

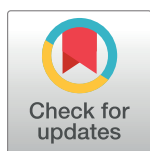
RESEARCH ARTICLE

# MPLasso: Inferring microbial association networks using prior microbial knowledge

Chieh Lo\*, Radu Marculescu

Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America

\* [chiehlcmu@gmail.com](mailto:chiehlcmu@gmail.com)



## Abstract

Due to the recent advances in high-throughput sequencing technologies, it becomes possible to directly analyze microbial communities in human body and environment. To understand how microbial communities adapt, develop, and interact with the human body and the surrounding environment, one of the fundamental challenges is to infer the interactions among different microbes. However, due to the compositional and high-dimensional nature of microbial data, statistical inference cannot offer reliable results. Consequently, new approaches that can accurately and robustly estimate the associations (putative interactions) among microbes are needed to analyze such compositional and high-dimensional data. We propose a novel framework called Microbial Prior Lasso (MPLasso) which integrates graph learning algorithm with microbial co-occurrences and associations obtained from scientific literature by using automated text mining. We show that MPLasso outperforms existing models in terms of accuracy, microbial network recovery rate, and reproducibility. Furthermore, the association networks we obtain from the Human Microbiome Project datasets show credible results when compared against laboratory data.

## OPEN ACCESS

**Citation:** Lo C, Marculescu R (2017) MPLasso: Inferring microbial association networks using prior microbial knowledge. *PLoS Comput Biol* 13(12): e1005915. <https://doi.org/10.1371/journal.pcbi.1005915>

**Editor:** Nicola Segata, University of Trento, ITALY

**Received:** June 25, 2017

**Accepted:** December 5, 2017

**Published:** December 27, 2017

**Copyright:** © 2017 Lo, Marculescu. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The MPLasso R package can be downloaded from here [https://github.com/ChiehLo/MPLasso\\_RPackage](https://github.com/ChiehLo/MPLasso_RPackage). All other relevant data are within the paper and its Supporting Information files.

**Funding:** This work was supported in part by the US National Science Foundation (NSF) [CCF-1514206], [https://www.nsf.gov/awardsearch/showAward?AWD\\_ID=1514206](https://www.nsf.gov/awardsearch/showAward?AWD_ID=1514206). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Author summary

Microbial communities exhibit rich dynamics including the way they adapt, develop, and interact with the human body and the surrounding environment. The associations among microbes can provide a solid foundation to model the interplay between the (host) human body and the microbial populations. However, due to the unique properties of compositional and high-dimensional nature of microbial data, standard statistical methods are likely to produce spurious results. Although several existing methods can estimate the associations among microbes under the sparsity assumption, they still have major difficulties to infer the associations among microbes given such high-dimensional data. To enhance the model accuracy on inferring microbial associations, we propose to integrate multiple levels of biological information by mining the co-occurrence patterns and interactions directly from large amount of scientific literature. We first show that our proposed method can outperform existing methods in synthetic experiments. Next, we obtain credible inference results from Human Microbiome Project datasets when compared against laboratory data. By creating a more accurate microbial association network, scientists in

**Competing interests:** The authors have declared that no competing interests exist.

this field will be able to better focus their efforts when experimentally verifying microbial associations by eliminating the need to perform exhaustive searches on all possible pairs of associations.

## Introduction

Microbes play an important role both in environment and human life. However, the way microbes affect the human health remains largely unknown. Knowledge of the microbial interactions can provide a solid foundation to model the interplay between the (host) human body and the microbial populations; this can serve as a key step towards precision medicine [1]. Unfortunately, understanding microbes interactions is difficult, as most microbes cannot be easily cultivated in standard laboratory settings. However, the recent increase of quality and reduced costs of sequencing technologies (e.g., shotgun or PCR directed sequencing [2]) enable researchers to collect information from the entire genome of all microbes under different environment conditions. As a result, various datasets ranging from earth ecosystem to human microbiome have been made publicly available under the Human Microbiome Project [3] or the Earth Microbiome Project [4].

In this paper, we aim at analyzing the networks of associations (putative interactions) among the microbes of human microbiome in order to understand how microbes can affect the human health. To this end, there exist several challenges: First, the amount of sequenced data that corresponds to human microbiome available from public websites is scarce. To date, one of the largest metagenomic datasets of human niches is the NIH Human Microbiome Project (HMP) [3] which only provides a few hundreds of healthy individual samples ( $n$ ) of various body sites, while the number of measured microbes ( $p$ ) usually ranges from hundreds to thousands. As a consequence, the number of associations ( $p(p-1)/2$ ) is much greater than the number of samples (i.e., high-dimensional data). Another big challenge stems from the nature of the data itself. Sequencing data only provides the *relative* abundance of various species; this is because the sequencing results are a function of sequencing depth and the biological sample size [5]. Therefore, from a statistical standpoint, the *relative* taxon abundance falls into the class of compositional data [6]; this causes statistical methods such as Pearson or Spearman correlations (which work with absolute values) to generate spurious results.

To infer microbe associations for both compositional and high-dimensional data, several algorithms have been developed. A pioneering method called SparCC [7] applies log-ratio transform on compositional data and directly approximates the correlation among microbes based on sparsity assumption of microbial associations. However, SparCC does not consider the influence of errors in compositional data; this may reduce the correlation estimation accuracy. More precisely, SparCC approximates the basis variance (i.e., the variance of compositional data) under the assumption that average correlations are small. Second, the iterative procedure used to estimate the magnitude of correlations can exceed value 1; this may cause poor approximations if one tries to remedy the problem by setting up the threshold value to 1 or -1 for the estimated correlations; these series of approximations may reduce the correlation estimation accuracy quite significantly. SPIEC-EASI [8] calculates the covariance of the log-ratio transformed data to approximate the covariance of the absolute abundance of microbes; then, it uses either neighborhood selection (mb) [9] or graphical Lasso (gl) [10] to estimate the conditional dependencies among microbes. CCLasso [11] is similar to SPIEC-EASI which applies log-ratio transform on compositional data and imposes a  $l_1$  penalty on the inverse covariance matrix of the microbes and then solves it to obtain a sparse covariance matrix.

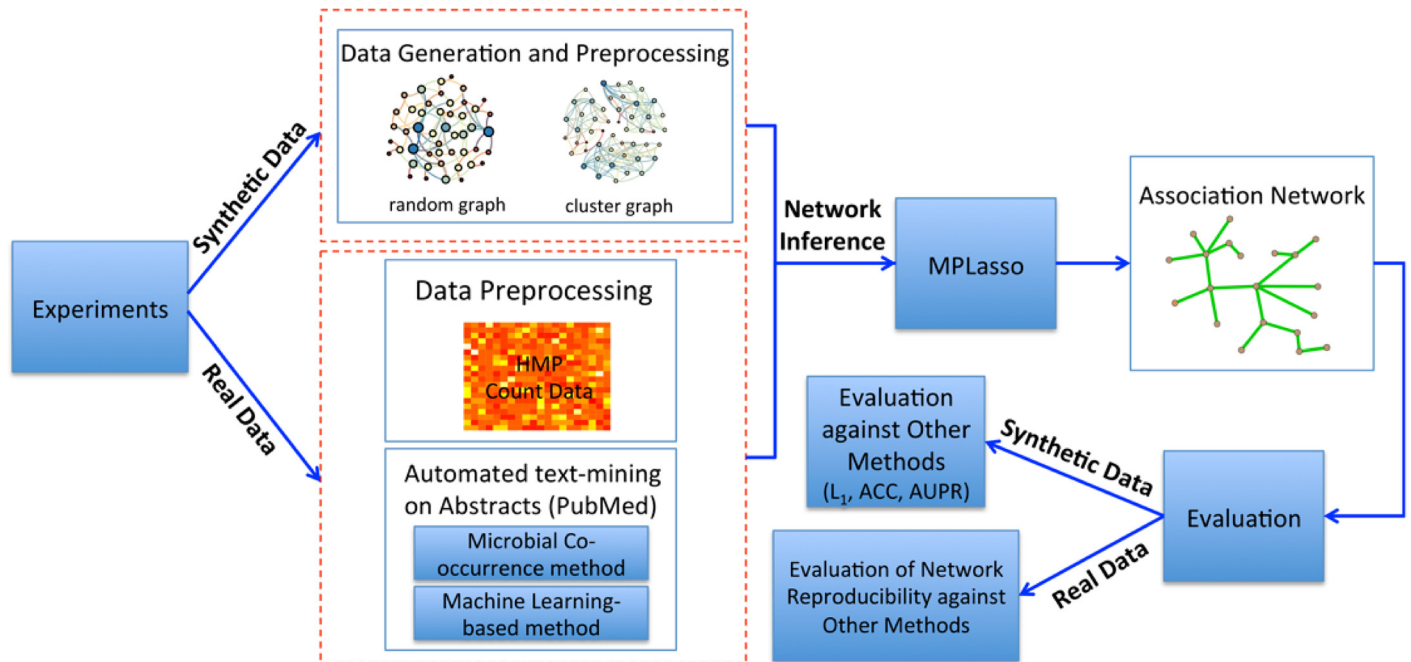
However, it is not clear whether or not CCLasso can obtain a consistent estimator on the inferred microbial covariance without showing consistency analysis (see consistency analysis for graphical Lasso in [S1 File](#) section 8).

We note that although the above methods can estimate the covariance among microbes under the sparsity assumption, they still have major difficulties to infer the associations among microbes given such high-dimensional data. To solve the problem caused by high-dimensional data, we propose to integrate multiple levels of biological information to enhance the model accuracy on inferring microbial associations. Indeed, an increasing amount of scientific literature provides a large amount of data which can be mined not only for the co-occurrence of microbes, but also to predict microbes associations directly. For instance, pioneering work [12] considers automated analysis of the co-occurrence of bacterial species through statistical testing approaches (e.g., Fisher's exact test). Recently, Lim et al. [13] incorporated machine learning techniques to automatically identify and extract microbial associations directly from the abstracts of scientific papers. Finally, Wang et al. [14] and Li et al. [15] use prior biological knowledge to reconstruct genes interaction networks.

To the best of knowledge, we are the first to consider experimentally verified biological knowledge as *a priori* information to derive microbial association networks. To this end, we transform the original problem of microbial associations estimation into a graph structure learning problem where nodes represent microbes and edges represent (pairwise) associations among microbes. With this new problem formulation, the graphical Lasso algorithm becomes suitable to infer the microbial association network. We also integrate the text mining results from the scientific literature as prior knowledge for inferring the microbes graph structure; the proposed algorithm *Microbial Prior Lasso (MPLasso)* turns out to be more accurate than other existing methods on inferring the microbial associations. The proposed MPLasso pipeline is shown in [Fig 1](#).

We assess the performance of MPLasso in the presence of prior knowledge by first comparing it against other previously proposed methods (e.g., CCLasso, REBACCA [16], SparCC, SPIEC-EASI, and CCREPE [17]) through synthetic data generated from different graph structures (run time comparisons of existing methods are summarized in [S1 File](#) section 1 and [S1 Table](#)). We show that our proposed MPLasso outperforms all these methods in terms of area under the precision-recall curve (AUPR) and accuracy (ACC) of network associations prediction. Next, we evaluate the HMP datasets of two different sequencing techniques (shotgun and 16S ribosomal RNA (rRNA)) at five different body sites and compare the reproducibility of the estimated results. Taken together, our contributions are three fold:

1. First, we integrate the graph learning algorithm (graphical Lasso) with *a priori* knowledge by mining the co-occurrence of microbes in literature. To the best of our knowledge, we are the first to integrate these two different algorithms to infer the association networks among microbes.
2. Second, we show that our proposed method, MPLasso in the presence of prior knowledge, outperforms all other previously proposed methods for inferring graph structures on different synthetic datasets. Additionally, the MPLasso accuracy on edge recovery rate is up to 95%, on average.
3. Third, we show that MPLasso can robustly and accurately estimate the associations among microbes with a reproducibility up to 90%. Additionally, the associations found with our approach correlate well with the experimental findings reported in the scientific literature.



**Fig 1. Our proposed framework of inferring microbial association network.** We conduct two different sets of experiments, namely, synthetic and real data. For the synthetic experiment, we generate data based on different graph structures and evaluate the performance of our proposed algorithm by using three performance metrics (i.e.,  $L_1$ , ACC, and AUPR). For the real data experiments, the prior information is obtained through automated text-mining. Since there is no “gold standard” network to evaluate performance, we evaluate the reproducibility of inferred networks instead.

<https://doi.org/10.1371/journal.pcbi.1005915.g001>

## Materials and methods

### Acquisition and transformation of microbial count data

In this paper, we consider high-throughput comparative metagenomic data obtained from the next-generation sequencing (NGS) platforms. More specifically, two types of gene sequencing data are considered: 16S rRNA and shotgun data. Shotgun data analyses are accomplished by unrestricted sequencing of the genome of all microorganisms present in a sample; on the contrary, the domain of 16S rRNA is restricted to bacteria and archaea. Data obtained from the human microbiome project (HMP) have a curated collection of sequence of microorganisms associated with the human body from both shotgun and 16S sequencing technologies.

For the 16S rRNA data, we consider the high-quality sequencing reads in 16S variable regions 3-5 (V35) of HMP healthy individuals from Phase one production study (May 1, 2010). The taxonomy classification of the 16S rRNA are performed using either mothur (HMMCP) [18] or QIIME (HMQCP) [19] pipelines. The resulting table for operational taxonomic units (OTUs) at each body site of the human samples can be obtained from <http://hmpdacc.org/HMMCP/> and <http://hmpdacc.org/HMQCP/>. For the shotgun data (HMASM), we obtain data from <http://hmpdacc.org/HMASM/> and use the trimmed sequences as inputs to the metaphlan2 [20] pipeline which can generate the OTU abundance for each sample.

The OTU table can be represented by a matrix  $\mathbf{D} \in \mathbb{N}^{n \times p}$  where  $\mathbb{N}$  represents the set of natural numbers.  $d^i = [d_1^i, d_2^i, \dots, d_p^i]$  denotes the  $p$ -dimensional row vector of OTU counts from the  $i_{th}$  sample ( $i = 1, \dots, n$ ). To account for different sequencing depths for each sample, the raw count data ( $d^i$ ) are typically transformed into *relative* abundances ( $x$ ) by using log-ratio transform [6]. Statistical inference on the log-ratio transform of the compositional data ( $x$ ) can

be shown to be equivalent to the log-ratio transform on the unobserved absolute abundance ( $d$ ) as:  $\log\left(\frac{x_i}{x_j}\right) = \log\left(\frac{d_i/m}{d_j/m}\right) = \log\left(\frac{d_i}{d_j}\right)$ . Here, we apply the centered log-ratio ( $clr$ ) transform as follows:

$$c = clr(x) = \left[\log\left(\frac{x_1}{m(x)}\right), \log\left(\frac{x_2}{m(x)}\right), \dots, \log\left(\frac{x_p}{m(x)}\right)\right] \quad (1)$$

where  $m(x) = \left(\prod_{i=1}^p x_i\right)^{\frac{1}{p}}$  is the geometric mean of the composition vector  $x$ . The resulting vector  $c$  is constrained to be a zero sum vector.

The covariance matrix of the  $clr$  transform  $C = Cov[clr(c)]$  can be related to the covariance matrix of the log-transformed absolute abundances  $\Gamma = Cov[\log \mathbf{D}]$  via the relationship [6, 8]  $C = \mathbf{U}\Gamma\mathbf{U}$ , where  $\mathbf{U} = \mathbf{I}_p - \frac{1}{p}\mathbf{J}$ , where  $\mathbf{I}_p$  is the  $p$ -dimensional identity matrix, and  $\mathbf{J}$  is the  $p$ -dimensional all-ones vector. For the case where  $p \gg 0$ , the finite sample estimator ( $\hat{C}$ ) serves as a good approximation of  $\hat{\Gamma}$ ; therefore, the finite sample estimator ( $\hat{C}$ ) serves as the basis on inferring the correlations among microbes. To account for the zero counts in samples, we add pseudo count to the original count data to avoid numerical issues when using the  $clr$  transform.

### Microbial Prior Lasso (MPLasso)

To infer the pairwise associations among microbes, we can transform the original inferring problem into a graph learning problem where each node represents an OTU (e.g., taxon) and each edge represents a pairwise association between microbes; the resulting graph is an undirected graph  $\mathcal{G} = (V, E)$ , where  $V$  and  $E$  represent the node and edge sets, respectively.

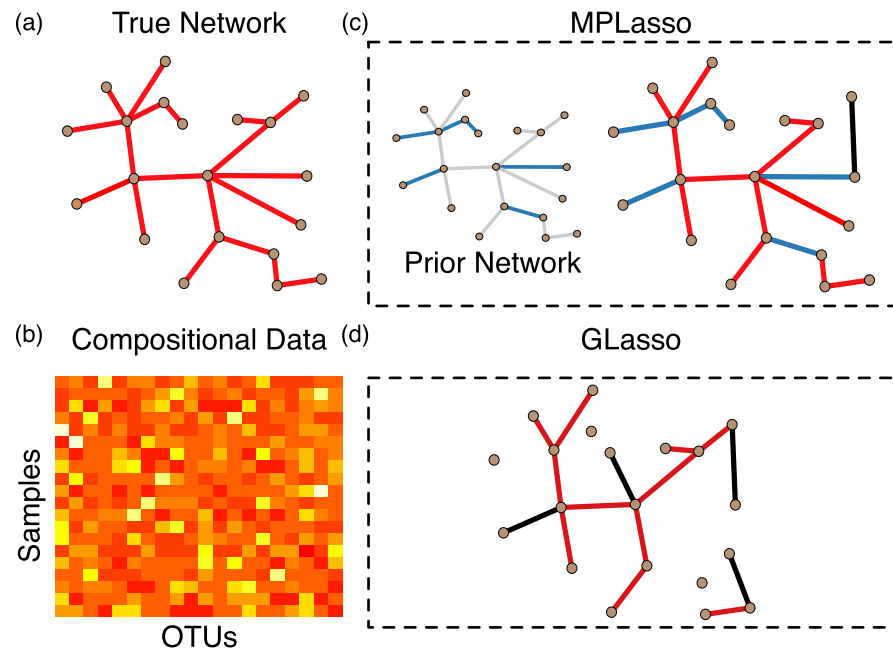
Suppose the observed data ( $d$ ) are drawn from a multivariate normal distribution  $N(d|\mu, \Sigma)$  with mean  $\mu$  and covariance  $\Sigma$ . The inverse covariance matrix (precision matrix)  $\Omega = \Sigma^{-1}$  encodes the conditional independence among nodes. More specifically, if the entry ( $i, j$ ) of the precision matrix  $\Omega_{i,j} = 0$ , then node  $i$  and node  $j$  are conditionally independent (given the other nodes) and there is no edge among them (i.e.,  $E_{i,j} = 0$ ).

However, microbial data usually come with a finite amount of samples ( $n$ ) but with high dimensionality ( $p$ ); this makes the graph inferring problem intractable since the number of variables  $\binom{p(p-1)}{2}$  is greater than  $n$ . To solve this problem, an important assumption that needs to be made is to assume that the underlying (true) graph is reasonably sparse. One suitable algorithm to select the precision matrix under sparsity assumption is to utilize the graphical Lasso proposed previously [8, 10].

As shown in Fig 2, we propose to utilize the information obtained from the scientific literature in order to construct the prior matrix  $\mathbf{P} \in \mathbb{R}^{p \times p}$ , where each entry  $\mathbf{P}_{i,j} \in [0, 1]$  represents the prior probability of associations between taxon  $i$  and taxon  $j$ . We can impose different amounts of penalties on the precision matrix; this is different from the standard formulation where the penalty ( $\rho$ ) imposed on the precision matrix is the same. Therefore, by incorporating the prior information into the penalty matrix ( $\mathbf{P}$ ), the proposed MPLasso can be formulated as follows:

$$\hat{\Omega} = \arg \max_{\Omega} \{ \log \det(\Omega) - \text{tr}(\Omega \hat{C}) - \rho |\mathbf{P} \otimes \Omega|_1 \} \quad (2)$$

where  $\hat{C}$  is the empirical covariance of the microbial data, and  $\Omega$  is the precision matrix of the estimated associations among microbes. Here  $\det$  and  $\text{tr}$  denote the determinant and the trace of a matrix, respectively.  $|\Omega|_1$  is the  $L_1$  norm, i.e., the sum of the absolute values of the elements of  $\Omega$  and  $\otimes$  represents the component-wise multiplication. When the value of  $\mathbf{P}_{i,j}$  is large, this



**Fig 2. Comparison of our proposed MPLasso and graphical Lasso (GLasso) on inferring the same compositional data in a small example.** (a) The edges of the true network are shown with red lines. (b) The entities of the compositional data matrix shown with denser colors represent higher values (c) Given the prior network where blue and black edges are correct and wrong information, respectively, the MPLasso can still accurately estimate the graph structure with one missing edge and only one wrongly estimated edge (black edge). (d) GLasso wrongly estimates several edges along with missing edges.

<https://doi.org/10.1371/journal.pcbi.1005915.g002>

directly puts a heavy penalty and represents a weaker association between taxa and vice versa. This way, by imposing the prior information, we can accurately infer the associations among microbes.

### Automated text-mining of microbial associations from literature

We extract two types of data to be used as priors for our model. One type of data is from the microbial co-occurrence in literature that examines the number of abstracts where two taxa appear together and compares this to random chance. The second type of data is from the machine learning-based method that extracts the full details of the interaction, including the sign and direction of the interaction.

To acquire the prior knowledge ( $\mathbf{P}$ ) of microbial associations from reported experiments and published papers, we utilize the PubMed database (<https://www.ncbi.nlm.nih.gov/pubmed/>) that contains a massive amount of papers with abstracts. For the 16S rRNA data where the taxonomy level can only be achieved at the genus level, we adopt the statistical testing method (i.e., Fisher's exact) [12] to identify the pairwise associations derived from the microbial co-occurrence in literature. On the other hand, for the shotgun data where the taxonomy level can be up to species level, we adopt both the microbial co-occurrence in literature and the machine-learning-based methods [13] to obtain such associations.

We modify the code available on <https://github.com/CSB5/atminter> that utilizes the Entrez search system to query all the possible combinations of taxon-taxon pairs from the data. More specifically, the query "taxon  $i$  AND taxon  $j$ " for genus (species) level are performed on PubMed database in order to obtain the *number* of papers that corresponds to this query term.



Acquisitions of abstract's *content* follow a similar way where the query term follows the format "species *i* AND species *j*" for each pair of species. Note that, all text-mining procedures are completely automated; that is, users only need to specify the species pairs and the tool will extract the information automatically (and comprehensively) from the PubMed database.

**Microbial co-occurrence in scientific literature.** We use Fisher's exact test, which only requires the *number* of abstracts, to examine the microbial co-occurrence in scientific literature. For example, the query "taxon *i* AND taxon *j*" returns four numbers: (1)  $n_i$ : the number of abstracts that contains only taxon *i*, (2)  $n_j$ : the number of abstracts that contains only taxon *j*, (3)  $n_{ij}$ : the number of abstracts that contain both taxa *i* and *j*, and (4)  $M$ : the number of abstracts that contain neither taxa *i* and *j* (see [S6 Table](#)). Next, by creating a 2-by-2 contingency table using the above four numbers (see [S7 Table](#) for an example), Fisher's exact test can be used to examine the probability that the number of abstracts where two taxa co-appear occurs at a higher rate than chance expectation. Note that we use the Bonferroni correction [21] to correct the *p*-value in order to deal with large amounts of candidate associations from the Fisher's exact test.

If taxa pair  $\langle i, j \rangle$  is rejected by the alternative hypothesis with high statistical significance (i.e., calculated *p*-value < 0.001), we put a larger penalty on entry  $(i, j)$  of the prior matrix  $\mathbf{P}$ . This way, we narrow down the solution space for candidate association pairs (see [S10 Table](#)); MPLasso can effectively select the associations from these candidate pairs within this restricted space. In this respect, prior information will not dominate the results, but rather improve the algorithm's accuracy and robustness.

**Machine learning-based approach.** In [13], the authors train the support vector machine [22] based on the manually curated abstracts and classify interactions into three categories: positive, negative, and no interaction. We use the pre-trained model provided by [13] to classify the abstracts of the species pairs obtained from the PubMed database. For example, for the (*Streptococcus mitis*, *Actinomyces naeslundii*) query, we obtain 65 abstracts that contain both taxa names. By concatenating these abstracts into a single file, the pre-trained classifier is able to classify this pair as either interacting or non-interacting. More specifically, if species pair  $\langle i, j \rangle$  is classified as interacting, then we put a smaller penalty on entry  $(i, j)$  of the prior matrix  $\mathbf{P}$ . In this respect, the species pair  $\langle i, j \rangle$  is more likely to be selected by MPLasso. Note that these experimentally validated interactions take precedence over (and we effectively ignore) the prior information obtained from the Fisher's exact test.

## Model selection

To select the optimal penalty parameter ( $\rho$ ), we use the Bayesian information criterion (BIC) [23] which is a standard method for model selection. The BIC for Gaussian graphical models takes the form:

$$BIC = -2l_n(\Omega) + |E| \log(n) \quad (3)$$

where  $|E|$  is the number of edges in the association network,  $n$  is the sample size, and  $l_n(\Omega) = \frac{n}{2} [\log(\det(\Omega)) - \text{tr}(\Omega \hat{C})]$ . Based on (3), we choose  $\rho$  that minimizes BIC.

## Results

### Experiments with synthetic data

To show the effectiveness of our proposed model, we first compare our model against several state-of-the-art models: CCREPE, SparCC, REBACCA, CCLasso, SPIEC (mb) and SPIEC (gl). All these codes have been implemented using the R language. We set up *p*-value at 0.05 for

CCREPE and the threshold of correlation for SparCC at 0.1 (see [S1 File](#) section 1 for precise simulation settings for each algorithm).

For MPLasso in real datasets, the true underlying network is only partially known and contains spurious information. To assess our algorithm performance with imperfect prior information, we consider prior information with different precision levels, where the precision level is defined as the number of true edges over the total number of edges in the prior information. The total number of edges in the prior network is set to be equal to the number of edges in the true underlying network. Therefore, a precision level of 0.1 indicates that 10% of the edges in the prior network are true edges, whereas the other 90% are spurious ones (see [S1 File](#) section 7 for details of introducing priors). We report the results we obtained for 0.5 precision level in the synthetic experiments while more results for different precision levels can be found in [S1 File](#) section 3 and [S2 Fig](#).

**Data generation.** We simulate the compositional data from the additive log normal distribution with a given mean and covariance matrix  $\ln d \sim N(\mu, \Sigma)$ ,  $x_i = \frac{d_i}{\sum_{i=1}^p d_i}$ , where  $\mu$  and  $\Sigma$  represent the mean and covariance, respectively;  $d$  is the sample generated from a multivariate logarithm normal distribution, and  $x$  is a compositional vector. To evaluate the performance of our model to recover different network structures, we report three representative network structures: cluster, band(4), and scale-free graph in [Table 1](#), and two other graphs (random and hub) in [S2 Table](#). Different sparsities on graph structure can strongly affect network recovery, and thus the network topologies we reported span a range of sparsity where band(4) is the least sparse followed by cluster and scale-free graphs.

We use the package in [\[24\]](#) to generate the precision matrix ( $\Theta$ ) and the positive definite covariance matrix  $\Sigma = \Theta^{-1}$  for each graph (see [S1 File](#) section 2 and [S1 Fig](#)). The covariance matrix is then computed to generate multivariate normal samples ( $d$ ). Since the number of samples can be around the same order as the number of OTU in real datasets, we generate a small number of samples to evaluate the performances of MPLasso and other methods. More specifically, we evaluate 6 different combinations, namely, ( $p = 50, n = (50, 100, 200)$ ) and ( $p = 100, n = (100, 200, 400)$ ). For each combination, we simulate 100 runs and calculate the mean value and standard deviation for all performance metrics. In addition to additive log normal distribution, we have included a new set of experiments to show that our proposed algorithm is able to deal with zero-inflated distributions. More precisely, we choose the negative binomial distribution that is suggested to be more suitable to model the microbial count data [\[25\]](#). Same experimental setting of parameters (as the additive log normal distribution) are considered, and experimental results are presented in [S6](#) and [S7 Figs](#), [S4](#) and [S5 Tables](#).

**Performance metrics.** We consider three different metrics as follows:

- Area Under the Precision-Recall Curve (AUPR): We compute the AUPR and ignore the sign of inferred edges. Precision is defined as the number of true positives, divided by the sum of true and false positives, while Recall is defined as the number of true positives, divided by the sum of true positive and false negatives.
- Accuracy (ACC): We estimate ACC as the number of true positives plus the true negatives, divided by total number of pairwise correlations.
- $L_1$  distance: The  $L_1$  distance is defined as the difference between estimated and true values. More specifically,  $L_1 = |\mathbf{R} - \hat{\mathbf{R}}|$ , where  $\mathbf{R}$  is the true correlation matrix and  $\hat{\mathbf{R}}$  is the estimated correlation matrix.



**Table 1. Performance comparison of different methods for additive log normal model.**

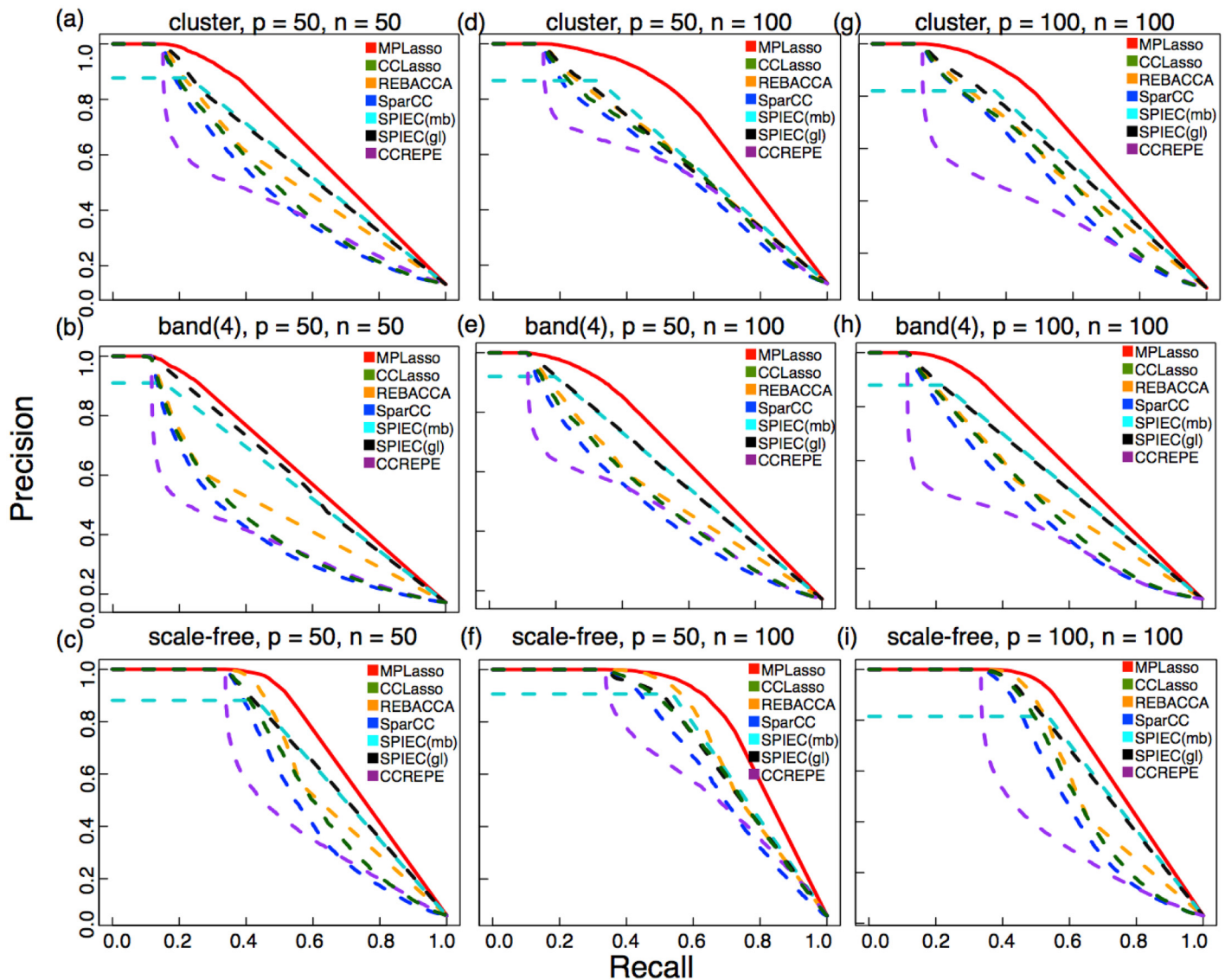
Method	$L_1$	ACC	AUPR	$L_1$	ACC	AUPR	$L_1$	ACC	AUPR
Cluster Graph									
MPLasso	0.059 (0.005)	<b>0.911 (0.010)</b>	<b>0.682 (0.024)</b>	0.052 (0.004)	<b>0.926 (0.009)</b>	<b>0.748 (0.029)</b>	0.028 (0.002)	<b>0.959 (0.004)</b>	<b>0.692 (0.023)</b>
CCLasso	0.080 (0.008)	0.893 (0.008)	0.526 (0.029)	0.068 (0.004)	0.903 (0.008)	0.614 (0.026)	0.053 (0.005)	0.950 (0.003)	0.562 (0.027)
SparCC	0.083 (0.004)	0.892 (0.009)	0.507 (0.028)	0.069 (0.003)	0.899 (0.010)	0.590 (0.030)	0.053 (0.002)	0.949 (0.003)	0.533 (0.027)
REBACCA	<b>0.055 (0.005)</b>	0.896 (0.010)	0.572 (0.027)	<b>0.042 (0.003)</b>	0.905 (0.010)	0.629 (0.031)	<b>0.025 (0.001)</b>	0.950 (0.004)	0.583 (0.027)
SPIEC (mb)	-	0.893 (0.010)	0.591 (0.030)	-	0.901 (0.012)	0.615 (0.030)	-	0.952 (0.004)	0.581 (0.026)
SPIEC (gl)	0.064 (0.006)	0.894 (0.010)	0.607 (0.024)	0.063 (0.006)	0.900 (0.011)	0.630 (0.024)	0.030 (0.003)	0.952 (0.004)	0.615 (0.026)
CCREPE	0.123 (0.011)	0.887 (0.009)	0.471 (0.022)	0.123 (0.011)	0.892 (0.009)	0.567 (0.025)	0.060 (0.005)	0.943 (0.003)	0.436 (0.022)
Band Graph									
MPLasso	0.093 (0.002)	<b>0.867 (0.007)</b>	<b>0.654 (0.018)</b>	0.087 (0.005)	<b>0.887 (0.007)</b>	<b>0.694 (0.019)</b>	0.048 (0.001)	<b>0.939 (0.002)</b>	<b>0.654 (0.013)</b>
CCLasso	0.092 (0.006)	0.853 (0.003)	0.468 (0.018)	<b>0.074 (0.004)</b>	0.863 (0.005)	0.551 (0.024)	0.062 (0.003)	0.929 (0.002)	0.506 (0.015)
SparCC	<b>0.087 (0.003)</b>	0.852 (0.003)	0.452 (0.020)	0.077 (0.003)	0.858 (0.004)	0.523 (0.019)	0.058 (0.001)	0.927 (0.001)	0.476 (0.015)
REBACCA	0.093 (0.002)	0.854 (0.004)	0.520 (0.027)	0.080 (0.002)	0.865 (0.005)	0.576 (0.024)	<b>0.044 (0.001)</b>	0.930 (0.002)	0.537 (0.016)
SPIEC (mb)	-	0.851 (0.004)	0.597 (0.039)	-	0.858 (0.007)	0.619 (0.025)	-	0.929 (0.002)	0.571 (0.020)
SPIEC (gl)	0.096 (0.000)	0.850 (0.004)	0.617 (0.027)	0.096 (0.000)	0.856 (0.007)	0.629 (0.016)	0.050 (0.000)	0.928 (0.002)	0.588 (0.013)
CCREPE	0.167 (0.004)	0.848 (0.000)	0.427 (0.017)	0.170 (0.003)	0.851 (0.004)	0.504 (0.019)	0.089 (0.001)	0.922 (0.000)	0.391 (0.012)
Scale-free Graph									
MPLasso	<b>0.066 (0.008)</b>	<b>0.970 (0.003)</b>	<b>0.750 (0.027)</b>	0.065 (0.008)	<b>0.976 (0.004)</b>	<b>0.817 (0.039)</b>	<b>0.033 (0.004)</b>	<b>0.985 (0.001)</b>	<b>0.758 (0.024)</b>
CCLasso	0.077 (0.008)	0.964 (0.002)	0.620 (0.046)	0.071 (0.010)	0.969 (0.004)	0.740 (0.063)	0.046 (0.005)	0.983 (0.001)	0.641 (0.041)
SparCC	0.078 (0.006)	0.963 (0.001)	0.594 (0.038)	0.067 (0.006)	0.967 (0.003)	0.697 (0.046)	0.050 (0.002)	0.982 (0.001)	0.610 (0.040)
REBACCA	0.069 (0.008)	0.966 (0.003)	0.668 (0.046)	<b>0.064 (0.010)</b>	0.973 (0.004)	0.758 (0.046)	0.034 (0.004)	0.984 (0.001)	0.673 (0.030)
SPIEC (mb)	-	0.962 (0.003)	0.646 (0.049)	-	0.969 (0.005)	0.710 (0.055)	-	0.982 (0.002)	0.630 (0.063)
SPIEC (gl)	0.068 (0.007)	0.963 (0.003)	0.695 (0.024)	0.069 (0.008)	0.969 (0.004)	0.747 (0.034)	<b>0.033 (0.004)</b>	0.983 (0.001)	0.712 (0.026)
CCREPE	0.072 (0.004)	0.961 (0.000)	0.549 (0.030)	0.070 (0.004)	0.962 (0.001)	0.660 (0.040)	0.035 (0.002)	0.980 (0.000)	0.515 (0.028)

We consider three different graph structures and three sets of parameters, namely, ( $p = 50, n = 50$ ), ( $p = 50, n = 100$ ), and ( $p = 100, n = 100$ ). For each experiment, we average over 100 simulation runs with standard deviations in round brackets. We use three metrics ( $L_1$ , ACC, AUPR) to quantify the performance as defined in Performance metrics. Bold numbers show best results.

<https://doi.org/10.1371/journal.pcbi.1005915.t001>

**Performance comparisons.** We report ( $p = 50, n = 50$ ), ( $p = 50, n = 100$ ), and ( $p = 100, n = 100$ ) in Table 1. For completeness, more experimental results are available in S2 and S3 Tables, S3 and S4 Figs. For  $L_1$  distance, all the methods are evaluated on the correlation matrix. For ACC and AUPR, in order to have a fair comparison among different methods, the microbial associations for correlation (covariance) based method (i.e., SparCC, CCREPE, REBACCA, and CCLasso) are obtained from the inferred microbial correlation (covariance), while for precision based methods (i.e., SPIEC (mb), SPIEC (gl), and MPLasso) is obtained from the inferred precision matrix. As it can be seen in both Fig 3 and Table 1, our proposed method (MPLasso) achieves the best AUPR on all the cases; this confirms that MPLasso can accurately identify associations among microbes. However, the  $L_1$  distance is greater than REBACCA due to the fact that MPLasso directly estimates the precision matrix of microbial associations, not on the correlation matrix.

As we increase the OTU numbers and fix the sample size, the performance for all methods degrades. For the case where ( $p = 100, n = 100$ ), MPLasso still outperforms all other methods in terms of ACC and AUPR. On the other hand, as we vary the sample size from 50 to 100 and fix the number of OTU to 50, the performance of all the metrics for MPLasso increases, as expected. When sample size equals 100, which is often the case in practice (e.g., HMP dataset),



**Fig 3. AUPR curves of different methods for additive log normal model.** Each set of experiment are averaged over 100 simulations. We compare three different sets of sample size ( $n$ ) and OTU numbers ( $p$ ) for three different graph structures. For ( $p = 50, n = 50$ ), (a) cluster, (b) band(4), and (c) scale-free. For ( $p = 50, n = 100$ ), (d) cluster, (e) band(4), and (f) scale-free. For ( $p = 100, n = 100$ ), (g) cluster, (h) band(4), and (i) scale-free. As can be seen, the MPLasso (red curve) performs better than all other methods.

<https://doi.org/10.1371/journal.pcbi.1005915.g003>

MPLasso can achieve outstanding performance in terms of both *average* ACC and AUPR (0.93 and 0.75, respectively). Also, when sample size equals 200 and 400 (see [S3 Table](#)), MPLasso can near-perfectly recover the network (i.e.,  $AUPR \approx 1$ ).

As shown in [Fig 3](#), we can see that most of the algorithms can achieve high precisions under low recalls, which means that they can accurately estimate the true edges. However, as the number of recalls increases, only MPLasso can still achieve high precision when comparing with other methods; this shows that MPLasso can recover edges with very low errors. Additionally, all methods show dependence on different graph structures; this is due to different sparsity of a particular type of graph encodes. Since scale-free graph is less sparse than band(4) and cluster graph, all methods achieve better performance in inferring edges. Additionally,

even when precision level is as low as 0.1 (i.e., only 10% of the edges in prior information are true edges, whereas the other 90% are spurious ones), MPLasso can still achieve up to an average 0.65 in AUPR for the case where ( $p = 50, n = 50$ ) (see S2 Fig).

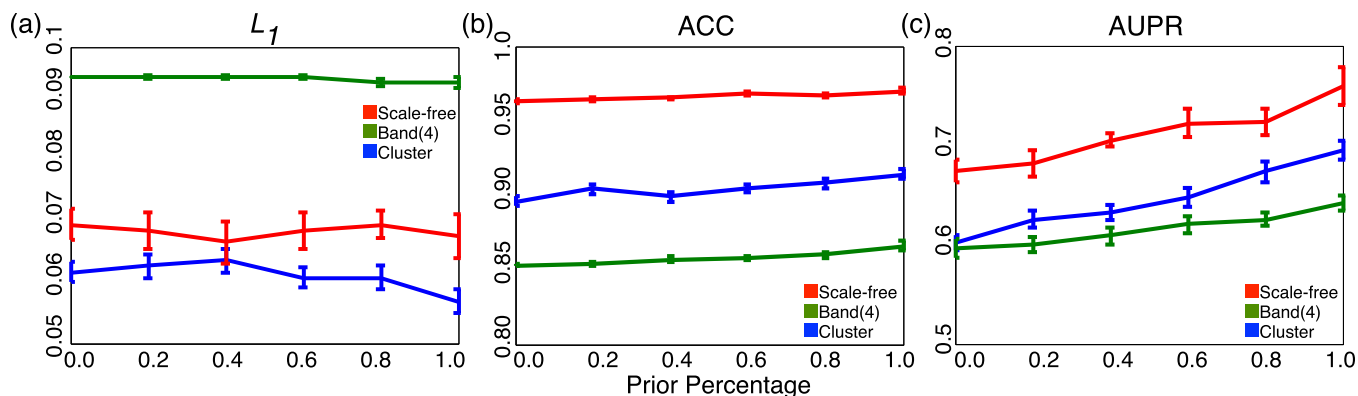
Since MPLasso is able to infer the sign of the edge, we also report the performance of different algorithms on accuracy of edge sign recovery, which is defined as the number of correctly inferred edge signs over the total number of inferred edges. As shown in S13 Table, all existing algorithms (including MPLasso) can successfully achieve a sign recovery accuracy above 0.9, on average. However, although algorithms have similar performance on edge sign recovery accuracy, MPLasso shows better performance in edge recovery accuracy (i.e., AUPR and ACC).

In addition to examining the impact of different precision levels, we also quantify the effect of different amount of prior information being used in the synthetic experiments for three graph structures (cluster, band(4), and scale-free graphs). As shown in Fig 4, if the amount of prior information increases, then the performance in terms of AUPR increases too, as expected. For  $L_1$ , which is evaluated based on correlations, different amounts of prior information have little effect due to the fact that MPLasso directly estimates the precision matrix. For ACC, since MPLasso has already achieved high performance, increasing the amount of prior information only brings a small increment of improvement in performance. When prior percentage = 100%, AUPR achieves around 20% improvements over the case without using any prior information. For cases without using any prior information, MPLasso can still achieve comparable results with other existing methods presented in Fig 3 and Table 1.

For the zero-inflated distribution (discussed in S1 File section 4 and S5 Fig), as it can be seen in S6 and S7 Figs, S4 and S5 Tables, the performance of our proposed method outperforms all the other methods except a few cases involving hub graphs; this is similar to the results for the additive log normal model. In summary, our results show that MPLasso works well with many different distributions and graph structures even in the cases with low precision levels and less prior information.

### HMP dataset

Emboldened by the success of our proposed algorithm on synthetic data, we have applied MPLasso to infer the associations among microbes for HMP data. Acquisitions and preprocessing for both 16S rRNA and shotgun sequencing data are described in **Material and methods** section. We report the same three body sites (i.e., buccal mucosa, supragingival plaque, and tongue dorsum) of each pipeline and filter out OTUs that appear in less than 10% of total



**Fig 4. The performance of different amount of prior information on three different graph structures. (a)  $L_1$  distance (b) ACC (c) AUPR.**

<https://doi.org/10.1371/journal.pcbi.1005915.g004>

samples—two more body sites (i.e., stool and anterior names) are reported in [S1 File](#) section 5 and [S8 Fig](#). The total number of samples and OTUs are summarized in [Table 2](#) and [S8 Table](#).

We use the *clr* transformation in (1) and add pseudo count 0.1 to all the samples, then normalize the counts to get compositional data. However, there is no true correlation network of taxon-taxon associations in real data as opposed to synthetic data. To assess and compare the performance among different methods in real data experiments, we measure the reproducibility of the resulting networks. More specifically, we define the “gold standard” network as the one that uses the full dataset. The reproducibility is defined as the number of edges that had been correctly estimated when using only *half* of the samples in the full dataset compared to the “gold standard” network. We randomly select half of the samples in the full dataset of each body site and then average over 20 independent simulations. We compare the reproducibility of the MPLasso against SPIEC (gl) which has a better performance than other existing algorithms on synthetic datasets as well as CCLasso which has a better performance than other correlation based methods in [11].

The reproducibility results are summarized in [Table 2](#). MPLasso has a better reproducibility over SPIEC (gl) and CCLasso; this implies that MPLasso is not only more robust, but also more accurate at inferring edges. We also consider reproducibility on different percentages of highly connected nodes in [S9 Table](#). Only when we consider as little as only 25% of high degree nodes, CCLasso has a better performance (but even so for 2% only, on average).

We also summarize the statistics of the non-associated pairs found by the Fisher’s exact test, potential associated pairs, associated pairs found by MPLasso, and recovered associated pairs in [S10 Table](#) (see also [S10 Fig](#) and [S1 File](#) section 9). As shown, the known associations obtained from Fisher’s exact test is around 50% over all possible pairs of associations (i.e., around 50% prior information). The recovery rate of associated pairs of MPLasso is around 80%. For comparison, we also include the recovery rate of associated pairs for CCLasso and SPIEC (gl) algorithms in [S11 Table](#). As we can see by comparing the [S10](#) and [S11](#) Tables, CCLasso tends to discover more edges than MPLasso and SPIEC (gl). Although CCLasso can obtain similar results on the recovery rate of associated pairs, it does not perform as well as MPLasso when considering the recovery rate (i.e., reproducibility) of both associated and non-associated pairs (see [Table 2](#)). In other words, CCLasso finds a greater amount of false positive

**Table 2. Reproducibility for MPLasso, SPIEC (gl), and CCLasso at different body sites of different types of HMP datasets.**

Body Site	( <i>n</i> , <i>p</i> )	MPLasso	SPIEC (gl)	CCLasso
<b>HMASM</b>				
BucMuc	(113, 73)	<b>0.963 (0.003)</b>	0.904 (0.013)	0.915 (0.005)
SupPla	(124, 129)	<b>0.942 (0.005)</b>	0.877 (0.009)	0.919 (0.005)
TonDor	(130, 103)	<b>0.948 (0.004)</b>	0.754 (0.030)	0.913 (0.015)
<b>HMMCP</b>				
BucMuc	(406, 74)	<b>0.923 (0.005)</b>	0.756 (0.033)	0.820 (0.014)
SupPla	(423, 84)	<b>0.923 (0.004)</b>	0.862 (0.007)	0.837 (0.012)
TonDor	(410, 77)	<b>0.934 (0.003)</b>	0.820 (0.014)	0.850 (0.012)
<b>HMQCP</b>				
BucMuc	(312, 75)	<b>0.876 (0.006)</b>	0.777 (0.023)	0.818 (0.011)
SupPla	(313, 51)	0.883 (0.010)	0.796 (0.015)	<b>0.896 (0.007)</b>
TonDor	(316, 45)	<b>0.860 (0.009)</b>	0.735 (0.020)	0.841 (0.024)

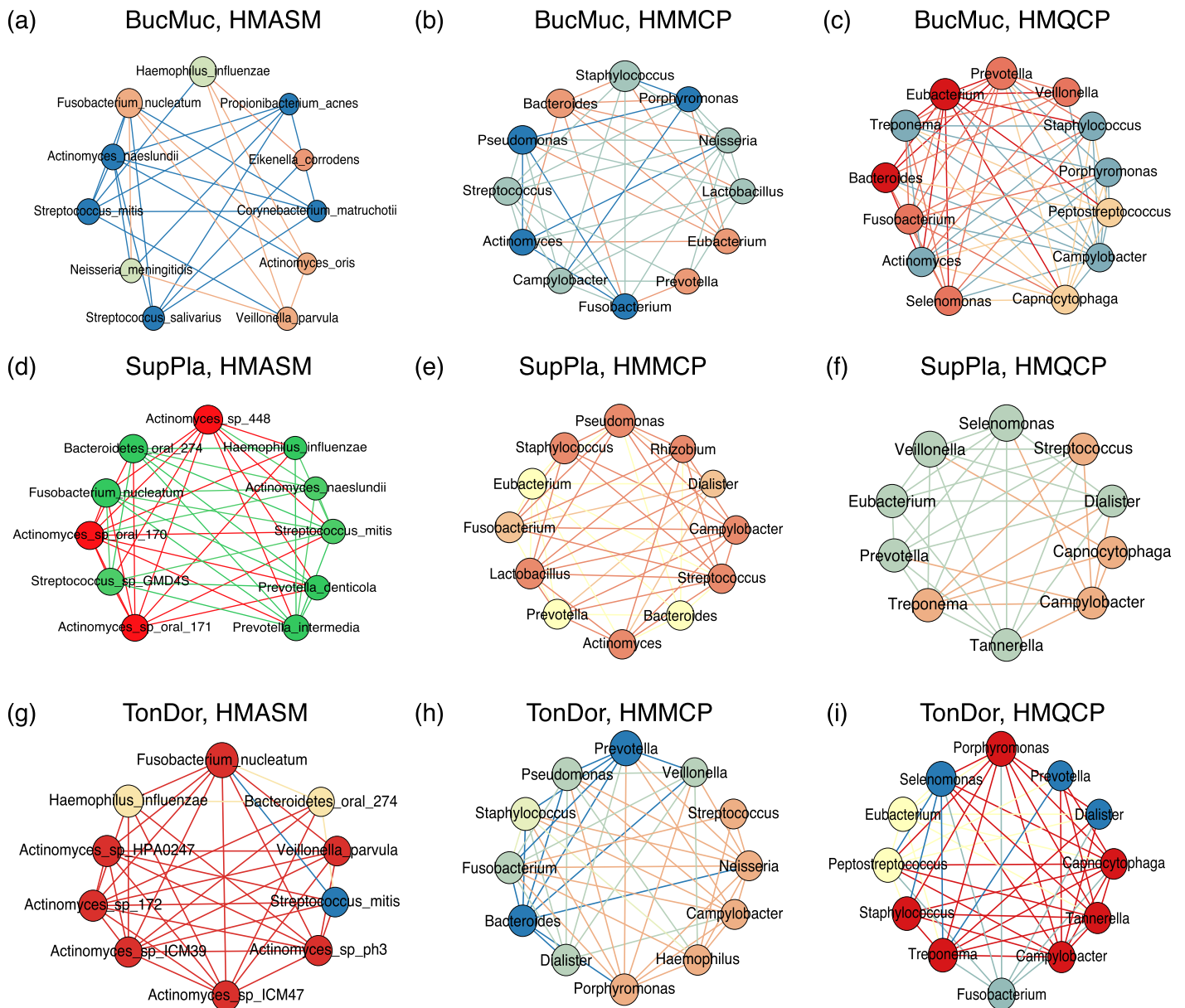
For each experiment, we average over 20 simulation runs with standard deviations in round brackets. Bold number shows best result. *n* and *p* represent sample size and taxa number, respectively.

<https://doi.org/10.1371/journal.pcbi.1005915.t002>



taxa pairs when compared to MPLasso; this is evaluated through AUPR in synthetic experiments shown in Fig 3.

To compare the estimated association networks at each body site for different pipelines (i.e., HMASM, HMMCP and HMQCP), we select the “top players” (i.e., high degree nodes) and arrange them using a counterclockwise layout as shown in Fig 5. For the genus level data, since we only utilize the Fisher’s exact test (that only requires the information of the *number* of abstracts), we can use *contents* of published scientific literature to validate the inferred



**Fig 5. Association network visualization of top degree nodes at different human body sites for different data types.** The same node colors represent the communities nodes belong to. For BucMuc: (a) HMASM, (b) HMMCP, and (c) HMQCP. For SupPla: (d) HMASM, (e) HMMCP, and (f) HMQCP. For TonDor: (g) HMASM, (h) HMMCP, and (i) HMQCP. As can be seen from species level data (HMASM), phylogenetically related OTUs fall in the same community. Node size represents the relative node degree within the association network with counterclockwise layout. The color of the edges is the same as the node color and does not have any special meaning. Abbreviations: BucMuc: Buccal mucosa, SupPla: Supragingival plaque, TonDor: Tongue dorsum.

<https://doi.org/10.1371/journal.pcbi.1005915.g005>

associations. In contrast, for the species level data, the machine learning-based approach has already used the contents of abstract to obtain the prior information; therefore, it is inappropriate to use any papers that appear in the PubMed search results to validate the inferred associations. To circumvent the potential circular validation, we only use the scientific literature that has *not* yet been used to create the prior information.

For the buccal mucosa (BucMuc), the association pair (*Streptococcus mitis*, *Actinomyces naeslundii*), which was found in HMASM (Fig 5(a)), has been shown to have associations [26]. Additionally, the associations are also detected at genus level data as shown in Fig 5(b). Note that the top degree nodes in HMMCP and HMQCP has 70% in common (i.e., belongs to same genus) which implies that the microbial composition of BucMuc is relatively robust.

For the supragingival plaque (SupPla), the “top players” in species level data (Fig 5(d)) mainly come from two genera: *Actinomyces* and *Prevotella* which can be widely found in SupPla and also correspond well with the HMMCP dataset (Fig 5(e)). Similarly, the species level associations in tongue dorsum (TonDor) is dominated by *Actinomyces* as shown in Fig 5(g); this is because *Actinomyces* possess 10 different strains out of the total 103 taxa, yet this does not imply that *all* members of a particular genus group should be associated. Although not seen in Fig 5(h) and 5(i), genus *Actinomyces* is also a high degree node in the association network of the genus data.

One noticeable observation in the species level dataset (HMASM) is that the same genus belongs to the same community which means that edges are mostly found within OTUs of the same taxonomic group. This phenomenon is called assortativity and it has been widely observed in other microbial network studies [17]. However, this does not imply that all members of the same taxon should be ecologically associated. To quantify the similarity of high degree nodes that are found both in HMMCP and HMQCP datasets, we compute the correlation between node degrees at different body sites by utilizing the Spearman correlation method (see S1 File section 6). We found that TonDor has lower correlations ( $\sim 0.5$ ) than other body sites ( $\sim 0.7$ ); this can be directly observed from Fig 5(h) and 5(i) that have a few high degree genera in common.

## Discussion

Inferring associations (putative interactions) among microbes and understanding their influence on the human body is an important step towards precision medicine. Advancements of high-throughput sequencing techniques enable us to gather metagenomic sequence data from different environment and human niches. The available high-throughput experimental data, however, are compositional and high-dimensional in nature.

Existing microbial network inferring methods focus on inferring the compositional data and use the graph sparsity assumption to overcome problems caused by high-dimensional data. However, all of these approaches do not consider the information that can be obtained from the scientific literature to directly describe the associations among microbes or their co-occurrence. By integrating multiple levels of biological information into the statistical models, we have shown that one can dramatically increase the model accuracy and edges recovery rate. To the best of knowledge, this is the first work to propose this automated pipeline to infer the associations on microbial data, show its feasibility, and measure performance metrics on both synthetic and real datasets.

We have also shown that our proposed algorithm *Microbial Prior Lasso* (MPLasso) outperforms all other existing methods when using synthetic data with different graph structures which simulate different levels of sparsity. We have evaluated several combinations of sample sizes and number of taxa to demonstrate the applicability of our approach under different



conditions and suggest rough guidelines for requisite sample size for the real data for the given assumption of the underlying graph structures.

Additionally, the use of prior information does not dominate the inferred results. Indeed, as summarized in [S10 Table](#), the prior information obtained by the microbial co-occurrence in literature is only used to restrict the search space in order to infer associations that are more plausible (i.e., more likely to be associated) than other candidate pairs of associations. More specifically, we first calculate the probability of association among taxa. Next, if two taxa are not associated, we penalize the associations among these two taxa when solving MPLasso. Consequently, MPLasso can effectively select taxa that are highly associated with high statistical confidence. In this respect, prior information will not dominate the results, but rather improve the algorithm's accuracy and robustness.

Our analyses on different levels of real HMP data show that MPLasso achieves better reproducibility than SPIEC (gl) and CCLasso; we have also found the assortativity at the species level data (HMASM) at different body sites. In other words, OTUs are more likely to interact with other phylogenetically related OTUs. Additionally, the detected genera at genus level (HMMCP and HMQCP) datasets show high correlations based on their node degrees (i.e., number of edges a node has to other nodes). Those high degree nodes (i.e., "top players") have been found experimentally as being ubiquitous at each body site; this confirms that MPLasso can accurately detect the "top players" and even correctly infer the associations among them. The resulting microbial association network can suggest credible directions for experimentalists to validate the results without exhausting search for all possible associations.

Recent studies report that people affected by microbiome related diseases show different microbiome profiles when compared to healthy individuals. For example, results show that individuals affected by the inflammatory bowel disease (IBD) have (30-50)% percent less biodiversity of commensal bacteria (e.g., *Firmicutes* and *Bacteroidetes*) when compared to healthy individuals. Another example shows that individuals with Type 2 diabetes (T2D) exhibit significant changes in 190 microbial OTUs, with particularly high abundance of *Enterobacteriaceae* compared to healthy individuals [27]. Therefore, by creating a more accurate microbial association network, scientists working in this field will be able to accurately identify the relationship between microbiome related diseases (such as T2D) and groups of taxa based on the inferred network. This way, scientists can develop new drugs or use probiotics to directly target identified groups of taxa.

Finally, the estimated microbial association networks of the real datasets can be used to understand why and how various eco-systems evolve over time. Recent studies use association networks to fit dynamic models, e.g., differential equation-based model of gut microbiome evolution of mice [28]. These microbe associations represent the putative microbial interactions that provide partial information about the true interaction network. Therefore, by incorporating the association network as additional information, we may be able to infer the microbial interaction networks more accurately [29]. Overall, MPLasso shows promising results and outperforms state-of-the-art methods. In the present framework, our proposed MPLasso creates the inferred association network to provide additional partial information; this can be useful to reveal the underlying dynamics (i.e., interactions) of microbial communities. However, MPLasso was not tested on a dynamic model of microbial communities. Inferring the dynamics or interactions among microbial communities would require a new algorithm which is left as future work.

## Software availability

The MPLasso R package can be downloaded from here [https://github.com/ChiehLo/MPLasso\\_RPackage](https://github.com/ChiehLo/MPLasso_RPackage)

## Supporting information

**S1 Fig. Different types of graph we considered to generate synthetic data.** (a) random (b) hub (c) cluster (d) band(4) and (e) scale-free graphs.  
(PDF)

**S2 Fig. Performance of AUPR of different precision levels.** Each point is averaged over 100 simulations. We compare 6 different sets of sample size and OTU numbers ( $(p = 50, n = (50, 100, 200))$  and  $(p = 100, n = (100, 200, 400))$ ).  
(PDF)

**S3 Fig. AUPR curves of different methods.** We compare three different sets of sample size and OTU numbers ( $(p = 50, n = 50)$ ,  $(p = 50, n = 100)$ , and  $(p = 100, n = 100)$ ). As can be seen the red curve (MPLasso) performs better than all other methods in random and hub graphs.  
(PDF)

**S4 Fig. AUPR curves of different methods.** Each set of experiment are averaged over 100 simulations. We compare three different sets of sample size and OTU numbers (i.e.,  $(p = 50, n = 200)$ ,  $(p = 100, n = 200)$ , and  $(p = 100, n = 400)$ ). As can be seen, the MPLasso (red curve) performs better than all other methods except the band(4) graph when  $(p = 100, n = 400)$ .  
(PDF)

**S5 Fig. The probability density distribution.** (a) real data (HMMCP, stool samples), (b) additive log-normal distribution, (c) negative binomial distribution.  
(PDF)

**S6 Fig. AUPR curves of different methods on negative binomial model.** Each set of experiment are averaged over 100 simulations. We compare three different sets of sample size and OTU numbers (i.e.,  $(p = 50, n = 50)$ ,  $(p = 50, n = 100)$ , and  $(p = 50, n = 200)$ ). As can be seen, the MPLasso (red curve) performs better than all other methods.  
(PDF)

**S7 Fig. AUPR curves of different methods on negative binomial model.** Each set of experiment are averaged over 100 simulations. We compare three different sets of sample size and OTU numbers (i.e.,  $(p = 100, n = 100)$ ,  $(p = 100, n = 200)$ , and  $(p = 100, n = 400)$ ). As can be seen, the MPLasso (red curve) performs better than all other methods.  
(PDF)

**S8 Fig. Association network visualization of top degree nodes at different human body sites for different data types.** The same node colors represent the communities nodes belong to. As can be seen from species level data (HMASM), phylogenetically related OTUs fall in the same community. Node size represents the relative node degree within the association network with counterclockwise layout. Abbreviations: AntNar: Anterior nares.  
(PDF)

**S9 Fig. Association network visualization for Global patterns dataset [30] for top 12 high degree taxa.** The Global Patterns dataset contains 26 environmental samples and 19216 OTUs. We first use the preprocessing criteria (i.e., OTU variance greater than  $10^{-5}$ ) to filter

out 32 species. Next, we obtain the prior information which contains 54 interacting taxa pairs from PubMed database. MPLasso finds 155 associated taxa pairs in total.

(PDF)

**S10 Fig. An illustration of (a) recovery rate of associated pairs and (b) text-mined pairs.**

(PDF)

**S1 Table. A comparison of correlation based methods adopted from [31].**

(PDF)

**S2 Table. Performance comparison of different methods for additive log normal model.**

We consider two additional graph structures (random and hub graph) and three sets of parameters, namely,  $(p = 50, n = 50)$ ,  $(p = 50, n = 100)$ , and  $(p = 100, n = 100)$ . For each experiment, we average over 100 simulation runs with standard deviations in round brackets. We use three metrics ( $L_1$ , ACC, AUPR) to quantify the performance. Bold number shows best result.

(PDF)

**S3 Table. Performance comparison of different methods for additive log normal model.**

We consider five different graph structures and three sets of parameters, namely,  $(p = 50, n = 200)$ ,  $(p = 100, n = 200)$ , and  $(p = 100, n = 400)$ . For each experiment, we average over 100 simulation runs with standard deviations in round brackets. Bold number shows best result.

(PDF)

**S4 Table. Performance comparison of different methods on negative binomial model.**

We consider five different graph structures and three sets of parameters, namely,  $(p = 50, n = 50)$ ,  $(p = 50, n = 100)$ , and  $(p = 50, n = 200)$ . For each experiment, we average over 100 simulation runs with standard deviations in round brackets. Bold number shows best result.

(PDF)

**S5 Table. Performance comparison of different methods on negative binomial model.**

We consider five different graph structures and three sets of parameters, namely,  $(p = 100, n = 100)$ ,  $(p = 100, n = 200)$ , and  $(p = 100, n = 400)$ . For each experiment, we average over 100 simulation runs with standard deviations in round brackets. Bold number shows best result.

(PDF)

**S6 Table. Entry for the 2-by-2 contingency table with the number of abstracts containing neither taxon A nor B in HMP datasets.** Abbreviations: AntNar: Anterior nares, BucMuc: Buccal mucosa, SupPla: Supragingival plaque, TonDor: Tongue dorsum.

(PDF)

**S7 Table. An illustration example of 2-by-2 contingency table that captures the information about how often two taxa appear together and separately.**

This example uses the taxa pair  $\langle i, j \rangle = \langle Escherichia, Citrobacter \rangle$ , where  $n_i$ : the number of abstracts that contains only taxon  $i$ ,  $n_j$ : the number of abstracts that contains only taxon  $j$ ,  $n_{ij}$ : the number of abstracts that contain both taxa  $i$  and  $j$ , and  $M$ : the number of abstracts that contain neither taxa  $i$  and  $j$ . For this particular contingency table, Fisher's exact test rejects the hypothesis that taxa  $i$  and  $j$  are associated. Therefore, these two taxa are *not* associated with high statistical significance (i.e., the calculated  $p$ -value  $< 0.001$ ) and a higher penalty is placed between taxa  $i$  and  $j$  in the prior matrix.

(PDF)

**S8 Table. Reproducibility for MPLasso, SPIEC (gl), and CCLasso at different body sites of different types of HMP datasets.** For each experiment, we average over 20 simulation runs

with standard deviations in round brackets. Bold number shows best result.  $n$  and  $p$  represent sample size and taxa number, respectively. Abbreviations: AntNar: Anterior nares.

(PDF)

**S9 Table. Different percentages of top degree nodes to calculate reproducibility for MPLasso, SPIEC (gl) and CCLasso at different body sites of different types of HMP datasets.** For each experiment, we average over 20 simulation runs with standard deviations in round brackets. Bold number shows best result. Abbreviations: AntNar: Anterior nares, BucMuc: Buccal mucosa, SupPla: Supragingival plaque, TonDor: Tongue dorsum.

(PDF)

**S10 Table. Prior information and the recovery rate of associated pairs found by MPLasso.**

(PDF)

**S11 Table. Recovery rate of associated pairs found by CCLasso and SPIEC (gl).**

(PDF)

**S12 Table. Interacting taxa pairs found by automated text-mining methods and association pairs suggested by MPLasso.**

(PDF)

**S13 Table. Accuracy of edge sign recovery accuracy for different algorithms on different synthetic graph structures.**

(PDF)

**S14 Table. Jaccard index of inferred edges among pairwise datasets.**

(PDF)

**S1 File. Supplementary text.** Contents: 1. Algorithms summaries, simulation settings and run time comparisons. 2. Graph generation process. 3. The impact of different precision levels on prior matrix and synthetic experiments. 4. Experiments with synthetic data generated from negative binomial distribution. 5. Experiments with HMP datasets with two more body sites. 6. Methods for calculating Spearman correlation of node degrees. 7. Prior knowledge introduction in synthetic experiment. 8. Consistency analysis for graphical Lasso algorithm. 9. Definitions for recovery rate of associated pairs and text-mined pairs.

(PDF)

## Acknowledgments

We thank Dr. Ryan Kim of CMU for valuable suggestions.

## Author Contributions

**Data curation:** Chieh Lo.

**Formal analysis:** Chieh Lo.

**Funding acquisition:** Radu Marculescu.

**Methodology:** Chieh Lo, Radu Marculescu.

**Software:** Chieh Lo.

**Supervision:** Radu Marculescu.

**Validation:** Chieh Lo.

**Visualization:** Chieh Lo.

**Writing – original draft:** Chieh Lo, Radu Marculescu.

## References

1. Lu Y, Goldstein D. Personalized medicine and human genetic diversity. *Cold Spring Harbor Perspectives in Medicine*. 2014; p. 1–12.
2. Jovel J, et al. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Frontiers in microbiology*. 2016; 7(April):459. <https://doi.org/10.3389/fmicb.2016.00459> PMID: 27148170
3. Turnbaugh P, et al. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*. 2007; 449(7164):804–810. <https://doi.org/10.1038/nature06244> PMID: 17943116
4. Gilbert J, Meyer F. The Earth Microbiome Project: meeting report of the “1 st EMP meeting on sample selection and acquisition” at Argonne National Laboratory October 6 th. *Standards in Genomic Sciences*. 2010; p. 249–253. <https://doi.org/10.4056/aigs.1443528> PMID: 21304728
5. Sims D, et al. Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews Genetics*. 2014; 15(2):121–32. <https://doi.org/10.1038/nrg3642> PMID: 24434847
6. Aitchison J. *The Statistical Analysis of Compositional Data*. London, UK, UK: Chapman & Hall, Ltd.; 1986.
7. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS computational biology*. 2012; 8(9):e1002687. <https://doi.org/10.1371/journal.pcbi.1002687> PMID: 23028285
8. Kurtz Z, et al. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology*. 2015; p. 1–25.
9. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*. 2006; 34(3):1436–1462. <https://doi.org/10.1214/009053606000000281>
10. Friedman J, et al. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008; p. 1–14.
11. Fang H, et al. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics*. 2015;(2009):1–8.
12. Freilich S, et al. The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic acids research*. 2010; 38(12):3857–68. <https://doi.org/10.1093/nar/gkq118> PMID: 20194113
13. Lim K, et al. @ MInter: automated text-mining of microbial interactions. *Bioinformatics*. 2016; 32(19): 2981–7. <https://doi.org/10.1093/bioinformatics/btw357> PMID: 27312413
14. Wang Z, Xu W, Lucas FAS, Liu Y. Incorporating prior knowledge into Gene Network Study. *Bioinformatics*. 2013; 29(20):2633–2640. <https://doi.org/10.1093/bioinformatics/btt443> PMID: 23956306
15. Li Y, Jackson S. Gene network reconstruction by integration of prior biological knowledge. *G3: Genes| Genomes| Genetics*. 2015; 5(6):1075–9. <https://doi.org/10.1534/g3.115.018127> PMID: 25823587
16. Ban Y, et al. Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics*. 2015; 31(20):3322–9. <https://doi.org/10.1093/bioinformatics/btv364> PMID: 26079350
17. Faust K, Sathirapongsasuti J. Microbial co-occurrence relationships in the human microbiome. *PLoS Computational Biology*. 2012; 8(7):e1002606. <https://doi.org/10.1371/journal.pcbi.1002606> PMID: 22807668
18. Schloss PD, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*. 2009; 75(23):7537–41. <https://doi.org/10.1128/AEM.01541-09> PMID: 19801464
19. Kuczynski J, Stombaugh J. Using QIIME to analyze 16S rRNA gene sequences from Microbial Communities. *Current Protocols in Bioinformatics*. 2012; p. 1–28.
20. Truong D, et al. MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nature Methods*. 2015; 12(10):902–3. <https://doi.org/10.1038/nmeth.3589> PMID: 26418763
21. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1998; 75(4): 800–802. <https://doi.org/10.1093/biomet/75.4.800>
22. Chang C, Lin C. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*. 2011; p. 1–39.
23. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*. 2007; 94(1):19–35. <https://doi.org/10.1093/biomet/asm018>
24. Zhao T, et al. The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*. 2012; 13:1059–1062. PMID: 26834510

25. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Computational Biology*. 2014; 10(4):1–12. <https://doi.org/10.1371/journal.pcbi.1003531>
26. Kreth J, et al. Bacterial and Host Interactions of Oral Streptococci. *DNA and Cell Biology*. 2009; 28(8): 397–403. <https://doi.org/10.1089/dna.2009.0868> PMID: 19435424
27. Zhang Y, Zhang H. Microbiota associated with type 2 diabetes and its related complications. *Food Science and Human Wellness*. 2013; 2(3–4):167–172. <https://doi.org/10.1016/j.fshw.2013.09.002>
28. Marino S, et al. Mathematical modeling of primary succession of murine intestinal microbiota. *Proceedings of the National Academy of Sciences of the United States of America*. 2014; 111(1):439–44. <https://doi.org/10.1073/pnas.1311322111> PMID: 24367073
29. Lo C, Marculescu R. Inferring Microbial Interactions from Metagenomic Time-series Using Prior Biological Knowledge. *ACM-BCB'17*. New York, NY, USA. 2017. p. 168–177.
30. Caporaso JG, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*. 2011; 108(Supplement 1):4516–4522. <https://doi.org/10.1073/pnas.1000080107>
31. Li C, et al. Predicting microbial interactions through computational approaches. *Elsevier Methods*. 2016; 102:12–9.