Dynamic Bayesian networks for integrating multi-omics time-series microbiome data

Daniel Ruiz-Perez[1⋆], Jose Lugo-Martinez[2⋆], Natalia Bourguignon[3,4,5], Kalai Mathee[4], Betiana Lerner[5], Ziv Bar-Joseph[2], and Giri Narasimhan[1]

[1] Bioinformatics Research Group (BioRG),
Florida International University, Miami, Florida, USA {druiz072,giri}@cs.fiu.edu
[2] Carnegie Mellon University, Pittsburgh, Pennsylvania, USA {jlugomar,zivbj}@cs.cmu.edu
[3] College of Engineering and Computing, Florida International University, Miami, Florida, USA
nataliabourguignon@gmail.com
[4] Herbert Wertheim College of Medicine, Florida International University, Miami, Florida, USA
kalai.mathee@fiu.edu
[5] National Technological University, Buenos Aires, Argentina blerner@frh.utn.edu.ar

Abstract. A key challenge in the analysis of longitudinal microbiomes data is to go beyond computing their compositional profiles and infer the complex web of interactions between the various microbial taxa, their genes, and the metabolites they consume and produce. To address this challenge, we developed a computational pipeline that first aligns multi-omics data and then uses dynamic Bayesian networks (DBNs) to integrate them into a unified model. We discuss how our approach handles the different sampling and progression rates between individuals, how we reduce the large number of different entities and parameters in the DBNs, and the construction and use of a validation set to model edges. Applying our method to data collected from Inflammatory Bowel Disease (IBD) patients, we show that it can accurately identify known and novel interactions between various entities and can improve on current methods for learning such interactions. Experimental validations support several predictions about novel metabolite-taxa interactions. The source code is freely available under the MIT Open Source license agreement and can be downloaded from https://github.com/DaniRuizPerez/longitudinal_multiomic_analysis_public.

Keywords: Dynamic interaction network inference · Longitudinal microbiome analysis · Multi-omic integration · Dynamic Bayesian networks · Temporal alignment.

⋆ equal contribution

1 Background

Microbiomes are communities of microbes inhabiting an environmental niche. The study of microbial communities offers a powerful approach for inferring their impact on the environment, specific diseases, and overall health. Initial work on microbiome analysis focused on *metataxonomics*, which involves sequencing reads from the 16S rRNA region to build a taxonomic profile of a microbiome [11]. While the profiling of the 16S gene provides some information about the set of microbes, it is not always possible to use it to fully determine the set of microbes present and their levels. This gave rise to studies that utilized *metagenomics* which sequenced the whole metagenome in a microbial community in order to determine their genetic profile and the detailed set of taxa and microbes [21]. More recently, additional types of biological data are being profiled in microbiome studies, including *metatranscriptomics*, which involves surveying the complete metatranscriptome of the microbial community [2], and *metabolomics*, which involves profiling the entire set of small molecules (metabolites) present in the microbiome's environmental niche [25].

Over the last decade, the NIH-funded Human Microbiome Project (HMP) [26] has been carried out in two phases with the aim to better understand the impact of the microbiome in human health and disease. In the second phase, referred to as integrative Human Microbiome Project (iHMP) [10], the goal has been to integrate multi-omics longitudinal data sets as a means to study the dynamics of the microbiome and host across select diseases, including preterm births, type 2 diabetes, and irritable bowel disorders.

Several studies have shed light on the microbiota living in environmental niches such as waters, soils, deep-sea vents, and Antarctic ice, as well as body sites of humans, animals, and plants [23,4,27]. However, most of these studies focused on characterizing the set of taxa and their levels and did not attempt to model the complex interactions between them. More recent studies, profiling additional types of data (including transcriptomics and metabolomics), mainly focus on separate analysis of each data type without linking all of them into a unified model [3]. In addition, microbiomes are inherently dynamic, and so to fully understand the complex interactions that take place within these communities, longitudinal microbiome data is required [7]. All the above points to a need for computational methods that can integrate diverse types of time series data to model the interactions among these entities and to help researchers understand systemic responses to biological events using multi-omics microbiome data.

More recently, a number of attempts have been made to analyze data from longitudinal studies. La Rosa et al. [13] studied patterns of changes in the abundance of important bacterial taxa in premature infant gut by sampling the gut microbiome over a period of time. However, their approach failed to capture interactions between taxa. An alternative approach involved the use of dynamical systems such as the generalized Lotka-Volterra (gLV) models [8,24]. gLV models use regression to study the stability of temporal bacterial communities. While gLV has been successfully applied to some datasets, the large set of parameters it uses is not well-suited for causality and probabilistic inference [16].

In previous work, we have shown that probabilistic graphical models (specifically Dynamic Bayesian Networks) can be used to study sequence microbiome data leading to models that can accurately predict future changes as well as identify interactions within the microbiome. While our DBN-based method was successful in modeling metagenomic sequence data from longitudinal microbiome studies, the method was not designed to incorporate useful and heterogeneous multi-omic data. We have extended our previous method so that it can utilize time series transcriptomics and metabolomics data in addition to the metagenomics information. The model also allows for a host of other multi-omics data, if available, to be integrated in a uniform manner.

The introduction of the multi-omic data implies a sizable increase in the number of nodes and edges in the DBN and requires the use of greater computational resources and parallelization techniques. However, the bigger challenge is overfitting of the large number of parameters. To overcome this challenge, we first restricted the set of allowable interactions between the entities based on simple biological assumptions. We also applied other common techniques, such as limiting the maximum number of parents of a node. The addition of multi-omics data allowed our methods to infer not just taxa-taxa interactions, but also to explain some of the mechanisms underlying these interactions. These include the impact of genes from various taxa on other taxa and how metabolite-taxa interactions can change the abundance at later time points. Statistical validations indicate that our DBNs correctly recover several known interactions. We have also performed new wet-lab experiments to test and validate novel model predictions.

2 Methods

Below we describe the processing pipeline that we propose.

2.1 Data

To test our proposed analysis pipeline, which combines temporal alignment, Bayesian network learning, and inference for multi-omics microbiome data, we used the Inflammatory Bowel Disease (IBD) cohort from the NIH Human Microbiome Project (HMP) [26]. The cohort included 132 individuals across five clinical centers [14]. During a period of one year, each subject was profiled every two weeks on average. This generated $1,785$ stool samples, 651 intestinal biopsies, and 529 blood samples. In addition to multi-omics data (e.g., metatranscriptomics [22] and metabolomics [6,17]), we also included clinical information.

2.2 Data pre-processing

The different data types were processed separately. First, the taxon, metabolite, and gene abundances were normalized to make each type separately add up to 1 for each subject. Then metabolites and genes were scaled to match the mean of the taxa because of the very distinct number of entities in each category. Metabolites without an HMDB correspondence or that had very low intensity values in most of the subjects were removed. Next, following [16], we performed temporal alignment of time series data from individuals. For this, we need to represent each discrete time series using a continuous function. Here we used B-splines for fitting continuous curves to the time-series multi-omic data profiled from each subject, including the microbial composition, gene expression, and metabolic abundance. To improve the accuracy of the reconstructed profiles, we removed any sample that had less than five measured time points in any of the multi-omics measurements. This led to the removal of 25 individuals from the cohort resulting in 107 individual multi-omic time series that were used for further analysis.

2.3 Temporal alignments

Given longitudinal samples from different subjects, we cannot expect that the rates at which various multi-omics levels change would be exactly the same between these individuals [1]. To facilitate the analysis of such longitudinal data across subjects, we first align the time series from the microbiome samples using the microbial composition profiles. As described earlier [16], these alignments use a linear time transformation function to warp one time series into a common, representative sample time series used as the reference.

To select an optimal reference sample from the 107 time series, we generated all possible pairwise alignments between them and selected the time series that resulted in the least total overall error in the alignments. Figure 1(a) shows the original (unaligned) time series of the abundance of the taxon, *Bacteroides dorei* (dashed orange), the warped (aligned) time series (solid orange) and the reference time series (blue).

Given an individual's warped/aligned time series of the abundance of a taxon of interest, the multi-omics data were incorporated as follows: the same transformation applied to the taxon abundance was applied to all the gene expression and metabolomic abundance time series. The resulting set used for the modeling comprised of 60 individual-wise heterogeneous alignments (after filtering out high alignment error individuals) involving 102 microbial taxa, 72 genes, and 70 metabolites.

2.4 Dynamic Bayesian network models

Using the aligned time series multi-omics data, we next learned graphical models that provide information about the relationships between the different omics (taxa, genes, and metabolites) and environmental (clinical) entities. In this work, we extend the DBN model proposed in Lugo-Martinez *et al.* [16] to account for multi-omics microbiome data with the goal of inferring the temporal relationships between the heterogeneous entities in a microbial community. A DBN is a directed acyclic graph where, at each time slice, nodes correspond to random variables of interest (e.g., taxa abundance, gene expression, age, etc.), and directed edges correspond to their conditional dependencies in the graph. These edges are defined as either: *intra*

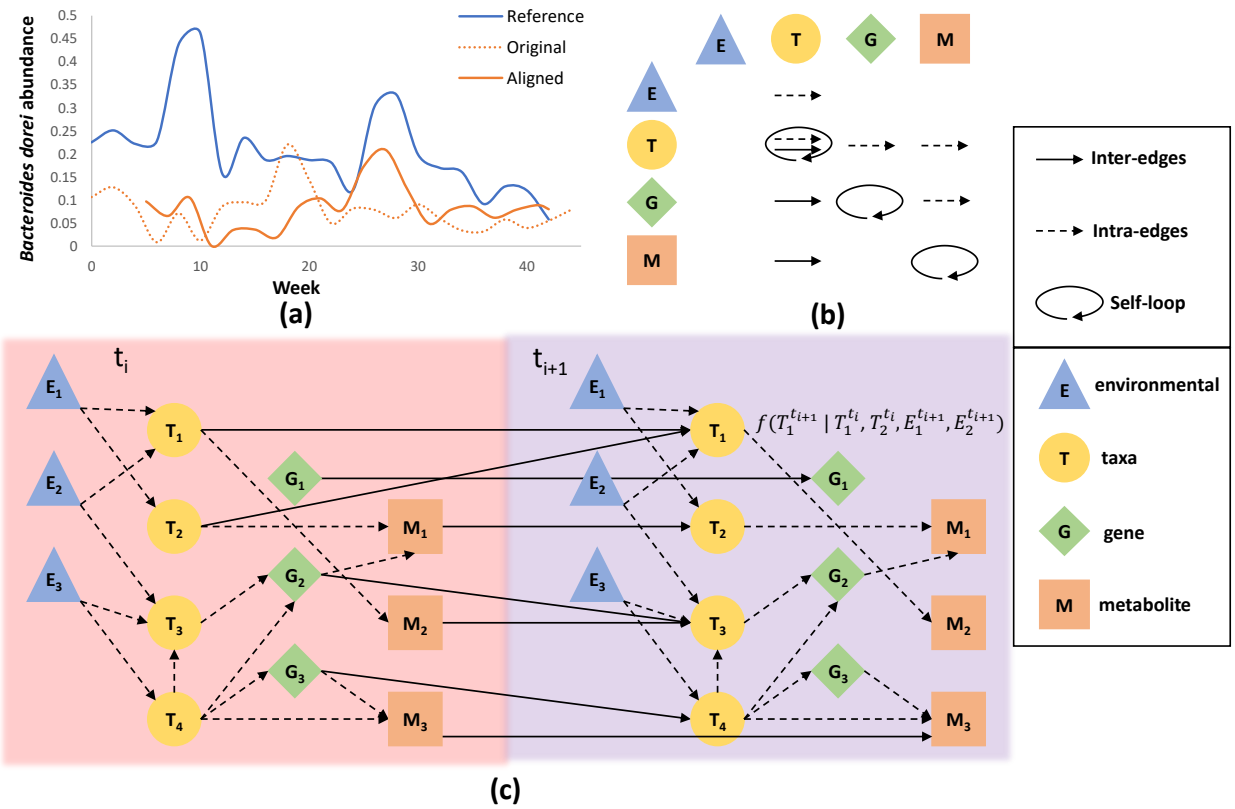

Fig. 1. Alignment and multi-omic framework. Figure shows the alignment process, the proposed structure of the DBN, and the interactions allowed to be learnt. **(a)** Alignment sample of the bacterial abundance of *Bacteroides dorei* in sample C3027 to the reference sample M2072. **(b)** Adjacency matrix encoding the metabolic framework of allowed interactions imposed (TGM). Cell (A, B) represents allowed interactions from parent A to child B. **(c)** Example of two consecutive DBN time slices.

edges connecting nodes from the same time slice, or *inter edges* connecting nodes between consecutive time slices. In a DBN model, only two slices are modeled and learned, as shown in Figure 1(c).

In this work, our DBN models encode four types of nodes: (i) taxon abundance, (ii) gene expression, (iii) metabolite concentration, and (iv) environmental factors. The first three types represent continuous variables, whereas the last type can be either discrete or continuous. The advantage of DBNs is their ability to seamlessly integrate discrete and continuous variables in a single probabilistic framework. Formally, let Θ denote the set of parameters for the DBN and G denote a specific network structure over discrete and continuous variables in the multi-omics microbiome study. The joint distribution can be decomposed as

$$P(\Delta)F(\Psi|\Delta) = \prod_{x \in \Delta} p(x \,|\, \mathbf{Pa}^G(x)) \prod_{y \in \Psi} f(y \,|\, \mathbf{Pa}^G(y)),$$

where P denotes a set of conditional probability distributions over discrete variables Δ, F denotes a set of linear Gaussian conditional densities over continuous variables Ψ, and $\mathbf{Pa}^G(X)$ denotes the set of parents for variable X in G [16,18]. In particular, continuous variables are modeled using a Gaussian with the mean set based on a regression model over the set of continuous parents as follows

$$f(y \,|\, u_1, \cdots, u_k) \sim N(\beta_0 + \sum_{i=1}^{k} \beta_i \times u_i, \sigma^2),$$

where $u_1, \ldots, u_k$ are continuous parents of y; β_0 is the intercept; $\beta_1, \ldots, \beta_k$ are the corresponding regression coefficients for $u_1, \cdots, u_k$; and σ^2 is the standard deviation. In the case that y has discrete parents then

we need to compute coefficients $B = \{\beta_i\}_{i=0}^{k}$ and standard deviation σ^2 for each discrete configuration. As highlighted in Fig. 1(c), the conditional linear Gaussian density function for variable $T_1^{t_{i+1}}$ denoted as $f(T_1^{t_{i+1}} \,|\, T_1^{t_i}, T_2^{t_i}, E_1^{t_{i+1}}, E_2^{t_{i+1}})$ is modeled by

$$N(\beta_0 + \beta_1 \times T_1^{t_i} + \beta_2 \times T_2^{t_i} + \beta_3 \times E_1^{t_{i+1}} + \beta_4 \times E_2^{t_{i+1}}, \sigma^2),$$

where $\beta_1, \beta_2, \beta_3$ and σ^2 are the DBN model parameters. Here, we infer the parameters Θ by maximizing the likelihood of the longitudinal multi-omics data D given our regression model and known structure G.

The problem of learning the DBN structure is expressed as finding the optimal structure and parameters

$$\max_{\Theta, G} P(D \,|\, \Theta, G) P(\Theta, G) = P(D, \Theta \,|\, G) P(G),$$

where $P(D \,|\, \Theta, G)$ is the likelihood of the data given the model. As in Lugo-Martinez *et al.* [16], we maximize $P(D, \Theta \,|\, G)$ for a given structure G using maximum log-likelihood estimation (MLE) combined with BIC score defined as

$$\mathrm{BIC}(G, D) = \log P(D \,|\, \Theta, G) - \frac{|\Theta|}{2} \log |D|,$$

where $|\Theta|$ is the number of DBN model parameters in structure G, and $|D|$ is the number of time points in D. This approach enables an effective way to search over the set of all possible DBN structures while favoring simpler structures.

2.5 Constraining the DBN structure

An important innovation in this work is how the structure of the DBN was constrained. Besides limiting the maximum number of possible parents (i.e., incoming edges) for each node to 3, we only allow edges between certain types of nodes in the network. These constraints (in the form of an matrix received as an input to the function) helped reduce the complexity of searching over possible structures and prevented over-fitting. Specifically, we allowed intra edges from environmental variables to microbial taxa (abundance) nodes, from taxa nodes to gene (expression) nodes and from gene nodes to metabolites (concentration) nodes. All other interactions within a time point (for example, direct gene to taxa) were disallowed. We also allowed inter edges from metabolites to taxa nodes in the next time point, and *self-loops* from any node $A_1^{t_i}$ to $A_1^{t_{i+1}}$, except for environmental variables for which no incoming edges were allowed. These restrictions referred to as the *TGM_Skeleton* reflect our understanding of the basic ways the different entities interact with each other, i.e., environmental variables are independent variables, taxa express genes, which are involved in metabolic pathways; finally, the metabolites impact the growth of taxa (in the next time slice).

We also learned DBNs using a less constrained framework, denoted *TGM*. Unlike *TGM_Skeleton*, in *TGM* we also allowed direct edges between taxa and metabolites. The TGM constraints are shown in the form of an adjacency matrix in Figure 1(b), which are checked during the structure learning step for the DBN. Note that other constraints such as requiring that taxa could only connect to genes present in their genome were not imposed since genomics reference databases are not always complete and so relying on these may lead to missing key interactions.

We used a greedy hill-climbing approach for structure learning where the search is initialized with a network that connects each node of interest at the previous time point to the corresponding taxa node at the next time point. Next, nodes are added as parents of a specific node via intra or inter edges depending on which valid edge leads to the largest increase of the log-likelihood function beyond the global penalty incurred by adding the parameters as measured by the BIC score approximation.

Every network was bootstrapped by randomly selecting with replacement as many subjects as in the dataset, and learning a different network 100 times. The networks were then combined, and the regression coefficient of the edges was averaged. Each edge was also labelled with the bootstrap support (percentage of times that edge appears). Each repetition was set to run independently on a separate processor using Matlab's Parallel Computing Toolbox. Other parallel implementations include parallelizing each repetition in the cross-validation computation of the inference error and each independent alignment error calculation using Python's Parallel library.

2.6 *In silico* validation of DBN edges

One of the biggest challenges in building models of biological interactions lies in developing methods to validate them and in providing confidence measures. Since DBNs are generative models, one approach is to predict time series using previous time points and thus to achieve cross validation [16]. Such technical validation, while informative, does not shed light on the accuracy of specific edges and interactions predicted by the model. To address this issue, a second approach is to match the edges against a database of known interactions between taxa and metabolites and/or taxa and genes. Unfortunately, no such comprehensive database exists. For example, highly curated databases such as HMDB [28] and MetaCyc [12] are inadequate since the intersection of their contents with the species and metabolites in our networks was very small.

To assist in the validation of taxa-metabolite ($T \rightarrow M$) edges in our networks, we used the tool MIMOSA [19]. MIMOSA calculates the metabolic potential of each species, i.e., the capability of a species to produce a metabolite under the conditions of the data set. The list of all taxon-metabolite pairs from our DBNs that resulted in a positive score in MIMOSA was used as a validation database.

For taxa-gene ($T \rightarrow G$) validations, we used KEGG to build a validation database of bacterial taxa and the genes present in their genomes. To keep this database small, we only used taxa and genes present in our network. If multiple strains were available for a bacterial species, then all genes from each strain were aggregated. The one-time creation of a local validation database also speeded up our computations considerably.

To calculate the statistical significance of validated interactions compared to a null model, a Poisson-Binomial distribution test was executed. The main reason that a simple binomial test cannot be performed is the differences in the in-degree distribution between different nodes in the network. Some nodes have many more validated interactions when compared to others, and so a uniform model for each edge does not accurately capture the null probability of selecting such an edge. This was done with the function *ppoisbinom* from the R package *poisbinom* [20], which gives the cumulative distribution function of the probability of validating by chance at least as many interactions as the number of true positives. The validation precision of the network was also calculated as the percentage of validated interactions from the ones predicted, even though this homogeneous metric ignores the differential significance of each interaction.

2.7 Laboratory validations of metabolites consumed by taxa

Wet lab experiments were carried out to validate predicted $M \rightarrow T$ interactions. This was done by growing relevant taxa in isolation, and adding the relevant metabolite to measure impact on growth. To select candidates for these experiments we used a network built using only metagenomic and metabolomic data. This was done to avoid the loss of strong interactions that only appear as indirect because of genes in the overall network. We next ranked the top predictions by sorting the metabolite-taxa interactions based on the ranking function $R(i,j) = |\lambda_{ij}^N| * \beta_{ij}$, where β_{ij} is the bootstrap score as defined by the percentage of repetitions in which that edge appears, and λ_{ij}^N is the normalized regression coefficient of the edge that goes from metabolite M_i to taxa T_j, using the normalization function introduced in [16]

$$\lambda_i^N = \frac{\lambda_i \times \bar{T}_i}{\sum_{j=1}^{k} \left|\lambda_j \times \bar{T}_j\right|},$$

where $\bar{T}_i$ is the mean abundance of taxa T_i across all samples.

We selected 5 of the top 10 interactions for further analysis based on the ready availability of the bacteria and metabolites. These metabolites were expected to affect taxa growth because of the edge between metabolite concentration and taxon abundance. After plotting the growth curves with the bacterium and metabolite in question, we assessed if each metabolite was enhancing/inhibiting the taxon growth using a two-tailed paired t-test when compared to growth without the metabolite.

3 Results

3.1 Resulting Dynamic Bayesian network models

We used the aligned gut microbiome data (Methods) to learn multi-omic dynamic Bayesian models (DBNs) that provide information about interactions between taxa, genes and metabolites, and the impact of clinical

variables on these entities over time. The DBN learned by our method for the IBD data set is presented in Figure 2. In this figure, each node represents either a bacterial taxon, a gene, a metabolite, or a clinical variable; directed edges represent inferred temporal relationships between the nodes. Since we realize that the network is complex and hard to read, we also provide a Cytoscape session with an interactive version that can be easily explored on the supporting website.

The learned network contains 244 nodes per time slice (101 microbial taxa, 72 genes, 70 metabolites, and 1 clinical variable). To determine the edges in the network we applied bootstrapping, rerunning the method 100 times with each execution using a new dataset created by selecting, with replacement, as many subjects as there were in the dataset. We next extracted all edges from all executions, resulting in 1033 distinct directed edges (434 inter edges and 599 intra edges), as shown in Figure 2. Note that while there was a very large overlap between edges learned in each iteration, since we used the union of all networks, the number of edges is larger than the number of possible edges for a single iteration (1033 vs. 244*3 = 732).

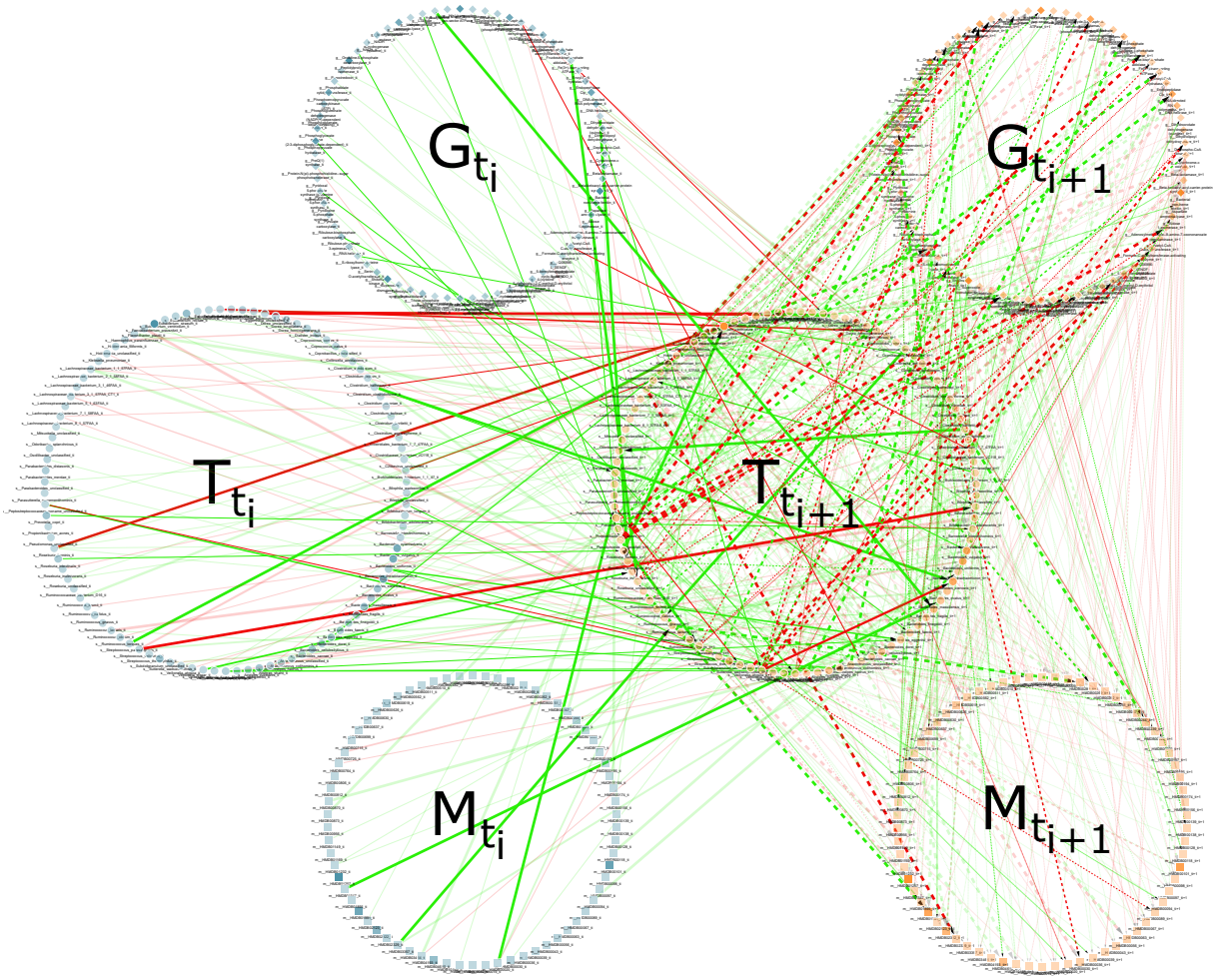

Fig. 2. Resulting DBN. Figure shows two consecutive time slices t_i (blue) and t_{i+1} (orange), where nodes are either taxa (circles), genes (diamonds), or metabolites (squares). The clinical variable didn't have any connection so it was removed from the graph. The different node types have been grouped in different circles, their transparency is proportional to their average abundance relative to that node type, and the two time slices were separated. Dotted lines denote *intra edges* (i.e., directed links between nodes in same time slice) whereas solid lines denote *inter edges* (i.e., directed links between nodes in different time slices). Edge color indicates positive (green) or negative (red) temporal influence and edge transparency indicates strength of bootstrap support. Edge thickness indicates statistical influence of regression coefficient after normalizing for parent values, as described in [16].

3.2 Evaluating the learned DBN model

We first performed a technical evaluation of the learned DBN model and compared it to models constructed by other existing methods [15]. The performance of each model was evaluated through per-subject cross-validation with the goal of predicting microbial composition using each learned model. In each iteration, the longitudinal microbial abundance profile of a single subject was selected as the test set, and the multi-omics data from all other subjects were used for building the network and learning model parameters. Next, starting from the second time point, we used the learned model to predict an abundance value for every taxon in the test set at each time point using the previous and current time points. Finally, we normalized the predicted values in order to represent the relative abundance of each taxon and measured the average predictive accuracy by computing the mean absolute error (MAE) for the selected taxon in the network. This process was repeated for different combinations of multi-omics data, which ranged from top 10, 20 up to all data, which includes taxa (T), genes (G), and metabolites (M) on the aligned data. The average MAE for predictions on the IBD data set for a sampling rate of two weeks is summarized in Figure 3.

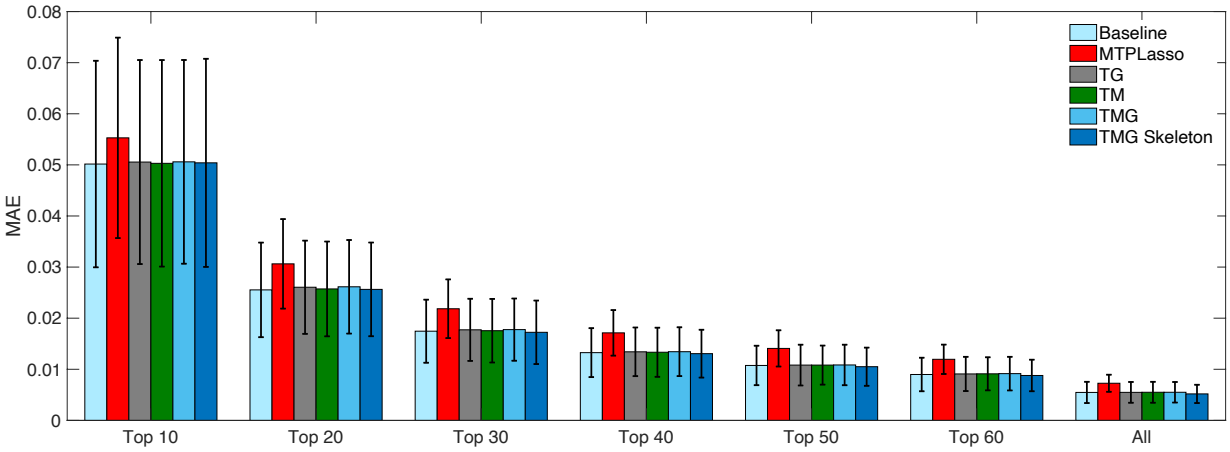

Fig. 3. Comparison of average predictive accuracy between methods on the IBD data sets. Figure shows the MAE of our proposed DBN models against a baseline method using only metagenomic data and a previously published approach for a sampling rate of two weeks which most closely resembles the originally measured time points. We have also compared the performance of our method when using all data types to a similar DBN approach that only uses subsets of them (T, G, and M correspond to taxa, genes and metabolites, respectively).

We used this process to compare the multi-omics DBN strategy to the one that used only metagenomic data [16] (Baseline) and to MTPLasso [15], which models time-series multi-omics microbial data using a gLV model. In both cases, we used the default setup and parameters, as described in the original publications. As shown by Figure 3 our method outperforms MTPLasso while it also performs favorably when compared to the Baseline (a DBN learned solely on taxon data) when all entities are used in the model.

3.3 Computationally validating predicted edges

We used the method described in Section 2.6 to validate the Taxa-Metabolite ($T \rightarrow M$) and Taxa-Gene ($T \rightarrow G$) edges predicted by the DBN. Each predicted interaction was either considered "validated" if it appears in the validation database, or "not validated" if it was not found, but the parent and child nodes were part of the database. Interactions predicted between taxa and/or metabolites not included in the database were not used in this analysis. We compared the results of different DBNs (where the differences were based on the subset of data types used for construction of the DBN or on the allowed interactions in the network) to a random baseline network. To generate the random network, we used the same nodes in the multi-omic network and assigned the same number of edges as in the learned DBN by randomly selecting a parent and

child from the possible interaction list (Figure 1(b)). This was repeated 1000 times, averaging the metrics over all random runs.

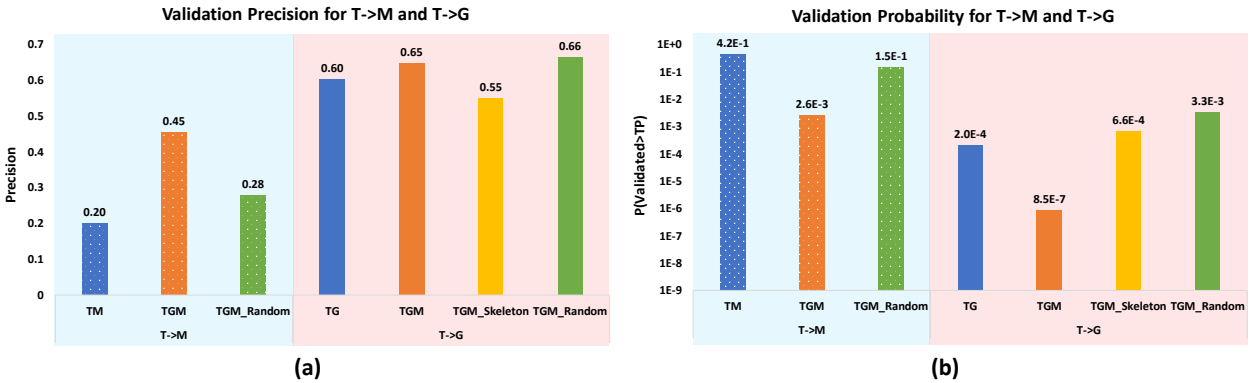

Fig. 4. ***In silico*** **validation results**. Two performance metrics for the different networks are shown. Edges that appeared in at least 90% of the bootstrap repetitions were used for this analysis. The blue area corresponds to the validation of taxa-produced metabolites, while the red area corresponds to the taxa-gene interactions. **(a)** Precision scores (higher is better) for the different networks. **(b)** Significant intersection score (lower is better) with the y axis showing the probability of validating by chance at least as many interactions as were actually validated.

Figure 4 shows the validation comparison for $T \rightarrow M$ and $T \rightarrow G$ interactions. For the former case, four different networks were compared; (a) a network that included only taxa and metabolites (TM), (b) a DBN learned by our model that uses taxa, genes, and metabolites (TGM), (c) a network with taxa, genes, and metabolites without allowing the auxiliary edges (TGM_Skeleton), and (d) the random baseline network (TGM_Random). For $T \rightarrow M$, the TGM_Skeleton cannot be evaluated since it does not allow for direct taxa to metabolites edges. Figure 4(a) shows the precision of each method in predicting validated interactions, while Figure 4(b) displays the significance of the overlap between the predicted and known interactions. As can be seen, the multi-omic TGM achieves a higher precision for predicting $T \rightarrow M$ interactions when compared to all other methods. TGM and baseline both performed best for the $T \rightarrow G$ predictions. On the other hand, in terms of significance, TGM performed much better than the random baseline and all the other networks, not only for $T \rightarrow G$, but also for $T \rightarrow M$ ($p = 8.5E-7$ vs. $p = 3.3E-3$ for TGM_Random). This makes TGM the best from the methods compared in terms of retrieving biologically relevant $T \rightarrow G$ and $T \rightarrow M$ interactions. Other bootstrap cutoffs lead to similar results (Figure 8).

3.4 Biological validation experiments

Since most of the edges selected by the DBN model were novel, we decided to experimentally test some of the top predictions. For this, we selected 5 of the top 10 interactions from the ranked prediction list (Section 2.7), based on availability of taxa and metabolites. The five predictions we tested were:

- 4-Methylcatechol (4-MC) $\rightarrow$ *Escherichia coli*
- 4-Hydroxyphenylacetate (4-HPA) $\rightarrow$ *Escherichia unclassified*
- D-Xylose $\rightarrow$ *Pseudomonas unclassified*
- 4-Guanidinobutanoate (4-GOB) $\rightarrow$ *Pseudomonas unclassified*
- 1-Methylnicotinamide (1-MNA) $\rightarrow$ *Pseudomonas unclassified*

P. aeruginosa ([9]) was used in the laboratory experiments instead of *Pseudomonas unclassified*, and *E. coli* ([5]) instead of *Escherichia unclassified*, but the results were expected to be the same. A standard Luria Bertani (LB 20%) culture media was used and the bacterial density was measured using a wavelength of 600 nm (OD600). Figure 5 shows the growth curves of the microbes before and after adding the metabolites, and a control case without adding the metabolites (LB 20%). As a positive control, we also profiled the impact of

known growth metabolites (the sugars Glucose and D-Xylose for *E. coli*, and Succinate for *P. aeruginosa*). All metabolites shown were recognized by our DBN as predictors for the target taxa and were expected to be used by them. Note that the metabolites were added at the stationary phase when the bacteria were no longer growing rapidly. For the *E. coli* predictions, 4-HPA, and 4-MC significantly enhanced the growth with higher potency for 4-MC. For *P. aeruginosa*, 4-GOB appeared to be inhibitory, but not statistically significant, while 1-MNA was significantly enhancing, as predicted. The p-values for all observations can be seen in Table 1, where a two-tailed paired t-test was executed for the 3 time points with the highest difference from the baseline. For more details on the experimental settings please check Appendix Section A.

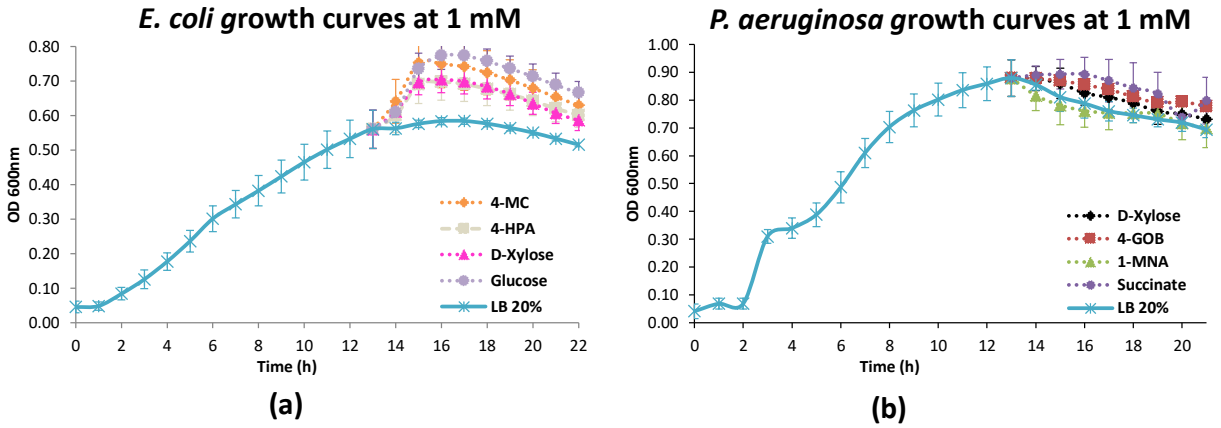

Fig. 5. Growth curves at 1mM. Different metabolites were introduced at 1mM concentration at the end of the exponential phase (0-14h). The resulting growth curves were plotted after all data points were averaged over 10 replicates. **(a)** *E. coli*, with Glucose and D-Xylose as controls. **(b)** *P. aeruginosa*, with Succinate as control.

***Escherichia coli* HB101**			
Met.	**16h**	**17h**	**18h**
LB	0.583	0.584	0.576
4-MC	0.750 (0.0002)	0.741 (0.0003)	0.724 (0.0003)
4-HPA	0.697 (0.0005)	0.692 (0.0005)	0.677 (0.0004)
D-Xylose	0.704 (0.0017)	0.699 (0.0016)	0.684 (0.0019)
Glucose	0.774 (0.0005)	0.773 (0.0003)	0.758 (0.0003)

***Pseudomonas aeruginosa* PAO1**			
Met.	**15h**	**16h**	**17h**
LB	0.811	0.787	0.760
D-Xylose	0.860 (0.0189)	0.825 (0.0395)	0.810 (0.0317)
4-GOB	0.870 (0.0185)	0.855 (0.0085)	0.838 (0.0124)
1-MNA	0.782 (0.0717)	0.761 (0.1551)	0.754 (0.7547)
Succinate	0.893 (0.0716)	0.893 (0.1195)	0.870 (0.0172)

Table 1. Metabolite effect at 1mM. Taxa density appears in black, while the p-values are inside parenthesis. Red p-values represent a significant difference compare to LB 20% ($p<0.05$). Green p-values represent a non-significant difference from LB 20%.

4 Discussion

Previous microbiome studies focused primarily on the set of taxa and their levels in each of the samples. More recent datasets are much richer, notably including gene expression and metabolomics data. The ability to integrate these multi-omics, longitudinal data remains a major challenge for microbiome analysis.

Here, we have presented a new approach based on continuous alignment followed by DBN modeling. Our method first represents each time series using continuous curves and then aligns them using a reference time series. Next, we sample uniformly the aligned curves and learn a DBN model that combines taxa, genes and metabolites. Edges in the DBN represent predicted interactions between the entities and can be used to explain changes in the microbiome over time.

We applied our method to recent data from IBD patients. As we show, models learned using the multi-omics data are able to successfully predict taxa abundance at future time points, improving on models that do not use all available data. We curated a taxa-metabolite and taxa-gene validation set and have shown that interactions predicted by the learned DBNs significantly intersect known interactions. Finally, we experimentally tested 5 of the top predictions of metabolite-taxa relationships and have been able to validate 4 of them.

While our computational pipeline successfully reconstructed the model for the data we analyzed, there are still a number of ways in which it can be improved. Our alignment method is based on a greedy approach, and uses a taxon abundance time series as a reference. Using other omics data for the alignment may lead to improved models. Another issue is the incompleteness of current microbiome interaction databases that are critical for evaluating learned networks. Without a more complete database of validated interactions it would be hard to compare different computational methods for this task. The laboratory validations show a viable way to validate some of the interactions. However, they could also be improved by attempting to recreate more realistic conditions for the experiments and could be enhanced to validate other omics observations as well.

Our alignment and DBN methods are implemented in Python and Matlab correspondingly. Source code and the dataset used in this paper can be obtained from the link on the cover page.

Acknowledgments. This work was partially supported by McDonnell Foundation program on Studying Complex Systems (awarded to ZB-J), National Science Foundation (award number DBI-1356505 to ZB-J), and National Institute of Health (award number 1R15AI128714-01 to GN). The authors thank Shekhar Bhansali and Maximiliano S. Perez for their support of the laboratory experiments.

References

1. Bar-Joseph, Z., Gitter, A., Simon, I.: Studying and modelling dynamic biological processes using time-series gene expression data. Nat Rev Genet **13**, 552–564 (2012)
2. Bashiardes, S., Zilberman-Schapira, G., Elinav, E.: Use of metatranscriptomics in microbiome research. Bioinformatics and biology insights **10**, BBI–S34610 (2016)
3. Beale, D.J., Karpe, A.V., Ahmed, W.: Beyond metabolomics: a review of multi-omics-based approaches. In: Microbial metabolomics, pp. 289–312. Springer (2016)
4. Bertilsson, S., Burgin, A., Carey, C.C., Fey, S.B., Grossart, H.P., Grubisic, L.M., Jones, I.D., Kirillin, G., Lennon, J.T., Shade, A., et al.: The under-ice microbiome of seasonally frozen lakes. Limnology and Oceanography **58**(6), 1998–2012 (2013)
5. Boyer, H.W., Roulland-dussoix, D.: A complementation analysis of the restriction and modification of dna in escherichia coli. Journal of molecular biology **41**(3), 459–472 (1969)
6. Franzosa, E.A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H.J., Reinker, S., Vatanen, T., Hall, A.B., Mallick, H., McIver, L.J., et al.: Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nature microbiology **4**(2), 293 (2019)
7. Gerber, G.K.: The dynamic microbiome. FEBS Lett **588**(22), 4131–4139 (2014)
8. Gibson, T.E., Gerber, G.K.: Robust and scalable models of microbiome dynamics. In: Proc. 35th International Conference on Machine Learning. pp. 1763–1772. PMLR 80 (2018)
9. Holloway, B.: Genetic recombination in pseudomonas aeruginosa. Microbiology **13**(3), 572–581 (1955)
10. Integrative, H.: The integrative human microbiome project. Nature **569**(7758), 641 (2019)
11. Janda, J.M., Abbott, S.L.: 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of clinical microbiology **45**(9), 2761–2764 (2007)
12. Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Paley, S.M., Pellegrini-Toole, A.: The ecocyc and metacyc databases. Nucleic acids research **28**(1), 56–59 (2000)
13. La Rosa, P.S., Warner, B.B., Zhou, Y., Weinstock, G.M., Sodergren, E., Hall-Moore, C.M., Stevens, H.J., Bennett, W.E., Shaikh, N., Linneman, L.A., Hoffmann, J.A., Hamvas, A., Deych, E., Shands, B.A., Shannon, W.D., Tarr, P.I.: Patterned progression of bacterial populations in the premature infant gut. Proc Natl Acad Sci **111**(34), 12522–12527 (2014)
14. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., et al.: Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature **569**(7758), 655 (2019)

15. Lo, C., Marculescu, R.: Inferring microbial interactions from metagenomic time-series using prior biological knowledge. In: Proc. 8th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. pp. 168–177. ACM-BCB '17 (2017)
16. Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G., Bar-Joseph, Z.: Dynamic interaction network inference from longitudinal microbiome data. Microbiome **7**(1), 54 (2019)
17. Mallick, H., Franzosa, E.A., Mclver, L.J., Banerjee, S., Sirota-Madi, A., Kostic, A.D., Clish, C.B., Vlamakis, H., Xavier, R.J., Huttenhower, C.: Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nature communications **10**(1), 3136 (2019)
18. McGeachie, M.J., Sordillo, J.E., Gibson, T., Weinstock, G.M., Liu, Y.Y., Gold, D.R., Weiss, S.T., Litonjua, A.: Longitudinal prediction of the infant gut microbiome with dynamic bayesian networks. Sci Rep p. 20359 (2016)
19. Noecker, C., Eng, A., Srinivasan, S., Theriot, C.M., Young, V.B., Jansson, J.K., Fredricks, D.N., Borenstein, E.: Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. MSystems **1**(1), e00013–15 (2016)
20. Olivella, S., Shiraito, Y.: poisbinom: A faster implementation of the poisson–binomial distribution. r package version 1.0. 1 (2017)
21. Riesenfeld, C.S., Schloss, P.D., Handelsman, J.: Metagenomics: genomic analysis of microbial communities. Annu. Rev. Genet. **38**, 525–552 (2004)
22. Schirmer, M., Franzosa, E.A., Lloyd-Price, J., McIver, L.J., Schwager, R., Poon, T.W., Ananthakrishnan, A.N., Andrews, E., Barron, G., Lake, K., et al.: Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. Nature microbiology **3**(3), 337 (2018)
23. Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., Arrieta, J.M., Herndl, G.J.: Microbial diversity in the deep sea and the underexplored "rare biosphere". Proceedings of the National Academy of Sciences **103**(32), 12115–12120 (2006)
24. Stein, R.R., Bucci, V., Toussaint, N.C., Buffie, C.G., Rätsch, G., Pamer, E.G., Sander, C., Xavier, J.B.: Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. PLoS Comput Biol **9**(12), 1–11 (2013)
25. Turnbaugh, P.J., Gordon, J.I.: An invitation to the marriage of metagenomics and metabolomics. Cell **134**(5), 708–713 (2008)
26. Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., Gordon, J.I.: The human microbiome project. Nature **449**(7164), 804 (2007)
27. Turner, T.R., James, E.K., Poole, P.S.: The plant microbiome. Genome biology **14**(6), 209 (2013)
28. Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., et al.: Hmdb 3.0—the human metabolome database in 2013. Nucleic acids research **41**(D1), D801–D807 (2012)