PREDICTING MICROBIAL INTERACTIONS WITH MODELLING APPROACHES

LI CHENHAO

(B. Sc. (Hons.), NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

May 2019

Supervisors:
Professor Wong Lim Soon, Main Supervisor
Associate Professor Niranjan Nagarajan, Co-Supervisor

Examiners:
Dr Swaine Lin Chen
Dr Rohan Benjamin Hugh Williams
Dr Jonathan Goke, Genome Institute of Singapore

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Signed:_____

Date:_____

LI CHENHAO

National University of Singapore

ABSTRACT

Predicting Microbial Interactions with Modelling Approaches

by

Li Chenhao

Computational modelling represents an attractive avenue for scalable data-driven analysis of microbial community function and dynamics. The practicability of utilising *in silico* modelling to directly learn ecological models is now conceivable with the vast amount of publicly available microbial community profiling data. Despite the promise, progress has been relatively muted as existing model inference algorithms relied on absolute abundances rather than the relative measurements generated with high-throughput microbial profiling.

We introduce a new algorithm for learning generalised Lotka-Volterra models (gLVMs) from longitudinal microbial profiling data by coupling **b**iomass **e**stimation and model inference in an **e**xpectation-**m**aximization-like algorithm (BEEM). We show that BEEM outperforms existing methods for inferring gLVMs, while simultaneously eliminating the need for absolute abundances as input. BEEM's application to previously inaccessible public datasets (due to lack of information on absolute abundances) allowed us for the first time to construct ecological models of microbial communities in the human gut on a per individual basis, revealing personalised dynamics and keystone species.

For cross-sectional microbial community profiles, correlation based strategies have been the most widely used approach to inferring microbial interactions. However, our benchmarking evaluations showed that correlation based methods have varied performance for predicting interactions. To better infer interactions and construct

models from cross-sectional data, we developed an extension of BEEM (BEEM-static). BEEM-static improves inference accuracy by automatically identifying samples that are close to steady states for training. In addition to interaction predictions, BEEM-static also enable instantaneous growth inference for each species member of the community. BEEM-static outperforms correlation based methods for modelling cross-sectional data, substantially improving the prediction accuracy of directed interactions.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisors Dr. Niranjan Nagarajan and Professor Wong Limsoon, and other members of my thesis advisory committee, Professors Lisa Tucker-Kellogg and Ken Sung, for their constant support and intellectual inputs on my projects. My special thanks go to Dr. Niranjan Nagarajan, who offered me the opportunity to pursue my PhD studies while working as a bioinformatics specialist at Genome Institute of Singapore (GIS).

I would like to extend my thanks to all our lab members from Computational and System Biology group 5 at GIS, thank you all for making my work and study a fantastic experience in my life.

Thank you to my best friend since I was 14 years old and now my girlfriend, Miss Ma Yuan, for all her love and understanding during the past five years of long-distance relationship.

Finally, my appreciations go to my family members and colleagues at NUS and GIS. I would not have made the way through this tough but fruitful journey without the understanding and help from all of you.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Appendices

1 INTRODUCTION

1.1 Background

The origin of life began in the form of microorganisms three billion years ago on our planet[1]. These tiny microbes have then witnessed the rise and fall of life forms, and they have survived and expanded to occupy almost every niche on earth. Our understanding of these microbes, nevertheless, has a short history, dating back only to the 1670s when the first microbe (a species of fungi) was observed under a microscope[2]. In the 19th century, studying the growth and function of a single microbe was made possible thanks to the invention of agar plates and microbial culturing techniques[3]. However, even until now, an estimated 90% of bacteria are not culturable in the laboratory, greatly hindering our understanding of the composition and function of complex communities formed by diverse microbial species on our planet[4].

To circumvent this challenge, researchers turn to molecular techniques to investigate the microbial world. The deoxyribonucleic acid (DNA) molecules which code the genetic information of all life forms can be extracted from any environment and read out on a sequencer[5]. Computational analysis of the sequencing data can help us to identify the microbial species present in the sample, using DNA sequences as molecular barcodes[6–8]. In addition, the relative abundance of each microbe in the

environment can be measured by counting the frequency of such barcodes. A new field of research, metagenomics – the study of total environmental DNA, has thus emerged and was developed into an essential tool to gain insights into microbial communities thanks to the advance in high-throughput sequencing technology[9]. Metagenomic techniques allow researchers to investigate the composition of all microbial species (also known as the microbiome) in a specific environment with sequencing data collected across different sites (cross-sectional data) or for one site through a period of time (longitudinal data)[10]. Intriguingly, many metagenomic studies suggest that microbial communities, though tiny in size, play vital roles in the environments they reside in.

Among different microbial communities of interest, the study of human associated microbiomes has drawn much attention recently. Thousands of microbial species that colonize human bodies have been identified and an increasing number of studies have shown that they are closely linked to our health status. Although researchers have linked various disorders or diseases with disrupted microbial community and identified specific disease related biomarkers, the mechanism for such associations is still poorly understood, thus limiting our ability to design microbiome-targeted intervention strategies.

One of the most critical aspects to understand the function of ecological communities formed by microbial species is to characterize interactions among them, with data-driven approaches based on metagenomic microbial abundance profiles becoming a widely used technique to identify candidate interactions. Analytical methods for this range from simply quantifying the association strength between microbial abundances to constructing mathematical models that describe the interplay between species. However, one of the critical challenges in the above analysis arises from the “compositional” nature of sequencing data, where the microbial abundances

measured are in proportions rather than absolute abundances. This compositional nature of the data can cause significant biases in analysis and preclude the use of ecological models for fitting the data[11,12].

In this thesis, we investigate the problem of predicting microbial interactions using computational approaches, focusing on the challenges in learning interactions from longitudinal and cross-sectional microbial profiling data. We start by providing a detailed survey of the literature on available methods for inferring microbial interactions (**Chapter 2**). We then describe a novel algorithm (BEEM) for accurately inferring interactions from longitudinal microbiome data and apply it to learn personalized microbiome models from densely sampled gut microbial profiles (**Chapters 3**). For cross-sectional data, we evaluate the performance of correlation based methods and establish that these methods perform poorly in terms of recapitulating ground truth microbial interactions (**Chapter 4**). Motivated by these findings, we further extend BEEM to work with cross-sectional microbiome data and showcase the significant improvements that it provides over existing methods (**Chapter 5**).

1.2 List of publications

This thesis contributes to the literature with the following publications (*: as first author):

1. C. Li*, K. M. K. Lim, K. R. Chng & N. Nagarajan. Predicting Microbial Interactions through Computational Approaches. *Methods* (2016)
2. T. V. Av-Shalom*, C. Li*, N. Nagarajan. Correlation based methods vary widely in their ability to correctly infer microbial interactions from microbiome survey data. Manuscript in preparation

3. C. Li*, K. R. Chng, T. V. Av-Shalom., L. Tucker-Kellogg & N. Nagarajan. An expectation-maximization-like algorithm enables accurate ecological modeling using longitudinal metagenome sequencing data. *bioRxiv* (2018)
4. C. Li*, N. Nagarajan. Accurate Inference of Ecological Models from Cross-Sectional Microbiome Sequencing Data. Manuscript in preparation

2 LITERATURE REVIEW – COMPUTATIONAL APPROACHES FOR PREDICTING MICROBIAL INTERACTIONS

2.1 Background

As an essential component in various ecosystems, microorganisms aggregate to form heterogeneous communities comprising of distinct proportions of diverse microbial entities, often referred to collectively as the *Microbiome*. Microorganisms in a microbiome do not live in isolation, but instead actively interact with other members within their community[13]. Taken as a whole, these interactions are a description of the overall function of the microbial community. As such, the characterization of microbial interactions is a key step towards the understanding of the community organization[14–16] and the engineering of microbial communities for biomedical[17,18] and industrial applications[19–21].

The pair-wise interaction between two microbes is the fundamental unit of microbial interactions. Such interactions can be categorized by their effect on the participants, i.e. positive, negative or neutral. In combination, there exist six core categories of interaction: mutualism (positive-positive), competition (negative-negative), antagonism (positive-negative), commensalism (positive-neutral), amensalism (negative-neutral) and neutralism (neutral-neutral)[22].

Traditionally, the investigation of microbial interactions required the use of laboratory experiments such as growth and co-culture assays[23–25]. However, the laborious nature of such methods renders them infeasible for large scale application. Computational approaches offer the opportunity to alleviate this issue by predicting interaction candidates for experimental validation[26]. These predictions can be based on various types of data such as the measured species abundances from high-throughput sequencing or reconstructed metabolic models for species communities. In addition, computational methods may also assist the collation of experimentally verified interactions from large compendiums of published literature. A graphical overview of computational approaches for predicting microbial interactions can be found in **Figure 2.1**.

Following the previous review on *in silico* microbial interaction inference methods[22], a number of new methods have since been proposed to address the various challenges in such a task. Here, we review the available computational approaches (grouped by the different types of data that they use) and the challenges that they address, discuss their advantages and limitations, and point out directions for future work in this area.

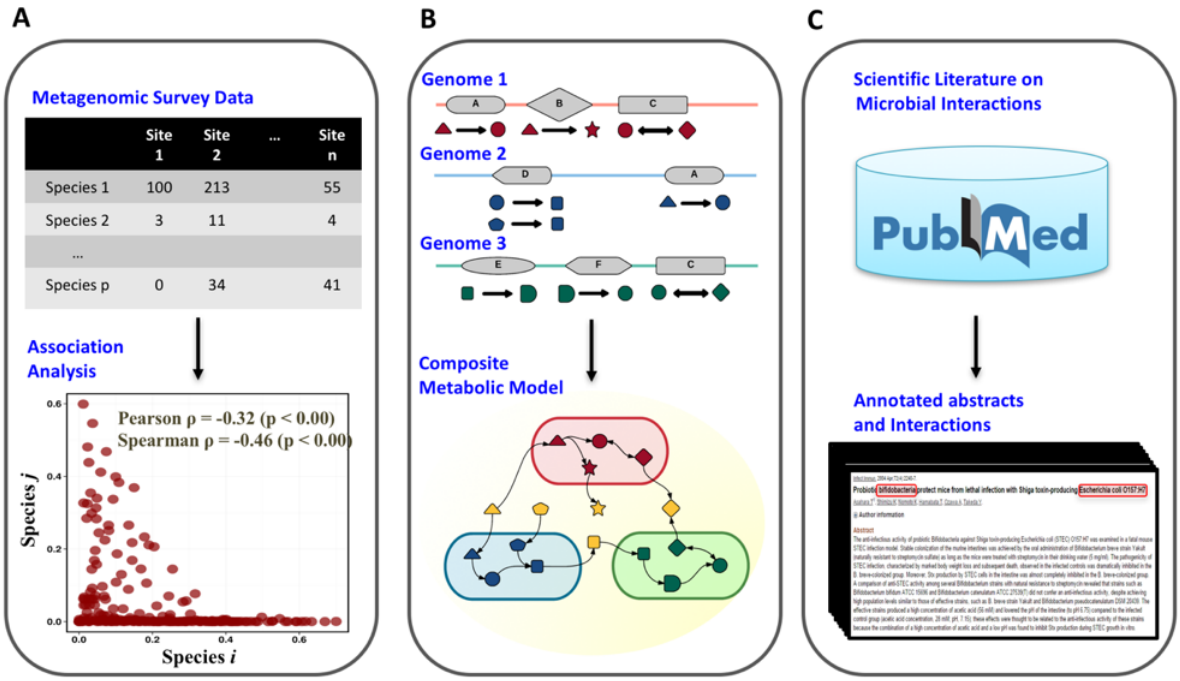

Figure 2.1 A graphical overview of computational approaches to predict microbial interactions.

(A) Microbial interactions are frequently inferred by observing correlations in species abundances in microbiome survey datasets as depicted here for a pair of species. (B) Interactions can also be predicted by reconstructing pathways from annotated genomes for each species and then jointly modeling community metabolism to identify metabolites that serve as interaction interfaces (show in yellow). Genes are depicted here as grey shapes while associated metabolites are shown as colored shapes. (C) Text mining of scientific literature databases (e.g. NCBI PubMed) is another approach for cataloguing microbial interactions that are experimentally validated and can serve as a gold-standard for the field.

2.2 Inferring interactions from microbiome survey data

Advances in high throughput sequencing technologies have made it possible to quantify the abundances of members in a microbial community in a relatively unbiased manner by sequencing marker genes or whole metagenomes (i.e. total DNA from a microbial community). The abundance of each species (or higher taxa) is then estimated by mapping raw reads to a reference database with taxonomically annotated complete or draft genomes and counting the reads assigned to the respective taxa. The data collected can then be further tabulated into a data matrix where each row represents read counts of a species across all the samples. To account for differences in sequencing depth (i.e. total number of reads generated for a sample), read counts are often normalized into proportions (relative abundances) by dividing by the column sums. Alternatively, the data matrix can be simplified to record only presence (1) or absence (0) information by setting a minimum threshold on read counts or relative abundances. Microbiome survey datasets can be collected across different sites or across different time points within the same site, with the techniques used to infer microbial interactions from them being somewhat distinct. In the next section we review methods that use survey data without a temporal component, referred to here as "cross-sectional" microbiome survey data.

2.2.1 Microbial interaction inference with cross-sectional microbiome survey data

Cross-sectional microbiome survey data provide a static view of the composition of microbiota across different sites. For human-associated microbiota, several recent studies have generated a significant amount of data across different patients and different body sites[27,28]. While many of these studies have focused on investigating the composition of microbial communities or identifying species associated with certain phenotypes, these datasets can also be used to infer interactions between species.

Although this can be a coarse-grained approach, inferring microbial interactions from available cross-sectional microbiome survey data can serve as the basis for understanding community structure and to generate useful hypotheses for further investigation[22]. The underlying rationale of the inference is that the observed community structure is driven by the ecological interactions between species, and therefore the non-random pattern of species distribution can be used to infer these interactions (**Figure 2.1A**). Such patterns include simple associations such as co-occurrence or co-exclusion and correlation, as well as more complex associations such as limited cycles in predator-prey systems[29].

2.2.1.1 Using co-occurrence or co-exclusion patterns

The simplest and yet interesting pattern that serves to inform about species interactions is the co-occurrence or co-exclusion of two species, providing evidence that there is strong dependency or competition between them. The detection of such patterns can be formulated into a statistical test of whether the species pair co-occurs or co-excludes each other more than random using the *Fisher's exact test*, which compares the co-occurrence pattern with the hypergeometric distribution to assign statistical significance[30].

In addition to the Fisher's exact test, it is possible to quantify the similarity between the distributions of two species across sites, with ecological distance (or similarity) scores. These scores, originally designed for comparing the overall species composition of two sites (e.g. the *Jaccard distance*[31] or *Bray-Curtis dissimilarity*[14]), have since been applied to compare the composition of two species across sites (using species abundances normalized across sites). Such distance scores obtain maximum (e.g. 1) when species are mutually exclusive and minimum (e.g. 0) when species have identical normalized abundances across all sites and can thus be used to assess the level

of co-occurrence or co-exclusion. The statistical significance of such scores can be assigned non-parametrically by comparing with an empirical null distribution generated by permuting the abundance matrix across sites for each species and re-calculating scores.

2.2.1.2 Using correlated abundances

Computing correlation between the abundance profiles of two species is another widely used approach to identify potential competitive or cooperative interactions in a microbial community[32–37]. Commonly used correlation coefficients for this include the Pearson and Spearman coefficients. The Pearson correlation coefficient between two variables is defined as the covariance of the two variables divided by their standard deviations and it captures linear dependencies. The Spearman correlation coefficient between two variables is the Pearson correlation coefficient between their rank orders, and thus detects monotonic relationships. However, the nature of microbiome survey data gives rise to several challenges in interaction inference using correlation coefficients, such as the compositionality effect, presence of indirect dependencies and data sparseness, and in the following sections we discuss these and the available approaches to account for them (See **Figure 2.2** for a graphical overview of the three challenges; See **Table 2-1** for a summary of algorithms).

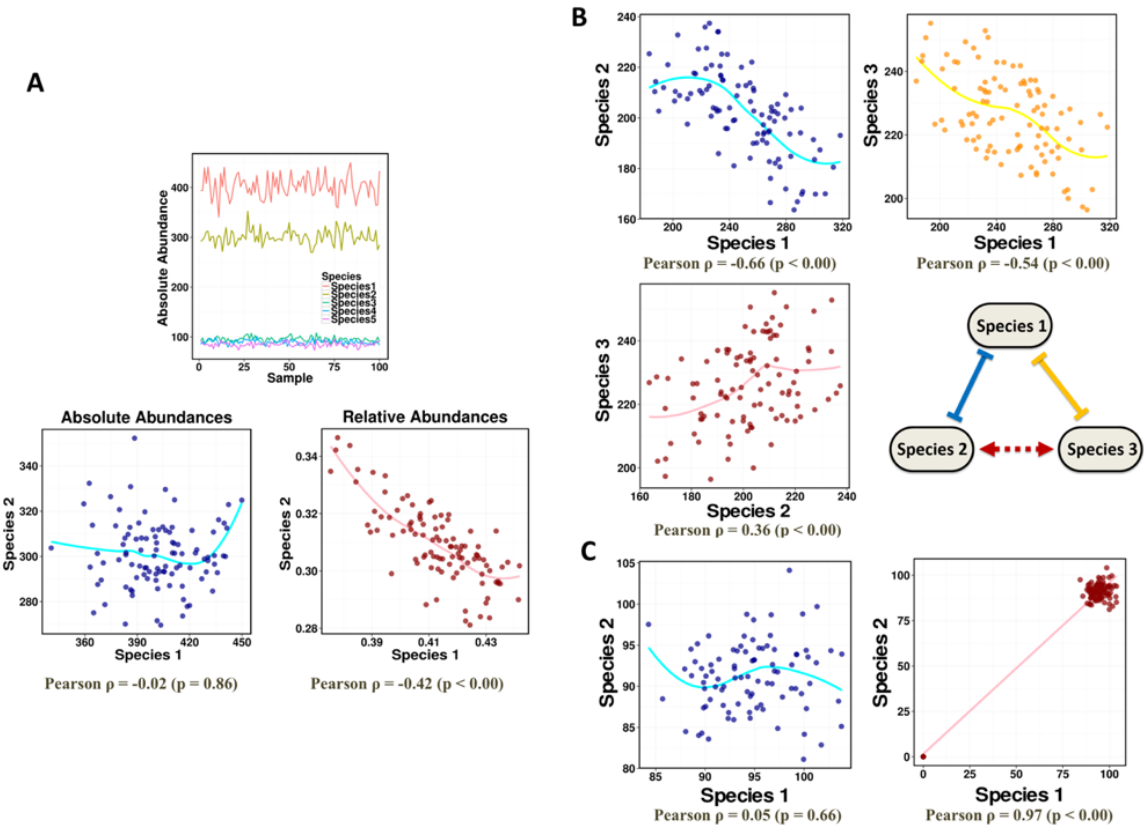

Figure 2.2 Graphical illustration of the challenges in using correlations from microbiome survey data to infer microbial interactions.

(A) Compositionality Effect: In a community with five species (top), where Species 1 and Species 2 have uncorrelated absolute abundances (bottom left), their abundances appear correlated after being normalized into relative abundances (bottom right). (B) Indirect Correlations (bottom right): The abundances of Species 2 and Species 3 are positively correlated (bottom left) not because the two species interact with each other, but because they both interact with Species 1 and are negatively correlated with it (top left and top right). (C) The abundances of Species 1 and Species 2 are not correlated (left). However, if there are sites where neither of the species is present, the two species can have an observed positive correlation (right).

Accounting for the compositionality effect

Intuitively, if the abundances of all species are constrained by a constant sum (e.g. one), an increase in the relative abundance of one species will cause a decrease in the abundance of all others. Therefore, even though the absolute abundances (i.e. the number of cells per unit volume, or the cell density) are independent, there is a tendency to get negative correlations using relative abundances and thus falsely predict interactions (**Figure 2.2A**). Such a negative bias, also known as the compositionality effect[14], is especially severe when correlations are calculated from cross-sectional microbiome survey data, because abundances of species are usually also uneven[11,38]. Therefore, the compositionality effect has to be corrected for in order to successfully use correlations to infer interactions.

CCREPE[14] is an algorithm that accounts for the compositionality effect by testing if a bootstrapping based distribution of correlation coefficients is sufficiently different from a null distribution generated using uncorrelated species profiles that are normalized to introduce a compositionality effect. Specifically, the permuted null distribution in CCREPE is generated by repeatedly shuffling the abundances of one species of interest, re-normalizing the abundance matrix into relative abundances, and finally re-computing correlations. The bootstrap distribution is generated by sampling the columns of the abundance matrix and re-computing correlation coefficients. By construction, CCREPE is designed to be conservative when the signal-to-noise ratio is low as this induces the bootstrap distribution to be wider and closer to zero.

Table 2-1 A comparison of correlation based methods

Methods	Similarity metric	Assumptions	Corrects for			Additional features	Availability	Implementation
			Compositionality effect	Indirect correlations	Data sparsity			
CCREPE	Pearson/Spearman correlation (or any similarity measure)	None	Yes	No	No		http://huttenhower.sph.harvard.edu/ccrepe	R
CCLasso	Pearson correlation between log absolute abundances	Edge density is no greater than 1/2-1/(p-1)	Yes	No	No		https://github.com/huayingfang/CCLasso	
SparCC		Average correlation between a species and others is zero	Yes	No	No		https://bitbucket.org/yonatanf/sparcc	Python
REBACCA		Each species interacts with less than p/4 other species	Yes	No	No		http://faculty.wcas.northwestern.edu/~hji403/REBACCA.htm	R
SPIEC-EASI	Partial correlation between log transformed absolute abundances	p is large. Number of interactions scales linearly with p.	Yes	Yes	No		https://github.com/zdk123/SpiecEasi	
MInt	Partial correlation between sequencing counts	Data follows a Poisson-multivariate normal hierarchical model	No	Yes	No	Corrects for known confounding variables	https://cran.r-project.org/web/packages/MInt/vignettes/MInt.html	

*The number of species in the dataset is denoted by p.

Techniques in compositional data analysis proposed by Aitchison[39,40] provide the mathematical basis for another group of algorithms based on correlation calculations. Such techniques are motivated by the observation that the ratio between the abundance of two species within a sample does not change, regardless of normalization. Therefore, we have the following formula:

$$t_{ij} = var\left[\log\left(\frac{X_i}{X_j}\right)\right] = var\left[\log\left(\frac{W_i}{W_j}\right)\right] = \omega_{ii} - \omega_{jj} - 2 \cdot \omega_{ij},$$

where X_i and W_i are random variables representing the relative and the absolute abundance of species i, respectively. The matrix $\boldsymbol{T} = (t_{ij})$ is called the *variance matrix* and can be computed from the data, while the matrix $\boldsymbol{\Omega} = (\omega_{ij})$, also known as the *basis covariance matrix*, is the covariance matrix of log transformed absolute abundances that we wish to compute. However, the linear system defined by the above formula has more variables than equations and thus has infinitely many solutions in general. In microbial interaction inference, it may typically be reasonable to assume that the basis covariance matrix is sparse, i.e. each species does not interact with a large number of other species. Using the sparseness assumption, the algorithm SparCC estimates the basis covariance matrix with an iterative approximation and refinement scheme[38]. Another algorithm REBACCA, adopting a similar assumption, solves the basis covariance matrix with sparse regression[41] (i.e. LASSO).

Besides the variance matrix based approach, Aitchison also proposed the *centered log-ratio (clr) transformation* as an alternative approach. Specifically, this involves computing the logarithm of the ratio between the relative abundance and the geometric mean of all relative abundances within site j:

$$\mathrm{clr}\left(\mathbf{x}^{(j)}\right) = \left[\log\frac{x_1^{(j)}}{\mathrm{g}(\mathbf{x}^{(j)})}, \log\frac{x_2^{(j)}}{\mathrm{g}(\mathbf{x}^{(j)})}, \ldots, \log\frac{x_p^{(j)}}{\mathrm{g}(\mathbf{x}^{(j)})}\right]^T = \mathbf{G} \cdot \log\left(\mathbf{x}^{(j)}\right),$$

where $\mathbf{x}^{(j)} = \left[x_1^{(j)}, \ldots, x_p^{(j)}\right]^T$ is a column vector representing the relative abundances of species at site j, $\mathrm{g}(\mathbf{x}^{(j)}) = (\prod_{i=1}^{p} x_i^{(j)})^{\frac{1}{p}}$, $\mathbf{G} = \mathbf{I}_\mathrm{p} - \frac{1}{\mathrm{p}}\mathbf{J}_\mathrm{p}$, $\mathbf{I}_\mathrm{p}$ is the p-dimentional identity matrix and $\mathbf{J}_\mathrm{p}$ is a p-dimentional matrix with all 1s. This function transforms the data from a constrained space with p dimensions to a $(p-1)$-dimensional Euclidian space, where analysis is free from the compositionality effect. The covariance matrix of the clr transformed variables ($\mathbf{\Gamma}$) can be mapped to many basis covariance matrices, but there is at most one with sparseness above a certain threshold[42]. CCLasso is a clr-transformation based algorithm that finds such a sparse basis covariance matrix using LASSO[42].

It is worth noting that the methods based on Aitchison's techniques all calculate Pearson correlation coefficients between log transformed abundances. Given the unevenness of microbiome survey data, log-transformation could result in more meaningful Pearson correlation coefficients. On the other hand, CCREPE can be generalized to Spearman correlation coefficient and the other distance scores mentioned above. However, correlation coefficient values from CCREPE must be interpreted with caution as they retain their intrinsic negative bias.

Removing the effect of indirect dependencies

Two species can be positively correlated because both of them are negatively correlated with a third species (**Figure 2.2B**). In order to infer direct interactions, we need to measure the dependency between two species conditioned on all other species, which is formally defined as computing a *partial correlation*. If Pearson correlation is used, partial correlations can be computed directly from the inverse of the covariance matrix (i.e. the *precision matrix*; note that the estimating the precision matrix direct inversion of the sample covariance matrix is in fact a challenging problem especially for large systems[44].):

$$\rho_{X_i X_j \cdot \boldsymbol{x} \setminus \{X_i, X_j\}} = -\frac{p_{ij}}{\sqrt{p_{ii} p_{jj}}},$$

where $\boldsymbol{x} = [X_1, \ldots, X_p]^T$ is a random vector of species abundances, $\mathbf{P} = (p_{ij})$ is the precision matrix and $\rho_{X_i X_j \cdot \boldsymbol{x} \setminus \{X_i, X_j\}}$ is the partial correlation between X_i and X_j conditioned on other variables in $\boldsymbol{x}$.

There are several existing algorithms that estimate partial correlations[45] or the precision matrix[46], and have been adapted to infer microbial interactions. MInt uses *graphical LASSO* to estimate the precision matrix under the sparseness assumption[47]. It also provides a method to remove the known confounding factors, such as measured biological covariates, experimental replicate and so on. SPIEC-EASI also employs graphical LASSO, and in addition accounts for the compositionality effect using the clr transformation[48]. However, unlike CCLasso, SPIEC-EASI directly uses the clr covariance matrix to replace the basis covariance matrix (i.e. assuming $\mathbf{G} = \mathbf{I_p}$), which is a reasonable approximation when the number of species is large. In addition, SPIEC-EASI provides an alternative algorithm to estimate partial correlations, which iteratively builds a sparse linear model for each species using the rest of the species as covariates.

Recently, Network Deconvolution[49], a *post hoc* approach to remove indirect associations, was applied to study oral[50] and plankton[51] microbial interactions. With an inferred interaction network, the network deconvolution framework removes edges attributed to the transitive effect. Network deconvolution is therefore a general framework and can be applied to a variety of measures including Spearman correlation and mutual information.

Dealing with the sparseness of microbiome survey data

A zero in microbiome survey data could be attributed to the physical absence of a species (i.e. a structural zero) or to insufficient sequencing depth to capture it (i.e. a

sampling zero), but it is non-trivial to distinguish the two and fill gaps caused by sampling depth[52,53]. Currently, none of the existing microbial interaction inference algorithms directly addresses this issue. Methods based on log transformation have to avoid directly transforming a zero by adding pseudo count to the observed counts, and thus implicitly assuming the presence of all species[38]. For CCREPE, bootstrapping can generate zero vectors for species with many zero values across sites. Therefore, before estimating correlations with these algorithms, species that are only present in a few samples are often removed. Moreover, when computing correlation between two species, sites where the abundances for both species are zero can become outliers and result in an apparent positive correlation, i.e. the "double-zero" problem[54] (**Figure 2.2C**).

2.2.1.3 Predicting complex associations

Complex associations are generally difficult to interpret, but they can serve as evidence for non-random ecological forces. Such associations can be captured by comparing the distributions of abundances. The *Mutual Information* measures the reduction in uncertainty of one variable when the other variable is given, and is widely used in inferring gene regulation networks from expression data[55–57]. In particular, it is of interest to derive a mutual information based measure, which can capture a wide range of different associations (i.e. general) and assign similar scores to them when noise levels are comparable (i.e. equitable). Reshef *et al.* proposed the Maximal Information Coefficient (MIC), which is meant to possess these two properties[58]. The MIC has been applied to detect novel non-linear associations using microbiome survey data[59,60]. However, the equitability and statistical power of MIC have been challenged in recent studies[61,62]. In addition, there is currently no mutual information based method that accounts for the compositionality effect. Another practical challenge is to obtain a sufficiently large sample size in order to estimate the marginal and joint distributions of

variables (the number of samples should be much larger than the number of bins used to as the discretize the data[63]) that is needed to accurately compute the MIC.

2.2.1.4 Benchmarking and integrating different methods

While there is a diversity of measures for capturing different patterns of microbial interactions, a comprehensive comparison of their performance is still lacking because of the lack of a gold-standard set of interactions. Nevertheless, it is possible to construct for validation purposes, a model for a species community with specified interactions, and generate species abundance data through simulations. The *generalized Lokta-Volterra models* (gLVM) are often used to model the growth of each species in response to the dynamics of its interacting partners:

$$\frac{\mathrm{d}}{\mathrm{d}t}x_i(t) = x_i(t)\left(\mu_i + \sum_{j=1}^{p}\beta_{ij}x_j(t)\right),$$

where $x_i(t)$ is the absolute abundance (density) of species i at time t, μ_i is the maximal growth rate of species i, and β_{ij} captures the influence of species j on the growth of species i. The gLVM explicitly encodes these pairwise interactions in a matrix (β_{ij}), which can be treated as the gold standard measure of interactions for a simulated community. Berry & Widder used the gLVM to generate simulated abundances in the equilibrium state of the community, and assessed different measures for their ability to capture the specified interactions[36]. However, there were several caveats in their design and analysis. Firstly, the parameters in the system were chosen arbitrarily without reference to any real microbial community profiles. In addition, the assumption of steady state profiles is unlikely to be valid, especially for human-associated microbiome data[64,65]. Moreover, a majority of their analysis was done assuming absolute species abundances, which is often unavailable for microbiome survey data. Finally, their assessment reported specificity and sensitivity of predictions, but ignored precision – a

metric that is of great practical value when predictions are tested using experimental approaches.

Another perspective to comparing different interaction prediction measures is to motivate the creation of an ensemble predictor that could potentially achieve superior performance. Limited attempts had been made to integrate various association measures so far. Both union[66] and intersection[50] of interaction networks inferred using different measures have been applied. Furthermore, Faust *et al.* combined the *p*-values for different measures with a multiple test correction approach to get their final predictions[14], however none of these approaches have been systematically tested for their utility and value compared to individual methods for predicting microbial interactions.

2.2.2 Interaction inference with temporal microbiome survey data

Despite the current lack of temporal microbiome survey data, we foresee in the near future, a rapid accumulation of such datasets given the increasing interest in understanding the dynamics of microbial communities[67–70]. Such time-series abundance data will also likely become a valuable resource for inferring species interactions as the time component allows for better modeling of the microbial ecosystem.

Compared to their cross-sectional counterpart, temporal datasets provides a dynamic view of species interactions, and can thus be used to infer interactions from shifted time series[71–73]. LSA (Local Similarity Analysis) is an algorithm that quantifies the similarity between shifted time series through dynamic programming[73]. Another approach employs the generalized Lotka Volterra equations to fit the data to a gLVM and directly solve for the interaction matrix through sparse regression[74–76]. Further details can be found in a recent review that provides an in-depth survey of methods specific to analyzing temporal microbiome survey data[77].

2.3 Predicting interactions from genomic information

Besides the compositional profile of microbial communities, another product of genome and whole metagenome sequencing is the fragmented or complete assembly of individual microbial genomes. The availability of genome sequences allows us to predict genes and then annotate them to assign function. Among these genomic elements, the enzyme coding genes are the essential functional carriers, and can serve as a proxy for the metabolic potential of the organism. The collection of enzymes together with the reactions they catalyze, can be linked using their shared metabolites and modeled as a network. The input and output compounds for the metabolic network of an organism represent the minimal set of units for metabolic "communication" with extra-cellular space and other organisms, and therefore can be used to bridge the metabolic networks of multiple species (**Figure 2.1B**). Such a framework that employs genomic information to predict ecological interactions has been termed the *Reverse Ecology* paradigm[78,79].

2.3.1 Microbial interaction inference using metabolic network topology

The key idea behind the reverse ecology paradigm is to identify the metabolites cycling between the organism and the environment in a metabolic network. A simple method is to ignore the stoichiometry of metabolic reactions and enzyme kinetics, and only consider the topology of the metabolic network[15,80]. Such a simplified model can be represented as a directed graph, in which edges are the metabolic reactions, while nodes are the metabolites involved. A further simplification can be made by splitting complex reactions into multiple substrate-product pairs (e.g. reaction A+B→C is split into A→C and B→C). It is then possible to apply algorithms based on graph theory to identify a set of metabolites that can only be obtained from the environment. Borenstein *et al.*

termed such a set of metabolites the "seed set" and implemented a graph topology based algorithm, NeetSeed for its determination[15,80].

Hypothetically, two species can compete for limited resources if their seed sets overlap, and the level of competition against species A when species B is present can be quantified by the fraction of metabolites in A's seed set that is covered by the seed set of B[31]. An alternative approach is to remove the metabolites (shared by species B) from species A and compute the proportion of metabolites that can be synthesized from the remaining metabolites. The competition level is then quantified by one minus the proportion. The web-based tool NetCmpt implements the latter method to quantify competition level using annotated genomes of two species as input[81]. On the flip side, two species can cooperate with each other where one species produces the metabolites in the seed set of its interacting partner. The metabolic complementarity of species B with species A can therefore be measured by the fraction of metabolites in A's seed set that is also in the non-seed set of B[31]. This measure has been implemented as a web-based tool called NetCooperate[82] to assist in identifying species pairs that could potentially have a cooperative relationship.

2.3.2 Predicting interactions with community constraint based models

Another approach to study metabolic interactions between species employs an extension of single-organism constraint based modelling (CBM). A CBM for a single species consists of a stoichiometry matrix representing all the metabolic reactions, a set of constraints that are imposed to bound flux values, and an objective function (usually biomass production) that should be optimized. The flux distribution in an equilibrium state which satisfies the constraints and optimizes the objective function, can be solved with linear programming (see Orth *et al.* for an introduction to flux balance analysis[83]). To extend this technique to model microbial communities, each species is treated as an

isolated compartment, and the set of transport reactions is represented as an additional compartment. All compartments are then combined into a single stoichiometry matrix. The objective function to be optimized for the community can be the total biomass production function[26,84,85], or a multilevel function which optimizes both individual and community growth[86,87]. In addition, enzyme kinetics[88] and diffusion models[89] have also been incorporated into community CBMs to capture temporal and cross-sectional dynamics in metabolism.

With a community CBM model, species interactions can be predicted from growth simulations, based on the difference in biomass production of a species growing in isolation compared to growing with its interacting partners[26]. Alternatively, metabolic cooperation can also be quantified as the difference in the minimum nutrients required to support community growth, with and without metabolic cross-feeding between community members[90].

2.4 Mining interactions from scientific literature

With many studies focused on generating high-throughput sequencing based profiles of microbial communities and inferring interactions from them, it is easy to overlook the existing body of literature describing microbial interactions identified through direct experimental approaches. Over the years, hundreds of studies have been conducted that have experimentally identified a large number of microbial interactions[91–96]. However, this information is currently spread across the literature and it is thus typically impractical to systematically use this information. The collation of experimentally validated interactions in the scientific literature into a reference, gold-standard database would create a valuable resource. However, with more than two million papers on bacteria alone, manual collation of this information is clearly infeasible. A potential

solution is the implementation of automated systems for information extraction from scientific literature (also referred to as *Text Mining*, **Figure 2.1C**).

Text mining techniques have been widely used for biomedical applications including the processing of patient medical records[97], classification of genetic variations in drug response[98], as well as in the identification of Protein-Protein Interactions (PPIs)[99,100]. However, for microbial interactions, a 2010 paper by Freilich *et al.* is currently the only reported study based on text mining techniques[16]. Freilich *et al.* used abstracts from the PubMed database to identify and quantify bacterial species co-occurrences. Significant interactions were then identified using a hypergeometric test for over-representation of species pair in the scientific literature. The putative interactions were then organized into a network and a clustering of species on the network was used to identify groups of organisms that serve as representatives of naturally occurring communities.

Two other categories of text mining techniques have been used in biomedical fields: Rule-based methods and Machine Learning (ML). Rule-based methods apply a set of precompiled rules to extract information from literature data. They generally improve on co-occurrence methods, boosting precision at the cost of recall[101,102] and can even outperform state of the art ML methods[103]. ML methods however, employ statistical algorithms capable of learning from annotated training data to accurately classify new, unseen datasets. *Bayesian networks*, *k-Nearest Neighbors*, and *Kernel methods* are ML techniques that have seen a degree of success in biomedical text mining tasks[101,103,104].

However, the application of text mining to microbial interactions comes with a number of specific challenges. Unlike other fields that possess an abundance of well-annotated corpora[105–108] for training and evaluating text mining systems, no similar resource exists for microbial interactions. Furthermore, while databases of various

bacterial species exist, these names are subject to change in the case of orthographic or typographic errors[109]. However, these changes are not retroactively executed and result in heterogeneity of nomenclature in the existing literature.

2.5 Concluding remarks

The three data types and the computational approaches discussed here are fundamentally different and yet complement each other in predicting and compiling microbial interactions. Species abundance patterns are assumed to be the outcome of microbial interactions, and in reverse can be used to infer the corresponding cause. Nevertheless, while there are various scores and approaches that can be used to quantitatively predict interactions, a systematic benchmarking study and an integration scheme is still lacking. In contrast, metabolic reconstruction approaches start from the mechanisms and predict interacting outcomes by modelling community metabolism. Such methods are usually less scalable and rely on accurately annotated genomes. The largely uncharacterized gene-metabolite relation in microbial communities also hinders the application of such methods. In addition, they are specific to metabolic interactions and therefore will miss interactions due to other mechanisms, such as bacteriocin production[95,96] and signaling processes[110–112]. A curated interaction database mined from the scientific literature could serve as a catalogue of gold-standard interactions. However, the application of text mining techniques to microorganisms is in its infancy and there is currently no such collection of validated interactions.

In principle, an integration of multiple approaches could improve the accuracy of microbial interaction prediction and provide a deeper understanding of the mechanisms involved. For example, predictions generated from metabolic reconstruction methods or literature mining can be compared with co-occurrence pattern to see if an “in-principle” interaction is also reflected in a specific kind of

community[16,31]. Conversely, co-occurrence pattern can be used as an initial filtering step before constructing all pair-wise metabolic models[90] or can be directly incorporated into a metabolic model[84,85]. It should be noted here that the nature of a microbial interaction between two species can be dynamic and depend on environmental context (nutritional sources[113] or other microbes[114]). Addressing these issues needs careful consideration of the biological context in which information about microbial interactions is applied and will likely require the integration of diverse data types and methods for studying and modeling microbial interactions.

2.6 Research objectives

Despite not as commonly used as correlation based approaches, methods for inferring interactions based on ecological models (i.e. gLVMs) from temporal microbiome data have a key advantage to be able to infer interactions with directionality and thus able to distinguish different types of ecological interactions (e.g. mutualism vs. commensalism). Extending the models to incorporate external factors including nutrient resources (e.g. host diet change for gut microbiome) and perturbations (e.g. antibiotics usage) make it possible to infer causal relationship between the microbiome and the environment[74–76].

However, learning ecological models requires measuring the density of each microbe (i.e. the absolute abundances), which is not available from high-throughput sequencing data. An assumption that the total number of cells per unit volume (i.e. total biomass) is the same across all samples have been made such that relative abundances are comparable for inferring interactions[115]. Nevertheless, the effect of ignoring the variation in total biomass on the accuracy of inference is not well understood. Another way to circumvent the lack of absolute quantification is to scale relative abundances with experimentally measured total biomass[74–76]. However, the total biomass measure is

often unavailable for existing studies, limiting the ability to learn accurate models leveraging large collections of public microbiome survey data.

In this thesis, we aim to investigate the performance of correlation based as well as ecological model based approaches for learning interactions. Importantly, we explore a novel angel – eliminating the need for biomass by learning ecological models from relative abundance data.

3 BEEM: An Expectation-Maximization-Like Algorithm Enables Accurate Ecological Modeling Using Longitudinal Microbial Profiling data

3.1 Background

A growing body of literature points to the important roles that different microbial communities play in diverse natural environments[116,117] and the human body[118]. This has particularly been aided by advances in next-generation sequencing technology, allowing for rapid, cost-effective taxonomic and functional profiling, combined with computational analysis that has helped associate the state of the microbiome with various environmental conditions[116,119] and human diseases[120–123]. Microbiomes are also constantly evolving and there is now a growing appreciation of the fact that complex interactions between community members[124,125] shape community dynamics[126,127] as well as overall function[128,129]. A systems view of the microbiome is thus essential for understanding and rationally manipulating it[74].

Because of its importance, there have been many approaches proposed to study microbial interactions and dynamics. Experimental approaches have ranged from simple two species co-culture experiments[130,131], all the way to complex, multi-stage reactor models[132]. Analytical approaches[133] frequently use simple correlations between the

abundances of various taxa in cross-sectional datasets to infer microbial interactions[48,134,135]. There are several challenges that need to be addressed in such analysis including the compositionality of sequencing data[48,134–136], low sensitivity and specificity of such methods[75,137], and the inability to infer directionality of interactions or dynamics of the system[133].

The most commonly used approach for modeling microbial ecology is based on classical predator-prey systems, also referred to as generalized Lotka-Volterra models (gLVMs). gLVMs are based on ordinary differential equations (ODE) that model the logistic growth of species, naturally capture predator-prey, amensalistic and competitive interactions, and have been applied to study dynamics of microbial ecosystems ranging from simple communities on cheese[138,139] to the human microbiome[74–76,140–142]. More importantly, from a practical perspective, gLVMs have been used for a range of applications including identifying potential probiotics against pathogens[74,76,140], forecasting changes in microbial density, characterizing important community members (e.g. keystone species[75]) and to analyze community stability[140,142].

Despite this, a key limitation of gLVMs that restricts applicability and wider use is the requirement for microbial abundance data on an absolute scale. Microbiome analysis using high-throughput sequencing naturally provides relative abundance estimates with what is often referred to as "compositionality bias"[134–136], and cannot be directly used to estimate gLVM parameters[141]. Scaling relative abundances to an absolute scale typically requires additional experimental data that is either not readily available (as is true for the vast proportion of publicly available datasets), is technically challenging to directly quantitate for different sample matrices and complex communities (e.g. using flow cytometry[143,144]), or can suffer from significant technical[145–147] and biological noise[148] (e.g. using 16S rRNA qPCR[74,76,140]).

We show that, surprisingly, scaling factors can be directly inferred from microbiome sequencing data, through an algorithm that also simultaneously estimates gLVM parameters (BEEM). This is achieved based on an expectation-maximization-like approach[149] that alternates between learning scaling factors and gLVM parameters, and thus obviates the need for experimental scaling factors which otherwise limit the use of many existing datasets. Based on synthetic data where biomass is precisely known, we show that BEEM estimated gLVM parameters are as accurate as those estimated with true biomass values, and significantly more accurate than what could be expected with commonly used (16S rRNA based) experimentally determined biomass estimates. Using data from a freshwater microbial community with flow cytometry based gold-standard cell counts, we show that biomass estimated using BEEM has good concordance with the gold-standard and improves significantly over existing techniques to normalize data. Leveraging BEEM's unique ability to learn gLVMs from relative abundance data, we analyzed publicly available datasets that represent the longest human gut microbiome time-series data available to-date[150–152]. This analysis revealed, for the first time, the personalized dynamics of gut microbial biomass in different individuals, with communities driven by distinct interaction networks and hub species. Our analysis suggests an emerging model for gut microbial dynamics where relatively low abundance species may play key roles in maintaining gut homeostasis.

3.2 Materials and Methods

3.2.1 The generalized Lotka-Volterra model (gLVM)

The gLV equations model the growth rate ($\frac{dx_i(t)}{dt}$) of each microbial species i as a function of absolute densities ($x_i(t)$) of all the p species in a community:

$$\frac{dx_i(t)}{dt} = \mu_i x_i(t) + \sum_{j=1}^{p} \beta_{ij} x_i(t) x_j(t) \qquad (1).$$

In the above model, the intrinsic growth rate parameter (μ_i) and self-interaction parameters (β_{ii}) define the logistic growth behavior of species i. In addition, the model also captures the impact of the absolute density of species j on the growth rate of species i through additional parameters (β_{ij}, $i \neq j$), assuming a linear and additive effects model. As high-throughput sequencing based approaches to analyze microbiomes only provide relative abundance estimates, scaling factors related to the total biomass for each sample are then needed to accurately fit gLVMs in practice.

3.2.2 The core algorithm of BEEM

In order to address the challenges of noisy experimental biomass data and, in general, to make gLVM modeling more widely applicable where biomass estimates are not available, we explored the idea of learning gLVM parameters directly from relative abundance data. To achieve this, we first note that model equation 1 can be expressed in terms of relative growth rates by dividing both sides of the equation by $x_i(t)$:

$$\frac{dx_i(t)}{dt}/x_i(t) = \frac{d\ln x_i(t)}{dt} = \mu_i + \sum_{j=1}^{p} \beta_{ij} x_j(t).$$

By explicitly introducing relative abundances ($\tilde{x}_i(t)$) and total biomass ($m(t)$, where $x_i(t) = m(t)\tilde{x}_i(t)$), we get:

$$\frac{d\ln m(t) + \ln \tilde{x}_i(t)}{dt} = \mu_i + m(t)\sum_{j=1}^{p} \beta_{ij}\tilde{x}_j(t)\ . \qquad (2)$$

To eliminate the biomass related term in the left-hand-side of the equation, we subtract the corresponding equation for a reference species r from both sides of the system, resulting in additive log ratio (ALR) transformed[39] relative abundances ($y_i(t) = \ln(\tilde{x}_i(t)/\tilde{x}_r(t))$) on the left-hand-side and a re-parameterized right-hand-side:

$$\frac{dy_i(t)}{dt} = a_i + m(t)\sum_{j=1}^{p} b_{ij}\tilde{x}_j(t)\,, i \neq r\,,$$

where $y_i(t) = \ln(\tilde{x}_i(t)/\tilde{x}_r(t))$ and the equations are re-parameterized by a_i and b_{ij}, which are related to the original parameters ($a_i = \mu_i - \mu_r$ and $b_{ij} = \beta_{ij} - \beta_{rj}$). This new system has the advantage that all unknowns are on the right-hand-side of the equation and the gradient term on the left-hand-side. The choice of the reference species is important as it could introduce noise to ALR transformed abundances for other species. Therefore, we select the species varying the least (with the lowest CV) in relative abundances as the default because such species is more likely to contain less experimental noise in its relative abundances.

An estimate for $dy_i(t)/dt$, denoted as Y_{it}, can be calculated as the derivative of a piece-wise polynomial spline fitted to the ALR transformed relative abundances ($y_i(t)$, see Section 3.2.3 for details). BEEM then estimates the model parameters $\boldsymbol{a}, \boldsymbol{b}$ and the biomass $\boldsymbol{m}$ using an EM-like algorithm with the following sum of squared error objective function:

$$\Theta(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{m}) = \sum_{i,t} \left(Y_{it} - \left(a_i + m_t \sum_{j=1}^{p} b_{ij} \tilde{X}_{jt} \right) \right)^2,$$

where $\tilde{X}_{it} = \tilde{x}_i(t)$ and $m_t = m(t)$ are the variables written in their matrix representations.

The EM-like algorithm in BEEM works by iterating two steps, an E-step and an M-step, to convergence as detailed below:

Model parameter estimation with Bayesian lasso (E-step): In iteration T, with estimated biomass from the previous iteration $\hat{m}_t^{(T-1)}$, BEEM estimates $\hat{a}_i^{(T)}$ and $\hat{b}_{ij}^{(T)}$ for each i ($i \neq r$) based on the following regression problem (also known as gradient matching):

$$Y_{it} \sim a_i^{(T)} + \hat{m}_t^{(T-1)} \sum_{j=1}^{p} b_{ij}^{(T)} \tilde{X}_{jt} .$$

Solving the above system is often limited by the amount of data available in practice. For microbial communities, it is usually assumed that the interaction vector (β_{ij}) is sparse (i.e. a species is only directly affected by a small number of other species). Consequently, the transformed matrix b_{ij} is also sparse and BEEM estimates it using a sparse regression technique based on a Bayesian approach (Bayesian lasso - BLASSO[140]; R package "monomvn" version 1.9-7; default parameters)[153].

Biomass estimation with linear regression (M-step): With $\hat{a}_i^{(T)}$ and $\hat{b}_{ij}^{(T)}$ from the E-step, the biomass $\hat{m}_t^{(T)}$ for each T can be computed as the coefficient of the following linear regression:

$U_{ti}^{(T)} \sim m_t^{(T)} V_{ti}^{(T)}, i \neq r$,

where $U_{ti}^{(T)} = Y_{it} - \hat{a}_i^{(T)}$ and $V_{ti}^{(T)} = \sum_{j=1}^{p} \hat{b}_{ij}^{(T)} \tilde{X}_{jt}$.

Initialization: For the initialization step in its EM-like algorithm, BEEM assumes that scaling factors inferred from a commonly used normalization approach for metagenomic data (Cumulative Sum Scaling - CSS[154]) provides a reasonable starting point for the algorithm to then learn better scaling factors. Note that, as expected, scaling factors from CSS normalization and BEEM cannot recapitulate the absolute scale corresponding to experimental measurements (e.g. by qPCR or flow cytometry), and so their estimates were scaled to the same median value across the time series as experimental measurements for subsequent comparisons. In practice, the true scale of all samples can be recovered by measuring the biomass for a single sample accurately.

Termination and parameter estimation: The E- and M-step in BEEM are run until convergence or a user specified maximal number of iterations. The search was

assumed to have reached convergence (to a local optimum) when the mean squared error (MSE) for the M-step starts to increase by more than 10% compared to the minimal MSE observed. In practice, on the datasets analyzed in this study, convergence takes a few hours using 4 CPUs. Estimates for $\hat{a}_i$, $\hat{b}_{ij}$ and $\hat{m}_t$ were calculated as the median of the values from all iterations whose MSE was within 10% of the minimal MSE.

3.2.3 Robust parameter estimation with BEEM

In our experiments with synthetic and real data, we noted that gLVM modelling can be sensitive to noise and outliers in the data, and this in turn could affect estimation of scaling factors with BEEM. To address this, we refined the core algorithm in BEEM with additional pre-processing steps that further enable robust parameter estimation.

Outliers in relative abundance data: We observed in our numerical analysis that outliers in the abundance data could notably affect the spline fitting procedure and lead to spurious gradient estimates. To obtain more robust spline fitting, an over-smoothed spline was first fitted to $y_i(t)$ (function "smooth.Pspline" from R package "pspline"[155] with maximal degree of five and a large smoothing parameter "spar=1e10") to calculate the absolute error in fitted values ($e_{it} = |y_i(t) - y_i(t)^{\text{smoothed}}|$), and points with absolute error larger than expected ($(e_{it} - \underset{j}{\text{median}}(e_{ij}))/\underset{j}{\text{MAD}}(e_{ij}) > \tau$, $\tau = 5$ by default) were then filtered out. The final smoothing spline was fitted (degree of five and smoothing parameter selected using cross validation) to the remaining data to calculate the estimated gradients Y_{it}. In addition, outliers in biomass estimated from the previous iteration ($\hat{m}_t^{(T-1)}$) were identified in the same way and replaced with interpolated values from the spline.

Outliers in estimated gradients: In practice, gradient matching based methods (including the various algorithms implemented in MDSINE) were found to be sensitive

to outliers in the estimated gradients (i.e. Y_{it}). To identify outliers in a time series (Y_{it}, for all t) a local regression (LOESS) smoother was fitted to de-trend Y_{it}, and the outliers were filtered out as described above.

Estimating constrained biomass values: For each time point, biomass was estimated as the slope of a linear regression ($U_{tk}^{(T)}$ against $V_{tk}^{(T)}$) where outliers in both $U_{tk}^{(T)}$ and $V_{tk}^{(T)}$ were identified and removed following a standard boxplot approach i.e. as deviations from the median by more than 1.5× inter-quartile range. In addition, the biomass was constrained to be positive by removing points where $U_{ti}^{(T)}$ and $V_{ti}^{(T)}$ had different signs.

3.2.4 Recovering gLVM parameters

Based on the previously stated assumption that the interaction matrix $\boldsymbol{\beta}$ is sparse, most entries in each column are expected to be zero and thus the median value for the j^{th} column in $\boldsymbol{b}$ would be expected to be $-\beta_{rj}$, allowing us to infer back all the other rows of $\boldsymbol{\beta}$ ($\beta_{ij} = b_{ij} + \beta_{rj}$). BEEM then assigns a Z-score like confidence value (s_{ij}) to each entry of $\boldsymbol{\beta}$, by dividing the estimated interaction strength by the column standard deviation ($s_{ij} = |\hat{\beta}_{ij}/\sigma_j|$). The growth rate vector $\boldsymbol{\mu}$ is not expected to be sparse but can be recovered by directly solving the original gLVM system (equation 2), using the already derived estimates for scaling factors and $\boldsymbol{\beta}$. For robustness, BEEM estimates the growth rate for each species as the median of positive estimates across all time points.

3.2.5 Datasets and evaluation metrics

Simulated datasets: MDSINE's Bayesian variable selection (BVS) algorithm (with spline smoothing option and minor bug fixes: https://bitbucket.org/chenhao_li/mdsine) was used to estimate parameters from the *C. difficile* infection dataset provided with the package[140]. Simulated datasets were then generated based on these estimated parameters

following the procedure described in Bucci *et al*[140] (excluding perturbations). Specifically, the learned growth rates, interaction strengths (inter- and intra-species) were used to parameterize the distributions (half normal for growth rates and intra-species interactions, zero-mean normal for inter-species interactions), from which ground truth gLVM parameters were sampled. The interaction structure was randomly generated. The gLVM was integrated numerically to generate time series microbiome samples based on initial abundances sampled from a normal distribution (parameterized by the dataset provided with MDSINE ("data_cdiff")). Noisy abundances were obtained by sampling from Poisson distributions[156] with means based on scaled abundances at each time point (sum = 5×10^4). Simulated qPCR and flow cytometry based values for total biomass were generated from log normal distributions with coefficients of variation (CV) that matched those seen in real datasets (qPCR=51%[140], flow cytometry=5%[143,144]).

Dataset from Props et al: This dataset was generated from a freshwater microbiome of a cooling system. The microbiome was profiled using 16S rRNA gene sequencing and the total biomass was measured with flow-cytometry. The original OTU table was obtained from the authors[144]. Samples for the "operation" stage, where the environment had roughly constant temperature were selected for BEEM analysis. OTUs with low mean relative abundances (<0.1%) were excluded, resulting in 26 OTUs across 57 time points from two replicates.

Dataset from Gibbons et al: This dataset included four long and dense (almost daily) gut microbiome time series collected by David *et al*[150] (two individuals with 180 and 311 samples) and Caporaso *et al*[151] (two individuals with 131 and 332 samples). The original OTU tables[152] were filtered to keep only top OTUs based on prevalence (>10 reads in most of the samples). In total, 26 and 22 OTUs were left for samples from David *et al* and Caporaso *et al*, respectively. In order to assess the robustness of the

inferred network, BEEM was run with 30 different seeds and edges with confidence score $s_{ij} \leq 1$ in more than 50% of the networks were kept. The final biomass was obtained by taking the geometric mean across all 30 runs.

Metrics for evaluation: The following metrics were used for evaluating inference algorithms:

1. Median relative error (MRE) for estimates $\widehat{\boldsymbol{\theta}}$ when the true values are $\boldsymbol{\theta}$: $\underset{\theta_i \neq 0}{\text{median}} \left|\frac{\widehat{\theta}_i - \theta_i}{\theta_i}\right|$.
2. Area under receiver operating characteristic curve (AUC-ROC) for the inferred microbial interactions: Absolute values of parameters was used to rank predicted edges for BLASSO and LIMITS (implemented in R package seqtime_0.1.1[157], default parameters), while confidence scores were used for Bayesian Variable Selection (BVS) in MDSINE and for BEEM.

3.3 Results

3.3.1 Experimentally obtained biomass estimates can lead to inaccurate gLVMs

High-throughput sequencing based approaches to analyze microbiomes only provide relative abundance estimates, and scaling factors related to the total biomass for each sample are then needed to accurately fit gLVMs in practice. The predominantly used approach to estimate total biomass is to quantify total copy number of the 16S rRNA gene using quantitative PCR (qPCR)[74,76,140]. However, 16S qPCR estimates have been reported to have high technical noise, with a coefficient of variation (CV) ranging from 11% to 75%[145–147]. To reconfirm this, we reanalyzed 16S qPCR data from a recent microbiome modeling study on *C. difficile* infections[140] and observed low concordance across technical replicates (Spearman ρ<0.22; **Figure 3.1A** and **Figure 8.1A**), as well

as high coefficient of variation (mean CV=51%). Another critical source of error with 16S qPCR based biomass estimates is biological, and arises due to the fact that bacteria can have widely varying number of copies of the 16S rRNA gene, even within the same ecological niche. For example, the 16S gene copy number of the four major gut bacterial phyla cover a broad spectrum (**Figure 3.1B**), ranging from a single copy to 15 copies[148]. Correspondingly, 16S qPCR estimated biomass of a community dominated by *Firmicutes* can be twice as much as that of a community dominated by *Bacteriodetes*, even if both communities have exactly the same cell density. Such large relative errors (~100%) can then have a significant impact on the accuracy of gLVMs estimated from the data, as we show below.

To test the impact of biomass estimation errors on model inference, we generated synthetic datasets (10 species community) based on parameters inferred from real datasets, similar to the approach in Bucci *et al*[140]. This framework allows us to carefully evaluate the impact of different levels of noise in a setting where model parameters are known. We noted that, given error-free biomass data, a state-of-the-art method (MDSINE[140]) was able to infer model parameters with median relative error <20% and with ~90% median AUC-ROC (area under the sensitive-specificity tradeoff curve) for interaction terms ($\boldsymbol{\beta}$; **Figure 3.1C**, True). However, as expected[141], directly using relative abundance estimates without scaling them increased median relative error for parameter estimates to >60% (**Figure 3.1C**, RA), with AUC-ROC for interaction terms being comparable to randomly generated parameters from the prior model for the simulation (**Figure 3.1C**, Random). Similar performance was obtained using another model fitting algorithm that works with relative abundance data and assumes small fluctuations in biomass values (LIMITS[75,157]; **Figure 8.1B**). Using biomass estimates with error profile similar to real qPCR data (CV=51%; without systematic errors due to varying copy number of the 16S rRNA gene), surprisingly, did not improve

performance substantially when one technical replicate was provided (**Figure 3.1C**, qPCR_rep1), and even with three technical replicates, growth rate parameter estimates (median relative error >70%) were comparable to random (**Figure 3.1C**, qPCR_rep3). These results highlight that experimental errors in biomass estimates can significantly impact gLVM parameter estimation even in a relatively well-controlled setting where model assumptions are strictly applied.

3.3.2 Jointly estimation of biomass and model parameters with BEEM

On the synthetic datasets used in previous section, we noted that despite not having any biomass data to work with, BEEM was a significant improvement over naïve analysis based on relative abundance data, as well as results based on scaled relative abundances with noisy biomass data (~3× reduction in relative error; **Figure 3.1C**, BEEM). In fact, BEEM estimated parameters were nearly as accurate as those obtained using noise-free biomass data (relative error for growth rate and interaction terms), except for a slight decrease in AUC-ROC for interaction terms (primarily due to rounding errors that provide non-zero estimates for zero terms). In comparison, other competing approaches (RA, qPCR, CSS) provided AUC-ROC performance similar to what is expected at random. Normalization approaches such as CSS[154] **(Figure 3.1C** and **Figure 8.1B**, CSS) and Trimmed Mean of M-values[158] (**Figure 8.1B**, TMM) were tested here as control analytical methods, but are not expected to work in general as they are designed to identify scaling factors that do not change across samples. We noted that BEEM's significant improvement over other experimental and computational approaches, and its ability to closely approximate analysis using true biomass estimates is a robust feature that remains valid even when experimental biomass estimates are significantly better (CV=5%, as expected from flow-cytometry data) and while using different parameter estimation approaches (**Figure 8.1B**).

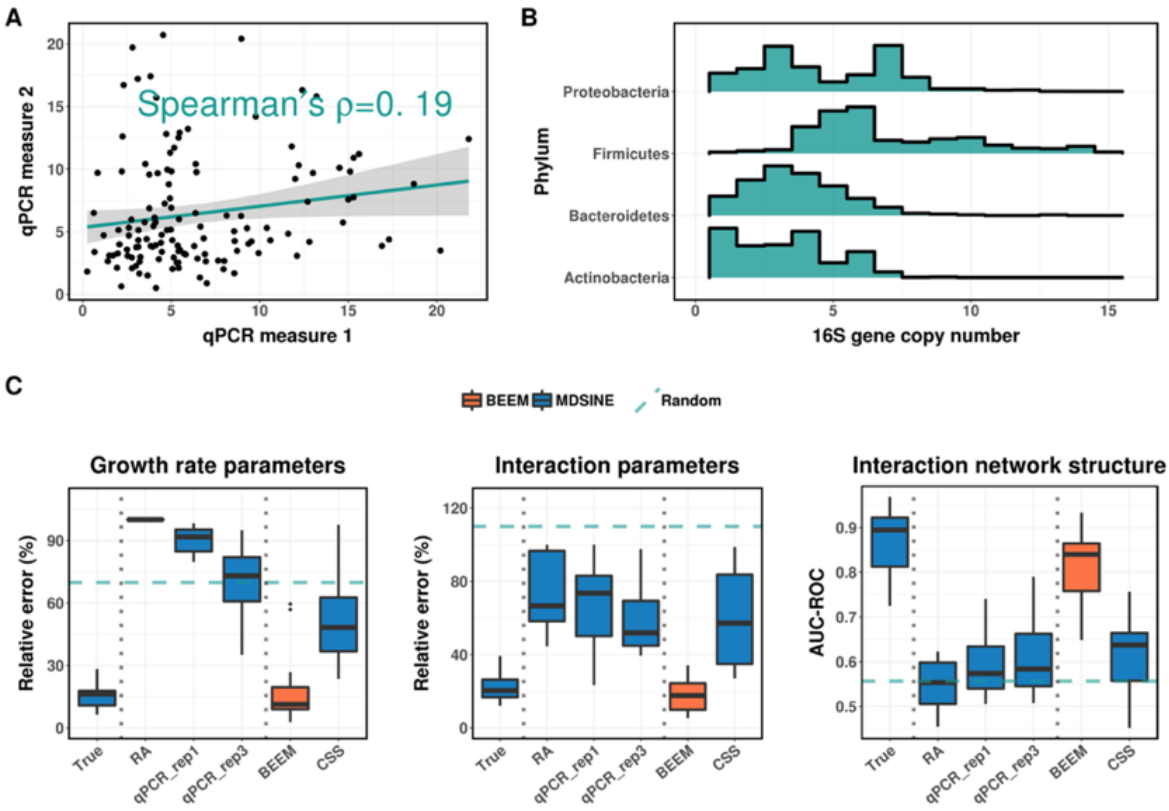

Figure 3.1 Noise in experimentally determined biomass severely distorts gLVM parameter estimation.

(A) Scatter plot with fitted linear regression line for two 16S qPCR technical replicates from Bucci *et al.* (B) Copy number variation for 16S rRNA genes in members of four major phyla of human gut bacteria. (C) Relative impact of different experimental (qPCR_rep1 – 1 qPCR replicate, qPCR_rep3 – mean of 3 qPCR replicates) and computational (RA – relative abundance, CSS – CSS normalization) data scaling approaches on gLVM parameter estimation (BVS algorithm for MDSINE), in comparison to using true biomass or using BEEM. Boxplots represent the summary of 15 simulations (10 species, 30 replicates with 30 time points each) and three different metrics are shown here including median relative error for growth rate ($\boldsymbol{\mu}$) and interaction ($\boldsymbol{\beta}$) parameters, and AUC-ROC for the interaction network. Dashed horizontal lines represent the performance of randomly generated parameters from the simulation model.

3.3.3 BEEM accurately estimates gLVM parameters and biomass in diverse model settings

As in any situation where parameters have to be estimated, a sufficient number of data points (multiple biological replicates) are needed to get accurate gLVM models and this in turn impacts BEEM's biomass estimates. In order to further study BEEM's performance characteristics, we generated synthetic datasets with varying number of species and data points, comparing BEEM's results to those obtained with noise-free biomass data and the same gradient matching algorithm (BLASSO) as used internally in BEEM. As expected, when the number of species increases but the number of data points remains constant (60 replicates with 30 timepoints), gLVM parameter estimation becomes harder (**Figure 3.2A**). However, despite the quadratic increase in the number of parameters, performance for both BLASSO (with true biomass) and BEEM seems to only degrade linearly (**Figure 3.2A**). In addition, even when the model has 25 species (650 model parameters) and can thus capture over 90% of the human gut microbiome[159] (**Figure 8.2**), interaction parameters estimated by BEEM were nearly as accurate as those with true biomass (**Figure 3.2A**), though growth rate parameters were more affected. We also noted that median relative error for biomass estimates from BEEM was generally well controlled (<10%; **Figure 3.2B**).

Increasing the number of data points available for model fitting for a fixed number of species (10) improved performance for both BLASSO with true biomass and BEEM, as expected. Performance improvements were most notable when going from 10 to 20 replicates and plateaued out after that (30 timepoints; **Figure 3.2C**). In general, after 20 replicates, differences between BLASSO and BEEM were small, especially in terms of estimating interaction parameters. Similarly, biomass estimates from BEEM had median relative error <5% when 20 replicates were available (**Figure 3.2D**). In

general, our analysis suggests that inherent limitations in gradient matching based on estimated gradients from data were a greater source of error for gLVM parameter estimation in many of our experiments than errors in BEEM estimated biomass values.

To assess BEEM's performance for biomass inference in real-world datasets we analyzed data from a recently published study on freshwater microbial communities[143,144], which to our knowledge is the only one to have sufficient longitudinal microbiome sequencing data as well as flow-cytometry based gold-standard biomass estimation. Notably, the flow cytometry data in this study was reported to have high reproducibility (CV<5%)[143], and therefore was suitable for use as the ground truth for total biomass. Surprisingly, with only 57 time points in total across two replicate experiments, BEEM was able to infer the total biomass for a 26-species community accurately solely based on 16S sequencing based relative abundances (Note that none of the species abundances had noticeable correlation with the true biomass, with a highest Spearman's ρ=0.41). BEEM estimated biomass trajectories closely tracked those obtained experimentally (**Figure 3.3A**), and showed strong correlation with ground truth (Spearman's ρ=0.72) while control values from CSS scaling exhibited weak correlation (Spearman's ρ=0.36; **Figure 3.3B**).

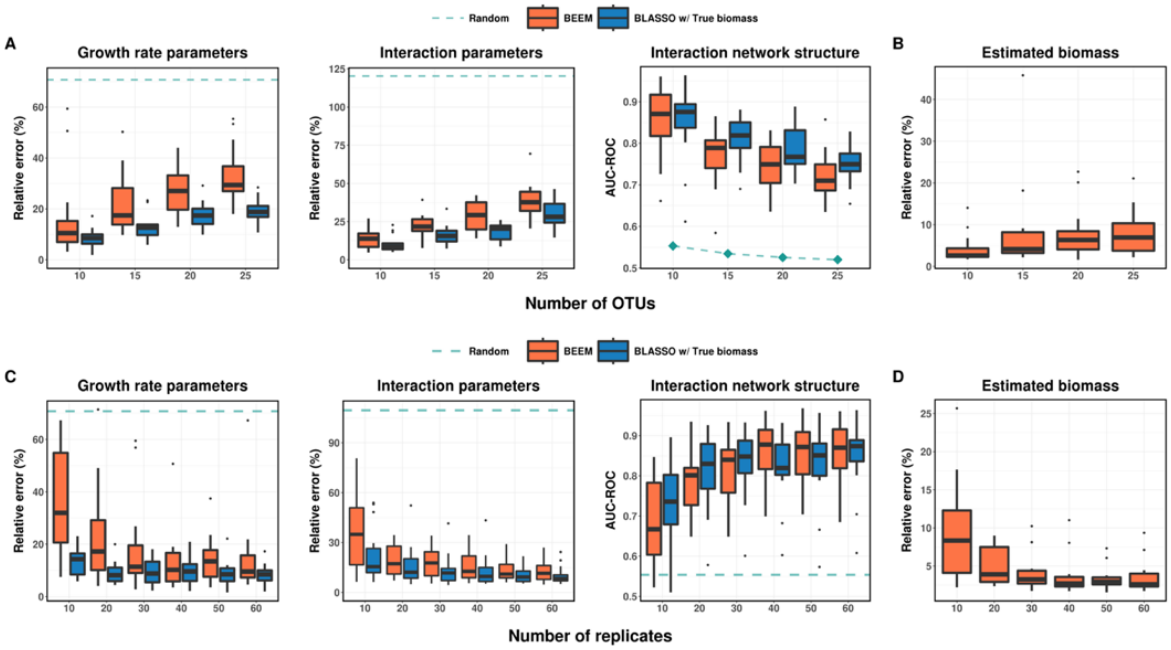

Figure 3.2 Robustness of parameter estimation with BEEM.

(A) Results with increasing number of species but fixed number of replicates (50). As expected, parameter estimation gets harder but BEEM's performance tracks the ideal case using BLASSO with true biomass values, especially for interaction parameters. (B) Median relative error in biomass estimates remains less than 10%. (C) Results with increasing number of replicates and fixed number of species (15). BEEM's performance converges to that of BLASSO with true biomass as the number of replicates increases. (D) Median relative error in biomass estimates reduces noticeably as the number of replicates increases.

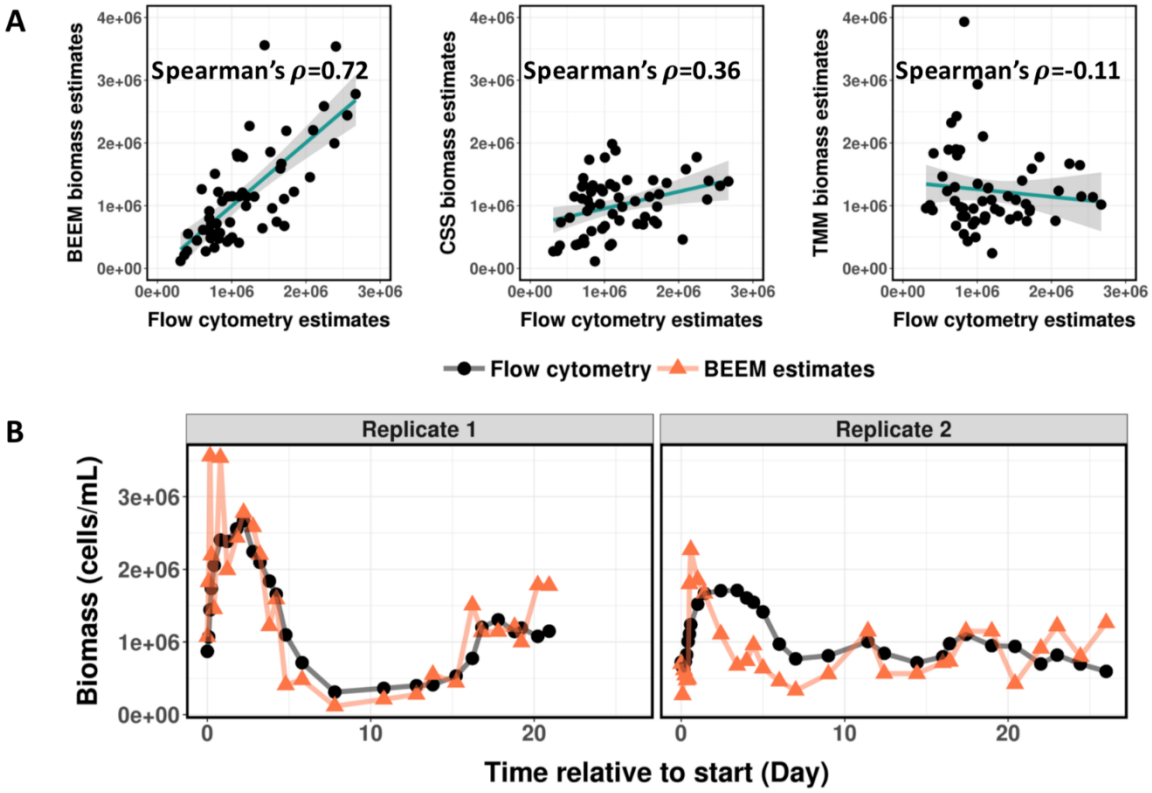

Figure 3.3 Concordance of BEEM estimated biomass with gold standard experimental measurements.

(A) BEEM estimated biomass values (orange) compared to gold standard measurements using flow cytometry (black). (B) Scatter plots with fitted linear regression line highlighting that BEEM's biomass estimates are notably more concordant with flow cytometry based values compared to CSS normalization based estimates.

3.3.4 Personalized gut microbial dynamics and keystone species

The development of BEEM allows us to analyze previously generated datasets in a gLVM framework, even when biomass measurements were not made in the original study. To showcase this capability, we applied BEEM to the longest (over one year) and most densely (almost daily) sampled human gut microbiome time-series datasets available to date (four individuals: DA, DB from David *et al*[151] and M3, F4 from Caporaso *et al*[150]). BEEM estimated models exhibited a good fit to the data, with predicted relative abundances for a day based on numerical integration from the previous day being in high concordance with observed data (median Spearman's ρ = 0.83). As BEEM directly infers daily biomass values, we plotted these and observed distinct individual-specific patterns: while subject DA's biomass was found to vary relatively smoothly, following an approximately cyclic pattern with a period of about three months (**Figure 3.4A**), subject M3's biomass fluctuated to a greater extent on a day to day basis with no clear trend (**Figure 3.4B**). Similar patterns were observed in parts for subjects DB and F4, which had a greater resemblance to DA overall (**Figure 8.3A, B**). The fluctuations predicted in M3's biomass were also found to be accompanied by frequent blooms of rare taxa that were otherwise not detected at other time points[152] and may be a consequence of this instability in the community. In contrast, the smoother progression of DA's biomass may be a reflection of the relative stability of the gut community in this individual, though the source of the observed cyclic patterns deserves to be explored further. As an initial hint, we noted that the strongest association between DA's biomass and reported metadata was a negative correlation with calcium intake (**Figure 8.4**).

Concordant with their distinct biomass dynamics, DA and M3 also exhibited microbial interaction networks that were unique to them (**Figure 3.4C, D**). DA's

network was defined by hub nodes for *Feacalibacterium prausnitzii* and *Bacteroides uniformis*, two species with many beneficial roles and frequent associations with a healthy gut[160,161]. The hubs were found to negatively affect the growth of *Enterobacteriaceae* species, consistent with previous reports for *B. uniformis*[162] and *F. prausnitzii*[163–165]. In comparison, the major hub nodes in M3's network were a *Blautia* and an *Oscillospira* species that were connected by a positive feed-forward loop. Additionally, we found that abundances of the *Blautia* and *Oscillospira* species were significantly negatively correlated with total biomass in M3's gut microbiome (**Figure 8.5**). Feed-forward loops have been implicated in destabilizing effects on ecosystems[142] and so these observations may explain the unstable behavior of M3's biomass as well as the corresponding susceptibility to invasive blooms of rare taxa[152]. *Oscillospira*'s protective role in M3's gut flora is further indicated by its parasitic relationship (negative-positive loop) with another hub species *B. fragilis*, an opportunistic pathogen that has been associated with diarrhea[166]. Interestingly, several of the transient species in M3's gut microbiome were observed to be at the periphery of the network, with a single incoming edge indicating that their abundances were being influenced by a hub species. For example, this was observed for several *Streptococcus* species that are primarily oral commensals and could be transient colonizers of the gut[167,168].

Despite differences in the identity of species in their interaction networks, the various individual-specific networks shared some common features, including the presence of a few hub nodes that negatively influenced many other species, and were generally not the most abundant species in the community (**Figure 3.4C, D** and **Figure 8.3C, D**). Overall, we also found that the ratio between out- and in- degree of species in the networks were negatively correlated with their median relative abundances (**Figure 8.6**), suggesting that the hub species in the interaction network, that are often considered

as keystone species for the community[75,169], are typically not the abundant species in the gut microbiome. We further confirmed this observation by analyzing a large collection (840 healthy individuals) of gut microbiome datasets[159], to find that the core species in the gut microbiome were also frequently not the most abundant species (**Figure 8.7**). Together, these observations suggest a model for the gut microbiome where relatively less abundant species in the community are more stable colonizers of the host, and by virtue of their impact on the growth of other species in the community, play an important role in defining its dynamics in different individuals.

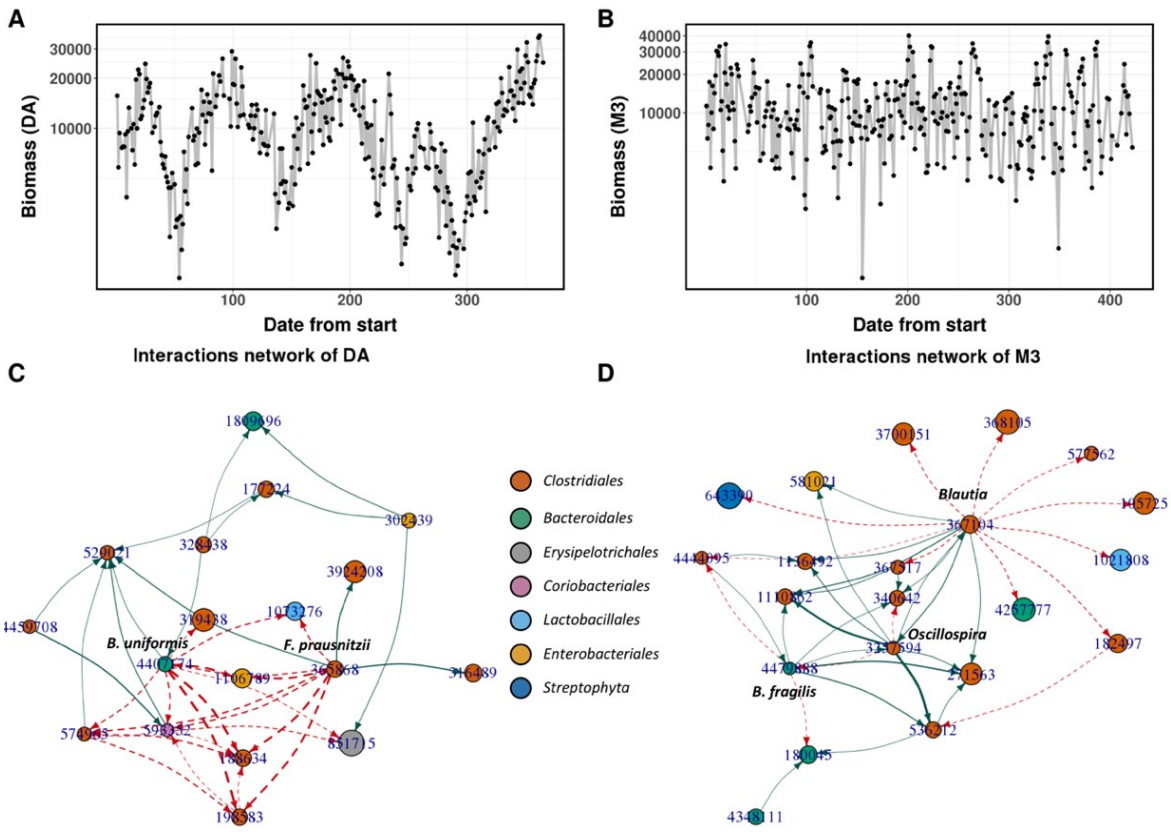

Figure 3.4 BEEM analysis of year long gut microbial time-series datasets.

(A,B) BEEM estimated biomass values for two individuals (DA and M3) with daily sampled, year long gut microbial time-series datasets from David et al[150] and Caporaso et al[151]. Interestingly, while M3's biomass fluctuates rapidly, DA's biomass seems to vary in a more defined fashion with a periodicity of around 3 months. (C, D) Graphs representing non-zero interaction terms in gLVM models learnt individually for DA and M3 using BEEM. Dashed and solid edges represent positive and negative interactions respectively. Edge widths are proportional to the interaction strength and node sizes are proportional to the log-transformed mean relative abundance of the corresponding species. Nodes are labeled with GreenGenes IDs and colored according to order level taxonomic annotations.

3.4 Discussion

A major limitation of most microbiome profiling datasets available to date is the restriction to relative abundances and the 'compositionality' of this data has led to significant challenges even when performing common statistical tests for correlated abundances[43]. These issues are amplified when considering systems models such as gLVMs, and our analysis here confirms that model parameter estimates can be severely distorted if relative abundances are not correctly scaled. In ecological models such as gLVMs, interactions between species are naturally a function of the absolute density of species in a community rather than their relative abundances. Correspondingly, while autoregression based methods such as sVar[152] and ARIMA[170] provide an alternative for model fitting with relative abundance data, their models and parameters are not ecologically interpretable. In addition, experimental approaches to measure scaling factors are generally seen as a laborious and occasionally feasible way to work with absolute abundances, but as we show here, this may not be the case if care is not taken to ensure that experimental noise is minimized and sufficient number of replicates are analyzed. By eliminating the need for additional experimental data, BEEM greatly expands the applicability of gLVMs to microbiome datasets, and its robustness could simultaneously improve the quality of models and scaling factor estimates, as observed in our synthetic and real datasets. Explicitly modelling microbial interactions through gLVMs has proven to be a powerful framework for studying microbial community dynamics[74–76,138–142], and the approach used in BEEM could also be extended (with minimal modifications) to time-series with external perturbations (e.g. antibiotics usage)[74,76,140], as well as systems models for gene expression regulation based on RNA-seq data[171].

Due to limited availability of absolute abundance data, gLVMs have generally been constructed by aggregating information across experiments and individuals[74,76,140]. We exploited the availability of year-long time series datasets and BEEM's facility with relative abundances to construct individual specific gut microbiome gLVMs for the first time. Intriguingly, we observed that our inferred scaling factors suggest that gut microbial biomass has distinct dynamics across different individuals. Consistent with a recent study on 20 individuals where human gut microbial biomass (measured via flow-cytometry) was found to have high variation (CV≈53% within a week)[43], we also noted high variability over time across the four individuals we analyzed (CV ranging from 49% to 76% over a year). Additionally, we observed cyclic behavior of biomass trajectories in multiple individuals, similar to the seasonal patterns reported in hunter-gatherers of western Tanzania[172], and the conserved patterns observed in other mammals across evolutionary timescales[173]. Similar patterns have not been reported before for western city dwellers, perhaps due to the confounding effects of aggregate analysis across individuals and the impact of highly diverse diets. BEEM analysis, however, suggests that the underlying patterns may still be conserved in urban subjects and may be more general than previously believed.

Our inference of gLVM models for each individual allows us to identify specific microbial species and the kinds of interactions that they have, to account for the distinct dynamics that were observed. For example, the positive feed forward loop observed between the hubs in M3's gut microbiome provides a specific, plausible and testable hypothesis to explain the instability observed there, and this capability can be valuable in future studies where targeted interventions are feasible. Despite differences in the microbial interaction networks observed for different individuals, a shared feature seems to be the presence of relatively lowly abundant species that act as hub nodes in

the network. A similar pattern was seen in cross-sectional data as well where frequently shared "core" gut microbiome species tend to not be the most abundant species in the community. These observations point to a model where species at low relative abundances stably colonize the gut (e.g. mucosa-associated ones) compared to abundant but transient (lumen-associated) bacteria and play an important role in defining gut microbiome dynamics. In particular, hub species were frequently found to negatively regulate more transient species in the community, in agreement with the known role of mucosa-associated species in providing colonization resistance against invasive pathogenic species[174].

An important point that we noted in the gut microbiome datasets that were analyzed here is the limited number of core species (prevalent in most time points for an individual) that are shared across individuals. This feature makes it infeasible to learn gLVM models by merging short time-series datasets across different individuals. Similar constraints might be present in other microbial communities as well, including specific challenges in measuring total biomass in complex matrices[43], and thus the development of BEEM makes it more feasible to generate the long and densely sampled datasets that are needed for such models. We also note that the analysis in BEEM can be directly extended to cross-sectional datasets if the corresponding communities are believed to be at equilibrium (i.e. $\frac{dx_i(t)}{dt} = 0$, for all species). This extension would significantly expand the amount of data that could be used and thus allow us to learn even more complex models in the future. As is the case for any modelling approach, no model is expected to be perfect, but as they capture more and more features of real systems, we can expect that their predictions become increasingly useful. BEEM's development therefore serves as an important step in expanding the use of modelling approaches to study microbial community dynamics and rationally identify appropriate perturbations.

4 UTILITY OF CORRELATION BASED METHODS TO INFER INTERACTIONS FROM MICROBIAL PROFILING DATA

4.1 Background

As discussed in previous chapters, interactions in microbial communities are often represented as weighted directed graphs based on their pairwise interactions (**Figure 4.1A**), where each node represents a microbial species and each weighted edge represents an interaction. The overall makeup of the microbiome is then determined by the interactions between different members in the community and the interaction these members have with their surrounding environment. A widely used high-throughput approach to study microbial interactions is to infer them computationally from microbial abundance profiles collected across different sites (i.e. cross-sectional microbiome survey data) by computing correlations between the abundances of different species. As reviewed in **Chapter 2**, many computational algorithms have been proposed to address various challenges in inferring microbial interactions using correlation based methods. However, the relative strengths and weaknesses of these methods, many of which have been recently developed, still remains unclear, particularly in the presence of data that is not derived from the models that these methods are based on. In this chapter we build upon previous benchmarking[175] and

review articles[133] as well as individual benchmarking techniques employed by publications of the different correlation based methods. We test the largest collection of different methods for inferring microbial interactions from cross-sectional microbiome survey data on both simulated and real datasets. Our evaluation procedures aimed to challenge the different methods by basing our simulations on both statistically modelled data with defined correlation structure, as well as ecologically modelled data using generalized Lotka-Volterra models (gLVMs).

Interestingly, our analysis shows that correlation-based methods predict microbial interaction networks reasonably well on statistically simulated data but perform poorly on data simulated from gLVMs. Additionally, we show that most correlation-based methods are not robust to variations in input features (e.g. number of samples, number of species etc.) and lack concordance with one another when applied to real datasets. These observations suggest that results from existing correlation based algorithms should be viewed with caution, especially when the goal is to infer ecological interactions from microbiome datasets.

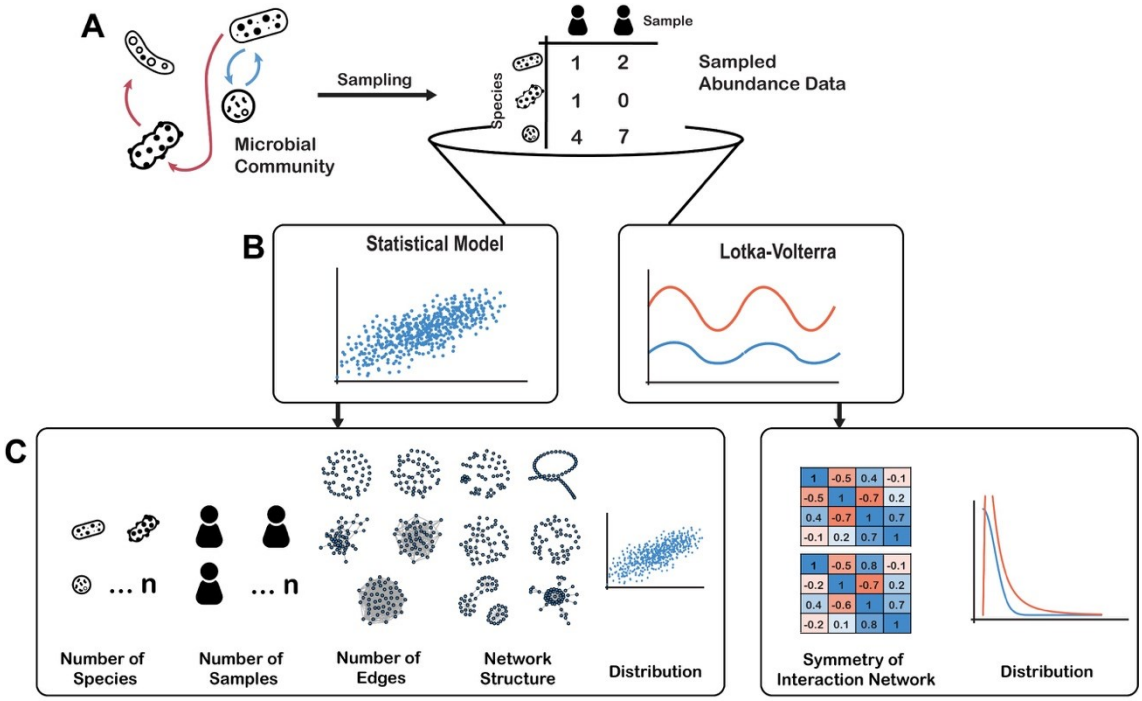

Figure 4.1 Overview of microbiome survey data and benchmarking data production process.

(A) Microbiome survey data is produced by sampling microbial communities and producing abundance tables where rows represent species and columns represent samples. These tables are analyzed using different algorithms to reconstruct correlation networks that represent the real underlying community make-up. (B) Microbiome survey data is fed into different models to produce synthetic data with known interactions. Two main models were used to create synthetic data: a statistical model, and an ecological model (generalized Lotka-Volterra model). (C) Synthetic data was generated with varying features to analyze which correlation detection tools were most robust to these different features.

4.2 Methods

4.2.1 Generation of synthetic datasets using statistical models

In order to capture the properties seen in real data, we generate simulated microbiome profiles using the normal-to-anything approach implemented in the SPIEC-EASI package[48], based on the real OTU count data (provided along with the package) and a defined correlation network topology as the ground truth. The simulator was run, varying input parameters to control for different features in the simulated dataset (**Figure 4.1C**). We set the default conditions in the simulation with the following features: 50 species (composite a median 69% of the real OTU data), 100 samples (an over-estimated number of samples for a typical microbiome study), a zero-inflated negative binomial distribution (shown to best fit the real OTU data[48]), one interaction per species and random (Erdős–Rényi model) graph structure. We then generated different simulations varying each feature one at a time (**Table 8-1**).

4.2.2 Test datasets based on synthetic microbial communities

Simulated data was generated following the same procedure as described in **Chapter 2** following the equation:

$$\frac{dx_i(t)}{dt} = \mu_i x_i(t) + \sum_{j=1}^{p} \beta_{ij} x_i(t) x_j(t).$$

To simulate samples with p species in total, we numerically integrated the gLVM equations with known interaction and growth parameters until a steady state was reached in each sample (all abundances change less than 10^{-5}) with initial abundances of species sampled from a uniform distribution (from 0.001 to the mean carrying capacity $-\mu_i/\beta_{ii}$ of all species). Absence-presence patterns as observed in real microbiome profiling data was introduced by setting the initial abundance of each species to zero

with a probability π in each sample, where π was estimated from the average absence rate of the top p species (ranked by the number of non-zero entries) in all healthy adult gut samples from the curatedMetagenomicData database (v1.7.92)[159]. The interaction matrix was forced to be "symmetric" by default ($\beta_{ij} = \beta_{ji}$) or "asymmetric" by randomly setting off-diagonal entries β_{ij} or β_{ji} to 0 when both of them were non-zero. To simulate an unevenly distributed community, instead of sampling self-interactions from a normal distribution (default), we sample from a log-normal distribution parameterized by inferred values from MDSINE using the *C. difficile* infection dataset provided with the software[140].

4.2.3 Evaluating predicted interaction networks against ground truth

Real microbial survey data was used to parameterize the data simulators described above (**Figure 4.1A**) to generate simulated datasets. In order to test the amount of data needed by the benchmarked methods for accurate inference, we varied the number of samples and the number of interaction parameters (by changing the number of species). We also examined the effect of the number of edges present in the interaction network as many of the methods rely on the edge sparsity assumption (**Figure 4.1C**). Besides, as real microbial communities have unknown properties on the structure of interaction network and the distribution of species abundances, the methods were tested for their robustness to varied condition of these properties.

To evaluate the performance of the algorithms on synthetic datasets, the inferred interactions (non-zero entries in the correlation or partial correlation matrix output by the algorithms) were ranked based on the absolute values of the correlation coefficients (partial correlation coefficients for MInt and SPIEC-EASI) or the confidence values (stability for SPIEC-EASI and -log(p-value) for CCREPE), and the area under the precision recall curve (AUPR) was calculated based on the ground truth correlation

(statistical model) or interaction network structure (gLVM). For gLVM simulated data with asymmetric interaction matrix, a predicted edge was considered correct if there is an interaction in either direction. To evaluate the ability of using partial correlation to address false positives due to transitive associations, we also calculated the AUPR using the partial correlation matrix directly computed from the correlation matrix using the R package 'corpcor'. The AUPR reported in the following sections is based on correlations if not specifically stated otherwise.

4.2.4 Evaluation of robustness of interaction networks inferred for a real microbial community

For a real dataset, it is not possible to calculate the accuracy of the methods due to lack of known interactions. However, a method should be robust and produce consistent results on random subsets of the dataset (given large enough sample size for each subset). Therefore, we use the consistency of inferred networks from subsampled datasets to evaluate the performance of the methods.

Count data was extracted from the American Gut Microbiome project OTU datasets[176]. We selected only samples from healthy individuals and the resulting abundance table was then filtered to remove species that are not commonly detected (present in <40% of samples), resulting in a data table with 62 species and 3233 samples. The dataset was randomly partitioned into four subsets (with 62 species and approximately 800 samples), on which each method was run to test for the consistency of its predictions. The top 300 edges (ranked as in **Section 4.2.4**) inferred from different partitions as well as the full dataset were compared to calculate the Jaccard similarity (the size of intersections over the size of union) as a measure of reproducibility.

4.2.5 Correlation based methods included

As summarized in **Table 4-1**, all correlation based methods reviewed in the previous chapter (**Section 2.2**) were tested with two new recently published methods BAnOCC[177] and gCoda[178]. Both BAnOCC and gCoda were developed to account for false positives caused by compositionality bias as well as indirect associations.

4.3 Results

4.3.1 Correlation based methods vastly vary in performance and robustness on data simulated from a parametric model

4.3.1.1 Using partial-correlation over correlation-based analysis does not improve performance as expected

Correlation based methods generally output either an inferred correlation matrix or a partial-correlation matrix (**Table 4-1**). In this analysis we evaluated both matrices for all methods, except for CCREPE as it does not provide meaningful correlation coefficients (**Chapter 2**). Theoretically the partial-correlation matrix removes correlations due to transitivity, and therefore we expect higher AUPR with it. However, we typically obtained lower AUPR values for our synthetic datasets when using partial correlation values, except for the methods BAnOCC and gCoda. This is likely due to the loss of true edges when converting correlation matrices to partial-correlation matrices (**Figure 8.8**), suggesting that partial correlation based methods can be too conservative under the conditions tested. For results reported in the rest of this chapter we use correlation matrix outputs where feasible (i.e. except for MInt and SPIEC-EASI), rather than partial-correlation matrix outputs as this provides better results in general.

4.3.1.2 Identification of key factors that influence performance variability in interaction network reconstruction

We examine different factors about the input data that potentially affect the performance of the algorithms, including number of samples, number of species, number of edges in

the network, network structure, and the distribution of species abundance across samples.

Kurtz *et al* noted previously that sample number has a large effect on the performance of correlation based methods[48]. As seen in **Figure 4.2**, CCREPE had the best performance when the sample size was less than 100, potentially attributed to the robust non-parametric test used to remove false positives. Most of the methods based on log-ratio transformation and sparse regression benefitted from increasing number of samples (CCLasso, REBACCA, SparCC, MInt, SPIEC-EASI, BAnOCC and gCoda). CCLasso and BAnOCC were found to have the best performance when the number of samples were large (1000). Interestingly, a simple approach using Spearman correlation had comparable or better performance than many methods on datasets with a medium sample size (500).

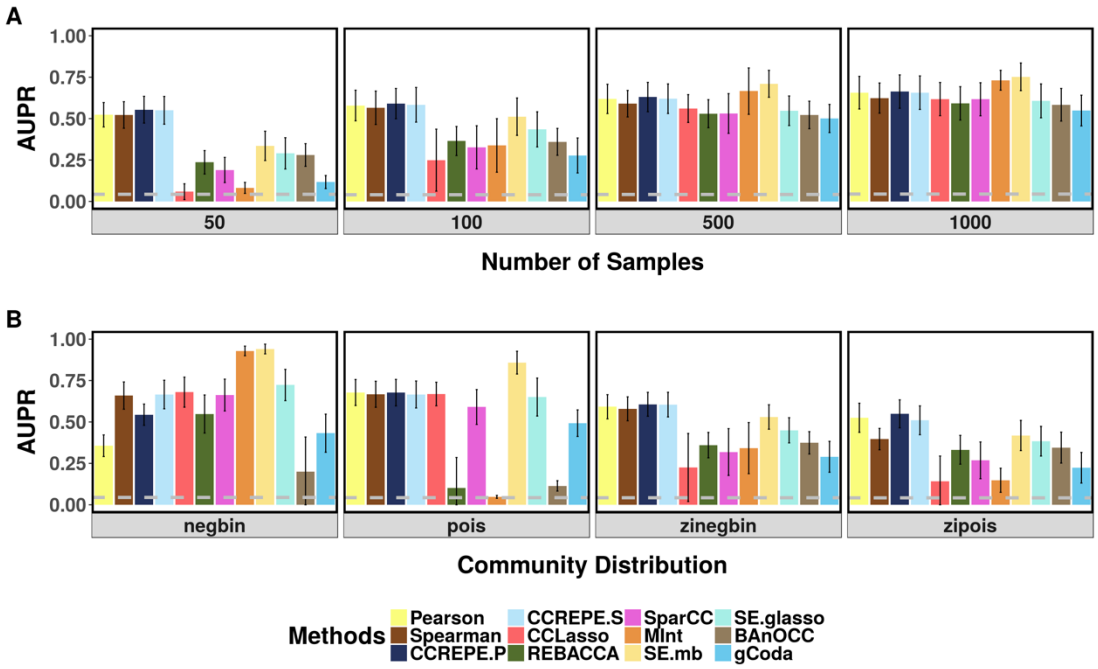

Figure 4.2 Performance of correlation-based methods on data simulated by statistical model.

Bars represent the mean area under precision recall of 30 replicates and the error bars represent the standard deviations comparing the reported correlation matrix with the ground truth precision matrix. Simulated data was generated with the following features: 50 species, 100 samples, 50 edges, random network structure, and zero inflated negative binomial distribution. (A) Results as a function of the number of samples in the dataset. (B) The impact of community distribution on performance. "negbin": negative binomial distribution, "pois": Poisson distribution, "zinegbin": zero inflated negative binomial distribution, "zipois": zero inflated Poisson distribution. CCREPE.P: CCREPE using Pearson's correlation, CCREPE.S: CCREPE using Spearman's correlation, SE.mb: SPIEC-EASI using its MB algorithm, SE.glasso: SPIEC-EASI using its glasso algorithm. Dashed lines represents the average performance of 30 random networks based on permuting the edges in the ground truth network.

The distribution of species abundance also had notable impact on the performance, with most methods performing best on data modelled under a negative binomial distribution. Most methods decrease in performance on data generated from zero inflated distributions. As discussed in the previous section (**Section 2.2**), frequently appeared zero values could become outliers that confuse the correlation based methods. Interestingly, we see a substantial drop in performance in REBACCA, MInt, and BAnOCC when the data was generated from a Poisson distribution, probably due to overfitting to over-dispersed models (variance greater than mean).

Factors that had a limited effect on performance were number of edges, number of species, and network topology (**Figure 8.9**). As expected, increasing the number of species generally decreased the performance of most methods due to the quadratic increase in the number of parameters (edges) to estimate, and SPIEC-EASI was the most robust to this effect. Increasing the number of edges increased the performance of most methods. However, this is likely an artefact caused by higher chance of inferring an edge correctly as can be seen from the increase in the performance of randomly permuted networks.

4.3.2 Correlation based methods fail to capture interactions in an ecological model

While some of the methods was able to estimate the correlation structure accurate as shown above, we are typically more interested in how well the correlation inferred can capture ecological interactions. To test this, we adopted the approach proposed by Berry & Widder[36] to generate simulated datasets from gLVMs – a simple yet widely used model for microbial ecology, and compare the inferred correlation matrix with the ground truth interaction matrix.

Note that in a gLVM, the effect of an interaction from a source species on a target species is a function of the interaction coefficient as well as the densities (absolute abundances) of both species. As the correlation based methods do not infer direction of interaction, they are expected to infer the interaction more accurately when the interaction between two species is symmetric (i.e. interacting in both directions with the same sign) and the abundances of the interacting species are close.

As we expect, all correlation based methods performed the best when the samples had even species abundance distribution and the interaction network was symmetric (**Figure 4.3**). However, performance dropped for all methods when the interactions were asymmetric and dropped further if species abundances were uneven in each sample. As for real microbial communities, it is expected that the species abundances are uneven and that the interactions are composed of different types including both symmetric and asymmetric. In addition, the parasitic interactions (i.e. positive-negative interacting pairs) could further complicate the observed species abundance pattern, making it difficult for correlation based methods to capture.

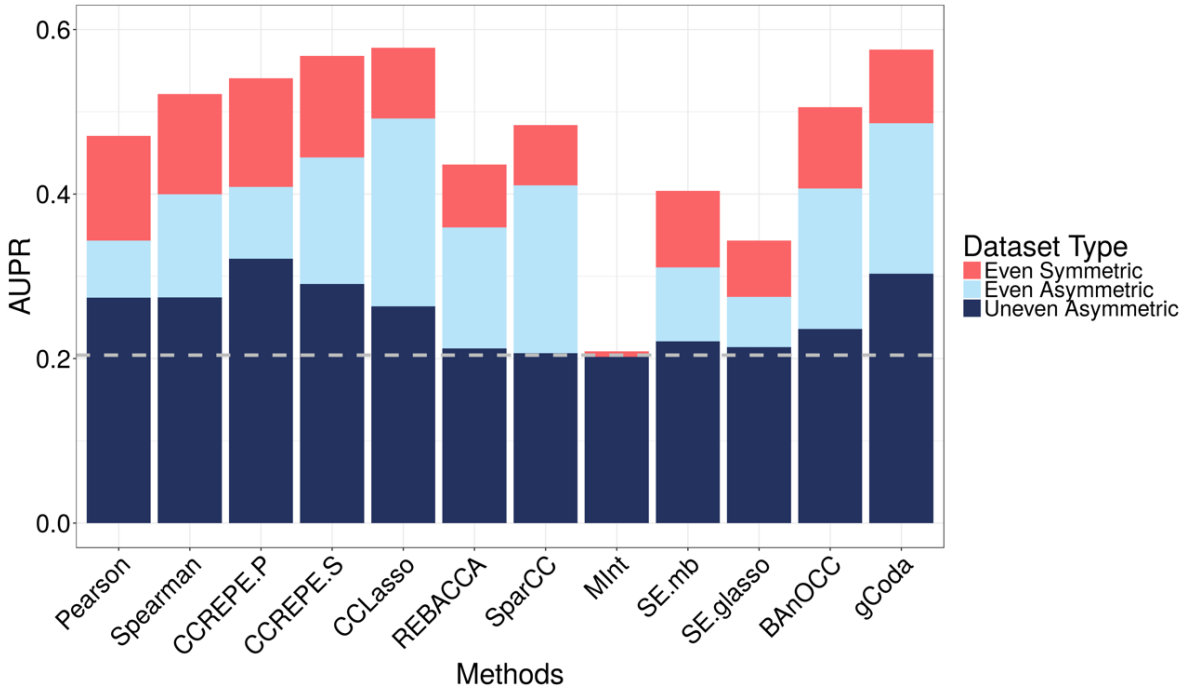

Figure 4.3 Performance of correlation-based methods on data simulated by gLV model.

Overlaid bar-plot where bars represent the AUPR values (mean of 30 replicates) of inferred interactions with correlation based methods (Symmetric/Asymetric: all interactions in the network are symmetric/asymetric, Even: evenly distributed species abudance in each sample, Uneven: log-normally distributed species abundance in each sample).

To compare the performance of the methods on our two different simulation schemes, we generated two similar datasets with the statistical and ecological models with the same features. Despite having a large sample size, we observe that all methods perform much worse on data generated using the simple ecological model (**Figure 8.10**), suggesting that correlation based measure may not be a proper indicator for ecological interactions if the underlying mechanism is close to the gLVMs.

4.3.3 Correlation based methods have low stability and concordance

4.3.3.1 Correlation based methods have low concordance with one another on both real and simulated data.

To investigate method concordance we ran each method on two different datasets: a real OTU dataset from the American Gut Microbiome project[176], and a simulated dataset with 1000 samples. We observe that most methods generally had low concordance with others (**Figure 4.4**). Methods based on variants of log-ratio transformation (**Chapter 2**) produced more similar results to each other on both simulated and real datasets. Notably, the most recent developed algorithms BAnOCC and gCoda had largely overlapped predictions likely due to the similar underlying log-normal latent structure assumed, but gCoda is much more scalable in computational resources (**Table 4-1**).

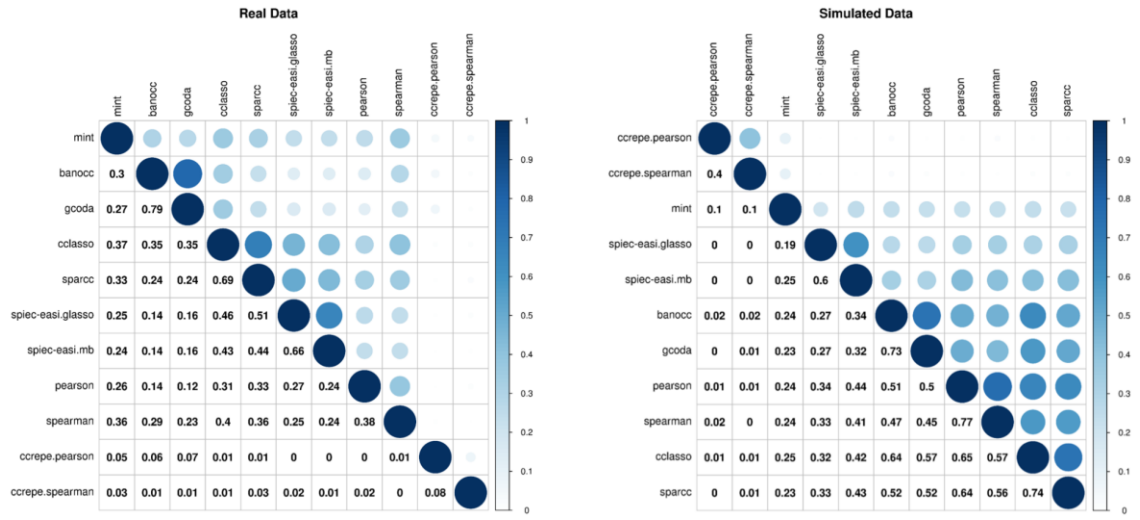

Figure 4.4 Correlation based methods have low stability and low concordance with one another.

Each circle on the upper corner represents the Jaccard similarity (proportional to both color and size of the circle) between two methods and the numbers on the bottom corner are the exact values. Methods were ran on a real dataset (left, 62 OTUs and 3233 samples) and a simulated dataset (right, 50 species, 1000 samples, 50 edges, random network and zero-inflated negative binomial distribution). REBECCA failed to run on the real dataset with 100GB of memory and was not included.

4.3.3.2 Correlation based methods have low concordance within themselves on subsampled data.

We evaluate the stability of inferred interactions by all the methods using randomly partitioned American Gut dataset (**Figure 8.11)**. Surprisingly, most of the methods had low concordance between partitions (Jaccard similarity around 0.5), with the exception of SPIEC-EASI, which particularly emphasize on the robustness to subsampling and reports edge confidence based on the stability of predictions across subsampled datasets. We also note that CCREPE had the worst stability (Jaccard similarity around 0.12) in general despite it is the most accurate method on small datasets.

4.4 Discussion

Currently, correlation based methods are the most widely used tool to infer microbial interaction networks from cross-sectional microbiome survey data. Confidence in these inferences is important considering the fact that studying these interactions in the lab is practically impossible given the work intensity and current culturing techniques. While newer and more sophisticated methods are being developed for this task, the baseline goal of uncovering microbial interaction networks through correlations is only valid if these methods actually have adequate performance. However, with the results from the systematic investigation on a large set of correlation based algorithms for detecting microbial interactions, we are greatly concerned about the varied performance of these methods on the different features present in the input data.

On datasets generated from statistical models, the most critical factor affecting the performance of correlation based methods was the sample size. Considering the large number of variables to infer, we found that hundreds of samples were needed for accurate inference of the correlation structure used to generate the data even though many of the methods employ sparse constraints (i.e. lasso regression). Nevertheless, the

sample size is usually limited by the resources, and caution must be taken when choosing and applying correlation based methods for analysis. Assuming microbial interactions resulted in correlated abundances, we have come up with recommendations on the usage of these methods as can be seen in **Table 4-1**.

Table 4-1 Summary of different correlation-based methods with recommendations on use

Method	Similarity Metric	Recommendations	Performance Metrics*		Availability (version)
			Average CPU runtime (s)	Average maximum memory usage	
CCREPE	Pearson/Spearman correlation (or any similarity measure)	Use if you have a small sample size (less than 100). Keep in mind that this method has low concordance with itself and other methods.	28.6±4.3	496.6 MB	http://huttenhower.sph.harvard.edu/ccrepe (v1.2.0)
SparCC	Linear Pearson correlations between the log-transformed components	Only use with large sample size (at least 500). Underperforms compared to CCLasso which uses similar methodology.	45.2±16.1	202.2 MB	https://bitbucket.org/yonatanf/sparcc
REBACCA	Correlations between the log-transformed elements	Only use with large sample size (at least 500). Underperforms compared to CCLasso which uses similar methodology.	54.7±11.7	355.9±7.5 MB	http://faculty.wcas.northwestern.edu/~hji403/REBACCA.htm
CCLasso	Pearson correlation between log absolute abundances	Only use with large sample size (at least 500). Outperforms SparCC and REBACCA which use similar methodology. Performs considerably well on gLV data.	60.3±11.3	242.3 MB	https://github.com/huayingfang/CCLasso (v1.0)
MInt	Partial correlation between sequencing counts	Only use with larger sample size (close to 500). Performs badly on ecologically modelled data.	653.0±100.5	189.2 MB	https://cran.r-project.org/web/packages/MInt/vignettes/MInt.html (v1.0.1)
SPIEC-EASI	Partial correlation between log transformed absolute abundance	Only use with larger sample size (greater than 100). Neighbor selection (mb) model performs best. Has greatest concordance.	106.3±20.3	343.1±4.3 MB	https://github.com/zdk123/SpiecEasi (v0.1.4)
BAnOCC	Log-basis correlation matrix and precision matrix	Requires most computational resources, not recommended if sample size is small (less than 500). Has relatively high concordance with itself and relatively good performance on gLV data.	.2792.4±131.5	3.6±0.1 GB	http://huttenhower.sph.harvard.edu/banocc (v1.0.1)
gCoda	Inverse covariance matrix	Requires large sample size (at least 500). Performs reasonably well on gLV data.	4.9±2.3	348.4±29.2 MB	https://github.com/huayingfang/gCoda (commit ID 584bd07)

*all methods were run using 1 core apart from BAnOCC which was run using 10 cores.

Intriguingly, we observe low concordance when running each method on subsampled datasets from the American Gut Microbiome project[176]. As previous analysis has suggested, the interactions are not likely to change across samples for human gut microbiome[179]. A plausible explanation is that the interactions among microbial members result in more complex association between species abundances than what correlation based methods can measure. On simulated data generated using simple ecological models (gLVMs), we confirmed that correlation based methods suffered from substantial drop in their accuracy in capturing the true interactions compared to the data simulated with a statistical model. Despite its conceptual simplicity, gLVM is known to be able to result in complex abundance patterns including the cyclic behaviour, where the correlation between abundances can be either

positive or negative depending on their abundances[180]. Assuming an ecological model like the gLVM closely captures the dynamics of the microbial community of interest, one potential way of inferring interactions was to learn the model directly from the data although theoretical and technical challenges still remain to be addressed[115].

5 BEEM-STATIC: EXTENDING THE BEEM FRAMEWORK TO LEARN ECOLOGICAL MODELS FROM CROSS-SECTIONAL MICROBIAL PROFILING DATA

5.1 Background

The interactions among members of a microbial community are important components of the behaviour and function of the microbial ecosystem. Such interactions are characterized by three basic factors: the direction (A affects B or B affects A or both ways), the sign (negative, positive or neutral) as well as the strength. Characterizing the microbial interactions is a key step towards understanding the ecology of a microbial community[137,181], forecasting microbial dynamics and designing potential interventions[182]. In a laboratory setup, microbial interactions are usually investigated using co-growth experiments, but such experiments suffer from its low throughput and the limitation that a large number of microbes cannot readily be grown in a lab[4].

As a high-throughput approach, computationally inferring microbial interactions has now become a popular way to provide a set of candidate interactions for further study. The generalized Lotka-Volterra models (gLVM) consist of sets of mathematical equations that capture the response of the growth of a species to the change in densities (absolute abundances) of other species, and are the most extensively used to study different types of pairwise microbial interactions[75,140,157]. The parameters of the gLVMs

(growth rates and interaction matrix) are often learned from long and dense longitudinal microbiome profiling data. In **Chapter 3**, we describe a novel expectation-maximization (EM) like approach named BEEM that overcomes the major challenge in gLVM inference due to the lack of accurate total biomass data. BEEM alternates between estimating gLVM parameters and biomass values to eliminate the needs for experimentally measured biomass scaling factors.

While longitudinal microbiome data is ideal for studying the dynamics of microbial abundances, it is much more scalable to generate cross-sectional microbiome data across different sites, especially different human subjects and a large amount of such data have been made available publicly[159]. Leveraging the easy availability of cross-sectional data, many computational methods had been developed to infer microbial interactions by calculating correlations between the abundances of microbial members, and such methods have become the most widely used approach for inferring microbial interactions (**Chapter 2**)[133]. However, one major limitation of correlation based approaches is that they are not able to reveal the directionality of the interactions, obscuring the relationship between the interacting microbes. In addition, correlation-based approaches have variable accuracy and reproducibility in inferring ecological interactions as was shown in our benchmarking experiments in **Chapter 4**.

On cross-sectional data, there is no information on the change rates of abundances comparing to the longitudinal data, so it is not possible to learn the gLVM parameters with the methods we described in **Chapter 3** directly. However, the abundance of one species is still expected to vary due to the presence-absence pattern of another species interacting with it and such variation contains the information to estimate the gLVM parameters if we assume that all the samples are at the steady states (i.e. the abundances of species will not change without external perturbation). The

theoretical feasibility of such idea has recently be demonstrated in a recent work[115]. Nevertheless, the proposed algorithm still requires absolute abundances, and has to make the additional assumption that the total biomass is constant across all the samples. Such assumption is very likely to contradict the presence-absence pattern required for the model inference. Furthermore, it is not clear if it is adequate to assume that all the species are at the steady state for every sample and if the inference accuracy will be affected if such assumption is violated.

In this chapter, we derive a new optimization problem from the gLVM assuming all the samples are at their steady states. We solve such problem with a novel algorithm for estimating both biomass and gLVM parameters from cross-sectional relative abundance data, named BEEM-static, inspired by the expectation maximization like framework adapted by BEEM. With extensive simulated data, we show that BEEM-static accurately infers biomass values as well as gLVM parameters and provides significant improvement over a family of methods (**Chapter 3**) for inferring interactions based on correlations. In addition, BEEM-static is robust to the presence of samples up to 50% that are perturbed from the steady states by automatically detecting and removing such samples. On a large human gut metagenomics dataset, BEEM-static successfully filtered out samples with perturbed microbiome by antibiotics. Meanwhile, the inferred biomass values for different age groups have consistent trend with experimental results. Furthermore, we highlight that gLVMs learned with BEEM-static from snapshot data have the potential to reveal dynamic information of the microbiome. We propose a novel analysis to predict species *in situ* growth with BEEM-static fitted model and showcase that the predicted growth trends are consistent with the change in DNA replication rates measured independently and offer a way to forecast the species abundances.

5.2 Methods

5.2.1 BEEM-static derivation

Generalized Lotka-Volterra models are typically written in the form of the following system of equations:

$$\frac{dx_i(t)}{dt} = \mu_i x_i(t) + \sum_{j=1}^{p} \beta_{ij} x_i(t) x_j(t) \,,$$

where $x_i(t)$ is the absolute density of species i at time point t, μ_i is the growth rate of species i and β_{ij} is the interaction term that defines the strength of the influence of species j's abundance on species i's growth.

At the non-trivial equilibrium ($\frac{dx_i(t)}{dt} = 0$ and $x_i > 0$):

$$\mu_i + \sum_{j=1}^{p} \beta_{ij} x_j = 0, \qquad (1)$$

where the time parameter t now becomes implicit in the equation. We divide by -β_{ii} and the biomass m on both sides, move the x_i term to the other side, and re-parameterize the equation to get:

$$\tilde{x}_i = \frac{a_i}{m} + \sum_{j=1, j\neq i}^{p} b_{ij} \tilde{x}_j \,, \quad (2)$$

where $a_i = -\frac{\mu_i}{\beta_{ii}}$, $b_{ij} = -\frac{\beta_{ij}}{\beta_{ii}}$, $\tilde{x}_i$ is the relative abundance of species i at equilibrium and the biomass $m = \sum_{i=1}^{p} x_i$ (thus m is fixed at the equilibrium). We then simultaneously estimate biomass and model parameters using an EM-like framework as detailed below:

Estimating model parameters (E-step): We estimate the model parameters for each species i with sparse regression (implemented with the 'glmnet' package in R) in iteration T:

$$\tilde{x}_i \sim a_i^{(T)} \cdot \frac{1}{m^{(T-1)}} + \sum_{j=1, j\neq i}^{p} b_{ij}^{(T)} \tilde{x}_j.$$

Estimating biomass (M-step): for a sample, the equation for each species i provides an estimate for the biomass, and we take the median of these estimates as a robust estimator for the biomass of the sample:

$$m^{(T)} = \text{median}\left(-\frac{a_i^{(T)}}{\sum_{j=1}^{p} b_{ij}^{(T)} \tilde{x}_j} \right).$$

Initialization and termination: the biomass values are initialized with normalized abundances using cumulative sum scaling (CSS[154]), with a user defined scaling constant used as the median of biomass values (kept constant through the EM iterations). The EM process is run until convergence when the median of relative changes in biomass values is smaller than 10^{-3}.

5.2.2 Detecting samples violating the equilibrium assumption

Samples that are not at equilibrium do not satisfy equation 1 and are therefore likely to result in inaccurate estimates for the biomass and model parameters. In each iteration, we calculate the median of squared error for each sample (including all the samples removed by the previous iteration) k in the E-step:

$$e_k = \underset{\tilde{x}_i \neq 0}{\text{median}} \left(\left(\tilde{x}_i - \hat{a}_i^{(T)} \cdot \frac{1}{m^{(T-1)}} + \sum_{j=1, j\neq i}^{p} \hat{b}_{ij}^{(T)} \tilde{x}_j \right)^2 \right).$$

We then remove samples for the next iteration's E-step to fit the regression if the median squared error is large ($\frac{e_k - \text{median}(e_k)}{\text{IQR}(e_k)} > \epsilon$, where IQR is the inter-quartile range and ϵ is a user defined parameter with 3 as its default value).

5.2.3 Selecting shrinkage parameters for sparse regression

The shrinkage parameter λ in the sparse regression penalizes the number of parameters to avoid overfitting and was selected based on 5-fold cross-validation in each iteration (selecting the value one standard error away from the best λ[183]). In the E-step of the first iteration ($T = 1$), a crude selection of $\lambda_c^{(1)}$ is made from a sparse sequence with large range (10^{-10} to 10^{-1}), then refined from a fine grain sequence from $\frac{\lambda_c^{(1)}}{10}$ to $10\lambda_c^{(1)}$. In latter iterations ($T > 1$), the $\lambda^{(T)}$ is adjusted from previous iterations with a sequence from $\frac{\lambda^{(T-1)}}{2}$ to $\lambda^{(T-1)}$ (the upper bound $\lambda^{(T)}$ is not updated to avoid progressively penalizing the parameters to reach an extremely sparse model).

5.2.4 Generating simulated data

Simulated data was generated following the procedure described in **Chapter 2** and **Chapter 3** by numerically integrating the model with known parameters. The off-diagonal interaction parameters were generated randomly without enforcing symmetry or asymmetry and the diagonal entries were generated following a normal distribution. To generate a sample which is not at steady state, a random time point along the numerical integration (excluding the first five time points) where more than 50% of species differ by >20% from steady state abundances was selected.

5.2.5 Analysis of gut microbiome data

Healthy gut microbiome profiles from the curatedMetagenomicData database were preprocessed and used as the standard dataset for learning gLVMs by removing (1)

replicate samples, (2) timepoints other than the first timepoint in longitudinal studies, (3) samples from antibiotic treatment timepoints and (4) samples from infants. In addition, we included three validation datasets to evaluate different aspects of the model learned by BEEM-static: (1) all samples from Raymond *et al*[184] to validate the ability of BEEM-static to filter out samples violating the model and growth estimation, (2) samples from healthy infants (only the first timepoint for each subject) with ages below 12 months to validate the biomass estimation and (3) all samples after fecal microbiome transplantation from Li *et al*[185] to evaluate the *in situ* growth estimation. To make the number of parameters tractable with the number of data points, we only kept core species that were present (>0.1%) in more than 30% of the samples and subsequently removed samples with the core species composite less than 30% of the total abundance in the standard dataset, resulting in 45 core species and 2962 samples.

5.2.6 Estimating *in situ* growth using BEEM-static and GRiD

With BEEM-static, the *in situ* growth rates are defined by the deviation from the equilibrium:

$$\hat{a}_i + \hat{m} \sum_{j=1}^{p} \hat{b}_{ij} \tilde{x}_j \,,$$

where $\hat{a}_i$, $\hat{m}$ and $\hat{b}_{ij}$ are estimated parameters. In addition, species replication rates for each sample in Raymon *et al* were estimated with the high-throughput mode of GRiD (v1.2.0; default parameters)[185]. GRiD values estimate the DNA replication rate as a proxy for the growth rate by computing the peak-to-trough ratio of the short read coverage of a genome. The genomes for stool samples provided with the software were used as the references and read reassignment using pathoscope2[186] was enabled (parameter "-p") to resolve ambiguous mappings.

5.2.7 Evaluation metrics

We compute median relative error to assess the accuracy of predicted parameters numerically as:

$$\text{median}\left(\frac{|\hat{\theta}-\theta|}{\max(|\hat{\theta}|,|\theta|)}\right),$$

where $\hat{\theta}$ and θ are the estimated and true parameters ($\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{m}$) respectively. The area under the receiver operating characteristic curve (AUC-ROC) was computed for the interaction matrix. The absolute values of $\widehat{\boldsymbol{b}}$ was used to rank the interactions (off-diagonal entries only) predicted for BEEM-static and the correlations was ranked the same as in **Chapter 4**. The accuracy of interaction signs was calculated as the fraction of interactions with correctly predicted signs in the true interaction matrix $\boldsymbol{b}$ (non-zero off-diagonal entries only).

5.3 BEEM-static accurately and robustly estimates biomass and model parameters on simulated datasets

We generated simulated datasets from gLVMs with different model parameters to evaluate the performance of BEEM-static. In general, BEEM-static estimated parameters have low relative error compared to the ground truth, with <5% in the biomass, <10% in the carrying capacity (a_i) and <20% in the interaction matrix (**Figure 5.1**). The amount of error for BEEM-static estimated parameters decreases and finally reach plateau with increasing number of samples as we expect (**Figure 5.1**), and we observe that the number of samples required for saturated performance (no significant decrease in relative errors) increases linearly with the number of species (**Figure 8.12** and **Table 5-1**). The increase in number of samples is mainly required for accurate inference of the interaction parameters, whose number increases quadratically with the number of species. On the other hand, inference of biomass values was found to benefit

from the increase in number of species, which is equivalent to increasing the number of data points for biomass estimation.

Table 5-1 Approximated number of samples for saturated performance

Number of species	**30**	**40**	**50**	**60**
Number of samples required	500	1000	1500	>2000

Furthermore, we compared the performance of BEEM-static with the correlation based methods benchmarked in **Chapter 4** (**Figure 5.2**). BEEM-static was found to have a median AUC-ROC about 87%, over 25% higher than the correlation based methods in general. Besides, BEEM-static predicted the signs of interactions much more accurately than the correlation based methods.

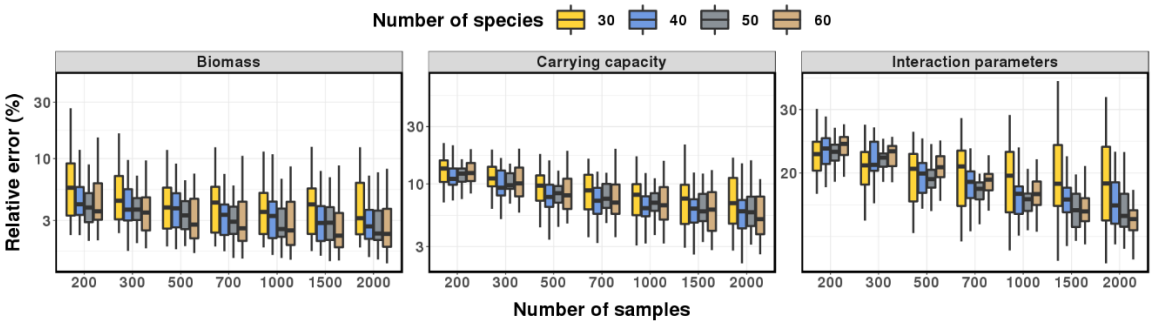

Figure 5.1 Median relative error of BEEM-static with varied number of samples and species with simulated data

Each boxplot represents 30 simulated datasets with different gLVM parameters.

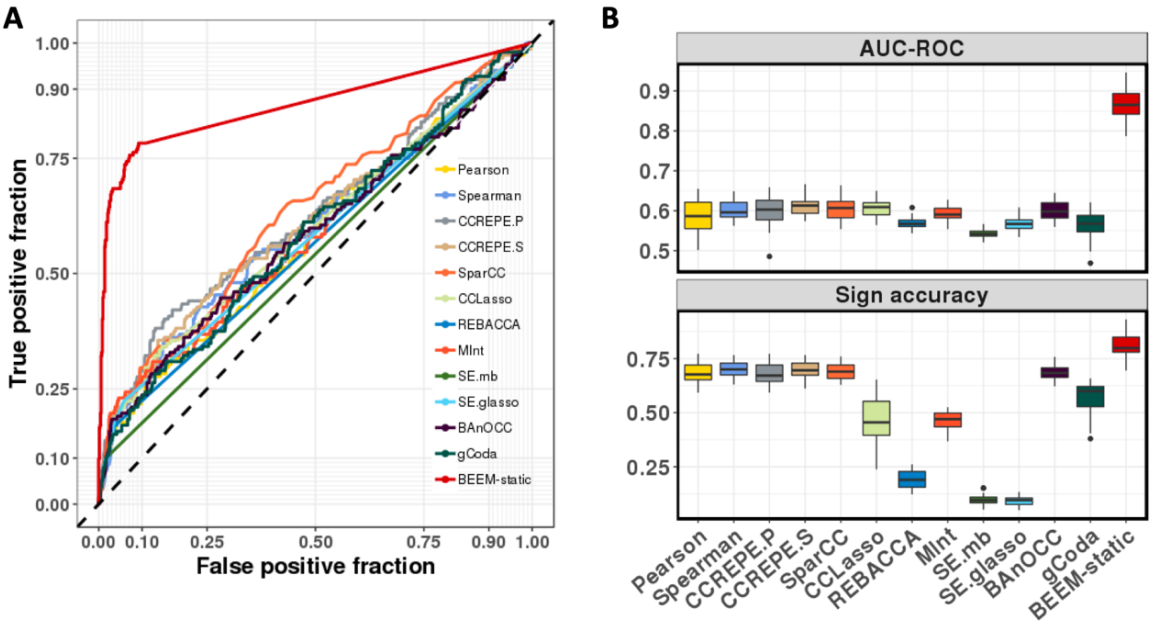

Figure 5.2 Accuracy of inferred interaction network by BEEM-static and correlation based methods

(A) Example receiver operating characteristic (ROC) curves. (B) Boxplots summarizing the area under ROC curve (AUC-ROC) values and the accuracy of inferred interactions signs (30 simulated datasets with 30 species and 500 samples for each boxplot). CCREPE.P: CCREPE corrected Pearson correlation, CCREPE.S: CCREPE corrected Spearman correlation, SE: SPIEC-EASI.

In a real microbiome dataset, the assumption that all samples are at the steady states is not likely valid due to short-term external perturbations (diet change, antibiotics usage etc). On simulated data, we observe that relative errors in all parameter types estimated by BEEM-static increases with larger percent of samples away from the steady states (**Figure 5.3A**). Despite the violation of the "steady state" assumption, BEEM-static notably outperforms correlation based methods (represented by spearman correlation corrected with CCREPE) in terms of predicting the presence of an interaction (higher AUC-ROC in **Figure 5.3B**). However, the presence of samples away from steady states has a much more profound impact on the accuracy of interaction signs predicted (**Figure 5.3B**). To overcome such challenge, we implemented a filter in BEEM-static to automatically remove samples that are detected to have poor fit to equation 1. The filter was found to greatly reduce the error rates and improve the AUC-ROC and sign accuracy of interaction parameters. Strikingly, the AUC-ROC of the interaction matrix predicted was about 80% even when 50% of samples were away from the equilibrium.

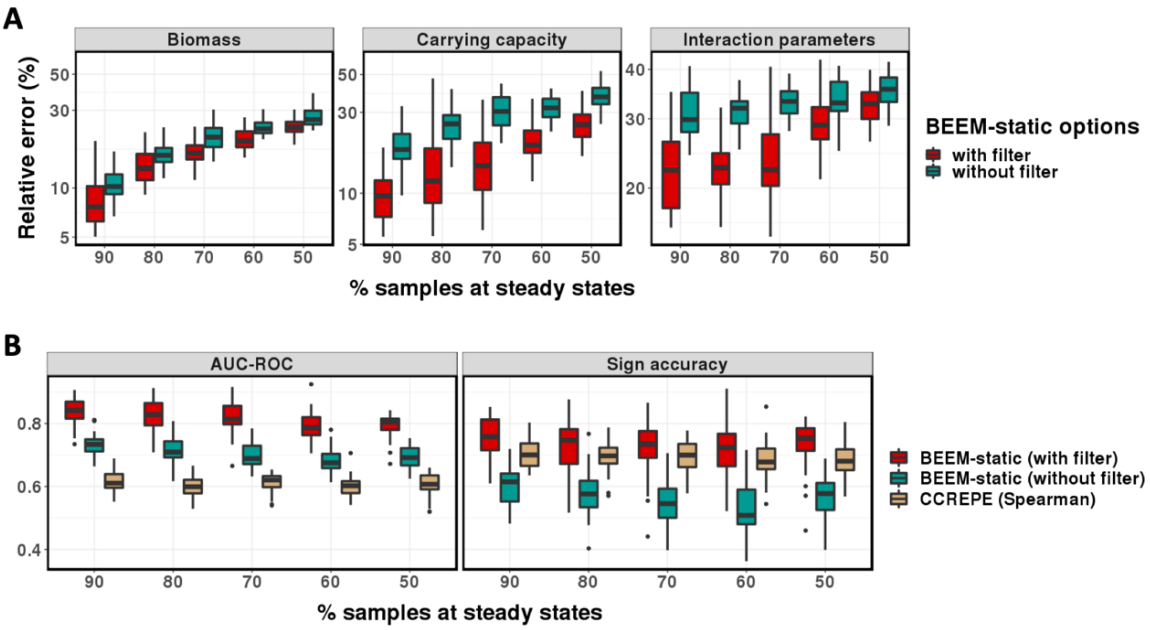

Figure 5.3 Effect of samples not at steady states

(A) The relative errors of BEEM-static with increasing percentage of samples not at steady states. Boxplots show the performance of BEEM-static with and without the filter to remove samples detected as not at steady states. (B) AUC-ROC and the accuracy of signs for the interactions inferred by BEEM-static and Spearman correlation corrected with CCREPE. Each boxplot represents 30 simulated datasets with 30 species and 500 samples at steady states.

5.4 Model learnt by BEEM-static recapitulates known biology of human gut microbiome

To validate the ability of BEEM-static to bring biological insights on real microbiome data, we ran BEEM-static on a large collection of gut microbiome profile of healthy adults as well as selected samples with special biological properties[159]. To assess the ability of BEEM-static for filtering out samples violating the "steady state" assumption, we included the samples from Raymond *et al*[184], where healthy volunteers were under antibiotics treatment. Such treatment is expected to kill most of the bacteria of the normal gut flora, thus greatly perturbing the microbial system away from the steady states. BEEM-static was able to successfully filter the antibiotics treated samples out with a sensitivity of 88% (**Figure 5.4A**).

An important advance of BEEM-static is to computationally infer comparable biomass values across all the samples (off by a single global scaling factor). Previous findings have shown that the total gut bacterial load of newborns younger than 12 months are significantly lower than that of adults[187,188]. Consistently, BEEM-static estimated biomass values for adults were also found to be significantly higher than the newborns (**Figure 5.4B**).

By fitting a dynamic model to the snapshot data, BEEM-static allows us to characterize the *in situ* growth of each species in each sample as the deviation from the steady state (0 in equation 1). The growth rate for a species is a net effect of cell replication rate and cell death rate. With genomic sequencing data, the DNA replication rate for bacteria can be estimated by the coverage of short reads across the genome, also known as the peak-to-trough ratio (PTR) and it is usually used as a proxy for the cell replication rate[189]. We estimated the PTR values for each species in the subjects not

currently taking antibiotics from Raymond *et al* using software GRiD[185], as orthogonal information for *in situ* growth. Species predicted to increase in absolute abundances by BEEM-static were found to have significantly higher PTR values than species predicted to decrease in absolute abundances (**Figure 5.4C**). To showcase the ability to forecast abundance changes, we trained BEEM-static including two time course datasets from Raymond *et al* (control samples not taking antibiotics) and Li *et al*[185] (samples after faecal microbiome transplantation), and compared the abundances change directions predicted by BEEM-static from earlier time points with the absolute abundance differences between adjacent timepoint pairs. Even though no prior information about the temporal links between samples were given to BEEM-static, we found that BEEM-static on average have an accuracy of 62% (Wilcoxon test for > 50% p-value = 5.7×10^{-6}) and some samples with over 80% prediction accuracy (**Figure 5.4D**). Ideally, daily sampled gut microbiome data from Fukuyama *et al*[190] would be a better choice for validating the forecasting accuracy of BEEM-static and is planned to be analysed as our future work.

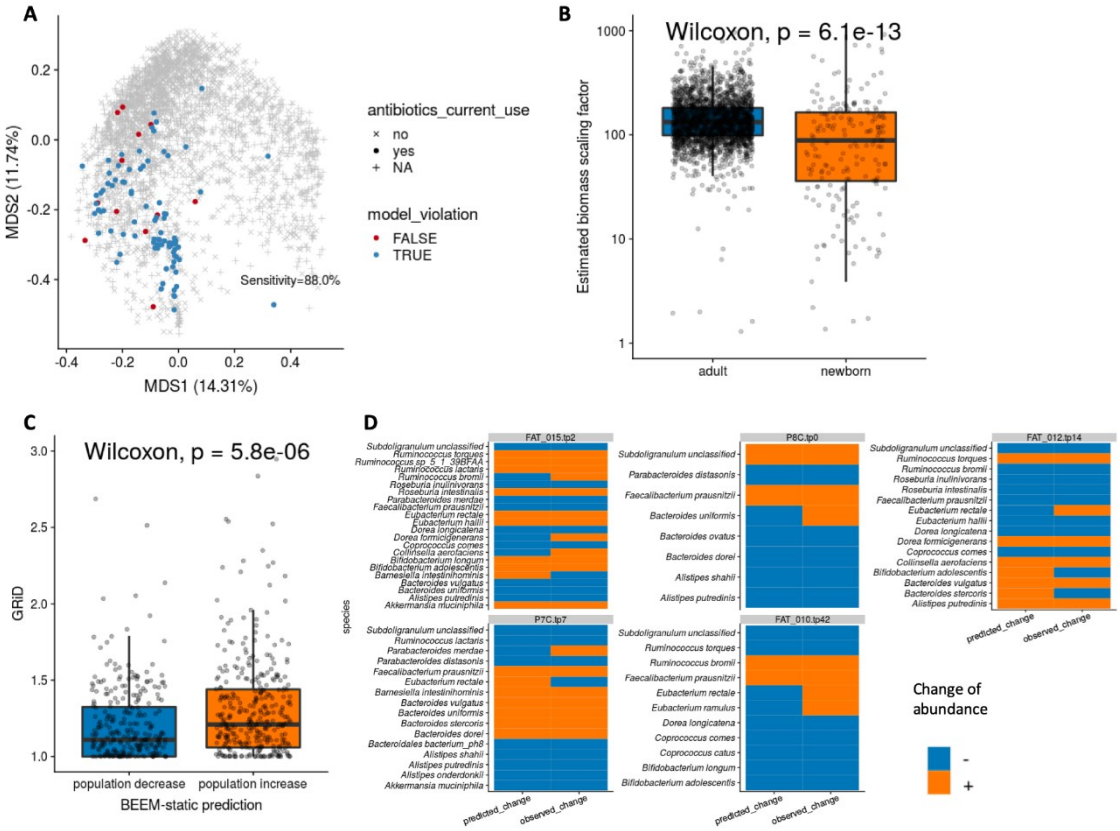

Figure 5.4 Analysis of a large gut microbiome dataset using BEEM-static

(A) Multidimensional scaling (MDS) plot using Bray-Curtis distance. The samples under antibiotics treatment were highlighted and colored by whether BEEM-static detected them as violating the model assumption (blue) or not (red). (B) Biomass scales (scaled to have a median of 100) estimated by BEEM-static for samples from adults and new-born infants. (C) Species *in situ* growth predicted by BEEM-static (increase or decrease) and GRiD (DNA replication rate). (D) Examples of predicting abudance changes in future time points using BEEM-static (samples with prediction accurary > 80% were shown).

6 DISCUSSION

As an important component of microbial community ecology and function, the interactions among microbial members is one of the most fundamental property yet remain poorly understood. The advances in high-throughput sequencing technologies have created new opportunities for the characterization of microbial interactions by generating large amount of microbiome profiling data.

In this thesis, we investigate how to utilise such rich source of data to infer microbial interactions. We leverage the generalized Lotka-Voterra model (gLVM) and propose a novel expectation maximization like framework, BEEM for learning accurate model parameters from "compositional" microbial profiling data by coupling the biomass estimation with the model inference. On longitudinal microbial profiling data, we demonstrated that BEEM significantly outperformed the existing approaches that relied on data scaled using experimentally measured biomass. We applied BEEM to infer biomass and interactions from four densely sampled longitudinal datasets and discovered personalized ecological interaction networks structures driving the distinct dynamics of microbial compositions in different individuals.

Comparing to longitudinal microbiome profiling data, there are currently much larger number of cross-sectional datasets available, which serve as great resource for studying microbial interactions at a population scale. However, the widely used

correlation based approaches for inferring interactions from such data could fail to capture ecological interactions, as we demonstrated with both synthetic and real microbiome data. Motivated by BEEM, we extended its core algorithm and developed BEEM-static to work with cross-sectional data and implemented an R package dedicating to the inference, exploration and visualisation of ecological models. Comparing to the theoretical approach based on ecological models proposed recently[115], BEEM-static provides greatly extended applicability by addressing two major challenges. BEEM-static is able to handle relative abundance data with the EM-like framework, and more importantly does not rely on the assumption that all samples are at the steady states. In addition, BEEM-static also allow us to infer instantaneous growth of each bacterium in a sample based on the deviation of abundance from its steady state.

Moving forward, we plan to develop BEEM-static into a toolbox for learning reusable ecological models from a large collection of cross-sectional data, which can be used to predict the dynamics of newly collected samples as well as their biomass values. In addition, the interactions learned with our two methods, especially BEEM-static provides important species candidates that carry interactions with major community members, and it will be interesting to see if BEEM-static can be used to guide the design of new probiotics manipulate the microbial community structure.

As presented, BEEM and BEEM-static utilize similar EM-like frameworks to solve the parameters although they have different derivations. In fact, we expect that our EM-like framework for handling relative abundance data to work on differential equation models beyond the gLVM. For example, the gLVM can be extended to include parameters for external perturbations (e.g. diet and antibiotics) by adding a set of constants to solve similar to the growth rate term in the model. In addition, it also worth to explore other complex ecological models involving high-order interactions or non-linear terms, although it should be demonstrated that they capture relevant dynamics of

microbial communities beyond the gLVM. In summary, we believe that our novel work makes significant contribution to both biological and computational sides of the field studying microbial function and ecology.

While this thesis mainly focuses on computational approaches for inferring microbial interactions, we appreciate the continuing efforts in investigating microbial interactions using experimental approaches in combination with computational or theoretical methods. For example, exhaustive co-growth experiments at medium scale (~10 species) have been conducted by different groups, systematically investigating different combinations of microbes to characterize their pairwise as well as higher-order interactions[131,191]. These studies not only provided insights into the interactions and assembly of simple microbial communities, but also generated well-controlled datasets for testing and benchmarking existing and newly developed algorithms. In the meantime, the development of artificial gut communities with bioreactors is expected to serve as platforms for generating high-quality datasets and testing interesting hypothesis[192].

7 REFERENCES

1. Djokic, T., Van Kranendonk, M. J., Campbell, K. A., Walter, M. R. & Ward, C. R. Earliest signs of life on land preserved in ca. 3.5 Ga hot spring deposits. *Nat. Commun.* **8**, 15263 (2017).
2. Chimileski, S. & Kolter, R. *Life at the Edge of Sight: A Photographic Exploration of the Microbial World*. (Harvard University Press, 2017).
3. Hesse, W. & Gr, D. H. M. *Walther and Angelina Hesse-Early Contributors to Bacteriology In an unassuming way,. they moved agar from the kitchen to the lab, revolutionizing bacteriology*. (1992).
4. Lok, C. Mining the microbial dark matter. *Nature* **522**, 270–273 (2015).
5. Lane, D. J. *et al.* Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci.* **82**, 6955–6959 (1985).
6. Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* **6**, 805–14 (2005).
7. Schloss, P. D. & Handelsman, J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.* **6**, 229 (2005).
8. Thomas, T., Gilbert, J. & Meyer, F. Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp.* **2**, 3 (2012).
9. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R245-9 (1998).
10. Blaser, M., Bork, P., Fraser, C., Knight, R. & Wang, J. The microbiome explored: recent insights and future challenges. *Nat. Rev. Microbiol.* **11**, 213–7 (2013).
11. Li, H. Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. *Annu. Rev. Stat. Its Appl.* **2**, 73–94 (2015).
12. Aitchison, J. The statistical analysis of compositional data. (1986).
13. Flemming, H.-C. & Wuertz, S. Bacteria and archaea on Earth and their abundance in biofilms. *Nat. Rev. Microbiol.* **17**, 247–260 (2019).
14. Faust, K. *et al.* Microbial co-occurrence relationships in the Human Microbiome. *PLoS Comput. Biol.* **8**, (2012).
15. Borenstein, E., Kupiec, M., Feldman, M. W. & Ruppin, E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 14482–14487 (2008).
16. Freilich, S. *et al.* The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Res.* **38**, 3857–3868 (2010).
17. Asahara, T. *et al.* Probiotic Bifidobacteria Protect Mice from Lethal Infection with

Shiga Toxin-Producing Escherichia coli O157:H7. *Infect. Immun.* **72**, 2240–2247 (2004).
18. Servin, A. L. & Coconnier, M.-H. Adhesion of probiotic strains to the intestinal mucosa and interaction with pathogens. *Best Pract. Res. Clin. Gastroenterol.* **17**, 741–754 (2003).
19. Brune, K. D. & Bayer, T. S. Engineering microbial consortia to enhance biomining and bioremediation. *Front. Microbiol.* **3**, 203 (2012).
20. Koch, C., Müller, S., Harms, H. & Harnisch, F. Microbiomes in bioenergy production: From analysis to management. *Curr. Opin. Biotechnol.* **27**, 65–72 (2014).
21. Brune, A. Symbiotic digestion of lignocellulose in termite guts. *Nat. Rev. Microbiol.* **12**, 168–80 (2014).
22. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10**, 538–50 (2012).
23. Zeidan, A. A., Rådström, P. & van Niel, E. W. J. Stable coexistence of two Caldicellulosiruptor species in a de novo constructed hydrogen-producing co-culture. *Microb. Cell Fact.* **9**, 102 (2010).
24. Kato, S., Haruta, S., Cui, Z. J., Ishii, M. & Igarashi, Y. Stable coexistence of five bacterial strains as a cellulose-degrading community. *Appl. Environ. Microbiol.* **71**, 7099–106 (2005).
25. Harcombe, W. Novel cooperation experimentally evolved between species. *Evolution* **64**, 2166–72 (2010).
26. Freilich, S. *et al.* Competitive and cooperative metabolic interactions in bacterial communities. *Nat. Commun.* **2**, 589 (2011).
27. Karlsson, F. H., Nookaew, I. & Nielsen, J. Metagenomic Data Utilization and Analysis (MEDUSA) and Construction of a Global Gut Microbial Gene Catalogue. *PLoS Comput. Biol.* **10**, (2014).
28. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
29. Shertzer, K. W., Ellner, S. P., Fussmann, G. F. & Hairston, N. G. Predator-prey cycles in an aquatic microcosm: Testing hypotheses of mechanism. *J. Anim. Ecol.* **71**, 802–815 (2002).
30. Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* **20**, 947–59 (2010).
31. Levy, R. & Borenstein, E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12804–9 (2013).
32. Zhou, J. *et al.* Functional molecular ecological networks. *MBio* **1**, e00169-10- (2010).
33. Maruyama, N. *et al.* Intraindividual variation in core microbiota in peri-implantitis and periodontitis. *Sci. Rep.* **4**, 6602 (2014).
34. Ju, F., Xia, Y., Guo, F., Wang, Z. & Zhang, T. Taxonomic relatedness shapes bacterial assembly in activated sludge of globally distributed wastewater treatment plants. *Environ. Microbiol.* **16**, 2421–2432 (2014).
35. Peng, X., Guo, F., Ju, F. & Zhang, T. Shifts in the microbial community, nitrifiers and denitrifiers in the biofilm in a full-scale rotating biological contactor. *Environ. Sci. Technol.* **48**, 8044–8052 (2014).
36. Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front. Microbiol.* **5**, 1–14 (2014).
37. Barberán, A., Bates, S. T., Casamayor, E. O. & Fierer, N. Using network analysis

to explore co-occurrence patterns in soil microbial communities. *ISME J.* **6**, 343–51 (2012).
38. Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. *PLoS Comput. Biol.* **8**, 1–11 (2012).
39. Aitchison, J. *The statistical analysis of compositional data*. (Chapman & Hall, Ltd., 1986).
40. Scealy,J.L. and Welsh,A.H. (2014) Colours and Cocktails: Compositional Data Analysis 2013 Lancaster Lecture. Aust. N. Z. J. Stat., 56, 145–169.
41. Ban, Y., An, L. & Jiang, H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics* btv364- (2015). doi:10.1093/bioinformatics/btv364
42. Fang, H., Huang, C., Zhao, H. & Deng, M. CCLasso: Correlation Inference for Compositional Data through Lasso. *Bioinformatics* btv349- (2015). doi:10.1093/bioinformatics/btv349
43. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507 (2017).
44. Oliver, D. S. Calculation of the inverse of the covariance. *Math. Geol.* **30**, 911–933 (1998).
45. Meinshausen, N. & Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* **34**, 1436–1462 (2006).
46. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–41 (2008).
47. Biswas, S., McDonald, M., Lundberg, D. S., Dangl, J. L. & Jojic, V. Learning microbial interaction networks from metagenomic count data. (2014).
48. Kurtz, Z. D. *et al.* Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
49. Feizi, S., Marbach, D., Médard, M. & Kellis, M. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.* **31**, 726–33 (2013).
50. Lee, M. S., Oh, S. & Tang, H. Characterization of microbial associations in human oral microbiome. *Biomed. Mater. Eng.* **24**, 3737–44 (2014).
51. Lima-Mendez, G. *et al.* Ocean plankton. Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).
52. Silverman, J. D., Roche, K., Mukherjee, S. & David, L. A. Naught all zeros in sequence count data are the same. *bioRxiv* 477794 (2018). doi: 10.1101/477794.
53. Dwivedi, A. K., Rao, M. B., Dwivedi, S. N., Deo, S. V. . & Shukla, R. On Classifying At Risk Latent Zeros Using Zero Inflated Models All India Institute of Medical Sciences University of Cincinnati. *J. Data Sci.* **12**, 307–323 (2014).
54. *Numerical Ecology*. *Developments in Environmental Modelling* **24**, (Elsevier, 2012).
55. Butte, A. J. & Kohane, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* 418–29 (2000).
56. Villaverde, A. F., Ross, J., Morán, F. & Banga, J. R. MIDER: network inference with mutual information distance and entropy reduction. *PLoS One* **9**, e96732 (2014).
57. Margolin, A. A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1**, S7 (2006).
58. Reshef, D. N. *et al.* Detecting novel associations in large data sets. *Science* **334**, 1518–24 (2011).

59. Wang, H., Masters, S., Edwards, M. a., Falkinham, J. O. & Pruden, A. Effect of disinfectant, water age, and pipe materials on bacterial and eukaryotic community structure in drinking water biofilm. *Environ. Sci. Technol.* **48**, 1426–1435 (2014).
60. Ren, T. *et al.* 16S rRNA survey revealed complex bacterial communities and evidence of bacterial interference on human adenoids. *Environ. Microbiol.* **15**, 535–547 (2013).
61. Kinney, J. B. & Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci.* **111**, 3354–3359 (2014).
62. Simon, N. & Tibshirani, R. Comment on 'Detecting Novel Associations In Large Data Sets' by Reshef Et Al, Science Dec 16, 2011. *arXiv* (2014).
63. Steuer,R. et al. (2002) The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18, S231–S240.
64. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl**, 4554–61 (2011).
65. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–63 (2014).
66. Lozupone, C. *et al.* Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. *Genome Res.* **22**, 1974–84 (2012).
67. David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15**, R89 (2014).
68. Benincà, E. *et al.* Chaos in a long-term experiment with a plankton community. *Nature* **451**, 822–5 (2008).
69. Gajer, P. *et al.* Temporal dynamics of the human vaginal microbiota. *Sci. Transl. Med.* **4**, 132ra52 (2012).
70. Chow, C.-E. T., Kim, D. Y., Sachdeva, R., Caron, D. A. & Fuhrman, J. A. Top-down controls on bacterial community structure: microbial network analysis of bacteria, T4-like viruses and protists. *ISME J.* **8**, 816–29 (2014).
71. Xia, L. C. *et al.* Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst. Biol.* **5 Suppl 2**, S15 (2011).
72. Xia, L. C., Ai, D., Cram, J., Fuhrman, J. a. & Sun, F. Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics* **29**, 230–237 (2013).
73. Ruan, Q. *et al.* Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* **22**, 2532–8 (2006).
74. Buffie, C. G. *et al.* Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature* **517**, 205–208 (2014).
75. Fisher, C. K. & Mehta, P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One* **9**, e102451 (2014).
76. 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).
77. Faust, K., Lahti, L., Gonze, D., de Vos, W. M. & Raes, J. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* **25**, 56–66 (2015).
78. Levy, R. & Borenstein, E. Reverse Ecology: from systems to environments and back. *Adv. Exp. Med. Biol.* **751**, 329–45 (2012).

79. Manor, O., Levy, R. & Borenstein, E. Mapping the inner workings of the microbiome: Genomic- and metagenomic-based study of metabolism and of metabolic interactions in the human gut microbiome. *Cell Metab.* **20**, 742–752 (2014).
80. Carr, R. & Borenstein, E. NetSeed: a network-based reverse-ecology tool for calculating the metabolic interface of an organism with its environment. *Bioinformatics* **28**, 734–5 (2012).
81. Kreimer, A., Doron-Faigenboim, A., Borenstein, E. & Freilich, S. NetCmpt: a network-based tool for calculating the metabolic competition between bacterial species. *Bioinformatics* **28**, 2195–7 (2012).
82. Levy, R., Carr, R., Kreimer, A., Freilich, S. & Borenstein, E. NetCooperate: a network-based tool for inferring host-microbe and microbe-microbe cooperation. *BMC Bioinformatics* **16**, 1–6 (2015).
83. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–8 (2010).
84. Stolyar, S. *et al.* Metabolic modeling of a mutualistic microbial community. *Mol. Syst. Biol.* **3**, 92 (2007).
85. Shoaie, S. *et al.* Understanding the interactions between bacteria in the human gut through metabolic modeling. *Sci. Rep.* **3**, 2532 (2013).
86. Zomorrodi, A. R. & Maranas, C. D. OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput. Biol.* **8**, e1002363 (2012).
87. El-Semman, I. E. *et al.* Genome-scale metabolic reconstructions of *Bifidobacterium adolescentis* L2-32 and *Faecalibacterium prausnitzii* A2-165 and their interaction. *BMC Syst. Biol.* **8**, 41 (2014).
88. Zomorrodi, A. R., Islam, M. M. & Maranas, C. D. d-OptCom: Dynamic multi-level and multi-objective metabolic modeling of microbial communities. *ACS Synth. Biol.* **3**, 247–57 (2014).
89. Harcombe, W. R. *et al.* Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep.* **7**, 1104–15 (2014).
90. Zelezniak, A. *et al.* Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc. Natl. Acad. Sci.* **112**, 201421834 (2015).
91. Bernet-Camard, M. F. *et al.* The human Lactobacillus acidophilus strain LA1 secretes a nonbacteriocin antibacterial subtance(s) active in vitro and in vivo. *Appl. Environ. Microbiol.* **63**, 2747–2753 (1997).
92. Forestier, C., De Champs, C., Vatoux, C. & Joly, B. Probiotic activities of *Lactobacillus casei rhamnosus*: *in vitro* adherence to intestinal cells and antimicrobial properties. *Res. Microbiol.* **152**, 167–173 (2001).
93. Mack, D. R., Michail, S., Wei, S., McDougall, L. & Hollingsworth, M. a. Probiotics inhibit enteropathogenic E. coli adherence in vitro by inducing intestinal mucin gene expression. *Am. J. Physiol.* **276**, G941–G950 (1999).
94. Chapman, C. M. C., Gibson, G. R. & Rowland, I. In vitro evaluation of single- and multi-strain probiotics: Inter-species inhibition between probiotic strains, and inhibition of pathogens. *Anaerobe* **18**, 405–413 (2012).
95. Klaenhammer, T. R. Bacteriocins of lactic acid bacteria. *Biochimie* **70**, 337–349 (1988).
96. Flynn, S. *et al.* Characterization of the genetic locus responsible for the production of ABP-118, a novel bacteriocin produced by the probiotic bacterium Lactobacillus salivarius subsp. salivarius UCC118. *Microbiology* **148**, 973–84 (2002).

97. Thomas, A. a. *et al.* Extracting data from electronic medical records: Validation of a natural language processing program to assess prostate biopsy results. *World J. Urol.* **32**, 99–103 (2014).
98. Thorn, C. F., Klein, T. E. & Altman, R. B. Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics* **11**, 501–505 (2010).
99. Cohen, A. M. & Hersh, W. R. A survey of current work in biomedical text mining. *Br. Bioinform* **6**, 57–71 (2005).
100. Dobrokhotov, P. B., Goutte, C., Veuthey, A. L. & Gaussier, E. Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation. *Bioinformatics* **19**, 91–94 (2003).
101. Chowdhary, R., Zhang, J. & Liu, J. S. Bayesian inference of protein-protein interactions from biological literature. *Bioinformatics* **25**, 1536–1542 (2009).
102. Temkin, J. M. & Gilder, M. R. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics* **19**, 2046–2053 (2003).
103. Tikk, D., Thomas, P., Palaga, P., Hakenberg, J. & Leser, U. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput. Biol.* **6**, e1000837 (2010).
104. Quan, C., Wang, M. & Ren, F. An unsupervised text mining method for relation extraction from biomedical literature. *PLoS One* **9**, 1–8 (2014).
105. Blaschke, C., Andrade, M. A., Ouzounis, C. & Al. Automatic extraction of biological information from scientific text: protein-protein interactions. *Interactions* (1999).
106. Kim, J. D., Ohta, T., Tateisi, Y. & Tsujii, J. GENIA corpus - A semantically annotated corpus for bio-textmining. *Bioinformatics* **19**, 180–182 (2003).
107. Pyysalo, S. *et al.* BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics* **8**, 50 (2007).
108. Fundel, K., Küffner, R. & Zimmer, R. RelEx - Relation extraction using dependency parse trees. *Bioinformatics* **23**, 365–371 (2007).
109. Lapage, S. *et al. International Code of Nomenclature of Bacteria*. (ASM Press, 1992).
110. Kaper, J. B. & Sperandio, V. Bacterial cell-to-cell signaling in the gastrointestinal tract. *Infect. Immun.* **73**, 3197–209 (2005).
111. Amin, S. A. *et al.* Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* **522**, 98–101 (2015).
112. Rutherford, S. T. & Bassler, B. L. Bacterial quorum sensing: its role in virulence and possibilities for its control. *Cold Spring Harb. Perspect. Med.* **2**, a012427-(2012).
113. Ben-Jacob, E. *et al.* Bacterial cooperative organization under antibiotic stress. *Phys. A Stat. Mech. its Appl.* **282**, 247–282 (2000).
114. Narisawa, N., Haruta, S., Arai, H., Ishii, M. & Igarashi, Y. Coexistence of antibiotic-producing and antibiotic-sensitive bacteria in biofilms is mediated by resistant bacteria. *Appl. Environ. Microbiol.* **74**, 3887–94 (2008).
115. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
116. Xiao, Y. *et al.* Mapping the ecological networks of microbial communities. *Nat. Commun.* **8**, 2042 (2017).
117. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biol.* **12**, 69 (2014).
118. Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449**, 804–10 (2007).

119. Hayat, R., Ali, S., Amara, U., Khalid, R. & Ahmed, I. Soil beneficial bacteria and their role in plant growth promotion: a review. *Ann. Microbiol.* **60**, 579–598 (2010).
120. Halfvarson, J. *et al.* Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.* **2**, 17004 (2017).
121. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).
122. Li, Q., Han, Y., Dy, A. B. C. & Hagerman, R. J. The Gut Microbiota and Autism Spectrum Disorders. *Front. Cell. Neurosci.* **11**, 120 (2017).
123. Chng, K. R. *et al.* Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. *Nat. Microbiol.* **1**, 16106 (2016).
124. Hol, F. J. H., Rotem, O., Jurkevitch, E., Dekker, C. & Koster, D. A. Bacterial predator-prey dynamics in microscale patchy landscapes. *Proceedings. Biol. Sci.* **283**, 20152154 (2016).
125. Miller, M. B. & Bassler, B. L. Quorum Sensing in Bacteria. *Annu. Rev. Microbiol.* **55**, 165–199 (2001).
126. Martin, M., Hölscher, T., Dragoš, A., Cooper, V. S. & Kovács, Á. T. Laboratory Evolution of Microbial Interactions in Bacterial Biofilms. *J. Bacteriol.* **198**, 2564–71 (2016).
127. Embree, M., Liu, J. K., Al-Bassam, M. M. & Zengler, K. Networks of energetic and metabolic interactions define dynamics in microbial communities. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15450–5 (2015).
128. Cordero, O. X. & Datta, M. S. Microbial interactions and community assembly at microscales. *Curr. Opin. Microbiol.* **31**, 227–234 (2016).
129. Fraune, S. *et al.* Bacteria–bacteria interactions within the microbiota of the ancestral metazoan Hydra contribute to fungal resistance. *ISME J.* **9**, 1543–1556 (2015).
130. Lim, K. M. K., Li, C., Chng, K. R. & Nagarajan, N. @MInter: Automated Text-mining of Microbial Interactions. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw357
131. Friedman, J., Higgins, L. M. & Gore, J. Community structure follows simple assembly rules in microbial microcosms. *Nat. Ecol. Evol.* **1**, 0109 (2017).
132. Blasche, S., Kim, Y., Oliveira, A. P. & Patil, K. R. Model microbial communities for ecosystems biology. *Curr. Opin. Syst. Biol.* **6**, 51–57 (2017).
133. Li, C., Kenneth, L. K. M., Chng, K. R. & Nagarajan, N. Predicting Microbial Interactions through Computational Approaches. *Methods* (2016). doi:10.1016/j.ymeth.2016.02.019
134. Faust, K. *et al.* Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Comput. Biol.* **8**, e1002606 (2012).
135. Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
136. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
137. Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front. Microbiol.* **5**, 219 (2014).
138. Maoz, A., Mayr, R. & Scherer, S. Temporal stability and biodiversity of two complex antilisterial cheese-ripening microbial consortia. *Appl. Environ. Microbiol.* **69**, 4012–8 (2003).
139. Mounier, J. *et al.* Microbial interactions within a cheese microbial community. *Appl. Environ. Microbiol.* **74**, 172–81 (2008).

140. Bucci, V. *et al.* MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biol.* **17**, 121 (2016).
141. Cao, H.-T., Gibson, T. E., Bashan, A. & Liu, Y.-Y. Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons. *BioEssays* **39**, 1600188 (2017).
142. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science* **350**, 663–6 (2015).
143. Props, R., Monsieurs, P., Mysara, M., Clement, L. & Boon, N. Measuring the biodiversity of microbial communities by flow cytometry. *Methods Ecol. Evol.* **7**, 1376–1385 (2016).
144. Props, R. *et al.* Absolute quantification of microbial taxon abundances. *ISME J.* **11**, 584–587 (2017).
145. Smith, C. J., Nedwell, D. B., Dong, L. F. & Osborn, A. M. Evaluation of quantitative polymerase chain reaction-based approaches for determining gene copy and gene transcript numbers in environmental samples. *Environ. Microbiol.* **8**, 804–815 (2006).
146. White, R. A., Blainey, P. C., Fan, H. C. & Quake, S. R. Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics* **10**, 116 (2009).
147. Sze, M. A., Abbasi, M., Hogg, J. C. & Sin, D. D. A Comparison between Droplet Digital and Quantitative PCR in the Analysis of Bacterial 16S Load in Lung Tissue Samples from Control and COPD GOLD 2. *PLoS One* **9**, e110351 (2014).
148. Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. K. & Schmidt, T. M. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res.* **43**, D593–D598 (2015).
149. Do, C. B. & Batzoglou, S. What is the expectation maximization algorithm? *Nat. Biotechnol.* **26**, 897–899 (2008).
150. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol.* **12**, R50 (2011).
151. David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15**, R89 (2014).
152. Gibbons, S. M., Kearney, S. M., Smillie, C. S. & Alm, E. J. Two dynamic regimes in the human gut microbiome. *PLOS Comput. Biol.* **13**, e1005364 (2017).
153. Gramacy, R. B. monomvn: Estimation for Multivariate Normal and Student-t Data with Monotone Missingness. (2017).
154. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200–1202 (2013).
155. Ramsey, J. & Ripley, B. pspline: Penalized Smoothing Splines. (2017).
156. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–17 (2008).
157. Faust, K. *et al.* Signatures of ecological processes in microbial community time series. *Microbiome* **6**, 120 (2018).
158. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
159. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
160. Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The impact of the gut microbiota on human health: An integrative view. *Cell* **148**, 1258–1270 (2012).
161. Munukka, E. *et al. Faecalibacterium prausnitzii* treatment improves hepatic

health and reduces adipose tissue inflammation in high-fat fed mice. *ISME J.* **11**, 1667–1679 (2017).
162. Gauffin Cano, P., Santacruz, A., Moya, Á. & Sanz, Y. *Bacteroides uniformis* CECT 7771 Ameliorates Metabolic and Immunological Dysfunction in Mice with High-Fat-Diet Induced Obesity. *PLoS One* **7**, e41079 (2012).
163. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14691–6 (2010).
164. Azcarate-Peril, M. A. *et al.* Impact of short-chain galactooligosaccharides on the gut microbiome of lactose-intolerant individuals. *Proc. Natl. Acad. Sci.* **114**, E367–E375 (2017).
165. Candela, M. *et al.* Unbalance of intestinal microbiota in atopic children. *BMC Microbiol.* **12**, 95 (2012).
166. Wexler, H. M. Bacteroides: the good, the bad, and the nitty-gritty. *Clin. Microbiol. Rev.* **20**, 593–621 (2007).
167. Derrien, M. & van Hylckama Vlieg, J. E. T. Fate, activity, and impact of ingested bacteria within the human gut microbiota. *Trends Microbiol.* **23**, 354–366 (2015).
168. Zhang, C. *et al.* Ecological robustness of the gut microbiota in response to ingestion of transient food-borne microbes. *ISME J.* **10**, 2235–45 (2016).
169. Agler, M. T. *et al.* Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation. *PLOS Biol.* **14**, e1002352 (2016).
170. Ridenhour, B. J. *et al.* Modeling time-series data from microbial communities. *ISME J.* **11**, 2526–2537 (2017).
171. Ay, A. & Arnosti, D. N. Mathematical modeling of gene expression: a guide for the perplexed biologist. *Crit. Rev. Biochem. Mol. Biol.* **46**, 137–51 (2011).
172. Smits, S. A. *et al.* Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science* **357**, 802–806 (2017).
173. Ren, T. *et al.* Seasonal, spatial, and maternal effects on gut microbiome in wild red squirrels. *Microbiome* **5**, 163 (2017).
174. Ashida, H., Ogawa, M., Kim, M., Mimuro, H. & Sasakawa, C. Bacteria and host interactions in the gut epithelial barrier. *Nat. Chem. Biol.* **8**, 36–45 (2012).
175. Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* **10**, 1669–81 (2016).
176. Schwager, E., Mallick, H., Ventz, S. & Huttenhower, C. A Bayesian method for detecting pairwise associations in compositional data. *PLOS Comput. Biol.* **13**, e1005852 (2017).
177. Fang, H., Huang, C., Zhao, H. & Deng, M. gCoda: Conditional Dependence Network Inference for Compositional Data. *J. Comput. Biol.* **24**, 699–708 (2017).
178. McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3**, (2018).
179. Bashan, A. *et al.* Universality of human microbial dynamics. *Nature* **534**, 259–262 (2016).
180. Vano, J. A., Wildenberg, J. C., Anderson, M. B., Noel, J. K. & Sprott, J. C. Chaos in low-dimensional Lotka-Volterra models of competition. *Nonlinearity* **19**, 2391–2404 (2006).
181. Marino, S., Baxter, N. T., Huffnagle, G. B., Petrosino, J. F. & Schloss, P. D. Mathematical modeling of primary succession of murine intestinal microbiota. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 439–44 (2014).
182. Buffie, C. G. *et al.* Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature* **517**, 205–208 (2014).
183. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized

Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).
184. Raymond, F. *et al.* The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J.* **10**, 707–720 (2016).
185. Li, S. S. *et al.* Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science.* **352**, 586–589 (2016).
186. Emiola, A. & Oh, J. High throughput in situ metagenomic measurement of bacterial replication at ultra-low sequencing coverage. *Nat. Commun.* **9**, 4956 (2018).
187. Hong, C. *et al.* PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* **2**, 33 (2014).
188. Tsuji, H., Matsuda, K. & Nomoto, K. Counting the Countless: Bacterial Quantification by Targeting rRNA Molecules to Explore the Human Gut Microbiota in Health and Disease. *Front. Microbiol.* **9**, 1417 (2018).
189. Hopkins, M. J., Macfarlane, G. T., Furrie, E., Fite, A. & Macfarlane, S. Characterisation of intestinal bacteria in infant stools using real-time PCR and northern hybridisation analyses. *FEMS Microbiol. Ecol.* **54**, 77–85 (2005).
190. Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science.* **349**, 1101–1106 (2015).
191. Fukuyama, J. *et al.* Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLOS Comput. Biol.* **13**, e1005706 (2017).
192. Venturelli, O. S. *et al.* Deciphering microbial interactions in synthetic human gut microbiome communities. *Mol. Syst. Biol.* **14**, e8157 (2018).
193. Silverman, J. D., Durand, H., Bloom, R. J., Mukherjee, S. & David, L. A. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *bioRxiv* 306597 (2018). doi:10.1101/306597

8 APPENDICES

APPENDIX 1 SUPPLEMENTARY TABLES & FIGURES FOR CHAPTER 2

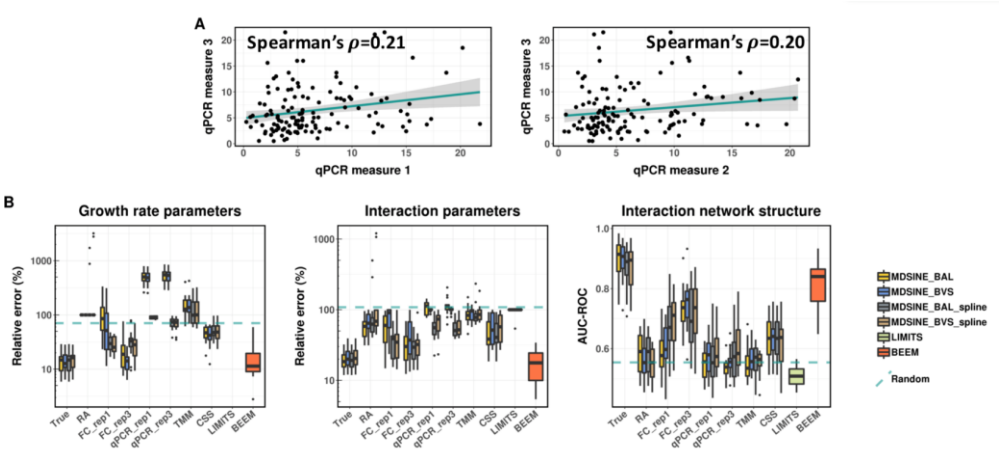

Figure 8.1. Noise in experimentally determined biomass severely distorts gLVM parameter estimation.

(A) Scatter plots with fitted linear regression lines for three 16S qPCR technical replicates from Bucci et al. (B) Relative impact of different gLVM parameter estimation algorithms (BAL, BVS, BAL_spline and BVS_spline as implemented in MDSINE) and data scaling approaches. Boxplots represent the summary of 15 simulations (10 species, 30 replicates with 30 time points each). Note that, in general, different data scaling approaches were found to impact performance more than the different estimation algorithms. Dashed horizontal lines represent the performance of randomly generated parameters from the simulation model. In general, scaling with true factors and using BEEM provided notably good results, and among competing experimental (FC_rep1, FC_rep3, qPCR_rep1, qPCR_rep3) and computational approaches (RA – relative abundance, CSS – CSS normalization, TMM – TMM normalization, LIMITS), having three replicates from flow cytometry was the closest (FC_rep3). Note that LIMITS does not compute growth rate parameters.

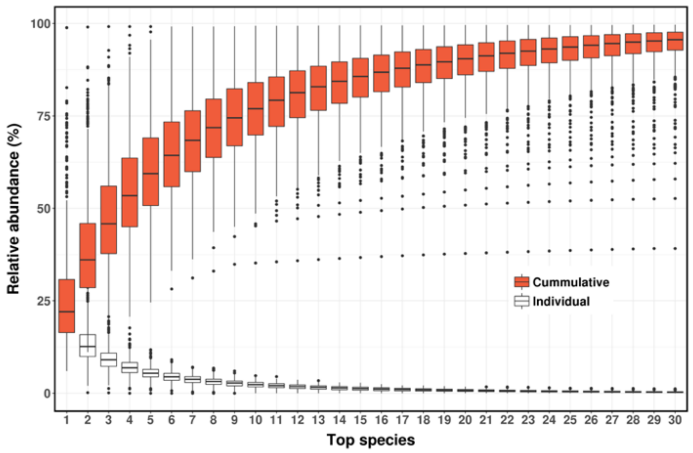

Figure 8.2. Relative abundance distribution of gut microbiome

Relative abundances observed for the most abundant species in 840 normal stool metagenomic samples

from Pasolli et al. Filled boxplots show cumulative values.

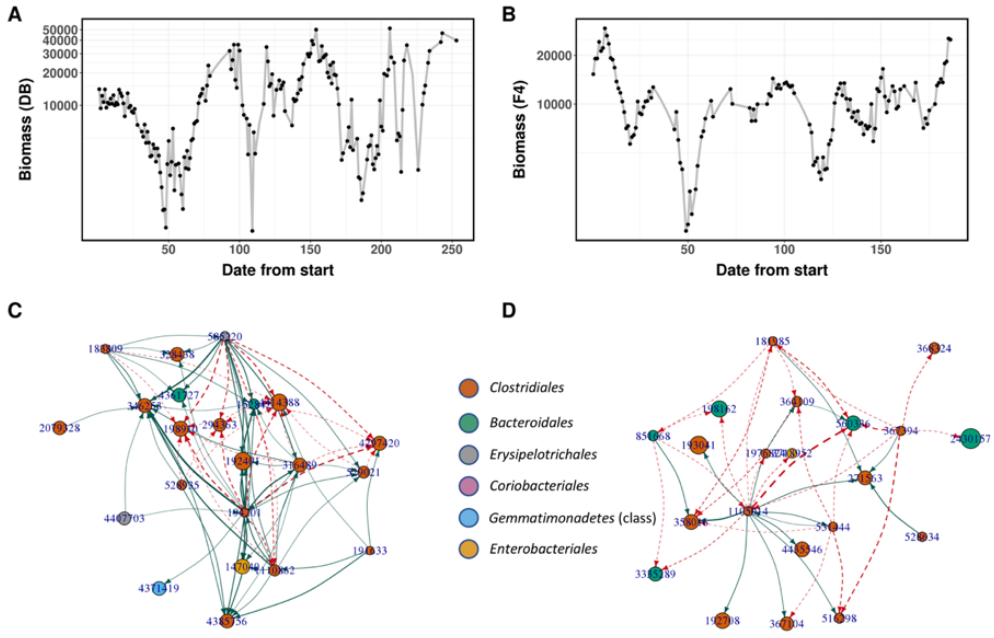

Figure 8.3 BEEM estimated biomass and interaction networks.

BEEM estimated biomass (A and B; scaled to median of 104) and interaction networks (C and D) from the two shorter gut microbial longitudinal profiles from David et al and Caporaso et al. Dashed and solid edges represent positive and negative interactions, respectively, in the networks. Edge widths are proportional to the interaction strength and node sizes are proportional to log-transformed mean relative abundance of the corresponding species (OTU). Nodes are labeled with GreenGenes IDs and colored according to order level of taxonomic annotation.

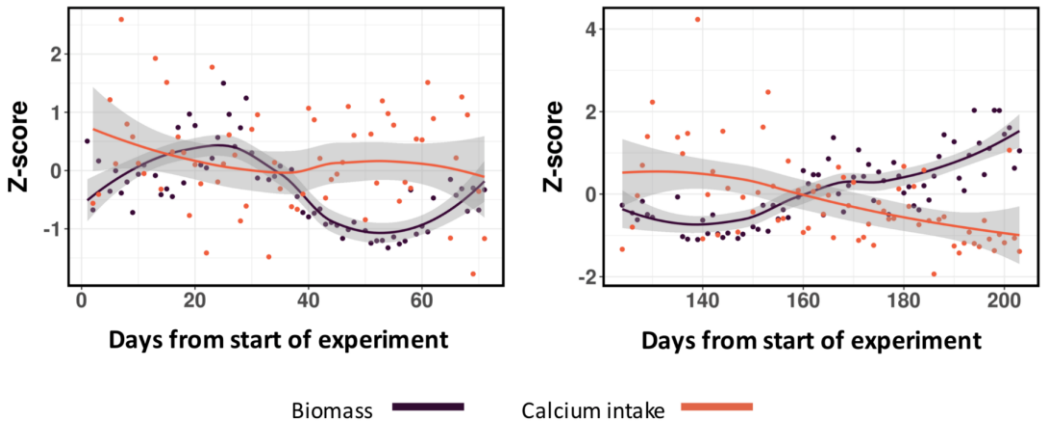

Figure 8.4 Association between biomass and dietary data

Changes in calcium intake (z-score normalized) for the preceding day (orange) in relation to BEEM-estimated biomass (z-score normalized) for subject DA's gut microbiome (purple; only time points with calcium intake data are shown). Lines represent loess smoothers and shaded regions depict 95% confidence intervals. Overall, the two variables were found to be significantly correlated (Spearman's ρ=-0.40, p-value=1.7×10-6)

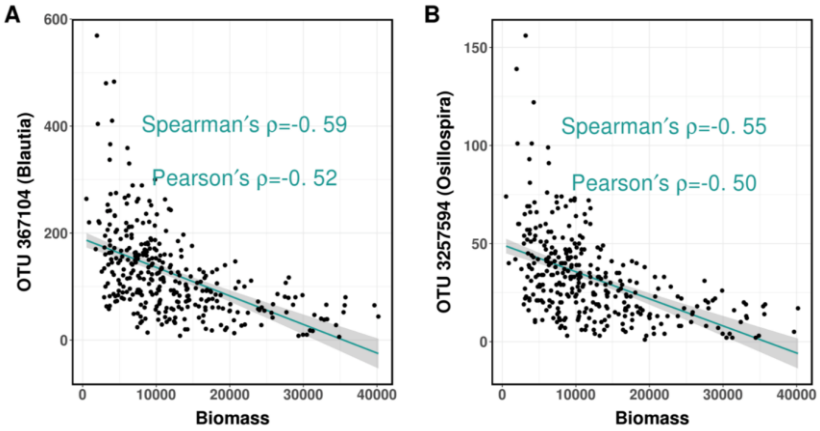

Figure 8.5 Biomass associated OTUs

Scatter plots with fitted linear regression lines between the two hub OTUs and the estimated biomass of M3's gut microbiome. All correlations are significant (p-value<2.2x10-16).

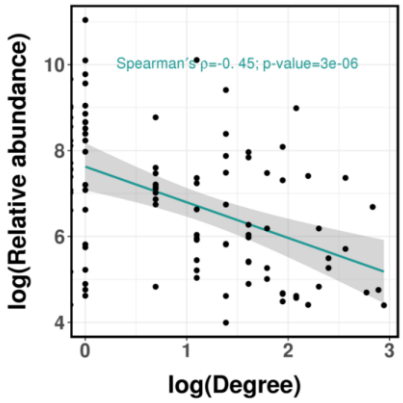

Figure 8.6 Association between hubness and relative abundance

Scatter plot with fitted linear regression line between the out- and in- degree of the OTU versus its mean relative abundance on log scale (based on networks for all 4 subjects).

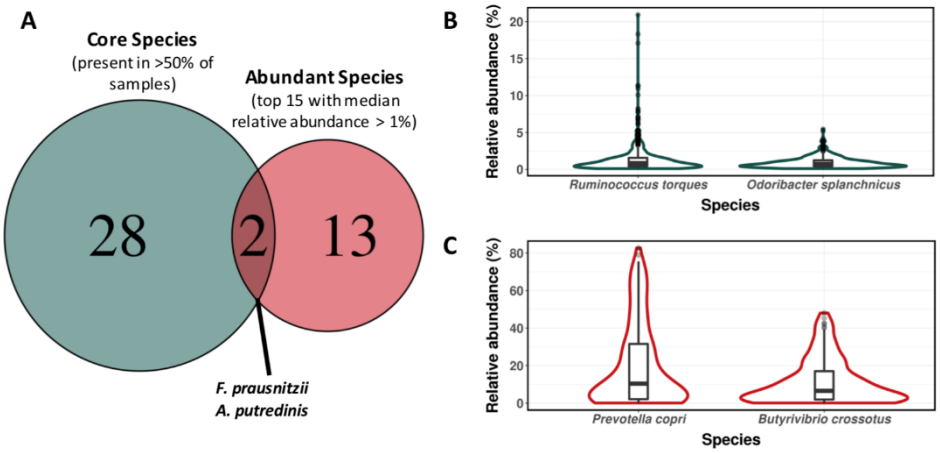

Figure 8.7 Core vs. abundant species

(A) Venn diagram for core species (present in >50% of samples) and abundant species (top 15 with median relative abundance > 1%) in healthy human gut microbiomes from Pasolli et al (N=840). (B) Examples of core gut species with low relative abundances. R. torques and O. splanchnicus were present in 95% and 69% of samples but both of them rarely have relative abundance >1%. (C) Examples of more abundant species that are only found in a fraction of individuals. P. copri and B. crossotus are frequently present at abundances >5%, but are only present in 34% and 15% of all samples, respectively.

APPENDIX 2 SUPPLEMENTARY TABLES & FIGURES FOR CHAPTER 3

Table 8-1 Summary of different variables in parametric modelling (n – number of species).

Feature	Control Condition	Test Conditions
Number of Species	50	10, 25, 50, 100
Number of Samples	100	50, 75, 100, 500, 1000
Number of Edges	1×n	0.5×n, 1×n, 2×n, 5×n, 10×n
Network Structure	Random (Erdős–Rényi)	Band, clock, cluster, Erdős–Rényi, hub, scale-free
Underlying Parametric Distribution	Zero-inflated negative binomial	negative binomial, poisson, zero-inflated negative binomial, zero-inflated poisson

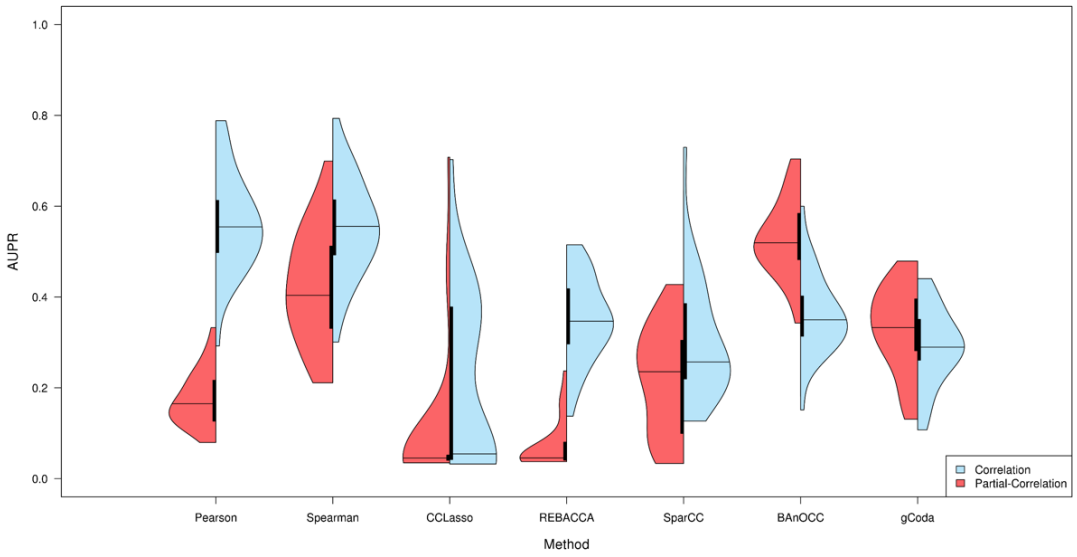

Figure 8.8 Performance of correlation-based methods on data simulated by a parametric model comparing outputted correlation and partial correlation matrices.

Split-violin plots for the AUPR of 30 replicates, comparing outputted correlation (left) or partial correlation (right) matrices with the known partial correlation matrix. Data simulated by the parametric model was modeled after American Gut Microbiome 16s data with the following control variables: 50 species, 100 samples, 1×number of species of edges, random (Erdős–Rényi) network structure, and zero inflated negative binomial distribution.

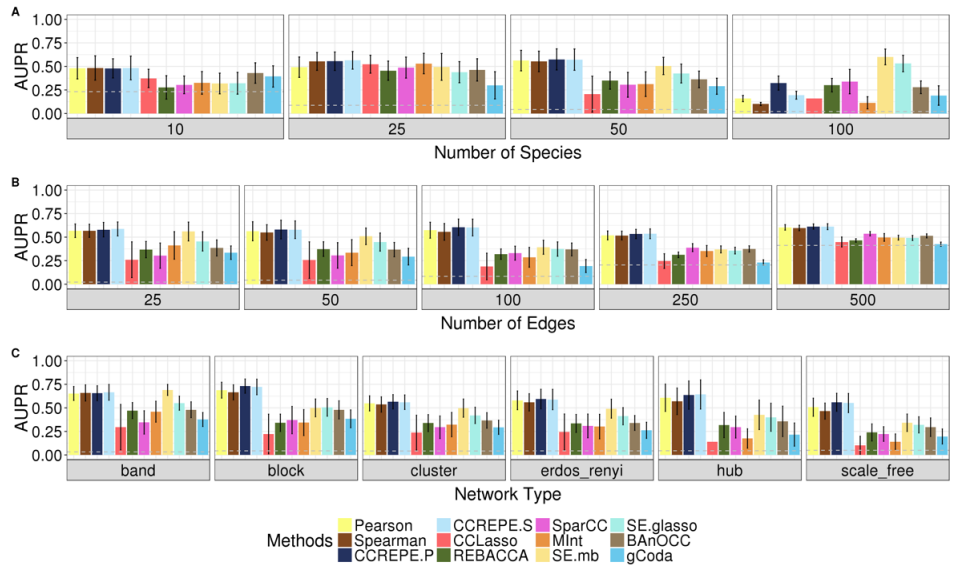

Figure 8.9 Performance of correlation-based methods on data simulated by statistical models.

Bars represent the mean AUPR of 30 replicates and the error bars represent the standard deviation comparing the outputted correlation matrix where available with the known partial correlation matrix. Simulated data was generated by vary one parameter (A – number of species, B – number of edges, C – network type) while keep other parameters the same as the control parameters described above in Figure 8.8.

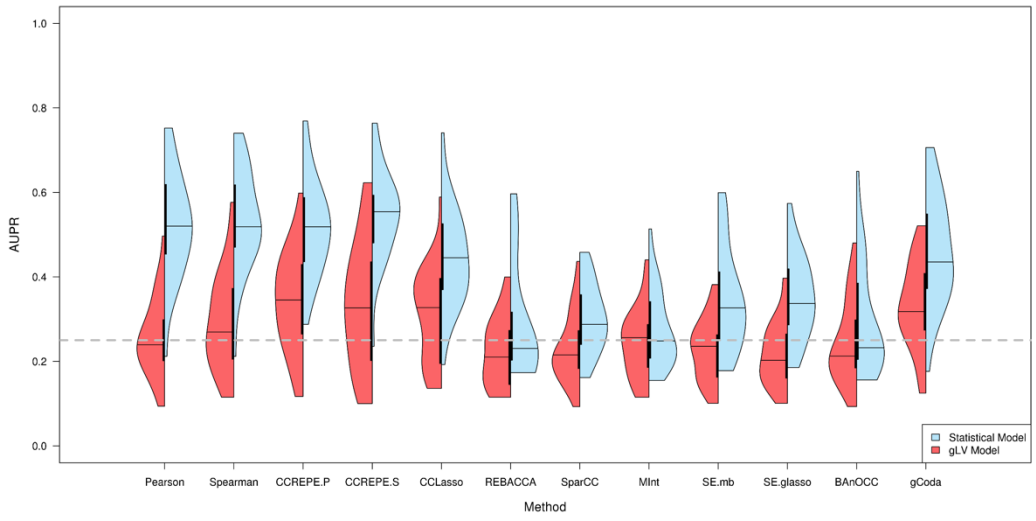

Figure 8.10 Performance of correlation-based methods on data simulated by gLVMs compared to data simulated by parametric models.

Split-violin plots representing the AUPR of 30 replicates, comparing outputted correlation matrix where available with the ground truth using simulated data generated from gLVMs (left) and parametric models (right). Simulated data was modeled after metagenomics sequencing data from stool samples of the curated metagenomics dataset data with the following parameters: 20 species, 500 samples, at equilibrium, asymmetric interaction network and species abundance in each sample following a log-normally distribution.

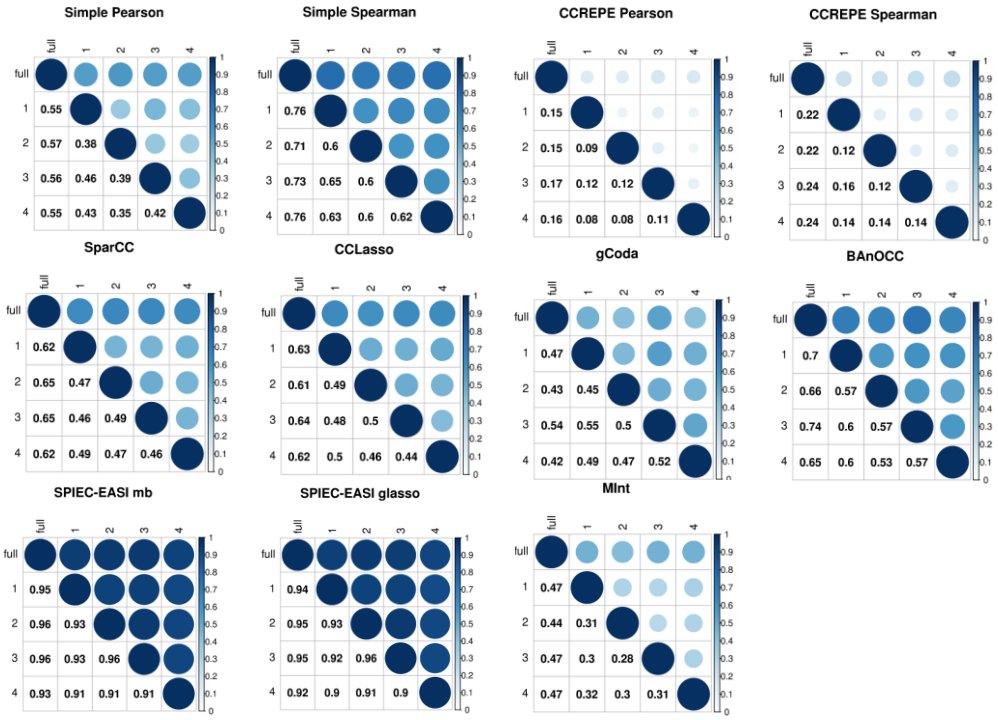

Figure 8.11 Jaccard similarity within individual methods on partitioned datasets.

Each circle on the upper corner represents the Jaccard similarity (proportional to both color and size of the circle) between two methods and the numbers on the bottom corner are the exact values.

APPENDIX 3 SUPPLEMENTARY TABLES & FIGURES FOR CHAPTER 4

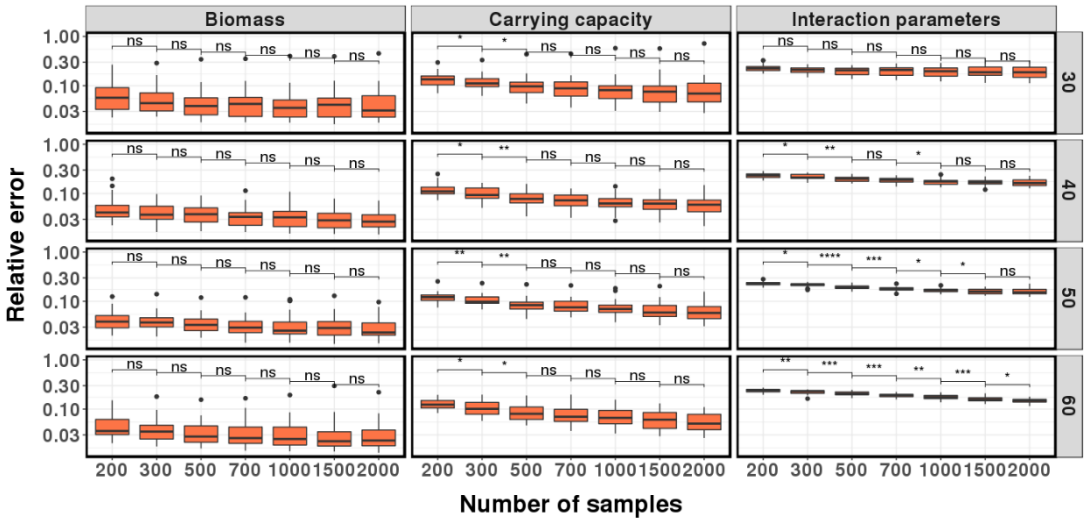

Figure 8.12 Median relative error of BEEM-static with varied number of samples and species (in rows) with simulated data

Each boxplot represents 30 simulated datasets with different gLVM parameters. The difference between adjacent "number of species" was tested with one-sided Wilcoxon rank sum test (ns: not significant, ****: p-value<10^{-3}, ***: p-value<10^{-2}, **: p-value<10^{-1}, *: p-value<0.05).