}{S}\sum_{k=1}^{S} v_j(t_k + \tau_{ij})$ is the mean of the time-delayed virus sample. As before, the correlation coefficients for all virus-host pairs is a bipartite matrix $\mathbf{R}^{\tau}$ of size $N_H \times N_V$.

Pearson correlation coefficients, as specified above, were computed using Matlab's native `corr` function with type="pearson". Alternate correlation types including Spearman and Kendall are also supported by the `corr` function and are utilized in the appendix.

2.4.6 eLSA networks

Extended Local Similarity Analysis (eLSA) is a correlation-based inference method which is widely used with *in situ* time-series of complex microbial communities (*e.g.* [34, 35, 36, 37, 38, 39, 40]). eLSA attempts to detect local correlations, that is, time-series which trend together for only a portion of the sample period. In addition, eLSA allows for time-delayed correlations (as described in the previous section subsection 2.4.5). To this end, a "local

similarity" (LS) score is computed for each pair of time-series. The LS score is analogous to computing the Pearson correlation for all possible subsections of the two time-series, with offsets up to a pre-decided length, and keeping the maximum absolute correlation. As an example, two time-series may trend strongly during the first half of the sample period but not during the second. For such a pair of time-series, the Pearson correlation would be low, but the LS score would be high.

To compute the LS score, the two time-series are first transformed to have normal distributions (we note that such a transformation is non-stationary and thus may induce spurious correlations). The LS score is the maximal sum of the product of the entries across all possible subsections, normalized by the time-series length. If a pre-defined delay is specified, the subsections are additionally offset from one another from zero up to to the delay amount [31, 32, 33].

We applied eLSA to our simulated time-series data. We used samples of all N_H host types and all N_V virus types with S regularly spaced sample times $t_1, \ldots, t_S$ as input. We used the `lsa-compute.py` Python script and set parameters to specify the number of sampled points (spotNum=S), number of replicates (repNum=1), number of bootstraps (b=0), and number of permutations (x=1). All other parameters were left with their default settings including the maximum allowed time delay (delayLimit=3). The `lsa-compute.py` script computes eLSA scores between all virus-host, host-host, and virus-virus pairs. We selected only the virus-host eLSA scores and arranged them in a bipartite matrix of size $N_H \times N_V$ analogous to the interaction network (see subsection 2.4.2). We used a custom Matlab script `write_elsa.m` to generate .csv data files in the format specified by the eLSA documentation. We used a custom bash script `elsa_compute_all.sh` to run the eLSA analysis on the ensemble of virus-microbe communities. Finally, we used a custom Matlab script `read_elsa.m` to import the results into Matlab for scoring (see subsection 2.4.8).

2.4.7 SparCC networks

Sparse Correlations for Compositional data (SparCC) is a correlation-based inference method for use with compositional time-series data. This is relevant for "-omics" data in which abundances are typically relative. It is well known that compositional data pose challenges for standard statistics, including Pearson and other types of correlation. Because the data sum to one, individual time-series are not independent. This biases correlations to be negative regardless the trend between the underlying absolute abundances. SparCC estimates the Pearson correlation between two time-series while taking into account these compositional dependencies. In particular, SparCC computes the variance of the log-transformed ratio of two time-series, and compares this quantity to the variances of the individual log-transformed time-series. SparCC assumes sparsity in the correlation matrix but is robust to violations of this assumption [42].

We applied SparCC to our simulated time-series data as a means to evaluate correlation-based inference in a scenario in which underlying viral and microbial densities can be measured only relatively. Given samples at S regularly spaced sample times $t_1, \ldots, t_S$, we first normalized the N_H host types and N_V virus types at each sample time t_k by

$$\mathcal{N}_{H,k} = \sum_{i=1}^{N_H} H_i(t_k) \tag{2.8}$$

for the hosts and by

$$\mathcal{N}_{V,k} = \sum_{j=1}^{N_V} V_j(t_k) \tag{2.9}$$

for the viruses. We used the normalized N_H host and N_V virus samples as input for the SparCC computation using the `SparCC.py` script. All parameters were left with their default settings. We used a custom Matlab script `write_sparcc.m` to generate .csv data files in the format specified by the SparCC documentation. We used a custom bash script `sparcc_compute_all.sh` to run the SparCC analysis on the ensemble of virus-

microbe communities. Finally, we used a custom Matlab script `read_sparcc.m` to import the results into Matlab for scoring (see subsection 2.4.8).

2.4.8 Scoring correlation network accuracy

To evaluate how well the Pearson correlation, eLSA, or SparCC (collectively referred to as "correlation") network $\mathbf{R}$ recapitulates the original interaction network $\tilde{\mathbf{M}}$, we compute the receiving operator curve (ROC). First, we binarize the interaction network $\tilde{\mathbf{M}}$ so that it is a boolean network $\mathbf{M}$ of zeros (non-interactions) and ones (interactions). Then we choose a threshold of interaction c between the minimum and maximum attainable values of the correlation network $\mathbf{R}$; for Pearson correlation these are -1 and +1. Correlations in $\mathbf{R}$ that are greater than or equal to c are categorized as interactions (ones), while those that are less are non-interactions (zeros). The true positive (TP) count is the number of interactions in $\mathbf{M}$ correctly predicted by the thresholded correlation network $\mathbf{R_c}$. The false positive (FP) count is the number of non-interactions in $\mathbf{M}$ incorrectly predicted by $\mathbf{R_c}$. The TP and FP counts are normalized by the number of interactions and non-interactions in $\mathbf{M}$ to obtain the true positive rate (TPR) and false positive rate (FPR). TPR and FPR are computed for all thresholds c to obtain the receiver operator curve (ROC).

The overall "score" of the correlation network $\mathbf{R}$ is the area under the curve (AUC). A perfect prediction results in AUC=1, since for some threshold TPR=1 and FPR=0. Random predictions result in AUC=1/2, since TPR=FPR across all possible thresholds. AUC values which are less than 1/2 indicate a misclassification of "interaction", that is, categorizing interactions and non-interactions in the opposite way would have resulted in a better prediction of $\tilde{\mathbf{M}}$.

2.5 Results

2.5.1 Standard Pearson correlation

We calculated the standard Pearson correlation networks for an ensemble *in silico* communities that varied in network size and network structure. For each network size $N = 10, 25, 50$, we generated 20 unique interaction networks. 10 of the networks were generated so that they were distributed along a range of nestedness values, and the other 10 were generated so that they were distributed along a range of modularity values (see subsection 2.4.2). For each interaction network, a single set of life history traits were generated to ensure coexistence using biologically feasible ranges according to subsection 2.4.3. The mechanistic model for the community dynamics is described in subsection 2.4.1. Time-series were simulated according to subsection 2.4.4 with $\delta = 0.3$, that is, the initial conditions were the fixed point values perturbed by 30% (for additional values of δ see Figure A.5 in the appendix). For $\delta = 0.3$, the mean coefficient of variation was 12% for host time-series and 4% for virus time-series (see Figure A.1 in the appendix). The time-series were sampled during the transient dynamics to represent *in situ* communities which are likely perturbed from equilibrium due to changing environmental conditions and intrinsic feedback. We sampled the time-series every 2 hours for 200 hours, that is, we took 100 samples (for additional sample frequencies see Figure A.8 in the appendix).

For each *in silico* community, we calculated the standard Pearson correlation network as described in subsection 2.4.5. Two example *in silico* communities of size $N = 10$ are shown in Figure 2.2 with their simulated time-series, log-transformed samples, and resulting correlation networks. The correlation networks were scored against the original interaction networks by computing AUC as described in subsection 2.4.8. The procedure for computing AUC is shown in Figure 2.3 for the two example *in silico* communities.

AUC values for all *in silico* communities are shown in Figure 2.4. Across varying network size and network structure, AUC is approximately $1/2$ implying that standard Pear-

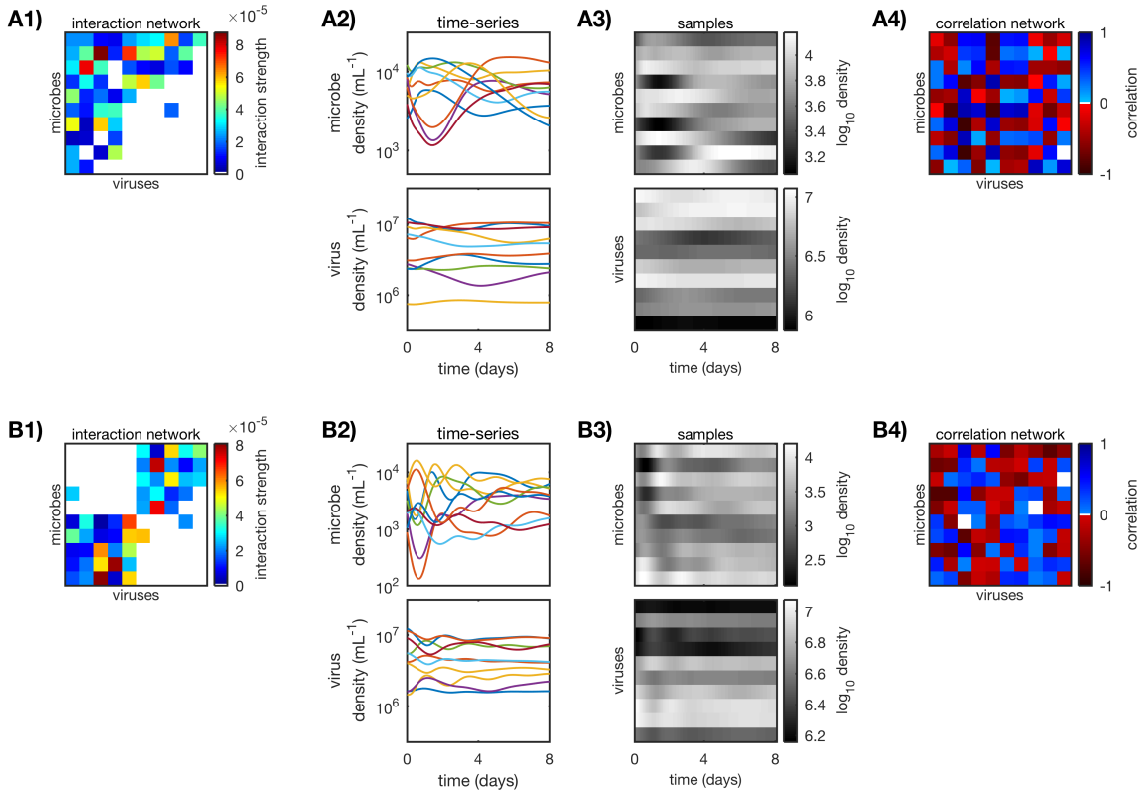

Figure 2.2: Calculating standard Pearson correlation networks for two *in silico* A) nested and B) modular communities ($N = 10$). A1-B1) Original weighted interaction networks, generated as described in subsection 2.4.2 and subsection 2.4.3. A2-B2) Simulated time-series of the virus-microbe dynamical system as described in subsection 2.4.4 ($\delta = 0.3$). A3-B3) Log-transformed samples, sampled every 2 hours for 200 hours from the simulated time-series. A4-B4) Pearson correlation networks, calculated from log-transformed samples as described in subsection 2.4.5.

son correlation networks lack predictive power. Similar results were found when varying the initial condition perturbation δ (Figure A.4) and the sampling frequency (Figure A.8). There are some cases for the smaller networks ($N = 10$) where AUC does deviate from $1/2$ although these deviations are small ($\approx \pm 10\%$). Interestingly these deviations tend to be negative indicating a misclassification of the interaction condition, that is, negative correlations are slightly better predictors of interaction than positive correlations. Overall however, the deviations disappear for larger networks ($N = 50$) implying that they are exceptions rather than the norm. We completed identical analyses for additional correlation metrics in particular Spearman correlation and Kendall correlation (see Figure A.2 in

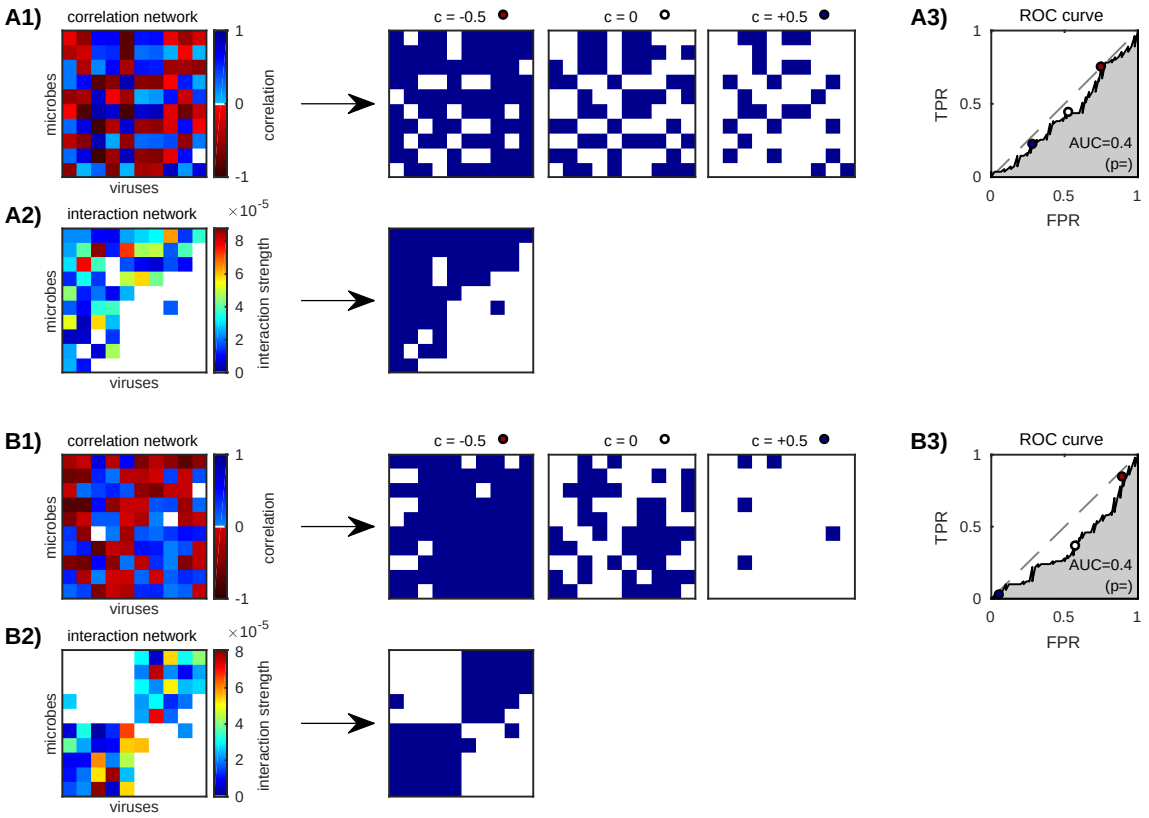

Figure 2.3: Scoring correlation network accuracy of the two *in silico* A) nested and B) modular communities ($N = 10$, see Figure 2.2) as described in subsection 2.4.8. A1-B1) Correlation networks are binarized according to thresholds c between -1 and $+1$, three of which are shown here ($c = -0.5, 0,$ and 0.5). A2-B2) Original interaction networks are also binarized. A3-B3) True positive rate (TPR) versus false positive rate (FPR) of the binarized correlation networks for each threshold c. Three example thresholds ($c = -0.5, 0,$ and 0.5) are marked (red, white, and blue circles). The "non-discrimination" line (grey dashed) is where TPR = FPR. The AUC or area under the ROC curve is a measure of relative TPR to FPR over all thresholds; AUC = 1 is a perfect result. Distributions for the reported p-values are shown in the appendix.

appendix). We found similar results reinforcing our conclusion that simple correlations between time-series are poor predictors of the underlying interaction network.

2.5.2 Time-delayed Pearson correlation

Given the results of the previous section subsection 2.5.1 – that standard correlations do not recapitulate interactions – we computed time-delayed correlation networks for the same ensemble of *in silico* communities. The addition of time-delays to standard correlation

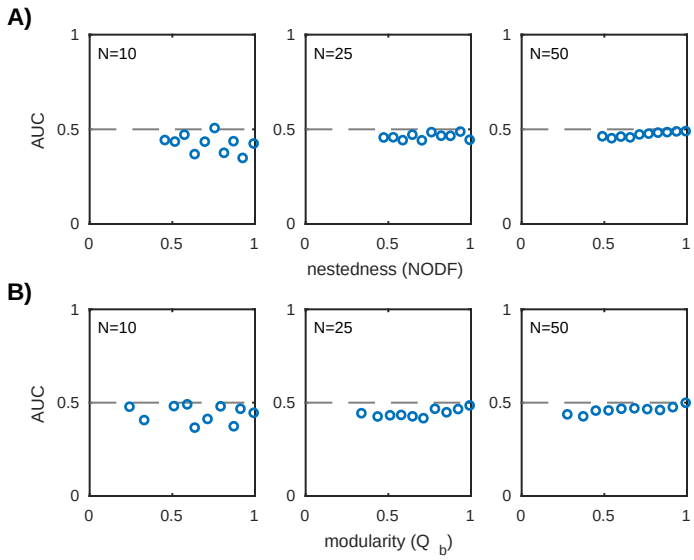

Figure 2.4: AUC values for standard Pearson correlation for the ensemble of A) nested and B) modular communities over three network sizes $N = 10, 25, 50$ (20 communities for each network size). AUC is computed as described in subsection 2.4.8. Each plotted point corresponds to a unique *in silico* community. Dashed line marks AUC=1/2 and implies the predicted network did no better than random guessing.

approaches is motivated by a large body of theoretical work on predator-prey dynamics, where both predator and prey populations oscillate but with a phase delay between them [56]. Similar results hold for the phase delay in simple phage-bacteria dynamics [9]. Time-delayed correlations are the basis of several existing correlation-based inference methods including eLSA [31, 32, 33].

For this analysis, we used the same ensemble of *in silico* communities (network sizes $N = 10, 25, 50$ of varying nestedness and modularity), simulated time-series ($\delta = 0.3$; see Figure A.5 in appendix), and sample frequency (2 hours; see Figure A.6 in appendix) as before (see subsection 2.5.1 for time-series). We calculated the time-delayed Pearson correlation networks as described in subsection 2.4.5, where for each virus-host pair the virus time-series is sampled later in time by some delay amount τ_{ij} relative to the host time-series (for Spearman and Kendall correlation, see Figure A.3 in appendix). Each delay is chosen such that the absolute value of the correlation for the virus-host pair is maximized. Since the optimal time-delay is not known in advance, delays between $0 < \tau_{ij} < t_S/2$, (0

hours and $t_S/2 = 100$ hours) were considered. The number of samples used to compute each correlation coefficient was kept fixed at $S = 100$ (sample duration 200 hours). Time-delayed Pearson correlation networks for the two example *in silico* communities of size $N = 10$ are shown in Figure 2.5A-B. AUC was computed as described in subsection 2.4.8.

AUC values for all *in silico* communities are shown in Figure 2.5C. For the small networks ($N = 10$) there are a few particular networks which have AUC scores greater than $1/2$. For the remaining small networks and the large networks ($N = 25, 50$), AUC $\approx 1/2$ implying time-delayed Pearson correlation lacks predictive power for these networks. Similar results were found for alternate correlation metrics (Spearman and Kendall; Figure A.3), initial condition perturbations δ (Figure A.5), and sampling frequencies (Figure A.6). Because AUC deviates from $1/2$ for only a few small networks and disappears for large networks, it should be considered an exceptional result rather than the norm for time-delayed Pearson correlation.

2.5.3 Correlation-based methods eLSA and SparCC

We performed a similar *in silico* analysis using eLSA [31, 32, 33] and SparCC [42], two established correlation-based inference methods which are widely used with *in situ* time-series data. We used the same ensemble of *in silico* communities as before (network sizes $N = 10, 25, 50$ of varying nestedness and modularity), along with the simulated time-series ($\delta = 0.3$; see Figure A.7), sample frequency (2 hours; see Figure A.9) and sample duration (200 hours). We implemented eLSA and SparCC as described in subsection 2.4.6 and subsection 2.4.7 respectively. eLSA and SparCC predicted networks for the two example *in silico* communities of size $N = 10$ are shown in Figure 2.6A-B. AUC was computed as before and as described in subsection 2.4.8.

AUC values for all *in silico* communities are shown in Figure 2.6C. We see the same trends as with standard correlation and time-delayed correlation (see Figs Figure 2.4 and Figure 2.5). Similar results hold for varying values of the initial condition perturbation δ

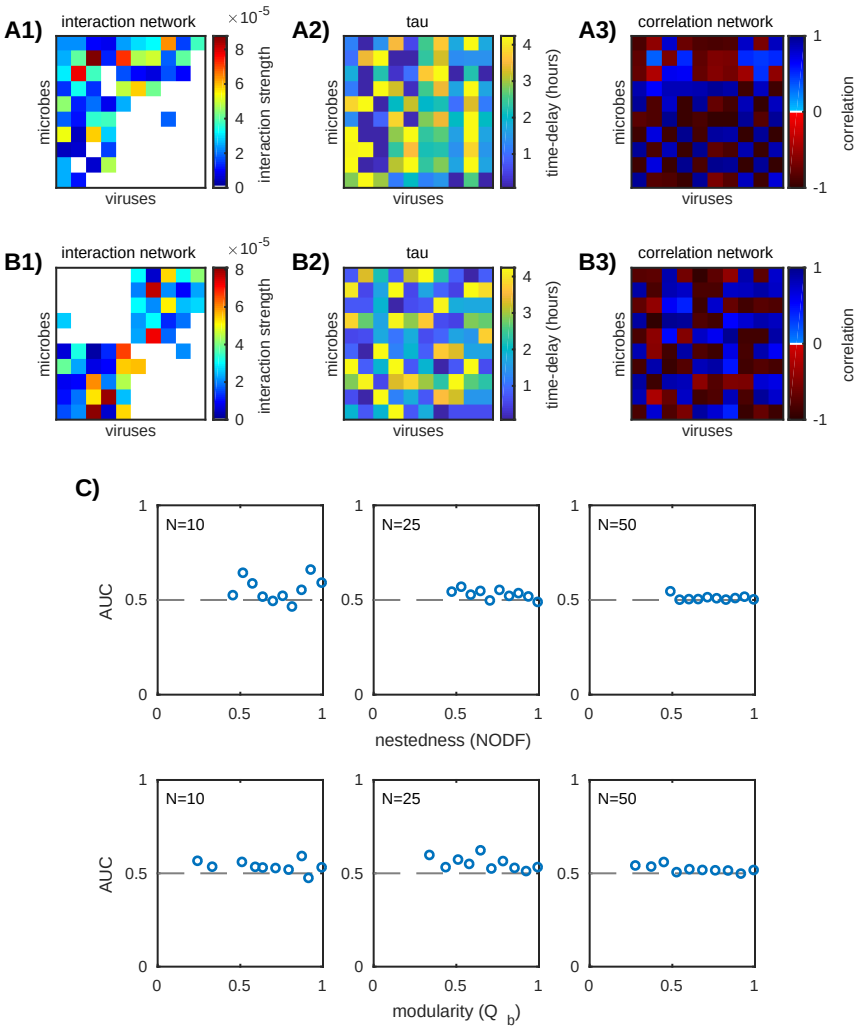

Figure 2.5: Performance of time-delayed Pearson correlation. A1-B1) Two example *in silico* interaction networks ($N = 10$). A2-B2) Time-delays τ_{ij} for each virus-host pair, chosen so that the absolute value of the correlation is maximized. A3-B3) Time-delayed Pearson correlation networks calculated as described in subsection 2.4.5. C) AUC values for the ensemble of nested (top row) and modular (bottom row) communities over three network sizes $N = 10, 25, 50$ (20 communities for each network size). Each plotted point corresponds to a unique *in silico* community. Dashed line marks AUC=1/2 and implies the predicted network did no better than random guessing.

(Figure A.7) and sampling frequency (Figure A.9). For small networks ($N = 10$), there are a few AUC scores which deviate weakly from 1/2 ($\approx \pm 10\%$). Interestingly, AUC scores for eLSA tend to be negative, implying a misclassification of interaction. AUC converges to 1/2 as network size increases ($N = 25, 50$) indicating that the AUC scores for small networks may themselves be spurious.

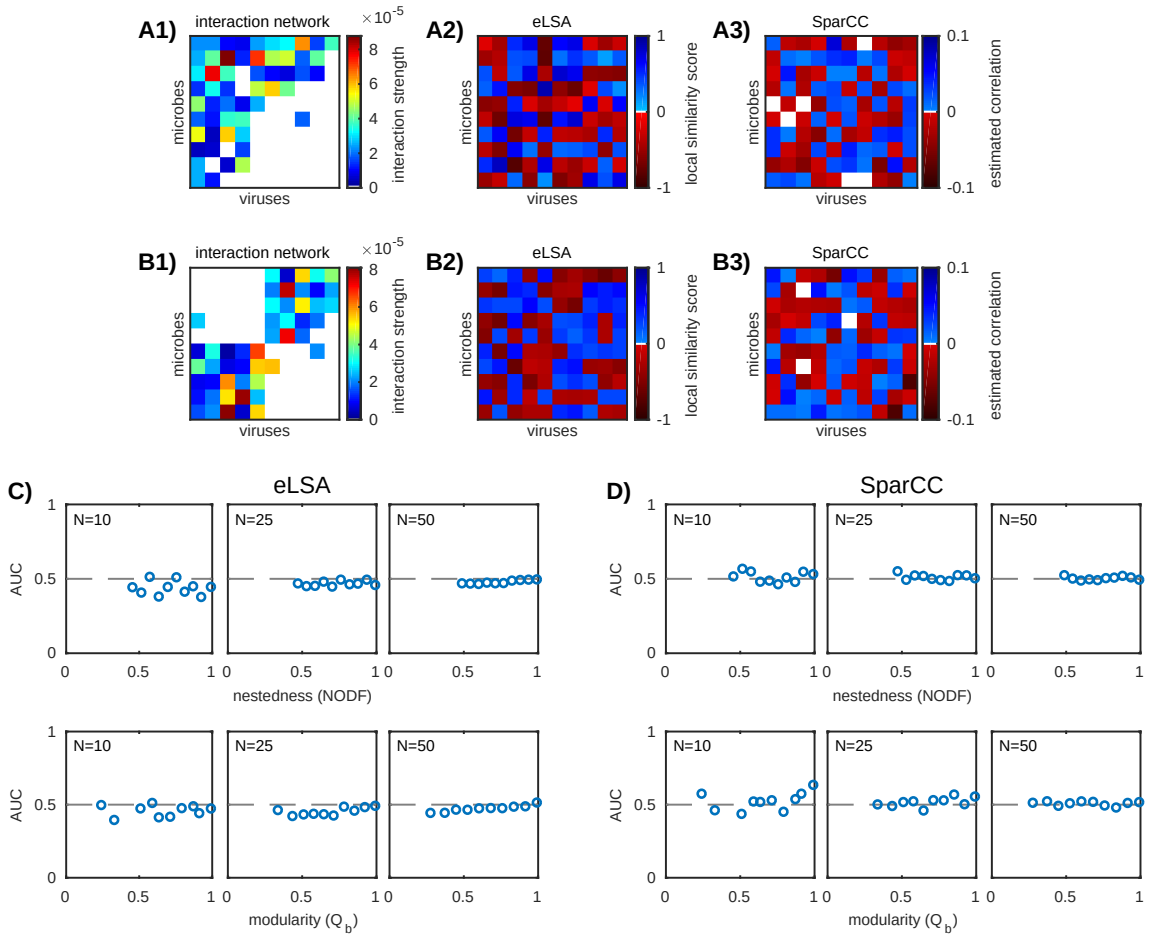

Figure 2.6: Performance of correlation-based inference methods eLSA and SparCC. A1-B1) Two example *in silico* interaction networks ($N = 10$). A2-B2) eLSA predicted network computed as described in subsection 2.4.6. A3-B3) SparCC predicted network computed as described in subsection 2.4.7 (color bar adjusted for visibility). C-D) AUC values for the ensemble of nested (top row) and modular (bottom row) communities over three network sizes $N = 10, 25, 50$ (20 communities for each network size). Each plotted point corresponds to a unique *in silico* community. Dashed line marks AUC=1/2 and implies the predicted network did no better than random guessing.

2.6 Discussion

Using *in silico* virus-microbe community dynamics, we calculated correlation networks among viral and microbial population time-series samples. We tested the accuracy of several different types of correlation and time-delayed correlation (Pearson, Spearman, and Kendall) and existing correlation-based inference methods (eLSA and SparCC). The correlation networks for all of these implementations failed to effectively predict the original

interaction networks, as quantified by the AUC score. Failure persisted across variation in network structure, network size, degree of initial condition perturbation (*i.e.* scaling the variability of dynamics), and sampling frequency. We therefore conclude these correlation-based inference methods do not meaningfully predict interactions given this mechanistic model of virus-microbe community dynamics.

Earlier, we stated the core assumption of correlation-based inference: that interacting populations are correlated and that non-interacting populations are not correlated. While this core assumption may sometimes hold in small microbe-only communities with simple interaction mechanisms [24], we find it does not necessarily hold in more complex virus-microbe communities. (Each inference method also faces challenges unique to its formulation: eLSA in particular uses a non-stationary data transformation which may induce additional spurious correlations.) We considered communities with microbes and viruses that interacted through a nonlinear mechanism (infection and lysis) across a spectrum of network sizes and network structure. We found that correlation-based inference performed poorly given variation in these network properties, but that there was greater variation in performance for small networks. Because this variation is relatively small and disappears for larger networks, successful predictions for small networks may themselves be spurious. Namely, for a small network (*e.g.* $N < 10$), there is a greater probability of randomly guessing the interactions correctly because the space of possible networks is smaller.

Our results raise concerns about the use of correlation-based methods on *in situ* datasets, since a typical community under consideration will have dozens or more interacting strains and therefore will not be in the low diversity microbe-only regime explored by [24]. Additional challenges such as external environmental drivers, measurement noise, and system stochasticity must also be carefully considered before applying correlation-based methods to *in situ* datasets. Although the degree of variability of dynamics had no effect on inference quality here, it may also be an important consideration for both experimental design and choice of inference method. For example, the model-based inference method developed by

[44] performs better when dynamics are highly variable. On the other hand, co-occurrence based inference methods, which require samples across space instead of time, may enable inference across different baseline environmental conditions even if the dynamics within a given environment are relatively stable.

In light of the poor performance of correlation-based methods, we advocate for increased studies of model-based inference. Model-based inference methods operate by first assuming an underlying dynamical model for the community (such as the one used in this chapter, Equation 2.1 and Equation 2.2). The dynamical model is then used to formulate an objective function for an optimization or regression problem, where the solution is the interaction network which best describes the sampled community time-series (for example, see [50, 41, 57, 44, 58, 59]). Unlike correlation-based methods which assume that similar trends in population indicate interaction, model-based inference has the potential to be tailored to complex communities and environments while leveraging existing knowledge about ecological mechanisms. Given favorable results of *in silico* benchmarking of model-based inference methods [50, 41, 57, 44, 58, 59], it will be important to investigate the efficacy of model-based inference methods for complex microbial and viral communities in practice.

2.7 Acknowledgments

We are grateful to Ben Bolduc, Stephen Beckett, and five anonymous reviewers for helpful comments and feedback. We thank both Yu-Hui Lin and David Demory for reviewing the code used in the analysis. This work was supported by the Simons Foundation (SCOPE award ID 329108, J.S.W.).

2.8 Availability of data and materials

Analysis was primarily performed in Matlab. All Matlab scripts, Matlab data files (also available as .csv files), and custom bash scripts for implementing eLSA and SparCC are

publicly available on GitHub (https://github.com/WeitzGroup/correlation_based_inference) and archived on Zenodo (DOI 10.5281/zenodo.844918). The BiMat Matlab package [51] used for characterizing bipartite networks is available on GitHub (https://github.com/cesar7f/BiMat). The eLSA Python package [31, 32, 33] is available on Bitbucket (https://bitbucket.org/charade/elsa/wiki/Home). The SparCC Python package [42] is available on Bitbucket (https://bitbucket.org/yonatanf/sparcc).

CHAPTER 3

INFERRING MULTIPLE INTERACTION NETWORKS IN VIRUS-BACTERIA COMMUNITIES FROM TIMESERIES DATA

Adapted from A R Coenen, A Bottu, S J Beckett, J Romberg, and J S Weitz, Inferring multiple interaction networks in virus-bacteria communities from timeseries data (in prep).

3.1 Abstract

Bacteria and their viruses coexist in complex communities across many different environments. Often we wish to know the interaction patterns of bacteria and viruses in these communities, such as which viruses infect which bacteria, or which bacteria have competitive or synergistic effects on one another. In general, it is difficult to measure such interactions directly due to technical limitations like culturability and scalability, as natural communities may have upwards of hundreds of coexisting populations. These limitations highlight the need for high-throughput, culture-independent methods for inferring interaction patterns of bacteria and viruses. Recent work has demonstrated proof-of-concept for inferring the virus-bacteria infection network from abundance timeseries data using model-based optimization. We extend this recent work and demonstrate how other interaction networks, in particular the bacteria-bacteria competition network, can be inferred simultaneously. We show that inference is robust to variation in network structure and characterize tradeoffs in inference quality, average interaction strength, and sampling. Finally, we find that feasibility – the prediction of long-term coexistence of all populations – drastically improves the quality of inference.

3.2 Methods

3.2.1 Model for bacteria-virus community ecology

We model the ecological dynamics of a virus-bacteria community with a system of nonlinear ordinary differential equations, building from previous work [44]. The community has multiple populations of bacterial hosts and viruses which are defined by their life history traits. The dynamics are given by

$$\begin{aligned} \text{bacteria} \quad \dot{H}_i &= \overbrace{r_i H_i \left(1 - \frac{1}{K}\sum_{i'=1}^{N_H} a_{ii'} H_{i'}\right)}^{\text{growth and competition}} - \overbrace{H_i \sum_{j=1}^{N_V} M_{ij}\phi_{ij} V_j}^{\text{lysis}} \\ \text{viruses} \quad \dot{V}_j &= \underbrace{V_j \sum_{i=1}^{N_H} M_{ij}\phi_{ij}\beta_{ij} H_i}_{\text{adsorption and lysis}} - \underbrace{m_j V_j}_{\text{decay}} \end{aligned} \tag{3.1}$$

where H_i and V_j are the densities of bacteria population i and virus population j respectively. There are N_H and N_V different bacteria and virus populations. Each bacteria population i has growth rate r_i, and there is a community-wide carrying capacity K. Bacteria within a single population and between other populations interact through the competition term $a_{ii'}$. Specifically, $a_{ii'}$ is the competitive (or synergistic) effect of bacteria in population i' on i. Each virus population j has decay rate m_j and interacts with bacteria populations with host-specific parameters. Viruses in population j infect bacteria in population i if $M_{ij} = 1$ and do not infect if $M_{ij} = 0$. Viruses in population j adsorb to hosts i with adsorption rate ϕ_{ij}, and after lysis create the burst size β_{ij} new virions.

3.2.2 Fixed points and requirements on M

The fixed point densities are obtained by setting all derivatives $\dot{H}_i$ and $\dot{V}_j$ to zero and solving for H_i and V_j, resulting in

$$\begin{aligned} \vec{H}^* &= \left((M \circ \phi \circ \beta)^T\right)^{-1} \vec{m} \\ \vec{V}^* &= (M \circ \phi)^{-1} \vec{r} \circ \left(\vec{1} - a\vec{H}^*/K\right) \end{aligned} \tag{3.2}$$

where $\circ$ is element-wise multiplication.

In order for the fixed points (H_i^*, V_j^*) to exist, the matrices $M \circ \phi \circ \beta$ and $M \circ \phi$ must be invertible. Since M is a boolean matrix of zeros and ones, it is sufficient that M be invertible. When sampling for a parameter set, we use MATLAB's `randi()` to repeatedly generate candidates for M until we obtain a matrix with no all-zero rows or columns. We apply this rule regardless of which parameter sampling strategy is used.

3.2.3 Parameter sampling

Plausible strategy

We consider two different strategies for sampling parameters for the model for host-virus ecology: plausible and feasible. In the plausible strategy, we designate biologically reasonable distributions for each parameter, then sample parameters independently for each population of bacteria and viruses. Here, all distributions are uniform, and ranges are given in Table 3.1.

Feasible strategy

For the feasible strategy, we follow the steps detailed in previous work [55]. In brief, we sample most of the parameters ($a_{ii'}, K, M_{ij}, \phi_{ij}, \beta_{ij}$) independently from biologically reasonable ranges (Table 3.1). In the same way, we also sample the steady state densities for hosts H_i^* and viruses V_j^* (Table 3.1). We then solve for the remaining, unknown parameters

Table 3.1: Sampling ranges and units for parameters of the model of bacteria-virus ecology (Equation 3.1).

	parameter	sampling range	units
r_i	host i growth rate		$1/\text{days}$
$a_{ii'}$	competitive effect of host i' on host i	$0-1$	
K	community-wide host carrying capacity	10^6	cells/mL
M_{ij}	interaction between host i and virus j	0 or 1	
ϕ_{ij}	adsorption rate of virus j into host i	$10^{-8}-10^{-7}$	$\text{mL}/(\text{virion}\cdot\text{day})$
β_{ij}	burst size of virus j per host i	$10-100$	$\text{virions}/\text{cell}$
m_j	virus j decay rate		$1/\text{days}$
H_i^*	host i steady-state density	10^3-10^4	cells/mL
V_j^*	virus j steady-state density	10^6-10^7	$\text{virions}/\text{mL}$

(r_i, m_j) using the fixed point equations (Equation 3.2)

$$\begin{aligned} \vec{m} &= (M \circ \phi \circ \beta)^T \vec{H}^* \\ \vec{r} &= \frac{(M \circ \phi)\vec{V}^*}{\vec{1} - a\vec{H}^*/K} \end{aligned} \tag{3.3}$$

where $\circ$ is element-wise multiplication, and division is also element-wise. We then check that all values (r_i, m_j) are positive and biologically plausible. If they are not, we re-sample the steady state densities (H_i^*, V_j^*) until plausibility is achieved. Because the steady state densities are positive, the resulting parameter set is feasible and results in local, short-term coexistence of all host and virus populations.

3.2.4 Timeseries simulations

We simulate the system of differential equations (Equation 3.1) using MATLAB's `ode45()`. We use a relative error tolerance of 10^{-8} and a timestep of 15 minutes.

We set initial host and virus densities according to which parameter sampling strategy was used. For plausible (non-feasible) parameter sets, initial host and virus densities are chosen randomly from the biologically reasonable ranges specified for H_i^* and V_j^* respectively (see Table 3.1). For feasible parameter sets, in which all fixed points (H_i^*, V_j^*) are

positive, we perturb the fixed points by a constant δ

$$H_i(0) = (1 \pm \delta)H_i^*, \quad V_j(0) = (1 \pm \delta)V_j^* \tag{3.4}$$

where either $+$ or $-$ is chosen randomly for each host and each virus population.

3.2.5 Objective function

We derive the objective function by discretizing the system of differential equations (Equation 3.1). We treat the host and virus equations separately, resulting in two distinct objective functions. The derivation of the virus objective function is given in detail in previous work [44].

Before discretizing the host equations, we apply the chain rule to obtain

$$\frac{d \ln H_i}{dt} = r_i \left(1 - \frac{\sum_{i'=1}^{N_H} a_{ii'} H_{i'}}{K}\right) - \sum_{j=1}^{N_V} M_{ij}\phi_{ij}V_j \tag{3.5}$$

We use the forward difference to approximate the derivative at time t_k

$$\frac{\Delta \ln H_i(t_k)}{\Delta t_k} \approx r_i - \sum_{i'=1}^{N_H} \tilde{a}_{ii'} H_{i'}(t_k) - \sum_{j=1}^{N_V} M_{ij}^{\phi} V_j(t_k) \tag{3.6}$$

For notational convenience, we define the approximated derivative as

$$G_{ik} \equiv \frac{\Delta \ln H_i(t_k)}{\Delta t_k} = \frac{\ln H_i(t_{k+1}) - \ln H_i(t_k)}{t_{k+1} - t_k} \tag{3.7}$$

and note that the analogous term W_{jk} for the approximated virus derivative is detailed in [44]. We also define the weighted host-host interactions

$$\tilde{a}_{ii'} \equiv \frac{r_i a_{ii'}}{K} \tag{3.8}$$

again noting the analogous host-virus interaction term $\tilde{M}_{ij} = M_{ij}\phi_{ij}\beta_{ij}$ as described in

[44]. We write N_T such equations for $N_T + 1$ measured time points $t_1 \dots t_{N_T+1}$. The N_T equations, for all N_H hosts, may be written compactly in matrix form

$$G \approx \vec{r} \otimes \vec{1} - \tilde{a}H - M^{\phi}V \tag{3.9}$$

where $\vec{1}$ is a $1 \times N_T$ vector of ones, such that the outer product $\vec{r} \otimes \vec{1}$ results in a $N_H \times N_T$ matrix. G, H, and V are the measured (and in the case of G, also transformed) host and virus density matrices with dimensions $N_H \times N_T$ (for G and H) and $N_V \times N_T$ (for V).

We use the discretized host matrix equation (Equation 3.9) as our objective function in an optimization problem to infer the unknown parameters $\vec{r}$, $\tilde{a}$, and M^{ϕ}

$$\begin{aligned} \underset{(\vec{r},\, \tilde{a},\, M^{\phi})}{\text{minimize}} \quad & \left\| G - \vec{r} \otimes \vec{1} + \tilde{a}H + M^{\phi}V \right\|_2 + \lambda_{\tilde{a}} \left\| \tilde{a} \right\|_1 + \lambda_{M^{\phi}} \left\| M^{\phi} \right\|_1 \\ \text{subject to} \quad & r_i \geq 0, \\ & \tilde{a}_{ii'} \geq 0, \\ & M^{\phi}_{ij} \geq 0. \end{aligned} \tag{3.10}$$

where $\|\bullet\|_2$ is the Frobenius norm and $\|\bullet\|_1$ is the 1-norm, that is, lasso regularization. The hyperparameters $\lambda_{\tilde{a}}$ and $\lambda_{M^{\phi}}$ are used to tune the sparsity of the interaction matrices $\tilde{a}$ and M^{ϕ}.

3.2.6 Convex optimization

The optimization problem (Equation 3.10) is convex and thus readily solvable with existing software packages. We use the MATLAB package CVX, which is open source and free to use [60].

3.2.7 Quantification of inference quality

To quantify the overall quality of the inference, we use the relative error

$$\epsilon = \frac{\left\| X - \tilde{X} \right\|_2}{\|X\|_2} \tag{3.11}$$

where X is the original matrix, $\tilde{X}$ is the inferred matrix, and $\|\bullet\|_2$ is the Frobenius norm. The relative error is 0 for a perfect match and grows larger when the inference is worse.

In addition, we quantify how well the inferred matrix performed as a binary classifier, that is, which host-host or virus-host pairs it correctly predicted as interacting (non-zero) versus not interacting (zero). Here we use AUC ("area under the curve") of an ROC ("receiver operator characteristic") curve, that is, the false positive rate versus the true positive rate. The AUC is 1 for a perfect classifier and becomes closer to 0.5 as the classifier performs no better than random guessing.

3.3 Results

3.3.1 Inference of heterogeneous microbe-microbe competition

To begin, we show an example of inference for a 10x10 *in silico* virus-host community with heterogeneous microbe-microbe competition (Figure 3.1). The ecological dynamics of the community are described with a system of differential equations, drawn from previous work [44] and modified to allow for heterogeneous microbe-microbe competition (Methods, Model for bacteria-virus community ecology). The life history traits for the community are sampled from biologically reasonable ranges (Table 3.1) using the feasible sampling strategy (Methods, Parameter sampling, Feasible strategy). By using the feasible strategy, we ensure that the fixed points of the system are all positive, thus leading to local coexistence of virus and host populations. We simulate timeseries for the virus-host community at high time-resolution, setting initial conditions to be the fixed points perturbed by

a constant $\delta = 0.7$ (Methods, Timeseries simulations). Simulated virus and host timeseries are shown in Figure 3.1a.

To infer the weighted host-host interaction network $\tilde{a}$ (Equation 3.8), we solve the convex optimization problem Equation 3.10 derived from the host equations in our model (Methods, Objective function and Convex optimization). The derivation is analogous to the virus-host inference problem described in [44]. In addition, we simultaneously infer the weighted virus-host interaction network by solving a separate convex optimization problem as described in [44]. The inclusion of the virus-host network $\tilde{M}$ provides a baseline expectation for assessing the performance of the host-host inference.

As shown in Figure 3.1, we are able to successfully infer the host-host network $\tilde{a}$ in the presence of heterogeneous interactions. The performance of the host-host inference is comparable to that of the virus-host inference, as quantified by relative error between the inferred network and the true network (0.056 and 0.028 respectively; Methods, Quantification of inference quality).

Next, we characterize variation in inference quality for the same 10x10 virus-host community given variation in the initial host and virus densities (Figure 3.2). Previous work [44] found that the amount of variability in the sampled dynamics strongly affected the inference quality. In particular, inference performed better when variability was high. Here, we tune the initial condition perturbation constant δ so that the community is simulated across a range of distances from its equilibrium; δ is thus a proxy for dynamic variability, with small δ typically leading to less variation.

In Figure 3.2, we present inference quality for both the host-host network $\tilde{a}$ and the virus-host network $\tilde{M}$ across a range of δ values. In contrast to previous work [44], we do not find a clear trend between δ and inference quality. Instead, we find that the relationship between δ and inference quality is distinctively unique across different virus-host communities (Appendix B), despite using the same parameterization framework (Methods, Parameter sampling, Feasible strategy). Furthermore, the trend does not emerge when

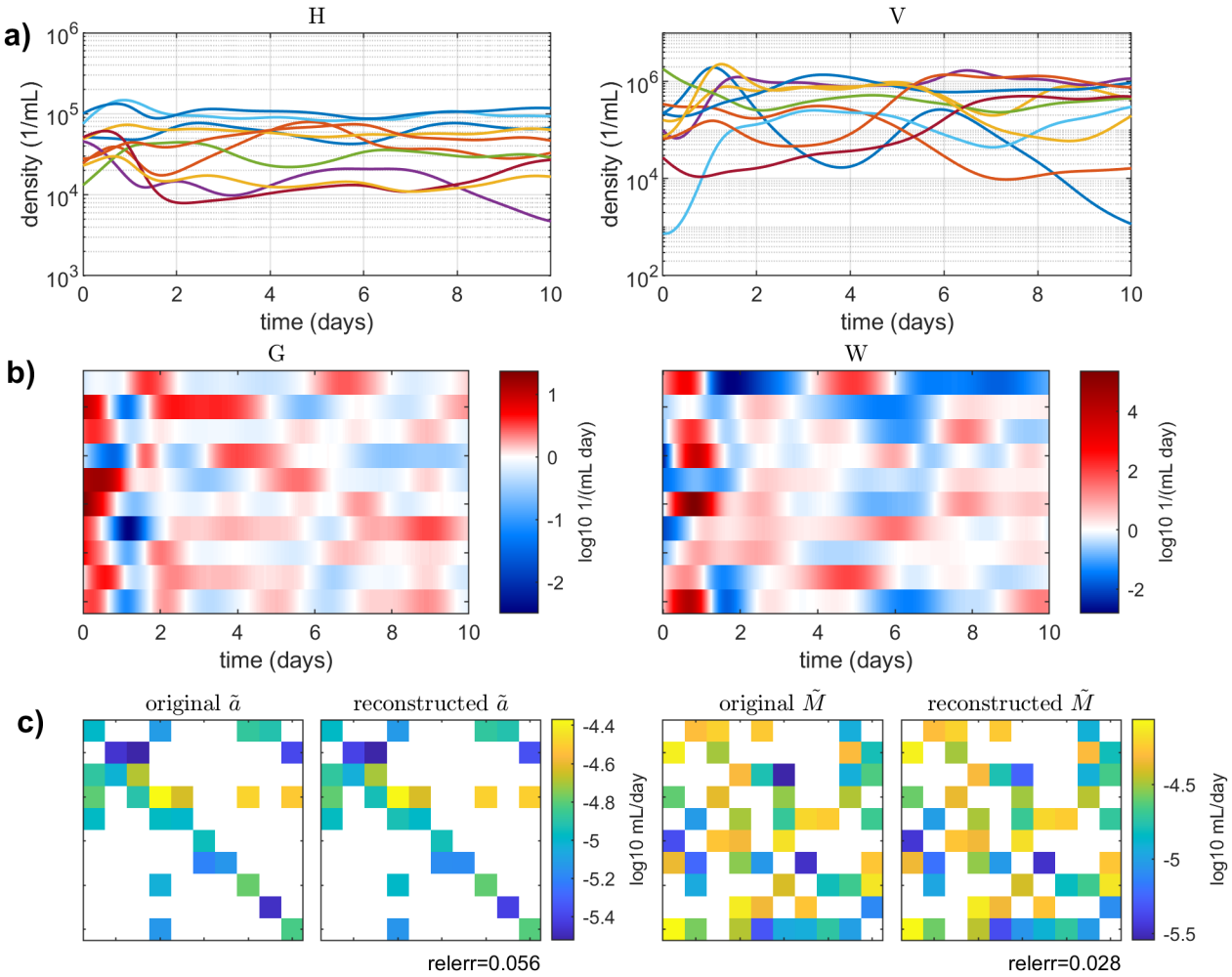

Figure 3.1: High-quality inference for a 10x10 virus-host community with heterogeneous host-host interactions. a) Simulated host H and virus V population timeseries. b) Transformed host G and virus W densities (Equation 3.7). c) Inference of both the heterogeneous host-host interaction matrix $\tilde{a}$ (left) and the host-virus interaction matrix $\tilde{M}$ (right; Equation 3.8). Both the original and reconstructed matrices are shown. Relative error is computed by Equation 3.11.

averaging inference quality across an ensemble of parameter sets (not included).

3.4 Discussion

In this chapter, we demonstrate that both the microbe-microbe and virus-host interaction networks can be inferred with high accuracy in the presence of heterogeneous microbe-microbe competition (Figure 3.1). This extends previous work [44] in which only the virus-host network was inferred in the presence of uniform microbe-microbe competition.

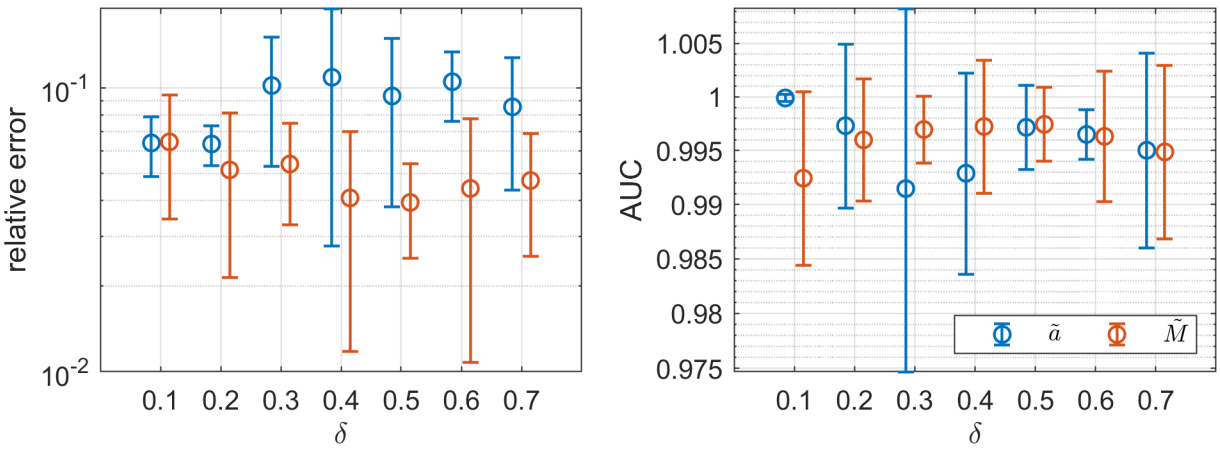

Figure 3.2: Variation in inference quality for a single 10x10 virus-host community (as shown in Figure 3.1). The initial condition perturbation δ is varied such that the community starts closer to ($\delta = 0.1$) or farther from ($\delta = 0.7$) its equilibrium (Equation 3.4). For each value of δ, an ensemble of $N = 10$ initial condition sets are simulated (Methods, Timeseries simulations). Inference quality is quantified by relative error (left) and AUC (right; see Methods, Quantification of inference quality) and is shown for both the host-host matrix $\tilde{a}$ (blue) and the virus-host matrix $\tilde{M}$ (orange; Equation 3.8).

Unlike previous work [44], we did not find a clear trend between initial condition perturbation δ (a proxy for variability in dynamics) and inference quality (Figure 3.2), and instead found that the relationship was unique for different parameter sets (Appendix B). Nonetheless, inference quality for the host-host network remained consistently high across an ensemble of parameter sets and was comparable in magnitude to (and occasionally lower than) inference quality of the virus-host network from previous work [44].

We have explored several intriguing directions for this model-based inference method, which we briefly describe here. First, we have seen evidence that inference quality for both the host-host and virus-host networks depends on the relative strength of the interactions. That is, when virus-host interactions are stronger on average than host-host interactions, inference of the virus-host network improves, while inference of the host-host network gets worse. We have observed similar behavior when host-host interactions are stronger. In future work, we will characterize this relationship quantitatively; we are interested to see where the transition from superior virus-host to superior host-host inference lies, as well as how quickly inference quality deteriorates for the weaker network. Second, we have

observed that the parameter sampling framework dramatically affects inference quality for both the host-host and virus-host networks. In particular, inference on feasible parameter sets performs much better than on plausible (and non-feasible) parameter sets. This has interesting implications for application to *in vitro* and *in situ* phage-bacteria communities. From an ecological perspective, we expect phage-bacteria communities to be feasible at distinct points across time and space, but with extinction and invasion events which effectively change the community make-up. Finally, unlike previous work [44], the convex optimization problem for inferring the host-host network has coupled parameters. Namely, the growth rates $\vec{r}$ are embedded in the weighted host-host interaction network $\tilde{a}$ (Equation 3.8. In this chapter, we have shown that the host-host network $\tilde{a}$ can be accurately inferred despite this coupling. It remains to be explored if the structure of $\tilde{a}$ imposed by $\vec{r}$ can be leveraged in some way for more accurate inference.

3.5 Acknowledgements

This work was supported in part through computing resources provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA.

This work was supported by the National Science Foundation under Grant 1829636 (J.S.W.).

CHAPTER 4

RECONSTRUCTING COMMUNITY DYNAMICS FROM PAIRWISE INTERACTIONS IN A COMPLEX BACTERIA-PHAGE COMMUNITY

Adapted from A R Coenen, N Solonenko, C L Sun, M Burris, G Domínguez-Huerta, A Mackey, L Chittick, S J Beckett, D Demory, D Muratore, A Davidson, M B Sullivan, and J S Weitz, Reconstructing community dynamics from pairwise interactions in a complex bacteria-phage community (in prep).

This chapter is in collaboration with the Sullivan Lab at the Ohio State University and is part of a larger collaboration on infection network inference. The work presented in this chapter integrates theory, computation, and experiment for a 5x5 phage-bacteria community. I had the opportunity to contribute to all three parts: I led theory and computation and assisted with the ∼16 hour experiment, led by members of the Sullivan Lab at OSU. Many thanks to Natalie and Christine for welcoming me and showing me the ropes!

Full author contributions are at the end of the chapter.

4.1 Abstract

Complex phage-bacteria networks shape microbiomes and biogeochemical cycles, hence quantifying "who interacts with whom" is essential for understanding downstream ecological, environmental, and evolutionary outcomes. Yet, quantifying infection networks at the population level remains difficult due to gaps in culturability and scalability of "-omics" methods. In this work, we assess the potential for combining trait data with high-throughput sequencing to infer quantitative interactions and dynamical outcomes in a synthetic 5x5 phage-bacteria community of marine *Cellulophaga baltica* and *Pseudoalteromonas* strains

and their associated phage. To do so, we sampled the *in vitro* community at high time- and strain- resolution over approximately 16 hours. We then integrated timeseries data and measured life history traits with nonlinear models of phage-bacteria community dynamics. We find that integrating realistic physiological features of infection is critical to recapitulate measured dynamics at community scales. Although our posterior estimates of quantitative life history traits are consistent with expectations, we also observe context-dependent shifts in life history traits. Follow-up experiments revealed that the virulence of certain phage-bacteria interactions may be changed, and even enhanced, in a community context, suggesting that *a priori* predictions of quantitative community outcomes from pairwise interaction data requires iterative model-data integration.

4.2 Introduction

Complex bacteria-phage communities are integral parts of human-associated, built environment, and environmental microbiomes [61, 62, 10, 63]. Phage infections of bacteria shape the fate of infected cells as well as population dynamics [64]. In the surface ocean, phage-induced lysis has been hypothesized to modulate the retentiveness of the microbial loop, redirecting carbon away from higher trophic layers [9, 2, 65, 66]. Integrated across surface oceans, this "viral shunt" has the potential to impact global biogeochemical cycles [67]. However, recent work has also shown that lysis may produce sticky aggregates which could enhance export of carbon out of the surface layers via a "viral shuttle" [2], a hypothesis supported by global surface ocean datasets that show a strong correlation between subsets of surface ocean phage and export events as measured in the water column [6]. Mechanistic models of the ecological feedbacks amongst diverse communities of marine phage and bacteria are needed to bridge the gap from interactions to ecosystem consequences [46].

Advances in high-throughput "-omics" technology allow for community-scale characterization of bacteria and phage diversity, gene transcription, and protein production.

Large-scale environmental "-omics" surveys have proved notable in expanding the range of viral sequence space, diversity, and even organizing viral sequence space into populations [68, 66, 69, 6]. However, despite efforts to link phenotype from genotype, high-throughput "-omics" provides limited information on quantitative life history traits at the level of individual phage or bacterial strains [70, 22]. In principle, traditional laboratory approaches allow for quantitative measurements of life history traits of bacteria and phage when isolated in culture [71]. Gaps in culturability and the absence of high-throughput life history trait assays means that the bulk of *in situ* interaction rates for phage-bacteria communities remain unknown. As a result, it remains challenging to predict and interpret mechanisms governing phage-bacteria dynamics *in situ*.

One approach to understanding complex communities is to first understand the parts, that is, the phage and bacteria strains or populations [10, 62], and next understand the qualitative nature of interactions, specifically whom interacts with whom [47, 72, 48]. Typically, the interactions between phage and bacteria are characterized in pairwise fashion given a particular culture condition, with sets of phage and bacteria derived from stock strains, evolution experiments, and environmental communities [23, 72]. In some cases, even the quantitative nature of interactions can be measured, e.g. adsorption rates, latent periods, and burst sizes [70, 73]. While labor intensive, such approaches have been used to predict outcomes of community competition experiments. For example, pairwise competition experiments have been used to accurately predict the outcomes of trio competitions for a set of 8 soil-associated bacteria species [74]. However, caution must be used when scaling up in this way because community effects are difficult to predict, in general [74, 75].

Pairwise effects and community outcomes may diverge from one another in several ways. First, life history traits of individual strains may be modified in the presence of additional strains. For example, net bacterial growth rates are reduced via competition for resources and, in other contexts, boosted through cross-feeding [76, 77]. Second, higher-order interactions which are absent from the individual or pairwise context may be pow-

erful drivers in the community context [74]. These 3-body interactions may depend on ecological conditions. For example, the outcomes of infection may differ depending on the multiplicity of infection, either in superinfection scenarios when multiple phage infect the same bacteria or in cross-infection scenarios when distinct phage infect the same bacteria [78, 79, 80].

A complementary approach to understanding complex communities is to use direct sequencing methods to characterize the community as a whole and infer pairwise interactions from community dynamics. Typically, a community is sampled over time and/or space through one or more data types, including cellular metagenomics, viral metagenomics, and cellular metatranscriptomics. The resulting data includes a strain-resolved timeseries or spatiotemporal series of relative abundances, absolute abundances, and transcript abundances. Then, statistical or mechanistic models are used to infer life history traits which best describe the timeseries. Many recent studies have taken this approach to inferring various life history traits for complex bacterial communities [81, 39, 38, 58, 50, 44, 59, 57, 82]. While these approaches are broadly applicable and thus relatively easy to implement, few of the results have been rigorously validated.

In this chapter, we present analysis of a complex marine bacteria-phage community as an attempt to bridge the gap between laboratory and high-throughput approaches. The community consists of 5 bacteria and 5 phage strains, all isolated from surface waters in the Baltic Sea [28]. The use of an intermediate complexity community represents an attempt to bridge the gap between culture-based and "-omics" approaches. In doing so we characterize individual and pairwise life history traits as well as use high-throughput "-omics" techniques for obtaining community-level timeseries. Notably all bacteria and phage strains coexisted over the course of our replicated $\sim$16-hour community experiments. As we show, dynamic community models can recapitulate observed community timeseries, even as our model-data integration reveals context-dependent shifts in some life history traits. In summary, this work provides a novel opportunity for reconciling labora-

tory and high-throughput approaches as well as comparing different scales of ecological complexity.

4.3 Experimental methods

4.3.1 Strains and growth conditions

Five bacterial strains were used: *Cellulophaga baltica* strains NN016038, #4 (1), and #18 (2) were isolated from the Baltic Sea in 1994 (NN016038, hereafter CBA 38) and (3) in 2000 (#4 and #18, hereafter CBA 4 and CBA 18; [83]), while *Pseudoalteromonas* sp. H100 (4) and 13-15 (5) were isolated from the North Sea in 1990 (hereafter PSA H100 and PSA 13-15; [84]). Five bacteriophages were used: 1) ϕ38:1, a podovirus isolated from the Baltic Sea in 2005 on CBA 38, 2) ϕ18:2, a siphovirus isolated from the Baltic Sea in 2000 on CBA 18, 3) ϕ18:3, a podovirus isolated from the Baltic Sea in 2005 on CBA 18 [83], 4) PSA HP1, a podovirus isolated from the North Sea in 1990 on PSA H100, and 5) PSA HS6, a siphovirus isolated from the North Sea in 1990 on PSA 11-68 (strain not included in this study; [85]).

Bacteria were grown on Pseudoalteromonas-Cellulophaga (PC) plates (20.5g Sigma Sea Salts, 1.5g peptone, 1.5g proteose peptone, 0.6g yeast extract, 13g agar/L) at room temperature (RT). Single colonies were inoculated and grown stationary at RT overnight in PC liquid growth medium (20.5g Sigma Sea Salts, 0.75g peptone, 0.5g yeast extract, 0.25g proteose peptone, 0.25g casamino acids, 1.5mL glycerol/L). Phages were stored in phage buffer (20.5g Sigma Sea Salts/L) and plaque forming units (PFUs) enumerated using the agar overlay method (Sambrook and Russell 2001) with 3.5mL molten soft agar (20.5g Sigma Sea Salts and 6g low melting point agarose/L) and 300ul overnight host culture per plate.

4.3.2 Doubling time experiments

Bacterial growth curves to determine doubling times were performed in triplicate in PC growth medium on cultures transferred 1:20 from overnight cultures into new media, grown stationary at RT. A regression of OD600 and colony forming units (CFUs) was constructed using exponential phase timepoints to estimate cell density from OD readings.

4.3.3 Adsorption rate experiments

Adsorption of phages to bacterial strains was characterized in triplicate by combining one phage with one bacterial strain and enumerating free and total phages over time. Pairs with a known interaction were combined at a multiplicity of infection (MOI) of 0.1 (1e7 phages and 1e8 cells/mL) and total and free phage PFUs were enumerated every 3-5 min for 24-25 min to calculate the adsorption constant. Pairs with no known interaction were combined at an MOI of 3 (3e8 phages and 1e8 cells/mL) and total and free phage PFUs were enumerated at time intervals for 3 hrs to examine evidence in support of a statistically significant decrease in free phage. All infections were performed with strains in mid-exponential phase.

4.3.4 One-step experiments

One-step growth curves were performed in triplicate to determine phage burst sizes and latent periods. Each phage–host pair was combined at an MOI of 0.1 (1e7 phages ad 1e8 cells/mL) and incubated for 15 min to allow phages to adsorb to cells. The infection was then diluted 1:100 in PC growth medium to reduce the chances of new adsorptions. Total and free phage PFUs were enumerated at regular intervals for 2-4 hrs. The latent period was determined as the length of time before a significant increase in phage concentration. The average burst size β was calculated as

$$< \beta >= \frac{< F_2 > - < F_1 >}{< V_1 > - < F_1 >} \tag{4.1}$$

where F is the free phage density before (F_1) and after (F_2) the burst event, and V_1 is the total phage density before the burst event. The averages $< \cdot >$ are taken over multiple time points as well as experiment replicates. All infections were performed with strains in mid-exponential phase.

4.3.5 Community experiment

The community experiment was performed in triplicate with all 5 host strains and all 5 phages together. Bacterial strains were inoculated and transferred individually; transfer cultures were pelleted at 4,000g for 10 min and 4e8 cells of each strain (2e9 total cells) were added to 200mL PC growth medium in each of 6x 2L flasks. Phages were then added at an MOI of 0.1 (4e7 each phage, 2e8 total phages) to 3 flasks while 3 contained no phage, and samples were taken every 35 min for 15 hrs 45 min. At each time point, we sampled for OD600, CFUs, metagenomic analysis (metaG), virome analysis, intracellular qPCR quantification (qINT), and extracellular qPCR quantification (qEXT). For metaG and qINT samples, 1mL was pelleted at 10,000g for 5 min and flash frozen. For virome and qEXT samples, 1mL was 0.2um filtered to remove cells and stored at 4 degrees C.

4.3.6 PSA HP1 community experiment

Experiments were conducted with all 5 bacterial strains and phage PSA-HP1 to determine whether the burst size of this phage was affected by the presence of multiple bacterial strains. This experiment was set up as the community experiment above except that only phage PSA-HP1 was added at an MOI of 0.1 (1e7 phages/mL). After a 15 min adsorption period, the infections were diluted 1:100 in PC growth medium to synchronize the infections. Total and free phage PFUs were enumerated every 10-20 min for 160 min. Latent period and average burst size were determined as above (see Experimental methods, One-step experiments).

4.3.7 Quantification of strain densities with qPCR

Primers for qPCR were designed to amplify 75-150bp portions of each bacterial strain or phage, with negligible amplification from other members of the community. The primers were designed using the complete genomes of each bacterial strain or phage. All host and phage genomes are sequenced and publicly available but 2 of the host genomes were not complete (*Cellulophaga baltica* 4 and *Pseudoalteromonas* sp. H100) and existed in genome fragments. To rectify this, we applied long read sequencing to resequence and complete those two genomes. In addition, we also resequenced *Pseudoalteromonas* sp. 13-15 since primer design was difficult for the two *Pseudoalteromonas* host strains as they are 99.9% identical genomically.

Primer pairs were tested for efficiency and mis-priming, and we used only primers with >85% efficiency and <10 copies/ul amplification from other community members. DNA was extracted from qINT samples using the DNeasy Blood and Tissue kit (Qiagen) following the manufacturer's instructions; qEXT samples were used as-is. qPCR was performed on an Eco Real-Time PCR System (Illumina) with PerfeCTa SYBR Green FastMix Reaction Mix (QuantaBio) in 13ul reactions. Per reaction, we used 6.5ul PerfeCTa master mix, 0.39ul forward primer, 0.39ul reverse primer (see Table C.1), 4.72ul nuclease-free water, and 1ul template. For qINT samples, extracted DNA was used as template, and for qEXT the 0.2um filtrate was used. Reactions were performed in technical duplicates with a standard curve consisting of 5x 10-fold dilution series of known concentration of the target strain or phage, used to calculate target sequence copies/ul. Cycling conditions were as follows: polymerase activation for 5 min at 95 degrees C; 40 cycles of 20 sec at 95 degrees C, 10 sec at primer annealing temperature (see Table C.1), and 20 sec at 72 degrees C, and a 55 degree – 95 degree C melt curve.

4.4 Theory and computational methods

4.4.1 Model for phage-host community ecological dynamics

We model phage-bacteria community ecological dynamics using a system of nonlinear differential equations, where state variables track the density of each strain of susceptible bacteria (S), phage (V), bacteria exposed to phage (E), and bacteria infected by phage (I). By convention, we use the index i to refer to bacteria strains, j for phage strains, and ij for bacteria-phage pairs (e.g. the bacteria strain i infected by phage strain j is denoted I_{ij}). Each bacteria strain, phage strain, and bacteria-phage pair also has unique parameters (life history traits) associated with it (see Table 4.1; see Table C.2 for strain names).

We assume susceptible bacteria are not nutrient-limited and model growth as a simple exponential term. Susceptible bacteria are infected by phage according to the interaction matrix M ($M_{ij} = 1$ if host i is infected by phage j and zero otherwise) in a density-dependent manner with adsorption rate ϕ. We assume that adsorption to a susceptible cell always leads to infection and that infections are 100% efficient. Once infected by a particular phage, a host passes through N_E different stages of infection ("exposed", $E^{(1)}$ to $E^{(N_E)}$) before reaching the final state ("infected", I) and lysing. We assume the transfer through infection stages is sequential and uninterruptible. We scale the transfer across infection stages such that it takes a newly infected host an average time of τ, the latent period, to be lysed (see Theory and computational methods, Latent period distributions). Once lysed, new virions are released with a burst size β. We assume free phage virions may adsorb to any host cell, regardless of its infection state (susceptible, exposed, or infected), although adsorption to an exposed or infected host has no effect other than removing that virion. Here, we assume phage cannot adsorb to cells which they cannot infect. The model

in full is

$$
\begin{aligned}
\dot{S}_i &= \overbrace{r_i S_i}^{\text{growth}} - \overbrace{S_i \sum_j^{N_V} M_{ij}\phi_{ij}V_j}^{\text{new infection}} \\
\dot{E}_{ij}^{(1)} &= \overbrace{M_{ij}\phi_{ij}S_iV_j}^{\text{new infection}} - \overbrace{\frac{N_E+1}{\tau_{ij}}E_{ij}^{(1)}}^{\text{delay}} \\
&\vdots \\
\dot{E}_{ij}^{(N_E)} &= \overbrace{\frac{N_E+1}{\tau_{ij}}E_{ij}^{(N_E-1)}}^{\text{delay}} - \overbrace{\frac{N_E+1}{\tau_{ij}}E_{ij}^{(N_E)}}^{\text{delay}} \\
\dot{I}_{ij} &= \overbrace{\frac{N_E+1}{\tau_{ij}}E_{ij}^{(N_E)}}^{\text{delay}} - \overbrace{\frac{N_E+1}{\tau_{ij}}I_{ij}}^{\text{lysis}} \\
\dot{V}_j &= \overbrace{\sum_i^{N_H} \beta_{ij}\frac{N_E+1}{\tau_{ij}}I_{ij}}^{\text{lysis}} - \overbrace{V_j \sum_i^{N_H} M_{ij}\phi_{ij}N_i}^{\text{adsorption}}
\end{aligned}
\tag{4.2}
$$

where N_i is the total density of bacteria strain i across the susceptible, exposed, and infected classes

$$
\begin{aligned}
N_i &= S_i + \overbrace{\sum_j^{N_V} E_{ij}^{(1)} + \ldots + \sum_j^{N_V} E_{ij}^{(N_E)}}^{\text{all exposed classes}} + \sum_j^{N_V} I_{ij} \\
&= S_i + \sum_k^{N_E}\sum_j^{N_V} E_{ij}^{(k)} + \sum_j^{N_V} I_{ij}
\end{aligned}
\tag{4.3}
$$

Table 4.1: Parameters, hyperparameters, and state variables for the phage-bacteria community model (Equation 4.2). (See Table C.2 for which bacteria and phage strains are assigned to which indices.)

hyperparameter	description
N_H	number of bacterial strains
N_V	number of phage strains
N_E	number of exposed classes

state variable	description
S_i	susceptible bacteria population i
$E_{ij}^{(k)}$	bacteria population i exposed to phage population j in the kth stage of infection
I_{ij}	bacteria population i infected by phage population j
N_i	sum of all of bacteria population i (Eqn Equation 4.3)
V_j	phage population j

parameter	description	units
r_i	bacteria i growth rate	1/hr
M_{ij}	interaction for host i - phage j (boolean)	-
ϕ_{ij}	adsorption rate for host i - phage j	mL/hr
β_{ij}	burst size for host i - phage j	-
τ_{ij}	latent period for host i - phage j	hr

4.4.2 Model for measurement bias

We assume a simple multiplicative error which acts on total bacteria and total phage density and is strain-dependent

$$\begin{aligned} \hat{N}_i &= \epsilon_i N_i \\ \hat{V}_j &= \epsilon_j V_j \end{aligned} \tag{4.4}$$

where $\hat{N}$ and $\hat{V}$ are the measured densities and N and V are the true densities. We define $\epsilon = 1$ to indicate no measurement bias, so that $\epsilon > 1$ indicates overestimation and $\epsilon < 1$ indicates underestimation.

4.4.3 Latent period distributions

The Linear Chain Trick allows us to formulate stochastic state transition models in terms of mean field ordinary differential equations and thus incorporate more nuanced assumptions

about dwell time distributions [86]. In the context of our model, we can manipulate the population-level distribution of latent periods for a given phage-host pair by adjusting the number of exposed classes N_E.

For host i and phage j, we define the latent period $\hat{\tau}_{ij}$ as the duration of time a cell spends from adsorption to lysis. We use $\hat{\tau}_{ij}$ to denote a random variable in contrast with the population average τ_{ij} which is a model parameter (Table 4.1) estimated from experiments (Experimental methods, One-step experiments). In our model, adsorption occurs when a cell transitions from the susceptible state S_i into the first exposed state $E_{ij}^{(1)}$, and lysis occurs when the cell transitions out of the final infection state I_{ij} (Equation 4.2). Because there is only a single influx into a non-branching transition chain, this a simple case of the Linear Chain Trick (see 1 in [86]) with inflow rate $M_{ij}\phi_{ij}S_iV_j$ and transition rate $\dfrac{N_E+1}{\tau_{ij}}$. Thus the latent period $\hat{\tau}_{ij}$ is Erlang distributed

$$\begin{aligned} \hat{\tau}_{ij} &\sim \text{Erlang}(\beta, \alpha) \\ \alpha &= N_E + 1 \\ \beta &= \frac{N_E+1}{\tau_{ij}} \end{aligned} \tag{4.5}$$

where α is the shape parameter and β is the rate parameter. Furthermore, the latent period distribution has mean

$$\langle \hat{\tau_{ij}} \rangle = \frac{\alpha}{\beta} = \tau_{ij} \tag{4.6}$$

4.4.4 Simulating timeseries

We use MATLAB's `ode45()` to numerically integrate the system of differential equations. We set a relative error tolerance of 10^{-8}. We convert all volumes to mL and times to hrs (see Table 4.1 for parameter units). We set initial values for susceptible bacteria S_i and free phage V_j equal to the initial densities measured in the experiment (Figure 4.2 and Table C.3). We simulate timeseries to 15 hours and 45 minutes, the duration of the

community experiment. For visualizations of dynamics, we use a timestep of 3 minutes (316 timepoints). For model-data fits, we use the experimental sample times which were every 35 minutes (27 timepoints, see Experimental methods: Community experiments).

The output of `ode45()` is a $(N_H + N_H \times N_V \times (N_E + 1) + N_V)$ by N_T array where N_T is the number of time points. We obtain the total density N_i for each strain of bacteria i by summing across susceptible, exposed, and infected states (Equation 4.3). For phage, we simply take V_j since we seek free phage density and not phage inside of cells.

To simulate the phage-host model (Equation 4.2) with measurement bias (Equation 4.4), we multiply the total bacteria densities N_i and free phage densities V_j by the corresponding measurement bias term ϵ.

4.4.5 Estimating posterior distributions with Markov-chain Monte Carlo

We use the MATLAB package `mcmcstat` to run the MCMC analyses [87, 88]. We used the default Metropolis-Hastings algorithm for sampling chains.

Possible parameters included in the MCMC runs were r, ϕ, β, η and the measurement bias ϵ (see Table 4.1 for parameter descriptions). We treated each element of each parameter separately (e.g. five separate growth rates r_i for the parameter r). For pairwise parameters (ϕ, β, η), we only included interacting pairs in the MCMC and kept non-interacting pairs fixed at zero. Parameters were sampled in linear space except for the adsorption rates ϕ which were sampled in log space.

For prior distributions, we used Gaussians with means equal to the values measured in the pairwise and single-strain experiments (see Experimental methods: One-step experiments, Adsorption curve experiments, and Doubling time experiments). We used standard deviations equal to the standard deviation within each parameter (for example, the standard deviation across all five growth rates r_i). For β, we decreased the standard deviation of the Gaussian to 70% of the standard deviation across the nine β_{ij}. For measurement bias ϵ, for which no experimental data were available, we chose a non-uniform prior with mean 0 and

standard deviation 0.2 (see Table C.5).

We set minimum and maximum sampling limits for each parameter. For parameters in linear space $(r, \beta, \eta, \epsilon)$, we set the minimum to zero. We set the maximum to ensure biological plausibility while remaining permissive enough to allow for exploration of the majority of the prior distribution. For ϕ which is log-transformed, we set biologically plausible minimum and maximum limits (see Table C.5).

For the error function, we use the sum-of-squares of differences between log-transformed model and data points

$$err = \sum_{t}^{N_T} \sum_{i}^{N_H} \left(\ln \hat{N}_i(t) - \ln N_{i,t} \right)^2 + \sum_{t}^{N_T} \sum_{i}^{N_V} \left(\ln \hat{V}_j(t) - \ln V_{j,t} \right)^2 \tag{4.7}$$

where $\hat{N}_i(t), \hat{V}_j(t)$ are the simulated total host and phage densities accounting for measurement bias (Theory and computational methods, Model for measurement bias) at sample time t. The data points $N_{i,t}, V_{j,t}$ are the qPCR samples for total host genomes and free phage genomes respectively (Experimental methods, Quantification with qPCR).

For starting values, we sampled randomly and uniformly within the corresponding sampling limits. We ran chains several times using different starting values and checked that posterior distributions were consistent. Chains were run for 10,000 steps and the first 2,000 steps, which encapsulated transient dynamics (i.e. the burn-in) for most of the runs, were removed. The remainder of the chain was taken as the posterior distribution.

We checked for chain convergence using two heuristics from the `mcmcstat`package, namely integrated autocorrelation time (IAT) and Geweke's diagnostic [87, 88].

4.4.6 Calculating timeseries envelopes

Timeseries envelopes were calculated by simulating timeseries for an ensemble of parameter sets. To generate the parameter sets, we took slices of the MCMC chain at randomly determined steps, excluding the transient. For this chapter, we drew 8,000 samples, that is,

the entire chain excluding the transient. For parameters not included in the MCMC run, we used values from the original parameter set obtained from experiments (Table C.4). We simulated timeseries for each parameter set using the same model and methods as described above (Theory and computational methods: Simulating timeseries). Then, we calculated percentiles for each time point across the ensemble.

To simulate the most likely timeseries (shown by dark blue lines in figures), we took the median of the chain, excluding the transient. We simulated timeseries as described above, using original parameter set values for parameters not included in the MCMC run (Table C.4).

4.5 Results

4.5.1 Experimental quantification of pairwise phage and bacteria life history traits

We characterized the life history traits of the 5 bacteria and 5 phage strains, including bacterial growth rates and host-specific phage burst sizes, adsorption rates, and latent periods. The bacteria and phage strains comprised *Cellulophaga baltica* (CBA) and *Pseudoalteromonas* (PSA) strains isolated from the Baltic Sea and the North Sea respectively. The combination of strains was chosen such that the 5x5 community infection network would be sparse but non-trivial (8 out of 25 phage-host pairs were previously known to interact), with a range of infection strengths, as per a previously measured quantitative host range network (Figure C.1). All life history trait experiments were performed in identical media, which was designed to allow for growth of all 5 host strains ("PC media"; see Theory and computational methods, Strains and Growth Conditions).

Life history traits for the 5 bacteria and 5 phage strains are summarized in Figure 4.1. Bacterial growth rates were measured for each strain separately (Experimental methods, Doubling time experiments; Figure 4.1a). Adsorption was checked for known non-interacting pairs, including between CBA and PSA (Experimental methods, Adsorption rate experiments). A pair which had initially been categorized as non-interacting via quantitative host

range was found to have significant adsorption (phage PSA HS6 and host PSA H100) and included in subsequent measurements. Adsorption rates were measured for the known interacting pairs (previously 8 of 25, now 9 of 25; Experimental methods, Adsorption rate experiments; Figure 4.1b). Burst sizes and latent periods were measured for the 9 interacting pairs via one-step growth curves (Experimental methods, One-step experiments; Figure 4.1c-d).

As anticipated, the phage-bacteria infection network had non-trivial structure, with 3 of 5 phage strains infecting multiple hosts (phage CBA 18:3, PSA HP1, and PSA HS6). Cross-infection occurred within the two genera (CBA and PSA) but not across (i.e. no CBA phage infected PSA hosts and vice versa). Infection life history traits varied significantly across phage-host pairs, with burst sizes ranging from $\sim$1 (phage PSA HS6 and host PSA 13-15) to $\sim$150 (phage CBA 13:8 and host CBA 4), latent periods ranging from 45 min (phage PSA HP1 and host PSA H100) to 95 min (phage CBA 18:3 and host CBA 18), and adsorption rates spanning an order of magnitude (1e-8 to 1e-7 mL/hr). The infection of host CBA 18 by phage CBA 38:1 was sufficiently inefficient that it was not possible to reliably estimate burst size or latent period. Growth rates also varied across host strains, ranging from 0.19 hr^{-1} (host CBA 4; doubling time of 3.7 hours) to 0.27 hr^{-1} (host PSA H100; doubling time of 2.6 hours). We also aimed to characterize bacterial growth dynamics but did not find evidence that bacteria were nutrient limited or near carrying capacity over the experimental time scales (Figure C.2). In addition, viral particles did not show evidence of decay on the time-scale of the community experiment ($\sim$16 hours). Because life history traits varied significantly among strains, we expected to observe fluctuations in transient dynamics that could potentially be used to infer pairwise interactions and fit nonlinear population models to community data (as in [81, 39, 38, 58, 50, 44, 59, 57, 82]).

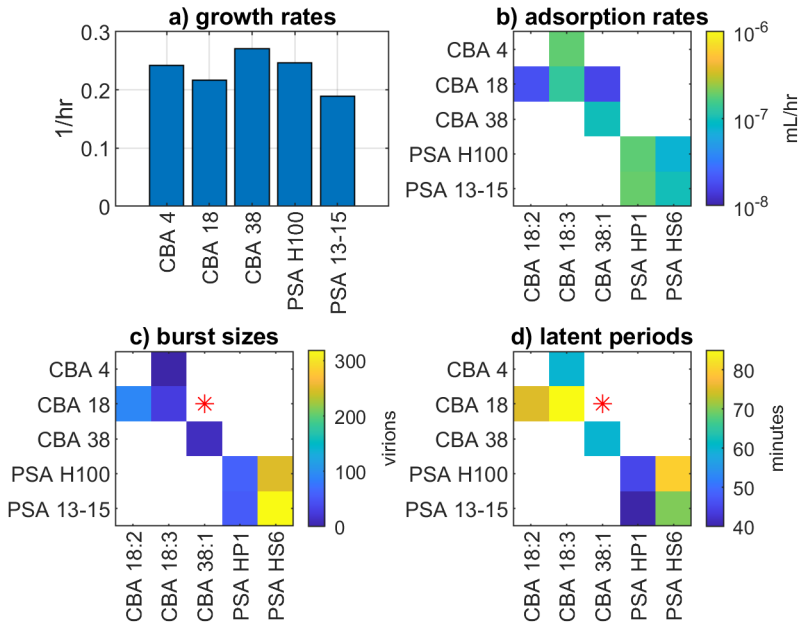

Figure 4.1: Individual and pairwise life history traits for 5 bacteria and 5 phage strains, all measured in PC media. a) Individual bacterial growth rates (Experimental methods, Doubling time experiments). b) Pairwise host-phage adsorption rates (Experimental methods, Adsorption rate experiments). c) Pairwise host-phage burst sizes (Experimental methods, One-step experiments). d) Pairwise host-phage latent periods (Experimental methods, One-step experiments). White denotes no interaction and red stars (on c, d) denote interaction but no data for the phage-host pair.

4.5.2 Time-resolved quantification of phage and bacteria coexistence in a complex community

We conducted an approximately 16-hour synthetic community experiment with 5 bacteria and 5 phage strains (Experimental methods, Community experiment). The community experiment was performed in identical PC media as used in the life history trait experiments (Experimental methods, Strains and growth conditions), allowing for direct comparison of predictions via life history traits and measurements from community experiment (next two Results sections). The community experiment was performed in triplicate. In addition, a separate, phage-free control experiment was performed, in which no lysate was added but otherwise followed the same protocols, allowing for measurement of bacteria-only community dynamics (Figure C.2). Initial densities for bacteria and phage strains were chosen to ensure coexistence for the duration of the experiment based on earlier trial experiments.

Each replicate was sampled every 35 minutes over 15 hours and 45 minutes, resulting in 28 time points for each flask. Multiple measurements were performed on each sample resulting in several data types including: optical density, colony forming units, bacterial metagenomes, metatranscriptomes, and viromes (data available upon request). Measurements were also performed for quantitative PCR ("qPCR") which we present below.

Quantification of bacteria and phage densities was done via qPCR, with separate reactions performed for intracellular and extracellular (filtered) samples (Experimental methods, Quantification of strain densities with qPCR). For the bacterial strains, intracellular qPCR ("qINT") was used as a proxy for density of live cells. For the phage strains, extracellular qPCR ("qEXT") was used as a proxy for density of free virions. Separate qINT reactions were also run for the 5 phage primers to quantify intracellular phage DNA (Figure C.3).

Bacteria and free phage densities from the $\sim$16-hour community experiment are shown in Figure 4.2. All 5 bacteria and 5 phage strains coexist at high density for the duration of the experiment, while simultaneously varying over several orders of magnitude. Distinct dips in viral densities between 0 and 4 hours can be seen indicating an initial round of adsorption for all 5 phage strains. All 5 phage strain densities stabilize by the end of the experiment; and all 5 bacteria strains remain at high density after $\sim$16 hrs. Relatively high sampling rates of 35 min reveal subtle transient dynamics, and all 10 populations fluctuate over several orders of magnitude, all while being highly repeatable across replicates.

4.5.3 Impacts of infection structure and latent period distributions on reconstructing community dynamics

We model community-scale phage-bacteria dynamics with a system of nonlinear differential equations, in which the population density of each bacteria and phage strain is tracked via a unique state variable. The phage in the experiments are known to be purely lytic, and the bacteria are expected to not be nutrient-limited. Thus, we chose a community-level built

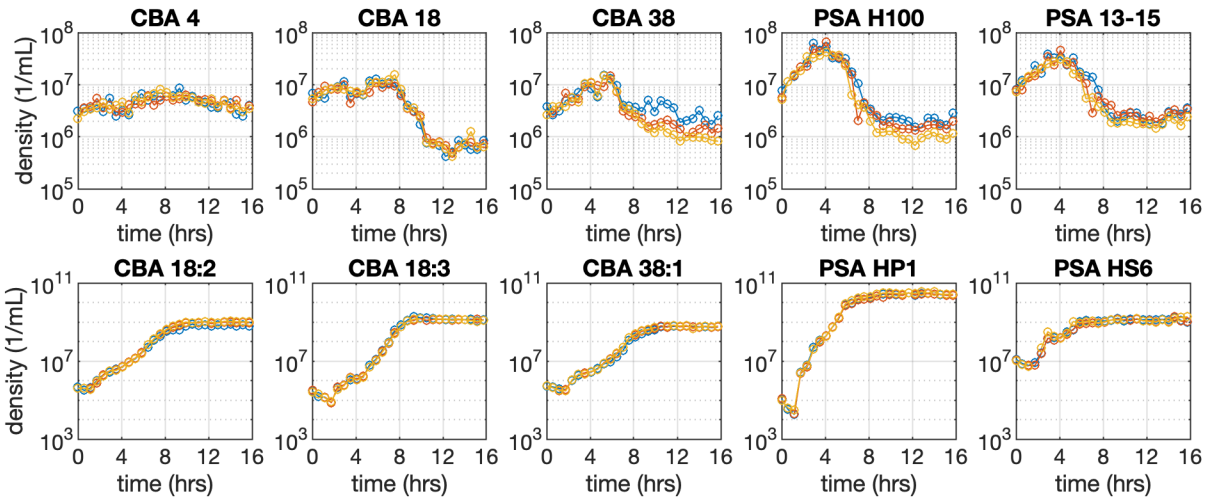

Figure 4.2: Measured densities of bacteria (top row) and free phage (bottom row) in the ∼16-hour community experiment. Measurements were taken every 35 minutes from three replicate community experiments (blue, yellow, and orange circles). Densities were quantified with intracellular and extracellular qPCR for bacteria and phage respectively (Experimental methods, Quantification of strain densities with qPCR).

on prior work [64, 44] that includes lytic infections without explicit accounting of nutrient dynamics (Theory and computational methods, Model for phage-bacteria community dynamics). In addition we introduced a compartmental model of infection in which uninfected bacteria ("susceptible") pass through different stages of infection (which we term "exposed") before reach the final state (which we term "infected") before lysing and release phage progeny (see e.g. [89] for use of such compartmental models in related contexts). In effect, variation in the number of stages and transition rates between stages controls both the mean and variance of the corresponding latent period distribution for each phage-host pair. For this model, the latent period distribution is an Erlang distribution with shape parameter $\alpha = N_E + 1$ and rate parameter $\beta = (N_E + 1)/\tau$ (Theory and computational methods, Latent period distribution, adapted from [86]). Thus, for a fixed latent period, increasing the number of exposed classes N_E preserves the mean but narrows the distribution and shifts the median slightly to the right (see Theory and computational methods, Latent period distribution; and see Figure C.4 for theoretical latent period distributions for 8 phage-host pairs).

All other strain-level parameters for our model correspond to the life history traits measured earlier: bacterial growth rates and host-specific adsorption rates, latent periods, and burst sizes (Results, Experimental quantification of pairwise phage and bacteria life history traits) which we represent as r, ϕ, τ, and β respectively. We use these measured life history traits, without modification, to initially parameterize our model and compare against measured outcomes (see Table 4.1 for parameter descriptions, Table C.2 for strain indices, and Table C.4 for parameter values). We used the measured densities at the first time point from the community experiment for the initial conditions for all simulations (see Table C.3).

In Figure 4.3, we show simulated timeseries for the phage-host community model for $N_E = 1$ and $N_E = 10$ compared against the qPCR data (Figure 4.2). Out of all the parameters, we found that model outcomes were particularly sensitive to deviations in the mean latent period (τ). In Figure 4.3, we show model simulations using the originally measured latent periods overlaid with simulations in which the latent periods are systematically decreased by a factor of 4 or increased by a factor of 4 (see Figure C.5 for sensitivity analysis of both N_E and latent period τ).

The simulated timeseries in Figure 4.3 capture many of the qualitative behaviors of qPCR measurements from the community experiment. Early phage dynamics are well-represented, such as the initial decline from adsorption, although the timing of initial declines tends to be earlier than the experiment. Late viral dynamics saturate at the correct order of magnitude for CBA 18:2 for a latent period multiplier of 4 and for PSA HS6 for all latent period multipliers. Saturation densities for CBA 18:3 and CBA 38:1 are too low by an order of magnitude, and saturation density for PSA HP1 is exceptionally lower than measured values. Simulated bacterial populations tend to crash early on when using experimentally measured life history traits, in contrast to the experimental findings. Notably however, extending the latent periods by a factor of 4 leads to the survival of host strains CBA 4, CBA 18, and CBA 38 for the entirety of the $\sim$16 hour experiment. In general, the x4 lengthened latent periods correspond to the experiment more closely, while baseline and

shortened latent periods lead to early bacterial crashes and early viral saturation. In addition, increasing the number of exposed classes N_E allows for subtly more complex viral dynamics, e.g. more distinct adsorption events (Figure C.6).

While there are notable qualitative similarities between the model simulations and the qPCR measurements, the differences listed above indicate a mismatch between experimental measurements and model predicted outcomes. The model was parameterized using the life history traits measured in individual and pairwise contexts, whereas the qPCR measurements correspond to a 5x5 coexisting community. The community context may lead, on one hand, to modified life history traits (e.g. growth rates decreased or increased by the presence of other strains). On the other hand, entirely new biology or ecology may occur in community contexts, potentially requiring a different model formulation (e.g. lysis inhibition). While our particular model may not capture higher order community effects, it agrees reasonably well with qPCR measurements, such that we consider it to be a "minimally complex" model and a starting point for explicit model-data fits.

4.5.4 Inferring life history traits from community phage and bacteria dynamics

The discrepancies between model simulations and the community experiment suggests the need to use model parameters as priors to formal model-data integration. To do so we used Markov-chain Monte Carlo (MCMC) to fit the qPCR community level measurements to a community phage-bacteria model with measurement error (Theory and computational methods, Model for phage-host community ecological dynamics and Model for measurement error).

In brief, we fixed the number of exposed classes N_E, then chose a set of model parameters to include in the MCMC analysis. For each parameter, we excluded non-interacting pairs, and we calculated prior distributions for the remaining parameters using the experimentally measured life history traits when available. For the error function, we used standard log-likelihood (Equation 4.7). We ran each MCMC for 10,000 steps and discarded the

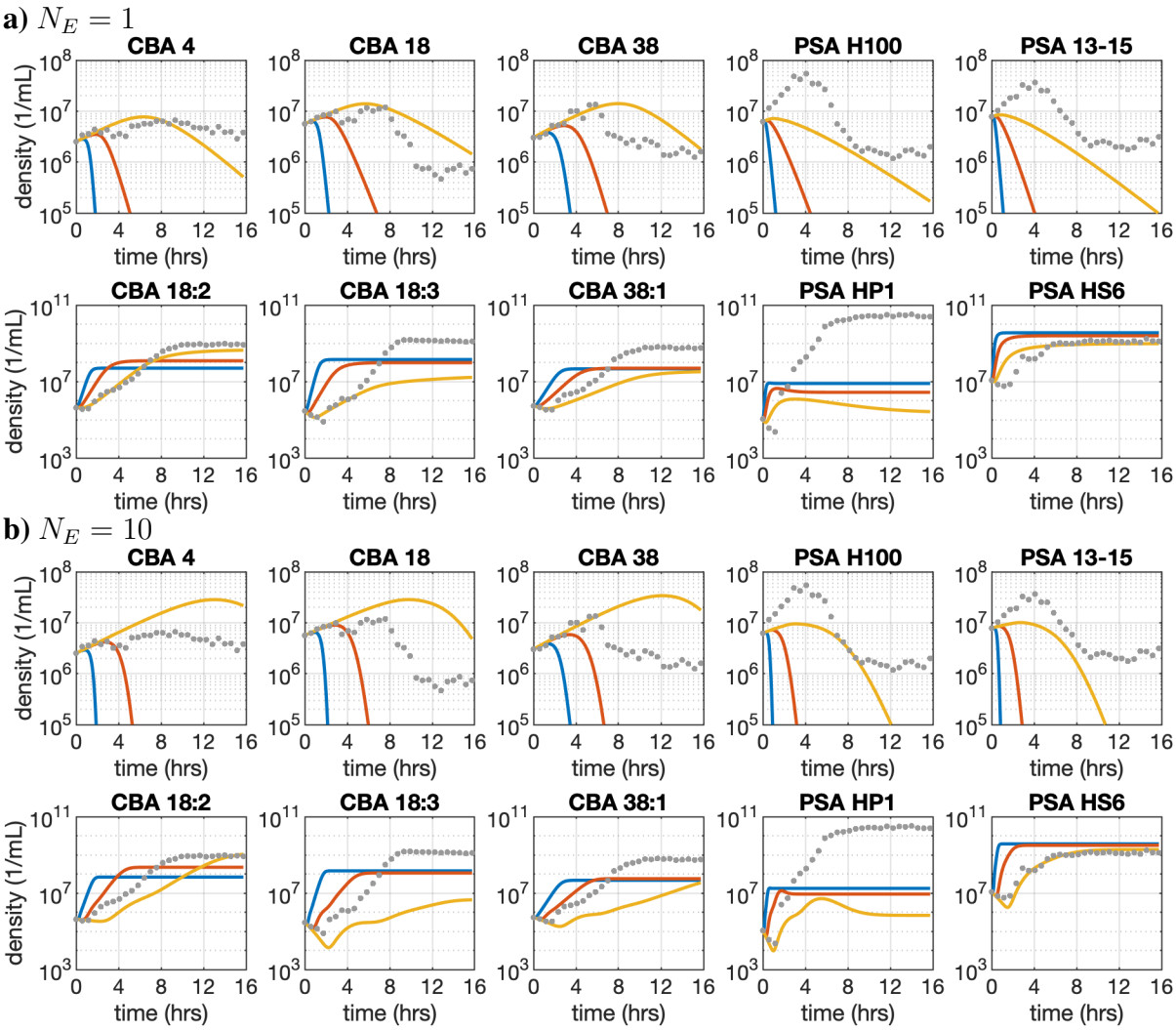

Figure 4.3: Simulated densities for bacteria (top row) and phage (bottom row) over ∼16 hours for the phage-bacteria community model (Equation 4.2) with a) $N_E = 1$ and b) $N_E = 10$. The models are parameterized with experimentally measured life history traits (Figure 4.1; see Table C.4 for all parameter values). Colors show the effect of shortening (blue, tau multiplier = 0.25) or lengthening (yellow, tau multiplier = 4) the baseline (orange, tau multiplier = 1) latent periods of the 9 phage-host pairs. Resulting theoretical latent period distributions are shown in Figure C.4. Average qPCR values from the community experiment are included for reference (grey circles).

transient, resulting in posterior distributions (Theory and computational methods, Markov chain monte carlo). From the posteriors, we simulated timeseries and corresponding envelopes at the 95th percentile (Theory and computational methods, Timeseries envelopes).

In Figure 4.4 and Figure 4.5 we show the result of MCMC model-data fits on the predicted phage and bacterial dynamical outcomes (Equation 4.2 and Equation 4.4) with

$N_E = 5$. The timeseries envelopes are in close agreement with the qPCR measurements (Figure 4.4), drastically improving upon the initial model simulations without any parameter tuning (Figure 4.3b). The root mean squared error of $\log_{10}$-transformed densities ("RMSE-$\log_{10}$", i.e. distance from qPCR data in terms of orders of magnitude) for the fitted model is 0.43 (95% range $0.56 - 1.57$) compared to the RMSE-$\log_{10}$ of 1.44 for the initial model (with x4 lengthened latent periods; Figure 4.3b yellow lines). The model-predicted dynamics of all phage except PSA HP1 are in close agreement to the measured outcomes (RMSE-$\log_{10}$ of $.75$ vs 0.37 for all other phage). Simulations for PSA HP1 are about an order of magnitude too low, which is nonetheless a significant improvement upon initial simulations (RMSE-$\log_{10}$ of 0.75 vs 3.94). Simulations for the 5 hosts capture early dynamics well in both trend and magnitude ($\sim$0-8 hours), but do not recapitulate dynamics at later times ($\sim$8-16 hours). As shown in Figure C.6, models with N_E were able to capture the complex adsorption dynamics for phage at early times (0-6 hours) but did not necessarily lead to better model fits overall as quantified by total squared error (total squared error $7.04 * 10^{21}$ vs. $8.26 * 10^{21}$).

The resulting posterior distributions for the model parameters (burst size β, adsorption rate ϕ, and measurement bias ϵ) are shown in Figure 4.5 (see Appendix C for settings, chains, and convergence heuristics; Table C.5, Figure C.7, Figure C.9, and Figure C.8). Where available, original measured values are marked for β and ϕ; though we do not have measured values for measurement bias ϵ, we use $\epsilon = 1$ as our baseline expectation, indicating no bias. We find that posterior distributions can diverge from the originally measured value in isolation. In particular, some burst sizes (CBA 18-CBA 18:2, CBA 18-CBA 18:3, and PSA H100-PSA HS6), and adsorption rates (CBA 18-CBA 38:1 and CBA 38-CBA 38:1) overlap. Otherwise, burst sizes estimated by MCMC tended to be higher than measured and adsorption rates tended to be lower Measurement bias parameters tended to be slightly greater than 1, with the exception of CBA 18:3.

Quantitative discrepancies between measured values and posterior distributions may

point to real biological, ecological, or environmental differences between the community and pairwise contexts, in addition to model misspecification or non-identifiability of parameter sets. Nonetheless, we found the discrepancy of phage PSA HP1 especially of note, as model-data fits consistently underestimated PSA HP1 density by several orders of magnitude in contrast to the other four phage. PSA HP1's density was also several orders of magnitude higher than the other four phage despite similar initial density (Figure 4.2) and unremarkable infection strength on two hosts (e.g. contrast with phage PSA HS6; Figure 4.1). We note that the posterior distributions for PSA HP1 burst size on both of its hosts are much higher than measured values (Figure 4.5a), which led us to investigate the potential that PSA HP1 infection traits may change between pairwise and community contexts.

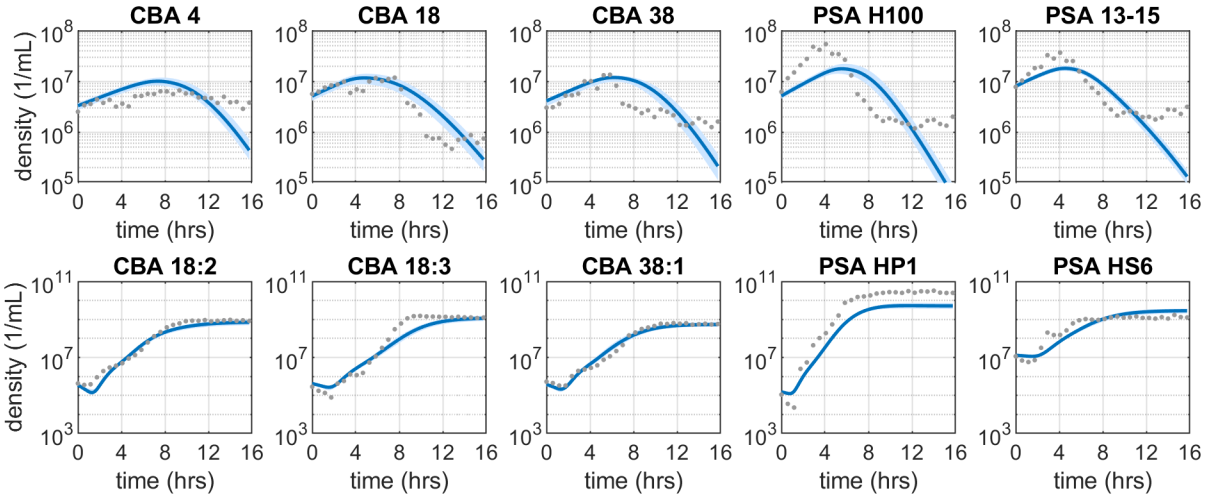

Figure 4.4: Simulated timeseries (dark blue) with 95th percentile envelopes (light blue) for phage-bacteria model with measurement bias (Equation 4.2 and Equation 4.4). Here, $N_E = 5$ (see Figure C.6 for alternate N_E). A subset of parameters (β, ϕ, ϵ) were fit to qPCR data (grey circles) using MCMC (Theory and computational methods, MCMC). Remaining parameters (r, τ) were set to originally measured experimental values (Table C.4). Timeseries and envelopes were calculated as described in the methods (Theory and computational methods, Timeseries envelopes) from the posterior distributions, which are shown in Figure 4.5.

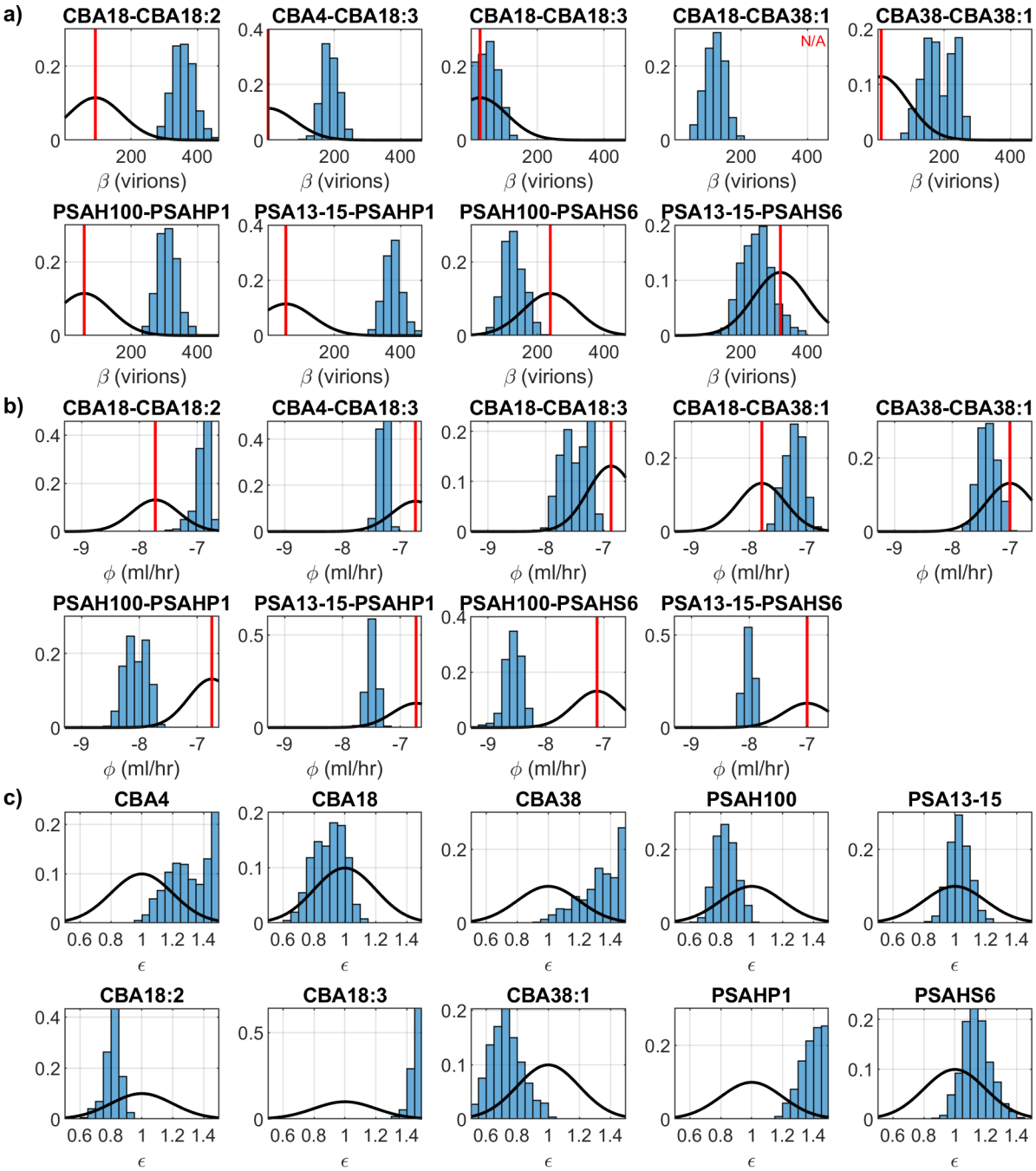

Figure 4.5: Posterior distributions (blue) from the MCMC run corresponding to Figure 4.4. Also shown: prior distributions (black lines) and measured experimental values where available (red lines). a) Burst sizes. b) Adsorption rates. c) Measurement bias. For more details on the MCMC run, see Appendix C: priors and sampling limits (Table C.5), chains (Figure C.7), convergence heuristics (Figure C.8), and covariance plots (Figure C.9).

4.5.5 Ecological impacts on pairwise life history traits

To assess the impact of community context on PSA HP1 life history traits, we performed additional experiments on phage PSA HP1 with all 5 bacteria strains present (Experimental methods, PSA HP1 community experiment). In particular, we performed one-step experiments to measure burst size (Experimental methods, One-step experiments). We contrast the "all host" one-step experiment with the pairwise one-steps in which PSA HP1 was paired with one of its two hosts, PSA H100 or PSA 13-15. The one-step timeseries, and resulting burst sizes, for these three different scenarios are shown in Figure 4.6. Surprisingly, the burst size of phage PSA HP1 increased when all 5 bacteria strains were present relative to the pairwise context in which only one host was present (123, SD=26.2 for all hosts; 54.2, SD=22.9 for PSA H100; 61.4, SD=17.3 for PSA 13-15). One-sided Wilcoxon rank sum tests indicated that the median burst size in the presence of all hosts 5 hosts was greater than the median burst size for PSA H100 (p=0.004) and for PSA 13-15 (p=0.01) across all experiment replicates. As of yet, we have not identified a mechanism underlying the change in PSA HP1's burst size in the community setting.

4.6 Discussion

In this chapter, we presented a 5x5 marine bacteria-phage community with well-characterized individual and pairwise life history traits measured in a common PC medium (Figure 4.1). The experimentally measured phage-host infection network had non-trivial structure, and infection strengths varied significantly across pairs. A $\sim$16 hour community experiment revealed that all 5 phage and 5 host strains were able to coexist at high densities, with notable fluctuations in population density and distinctive transient dynamics (Figure 4.2). Strain-resolved densities were obtained with quantitative PCR at relatively short time intervals (35 minutes), and replicate experiments were highly consistent. Though not presented here, bacterial metagenomes, metatranscriptomes, and viromes were also obtained. These

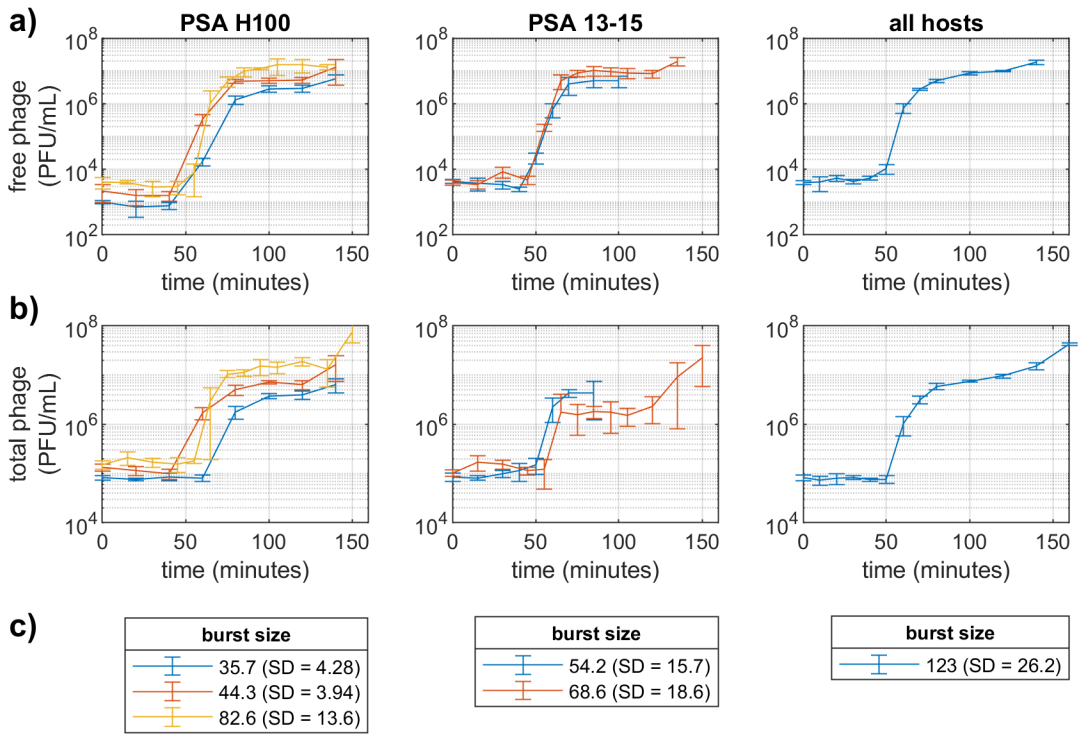

Figure 4.6: One-step experiments for phage PSA HP1 on 3 different combinations of host strains: PSA H100 only (left column), PSA 13-15 only (middle column), and all 5 hosts (right column; see Experimental methods, PSA HP1 community experiment). a) Free phage density and b) total phage density were used to calculate c) burst size (see Equation 4.1 in Experimental methods, One-step experiments). Colors denote unique experiments (3 for PSA H100, 2 for PSA 13-15, and 1 for all hosts). Experiments were run in triplicate; error bars denote standard deviation.

efforts thus resulted in a highly-resolved and data-rich timeseries of community ecological dynamics at a scale of intermediate ecological complexity which will help bridge gaps between culture-based and high-throughput methods.

Dynamical models of phage-bacteria ecology successfully recapitulated the population dynamics observed in the community experiment. Models parameterized with the experimentally measured life history traits recreated some but not all qualitative behaviors (Figure 4.3) and, furthermore, were highly sensitive to variation in latent period and latent period distribution (Figure C.4 and Figure C.5) indicating the importance of model structure and compartmental-type models in representing phage-bacteria infection dynamics. Model-data integration using Markov-chain Monte Carlo further improved agreement

between simulation and data (Figure 4.4) but, in some cases, resulted in large discrepancies between predicted and measured life history traits (Figure 4.5). Additional one-step experiments performed with phage PSA HP1 revealed that burst size changed between pairwise and community contexts (Figure 4.6), possibly indicating higher-order interactions between its two hosts PSA H100 and PSA 13-15 or interactions among the broader community including non-hosts CBA 4, CBA 18, and CBA 38.

In summary, this work establishes a data-rich and highly-resolved dataset for a marine bacteria-phage community of intermediate ecological complexity. Dynamic models of community ecology successfully recapitulated observed dynamics while also predicting context-dependent shifts in some life history traits, which remain to be tested experimentally and interrogated for underlying biological mechanisms. Quantitative discrepancies remain between simulated and observed population dynamics, which may point to biological, ecological, or environmental effects still missing from or misspecified in the model. Additional "-omics" data (bacterial metagenomes, metatranscriptomes, and viromes) from the community experiment present further opportunities for model-data integration and inference of phage-bacteria infection networks. In closing, this work compares different scales of ecological complexity and provides a gold standard phage-bacteria community for reconciling laboratory and high-throughput methods for studying microbial communities.

4.7 Author contributions

MBS and JSW conceptualized and led the project. ARC led theory/computation with support from SJB, DD, DM, and JSW. NS led experiment design/implementation and data acquisition with support from ARC, CLS, MB, GDH, AM, and LC. CLS led sequencing and primer design. ARC wrote the chapter with support from NS, CLS, MBS, and JSW. ARC, NS, CLS, and AD created tables and figures.

4.8 Acknowledgements

We are grateful to the entire BioOceans team for feedback and discussion over the course of this project, which has spanned several years. In particular, we thank Steven Wilhelm, Gary LeCleir, Naomi Gilbert, and Debbie Lindell for thoughtful and spirited engagement. We also thank members from the Weitz Group for helpful feedback and Jessica Irons for administrative support.

This work was supported in part through computing resources provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA.

This work was supported by the National Science Foundation under Grant 1829636 (J.S.W.).

CHAPTER 5

A PRIMER FOR MICROBIOME TIME-SERIES ANALYSIS

Adapted from A R Coenen, S K Hu*, E Luo*, D Muratore*, and J S Weitz, A Primer for Microbiome Time-Series Analysis, Frontiers in Genetics, (2020) [45]. *Equal contributors.*

5.1 Abstract

Time-series can provide critical insights into the structure and function of microbial communities. The analysis of temporal data warrants statistical considerations, distinct from comparative microbiome studies, to address ecological questions. This primer identifies unique challenges and approaches for analyzing microbiome time-series. In doing so, we focus on (1) identifying compositionally similar samples, (2) inferring putative interactions among populations, and (3) detecting periodic signals. We connect theory, code and data via a series of hands-on modules with a motivating biological question centered on marine microbial ecology. The topics of the modules include characterizing shifts in community structure and activity, identifying expression levels with a diel periodic signal, and identifying putative interactions within a complex community. Modules are presented as self-contained, open-access, interactive tutorials in R and Matlab. Throughout, we highlight statistical considerations for dealing with autocorrelated and compositional data, with an eye to improving the robustness of inferences from microbiome time-series. In doing so, we hope that this primer helps to broaden the use of time-series analytic methods within the microbial ecology research community.

5.2 Introduction

Microbiomes encompass biological complexity from molecules to genes, metabolisms, and community ecological interactions. Understanding this complexity can be difficult due to domain- or location- specific challenges in sampling and measurement. The application of sequencing technology has revolutionized almost all disciplines of microbial ecology, by allowing researchers the opportunity to access the diversity, functional capability, evolutionary history, and spatiotemporal dynamics of microbial communities rapidly and at a new level of detail [90, 91]. Increasingly, it is now possible to sample at the time-scale at which those processes occur, resulting in the collection of microbiome time-series data. While such high-resolution sampling opens new avenues of inquiry, it also presents new challenges for analysis [92, 93, 94, 24, 95].

One of the first challenges in analyzing microbiome data is to categorize sequences in terms of taxa or even "species" [96, 97]. Many methods have been developed to perform this categorization [98, 91, 99, 100, 101, 102, 103, 104, 105, 106, 30, 107]. Particular choices used to define species-level units may alter downstream estimations of diversity and other parameters of interest [108, 109, 110]. Indeed, even the procedures for estimating common diversity parameters are impacted by the properties of read count data [111]. However, some definition of taxa is often necessary for characterizing the composition of microbial communities. In this primer, we use the term *taxon* to denote approximately species-level designations such as operational taxonomic unit (OTU) or amplicon sequence variant (ASV).

Once sequences have been categorized to approximate species-level groups, the interpretation of their read count abundances is accompanied by assumptions that violate many standard parametric statistical analyses. For example, zero reads from a sample mapping to a particular taxon is commonplace in microbiome sequence results, yet it typically remains unclear if a zero indicates evidence of absence (e.g., taxon not present in sample,

incapable of transcribing a gene) or absence of evidence (e.g., below detection, inadequate sequencing depth) [112, 94]. In addition, sequence data is compositional, and therefore does not include information on absolute abundances [113]. As a result, compositional data has an intrinsic negative correlation structure, meaning that the increase in relative abundance of one community member necessarily decreases the relative abundances of all other members [114].

The issues of categorization and sampling depth apply to all kinds of microbiome data sets. In particular, temporal autocorrelation presents an additional complexity to microbiome time-series, in that each observation is dependent on the observations previous to it in time. Autocorrelation also precludes the use of many standard statistical techniques, which assume that observations are independent. In Figure 5.1, we show how autocorrelation leads to high incidences of spurious correlations among independent time-series.

Complex microbiome data demand nuanced analysis. In this chapter, we provide a condensed synthesis of principles to guide microbiome time-series analysis in practice. This synthesis builds upon and is complementary to prior efforts that established the importance of analyzing temporal variation for understanding microbial communities (e.g., [115]). Here, we introduce core statistical methods for microbiome time-series analysis as a starting point and suggest further reading on other possible methods. Our process is described in detail via several code tutorials at https://github.com/WeitzGroup/analyzing_microbiome_timeseries that include analytic tools and microbiome time-series data, and provide a software skeleton for the custom analysis of microbiome time-series data. These tutorials include the basics of discovering underlying structure in high-dimensional data via statistical ordination and divisive clustering, nonparametric periodic signal detection in temporal data, and model-based inference of interaction networks using microbiome time-series.

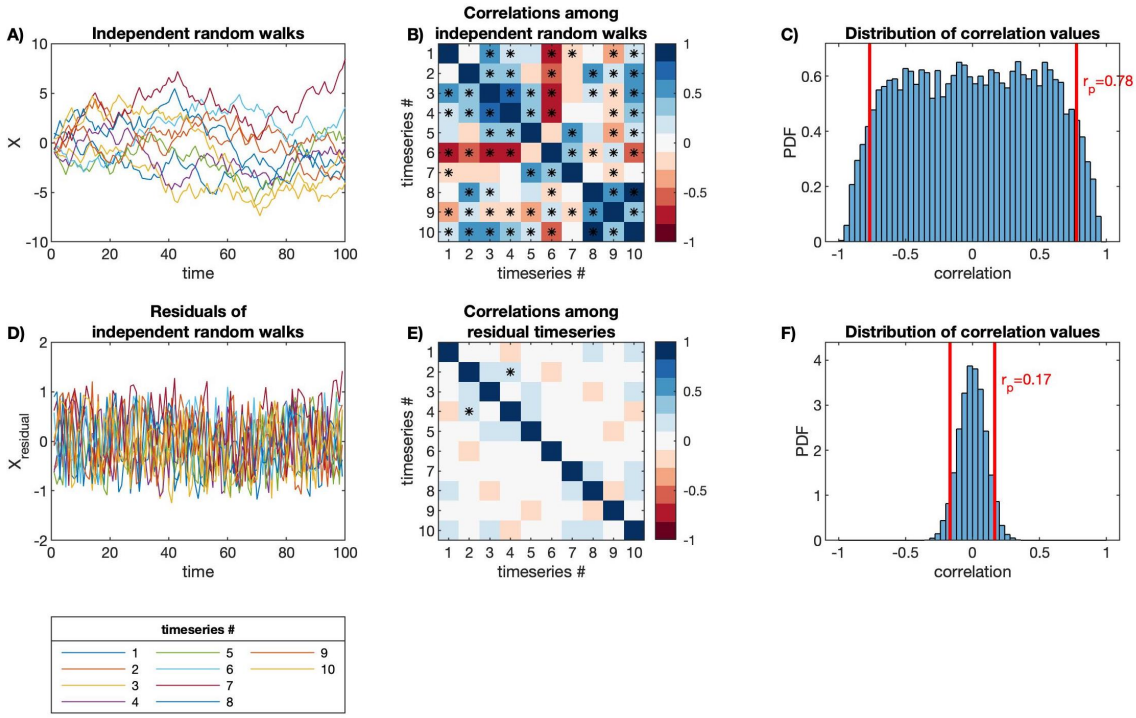

Figure 5.1: Independent random walks yield apparently significant correlations (when evaluated as independent pairs) despite no underlying interactions, in contrast to residuals (i.e., point-to-point differences). A) Time-series of independent random walks, $x_i(t)$. B) Correlation structure of independent random walks. C) Distribution of correlation values for an ensemble of independent random walks, with p-value $= 0.05$ marked (red lines). D) Time-series of the residuals of independent random walks, i.e., $\Delta x_i(t) = x_i(t + \Delta t) - x_i(t)$. E) Correlation structure of residual time-series. F) Distribution of correlation values for the same ensemble as (C) but taken between the residual time-series, with p-value $= 0.05$ marked (red lines).

5.3 Methods

5.3.1 Overview of tutorials

We describe three distinct categories of time-series analyses: clustering, identifying periodicity, and inferring interactions. For each category, we demonstrate analyses that answer an ecologically motivated question (Figure 5.2). Each tutorial emphasizes normalization methods specifically developed for the analysis of compositional data. Each tutorial also addresses challenges related to multiple hypothesis testing, overdetermination, and measurement noise. Interactive, self-contained tutorials that execute the workflows described in the chapter are available in R and Matlab https://github.com/WeitzGroup/analyzing_microbiome_timeseries.

5.3.2 Dataset Sources

For modules I and II, time-series data are derived from an 18S rRNA gene amplicon data set from [116], in which samples were collected at 4 hour intervals for a total of 19 time points (Lagrangian sampling approach). Input data are in the form of sequence count tables, where samples are represented as columns and each row is a taxonomic designation (OTU or transcript ID) with sequence counts or read coverage abundance per taxon (here we use "taxon" as shorthand). The code in each of these modules can be customized for use on other data, although for the purposes of analyzing any temporal-scale variability, samples must be taken at a frequency sufficiently shorter than the temporal scale of interest (e.g., daily temporal variability requires sub-daily sampling, seasonal temporal variability requires sub-seasonal sampling).

For module III, time-series data are simulated from a synthetic microbial community, for which the "true" network is known. The techniques in this module can be applied to time-series data as has been done in a handful of studies [117, 44, 50, 58, 118, 119, 57, 120, 121, 59].

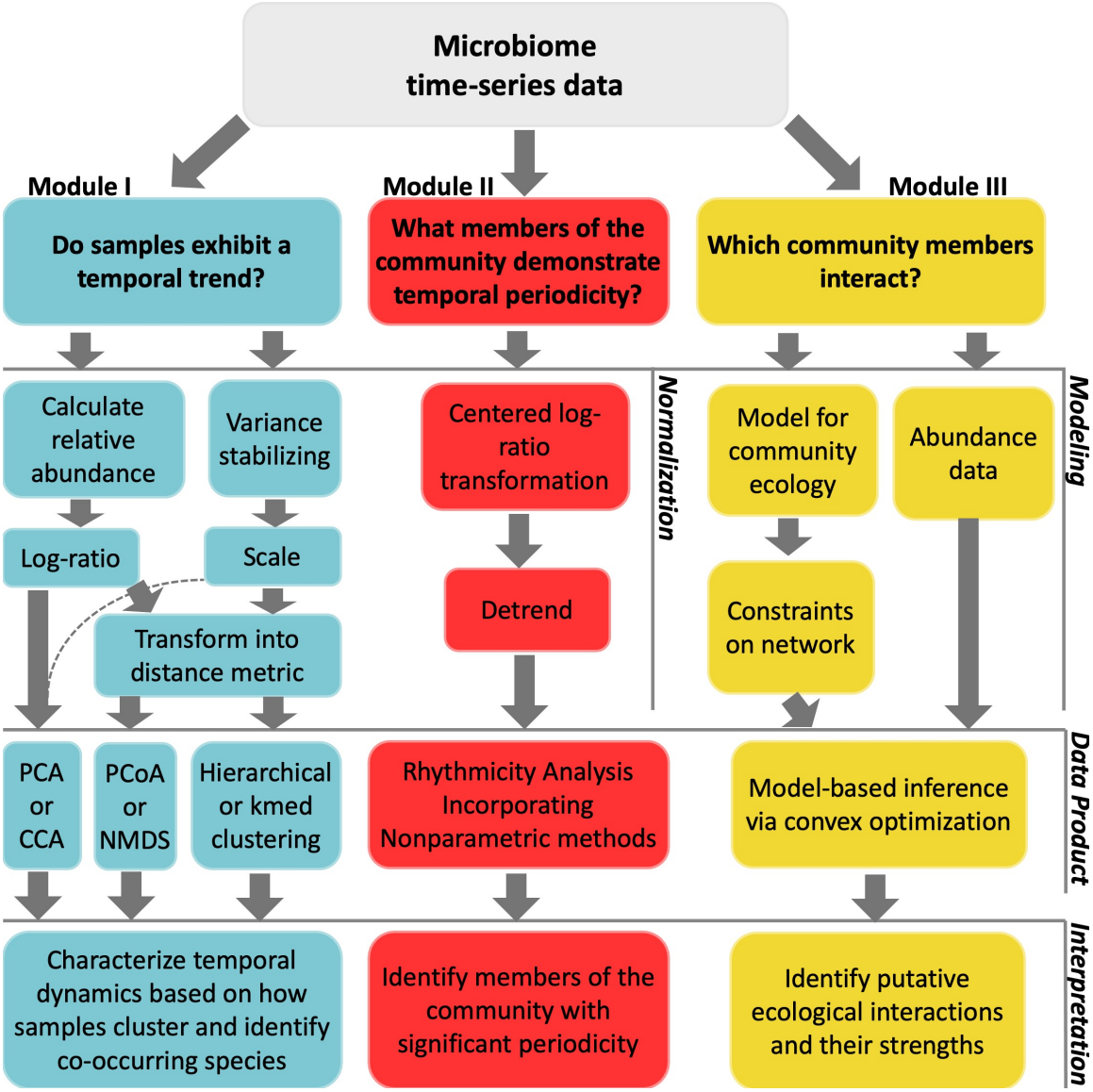

Figure 5.2: Workflow of techniques implemented in each module. The top layer considers questions of interest for a particular study. In the second layer, data normalizations are listed as implemented in module I and module II. For module III, we use synthetic data and instead list modelling inputs. The third layer shows the analytical techniques used in this primer, which we note is not exhaustive. These techniques provide some insight into the initial question asked, as described in the fourth layer.

5.3.3 Normalization

Log-ratio transformations

Microbiome data tend to have three properties: (1) they are sum-constrained (all reads sum to the sequencing depth), (2) they are nonnegative, and (3) they are prone to heteroskedasticity (the variance of the data is not equal across its dynamic range). These attributes of microbiome data violate some underlying assumptions of traditional statistical techniques. Transforming microbiome data into log-ratios [122] can mitigate these problems by stabilizing variance and distributing values over all real numbers, as well as mitigating statistical bias related to sequencing protocols [123].

The simplest log-ratio transformation requires selecting some particular focal variable/taxon in the composition, dividing all other variables in each sample by the abundance of the focal taxon, and taking the natural logarithm. Mathematically:

$$LR_i = ln(x_i) - ln(x_{focal}) \tag{5.1}$$

This kind of log-ratio transformation eliminates negative constrained covariances, but all variables become relative to the abundance of an arbitrary focal taxon. Instead of selecting a focal taxon, the *Centered Log-Ratio Transformation* constructs ratios against the geometric average of community abundances [124].

$$CLR_i = ln(x_i) - \frac{1}{n}\sum_{k=1}^{n} ln(x_k) \tag{5.2}$$

This transformation retains the same dimensionality as the original data, but is also still

sum constrained:

$$\sum_{k=1}^{n} CLR_k = \sum_{k=1}^{n} \left(ln(x_k) - \frac{1}{n} \sum_{k=1}^{n} ln(x_k) \right) \tag{5.3}$$

$$\sum_{k=1}^{n} CLR_k = \sum_{k=1}^{n} ln(x_k) - \frac{n}{n} \sum_{k=1}^{n} ln(x_k) \tag{5.4}$$

$$= 0 \tag{5.5}$$

Log-based transformations require some caution when dealing with data sets with large numbers of zeros, namely because the logarithm of zero is undefined. To overcome this problem, implementations usually employ some pseudocount method, i.e., adding a small number to all observations to make the log of zero observations calculable. Adding a pseudocount disproportionately affects rare taxa, where the magnitudes of differences between samples may be similar to the magnitude of the added pseudocount and therefore obscured [125].

Z-Score Transformation

Another transformation that converts data from counts to a continuous real-valued number is the z-score transformation, achieved by applying this relationship:

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \tag{5.6}$$

where x_i is an observation, μ_x is the mean of population x, and σ_x is the standard deviation of x. Often, μ_x and σ_x are estimated by the sample mean and standard deviation. The z-score is how far, in terms of number of standard deviations, a given observation is from the sample mean [126]. Of note, this transformation places variables of different magnitudes on a scale with the same range.

Variance Stabilizing Transformation

Log-ratio-based transformations in microbiome applications broadly serve the purpose of making the data more compatible with statistical methods that assume continuous/real-valued data and errors with equal variances. Such transformations are necessary because of the heteroscedasticity of sequence count data. A different approach to circumvent heteroscedastic data is to directly estimate a function which describes how the variance in the data increases as a function of the mean. Alternatively, it is possible to use a variance-stabilizing transformation, e.g. as implemented by the DESeq2 software package [127]. While the variance-stabilizing transformation is similar to a log transformation in the case of large counts, it is better suited to deal with zeros and does not rely on a pseudocount.

Distance metric

Multivariate microbiome data is not necessarily easy to summarize or visualize in two or three dimensions. Therefore, to summarize and explore data, we want to recapitulate the high-dimensional properties of the data in fewer dimensions. Such low-dimensional representations are distance-based. A distance matrix is obtained by applying a distance metric to all pairwise combinations of observations. For example, given data matrix X, the Euclidean distance between observations X_i and X_j is:

$$d(X)_{ij} = \sqrt{(x_i - x_j)^2} \tag{5.7}$$

Different metrics measure distance using different attributes of the data (for comprehensive reviews of ecological distance metrics we recommend [128, 129]). For example, only presence/absence of different community members is used to calculate Jaccard distance [130] and unweighted Unifrac [131], which also takes into account phylogenetic relationships between taxa. These metrics can be calculated on count data without transformation, and capture changes in the presence of rare taxa. On the other hand, Euclidean distance empha-

sizes changes in relative composition. Weighted Unifrac distance incorporates phylogenetic information as well as changes in relative abundances. Euclidean distance performed on log-ratio transformed data is analogous to Aitchinson's distance [132], which is recommended for the analysis of the difference of compositions.

In addition to distance metrics, sample-to-sample difference can also be compared by dissimilarities, such as the Bray-Curtis dissimilarity, which is defined between sample i and sample j as:

$$BC_{ij} = 1 - \frac{2\sum_{k=1}^{n} \min(s_{i,k}, s_{j,k})}{\sum_{k=1}^{n} s_{i,k} + \sum_{k=1}^{n} s_{j,k}} \tag{5.8}$$

where n is the total number of unique taxon observed between both samples, and $s_{i,k}$ is the abundance of taxon k in sample i. Bray-Curtis is widely used in ecological studies to measure differences in community composition [133]. A dissimilarity score of 0 means the two samples had identical communities, and a dissimilarity score of 1 means the two samples had no taxa in common. However, Bray-Curtis dissimilarity does not obey the triangle inequality [134], which means that multivariate methods that assume distance matrices as input (e.g., NMDS) may yield uninterpretable results. For example, two samples that each have a Bray-Curtis dissimilarity of 0.05 from a third sample may have a Bray-Curtis dissimilarity of 1 from each other.

5.3.4 Ordination

Covariance-Based Ordination

Statistical ordination can be used to explore multivariate microbiome data. An ordination is a transformation that presents data in a new coordinate system, e.g. making high-dimensional data visualizable in two or three dimensions. Principal Components Analysis (PCA) is a method which selects this coordinate system via the eigendecomposition of the

sample covariance matrix, i.e., which is equivalent to solving the factorization problem:

$$Q_{m\times m} = U_{m\times m} D_{m\times m} U^T_{m\times m}. \tag{5.9}$$

Here, Q is the sample by sample covariance matrix, D is a diagonal matrix containing the eigenvalues of Q, and U is a matrix of the eigenvectors associated with those eigenvalues. For PCA, the eigenvectors (or principal axes) are interpreted as new, uncorrelated variables, which are an orthogonal linear combination of the original m variables [135]. Each of the eigenvalues corresponds to one of the eigenvectors and refers to its magnitude, which is proportional to the amount of variance in the data explained by that eigenvector. To plot a PCA, we select a subset of eigenvectors with the largest associated eigenvalues, apply the linear combination of variables contained in those eigenvectors to each observation, and then plot the observations with the resulting coordinates. Importantly, basic PCA relies on a least-squares approach for solving a linear model with the observed variables, which poorly models heteroscedastic nonnegative data such as taxon sequence counts. Nonlinear PCA [136] is one extension of PCA that can discover more sophisticated correlation structure between observed variables.

Principal Coordinates Analysis (PCoA), based on PCA, is another technique that allows for more flexibility in ordination modeling [113, 129]. PCoA, on the other hand, uses the same procedure as PCA, except on a sample by sample distance matrix is decomposed instead of the sample covariance matrix [137], using the statistical properties of the distances instead of the original observed data. The choice of distance metric allows for the implementation of PCoA on either transformed (in which distance such as euclidean may be suitable) or raw count (in which distance such as Jaccard or unweighted Unifrac may be suitable) microbiome data. For both PCA and PCoA, scaling the data, for example with a z-score transformation, is recommended so that no one variable disproportionately influences the ordination [138].

Nonmetric Multidimensional Scaling

Nonmetric Multidimensional Scaling (NMDS) is an alternative ordination method which forces data to be projected into a prespecified number of dimensions [139]. NMDS projects high-dimensional data into a lower-dimensional space such that all pairwise distances between points are preserved. To implement NMDS, we solve the optimization problem:

$$\hat{X}' = \arg\min \|d(X) - d(X')\|_2 \tag{5.10}$$

where X is the original data matrix and X' is the data in the lower-dimensional space. Here d is a distance metric (see Distance section). Because the sum of pairwise distances is the quantity being minimized by NMDS, this method is strongly affected by outliers, so data should be examined for outliers prior to NMDS ordination. Additionally, unlike PCA and PCoA, where the new sample coordinates are directly related to the measured variables, NMDS coordinates have no meaning outside of their pairwise distances. Another important difference between NMDS and PCA is that the NMDS is enforced to fit the ordination to a fixed number of dimensions, which means the projection is not guaranteed to be a good fit. *Stress* [139] is the quantification of how well the NMDS projection recapitulates the distance structure of the original data:

$$Stress = \sqrt{\frac{\sum(d(X) - d(X'))^2}{\sum d(X)^2}} \tag{5.11}$$

The closer the stress is to 0, the better the NMDS performed.

Clustering

Clustering defines relationships between individual data points, identifying a collection of points that are more similar to each other than members of other groups. Many clustering algorithms have been developed for the analysis of time series data (comprehensively

reviewed in [140]). These algorithms include hierarchical methods, such as agglomerative clustering and k-medoids [141, 95], topological methods such as self-organizing maps [142, 143],and density-based methods such as the DBSCAN algorithm [144]. As a working example, we implement two types of hierarchical distance-based clustering algorithms, the partitioning about medoids (PAM or k-medoid) algorithm [145], and hierarchical agglomerative clustering [146]. A hierarchical clustering method is one which works by partitioning the data into groups with increasingly similar features. The number of groups to divide the taxa into is determined prior to calculation, which begs the question: how many groups? This question can be quantitatively assessed using several indices. A clustering algorithm can be implemented using a range of possible numbers of clusters, and then comparison of these indices will indicate which number has a high degree of fit without over-fitting. These indices can also be used to help choose between clustering algorithms.

One such index is sum of squared differences, which is related to the total amount of uniformity in all clusters, defined as

$$SSE = \sum_{k=0}^{n_{\text{clusters}}} \sum_{i=0}^{n_{\text{members}}} \left(\overbrace{x_{i,k}}^{\text{Cluster member}} - \overbrace{c_k}^{\text{Cluster center}} \right)^2 \tag{5.12}$$

A common heuristic to identifying an optimal number of clusters is to plot SSE vs. k and look for where the curve "elbows", or where the decrease slows down [147, 141] (see clustering tutorial).

Another way to evaluate the efficacy of clustering is via the Calinski-Harabasz index [148], which is the ratio of the between-cluster squared distances to the within-cluster squared differences [147]:

$$CH = \frac{\frac{B(x)}{k-1}}{\frac{W(x)}{n-k}} \tag{5.13}$$

where $B(x)$ is the between cluster sum of square differences, $W(x)$ is the within cluster

sum of square differences, n is the number of taxa, and k is the number of clusters. This index accounts for the number of clusters the data are partitioned into as well as the overall variation in the data as a whole. A large value of CH indicates that the between-cluster differences are much higher than the average differences between the dynamics of any pair of taxa in the data, so a maximum value of CH indicates maximum clustering coherence.

The "Silhouette width" is another index which allows for fine-scale examination of the coherence of individual taxon to their cluster. Silhouette width is therefore helpful for identifying outliers in clusters [147]. The silhouette width for any given clustering of data is calculated for each taxon by taking the ratio of the difference between that taxon's furthest in-cluster neighbor and nearest out-of-cluster neighbor to the maximum of the two, such that

$$SW_i = \frac{\overbrace{min(d(x_i, x_{j \notin C}))}^{\text{sum square diff out of cluster}} - \overbrace{max(d(x_i, x_{j \in C}))}^{\text{sum square diff in cluster}}}{max(min(d(x_i, x_{j \notin C})), max(d(x_i, x_{j \in C})))} \tag{5.14}$$

where C is all taxa in the cluster, and d is the sum square difference operator. The widths can range from -1 to 1. Silhouette widths above 0 indicate taxa which are closer to any of their in-cluster neighbors than any out-of-cluster taxa, so having as many taxa with silhouette widths above 0 as possible is desirable. Any taxon with particularly low silhouette widths compared to the rest of their in-cluster neighbors should be investigated as potential outliers.

5.3.5 Periodicity Analysis

Periodicity analysis reveals whether or not a signal exhibits a cyclical periodic change in abundance. Approaches to identifying periodic signals include parametric methods and non-parameteric methods. The multi-taper method is an example of a parameteric method, which uses autoregression to find periodic signals in low signal-to-noise data [149] (for a software implementation in R https://cran.r-project.org/web/packages/ssa/index.html). Other

examples of parameteric methods include harmonic regression [150, 151], methods based on frequency spectral decomposition [152], and a widely used [153, 116, 154, 155] non-parametric method, 'Rhythmicity Analysis Incorporating Nonparametric methods' (RAIN) [156].

The RAIN method identifies significant periodic signals given a pre-specified period and sampling frequency. RAIN then conducts a series of Mann-Whitney U tests (rank-based difference of means [157]) between time-points in the time-series over the course of one period. For example, one such series of tests might answer the question: are samples at hours 0, 24, 48 higher in rank than the samples at hours 4, 28, 52. Then, the sequence of ranks is examined to determine if there is a consistent rise and fall about a peak time. For this procedure to work, RAIN relies on the assumption that time-series are stationary, or have the same mean across all sampled periods. One way to normalize microbiome time-series to better fit this assumption is detrending, or regression normalization, which removes longer-term temporal effects such as seasonality. A first approximation of non-stationary linear processes can be made by taking the linear regression of all time-points with time as the independent variable, then subtracting this regression from the time-series. This operation stabilizes the data to have a similar mean across all local windows.

In order to assess periodicity for an entire microbial community, we may conduct many hypothesis tests. The more tests that are performed at once, the higher the probability of finding a low p-value due to chance alone [158]. Some form of multiple testing correction is therefore encouraged. False Discovery Rate (FDR) based methods are recommended for high-throughput biological data over more stringent Familywise Error Rate corrections [159, 160]. The method employed here is the Benjamini-Hochberg step-up procedure [161] (for graphical demonstration see the 'periodicity' tutorial in the associated software package). P-values are ranked from smallest to largest, and all null hypotheses are sequentially rejected until test k where:

$$p_k \geq \frac{k}{m}\alpha \tag{5.15}$$

where m is the total number of tests conducted, and α is the desired false discovery rate amongst rejected null hypotheses. Alternative p-value adjustment methods designed for sequencing data have been proposed [162] which take into account correlation between tests, although simulations [163] demonstrate that for moderate effect sizes, methods such as Benjamini-Hochberg generally control false discoveries as expected, if not slightly more conservatively.

5.3.6 Inferring interactions

Model specification of ecological dynamics

Inferring interactions using a model-based approach requires the specification of ecological (or eco-evolutionary) dynamics. Model specification requires extensive knowledge of the system of interest. Furthermore, models can be specified at different levels of abstraction regarding taxonomic resolution (e.g. [164]) and biological mechanisms (e.g. [165]), leading to challenges in interpretability [166]. Alternatively, data-driven identification of dynamical systems is an active area of research (e.g. [82, 167, 168]), providing a possible way forward when an appropriate model is not known *a priori*.

Currently, widely used models include some variation of Lotka-Volterra dynamics where each taxon is represented as a population whose abundances vary in time given density-dependent feedback with other populations [117, 44, 50, 58, 118, 119, 57, 120, 121, 59]. Here, we focus on a variant of this class of problem, i.e., virus-microbe dynamics.

The microbe-virus ecological dynamics are modeled via a system of differential equations

$$\dot{H}_i = r_i H_i \left(1 - \frac{1}{K} \sum_{i'}^{N_H} H_{i'}\right) - H_i \sum_{j}^{N_V} M_{ij} \phi_{ij} V_j \tag{5.16}$$

$$\dot{V}_j = V_j \sum_{i}^{N_H} M_{ij} \phi_{ij} \beta_{ij} H_i - m_j V_j \tag{5.17}$$

where H_i and V_j denote the densities of host (i.e. microbe) type i and virus type j as they change over time. There are N_H host types and N_V virus types, each defined by their life history traits: growth rate r_i for host type i, decay rate m_j for virus type j, and a community-wide host carrying capacity K. The interactions between hosts and viruses are modeled as antagonistic infections culminating in the lysis (i.e., death) of the host cell and release of new viruses. For each pair host type i and virus type j, the infection is quantified by the interaction coefficient M_{ij}, adsorption rate ϕ_{ij} and burst size β_{ij}. The interaction coefficient is either 1 (the virus infects the host) or 0 (the virus does not infect the host) [169, 170].

We randomly sample the life history traits and interaction parameters such that they are biologically plausible and guarantee local coexistence of all host and virus types (as described in [44]). We simulate the time-series of the resulting dynamical system using ODE45 in Matlab.

Objective function for model-based inference

We seek the interaction network that minimizes the difference between observed dynamics in densities and those predicted by the dynamical model. We use the virus equations (Equation 5.17) to derive the objective function

$$\min \quad \left\| W - \begin{pmatrix} \tilde{M}^T & -\vec{m} \end{pmatrix} \begin{pmatrix} H \\ \vec{1} \end{pmatrix} \right\|_2 + \lambda \left\| \tilde{M} \right\|_1 \tag{5.18}$$

$$F \text{subject to} \quad \tilde{M}_{ij} > 0 \tag{5.19}$$

$$m_j > 0 \tag{5.20}$$

where W_{jk} is the per-capita derivative estimate of virus type j at sampled time t_k, H_{ik} is the density of host type i at sampled time t_k, $\tilde{M}_{ij}^T = M_{ij}\phi_{ij}\beta_{ij}$ is the weighted infection coefficient between virus type j and host type i, and m_j is the decay rate of virus type j (as

described in [44]). We seek to estimate the unknown weighted infection network $\tilde{M}$, using sampled densities of hosts H and viruses W over time.

To prevent over-fitting, we introduce a hyper-parameter λ, which can be tuned to control the sparsity of the inferred network $\tilde{M}$. Other approaches can also be used to identify a balance between goodness of fit and model complexity, such as k-crossfold validation or information criterion (e.g. AIC). For an example of using k-crossfold validation, see [57].

Interaction inference via convex optimization

In practice, we can solve the minimization problem (Equation 5.20) and infer the interaction network $\tilde{M}$ using convex optimization. Convex optimization is a well-developed technology for efficiently and accurately solving minimization problems of a particular form which are guaranteed to have a global minimum. Here, we use a freely available third-party software package for Matlab available for download at http://cvxr.com/cvx/ [60, 171] (also available for implementation in Python at https://www.cvxpy.org [172, 173]). The details of implementation are described in [44] and in the accompanying code tutorial.

In addition to convex optimization, there are many methods for inferring the interaction network, and dynamical systems parameters in general, from time-series. Two recent examples include MCMC fitting [174, 175] and Tikhonov regularization [57].

5.4 Results and Discussion

5.4.1 Exploring Shifts in Daily Protistan Community Activity

The North Pacific Subtropical Gyre (NPSG) is widely studied as a model ocean ecosystem. Near the surface, the NPSG undergoes strong daily changes in light input. Abundant microorganisms in the NPSG surface community, such as the cyanobacteria *Prochlorococcus* and *Crocosphaera*, adapt metabolic activities such as cell growth and division to particular times of day [176, 154, 177]. However, the extent to which these daily cycles and the timings of particular metabolic activities extend to protistan members of the NPSG sur-

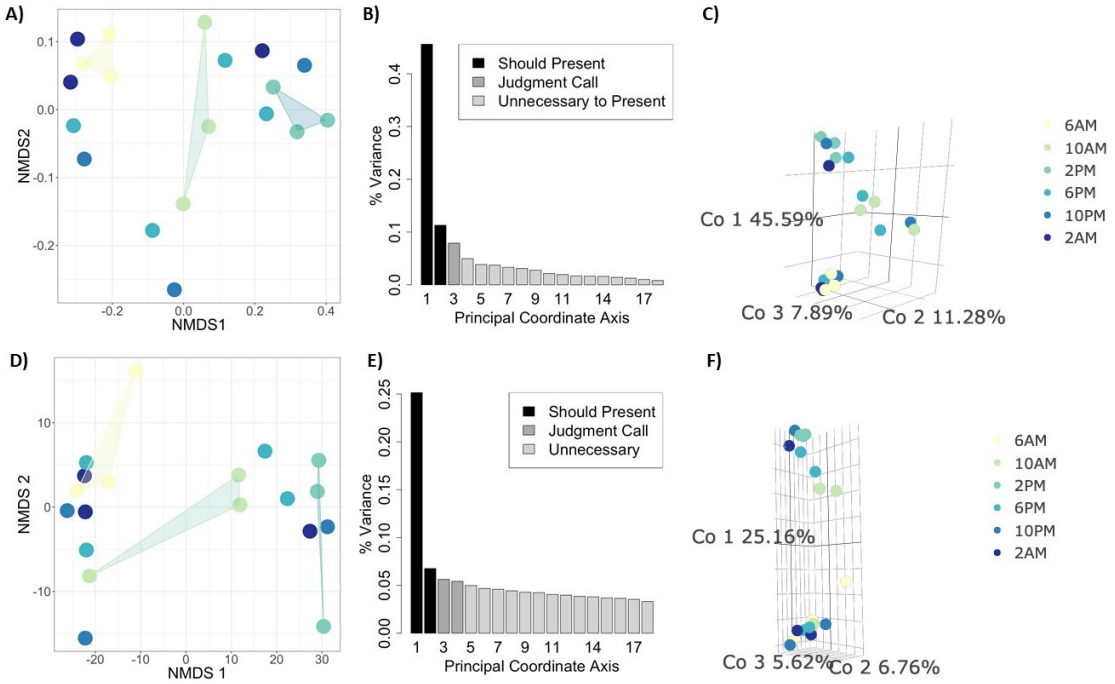

Figure 5.3: Comparing statistical ordination techniques for 18S community compositions across samples. Top row: Ordinations using Jaccard distance for comparison of presence/absence of community members between samples. Bottom row: Ordinations using Euclidean distance on isometric log-ratio transformed data. (A,D) Non-metric Multidimensional Scaling (NMDS) projection in two dimensions, arbitrary units. Convex hulls have been drawn to emphasize ordinal separation of 6AM (yellow), 10AM (light green), and 2PM (teal) samples. (B,E) Scree plots for PCoA ordinations. Each bar corresponds to one axis of the PCoA, the height is proportional to the amount of variance explained by that axis. We decided the first 3 axes were necessary to summarize the data in these cases (explaining a total of (B) 64.76% and (E) 37.54% of the variance). Shading of bars indicate our interpretations of which axes are important to show (black), which are unimportant (light grey), and which are intermediate cases (medium grey). (C,F) PCoA ordinations using the selected axes after scree plot examination. Each point is one sample, the color of the point indicates the time of day at which the sample was taken (colors correspond to NMDS projections).

face ecosystem remains less characterized. To this end, we examined an 18S rRNA gene diel dataset from a summer 2015 cruise sampled every 4 hours for 3 days on a Lagrangian track near Station ALOHA [116]. In this expedition, both rRNA and rDNA were sampled to explore differences in metabolic activity for particular community members at different times of day [178]. Previous work [116] found shifts in the metabolically active protis-

tan community, including phototrophic chlorophytes and haptophytes as well as parasitic Syndiniales.

In this analysis, we asked whether or not the metabolically active component of the microbial community is unique to different times of day. Therefore, we focused specifically on the 18S rRNA gene data as a proxy for overall functional activity of protistan taxa [178, 179, 180]. We used statistical ordination to explore underlying sample covariance. Samples that appear near each other in a statistical ordination have similar multivariate structure. In the clustering tutorial we present several methods for performing ordination, e.g., NMDS and PCoA (see Methods: Ordination). In Figure 5.3 (B) and (C), we construct a PCoA using Jaccard distance to emphasize changes in presence/absence of rRNA signatures, and find that the first 3 Principal Coordinates explain 64.76% of the variation amongst all samples. Samples from 2PM and 6AM strongly differentiate along the first coordinate axis, while samples at 10AM settle between them. The ordination suggests that the taxa which are transcribing the 18S rRNA gene at 2PM are fairly distinct from those transcribing at 6AM, while 10AM is intermediate between the two. We also constructed a corresponding NMDS ordination using the same distance matrix that we constrained to two dimensions. The pattern of separation between 2PM and 6PM is maintained in this projection, reinforcing its importance as an underlying structural feature of these data. Next, we constructed an additional PCoA ordination on the Euclidean distance matrix of isometric log-ratio transformed 18S rRNA counts (see clustering tutorial for implementation). We select the isometric log-ratio transformation to alleviate the constraint of compositionality and to scale the data to a similar range of magnitudes, making Euclidean distance a suitable choice of distance metric. As seen in the scree plot in Figure 5.3 (E), while the first Principal Coordinate explained about 25% of the variation between samples, the following four Principal Coordinates each explained around 5% of the variation. Despite the low proportion of total variance explained, strong separation emerges between 2PM and 6AM samples along the largest coordinate axis. This ordination qualitatively agrees with a corresponding NMDS

ordination (Figure 5.3 (D)) forced into two dimensions.

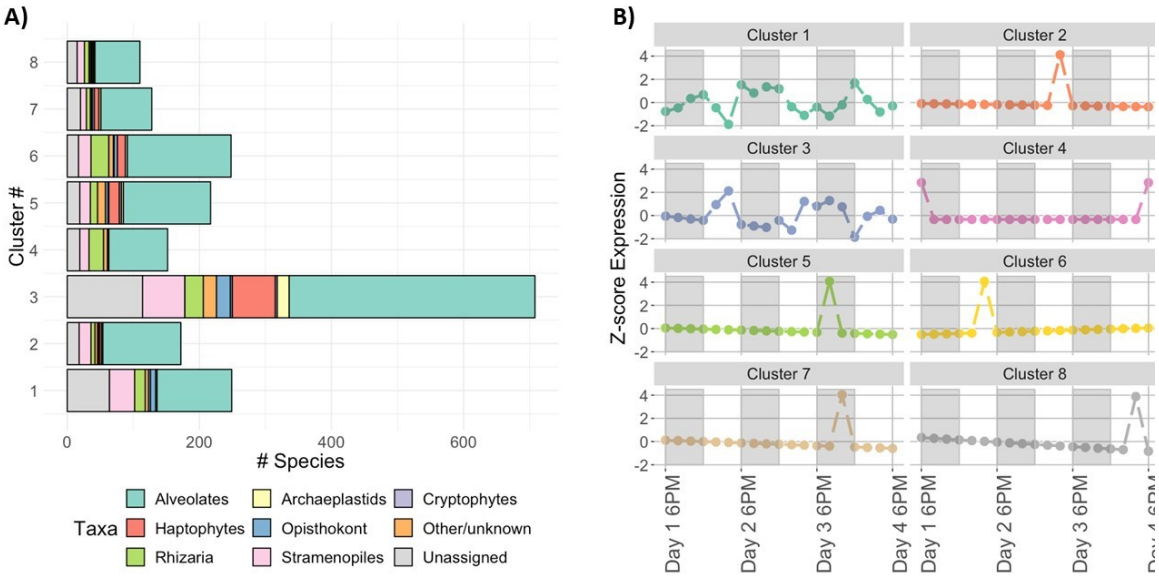

Figure 5.4: Characterization of protist clusters. (A) Cluster membership based on the phylum or class level protistan taxonomy. The 'Other/unknown' category includes sequences with non-specific identity such as 'uncultured eukaryote' and 'Unassigned' denotes sequences with no taxonomic hit ($< 90\%$ similar to reference database). (B) Representative taxon time-series for each cluster. Y-axis is z-score (see Methods: Normalizations), so a value of 0 corresponds to mean expression level. White and shaded regions represent samples taken during the light (white) dark cycle (shaded).

Noting the differences in active community members between 2PM and 6AM, we identified co-occurring taxa by clustering their temporal dynamics after variance-stabilization and scaling normalizations (see clustering tutorial for discussion). Based on comparisons of sum squared errors and the CH index introduced in Methods, we opted to divide the OTUs into eight clusters (Figure 5.4 for composition and representative temporal signature, tutorial for details on cluster selection). After comparing cluster evaluation metrics for hierarchical agglomerative clustering and a k-medoids algorithm, we conducted this clustering with k-medoids (see clustering tutorial for implementation). This method allows us to identify the time-series of the median taxon for each cluster as a representative shape for the cluster's temporal dynamics. We observe 2PM peaks associated with clusters 2,3,6, and 8 and increased nighttime expression levels in cluster 1. These temporal patterns co-

incide with those surmised during our exploratory ordination of the community sampled at each time point (where 2PM and 6AM samples formed distinct clusters, Figure 5.3). Upon closer inspection of cluster membership (bar plots in Figure 5.4 (A)), we find cluster 3 contains 65/105 (62%) of haptophyte OTUs and 18/33 (55%) of archaeplastids, including members of chlorophyta.

These results suggest temporal niche partitioning within the complex protistan community, consistent with the findings of [116]. By clustering results with respect to temporal patterns, we were able to parse the complex community to reveal the identities of key taxonomic groups driving the observed temporal patterns. The taxonomic composition of cluster 3 was made up of haptophytes and chlorophytes. Photosynthetic chlorophytes have previously been found to be correlated with the light cycle [176, 181] and the temporal pattern found in [116] was similar to the standardized expression level (Figure 5.4B), as was the inferred relative metabolic activity of haptophytes.

5.4.2 Identifying Protists with Diel Periodicity in 18S Expression Levels

The metabolic activity of microbes is a critical aspect of the basis of marine food webs [182]. In the euphotic zone, microbial populations are inherently linked to the light cycle as the energy source for metabolism. Identifying diel patterns in protists is particularly interesting due to widespread mixotrophy, where a mixotroph may ingest prey during periods of limiting inorganic nutrients or light [183, 184, 185]. Additionally, protistan species encompass a wide range of cell sizes, thus the synchronization of light among photoautotrophs may reflect species-specific differences in nutrient uptake strategies [186, 187]. Based on the observation of sample differentiation between the middle of the day (2PM) and dawn (6AM) from exploratory ordination and clustering analyses described in 4.1, we further investigated the hypothesis that some protists may exhibit a 24-hour periodicity in their 18S rRNA gene expression levels.

The high-resolution nature of the sequencing effort in this study enabled us to ask which

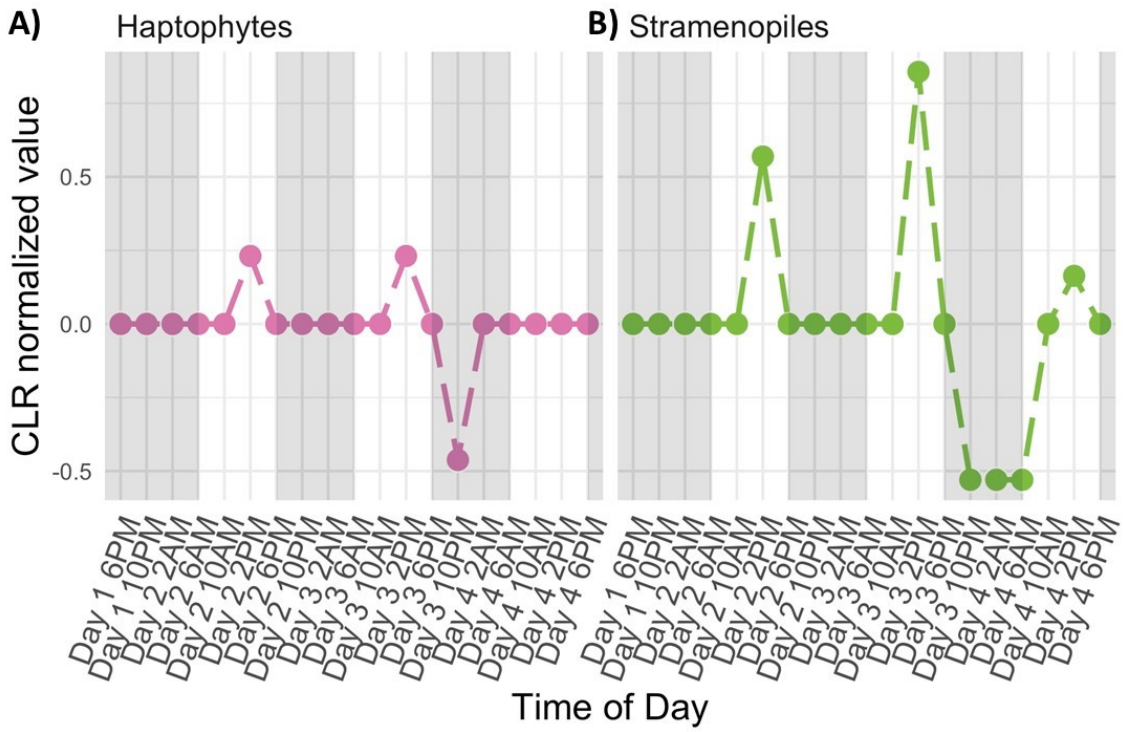

Figure 5.5: Centered Log Ratio (CLR)-transformed, detrended 18S rRNA gene levels (y-axes) over time (x-axes) for a subset of OTUs found to have significant diel periodicity (RAIN analysis). A value of 0 denotes the mean expression level for a given OTU. Included OTUs belong to the (A) Haptophyte and (B) Stramenopile groups. White and shaded regions represent samples taken during the light (white) dark cycle (shaded).

members of the protistan community had 24-hour periodic signals. Following normalization (CLR, Eq 2) and detrending to center mean expression levels across the entire time series (see Periodicity tutorial and Methods: Periodicity Analysis), we used RAIN to assess the periodic nature of each OTU over time. Results from RAIN analysis reported p-values for each OTU at the specified period as well as estimates of peak phase and shape. The null hypothesis tested by RAIN is that the observations do not consistently increase, then decrease (or vice-versa) once over the course of a period. Rejecting the null hypothesis, then, asserts a time-series has one peak during the specified period. To determine which OTUs were found to have significant periodicity, we rejected the null hypothesis at 5% FDR level (Eq. 13). Figure 5.5 illustrates examples of two protistan OTUs with significant diel periodicity, a haptophyte and stramenopile. Trends in CLR normalized values for each

OTU indicated that there was a repeated and temporally coordinated relative increased in the metabolic activity of both taxa at 2PM Figure 5.5. Both groups have previously been found to respond to day-night environmental cues, which is also supported by [116].

Identities of OTUs found to have significant diel periodicity included taxa with known phototrophic and/or heterotrophic feeding strategies. This suggests that taxa with diel changes in metabolic activity may be responding to light or availability of prey. More specifically, several known phototrophs or mixotrophs, including dinoflagellates, haptophytes, and stramenopiles were found to have significant diel periodicity. Interestingly, there were a number of OTUs identified as belonging to the Syndiniales group (Alveolates) which are obligate parasites. Diel rhythmicity among these parasites suggests that they may be temporally coordinated to hosts that also have a periodic signal, which includes dinoflagellates.

5.4.3 Inferring interactions in a synthetic microbial community

The goal of an inference method is to quantify ecological interactions between pairs of populations or taxonomic designation of interest. The result of such analysis is an interaction network for the community of interest. In the context of microbial communities, "interaction" can be broadly defined and include, for example, direct competition for a nutrient, toxin-mediated attacks, or cooperation via exchange of secondary metabolites. Besides pairwise interactions between microbes, other interactions may be of interest, such as higher-order interactions (e.g. three-way microbial exchanges [188, 189, 50]), pressures from other trophic levels (e.g. grazers, viruses), or driving via environmental variables (e.g. antibiotics, nutrient flux). Inferring interaction networks is a challenging task, in part due to autocorrelation inherent in time-series data. Time-series which are highly autocorrelated appear correlated with one another, even when there is no underlying causal relationship (see Figure 5.1). This leads to high false-positive rates of inferred interactions, particularly for correlation-based inference methods [41, 190, 191, 43, 24, 192, 193].

Model-based inference methods are built from dynamical models of micarobial community ecology. As such, temporal variation and structure is incorporated into any model-based inference framework, accounting for potentially difficult statistical properties such as autocorrelation. Model-based inference has been shown to perform favorably in *in silico* studies [117, 44, 50, 58, 118, 119, 57, 120, 121, 59]. Major challenges remain for implementing model-based inference in practice, including requirements of high time-resolution data and a detailed understanding of the biological and ecological mechanisms at play in order to correctly specify the underlying model. Futhermore, evaluating accuracy of inferred networks remains difficult, in part because different networks can produce similar patterns of ecological dynamics [166]. Despite challenges, model-based inference has shown potential to accurately infer interaction networks in a computationally efficient and scalable manner (see one such application in [57]).

Here, we demonstrate the use of a model-based inference method on a synthetic microbial community with viruses (methods and code adapted from [44]). We use a synthetic community so that the inferred network can be compared to the original, "ground-truth" network. Using our model for microbe-virus ecological dynamics (Equation 5.17), we simulate population time-series of the community over the course of several days. We sample the simulated time-series to use as data inputs into the minimization problem (Equation 5.20), from which we estimate the weighted microbe-virus infection network $\tilde{M}$. Simulated time-series, data inputs, original and reconstructed networks are shown in Figure 5.6). As shown, the reconstructed network closely resembles the original, with only minor quantitative differences (i.e. in the strengths of the interactions). We caution that the choice (and parameterization) of ecological dynamics is critical to developing a model-based approach, for alternative examples see [117, 44, 50, 58, 118, 119, 57, 120, 121, 59].

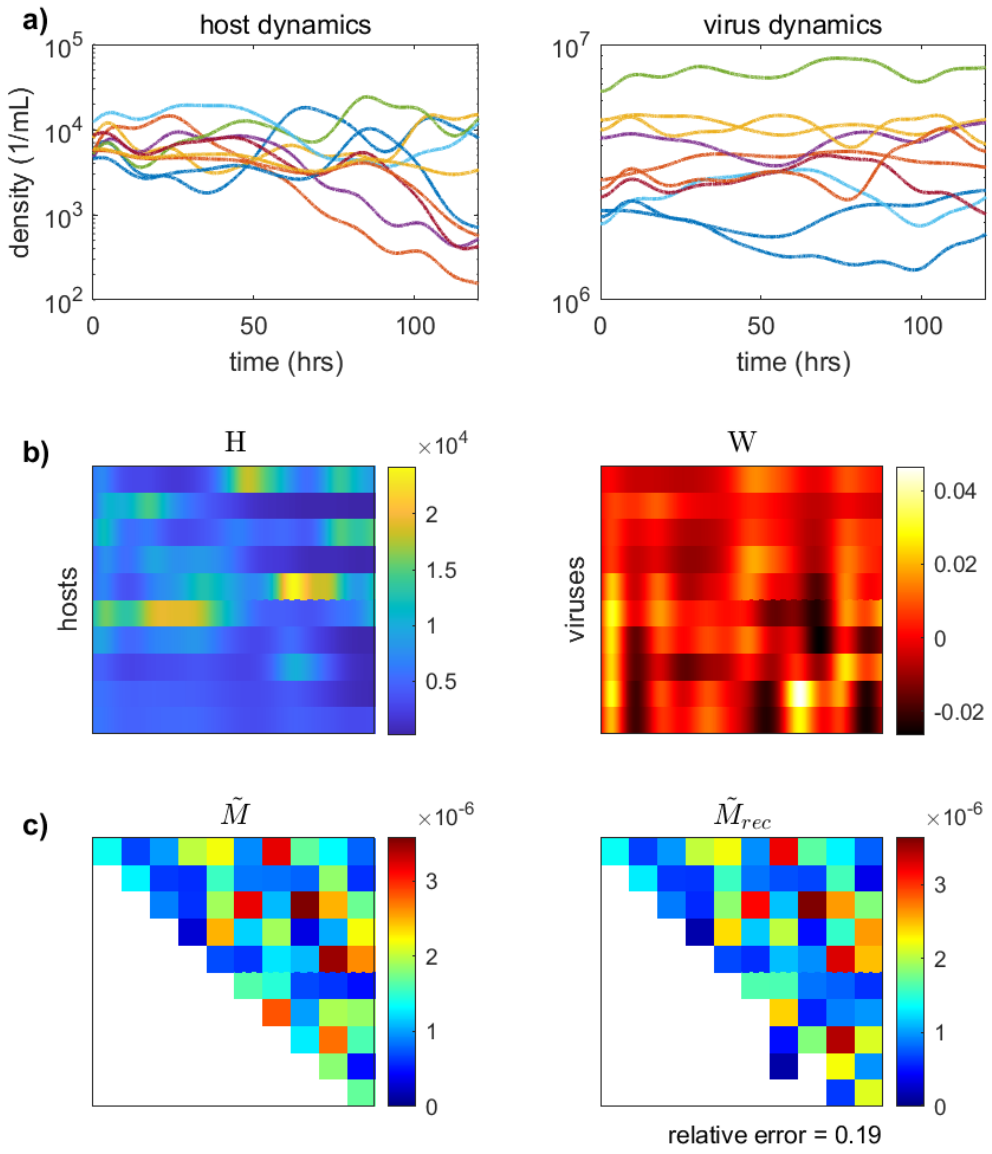

Figure 5.6: Inferring the microbe-virus infection network from time-series data for a 10 by 10 synthetic microbe-virus community. a) Simulated host (left) and virus (right) densities over time. b) Host densities (left, H) and transformed virus differences (right, W), for input into the objective function (Equation 5.20). c) The original "ground-truth" interaction network (left) and the reconstructed network (right). In the interaction matrix, the rows denote hosts, the columns represent viruses, and the colors denote the scaled intensity of interactions (where white denotes no interaction).

5.5 Conclusion

The aim of this primer was to integrate analytic advances together to serve practical aims, so that they can be transferred for analysis of other high resolution temporal data sets. Conducting high-resolution temporal analyses to understand microbial community dynamics has become more feasible in recent years with continued advances in sequence technology. Accordingly, specific statistical considerations should be taken into account as a precursor for microbiome analysis. In this primer, we summarized challenges in analyzing time-series data and present examples which synthesize practical steps to manage these challenges. For further reading on the topics addressed here, we recommend: normalizations and log-ratios [114, 124], distance calculations [194], clustering [41, 195], statistical ordination [196, 197], regression [198], vector autoregression [199], periodicity detection [200], general best practices [138] and an in-depth review of multivariate data analysis [129]. For inferring interactions from time-series, model-based inference approaches have significant potential [117, 44, 50, 58, 118, 119, 57, 120, 121, 59]. Although correlation-based methods have been widely used for inferring interactions, recent literature suggests that correlation-based methods are poor indicators of interaction [190, 191, 43, 24, 192, 193]. Other model-free methods, such as Granger causality [193] and cross-convergent mapping [201], may be useful alternatives for inference although care should be taken that data do not violate the methods' assumptions [202, 203]. In closing, we hope that the consolidated methods and workflows in both R and Matlab help researchers from multiple disciplines advance the quantitative *in situ* study of microbial communities.

5.6 Data Availability

For the 18S rRNA gene-based survey, data originated from [116]. The raw sequence data can also be found under SRA BioProject PRJNA393172. Code to process this 18S rRNA tag-sequencing data can be found at https://github.com/shu251/18Sdiversity_diel and

quality checked reads and final OTU table used for downstream data analysis is available (10.5281/zenodo.1243295), as well as in the GitHub https://github.com/WeitzGroup/analyzing_microbiome_timeseries.

5.7 Conflict of Interest Statement

The authors declare no conflict of interest.

5.8 Author Contributions

AC, SH, EL, DM, and JSW conceptualized the work. SH provided data for analysis. AC, DM, and JSW designed the methods and analyses. SH and DM wrote code for the clustering and periodicity tutorials, AC wrote code for the inference tutorial. AC, SH, EL, DM, and JSW co-wrote the chapter. All authors approve of this chapter.

5.9 Funding

This work was supported by the Simons Foundation (SCOPE award ID 329108) and the National Science Foundation (NSF Bio Oc 1829636).

5.10 Acknowledgments

We thank Dave Caron for helpful feedback and multiple reviewers for their feedback on this chapter.

CHAPTER 6
CONCLUSIONS

This thesis focused on the inference of ecological interactions from time-series data in virus-microbe communities. The ability to infer interactions in a high-throughput and culture-independent manner is critical for fully understanding the role of viruses in microbial communities. In existing literature, many inference methods have been proposed and implemented on "-omics" time-series in natural communities, yet few have been rigorously tested on communities with known interaction networks. This work bridges some of those gaps by utilizing *in silico* and *in vitro* virus-microbe communities.

We have shown that correlation and correlation-based inference methods, widely used in existing literature, perform poorly in complex communities in Chapter 2 (published, [43]). In contrast, we have demonstrated the flexibility and robustness of model-based inference, extending previous work [44] in Chapter 3 (in prep). At the same time, we have found that performing inference in a realized, *in vitro* phage-bacteria community requires more complex models to account for viral latent periods as shown in Chapter 4 (in prep). We utilized a different method, namely Markov-chain Monte Carlo, to compare inferred parameters to experimentally measured parameters in pairwise contexts. In Chapter 5, we proposed a suite of methods to predict interactions, under limited assumptions of normality and independence among time-series samples, when a dynamical model is not explicitly known (published, [45]).

The research presented in this thesis advances our understanding of the successes and limitations of inference, in comparing predicted interactions against known networks. Yet, work remains before we can apply model-based inference to natural communities. We briefly identify five major areas where research is still needed: data type, sampling frequency, parameter magnitude and variation, environment, and model specificity and iden-

tifiability.

In this thesis, we utilized time-series data of absolute abundances of bacteria and viruses. However, absolute abundances are difficult to obtain without targeted methods like qPCR. Currently "-omics" data, such as metagenomes and viromes, is the most promising source of high-throughput and culture-independent data for natural communities. Such data is compositional; the ramifications of compositional data have been well studied for correlation-based and some other statistical methods (Chapter 5) but remains to be studied for model-based methods. In addition, "-omics" data presents challenges with measurement bias, detection limit, and taxonomic resolution. Taxonomic resolution especially poses challenges for model-based inference methods, which delineate populations based on phenotypic traits. Finally, "-omics" data presents the opportunity to integrate entirely new markers of community activity, such as transcriptional activity.

Here, we have worked with high time-resolution data sampled on the order of minutes and hours (Chapters 2, 3, and 4) as opposed to longitudinal data on the order of weeks, months, and years (Chapter 5). We have shown that sampling frequency is an important aspect of experimental design and can strongly affect inference accuracy (Chapter 3). In particular, sampling must be on a similar, or shorter, time scale as the life history traits to be inferred. In the *in vitro* community in Chapter 4, doubling times and latent periods were on the order of hours. In natural communities, doubling times and latent periods are typically on the order of hours and days. Yet, sub-daily sampling is still extremely labor-intensive in ocean environments which are more typically sampled monthly or yearly. In addition, natural communities are often highly diverse, with hundreds or thousands of unique ecotypes. Although not explored in this thesis, large communities will require more intensive sampling to avoid underfitting. Finally, sampling communities over long time scales, on the order of weeks and months, will need to grapple with the effects of evolution. Parallel experiments, as proposed in previous work [44], may help mitigate such effects.

In Chapter 3, we demonstrated that large discrepancies in parameter magnitude nega-

tively effects inference accuracy. For example, when one bacterial population had a growth rate that was larger than the others by an order of magnitude, inference effectively failed. If prior knowledge of parameter ranges exist, it can be incorporated into the inference method, such as by tightening constraints in convex optimization or by setting priors in Markov-chain Monte Carlo. We also expect life history traits to vary *within* populations. Variation within populations is easily accounted for by some inference methods like Markov-chain Monte Carlo (Chapter 4) but is not as well developed for inference methods like convex optimization (Chapter 3). Variation of life history traits within populations is also closely tied to the issue of taxonomic resolution and evolution.

Environmental effects pose additional challenges for inference in natural communities. On long time-scales of months and years, seasonal effects and major events like algal blooms can drastically shift microbial community structure and viral infection mode. Microscopic effects, such as particle-association or sticky lysate, may be strong enough to violate the well-mixed assumption of mean-field ODE models. Physical oceanographic effects, such as upwelling, need also be considered; Lagrangian sampling in a single ocean layer should be considered in experimental design. Finally, changing nutrient conditions can also shift microbial community structure. Although not employed in this thesis, ecological models can consider nutrient dynamics explicitly and can be integrated into model-based inference.

Finally, model specificity remains a challenge for modeling behavior of ecological communities. In Chapter 4, we found that our original models from Chapter 3 were not sufficient to recapitulate *in vitro* community dynamics. The effects of model mis-specification on model-based inference is an open area of research; characterizing effects on inference accuracy for specific combinations of models can be easily done *in silico*. On the other hand, inference of the model equations themselves from data is an active area of research and is well-suited to high-time resolution datasets [82, 204]. Model identifiability – whether one or multiple parameter sets result in the same output – poses further challenges

for model-data integration efforts and interpretability of parameters.

Viruses and microbes shape human and ecosystem health. Yet, we still lack knowledge on how the interactions between viruses and microbes reshape their mutual fates, as well as the fate of the surrounding environment, over ecological and evolutionary time scales. This work is a step towards understanding the implications of virus-microbe interactions – because in order to achieve this goal, we must first be able to identify which viruses and microbes are interacting. By focusing on culture-independent, high-throughput data, we ensure that the inference methods are widely applicable across varying environments and organisms.

Appendices

APPENDIX A

SUPPLEMENTARY INFORMATION FOR "THE LIMITATIONS OF CORRELATION-BASED INFERENCE IN COMPLEX VIRUS-MICROBE COMMUNITIES"

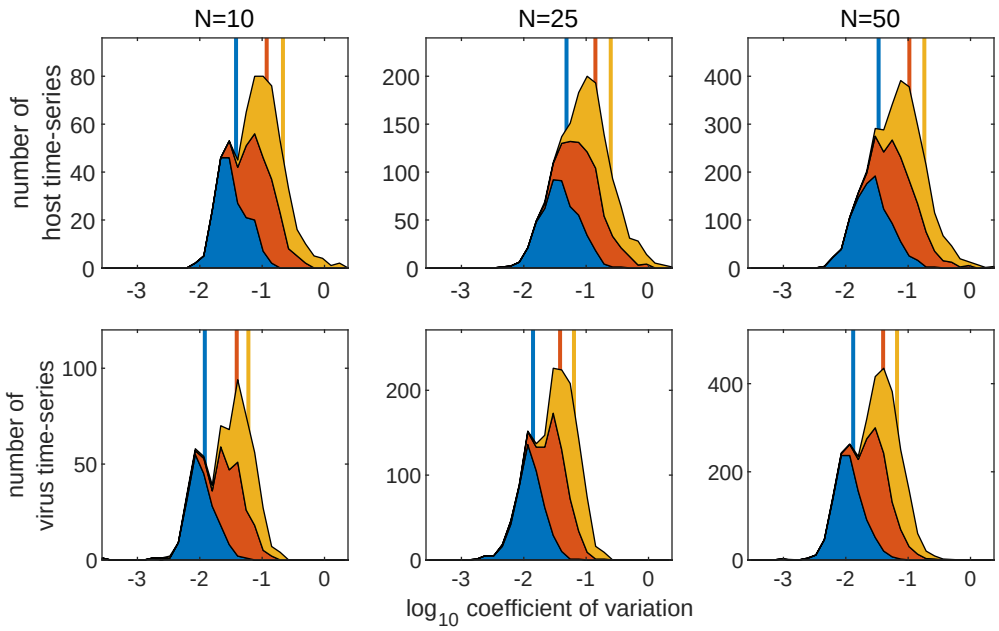

Figure A.1: Distributions of coefficients of variation for each simulated host time-series (top row) and virus time-series (bottom row) for the ensemble of communities over three network sizes ($N = 10, 25, 50$ with 20 communities for each N). The coefficient of variation for an individual time-series is $CV = \sigma/\mu$ where σ is the standard deviation and μ is the mean of the time-series from $t = 0$ hours to $t = 200$ hours (the sample duration used in the main text). The colors correspond to time-series with different initial condition perturbation amounts $\delta = 0.1$ (blue), 0.3 (orange), and 0.5 (yellow); the three distributions are plotted cumulatively here. Solid vertical lines correspond to distribution means. For both hosts and viruses, CV scales with δ but does not scale with N. The mean CVs for host time-series for $\delta = 0.1,\ 0.3,\ 0.5$ (averaged across network sizes) are 0.04 $(10^{-1.40})$, 0.12 $(10^{-0.92})$, and 0.22 $(10^{-0.67})$ respectively. For viruses time-series, they are 0.01 $(10^{-1.88})$, 0.04 $(10^{-1.41})$, and 0.06 $(10^{-1.20})$. Notably, increasing δ (and thus CV) did not improve AUC for any of the correlation-based inference methods.

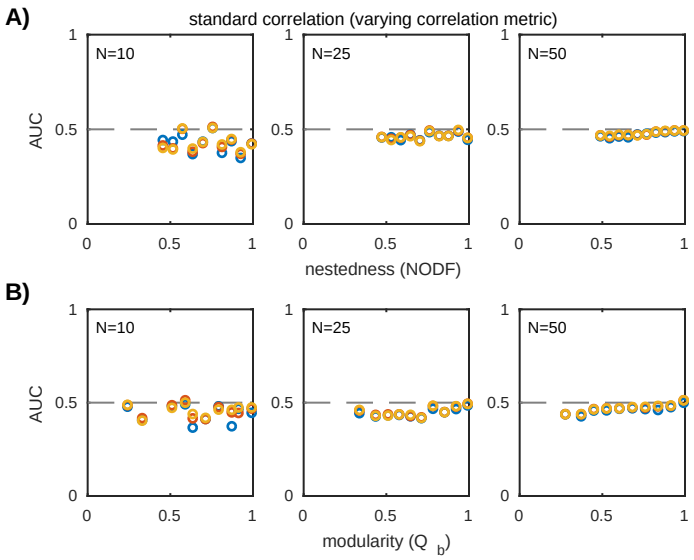

Figure A.2: AUC values for standard correlation of various types (blue=Pearson, orange=Spearman, yellow=Kendall) for the ensemble of A) nested and B) modular communities over three network sizes $N = 10, 25, 50$. Dashed line marks AUC=1/2 and implies the predicted network did no better than random guessing. This figure corresponds to Figure 2.4.

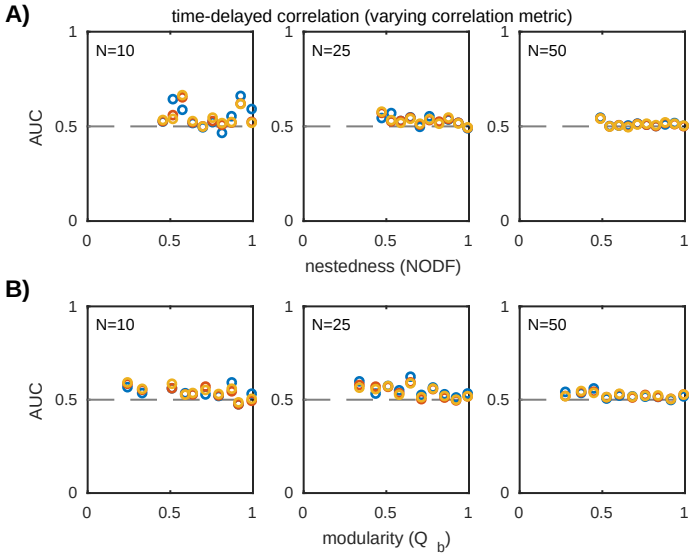

Figure A.3: AUC values for time-delayed correlation of various types (blue=Pearson, orange=Spearman, yellow=Kendall) for the ensemble of A) nested and B) modular communities over three network sizes $N = 10, 25, 50$. Dashed line marks AUC=1/2 and implies the predicted network did no better than random guessing. This figure corresponds to Figure 2.5C.

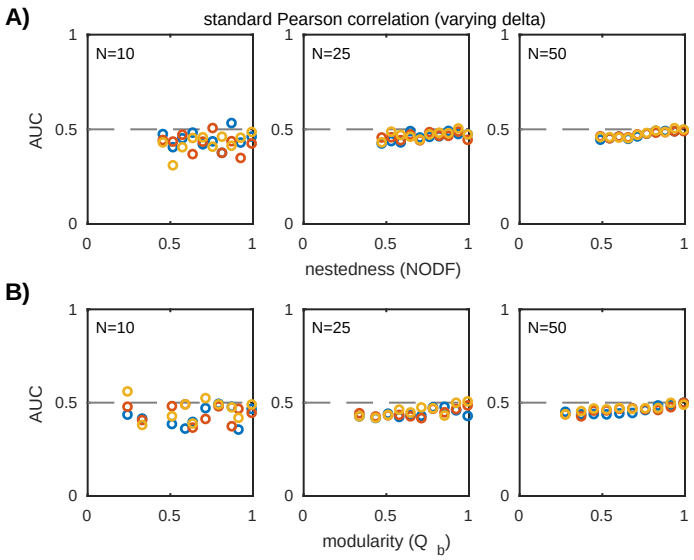

Figure A.4: AUC values for standard Pearson correlation with varying δ values (blue=0.1, orange=0.3, yellow=0.5) for the ensemble of A) nested and B) modular communities over three network sizes $N = 10, 25, 50$. Dashed line marks AUC=1/2 and implies the predicted network did no better than random guessing. This figure corresponds to Figure 2.4.

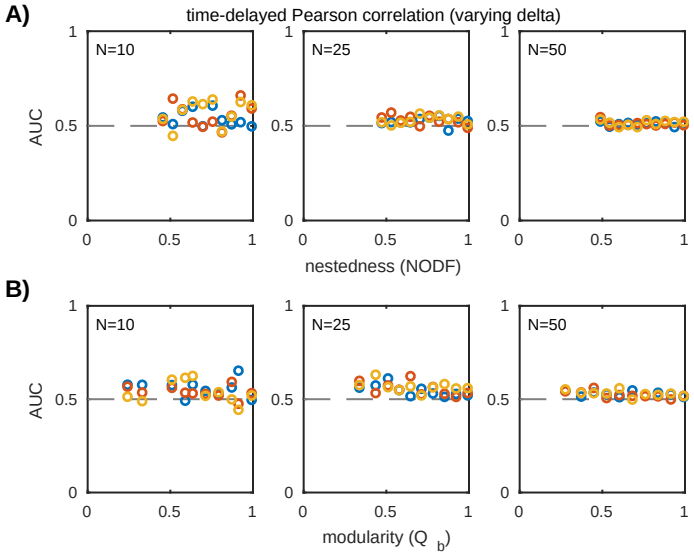

Figure A.5: AUC values for time-delayed Pearson correlation with varying δ values (blue=0.1, orange=0.3, yellow=0.5) for the ensemble of A) nested and B) modular communities over three network sizes $N = 10, 25, 50$. Dashed line marks AUC=1/2 and implies the predicted network did no better than random guessing. This figure corresponds to Figure 2.5C.

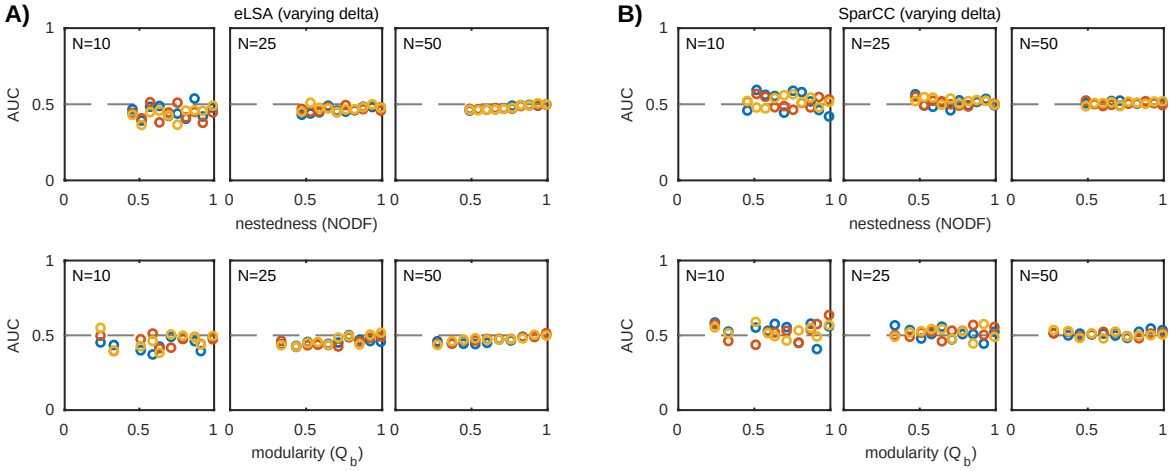

Figure A.6: AUC values for eLSA and SparCC with varying δ values (blue=0.1, orange=0.3, yellow=0.5) for the ensemble of A) nested and B) modular communities over three network sizes $N = 10, 25, 50$. Dashed line marks AUC=1/2 and implies the predicted network did no better than random guessing. This figure corresponds to Figure 2.6C.

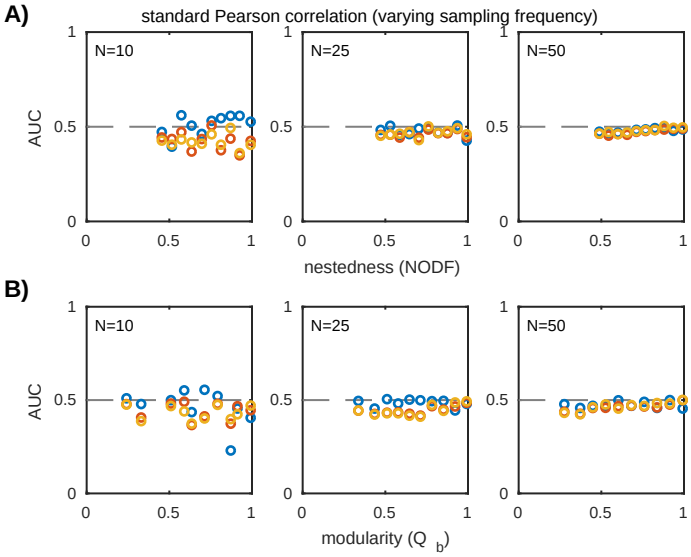

Figure A.7: AUC values for standard Pearson correlation with varying sample frequencies (blue=0.5 hrs, orange=2 hrs, yellow=4 hrs) for the ensemble of A) nested and B) modular communities over three network sizes $N = 10, 25, 50$. Dashed line marks AUC=1/2 and implies the predicted network did no better than random guessing. This figure corresponds to Figure 2.4.

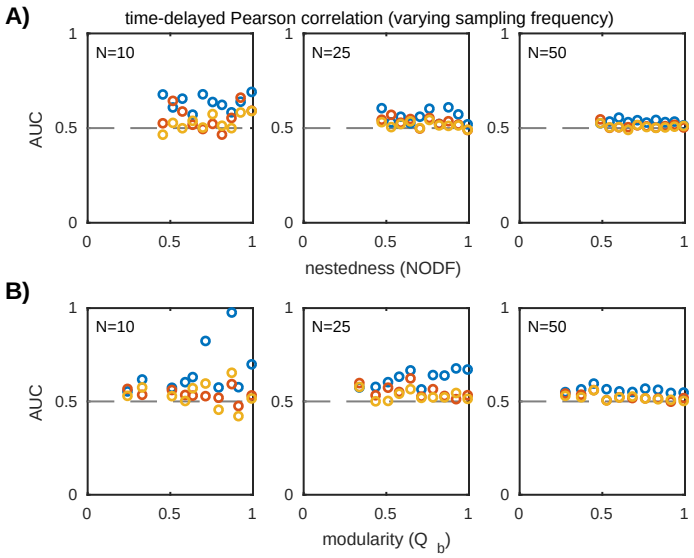

Figure A.8: AUC values for time-delayed Pearson correlation with varying sample frequencies (blue=0.5 hrs, orange=2 hrs, yellow=4 hrs) for the ensemble of A) nested and B) modular communities over three network sizes $N = 10, 25, 50$. Dashed line marks AUC=1/2 and implies the predicted network did no better than random guessing. This figure corresponds to Figure 2.5C.

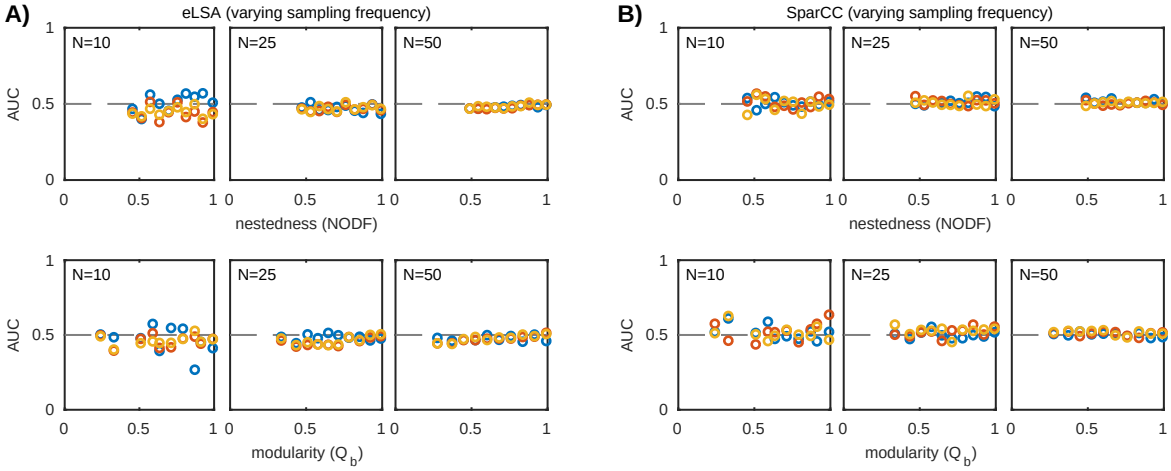

Figure A.9: AUC values for eLSA and SparCC with varying sample frequencies (blue=0.5 hrs, orange=2 hrs, yellow=4 hrs) for the ensemble of A) nested and B) modular communities over three network sizes $N = 10, 25, 50$. Dashed line marks AUC=1/2 and implies the predicted network did no better than random guessing. This figure corresponds to Figure 2.6C.

APPENDIX B

SUPPLEMENTARY INFORMATION FOR "INFERRING MULTIPLE INTERACTION NETWORKS IN VIRUS-BACTERIA COMMUNITIES FROM TIMESERIES DATA"

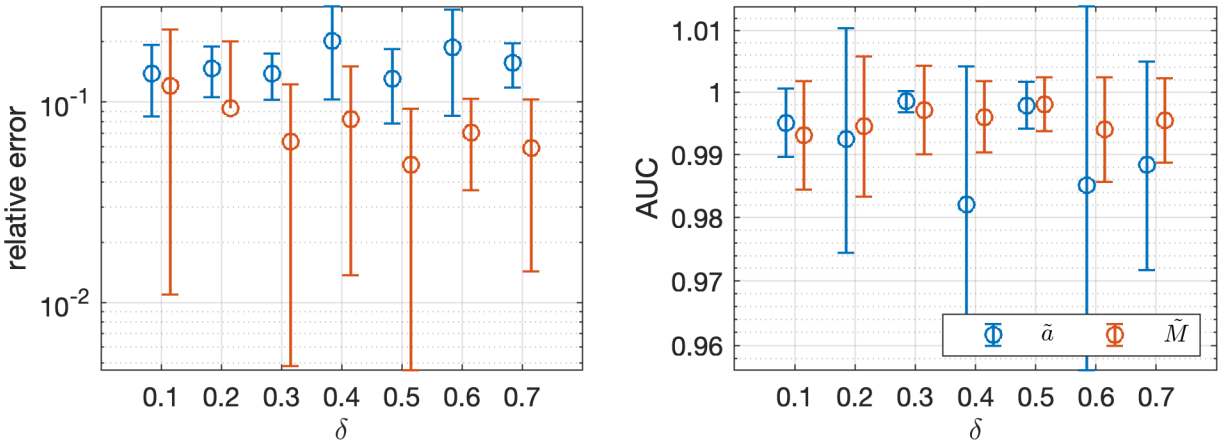

Figure B.1: An example of inference quality vs perturbation constant δ for a uniquely sampled 10x10 virus-host community.

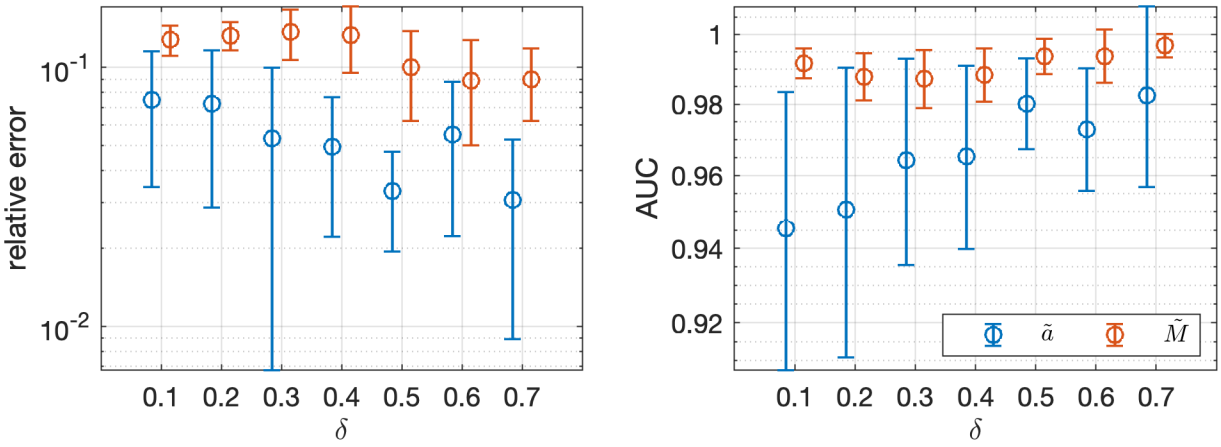

Figure B.2: An example of inference quality vs perturbation constant δ for a uniquely sampled 10x10 virus-host community.

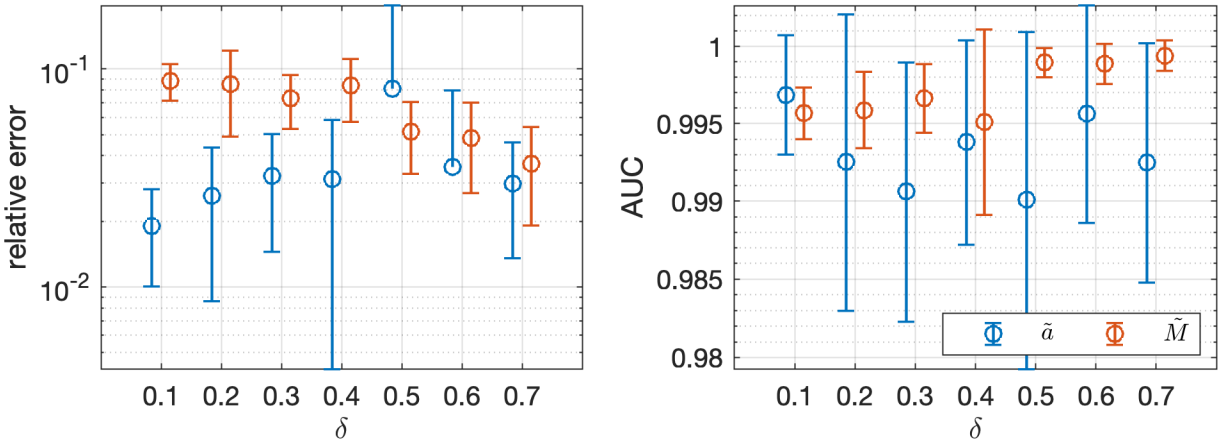

Figure B.3: An example of inference quality vs perturbation constant δ for a uniquely sampled 10x10 virus-host community.

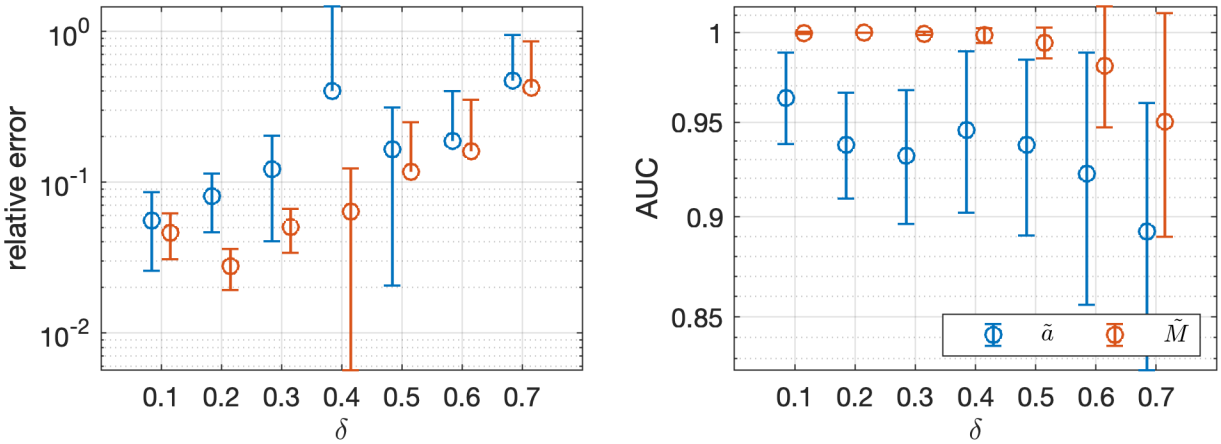

Figure B.4: An example of inference quality vs perturbation constant δ for a uniquely sampled 10x10 virus-host community.

APPENDIX C

SUPPLEMENTARY INFORMATION FOR "RECONSTRUCTING COMMUNITY DYNAMICS FROM PAIRWISE INTERACTIONS IN A COMPLEX BACTERIA-PHAGE COMMUNITY"

	CBA 4	CBA 18	CBA 38	PSA H100	PSA 13-15
CBA 18:2		4.97E+11			
CBA 18:3	2.19E+09	7.35E+10			
CBA 38:1		3.90E+08	1.66E+09		
PSA HP1				9.50E+10	1.33E+11
PSA HS6					4.00E+10

1.00E+08
1.00E+09
1.00E+10
1.00E+11

Figure C.1: Previously measured quantitative host range (qHR) network. Originally, 8 pairs were measured as interacting. A 9th pair (PSA HS6 - PSA H100) was found to interact weakly after performing pairwise adsorption assays.

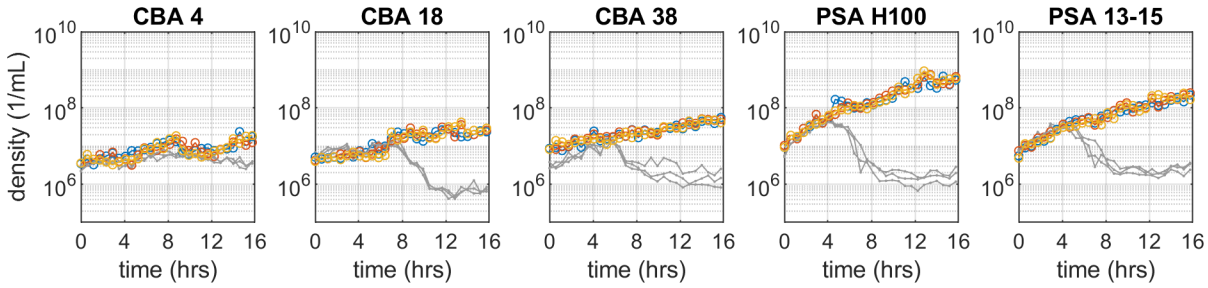

Figure C.2: qPCR measurements from the host-only community experiment. The three replicates are shown in different colors (blue, orange, and yellow). For reference, the qPCR measurements from the host-phage community experiment (Figure 4.2) are also shown (grey lines).

Error for sensitivity analysis in Figure C.5

$$err_i = \sum_{t}^{N_T} \left(\ln \hat{N}_{i,t} - \ln N_{i,t} \right)^2 \tag{C.1}$$

where $\hat{N}_{i,t}$ is the model output for channel i at time t and $N_{i,t}$ is the qPCR data for channel i at time t.

Table C.1: Primer sequences and annealing temperatures (T_m) used for qPCR for each bacterial strain and phage.

Target	**Name**	**F sequence**	**R sequence**	**Tm (degrees C)**
PSA H100	H100 2	GGTGAACTAATTCAATTGGGCGAT	ATCGGTGAGTGTCGAGGGT	54
PSA 13-15	13-15 1	GAGTTTGTGTCGTTGGATCGT	CCCAACTAGTAAACCACCAATCA	54
Cba4	4 set A	TGTCGAGTTTTCTTTCAGTAGCGTG	ACGCCGCACTAATCATAGCCT	62
Cba18	18 set A	TTTTACGAGAACGCCATCTTTCCAC	TGATGTAAGAGGGTTGAGGGCT	62
Cba38	38 set B	CTAGCTCGTAACCCGTCAACCT	ATGGTGCTATTCACTTACTTCCTGC	62
PSA-HP1	HP1 set 2	TGAGCGATATGAGTGTCCGC	ACAAGCTTCCGACCAGAGTG	62
PSA-HS6	HS6 set A	AACCCCGCCTTAGCAACTGT	TGATCAGCGGAGCCACTACG	62
ϕ18:2	18:2 set 1	AAAGGAATGCCCGGAGTCAG	GGGCCGCTGCATGATTAAAG	63
ϕ18:3	18:3 set D	GGATAGGCCACGACGGAGAC	CCAATGCTTTCCGCATCCTTGA	64
ϕ38:1	38:1 set B	AGACAGCCGACGATAACGCA	GCGTCGAAAGGTGTGAACCC	62

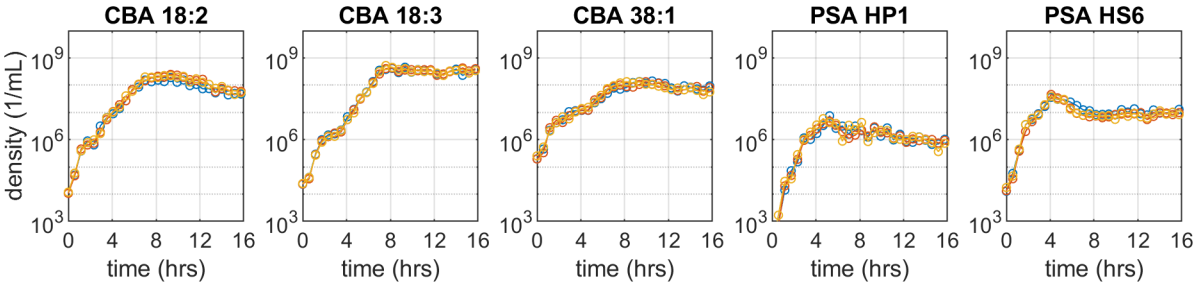

Figure C.3: Intracellular phage DNA measured with qPCR ("qINT") from the community experiment. Three replicates are shown in different colors (blue, orange, yellow).

Table C.2: Index assignment for the 5 bacteria and 5 phage strains. By convention, i refers to bacteria strains and j refers to phage strains. Model parameters with a double index $_{ij}$ refer to the pair of host i and phage j.

i	bacteria strain	j	phage strain
1	CBA 4	1	CBA 18:2
2	CBA 18	2	CBA 18:3
3	CBA 38	3	CBA 38:1
4	PSA H100	4	PSA HP1
5	PSA 13-15	5	PSA HS6

Table C.3: Initial densities of susceptible bacteria S and phage V, used for all model simulations in Chapter 4 and Appendix C. All other state variables (E and I) were initialized at zero. Initial densities were computed from the community experiment as the average across the 3 replicates at the first time point (Figure 4.2).

variable	initial density (1/mL)	variable	initial density (1/mL)
S_1	$2.51 * 10^6$	V_1	$4.29 * 10^5$
S_2	$5.64 * 10^6$	V_2	$2.87 * 10^5$
S_3	$3.03 * 10^6$	V_3	$5.28 * 10^5$
S_4	$6.21 * 10^6$	V_4	$1.10 * 10^5$
S_5	$7.75 * 10^6$	V_5	$1.15 * 10^7$

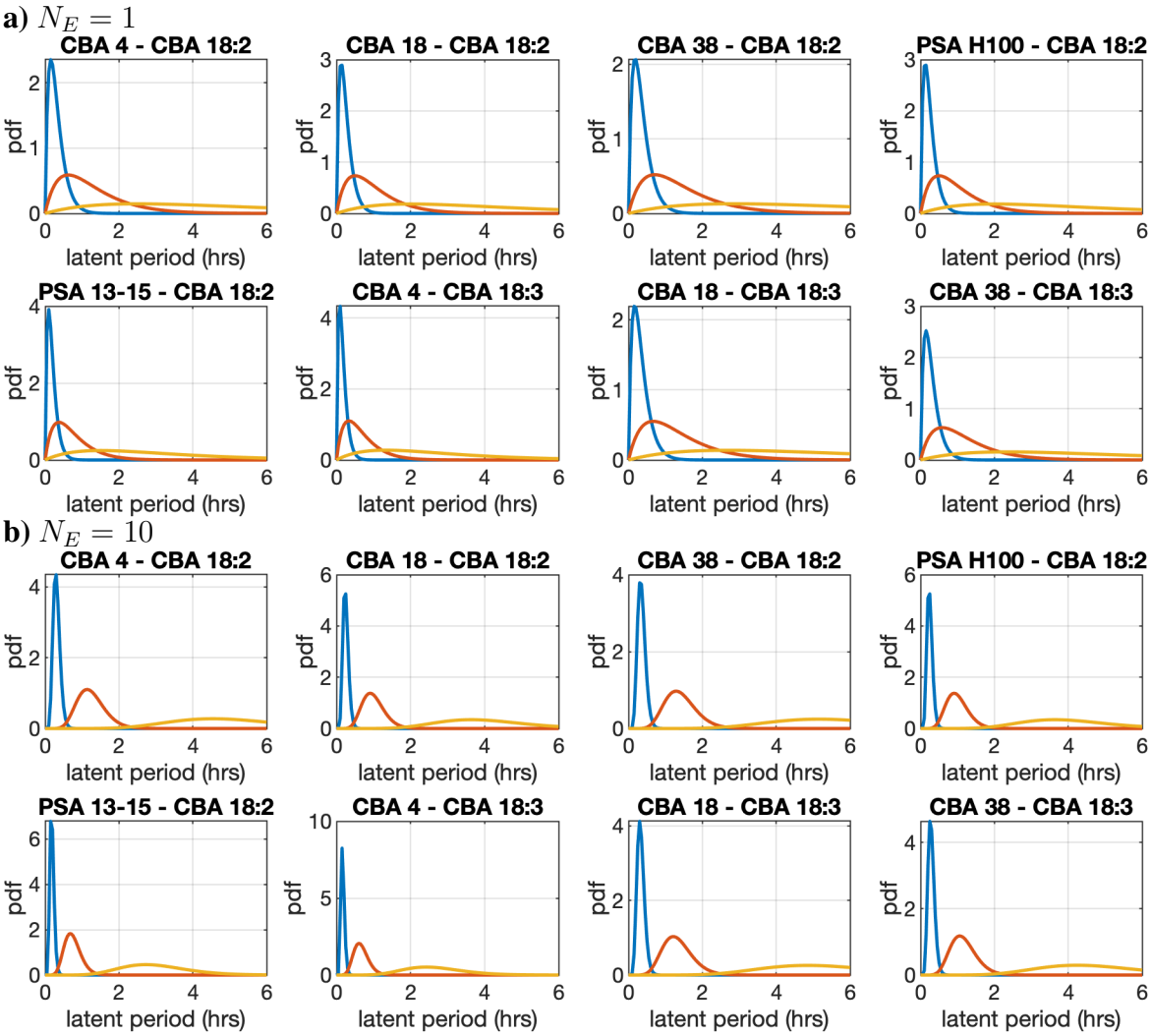

Figure C.4: Theoretical latent period distributions (Methods, Latent period distributions) for 8 phage-host pairs for which latent period was measured (Figure 4.1d; the pair CBA 18-CBA 38:1 does not have latent period data). Distributions correspond to the phage-bacteria model (Equation 4.2) with a) $N_E = 1$ and b) $N_E = 10$. Colors correspond to shortened (blue, latent period multiplier=0.25), baseline (orange, latent period multiplier=1), and lengthened (yellow, latent period multiplier=4) latent periods, as used in Figure 4.3 in the main text.

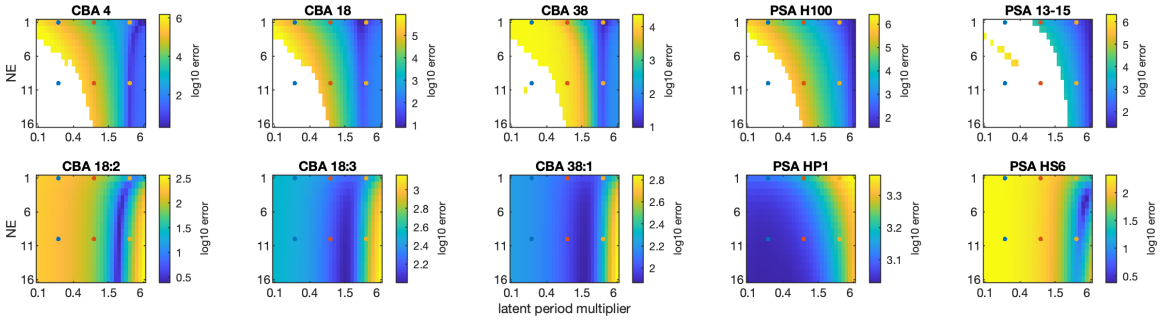

Figure C.5: Sensitivity analysis of latent period multiplier and number of exposed classes N_E for the phage-host model (Equation 4.2), corresponding to Figure 4.3 in the main text. For each pair (latent period multiplier, N_E), the phage-host model is simulated with original parameter values (Table 4.1) but with every latent period multiplied by the latent period multiplier. Error between simulation and data is calculated separately for each host and phage channel (Equation C.1). White denotes infinite error, due to extinction in the simulation. Blue, orange, and yellow dots (NE=1,10 and latent period multiplier=0.25, 1, 4) correspond to the simulated time-series shown in Figure 4.3 in the main text. Note x-axis (latent period multiplier) is equally spaced in log-space; also note non-uniform color bar limits.

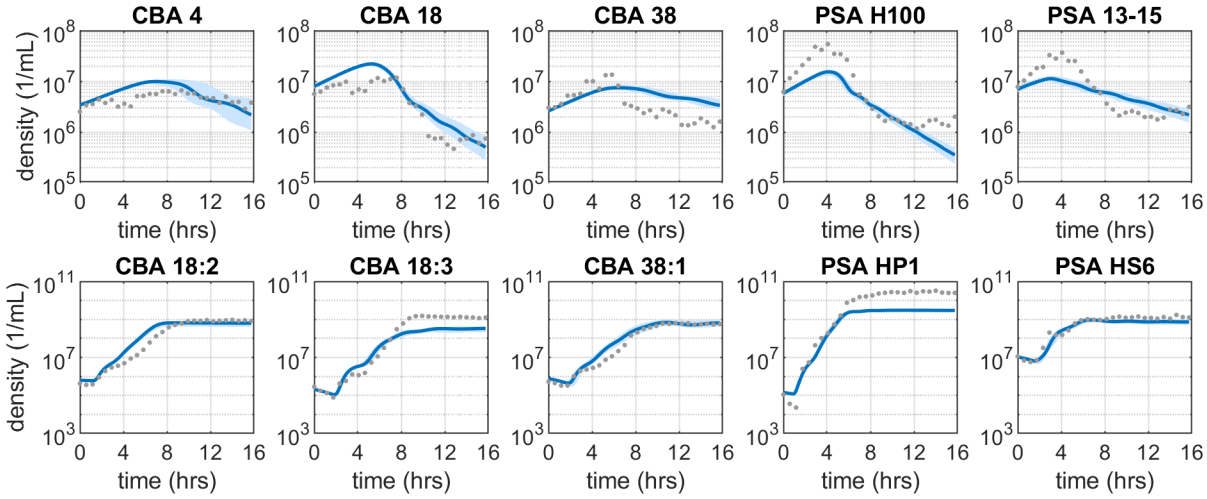

Figure C.6: Example of an MCMC run using a slightly different model than in Figure 4.4. Here, $N_E = 50$. 95th percentile timeseries envelopes are shown (Theory and computational methods, Timeseries envelopes).

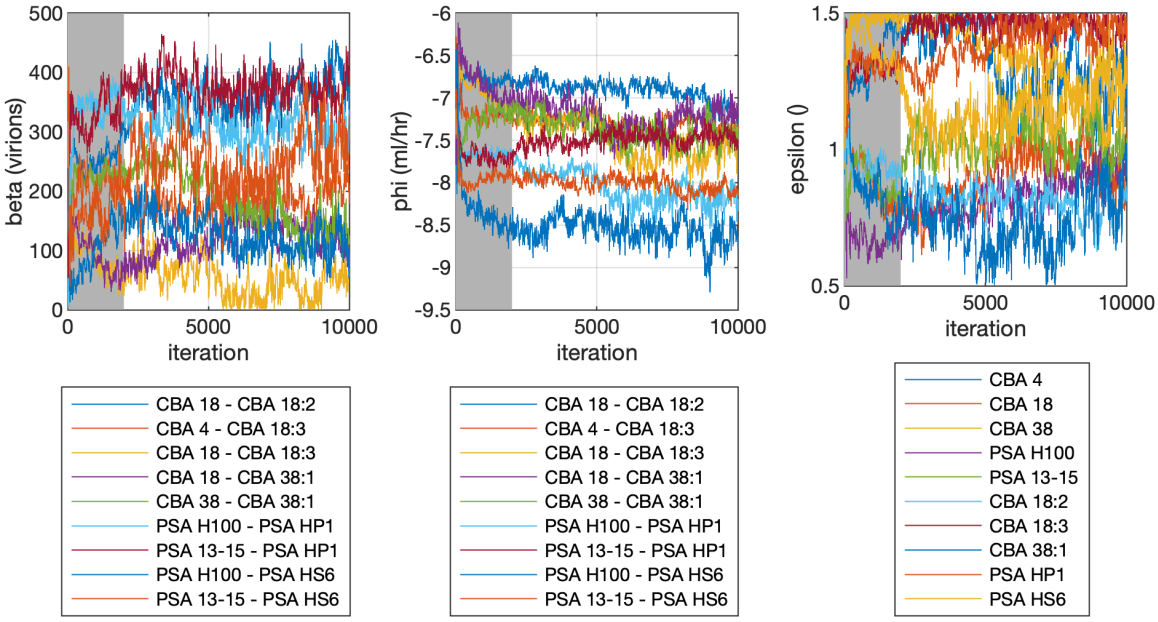

Figure C.7: Chains for the MCMC run corresponding to Figure 4.4 and Figure 4.5 in the main text. Gray shaded area denotes transient. See Table C.5 for MCMC settings.

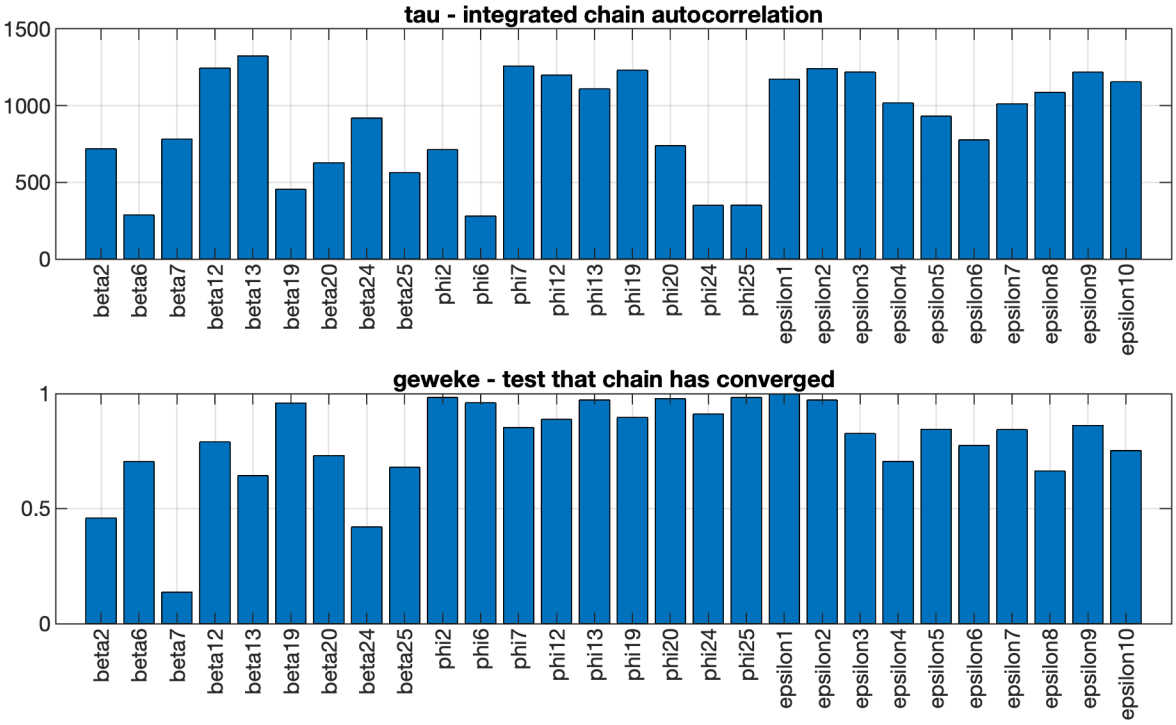

Figure C.8: Convergence statistics for the MCMC run corresponding to Figure 4.4 and Figure 4.5 in the main text. Note linear indexing e.g. "beta25" refers to host 5 and phage 5. See Table C.2 for strain names. See Table C.5 for MCMC settings.

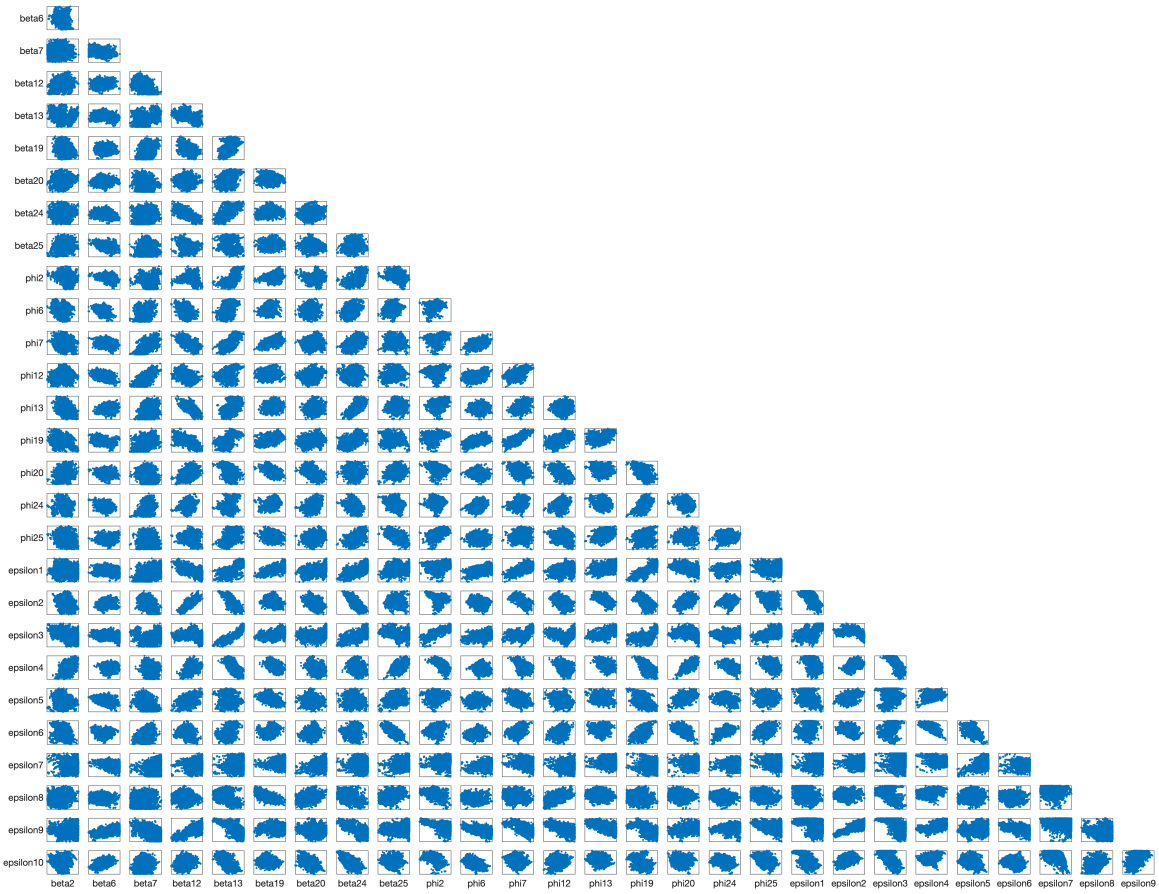

Figure C.9: Covariance plots for the MCMC run corresponding to Figure 4.4 and Figure 4.5 in the main text. Note linear indexing e.g. "beta25" refers to host 5 and phage 5. See Table C.2 for strain names. See Table C.5 for MCMC settings.

Table C.4: Parameter values for simulations in Figure 4.3. See Table 4.1 for parameter descriptions, Table C.2 for strain names, and Table C.3 for initial conditions.

par	value	units
N_H	5	-
N_V	5	-
N_E	a) 1, b) 10	-
r_1	0.2425	1/hr
r_2	0.2163	1/hr
r_3	0.2707	1/hr
r_4	0.2462	1/hr
r_5	0.1888	1/hr

par	value	units
M_{11}	0	-
M_{21}	1	-
M_{31}	0	-
M_{41}	0	-
M_{51}	0	-
M_{12}	1	-
M_{22}	1	-
M_{32}	0	-
M_{42}	0	-
M_{52}	0	-
M_{13}	0	-
M_{23}	1	-
M_{33}	1	-
M_{43}	0	-
M_{53}	0	-
M_{14}	0	-
M_{24}	0	-
M_{34}	0	-
M_{44}	1	-
M_{54}	1	-
M_{15}	0	-
M_{25}	0	-
M_{35}	0	-
M_{45}	1	-
M_{55}	1	-

par	value	units
ϕ_{11}	0	mL/hr
ϕ_{21}	1.884e-08	mL/hr
ϕ_{31}	0	mL/hr
ϕ_{41}	0	mL/hr
ϕ_{51}	0	mL/hr
ϕ_{12}	1.83e-07	mL/hr
ϕ_{22}	1.332e-07	mL/hr
ϕ_{32}	0	mL/hr
ϕ_{42}	0	mL/hr
ϕ_{52}	0	mL/hr
ϕ_{13}	0	mL/hr
ϕ_{23}	1.651e-08	mL/hr
ϕ_{33}	9.9e-08	mL/hr
ϕ_{43}	0	mL/hr
ϕ_{53}	0	mL/hr
ϕ_{14}	0	mL/hr
ϕ_{24}	0	mL/hr
ϕ_{34}	0	mL/hr
ϕ_{44}	1.794e-07	mL/hr
ϕ_{54}	1.878e-07	mL/hr
ϕ_{15}	0	mL/hr
ϕ_{25}	0	mL/hr
ϕ_{35}	0	mL/hr
ϕ_{45}	7.62e-08	mL/hr
ϕ_{55}	1.002e-07	mL/hr

par	value	units
β_{11}	0	-
β_{21}	91.21	-
β_{31}	0	-
β_{41}	0	-
β_{51}	0	-
β_{12}	0.94	-
β_{22}	27.29	-
β_{32}	0	-
β_{42}	0	-
β_{52}	0	-
β_{13}	0	-
β_{23}	1	-
β_{33}	10.5	-
β_{43}	0	-
β_{53}	0	-
β_{14}	0	-
β_{24}	0	-
β_{34}	0	-
β_{44}	58.69	-
β_{54}	54.2	-
β_{15}	0	-
β_{25}	0	-
β_{35}	0	-
β_{45}	248.2	-
β_{55}	318.8	-

par	value	units
τ_{11}	0	hr
τ_{21}	1.25	hr
τ_{31}	0	hr
τ_{41}	0	hr
τ_{51}	0	hr
τ_{12}	1	hr
τ_{22}	1.417	hr
τ_{32}	0	hr
τ_{42}	0	hr
τ_{52}	0	hr
τ_{13}	0	hr
τ_{23}	1	hr
τ_{33}	1	hr
τ_{43}	0	hr
τ_{53}	0	hr
τ_{14}	0	hr
τ_{24}	0	hr
τ_{34}	0	hr
τ_{44}	0.75	hr
τ_{54}	0.6667	hr
τ_{15}	0	hr
τ_{25}	0	hr
τ_{35}	0	hr
τ_{45}	1.333	hr
τ_{55}	1.167	hr

Table C.5: Settings for MCMC run shown in Figure 4.4 and Figure 4.5. Only non-zero parameters were included in the MCMC run. Adsorption rates ϕ were $\log_{10}$ transformed before running the chain. Prior distributions were Gaussians with mean μ and standard deviation σ. See Theory and computational methods: Estimating posterior distributions with Markov-chain Monte Carlo.

parameter	**transform**	**sampling min**	**sampling max**	**prior μ**	**prior σ**	**units**
β_{21}	linear	0	1000	92.21	81.106	-
β_{12}	linear	0	1000	0.94	81.106	-
β_{22}	linear	0	1000	27.29	81.106	-
β_{23}	linear	0	1000	NaN	81.106	-
β_{33}	linear	0	1000	10.5	81.106	-
β_{44}	linear	0	1000	58.7	81.106	-
β_{54}	linear	0	1000	54.2	81.106	-
β_{45}	linear	0	1000	238.3	81.106	-
β_{55}	linear	0	1000	318.8	81.106	-
ϕ_{21}	$\log_{10}$	-10	-5	-7.7249	0.40602	ml/hr
ϕ_{12}	$\log_{10}$	-10	-5	-6.7375	0.40602	ml/hr
ϕ_{22}	$\log_{10}$	-10	-5	-6.8755	0.40602	ml/hr
ϕ_{23}	$\log_{10}$	-10	-5	-7.7825	0.40602	ml/hr
ϕ_{33}	$\log_{10}$	-10	-5	-7.0044	0.40602	ml/hr
ϕ_{44}	$\log_{10}$	-10	-5	-6.7462	0.40602	ml/hr
ϕ_{54}	$\log_{10}$	-10	-5	-6.7263	0.40602	ml/hr
ϕ_{45}	$\log_{10}$	-10	-5	-7.118	0.40602	ml/hr
ϕ_{55}	$\log_{10}$	-10	-5	-6.9991	0.40602	ml/hr
ϵ_1	linear	0.5	1.5	1	0.2	-
ϵ_2	linear	0.5	1.5	1	0.2	-
ϵ_3	linear	0.5	1.5	1	0.2	-
ϵ_4	linear	0.5	1.5	1	0.2	-
ϵ_5	linear	0.5	1.5	1	0.2	-
ϵ_6	linear	0.5	1.5	1	0.2	-
ϵ_7	linear	0.5	1.5	1	0.2	-
ϵ_8	linear	0.5	1.5	1	0.2	-
ϵ_9	linear	0.5	1.5	1	0.2	-
ϵ_{10}	linear	0.5	1.5	1	0.2	-

REFERENCES

[1] M. Breitbart and F. Rohwer, "Here a virus, there a virus, everywhere the same virus?" *Trends in Microbiology*, vol. 13, no. 6, pp. 278–284, 2005.

[2] J. S. Weitz and S. W. Wilhelm, "Ocean viruses and their effects on microbial communities and biogeochemical cycles," *F1000 Biol Rep*, vol. 4, p. 17, 2012.

[3] C. A. Suttle, "Viruses in the sea," *Nature*, vol. 437, no. 7057, pp. 356–61, 2005.

[4] K. D. Bidle and A. Vardi, "A chemical arms race at sea mediates algal host-virus interactions," *Curr Opin Microbiol*, vol. 14, no. 4, pp. 449–57, 2011.

[5] D. Lindell, M. B. Sullivan, Z. I. Johnson, A. C. Tolonen, F. Rohwer, and S. W. Chisholm, "Transfer of photosynthesis genes to and from Prochlorococcus viruses," *Proc Natl Acad Sci U S A*, vol. 101, no. 30, pp. 11 013–8, 2004.

[6] J. R. Brum, I.-E. J. Cesar, R. Simon, D. Guilhem, A. S. G., A. Adriana, C. Samuel, C. Corinne, de Vargas Colomban, G. J. M., G. Gabriel, G. A. C., G. Lionel, H. Pascal, I. Daniele, N. Fabrice, O. Hiroyuki, P. Stephane, P. B. T., S. S. M., S. Sabrina, D. Celine, K.-L. Stefanie, P. Marc, S. Sarah, null null, B. Peer, B. Chris, S. Shinichi, W. Patrick, K. Eric, and S. M. B., "Patterns and ecological drivers of ocean viral communities," *Science*, vol. 348, no. 6237, p. 1 261 498, 2015.

[7] M. Breitbart, "Marine viruses: Truth or dare," *Annual Review of Marine Science*, 2012.

[8] C. Brussaard, "Viral control of phytoplankton populations – a review," *Journal of Eukaryotic Microbiology*, 2005.

[9] J. S. Weitz, C. A. Stock, S. W. Wilhelm, L. Bourouiba, M. L. Coleman, A. Buchan, M. J. Follows, J. A. Fuhrman, L. F. Jover, J. T. Lennon, M. Middelboe, D. L. Sonderegger, C. A. Suttle, B. P. Taylor, T. Frede Thingstad, W. H. Wilson, and K. Eric Wommack, "A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes," *The ISME Journal*, vol. 9, no. 6, pp. 1352–1364, 2015.

[10] F. Rohwer and R. V. Thurber, "Viruses manipulate the marine environment," *Nature*, vol. 459, no. 7244, pp. 207–12, 2009.

[11] L. D. McDaniel, E. Young, J. Delaney, F. Ruhnau, K. B. Ritchie, and J. H. Paul, "High frequency of horizontal gene transfer in the oceans," *Science*, vol. 330, no. 6000, p. 50, 2010.

[12] L. Deng, A. Gregory, S. Yilmaz, B. T. Poulos, P. Hugenholtz, and M. B. Sullivan, "Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as revealed by viral tagging," *MBio*, vol. 3, no. 6, 2012.

[13] L. Deng, J. C. Ignacio-Espinoza, A. C. Gregory, B. T. Poulos, J. S. Weitz, P. Hugenholtz, and M. B. Sullivan, "Viral tagging reveals discrete populations in Synechococcus viral genome sequence space," *Nature*, vol. 513, no. 7517, pp. 242–5, 2014.

[14] A. D. Tadmor, E. A. Ottesen, J. R. Leadbetter, and R. Phillips, "Probing individual environmental bacteria for viruses by using microfluidic digital PCR," *Science*, vol. 333, no. 6038, pp. 58–62, 2011.

[15] S. Roux, A. K. Hawley, M. T. Beltran, M. Scofield, P. Schwientek, R. Stepanauskas, T. Woyke, S. J. Hallam, and M. B. Sullivan, "Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics," *eLIFE*, 2014.

[16] J. M. Labonte, B. K. Swan, B. Poulos, H. Luo, S. Koren, S. J. Hallam, M. B. Sullivan, T. Woyke, K. E. Wommack, and R. Stepanauskas, "Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton," *Isme J*, 2015.

[17] J. H. Munson-McGee, S. Peng, S. Dewerff, R. Stepanauskas, R. J. Whitaker, J. S. Weitz, and M. J. Young, "A virus or more in (nearly) every cell: Ubiquitous networks of virus-host interactions in extreme environments," *Isme J*, 2018.

[18] M. Breitbart, P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer, "Genomic analysis of uncultured marine viral communities," *Proc Natl Acad Sci U S A*, vol. 99, no. 22, pp. 14 250–5, 2002.

[19] R. A. Edwards and F. Rohwer, "Viral metagenomics," *Nat Rev Microbiol*, vol. 3, no. 6, pp. 504–10, 2005.

[20] M. R. Clokie, A. D. Millard, A. V. Letarov, and S. Heaphy, "Phages in nature," *Bacteriophage*, vol. 1, no. 1, pp. 31–45, 2011.

[21] S. Roux, F. Enault, B. L. Hurwitz, and M. B. Sullivan, "VirSorter: Mining viral signal from microbial genomic data," *PeerJ*, 2015.

[22] R. A. Edwards, K. McNair, K. Faust, J. Raes, and B. E. Dutilh, "Computational approaches to predict bacteriophage-host relationships," *FEMS Microbiology Reviews*, 2015.

[23] M. Layeghifard, D. M. Hwang, and D. S. Guttman, "Disentangling interactions in the microbiome: A network perspective," *Trends Microbiol*, vol. 25, no. 3, pp. 217–228, 2017.

[24] S. Weiss, W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, A. Birmingham, J. A. Cram, J. A. Fuhrman, J. Raes, F. Sun, J. Zhou, and R. Knight, "Correlation detection strategies in microbial data sets vary widely in sensitivity and precision," *Isme J*, vol. 10, no. 7, pp. 1669–81, 2016.

[25] K. Faust, L. Lahti, D. Gonze, W. M. de Vos, and J. Raes, "Metagenomics meets time series analysis: Unraveling microbial community dynamics," *Curr Opin Microbiol*, vol. 25, pp. 56–66, 2015.

[26] K. Faust and J. Raes, "Microbial interactions: From networks to models," *Nat Rev Microbiol*, vol. 10, no. 8, pp. 538–50, 2012.

[27] J. A. Fuhrman, "Microbial community structure and its functional implications," *Nature*, vol. 459, no. 7244, pp. 193–9, 2009.

[28] K. Holmfeldt, N. Solonenko, C. Howard-Varona, M. Moreno, R. R. Malmstrom, M. J. Blow, and M. B. Sullivan, "Large-scale maps of variable infection efficiencies in aquatic Bacteroidetes phage-host model systems," *Environmental Microbiology*, vol. 18, no. 11, pp. 3949–3961, 2016.

[29] C. Howard-Varona, K. R. Hargreaves, S. T. Abedon, and M. B. Sullivan, "Lysogeny in nature: Mechanisms, impact and ecology of temperate phages," *The ISME Journal*, vol. 11, no. 7, pp. 1511–1520, 2017.

[30] E. Luo, F. O. Aylward, D. R. Mende, and E. F. DeLong, "Bacteriophage distributions and temporal variability in the ocean's interior," *mBio*, vol. 8, no. 6, M. A. Moran, E. Allen, and A. Culley, Eds., e01903–17, 2017.

[31] Q. Ruan, D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman, and F. Sun, "Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors," *Bioinformatics*, vol. 22, no. 20, pp. 2532–8, 2006.

[32] L. C. Xia, J. A. Steele, J. A. Cram, Z. G. Cardon, S. L. Simmons, J. J. Vallino, J. A. Fuhrman, and F. Sun, "Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates," *BMC Syst Biol*, vol. 5 Suppl 2, S15, 2011.

[33] L. C. Xia, D. Ai, J. Cram, J. A. Fuhrman, and F. Sun, "Efficient statistical significance approximation for local similarity analysis of high-throughput time series data," *Bioinformatics*, vol. 29, no. 2, pp. 230–7, 2013.

[34] C. E. Chow, R. Sachdeva, J. A. Cram, J. A. Steele, D. M. Needham, A. Patel, A. E. Parada, and J. A. Fuhrman, "Temporal variability and coherence of euphotic zone bacterial communities over a decade in the Southern California Bight," *Isme J*, vol. 7, no. 12, pp. 2259–73, 2013.

[35] J. A. Gilbert, J. A. Steele, J. G. Caporaso, L. Steinbruck, J. Reeder, B. Temperton, S. Huse, A. C. McHardy, R. Knight, I. Joint, P. Somerfield, J. A. Fuhrman, and D. Field, "Defining seasonal marine microbial community dynamics," *Isme J*, vol. 6, no. 2, pp. 298–308, 2012.

[36] L. Liu, J. Yang, H. Lv, and Z. Yu, "Synchronous dynamics and correlations between bacteria and phytoplankton in a subtropical drinking water reservoir," *FEMS Microbiol Ecol*, vol. 90, no. 1, pp. 126–38, 2014.

[37] S. F. Paver, K. R. Hayek, K. A. Gano, J. R. Fagen, C. T. Brown, A. G. Davis-Richardson, D. B. Crabb, R. Rosario-Passapera, A. Giongo, E. W. Triplett, and A. D. Kent, "Interactions between specific phytoplankton and bacteria affect lake bacterial community succession," *Environmental Microbiology*, vol. 15, no. 9, pp. 2489–2504, 2013.

[38] D. M. Needham, C. E. Chow, J. A. Cram, R. Sachdeva, A. Parada, and J. A. Fuhrman, "Short-term observations of marine bacterial and viral communities: Patterns, connections and resilience," *Isme J*, vol. 7, no. 7, pp. 1274–85, 2013.

[39] C.-E. T. Chow, D. Y. Kim, R. Sachdeva, D. A. Caron, and J. A. Fuhrman, "Top-down controls on bacterial community structure: Microbial network analysis of bacteria, T4-like viruses and protists," *The ISME Journal*, vol. 8, no. 4, pp. 816–829, 2014.

[40] J. A. Steele, P. D. Countway, L. Xia, P. D. Vigil, J. M. Beman, D. Y. Kim, C. E. Chow, R. Sachdeva, A. C. Jones, M. S. Schwalbach, J. M. Rose, I. Hewson, A. Patel, F. Sun, D. A. Caron, and J. A. Fuhrman, "Marine bacterial, archaeal and protistan association networks reveal ecological linkages," *Isme J*, vol. 5, no. 9, pp. 1414–25, 2011.

[41] Z. D. Kurtz, C. L. Muller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau, "Sparse and compositionally robust inference of microbial ecological networks," *PLoS Comput Biol*, vol. 11, no. 5, e1004226, 2015.

[42] J. Friedman and E. J. Alm, "Inferring correlation networks from genomic survey data," *PLoS Comput Biol*, vol. 8, no. 9, e1002687, 2012.

[43] A. R. Coenen and J. S. Weitz, "Limitations of correlation-based inference in complex virus-microbe communities," *mSystems*, vol. 3, no. 4, S. Bordenstein, Ed., e00084–18, 2018.

[44] L. F. Jover, J. Romberg, and J. S. Weitz, "Inferring phage–bacteria infection networks from time-series data," *Royal Society Open Science*, vol. 3, no. 11, p. 160 654, 2016.

[45] A. R. Coenen, S. K. Hu, E. Luo, D. Muratore, and J. S. Weitz, "A primer for microbiome time-series analysis," *Frontiers in Genetics*, vol. 11, p. 310, 2020.

[46] S. Roux and J. R. Brum, "A viral reckoning: Viruses emerge as essential manipulators of global ecosystems.," *Environ Microbiol Rep*, vol. 11, no. 1, pp. 3–8, 2019.

[47] J. S. Weitz, T. Poisot, J. R. Meyer, C. O. Flores, S. Valverde, M. B. Sullivan, and M. E. Hochberg, "Phage-bacteria infection networks," *Trends Microbiol*, vol. 21, no. 2, pp. 82–91, 2013.

[48] C. O. Flores, S. Valverde, and J. S. Weitz, "Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages," *The ISME journal*, vol. 7, no. 3, pp. 520–532, 2013.

[49] M. B. Sullivan, J. B. Waterbury, and S. W. Chisholm, "Cyanophages infecting the oceanic cyanobacterium Prochlorococcus," *Nature*, vol. 424, no. 6952, pp. 1047–51, 2003.

[50] C. K. Fisher and P. Mehta, "Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression," *PLoS One*, vol. 9, no. 7, e102451, 2014.

[51] C. Flores, T. Poisot, S. Valverde, and J. Weitz, "BiMAT: A MATLAB(R) package to facilitate the analysis and visualization of bipartite networks," 2014.

[52] M. Almeida-Neto, P. Guimaraes, P. R. Guimaraes, R. D. Loyola, and W. Ulrich, "A consistent metric for nestedness analysis in ecological systems: Reconciling concept and measurement," *Oikos*, vol. 117, no. 8, pp. 1227–1239, 2008.

[53] M. J. Barber, "Modularity and community detection in bipartite networks," *Phys. Rev. E*, vol. 76, p. 066 102, 6 2007.

[54] S. J. Beckett, "Improved community detection in weighted bipartite networks," *R Soc Open Sci*, vol. 3, no. 1, p. 140 536, 2016.

[55] L. F. Jover, M. H. Cortez, and J. S. Weitz, "Mechanisms of multi-strain coexistence in host-phage systems with nested infection networks," *J Theor Biol*, vol. 332, pp. 65–77, 2013.

[56] M. Rosenzweig and R. H. MacArthur, "Graphical representation and stability conditions of predator-prey interactions," vol. 97, pp. 209–223, 1963.

[57] R. R. Stein, V. Bucci, N. C. Toussaint, C. G. Buffie, G. Ratsch, E. G. Pamer, C. Sander, and J. B. Xavier, "Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota," *PLoS Comput Biol*, vol. 9, no. 12, e1003388, 2013.

[58] P. Dam, L. L. Fonseca, K. T. Konstantinidis, and E. O. Voit, "Dynamic models of the complex microbial metapopulation of Lake Mendota," *NPJ systems biology and applications*, vol. 2, pp. 16 007–16 007, 2016.

[59] S. Marino, N. T. Baxter, G. B. Huffnagle, J. F. Petrosino, and P. D. Schloss, "Mathematical modeling of primary succession of murine intestinal microbiota," *Proc Natl Acad Sci U S A*, vol. 111, no. 1, pp. 439–44, 2014.

[60] M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming, version 2.1*, http://cvxr.com/cvx, 2014.

[61] G. P. C. Salmond and P. C. Fineran, "A century of the phage: Past, present and future," *Nature Reviews Microbiology*, vol. 13, no. 12, pp. 777–786, 2015.

[62] C. A. Suttle, "Marine viruses —major players in the global ecosystem," *Nature Reviews Microbiology*, vol. 5, no. 10, pp. 801–812, 2007.

[63] M. Breitbart, C. Bonnain, K. Malki, and N. A. Sawaya, "Phage puppet masters of the marine microbial realm," *Nature Microbiology*, vol. 3, no. 7, pp. 754–766, 2018.

[64] J. S. Weitz, *Quantitative Viral Ecology: Dynamics of Viruses and Their Microbial Hosts*, ser. Monographs in Population Biology. Princeton Press, 2016.

[65] J. A. Fuhrman, "Marine viruses and their biogeochemical and ecological effects," *Nature*, vol. 399, no. 6736, pp. 541–548, 1999.

[66] S. Roux, J. R. Brum, B. E. Dutilh, S. Sunagawa, M. B. Duhaime, A. Loy, B. T. Poulos, N. Solonenko, E. Lara, J. Poulain, S. Pesant, S. Kandels-Lewis, C. Dimier, M. Picheral, S. Searson, C. Cruaud, A. Alberti, C. M. Duarte, J. M. Gasol, D. Vaque, P. Bork, S. G. Acinas, P. Wincker, M. B. Sullivan, and T. O. Coordinators, "Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses," *Nature*, vol. 537, no. 7622, pp. 689–693, 2016.

[67] L. F. Jover, T. C. Effler, A. Buchan, S. W. Wilhelm, and J. S. Weitz, "The elemental composition of virus particles: Implications for marine biogeochemical cycles," *Nature Reviews Microbiology*, vol. 12, no. 7, pp. 519–528, 2014.

[68] B. L. Hurwitz and M. B. Sullivan, "The Pacific Ocean Virome (POV): A marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology," *Plos One*, vol. 8, no. 2, e57355–, 2013.

[69] P. Simmonds, M. J. Adams, M. BenkHo, M. Breitbart, J. R. Brister, E. B. Carstens, A. J. Davison, E. Delwart, A. E. Gorbalenya, B. Harrach, R. Hull, A. M. Q. King, E. V. Koonin, M. Krupovic, J. H. Kuhn, E. J. Lefkowitz, M. L. Nibert, R. Orton, M. J. Roossinck, S. Sabanadzovic, M. B. Sullivan, C. A. Suttle, R. B. Tesh, R. van der Vlugt, A. Varsani, and F. M. Zerbini, "Virus taxonomy in the age of metagenomics," *Nature Reviews Microbiology*, vol. 15, no. 3, pp. 161–168, 2017.

[70] P. Turner, J. Draghi, and R. Wilpiszeski, "High-throughput analysis of growth differences among phage strains," *Journal of microbiological methods*, vol. 88, pp. 117–21, 2011.

[71] S. T. Abedon, *Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses*. Cambridge: Cambridge University Press, 2008.

[72] C. O. Flores, J. R. Meyer, S. Valverde, L. Farr, and J. S. Weitz, "Statistical structure of host–phage interactions," *Proceedings of the National Academy of Sciences*, vol. 108, no. 28, E288, 2011.

[73] Y. Shao and I.-N. Wang, "Bacteriophage adsorption rate and optimal lysis time," *Genetics*, vol. 180, no. 1, pp. 471–482, 2008.

[74] J. Friedman, L. M. Higgins, and J. Gore, "Community structure follows simple assembly rules in microbial microcosms," *Nat Ecol Evol*, vol. 1, no. 5, p. 109, 2017.

[75] J. M. Levine, J. Bascompte, P. B. Adler, and S. Allesina, "Beyond pairwise mechanisms of species coexistence in complex communities," *Nature*, vol. 546, no. 7656, pp. 56–64, 2017.

[76] E. C. Seth and M. E. Taga, "Nutrient cross-feeding in the microbial world," *Frontiers in Microbiology*, vol. 5, p. 350, 2014.

[77] S. Pande, H. Merker, K. Bohl, M. Reichelt, S. Schuster, L. F. de Figueiredo, C. Kaleta, and C. Kost, "Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria," *The ISME Journal*, vol. 8, no. 5, pp. 953–962, 2014.

[78] J. Bondy-Denomy, J. Qian, E. R. Westra, A. Buckling, D. S. Guttman, A. R. Davidson, and K. L. Maxwell, "Prophages mediate defense against phage infection through diverse mechanisms," *The ISME journal*, vol. 10, no. 12, pp. 2854–2866, 2016.

[79] T. N. Mavrich, G. F. Hatfull, N. L. Hiller, S. Abedon, C. Igler, and J. Bondy-Denomy, "Evolution of superinfection immunity in cluster a mycobacteriophages," *mBio*, vol. 10, no. 3, e00971–19, 2019.

[80] T. W. Berngruber, F. J. Weissing, and S. Gandon, "Inhibition of superinfection and the evolution of viral latency," *Journal of Virology*, vol. 84, no. 19, pp. 10 200–10 208, 2010.

[81] G. D. Hannigan, M. B. Duhaime, D. Koutra, and P. D. Schloss, "Biogeography and environmental conditions shape bacteriophage-bacteria networks across the human microbiome," *PLOS Computational Biology*, vol. 14, no. 4, e1006099–, 2018.

[82] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, p. 3932, 2016.

[83] H. Karin, M. Mathias, N. Ole, and R. Lasse, "Large variabilities in host strain susceptibility and phage host range govern interactions between lytic marine phages and their flavobacterium hosts," *Applied and Environmental Microbiology*, vol. 73, no. 21, pp. 6730–6739, 2007.

[84] A. Wichels, S. S. Biel, H. R. Gelderblom, T. Brinkhoff, G. Muyzer, and C. Schutt, "Bacteriophage diversity in the north sea," *Applied and Environmental Microbiology*, vol. 64, no. 11, pp. 4128–4133, 1998.

[85] D. Melissa, W. Antje, and S. Matthew, "Six Pseudoalteromonas strains isolated from surface waters of Kabeltonne, offshore Helgoland, North Sea," *Genome Announcements*, vol. 4, no. 1, e01697–15, 2016.

[86] P. J. Hurtado and A. S. Kirosingh, "Generalizations of the 'linear chain trick': Incorporating more flexible dwell time distributions into mean field ODE models," *Journal of Mathematical Biology*, vol. 79, no. 5, pp. 1831–1883, 2019.

[87] H. Haario, M. Laine, A. Mira, and E. Saksman, "Dram: Efficient adaptive mcmc," *Statistics and Computing*, vol. 16, no. 4, pp. 339–354, 2006.

[88] H. Haario, E. Saksman, and J. Tamminen, "An adaptive metropolis algorithm," vol. 7, no. 2, pp. 223–242, 2001.

[89] N. Mitarai, S. Brown, K. Sneppen, and T. J. Silhavy, "Population dynamics of phage and bacteria in spatially structured habitats using phage lambda and escherichia coli," *Journal of Bacteriology*, vol. 198, no. 12, pp. 1783–1793, 2016.

[90] D. A. Caron, "Towards a molecular taxonomy for protists: Benefits, risks, and applications in plankton ecology," *Journal of Eukaryotic Microbiology*, vol. 60, no. 4, pp. 407–413, 2013.

[91] S. M. Huse, L. Dethlefsen, J. A. Huber, D. M. Welch, D. A. Relman, and M. L. Sogin, "Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing," *PLOS Genetics*, vol. 4, no. 11, e1000255–, 2008.

[92] R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L.-I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P. C. Dorrestein, "Best practices for analysing microbiomes," *Nature Reviews Microbiology*, vol. 16, no. 7, pp. 410–422, 2018.

[93] S. Widder, R. J. Allen, T. Pfeiffer, T. P. Curtis, C. Wiuf, W. T. Sloan, O. X. Cordero, S. P. Brown, B. Momeni, W. Shou, H. Kettle, H. J. Flint, A. F. Haas, B. Laroche, J.-U. Kreft, P. B. Rainey, S. Freilich, S. Schuster, K. Milferstedt, J. R. van der Meer, T. Grobkopf, J. Huisman, A. Free, C. Picioreanu, C. Quince, I. Klapper, S. Labarthe, B. F. Smets, H. Wang, I. N. I. Fellows, and O. S. Soyer, "Challenges in microbial ecology: Building predictive understanding of community function and dynamics," *The Isme Journal*, vol. 10, 2557 Ep -, 2016.

[94] S. Weiss, Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vazquez-Baeza, A. Birmingham, E. R. Hyde, and R. Knight, "Normalization and microbial differential abundance strategies depend upon data characteristics," *Microbiome*, vol. 5, no. 1, p. 27, 2017.

[95] P. J. McMurdie and S. Holmes, "Waste not, want not: Why rarefying microbiome data is inadmissible," *PLOS Computational Biology*, vol. 10, no. 4, pp. 1–12, 2014.

[96] K. T. Konstantinidis, A. Ramette, and J. M. Tiedje, "The bacterial species definition in the genomic era," *Philosophical Transactions of the Royal Society B*, no. October, pp. 1929–1940, 2006.

[97] D. A. Caron and S. K. Hu, "Are we overestimating protistan diversity in nature?" *Trends in Microbiology*, vol. 27, no. 3, pp. 197–205, 2019.

[98] M. Blaxter, J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd, and E. Abebe, "Defining operational taxonomic units using DNA barcode data," *Philosophical*

transactions of the Royal Society of London. Series B, Biological sciences, vol. 360, no. 1462, pp. 1935–1943, 2005.

[99] B. J. Callahan, P. J. McMurdie, and S. P. Holmes, "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis," *The Isme Journal*, vol. 11, 2639 Ep -, 2017.

[100] A. M. Eren, H. G. Morrison, P. J. Lescault, J. Reveillaud, J. H. Vineis, and M. L. Sogin, "Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences," *The Isme Journal*, vol. 9, 968 Ep -, 2014.

[101] F. Mahe, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn, "Swarm v2: Highly-scalable and high-resolution amplicon clustering," *PeerJ*, vol. 3, e1420, e1420–e1420, 2015.

[102] P. Katsonis, A. Koire, S. J. Wilson, T.-K. Hsu, R. C. Lua, A. D. Wilkins, and O. Lichtarge, "Single nucleotide variations: Biological impact and theoretical interpretation," *Protein science : a publication of the Protein Society*, vol. 23, no. 12, pp. 1650–1666, 2014.

[103] S. Sunagawa, D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. a. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, S. Rasmussen, S. Brunak, O. Pedersen, F. Guarner, W. M. de Vos, J. Wang, J. Li, J. Dore, S. D. Ehrlich, a. Stamatakis, and P. Bork, "Metagenomic species profiling using universal phylogenetic marker genes," *Nat Methods*, vol. 10, no. 12, pp. 1196–1199, 2013.

[104] D. R. Mende, S. Sunagawa, G. Zeller, and P. Bork, "Accurate and universal delineation of prokaryotic species," *Nature Methods*, vol. 10, p. 881, 2013.

[105] N. J. Varghese, S. Mukherjee, N. Ivanova, T. Konstantinidis, K. Mavrommatis, N. C. Kyrpides, and A. Pati, "Microbial species delineation using whole genome sequences," *Nucleic Acids Research*, vol. 43, no. 14, pp. 6761–6771, 2015.

[106] S. Roux, J. B. Emerson, E. A. Eloe-Fadrosh, and M. B. Sullivan, "Benchmarking viromics: an ¡i¿in silico¡/i¿ evaluation of metagenome-enabled estimates of viral community composition and diversity," *PeerJ*, vol. 5, e3817, 2017.

[107] K. T. Konstantinidis and J. M. Tiedje, "Genomic insights that advance the species definition for prokaryotes," *Proceedings of the National Academy of Sciences*, vol. 102, no. 7, pp. 2567–2572, 2005.

[108] S. K. Hu, Z. Liu, A. A. Y. Lie, P. D. Countway, D. Y. Kim, A. C. Jones, R. J. Gast, S. C. Cary, E. B. Sherr, B. F. Sherr, and D. A. Caron, "Estimating protistan diversity

using high-throughput sequencing," *Journal of Eukaryotic Microbiology*, vol. 62, no. 5, pp. 688–693, 2015.

[109] M. Kim, M. Morrison, and Z. Yu, "Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes," *Journal of Microbiological Methods*, vol. 84, no. 1, pp. 81–87, 2011.

[110] N. Youssef, C. S. Sheik, L. R. Krumholz, F. Z. Najar, B. A. Roe, and M. S. Elshahed, "Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys," *Applied and Environmental Microbiology*, vol. 75, no. 16, p. 5227, 2009.

[111] A. D. Willis, "Rigorous statistical methods for rigorous microbiome science," *mSystems*, vol. 4, no. 3, 2019.

[112] J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop, "Differential abundance analysis for microbial marker-gene surveys," *Nature Methods*, vol. 10, 1200 Ep -, 2013.

[113] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome datasets are compositional: And this is not optional," *Frontiers in Microbiology*, vol. 8, p. 2224, 2017.

[114] J. D. Silverman, A. D. Washburne, S. Mukherjee, and L. A. David, "A phylogenetic transform enhances analysis of compositional microbiota data," *eLife*, vol. 6, A. Fodor, Ed., e21887, 2017.

[115] K. Faust, L. Lahti, D. Gonze, W. M. de Vos, and J. Raes, "Metagenomics meets time series analysis: Unraveling microbial community dynamics," *Current Opinion in Microbiology*, vol. 25, pp. 56–66, 2015.

[116] S. K. Hu, P. E. Connell, L. Y. Mesrop, and D. A. Caron, "A hard day's night: Diel shifts in microbial eukaryotic activity in the North Pacific Subtropical Gyre," *Frontiers in Marine Science*, vol. 5, p. 351, 2018.

[117] K. Faust, F. Bauchinger, B. Laroche, S. de Buyl, L. Lahti, A. D. Washburne, D. Gonze, and S. Widder, "Signatures of ecological processes in microbial community time series," *Microbiome*, vol. 6, no. 1, p. 120, 2018.

[118] Y. Xiao, M. T. Angulo, J. Friedman, M. K. Waldor, S. T. Weiss, and Y.-Y. Liu, "Mapping the ecological networks of microbial communities," *Nature Communications*, vol. 8, no. 1, p. 2042, 2017.

[119] O. S. Venturelli, A. C. Carr, G. Fisher, R. H. Hsu, R. Lau, B. P. Bowen, S. Hromada, T. Northen, and A. P. Arkin, "Deciphering microbial interactions in synthetic

human gut microbiome communities," *Molecular systems biology*, vol. 14, no. 6, e8157–e8157, 2018.

[120] O. Ovaskainen, G. Tikhonov, D. Dunson, V. Grotan, S. Engen, B.-E. Saether, and N. Abrego, "How are species interactions structured in species-rich communities? a new method for analysing time-series data," *Proceedings. Biological sciences*, vol. 284, no. 1855, p. 20 170 768, 2017.

[121] J. Mounier, C. Monnet, T. Vallaeys, R. Arditi, A.-S. Sarthou, A. Helias, and F. Irlinger, "Microbial interactions within a cheese microbial community," *Applied and Environmental Microbiology*, vol. 74, no. 1, pp. 172–181, 2008.

[122] J. Aitchison, "The statistical analysis of compositional data," *Journal of the International Association for Mathematical Geology*, vol. 44, no. 2, pp. 139–177, 1983.

[123] M. R. McLaren, A. D. Willis, and B. J. Callahan, "Consistent and correctable bias in metagenomic sequencing measurements," *bioRxiv*, p. 559 831, 2019.

[124] J. J. Egozcue, V. Pawlowsky-Glahn, G. Figueras, and C. Vidal, "Isometric logratio transformations for compositional data analysis," *Mathematical Geology*, vol. 35, pp. 279–300, 2003.

[125] M. C. Tsilimigras and A. A. Fodor, "Compositional data analysis of the microbiome: Fundamentals, tools, and challenges," *Annals of Epidemiology*, vol. 26, no. 5, pp. 330–335, 2016, The Microbiome and Epidemiology.

[126] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker, "Analysis of microarray data using Z score transformation," *The Journal of Molecular Diagnostics*, vol. 5, no. 2, pp. 73–81, 2003.

[127] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 12, p. 550, 2014.

[128] J. Kuczynski, Z. Liu, C. Lozupone, D. McDonald, N. Fierer, and R. Knight, "Microbial community resemblance methods differ in their ability to detect biologically relevant patterns," *Nature methods*, vol. 7, 10 2010.

[129] P. L. Buttigieg and A. Ramette, "A guide to statistical analysis in microbial ecology: A community-focused, living review of multivariate data analyses," *FEMS Microbiology Ecology*, vol. 90, no. 3, pp. 543–550, 2014.

[130] P. Jaccard, "The distribution of the flora in the alpine zone.1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[131] C. Lozupone and R. Knight, "UniFrac: A new phylogenetic method for comparing microbial communities," *Applied and Environmental Microbiology*, vol. 71, no. 12, pp. 8228–8235, 2005.

[132] J. A. Aitchison, C. Vidal, J. Martin-Fernandez, and V. Pawlowsky-Glahn, "Logratio analysis and compositional distance," *Mathematical Geology*, vol. 32, pp. 271–275, 2000.

[133] J. R. Bray and J. T. Curtis, "An ordination of the upland forest communities of southern Wisconsin," *Ecological monographs*, vol. 27, no. 4, pp. 325–349, 1957.

[134] J. C. Gower and P. Legendre, "Metric and Euclidean properties of dissimilarity coefficients," *Journal of classification*, vol. 3, no. 1, pp. 5–48, 1986.

[135] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[136] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991.

[137] D. Borcard and P. Legendre, "All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices," *Ecological Modelling*, vol. 153, no. 1, pp. 51–68, 2002.

[138] S. Holmes and W. Huber, *Modern Statistics for Modern Biology*. Cambridge, England: Cambridge University Press, 2019.

[139] J. B. Kruskal, "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, 1964.

[140] T. W. Liao, "Clustering of time series data—a survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[141] F. K. Gulagiz and S. Sahin, "Comparison of hierarchical and non-hierarchical clustering algorithms," *International Journal of Computer Engineering and Information Technology*, vol. 9, no. 1, p. 6, 2017.

[142] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[143] M. T. Kavanaugh, B. Hales, M. Saraceno, Y. H. Spitz, A. E. White, and R. M. Letelier, "Hierarchical and dynamic seascapes: A quantitative framework for scaling pelagic biogeochemistry and ecology," *Progress in Oceanography*, vol. 120, pp. 291–304, 2014.

[144] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "DBSCAN: Past, present and future," in *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, Nicosia, Cyprus, 2014, pp. 232–238.

[145] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. New York City, NY: John Wiley & Sons, 2009, vol. 344.

[146] F. Murtagh, "Multidimensional clustering algorithms," *Compstat Lectures, Vienna: Physika Verlag, 1985*, 1985.

[147] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Proceedings of the 2010 IEEE International Conference on Data Mining*, ser. Icdm '10, Washington, DC, USA: IEEE Computer Society, 2010, pp. 911–916.

[148] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[149] M. E. Mann and J. M. Lees, "Robust estimation of background noise and signal detection in climatic time series," *Climatic change*, vol. 33, no. 3, pp. 409–445, 1996.

[150] E. A. Ottesen, C. R. Young, S. M. Gifford, J. M. Eppley, R. Marin, S. C. Schuster, C. A. Scholin, and E. F. DeLong, "Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages," *Science*, vol. 345, no. 6193, pp. 207–212, 2014.

[151] R. Yang and Z. Su, "Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation," *Bioinformatics*, vol. 26, no. 12, pp. i168–i174, 2010.

[152] R. Yang, C. Zhang, and Z. Su, "LSPR: An integrated periodicity detection algorithm for unevenly sampled temporal microarray data," *Bioinformatics*, vol. 27, no. 7, pp. 1023–1025, 2011.

[153] F. O. Aylward, D. Boeuf, D. R. Mende, E. M. Wood-Charlson, A. Vislova, J. M. Eppley, A. E. Romano, and E. F. DeLong, "Diel cycling and long-term persistence of viruses in the ocean's euphotic zone," *Proceedings of the National Academy of Sciences*, vol. 114, no. 43, pp. 11 446–11 451, 2017.

[154] S. T. Wilson, F. O. Aylward, F. Ribalet, B. Barone, J. R. Casey, P. E. Connell, J. M. Eppley, S. Ferron, J. N. Fitzsimmons, C. T. Hayes, A. E. Romano, K. A. Turk-Kubo, A. Vislova, E. V. Armbrust, D. A. Caron, M. J. Church, J. P. Zehr, D. M. Karl, and E. F. DeLong, "Coordinated regulation of growth, activity and transcription in nat-

ural populations of the unicellular nitrogen-fixing cyanobacterium Crocosphaera," *Nature Microbiology*, vol. 2, 17118 Ep -, 2017.

[155] M. E. Hughes, K. C. Abruzzi, R. Allada, R. Anafi, A. B. Arpat, G. Asher, P. Baldi, C. de Bekker, D. Bell-Pedersen, J. Blau, S. Brown, M. F. Ceriani, Z. Chen, J. C. Chiu, J. Cox, A. M. Crowell, J. P. DeBruyne, D.-J. Dijk, L. DiTacchio, F. J. Doyle, G. E. Duffield, J. C. Dunlap, K. Eckel-Mahan, K. A. Esser, G. A. FitzGerald, D. B. Forger, L. J. Francey, Y.-H. Fu, F. Gachon, D. Gatfield, P. de Goede, S. S. Golden, C. Green, J. Harer, S. Harmer, J. Haspel, M. H. Hastings, H. Herzel, E. D. Herzog, C. Hoffmann, C. Hong, J. J. Hughey, J. M. Hurley, H. O. de la Iglesia, C. Johnson, S. A. Kay, N. Koike, K. Kornacker, A. Kramer, K. Lamia, T. Leise, S. A. Lewis, J. Li, X. Li, A. C. Liu, J. J. Loros, T. A. Martino, J. S. Menet, M. Merrow, A. J. Millar, T. Mockler, F. Naef, E. Nagoshi, M. N. Nitabach, M. Olmedo, D. A. Nusinow, L. J. Ptav cek, D. Rand, A. B. Reddy, M. S. Robles, T. Roenneberg, M. Rosbash, M. D. Ruben, S. S. Rund, A. Sancar, P. Sassone-Corsi, A. Sehgal, S. Sherrill-Mix, D. J. Skene, K.-F. Storch, J. S. Takahashi, H. R. Ueda, H. Wang, C. Weitz, P. O. Westermark, H. Wijnen, Y. Xu, G. Wu, S.-H. Yoo, M. Young, E. E. Zhang, T. Zielinski, and J. B. Hogenesch, "Guidelines for genome-scale analysis of biological rhythms," *Journal of Biological Rhythms*, vol. 32, no. 5, pp. 380–393, 2017, Pmid: 29098954.

[156] P. F. Thaben and P. O. Westermark, "Detecting rhythms in time series with RAIN," *Journal of biological rhythms*, vol. 29, no. 6, pp. 391–400, 2014.

[157] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[158] D. L. Streiner, "Best (but oft-forgotten) practices: The multiple problems of multiplicity–whether and how to correct for many statistical tests," *The American Journal of Clinical Nutrition*, vol. 102, no. 4, pp. 721–728, 2015.

[159] M. E. Glickman, S. R. Rao, and M. R. Schultz, "False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies," *Journal of Clinical Epidemiology*, vol. 67, no. 8, pp. 850–857, 2014.

[160] W. S. Noble, "How does multiple testing correction work?" *Nature Biotechnology*, vol. 27, 1135 Ep -, 2009.

[161] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[162] K. N. Conneely and M. Boehnke, "So many correlated tests, so little time! rapid adjustment of P values for multiple correlated tests," *The American Journal of Human Genetics*, vol. 81, no. 6, pp. 1158–1168, 2007.

[163] J. R. Stevens, A. Al Masud, and A. Suyundikov, "A comparison of multiple testing adjustment methods with block-correlation positively-dependent tests," *Plos One*, vol. 12, no. 4, pp. 1–12, 2017.

[164] D. Storch and A. L. vSizling, "The concept of taxon invariance in ecology: Do diversity patterns vary with changes in taxonomic resolution?" *Folia Geobotanica*, 2008.

[165] S. Vincenzi, A. J. Crivelli, S. Munch, H. J. Skaug, and M. Mangel, "Trade-offs between accuracy and interpretability in von Bertalanffy random-effects models of growth," *Ecological Applications*, vol. 26, no. 5, pp. 1535–1552, 2016.

[166] H.-T. Cao, T. E. Gibson, A. Bashan, and Y.-Y. Liu, "Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons," *BioEssays*, vol. 39, no. 2, p. 1 600 188, 2017.

[167] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Inferring biological networks by sparse identification of nonlinear dynamics," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 2, no. 1, pp. 52–63, 2016.

[168] N. M. Mangan, J. N. Kutz, S. L. Brunton, and J. L. Proctor, "Model selection for dynamical systems via sparse regression and information criteria," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 473, no. 2204, p. 20 170 009, 2017.

[169] L. F. Jover, M. H. Cortez, and J. S. Weitz, "Mechanisms of multi-strain coexistence in host–phage systems with nested infection networks," *Journal of Theoretical Biology*, vol. 332, pp. 65–77, 2013.

[170] D. A. Korytowski and H. L. Smith, "Persistence in phage-bacteria communities with nested and one-to-one infection networks," *Discrete and Continuous Dynamical Systems - B*, vol. 22, 2017.

[171] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds., Springer-Verlag Limited, 2008, pp. 95–110.

[172] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[173] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd, "A rewriting system for convex optimization problems," *Journal of Control and Decision*, vol. 5, no. 1, pp. 42–60, 2018.

[174] K. Thamatrakoln, D. Talmy, L. Haramaty, C. Maniscalco, J. R. Latham, B. Knowles, F. Natale, M. J. L. Coolen, M. J. Follows, and K. D. Bidle, "Light regulation of coccolithophore host–virus interactions," *New Phytologist*, vol. 221, no. 3, pp. 1289–1302, 2019.

[175] J. M. Zobitz, A. R. Desai, D. J. P. Moore, and M. A. Chadwick, "A primer for data assimilation with ecological models using Markov Chain Monte Carlo (MCMC)," *Oecologia*, 2011.

[176] F. O. Aylward, J. M. Eppley, J. M. Smith, F. P. Chavez, C. A. Scholin, and E. F. DeLong, "Microbial community transcriptional networks are conserved in three domains at ocean basin scales," *Proceedings of the National Academy of Sciences*, vol. 112, no. 17, pp. 5443–5448, 2015.

[177] F. Ribalet, J. Swalwell, S. Clayton, V. Jimenez, S. Sudek, Y. Lin, Z. I. Johnson, A. Z. Worden, and E. V. Armbrust, "Light-driven synchrony of Prochlorococcus growth and mortality in the subtropical Pacific gyre," *Proceedings of the National Academy of Sciences*, vol. 112, no. 26, pp. 8008–8012, 2015.

[178] S. K. Hu, V. Campbell, P. Connell, A. G. Gellene, Z. Liu, R. Terrado, and D. A. Caron, "Protistan diversity and activity inferred from RNA and DNA at a coastal ocean site in the eastern North Pacific," *FEMS Microbiology Ecology*, vol. 92, no. 4, 2016.

[179] S. Charvet, W. F. Vincent, and C. Lovejoy, "Effects of light and prey availability on Arctic freshwater protist communities examined by high-throughput DNA and RNA sequencing," *FEMS Microbiology Ecology*, vol. 88, no. 3, pp. 550–564, 2014.

[180] D. Xu, R. Li, C. Hu, P. Sun, N. Jiao, and A. Warren, "Microbial eukaryote diversity and activity in the water column of the South China Sea based on DNA and RNA high throughput sequencing," *Frontiers in Microbiology*, vol. 8, p. 1121, 2017.

[181] R. S. Poretsky, I. Hewson, S. Sun, A. E. Allen, J. P. Zehr, and M. A. Moran, "Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre," *Environmental Microbiology*, vol. 11, no. 6, pp. 1358–1375, 2009.

[182] D. M. Karl, "Hidden in a sea of microbes," *Nature*, vol. 415, no. 6872, pp. 590–591, 2002.

[183] K. Nygaard and A. Tobiesen, "Bacterivory in algae: A survival strategy during nutrient limitation," *Limnology and Oceanography*, vol. 38, no. 2, pp. 273–279, 1993.

[184] Z. M. McKie-Krisberg, R. J. Gast, and R. W. Sanders, "Physiological responses of three species of Antarctic mixotrophic phytoflagellates to changes in light and dissolved nutrients," *Microbial ecology*, vol. 70, no. 1, pp. 21–29, 2015.

[185] Z. V. Finkel, J. Beardall, K. J. Flynn, A. Quigg, T. A. V. Rees, and J. A. Raven, "Phytoplankton in a changing world: Cell size and elemental stoichiometry," *Journal of Plankton Research*, vol. 32, no. 1, pp. 119–137, 2009.

[186] M. Hein, M. F. Pedersen, and K. Sand-Jensen, "Size-dependent nitrogen uptake in micro-and macroalgae," *Marine ecology progress series. Oldendorf*, vol. 118, no. 1, pp. 247–253, 1995.

[187] M. Gerea, C. Queimalinos, and F. Unrein, "Grazing impact and prey selectivity of picoplanktonic cells by mixotrophic flagellates in oligotrophic lakes," *Hydrobiologia*, vol. 831, no. 1, pp. 5–21, 2019.

[188] E. Bairey, E. D. Kelsic, and R. Kishony, "High-order species interactions shape ecosystem diversity," *Nature Communications*, 2016.

[189] J. Grilli, G. Barabas, M. J. Michalska-Smith, and S. Allesina, "Higher-order interactions stabilize dynamics in competitive network models," *Nature*, 2017.

[190] H. Hirano and K. Takemoto, "Difficulty in inferring microbial community structure based on co-occurrence network approaches," *BMC Bioinformatics*, vol. 20, 1 2019.

[191] A. Carr, C. Diener, N. S. Baliga, and S. M. Gibbons, "Use and abuse of correlation analyses in microbial ecology," *The ISME Journal*, 2019.

[192] L. L. Thurman, A. K. Barner, T. S. Garcia, and T. Chestnut, "Testing the link between species interactions and species co-occurrence in a trophic network," *Ecography*, 2019.

[193] K. Mainali, S. Bewick, B. Vecchio-Pagan, D. Karig, and W. F. Fagan, "Detecting interaction networks in the human microbiome with conditional Granger causality," *PLOS Computational Biology*, vol. 15, no. 5, pp. 1–21, 2019.

[194] A. D. Willis and B. D. Martin, "DivNet: Estimating diversity in networked communities," *bioRxiv*, p. 305 045, 2018.

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

[196] B. Ren, S. Bacallado, S. Favaro, S. Holmes, and L. Trippa, "Bayesian nonparametric ordination for the analysis of microbial communities," *Journal of the American Statistical Association*, vol. 112, no. 520, pp. 1430–1442, 2017.

[197] J. T. Morton, J. Sanders, R. A. Quinn, D. McDonald, A. Gonzalez, Y. Vazquez-Baeza, J. A. Navas-Molina, S. J. Song, J. L. Metcalf, E. R. Hyde, M. Lladser, P. C. Dorrestein, and R. Knight, "Balance trees reveal microbial niche differentiation," *mSystems*, vol. 2, no. 1, J. K. Jansson, Ed., e00162–16, 2017.

[198] B. D. Martin, D. Witten, and A. D. Willis, "Modeling microbial abundances and dysbiosis with beta-binomial regression," *arXiv e-prints*, arXiv:1902.02776, arXiv:1902.02776, 2019.

[199] R. Opgen-Rhein and K. Strimmer, "Learning causal networks from systems biology time course data: An effective model selection procedure for the vector autoregressive process," *BMC Bioinformatics*, vol. 8, 2 2007.

[200] J. Ernst and Z. Bar-Joseph, "STEM: A tool for the analysis of short time series gene expression data," *BMC Bioinformatics*, vol. 7, no. 1, p. 191, 2006.

[201] G. Sugihara, R. May, H. Ye, C. H. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," *Science*, vol. 338, no. 6106, pp. 496–500, 2012.

[202] E. B. Baskerville and S. Cobey, "Does influenza drive absolute humidity?" *Proceedings of the National Academy of Sciences*, vol. 114, no. 12, E2270–e2271, 2017.

[203] J. M. McCracken and R. S. Weigel, "Convergent cross-mapping and pairwise asymmetric inference," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 90, no. 6, p. 062 903, 2014.

[204] J. E. Middleton, J. Martinez Martinez, W. H. Wilson, and N. R. Record, "Functional dynamics of emiliania huxleyi virus-host interactions across multiple spatial scales," *Limnology and Oceanography*, vol. 62, no. 3, pp. 922–933, 2017.