Inferring phage-bacteria infection networks from time series data

Luis F. Jover,[1] Justin Romberg,[2] and Joshua S. Weitz[3, 1]

[1] *School of Physics, Georgia Institute of Technology, Atlanta, GA, USA*
[2] *School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA*
[3] *School of Biology, Georgia Institute of Technology, Atlanta, GA, USA*

(Dated: May 3, 2016)

In communities with bacterial viruses (phage) and bacteria, the phage-bacteria infection network establishes which virus types infects which host types. The structure of the infection network is a key element in understanding community dynamics. Yet, this infection network is often difficult to ascertain. Introduced over 60 years ago, the plaque assay remains the gold-standard for establishing who infects whom in a community. This culture-based approach does not scale to environmental samples with increased levels of phage and bacterial diversity, much of which is currently unculturable. Here, we propose an alternative method of inferring phage-bacteria infection networks. This method uses time series data of fluctuating population densities to estimate the complete interaction network without having to test each phage-bacteria pair individually. We use *in silico* experiments to analyze the factors affecting the quality of network reconstruction and find robust regimes where accurate reconstructions are possible. In addition, we present a multi-experiment approach where time series from different experiments are combined to improve estimates of the infection network and mitigate against the possibility of evolutionary changes to infection during the time-course of measurement.

I. INTRODUCTION

Bacterial viruses are ubiquitous and play an important ecological role at the global scale. In the oceans, viruses are responsible for a significant fraction of bacterial mortality and as a result have an effect on global geobiochemical cycles [1–4]. By killing bacteria, they redirect resources from higher trophic levels and back into the microbial resource pool. Yet, not all bacteria types are susceptible to all virus types. Each phage type potentially infects subset of hosts which can be presented as complex networks of infection [5]. Quantifying who infects whom remains essential to understanding how individual-based traits affect ecosystem-wide properties in complex environments.

For more than 60 years, the host range of phage, i.e., the types of host that a phage type infects, has been measured using plaque assays [6]. A plaque assay is an experimental method in which a growing culture of bacteria on an agar surface are exposed to phage. Clear "plaques" are formed whenever the phage can infect and lyse the target host. Plaque assays are considered the gold-standard for determining infection but are hard to scale-up to community levels. The principal reason is that the majority of phage and bacteria in a community sample are not yet available in culture. In response, a number of (partially) culture-independent methods have been proposed, including viral tagging [7, 8], PhageFISH [9], and polonies [9]. Each of these methods requires some degree of culturing or co-visualization of labeled particles, which also presents challenges for scaling-up to complex communities. Moreover, none of these methods leverage the information contained in the temporal dynamics of virus-bacteria systems.

The inference of interaction networks from system dynamics is a field of study with wide-spread applications from inference of gene regulatory networks [10, 11], and chemical reaction [12], to neural networks [13]. The key insights from one class of inference methods is that statistical patterns in dynamics, including cross-correlation and mutual-information, can be leveraged to infer interaction [14]. However, such correlation-based approaches can be of limited value when applied to high dimensional systems with nonlinear interaction. As an alternative, Shandilya et al. [15] showed a method for reconstructing interaction networks from discrete measurements of the time series in systems where the underlying functional form of the interactions is known. Similarly, Stein et al. [16] following the work of Monier et al. [17] used discretized Lotka-Volterra equations to estimate interaction networks, model parameters, and time dependent perturbations in competitive microbial communities.

Here, we extend the approach of Stein et al. [16] to phage-bacteria systems with antagonistic interactions. We derive the principles underlying the method and test its validity using *in silico* experiments. As we show, inferring realistic phage-bacteria infection networks in complex communities may be possible given appropriate deployment of existing technologies already available to estimate changing genotype densities over time.

II. METHOD

A. Model

We model the interaction between N_h host types and N_v virus types using a generalization of the Lotka-Volterra predator-prey equations [18, 19]. The densities of multiple host and virus types are described by a sys-

tem of differential equations that include the effect of competition between host types and the infection of host by multiple virus types [20, 21]:

$$\frac{dh_i}{dt} = r_i h_i \left(1 - \frac{\sum_{i'}^{N_h} a_{ii'} h_{i'}}{K}\right) - h_i \sum_{j}^{N_v} M_{ij}\phi_{ij} v_j, \quad (1)$$

$$\frac{dv_j}{dt} = v_j \sum_{i}^{N_h} \beta_{ij}\phi_{ij} M_{ij} h_i - m_j v_j, \quad (2)$$

The model consists of N_H equations of the form (1) for the density of each host type, h_i, and N_V equations of the form (2) for the virus densities, v_j. In this system: r_i is the growth rate of host i in the absence of viruses and other hosts, $a_{ii'}$ is the competitive effect of host i' on host i, K is the system-wide carrying capacity, ϕ_{ij} is the adsorption rate of virus j when attaching to host i, β_{ij} is the burst size of virus j when infecting host i, m_j is the decay rate of virus j. Finally M_{ij} is the infection matrix, i.e., a matrix representation of the infection network, which takes a value of 1 if host i is infected by virus j and zero otherwise.

B. Numerical simulations of the dynamics; infection network ensembles and model parameters

To study the performance of our reconstruction method, we simulated time series of systems where several hosts and virus types interact. We used MATLAB's ODE45 to numerically integrate systems of equations of the form described in Section II A. In doing so, we utilize both random infection networks and nested infection networks. Nested interaction networks are commonly observed in culture-based analyses, such that the host range of phage and the phage range of hosts form ordered subsets [22]. Following Jover et al. 2015 [23] we generated an ensemble of 100 infection matrices, each one with 10 host types and 10 virus types, spanning a spectrum of nestedness values. The infection matrices were generated by starting with a modular matrix and shifting interactions, through a random process, to regions that increase nestedness [23]. We also found feasible parameter sets (i.e., parameters with positive steady state densities) for each one of the infection matrices. We followed the procedure described in [23] to find feasible parameter sets. Namely, we select a subset of the model parameters and target densities (Table I) and use the steady state equations to solve for the rest of the parameters obtaining a feasible parameter set.

C. Infection network reconstruction

Our method for reconstructing infection networks requires discrete measurements of the dynamics resulting from the interaction of different host and virus types. This method extends the approach described in [16] to host-phage systems. We will use only the equations describing the dynamics of the viruses (equations of the form (2)). . We start by rewriting equation (2) in the form:

Parameter (unit)	Range\Value
ϕ_j (ml/(virus· d)	10^{-8} - 10^{-7}
β_j (viruses/cell)	10 - 50
H_i^* (cell/ml)	10^3 - 10^4
V_j^* (virus/ml)	10^6 - 10^7
K (ml)	$max(H_i^*) \times 100 = 10^6$

TABLE I: Parameter and target steady state density ranges used to find feasible parameter sets. Bacteria growth rates, r_i, and virus decay rates, m_j, were derived using the steady state equations and the parameters presented in this table (see Methods, given feasibility-based framework). The range denotes the limits of the uniform distributions used to generate parameters.

$$\frac{d\ln(v_j)}{dt} = \sum_{i}^{N_h} \beta_{ij}\phi_{ij} M_{ij} h_i - m_j. \quad (3)$$

We assume that we have $N+1$ measurements of the densities of all virus and host types in the system at times $[t_1, t_2, ..., t_{N+1}]$. For time step, $\Delta t_n = t_{n+1} - t_n$, we can write a discretized form of equation (3):

$$\frac{\Delta\ln(v_j(t_n))}{\Delta t_n} \approx \sum_{i}^{N_h} \tilde{M}_{ij} h_i(t_n) - m_j, \quad (4)$$

where we define the quantitative infection network $\tilde{M}_{ij} \coloneqq M_{ij}\phi_{ij}\beta_{ij}$, and $\frac{\Delta\ln v_j(t_n)}{\Delta t_n} \coloneqq \frac{\ln(v_j(t_{n+1})) - \ln(v_j(t_n))}{t_{n+1} - t_n}$. We can write an analogous equation to equation (4) for all time steps and all virus types in the system. All of these equations can be written in a compact form using a single matrix equation:

$$W \approx \left(\tilde{M}^{\intercal}\, \vec{m}\right) \begin{pmatrix} H \\ \vec{1} \end{pmatrix}, \quad (5)$$

where W and H are matrices with elements $W_{ij} = \frac{\Delta\ln v_i(t_j)}{\Delta t_j}$ and $H_{ij} = h_i(t_j)$, $\vec{m}$ is the column vector of decay rates with elements m_i, and $\vec{1}$ is a vector of ones with dimensions $1 \times N$. Given density measurements of the hosts and viruses we can reconstruct the quantitative infection network using equation (5). We solve the following minimization problem to obtain approximations $\tilde{M}_{rec}$ and $\vec{m}_{rec}$ of the quantitative infection matrix, $\tilde{M}$, and the decay rate vector $\vec{m}$:

$$\begin{aligned} \underset{(\tilde{M}^\intercal \vec{m})}{\arg\min} \quad & \left\| W - \left(\tilde{M}^\intercal \vec{m}\right) \begin{pmatrix} H \\ \vec{1} \end{pmatrix} \right\|_2 \\ \text{subject to} \quad & M_{ij} >= 0, \\ & m_i > 0. \end{aligned} \tag{6}$$

To solve this problem we used CVX, a package for specifying and solving convex problems [24, 25]. In this study we focus on the reconstruction of the quantitative infection network, but the method also infers decay rates for all virus types in the system. We use a normalized Frobenius distance between the original and reconstructed infection matrices as a metric of the quality of reconstruction, namely:

$$\text{Error}_{\text{rec}} = \frac{\|\tilde{M} - \tilde{M}_{\text{rec}}\|_2}{\|\tilde{M}\|_2}. \tag{7}$$

III. RESULTS

A. Reconstruction quality depends on the variability of the dynamics

We begin with an example in which there are 10 host types, 10 virus types and 20 virus-bacteria interactions. The effective infection rates ($\phi * \beta$) vary from 10^{-7} to 5×10^{-6}. Figure 1 shows an example of a successful infection network reconstruction using the method described in Section II C. The matrices W and H were calculated using measurements of the dynamics every 6 min for a total of 96 hours. This results in a reconstruction error $Error_{rec} = 0.01$. The method is able to correctly identify all of the interactions. The small error arises from differences in the inferred quantitative values.

In general, there are multiple factors affecting reconstruction quality. One important factor is the variability of the dynamics. For example, if the dynamics start at a fixed point, there would be no variability in the dynamics, the columns of the matrix H would all be identical and it would not be possible to infer the infection network. We test the effect of variability systematically by performing matrix reconstruction for an ensemble of matrices and different levels of variability. To control variability in the dynamics we change how far the initial densities are from the equilibrium densities. We initialize density of each host and virus type in the system at $x_0 = x_{eq}(1\pm\delta)$, where x_{eq} is the equilibrium density of a given type and δ is a free parameter that controls the distance from its equilibrium density. We calculated the mean reconstruction error for an ensemble of 100 matrices (Figure 2). The reconstruction error has a maximum at $\delta = 0$ (not shown for visualization purposes), which corresponds to starting the system at the equilibrium densities. The quality of the reconstruction increases as the initial conditions move away from the equilibrium densities.

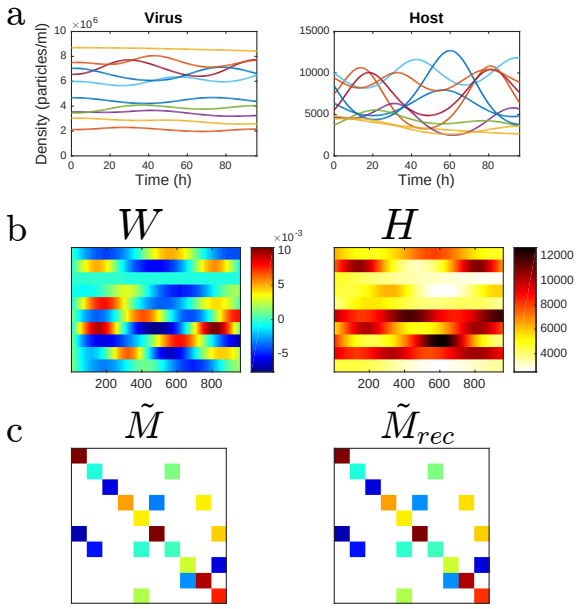

FIG. 1: Example of infection network reconstruction. (a) Virus and host dynamics for 96 hours. (b) Matrices W and H constructed by taking measurements of virus and host densities every 6 min as described in Section II C. (c) Original and reconstructed infection matrices ($Error_{rec} = 0.01$). A feasible parameter set was used in the simulation as described in Section II B

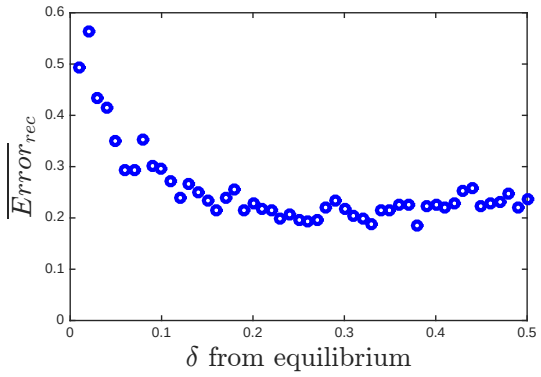

FIG. 2: Mean reconstruction error as a function of the fraction away from the equilibrium densities, δ, for an ensemble of 100 matrices. Feasible parameter set were used in the simulation as described in section II B

B. Reconstruction from multiple experiments: an alternative approach

We propose an improvement to the single experiment approach for reconstruction. In this alternative approach we combine measurements from different experiments to increase reconstruction quality. One key advantage of this approach is that, by increasing the number of exper-

iments used for reconstruction, we can reduce the total time and number of measurements per experiment. This is a crucial advantage in virus-bacteria systems, which are known to evolve rapidly [26–28]. In the multiple-experiment approach we generate a host matrix H and a virus matrix W by combining matrices from multiple experiments that differ only in their initial conditions (Figure 3). This extends equation (5) to include information from multiple experiments. Specifically, assuming that we perform p different experiments and calculate matrices $\{H_1, H_2, \ldots, H_p\}$ and $\{W_1, W_2, \ldots, W_p\}$ for each experiment, we can write the system:

$$(W_1 W_2 \ldots W_p) \approx \left(\tilde{M}^{\intercal}\, \vec{m}\right) \begin{pmatrix} H_1 & H_2 & \ldots & H_p \\ & & \vec{1} & \end{pmatrix}, \quad (8)$$

where $\vec{1}$ is a vector of ones with dimensions $1 \times (N_1 + N_2 + \ldots + N_p)$, assuming that we take N_i measurements from experiment i. Using the same minimization process presented in Section II C we can obtain an approximation, $\tilde{M}_{rec}$, of $\tilde{M}$.

Figure 4 compares the single and multiple experiments approach for three matrices with different nestedness values. We see how the multiple experiment approach results in lower reconstruction error for the three different cases. Figure 5 extends the comparison to an ensemble of 100 different matrices. We compare the multiple experiment approach to the average result of the single experiment approach. For a given matrix we performed 20 different experiments. Each experiment has the same infection matrix and the same model parameters but different initial conditions. We compare the performance of the reconstruction using each experiment individually vs. combining the measurements of the 20 experiments as described in equation (8). In this comparison we fix the total number of measurements; We compare the reconstruction error when using 960 measurements from a single experiment (measuring the dynamics every 6 minutes for 96 hours), against the performance when combining the first 48 measurement of all 20 experiments (every 6 minutes for 4.8 hours).

We performed the comparison for 100 different matrices (Figure 5). Multiple-experiment reconstruction results in lower error than the average single experiment reconstructions across a wide range of nestedness values. The multiple experiment approach is also more robust; it results in smaller variance in the reconstruction error. Performing more than a few experiments not only decreases the mean reconstruction error, but also decreases the standard deviation significantly (Figure 6). For the specific configuration studied here reconstruction error minimizes around 18 experiments.

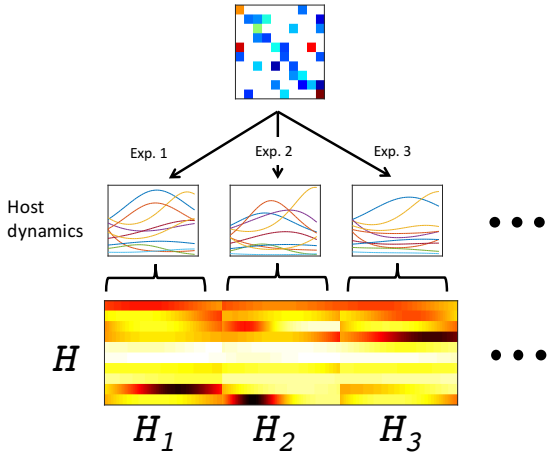

FIG. 3: Schematic representation of how H is calculated in the multiple-experiment approach. Multiple experiments are performed with the same matrix $\tilde{M}$ and different initial conditions.

C. Robustness of inference given noise in measurement

Here we evaluate the effect of measurement of white Gaussian noise on the quality of the inference. We follow the same procedure as in the noiseless case to reconstruct infection networks using multiple experiments. Figure 7 shows mean reconstruction error for an ensemble of 100 matrices as a function of the signal-to-noise ratio (SNR). We see that using 20 experiments and 48 measurements per experiment, network inference is possible for large signal-to-noise ratio, but reconstruction error increases significantly when the noise approaches 10% of the signal (SNR = 10 dB).

IV. DISCUSSION

We presented a theory-driven method to estimate host-phage infection networks in an community with multiple virus and host types. Current experimental techniques to measure such networks are difficult to scale to large systems. In addition, techniques that depend on isolation of viruses and/or hosts capture only a subset of potential interactions present in natural environments. Our approach addresses these limitations by using time-series measurements of experiments involving the whole virus-bacteria community. We also presented an improvement over the single experiment approach for infection network reconstruction. In the multiple-experiment approach we combined measurements from multiple experiments increasing the variability and lowering the reconstruction error. The multiple-experiment approach has the additional advantage of requiring shorter measurement time

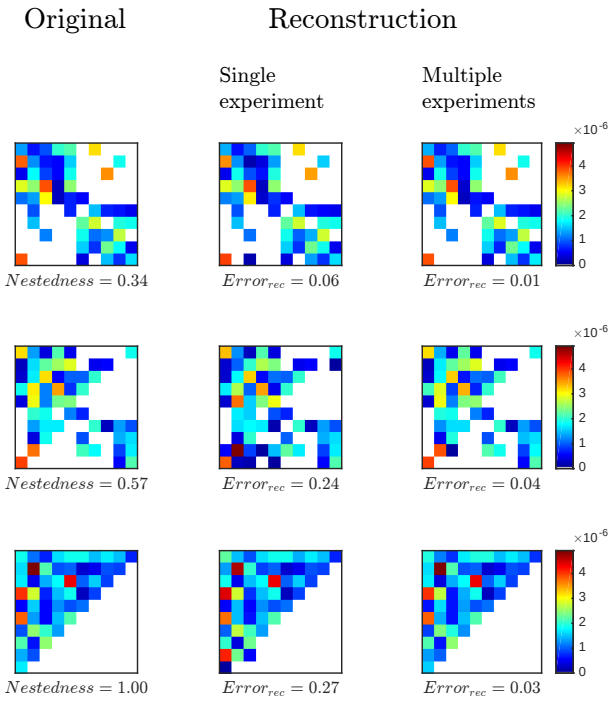

FIG. 4: Examples of reconstruction for three different matrices and two different methods. Each row shows the original matrix and the resulting reconstruction for each method. The first column shows the original matrices with values of nestedness (NODF): 0.34, 0.55, and 1 respectively. The middle column shows the reconstructed matrices and corresponding reconstruction errors for the single experiment approach using 960 measurements. The last column from the right shows the reconstructed matrices and corresponding errors for the multiple experiment approach using 20 experiments and 48 measurements per experiment. The total number of measurements is the same in the three different methods. The time between measurements is, $\Delta t = 6min$.

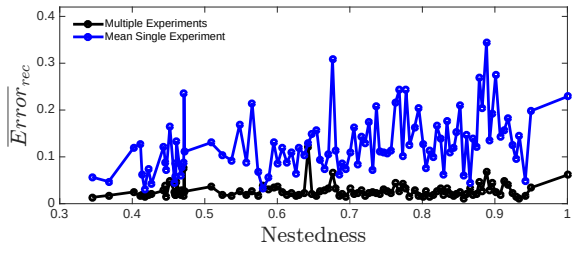

FIG. 5: Reconstruction error vs Nestedness for two different methods. Black line denotes the reconstruction error, $Error_{rec}$, using the multiple-experiments approach. Blue line describes the mean reconstruction error for the same 20 experiments used in the multiple-experiment approach but using each experiment separately. The total number of measurements is the same in both approaches.

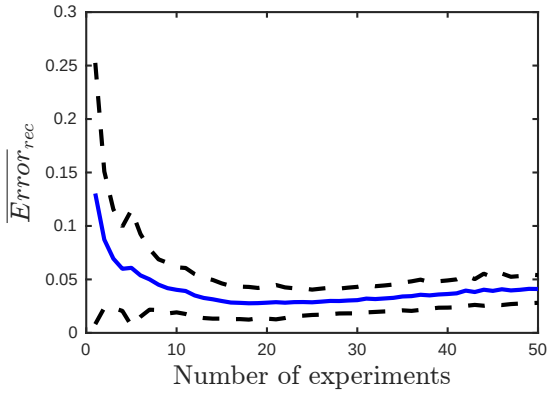

FIG. 6: Mean (blue line) and standard deviation (dotted line) of the reconstruction error for 100 infection matrices as a function of the number of experiments used in the multiple-experiment approach. Fixed number of total measurements (960). $\Delta t = 6min$.

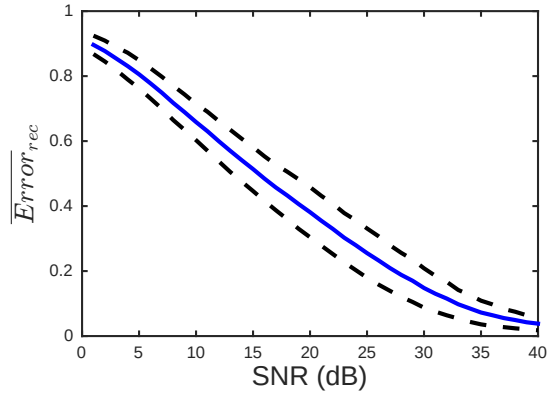

FIG. 7: Mean (blue line) and standard deviation (dotted line) of the reconstruction error for 100 different matrices as a function of the signal-to-noise ratio. The multiple experiment approach was used to reconstruct the matrix $\tilde{M}$. For each reconstruction, the matrices H and W were constructed using 20 runs with different initial conditions and 48 measurements per run. $\Delta t = 6$ min

per experiment. As a consequence, there is a lower probability of a host gaining resistance to a virus type or a virus developing the ability to infect a new host, increasing the chances of reconstructing the infection network of the target community.

The current method takes as input the measured densities of bacteria and phage in an environmental sample. Next-generation high-throughput sequencing techniques provide a means to characterize bacterial and viral communities in a variety of environmental samples [29–33]. In the past, such characterization has focused on phylogenetics groups, by using RNA and other marker genes. Such markers are insufficiently resolved with respect to

differences in relevant phenotypes, e.g., phage-bacteria infectivity. However, new computational approaches are increasingly able to resolve strain-level dynamics from metagenomic datasets [34, 35]. The increased used of quantitative piplines from sample to strain *density* for both bacteria and viruses will enable the kind of inference proposed here.

Our present approach uses the nonlinear dynamics of virus populations, to infer virus-bacteria infection networks. Nonetheless, this method can be expanded by including nonlinear bacterial population dynamics to infer competitive interactions between bacteria types and bacterial growth rates. In developing this method, it is important to keep in mind that the present approach is adapted to a specific functional form of the interactions in a virus-bacteria communities. Experimental verification (e.g., see Stein et a l. [16]) is necessary to test whether or not the dynamical model is a sufficiently robust representation of naturally occurring systems. Nevertheless, this study presents key steps towards an alternative way of determining who infects whom in a virus-bacteria community. This view has the potential to significantly reduce the experimental burden, e.g., we are able to infer $n_h \times n_v$ interactions by measuring the dynamics of $n_h + n_v$ organisms, and to overcome the limitations of culture-based approaches by inferring interactions without culturing.

V. ACKNOWLEDGMENTS

The authors thank Sam Brown and Joey Leung for their comments and suggestions, and acknowledge the support of the NSF grant No.OCE-1233760.

[1] Wilhelm, S. W. & Suttle, C. A. Viruses and nutrient cycles in the sea. *BioScience* **49**, 781–788 (1999).

[2] Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541–548 (1999).

[3] Suttle, C. A. Marine viruses - major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).

[4] Jover, L. F., Effler, T. C., Buchan, A., Wilhelm, S. W. & Weitz, J. S. The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nat. Rev. Microbiol.* **12**, 519–528 (2014).

[5] Weitz, J. S. *et al.* Phage–bacteria infection networks. *Trends Microbiol.* **21**, 82–91 (2013).

[6] Dulbecco, R. & Vogt, M. Plaque formation and isolation of pure lines with poliomyelitis viruses. *J. Exp. Med.* **99**, 167–182 (1954).

[7] Ohno, S. *et al.* A method for evaluating the host range of bacteriophages using phages fluorescently labeled with 5-ethynyl-2-deoxyuridine (edu). *Appl. Microbiol. Biotechnol.* **95**, 777–788 (2012).

[8] Deng, L. *et al.* Contrasting life strategies of viruses that infect photo-and heterotrophic bacteria, as revealed by viral tagging. *MBio* **3**, e00373–12 (2012).

[9] Allers, E. *et al.* Single-cell and population level viral infection dynamics revealed by phagefish, a method to visualize intracellular and free viruses. *Environ. Microbiol.* **15**, 2306–2318 (2013).

[10] Gardner, T. S., Di Bernardo, D., Lorenz, D. & Collins, J. J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105 (2003).

[11] Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E. & Guthke, R. Gene regulatory network inference: data integration in dynamic modelsa review. *Biosystems* **96**, 86–103 (2009).

[12] Arkin, A. & Ross, J. Statistical construction of chemical reaction mechanisms from measured time-series. *J. Phys. Chem.* **99**, 970–979 (1995).

[13] Hu, T., Leonardo, A. & Chklovskii, D. B. Reconstruction of sparse circuits using multi-neuronal excitation (rescume). In *Adv. Neural Inf. Process. Syst.*, 790–798 (2009).

[14] Rubido, N. *et al.* Exact detection of direct links in networks of interacting dynamical units. *New J. Phys.* **16**, 093010 (2014).

[15] Shandilya, S. G. & Timme, M. Inferring network topology from complex dynamics. *New Journal of Physics* **13**, 013004 (2011).

[16] Stein, R. R. *et al.* Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS. Comput. Biol.* **9**, e1003388 (2013).

[17] Mounier, J. *et al.* Microbial interactions within a cheese microbial community. *Appl. Environ. Microbiol.* **74**, 172–181 (2008).

[18] Lotka, A. J. *Elements of Physical Biology*, 92–94 (Williams & Wilkins Company, Baltimore, USA, 1925).

[19] Volterra, V. Fluctuations in the abundance of a species considered mathematically. *Nature* **118**, 558–560 (1926).

[20] Jover, L. F., Cortez, M. H. & Weitz, J. S. Mechanisms of multi-strain coexistence in host–phage systems with nested infection networks. *J. Theor. Biol.* **332**, 65–77 (2013).

[21] Weitz, J. S. *Quantitative Viral Ecology: Dynamics of Viruses and Their Microbial Hosts* (Princeton University Press, 2016).

[22] Flores, C. O., Meyer, J. R., Valverde, S., Farr, L. & Weitz, J. S. Statistical structure of host–phage interactions. *P. Natl. Acad. Sci.* **108**, E288–E297 (2011).

[23] Jover, L. F., Flores, C. O., Cortez, M. H. & Weitz, J. S. Multiple regimes of robust patterns between network structure and biodiversity. *Sci. Rep.* **5** (2015).

[24] Grant, M. & Boyd, S. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx` (2014).

[25] Grant, M. & Boyd, S. Graph implementations for nonsmooth convex programs. In Blondel, V., Boyd, S. & Kimura, H. (eds.) *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, 95–110 (Springer-Verlag Limited, 2008). `http://stanford.edu/~boyd/graph_dcp.html`.

[26] Middelboe, M. *et al.* Effects of bacteriophages on the population dynamics of four strains of pelagic marine

bacteria. *Microb. Ecol.* **42**, 395–406 (2001).

[27] Meyer, J. R. *et al.* Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science* **335**, 428–432 (2012).

[28] Weitz, J. S. & Wilhelm, S. W. Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol. Rep.* **4**, 17 (2012).

[29] Hannigan, G. D. & Grice, E. A. Microbial ecology of the skin in the era of metagenomics and molecular microbiology. *Cold Spring Harbor Perspect. Med.* **3**, a015362 (2013).

[30] Wood-Charlson, E. M., Weynberg, K. D., Suttle, C. A., Roux, S. & Oppen, M. J. Metagenomic characterization of viral communities in corals: mining biological signal from methodological noise. *Environ. Microbiol.* **17**, 3440–3449 (2015).

[31] Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).

[32] Minot, S. *et al.* The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* **21**, 1616–1625 (2011).

[33] Angly, F. E. *et al.* The marine viromes of four oceanic regions. *PLoS Biol.* **4**, e368 (2006).

[34] Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the illumina hiseq and miseq platforms. *ISME J.* **6**, 1621–1624 (2012).

[35] Luo, C. *et al.* Constrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* **33**, 1045–1052 (2015).