Inferring the quasipotential landscape of microbial ecosystems with topological data analysis

William K. Chang, Libusha Kelly

March 19, 2019

Abstract

The dynamics of high-dimensional, nonlinear systems drive biology at all scales, from gene regulatory networks to ecosystems. Microbial ecosystems ('microbiomes') exemplify such systems due to their richness and the small length- and time-scales of complex ecological and evolutionary dynamics. Microbes inhabit, respond to, and alter environments ranging from the human gut to the ocean. Here, using information theory and topological data analysis [1] (TDA), we model microbiome dynamics as motion on a potential energy-like landscape, called the quasipotential, identifying attractor states and trajectories that characterize ecological processes including disease progression in the human microbiome and geochemical cycling in the oceans. Our approach allows holistic analysis and prediction of large-scale dynamics in generalized complex systems that are difficult to reduce to their underlying interactions.

Numerous variables define the state of a microbiome, from the frequencies of microbial operational taxonomic units (OTUs) and their genetic alleles, which are decoupled due to genomic plasticity and horizontal gene transfer [2, 3]; to environmental conditions such as temperature, pH, and biochemical concentrations. A microbiome thus has a vast number of potential configurations. By contrast, systemic phenotypes, such as human gut infections or aquatic algal blooms, persist for much longer than bacterial generation time, with fewer degrees of freedom and configurations. Whether a human host is 'sick' or 'healthy' constitutes one degree of freedom with two possible configurations, for example. Within each phenotype or phase of a periodic process, community compositions can be diverse [4]. It is thus challenging to infer systemic phenotypes from community structure, or to predict community dynamics associated with events such as disease and ocean acidification.

One approach to analyzing microbiome dynamics has been to approximate the network of underlying pairwise interactions between OTUs by calculating the inverse covariance matrix from time series data, often as a basis for modeling population dynamics [5–7]. Due to the compositional nature of much microbiome data, direct fitting of population dynamics models can be misleading [8]. It is also challenging to experimentally characterize microbe-microbe interactions due to the unculturability of many bacteria [9]. Even where pairwise interactions between OTUs could be experimentally validated, higher-order interactions may be significant. As the number of OTUs increases, the combinatorial explosion of potential higher-order interactions renders interaction-based frameworks infeasible.

We adopted an alternative *quasipotential landscape* approach, which represents the configurations of a dynamical system—here, the possible compositions of a microbiome—as coordinates in phase space, where similar configurations are located close together. By analogy to statistical physics, the density of observations around each point in phase space is assumed to be inversely related to the value of a potential energy-like function, called the quasipotential. The system dynamics are considered as stochastic motion on the resultant manifold, with topological features corresponding to the probable configurations of the system and trajectories between them. Thus, the landscape encodes the underlying interactions between components without explicit assumptions (Fig. 1A). Related approaches have been used to investigate questions in ecology [10] and molecular biology [11, 12].

We used the TDA algorithm Mapper [13, 14] to infer the quasipotential landscapes for three published microbial time series data sets, two human gut microbiomes—one collected from seven cholera patients from disease through recovery [15], one from two mostly healthy adult males [16]—and one of marine *Prochlorococcus* communities spanning multiple depths collected from one site in the Atlantic Ocean (BATS) and one in the Pacific (HOT) [17]. Mapper represents

the underlying distribution of data in a metric space as an undirected graph, where each vertex comprises a non-exclusive subset of data points spanning a patch of phase space. An edge is drawn between each two vertices that share at least one point (Fig. 1A), representing connectivity between patches. As microbiome compositions are probability distributions, we used the square root of the Jensen-Shannon divergence as a metric [18]. For specifics of inputs to Mapper used, see Methods.

We estimated the quasipotential for each vertex by calculating the k-nearest neighbors (kNN) density [19] for each constituent data point i:

$$\mathrm{kNN}(i,k) = \frac{\sum_j^k d_{ij}}{k} \tag{1}$$

where d_{ij} is the distance between points i and j, choosing k equal to 10% of the number of samples in each data set, rounded to the nearest integer. kNN varies inversely with density, making it a proxy for the quasipotential. For a vertex V representing n points, we define its quasipotential as

$$Q(V) = \frac{\sum_{i\in V}^n \mathrm{kNN}(i,k)}{n^2} \tag{2}$$

The n^2 term in the denominator compensates for the differing sizes of vertices.

We then defined basins of attraction on the landscape. We designated each vertex with lower Q than its neighbors to be a local minimum of the quasipotential. Connected vertices tied for minimum Q were each assigned to be a local minimum. To approximate a gradient, we converted the undirected Mapper graph to a directed graph, with each edge pointing from the the vertex with greater Q to the one with lower Q. For each non-minimum vertex, we found the graph distance d_g to each local minimum constrained by edge direction. We defined the *basin of attraction* B_x of a minimum V_x as the set of vertices V with uniquely shortest graph distance to V_x:

$$V \in B_x \text{ if } d_g(V,V_x) < d_g(V,V_y) \tag{3}$$

for all $x \neq y$ and $V_y \in M$, where M is the set of all local minima (Fig 1B).

We found the cholera phase space to be partitioned by clinical phenotype, i.e. diarrhea or recovery (Fig. 2A). The diarrhea region was further subdivided into two basins, 2 and 7 (Fig. 2B), with patients (except D) following a succession of 2 to 7 during diarrhea (Fig. 2C), suggesting that cholera presents a strong deterministic perturbation to the gut microbiome, and has a universal 'early' state (basin 2) and a universal 'late' state (basin 7) with distinct communities in these patients. Generally, patients occupied basin 7 for longer than they did basin 2, suggesting that the stability of the late state in a given patient influences disease duration.

To quantify stability, we calculated a temporal correlation function for each basin-patient pair during the diarrhea phase. Given that a system occupied basin B_x at time t, we defined the temporal correlation to be the expectation that it will still (or again) occupy basin B_x at time $t+\tau$:

$$f_x(t) = \begin{cases} 1 & \text{if system is associated with basin } B_x \text{ at time } t \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

$$\mathrm{corr}_x(\tau) = \langle f_x(t+\tau) \rangle \tag{5}$$

for all sampled intervals of length τ, where $f_x(t) = 1$. Where a data point is associated with multiple basins, we weigh the association with each basin as $f'_x(t) = \frac{1}{p} f_x(t)$, with p the total number of unique basins associated with the system at time t, with the unassigned/unstable state regarded as a single distinct basin. Monotonically decreasing correlation functions indicate metastability; slopes become more negative with decreasing stability. We found non-monotonic correlation functions for basin 7 in patients A, C, and E, coinciding with increased duration of diarrhea, with patients B and F exhibiting the expected monotonicity (Fig. 2D). This indicated that patients A, C, and E repeatedly entered and exited basin 7, and that prolonged diarrhea in these three patients may not be due to stabilization of basin 7, but instability or inaccessibility of alternative, healthy states.

In contrast to the cholera data set, the two healthy adult gut microbiome time series from David *et al.* [16] were separated by subject (Fig. 3A). Both subjects' microbiomes experienced perturbations: subject A traveled from his residence in the United States to southeast Asia, twice experiencing traveller's diarrhea; and subject B, also based in the US, suffered an acute infection

by *Salmonella*. Previous studies [16, 20] noted that, while the microbiome of A returned to its original state after travel, recovery from *Salmonella* left the microbiome of B in an alternative state. Confirming this, we found that subject A occupied the same regions of phase space before and after travel, while subject B occupied disjoint regions before and after infection (Fig. 3B). The regions each constituted several basins of attraction (Fig. 3C), suggesting that the clinical 'healthy' phenotype of an individual is a probability over multiple states.

The coarse-graining of microbiome compositions by basins intuitively suggests a notion of dynamic stability, defined as a stationary probability across basins between time windows. Subject A occupied basins with stationary probability before and after travel, exhibiting dynamic stability (Fig. 3D). Temporal correlations showed that subject A, as well as subject B before infection, repeatedly visited the same set of basins; in contrast, subject B after infection transiently occupied several basins without repetition, indicating clinical recovery without restoration of dynamic stability (Fig. 3D).

In contrast to the human gut, the *Prochlorococcus* phase space was organized by gradients of depth (Fig. 4A) and temperature (Supporting Fig. 4), indicating that, in these environments, small changes to environmental conditions result in small changes to community structure. The phase space possessed multiple basins of attraction (Fig 4B), with basin 4 largely representing shallow fractions of the water column $\leq$ 100m; basins 2, 3, and 6 deeper fractions; and basin 1 intermediate depths. Basin 5 represented an infrequently-occupied region sampled only by the 140m fraction at BATS on January 27, 2004, and by the 125m fraction at HOT on January 31, 2008. As such, basin 5 possibly constitutes an alternative state for deep water fractions in mid-winter. Communities differing in depth rarely shared compositions, and transitioned between basins, in many cases periodically across calendar years (Fig. 4C), suggesting that some communities experienced abrupt periodic shifts in environmental conditions due to geochemical events.

It is known that the BATS water column undergoes an annual late winter upwelling [17], intermixing communities that otherwise inhabit different depth depths, and homogenizing environmental conditions across depths. We predicted that mixing would drive communities at all depths at BATS to converge on a common state, while no convergence would be observed at HOT. Accordingly, we observed a transition to basin 1 by all depths at BATS in January of each year. After June, depths 1-20m and 120-200m relax toward basins characteristic of shallow and deep depth fractions, respectively, while basin 1 persists longer in intermediate depths 40-100m. By contrast, the probability of a given depth fraction at HOT occupying any basin remains uniform over the calendar year; the distribution is especially stationary for shallow depths (Fig. 4C). This periodicity was also evident in periodic correlation functions for BATS, and non-periodic for HOT (Fig. 4D).

In conclusion, our use of TDA to map the microbial quasipotential landscape revealed the role of latent clinical and environmental variables [21], such as disease state and annual phase, in organizing microbiomes over time. The quasipotential model assumes ergodicity with symmetrical noise weak relative to the drive of gradient descent; where perturbations are strong and the system is kept far from metastability, the quasipotential would fail to predict the dynamics. Nolting [10] and Abbott [22] discuss these caveats in detail. TDA inference of the quasipotential naturally depends on the quantity and quality of data; subsampling analyses suggest that our biological conclusions here are robust (Supplemental Material). Finally, we note that Mapper is a recently developed method targeted at high-volume, high-dimensional data and, as such, the theoretical limits of its robustness and usefulness regarding smaller data sets remain untested. We recommend further development of topology-related methods using population dynamics simulations with known ground truths. Regardless, we have shown topological methods to be useful for holistic analysis of dynamics in complex biological systems where mechanistic details are unclear. We expect topological methods to facilitate use of quantitative methods such as Markov models [23, 24] and critical transition theory [25–29] in predicting large-scale dynamics of microbiomes and other biological systems.

Methods

Details of Mapper analysis

Briefly, Mapper bins data using combinations of overlapping intervals for a set of filter values. Then, for each bin, it performs hierarchical clustering for all pairwise distances between data

points within that bin. It creates a histogram of branch lengths using a predefined number of bins, and uses the first empty bin in the histogram as a cutoff, separating the hierarchical tree into single-linkage clusters. These clusters are represented as vertices in the Mapper graph.

The algorithm requires three types of inputs:

1. a pairwise distance matrix between each pair of data points;
2. the output of a set of filter functions, each of which maps each data point to a scalar value;
3. a set of hyperparameters, specifying the number of intervals for each filter function; the percent overlap between adjacent intervals; and the number of bins used to determine a cutoff within each combination of filter intervals.

For the filter functions used by Mapper to bin data points, we perform principal coordinate analysis (PCoA, also known as classical multidimensional scaling) in two dimensions on the pairwise distance matrix, and used the ranked values of principal coordinates 1 and 2 as the first and second filter values for Mapper, following Rizvi *et al.* [13].

As the Mapper algorithm is relatively new, there are currently no standard protocols to optimize the values of the hyperparameters. For our purposes, it was important that the algorithm achieved a sufficiently high resolution in partitioning data, but also adequately represented connections between regions of phase space. We set hyperparameter values for each data set according to the following heuristic:

1. the largest vertex in the resultant Mapper graph should represent no more than $\approx 10\%$ of the total number of data points in the set;
2. the number of connected components representing only one data point should be minimized.

Table 1 lists hyperparameter values used to analyze each data set.

Data set	# intervals for (rank(PCo1), rank(PCo2))	% overlap	# bins
Cholera	$(15, 15)$	70	10
Two healthy adult males	$(30, 30)$	50	10
Prochlorococcus	$(20, 20)$	60	10

Table 1: Hyperparameters used to generate the Mapper representation of each data set.

We wish to note that, while performing 2D PCoA on a high-dimensional dataset such as 16S relative abundances will almost invariably lead to loss of information, the local clustering performed by Mapper uses the original distance matrix, and thus uses all available information in inferring local structure.

Details of basin assignment

Vertices equidistant to multiple minima were defined to be unstable regions unassigned to any basin. Multiple connected minima were defined as belonging to the same basin. Notably, one data point may be associated with multiple vertices and basins, or an unstable region and at least one basin: we interpreted this to mean that the point is near a saddle point separating basins, and as the 'true' coordinates of the saddle point are unknown, the data point is assigned to *all* such basins and/or an unstable region with uniform weight.

Software

All analysis and visualization was performed in the open-source programming language R (`http://cran.r-project.org`). The main repository for the study can be found on GitHub, at `http://github.com/kellylab/microbial-landscapes`.

An open-source implementation of Mapper in R, `TDAmapper`, was used for the main analysis and can be found at `http://github.com/wkc1986/TDAmapper`. This package was forked from the original implemented by Daniel Müllner which is maintained by Paul T. Pearson and can be found at `https://github.com/paultpearson/TDAmapper`. Other R packages used include `cowplot`, `data.table`, `ggplot2`, `ggraph`, `igraph`, `philentropy`, `tidygraph`, and `tidyverse`.

Acknowledgements

We thank Dave van Insberghe of the Martin Polz lab at the Massachusetts Institute of Technology for processing and performing OTU calling on the data from Hsiao *et al.* [15] and David *et al.* [16].

References

1. Wasserman, L. Topological Data Analysis. *arXiv:1609.08227 [stat]*. arXiv: 1609.08227. http://arxiv.org/abs/1609.08227 (2018) (Sept. 26, 2016).
2. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299–304. ISSN: 0028-0836 (May 18, 2000).
3. Polz, M. F., Alm, E. J. & Hanage, W. P. Horizontal Gene Transfer and the Evolution of Bacterial and Archaeal Population Structure. *Trends in genetics : TIG* **29,** 170–175. ISSN: 0168-9525 (Mar. 2013).
4. Fuhrman, J. A., Cram, J. A. & Needham, D. M. Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology* **13,** 133–146. ISSN: 1740-1526 (Mar. 2015).
5. Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology* **8,** e1002687. ISSN: 1553-7358 (Sept. 20, 2012).
6. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *arXiv preprint arXiv:1408.4158*. http://arxiv.org/abs/1408.4158 (2015) (2014).
7. Stein, R. R. *et al.* Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. *PLoS Comput Biol* **9,** e1003388 (Dec. 12, 2013).
8. Silverman, J. D., Washburne, A. D., Mukherjee, S. & David, L. A. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* **6,** e21887. ISSN: 2050-084X (Feb. 15, 2017).
9. Pande, S. & Kost, C. Bacterial Unculturability and the Formation of Intercellular Metabolic Networks. *Trends in Microbiology*. ISSN: 0966-842X. doi:10.1016/j.tim.2017.02.015. http://www.sciencedirect.com/science/article/pii/S0966842X17300525 (2017) (2017).
10. Nolting, B. C. & Abbott, K. C. Balls, cups, and quasi-potentials: quantifying stability in stochastic systems. *Ecology* **97,** 850–864. ISSN: 1939-9170 (Apr. 1, 2016).
11. Gu, S. *et al.* The Energy Landscape of Neurophysiological Activity Implicit in Brain Network Structure. *Scientific Reports* **8,** 2507. ISSN: 2045-2322 (Feb. 6, 2018).
12. Zhou, J. X., Aliyu, M. D. S., Aurell, E. & Huang, S. Quasi-potential landscape in complex multi-stable systems. *Journal of The Royal Society Interface* **9,** 3539–3553. ISSN: 1742-5689, 1742-5662 (Dec. 7, 2012).
13. Rizvi, A. H. *et al.* Single-cell topological RNA-Seq analysis reveals insights into cellular differentiation and development. *Nature biotechnology* **35,** 551–560. ISSN: 1087-0156 (June 2017).
14. Singh, G., Mémoli, F. & Carlsson, G. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *Eurographics Symposium on Point-Based Graphics,* 11 (2007).
15. Hsiao, A. *et al.* Members of the human gut microbiota involved in recovery from Vibrio cholerae infection. *Nature* **515,** 423–426. ISSN: 0028-0836 (Nov. 20, 2014).
16. David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biology* **15,** R89. ISSN: 1474-760X (2014).
17. Malmstrom, R. R. *et al.* Temporal dynamics of Prochlorococcus ecotypes in the Atlantic and Pacific oceans. *The ISME Journal* **4,** 1252–1264. ISSN: 1751-7362 (Oct. 2010).
18. Koren, O. *et al.* A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. *PLOS Computational Biology* **9,** e1002863. ISSN: 1553-7358 (Jan. 10, 2013).
19. Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern classification* Google-Books-ID: YoxQAAAAMAAJ. 688 pp. ISBN: 978-0-471-05669-0 (Wiley, 2001).

20. Gonze, D., Coyte, K. Z., Lahti, L. & Faust, K. Microbial communities as dynamical systems. *Current Opinion in Microbiology* **44,** 41–49. ISSN: 1369-5274 (Aug. 1, 2018).

21. Nguyen, L. H. & Holmes, S. Bayesian Unidimensional Scaling for visualizing uncertainty in high dimensional datasets with latent ordering of observations. *BMC Bioinformatics* **18,** 394. ISSN: 1471-2105 (Sept. 13, 2017).

22. Abbott, K. C. & Nolting, B. C. Alternative (un)stable states in a stochastic predator–prey model. *Ecological Complexity* **32,** 181–195. ISSN: 1476945X (Dec. 2017).

23. DiGiulio, D. B. *et al.* Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences* **112,** 11060–11065. ISSN: 0027-8424, 1091-6490 (Sept. 1, 2015).

24. Faure, M. & Schreiber, S. J. Quasi-stationary distributions for randomly perturbed dynamical systems. *The Annals of Applied Probability* **24,** 553–598. ISSN: 1050-5164 (Apr. 2014).

25. Dai, L., Vorselen, D., Korolev, K. S. & Gore, J. Generic Indicators for Loss of Resilience Before a Tipping Point Leading to Population Collapse. *Science* **336,** 1175–1177. ISSN: 0036-8075, 1095-9203 (June 1, 2012).

26. Dakos, V. & Bascompte, J. Critical slowing down as early warning for the onset of collapse in mutualistic communities. *Proceedings of the National Academy of Sciences* **111,** 17546–17551. ISSN: 0027-8424, 1091-6490 (Dec. 9, 2014).

27. Leemput, I. A. v. d. *et al.* Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences* **111,** 87–92. ISSN: 0027-8424, 1091-6490 (Jan. 7, 2014).

28. Scheffer, M. *et al.* Anticipating Critical Transitions. *Science* **338,** 344–348. ISSN: 0036-8075, 1095-9203 (Oct. 19, 2012).

29. Scheffer, M., Carpenter, S. R., Dakos, V. & Nes, E. v. Generic Indicators of Ecological Resilience: Inferring the Chance of a Critical Transition. *Annual Review of Ecology, Evolution, and Systematics* **46,** null (2015).

1 Figures

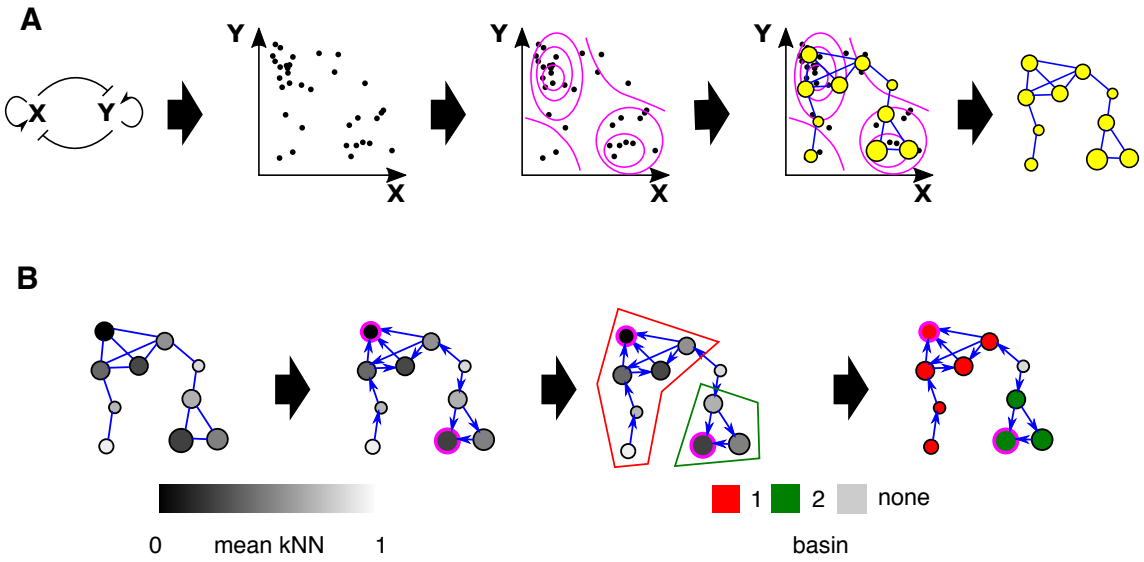

Figure 1: A. Using the Mapper algorithm to infer the quasipotential landscape of a toy ecosystem. The mutually antagonistic interaction between species X and Y leads to denser sampling of the phase space where either X or Y is abundant and the other is rare than in other regions; configurations in which X and Y are similar in density are unstable, as small uncertainties in numerical advantage will eventually lead to the dominance of one species over the other. This probability density is analogous to an inverse of the quasipotential landscape. Mapper infers a 'skeleton' of density from the data represented as a point cloud. This representation preserves major features of the landscape such as the two densely-sampled clusters, representing attractor states and their basins of attraction, separated by a sparsely-sampled unstable region. B. Identification of local minima and basins of attraction in the Mapper graph shown in A. Data density for each vertex is calculated as the mean kNN distance for samples associated with that vertex. The graph is converted to a directed graph, with each edge pointing in the direction of increasing kNN. A local minimum, highlighted in pink, is defined as a vertex that has lower kNN than all its neighbors. Finally, the basin of attraction associated with a local minimum is defined as the set of vertices that have uniquely shortest directed graph distance to that minimum. Non-minima vertices with equal graph distances to multiple local minima are unassociated with any basin (grey).

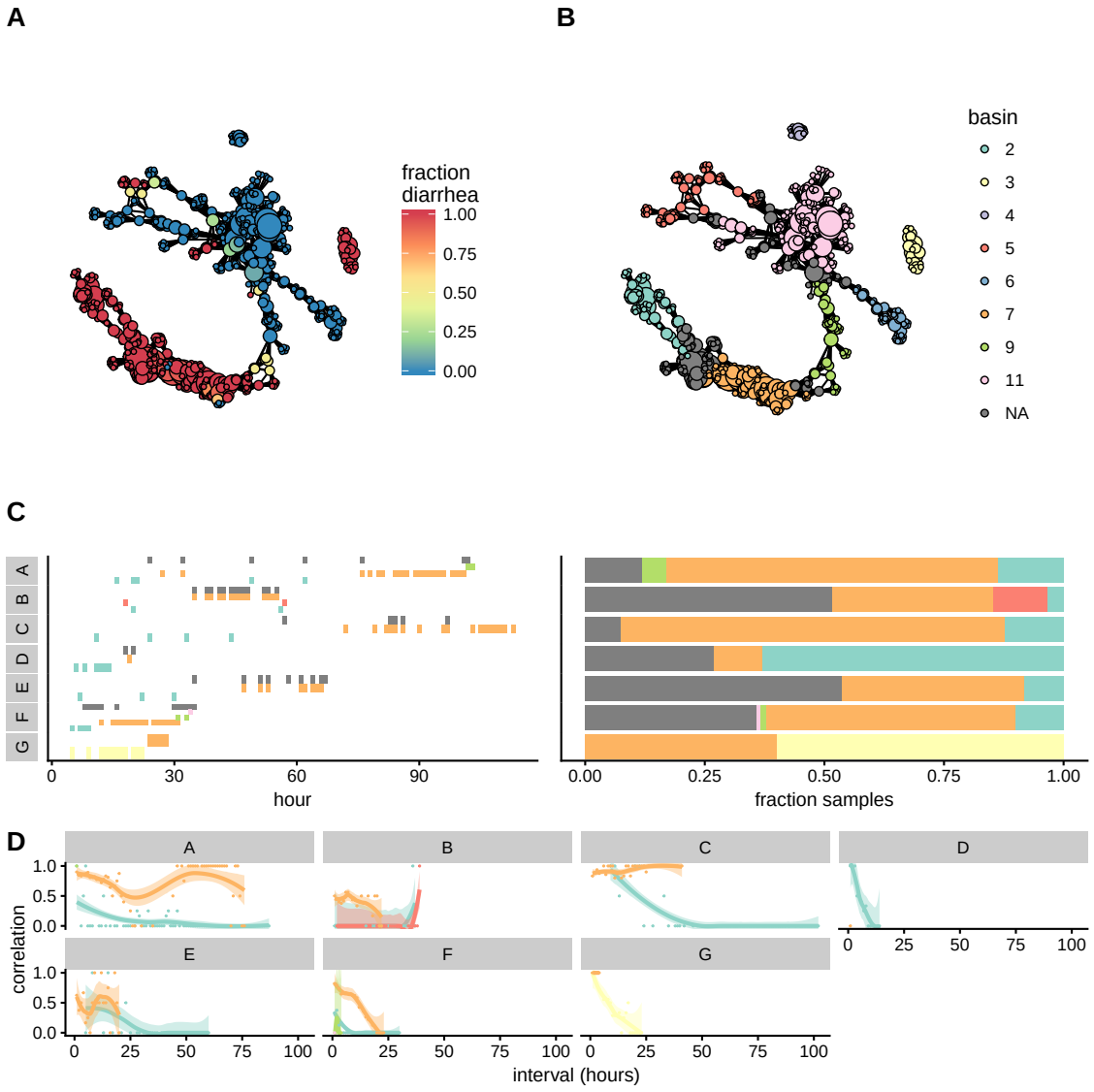

Figure 2: A. Mapper representation of the combined cholera data reveals disease- and healthy-associated neighborhoods of the phase space. Connected components of the Mapper graph representing only one sample are not shown. Disjoint regions of phase space are represented as separate connected components. B. Partitioning of the phase space into basins of attraction. Vertices unassigned to any basin are colored in grey. C. Left: progression of subject compositions during the diarrhea phase by basin of attraction, showing persistence of basin occupation over time. Y axis and color indicate basin index, with color indexing as in B. Where a sample was associated with multiple basins, all were included. Right: occupancy of basins during the diarrhea phase for each subject. D. Temporal correlation function for the diarrhea phase of each subject. Lines: smoothed empirical mean; ribbons: standard error of the mean. Values outside the range of $0 \leq y \leq 1$ omitted.

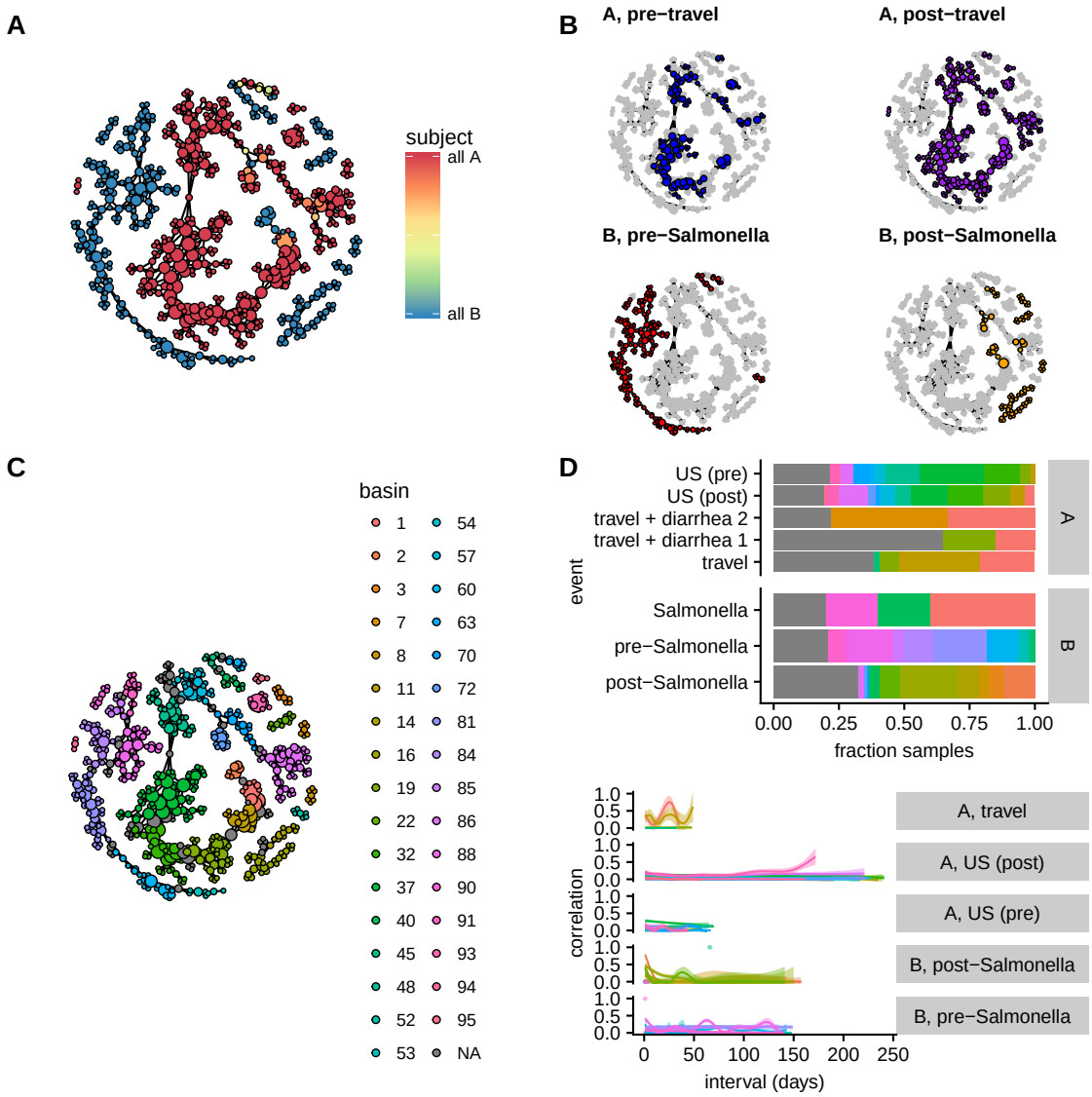

Figure 3: A. Mapper representation of the combined daily time series of two healthy adult human gut microbiomes shows that the phase space is strongly partitioned by individual. Connected components of the Mapper graph representing only one sample are not shown. B. The gut microbiome of subject A occupies the same neighborhoods of phase space before and after perturbation, while that of B occupies two nearly-disjoint neighborhoods. A. Basins of attraction in the phase space spanned by two mostly healthy adult male gut microbiomes. B. Top: occupancy of basins for different events. A similar occupancy distribution is observed for subject A for days spent in the US pre- and post-travel, indicating reversion to a reproducible healthy state. Similarly, occupancy distribution is similar between the two instances of diarrhea while traveling. In contrast, subject B shows different occupancy distributions for healthy state samples pre- and post-*Salmonella* infection, in agreement with prior analyses suggesting that infection drove an irreversible transition to an alternate stable state. Bottom: temporal correlation function for each basin during each event in the 'healthy' phases of each subject. Lines: smoothed empirical mean; ribbons: standard error of the mean.

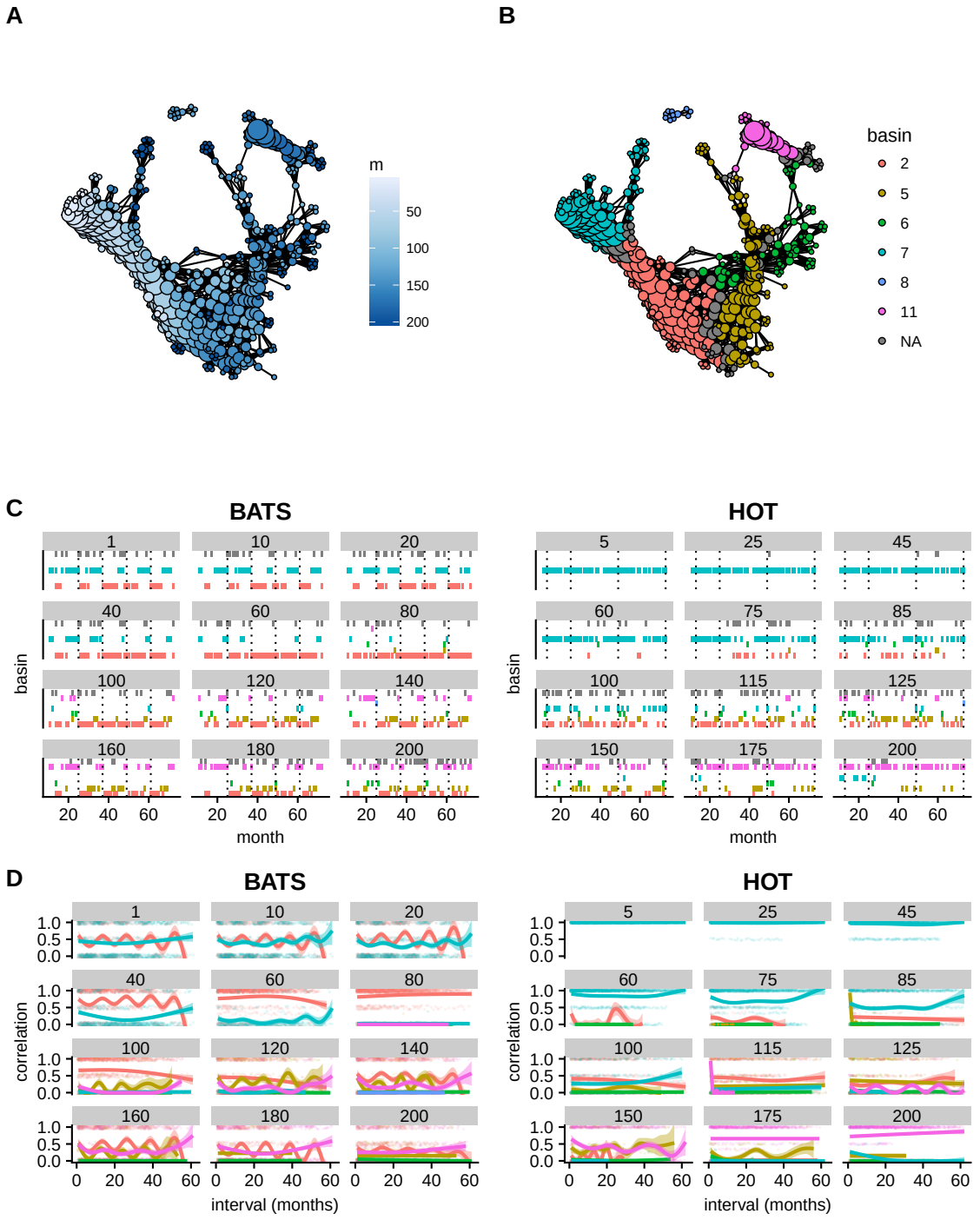

Figure 4: The combined phase space of two *Prochlorococcus* communities inhabiting the Atlantic and Pacific Oceans, respectively. Connected components of the Mapper graph representing only one sample are not shown. A. Mean depth varies continuously across the phase space. B. Partitioning of the phase space into basins of attraction. C. Time series per site-depth fraction. Dotted lines indicate samples during January. Colors indicate basins as in B. D. Temporal correlation functions for each basin per site-depth fraction.