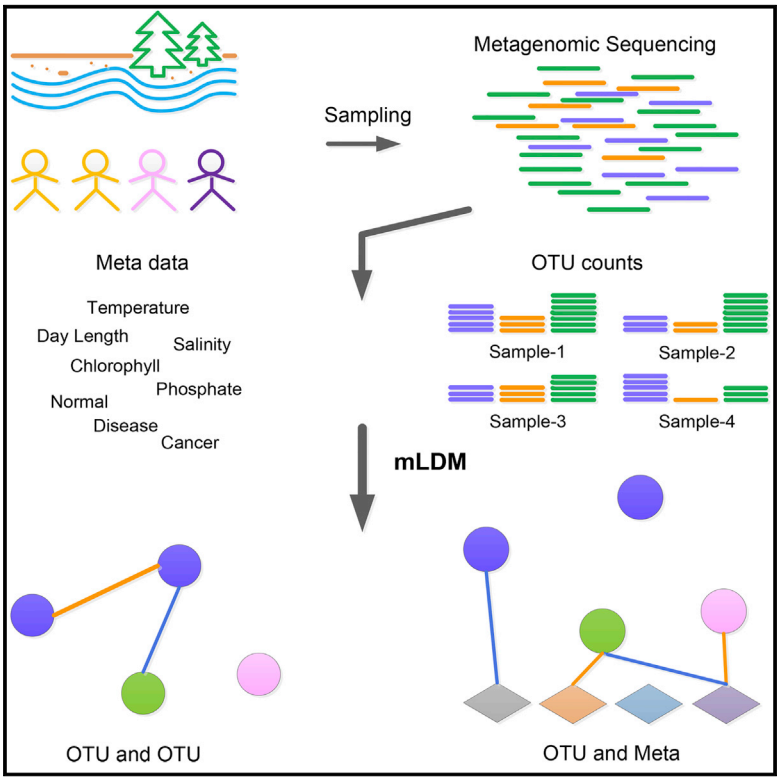
Report

Cell Systems

Inference of Environmental Factor-Microbe and Microbe-Microbe Associations from Metagenomic Data Using a Hierarchical Bayesian Statistical Model

Graphical Abstract



Highlights

- mLDM infers microbe and environmental associations in metagenomic datasets
- It removes indirect associations among OTUs correlated with a common factor
- The model accounts for compositional bias and variance in metagenomic sequencing data
- We apply mLDM to marine and human gut microbial datasets

Authors

Yuqing Yang, Ning Chen, Ting Chen

Correspondence

ningchen@tsinghua.edu.cn (N.C.),
tingchen@tsinghua.edu.cn (T.C.)

In Brief

A new statistical model for analyzing metagenomics data reveals associations among microbes and between microbes and environmental factors.

Yang et al., 2017, Cell Systems *4*, 129–137
January 25, 2017 © 2016 Elsevier Inc.
http://dx.doi.org/10.1016/j.cels.2016.12.012

Cell Systems
Report

CellPress

Inference of Environmental Factor-Microbe and Microbe-Microbe Associations from Metagenomic Data Using a Hierarchical Bayesian Statistical Model

Yuqing Yang,[1,2] Ning Chen,[1,*] and Ting Chen[1,2,3,4,*]
[1]MOE Key Lab of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, Center for Brain Inspired Computing Research (CBICR), TNLIST, Beijing 100084, China
[2]Department of Computer Science and Technology, State Key Lab of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China
[3]Program in Computational Biology and Bioinformatics, University of Southern California, CA 90089, USA
[4]Lead Contact
*Correspondence: ningchen@tsinghua.edu.cn (N.C.), tingchen@tsinghua.edu.cn (T.C.)
http://dx.doi.org/10.1016/j.cels.2016.12.012

SUMMARY

The inference of associations between environmental factors and microbes and among microbes is critical to interpreting metagenomic data, but compositional bias, indirect associations resulting from common factors, and variance within metagenomic sequencing data limit the discovery of associations. To account for these problems, we propose metagenomic Lognormal-Dirichlet-Multinomial (mLDM), a hierarchical Bayesian model with sparsity constraints, to estimate absolute microbial abundance and simultaneously infer both conditionally dependent associations among microbes and direct associations between microbes and environmental factors. We empirically show the effectiveness of the mLDM model using synthetic data, data from the TARA Oceans project, and a colorectal cancer dataset. Finally, we apply mLDM to 16S sequencing data from the western English Channel and report several associations. Our model can be used on both natural environmental and human metagenomic datasets, promoting the understanding of associations in the microbial community.

INTRODUCTION

Understanding interactions among microbes and between microbes and their environment is a key research topic in microbial ecology (Konopka, 2009). Most microbes cannot be cultured in laboratories, making it difficult to gain an understanding of their interactions with existing technologies. However, with the advancement of high-throughput sequencing technology, we are able to sequence 16S rRNA genes or whole metagenome of uncultured microbes directly from samples at diverse times or spots and, as a result, obtain microbial abundance information (Wooley et al., 2010) for further exploration. Various microbial datasets from different environments, such as oceans, soils, and humans have been published (Barberán et al., 2012; Proctor, 2015; Sogin et al., 2006) over the last few years. One of the major challenges is to discover associations, usually referred to as positive and negative relationships, among microbes and between microbes and environmental factors, or EFs. Such associations could help us to unravel real interactions, including, for example, commensalism, parasitism and competition in a community, resulting in a broad understanding of community-wide dynamics.

Associations can be measured by different statistical methods to investigate underlying relationships. Existing association studies can be classified into two main categories. The first is pairwise association calculation, such as Pearson's correlation coefficient (PCC) and Spearman's rank correlation coefficient (SCC), which directly computes the correlation between two species. Local similarity association (LSA) also computes pairwise association, but its mechanism differs from the others, and it calculates associations using the dynamic programming (Ruan et al., 2006). The second is complex association calculation that estimates the relationships between one species and the remaining species and/or EFs via multivariate regression-based methods (Chen and Li, 2013; Faust and Raes, 2012; Warton et al., 2015). Methods of calculating pairwise association are simple, fast and widely adopted (Chow et al., 2014; Eiler et al., 2012; Gilbert et al., 2012; Qin et al., 2012; Schwab et al., 2014; Steele et al., 2011), but such methods are not suitable for metagenomic datasets for the following two reasons. First, their calculated values may not indicate real associations because of compositional bias introduced when association is computed using methods that assume data are unconstrained, while ignoring dependence among the elements of compositional data (Aitchison, 1982). More specifically, the abundance of each microbe in metagenomic samples is usually normalized as the compositional relative abundance by dividing its read count over the total read count of a particular sample. Thus, after normalization, the relative abundance x_i is not independent from the relative abundance of the rest of the microbes, regardless of their underlying relationships as:

$$\sum_i x_i = 1 \rightarrow \sum_{j \neq i} \mathrm{cov}(x_i, x_j) = -Var(x_i).$$

Compositional bias tends to be more severe when some dominant species exist. This is particularly widespread in the marine microbial community (Caporaso et al., 2012; Chow et al., 2013). Consequently, for association studies, it is desirable to develop computational methods that bypass compositional bias in order to enable the inference of associations in metagenomic sequencing data. Second, the observed read count of one microbe may deviate from its true abundance based on a given experimental protocol, in which a series of sample preparation, amplification (Acinas et al., 2005), and sequencing steps, can lead to large variance of read counts. This variance within metagenomic sequencing data is also ignored by pairwise association calculation methods.

Recent advances have been made in the development of statistical methods to study associations using sequencing data, while taking compositional bias into account. For example, CCREPE (Faust et al., 2012) estimates the compositionally corrected p value for every association, allowing the extraction of significant associations via pairwise association calculation. Permutation and bootstrapping have also been used to generate the null distribution of the association while considering compositional bias, and the corrected p value is obtained by the pooled-variance Z-test. However, the limited number of data samples results in unreliable null distribution and corrected p values that are sensitive to noise. SparCC (Friedman and Alm, 2012) infers correlations among microbes by utilizing log-ratio transformation to eliminate the effect of the total number of read counts, while imposing sparsity of correlations among microbes. SPIEC-EASI (Kurtz et al., 2015) uses the covariance of the centered log-ratio-transformed data to approximate the covariance of log-transformed absolute abundance of microbes and obtains conditionally dependent associations among microbes. Similar to SPIEC-EASI, CCLasso (Fang et al., 2015) estimates the covariance matrix via an alternating direction algorithm instead of the graphical lasso. However, without considering environmental factors, many associations between and among microbes, as determined by these methods, may not be real. For example, Figure 1C shows that two unrelated microbes (OTU-1 and OTU-2) may appear to be associated just because they both respond to the same environmental perturbation (EF-1). Lima-Mendez et al. (2015) considered the effect of environmental factors by filtering out indirect associations among OTUs. However, this method is limited since only triples, i.e., two OTUs and one EF, are included each time. The influence of EFs on OTU-OTU associations was previously explored by Pascual García et al. (2014) by testing the statistical significance of associations among OTUs. A null model based on the assumption that associations are independent from either taxa or locations is constructed via the binary presence-absence matrix, which records the presence or absence of taxa in samples. However, the mere presence-absence binary information in taxa of samples, not abundance, is utilized; therefore, the results may be restricted.

RESULTS

Controlling for Bias, Indirect Effects, and Variance Using a Hierarchical Bayesian Model

To address the shortcomings of the methods noted above, we propose the metagenomic Lognormal-Dirichlet-Multinomial (mLDM) model in this study (Figure 1). It is a typical hierarchical Bayesian model (Agresti and Hitchcock, 2005; Ovaskainen et al., 2010) that learns complex relationships underlying the data. The sequencing process in which millions of DNA molecules are randomly sampled from a DNA library for sequencing (Metzker, 2010) can be modeled by a multinomial distribution. In metagenomics, a DNA library consists of a large number of amplified 16S or 18S rRNA gene sequences, and the relative abundances of OTUs in a library, which are determined by their real abundances in the environmental samples, can be modeled by a Dirichlet distribution. Thus mLDM models read counts of OTUs via Dirichlet-Multinomial distributions to estimate associations among OTUs considering the effect of EFs (Figure 1B). The real abundances of OTUs are determined by associations, both among OTUs and between OTUs and EFs, and consequently, mLDM applies a lognormal distribution to parameterize these two kinds of associations using two matrices, one denoting conditionally dependent associations among microbes and the other representing direct associations between microbes and EFs. Finally, the two estimated parameters are visualized as an OTU-OTU association network (Figure 1D) and an EF-OTU association network (Figure 1E), respectively.

Simulated Experiment on Synthetic Dataset

To show the effectiveness of the proposed mLDM model, we conducted several experiments and compared mLDM with several state-of-the-art models, including eight programs: PCC, SCC, CCREPE, SparCC, CCLasso, glasso (graphical lasso) (Friedman et al., 2008), SPIEC (multiple lasso [ml]), and SPIEC (graphical lasso [gl]). SPIEC (ml) and SPIEC (gl) are two different modules within SPIEC-EASI. The first five methods estimate associations via the calculation of correlations with PCC as the baseline, and the last three compute the conditional dependence with glasso as the baseline. In the next experiment, we will estimate the following: (1) OTU-OTU associations among all microbes (or OTUs) and (2) EF-OTU associations between environmental factors and microbes. Synthetic data can be naturally produced via our generative process, as shown in Figure 1B, based on five different graphical structures (random, cluster, scale-free, hub, and band) for microbes parameterized by Θ, and randomly sparse association structure between microbes and EFs parameterized by B. Receiver operating characteristic (ROC) curves, area under the curve (AUC) scores, and Δ_1 distance, which is defined as the L_1 distance between estimated results and ground truth, are used for evaluation.

First, we compared performances of all nine methods based on the estimation of OTU-OTU and EF-OTU associations with simulation parameters p = 50, Q = 5, and n = 500, corresponding to 500 samples with 50 OTUs and five EFs. Figure 2A shows the ROC curves of OTU-OTU association studies for five different types of graphical structures. The corresponding AUC scores and $\Delta_1^{(1)}$ distances are summarized in Table S1. From the ROC curves, we learn that mLDM has larger true-positive rates than any of the other methods when false-positive rates are small. The AUC scores of mLDM are superior to those of all other state-of-the-art methods across all five different graphical structures. A direct comparison between mLDM and glasso and SPIEC-EASI, which both estimate conditionally dependent

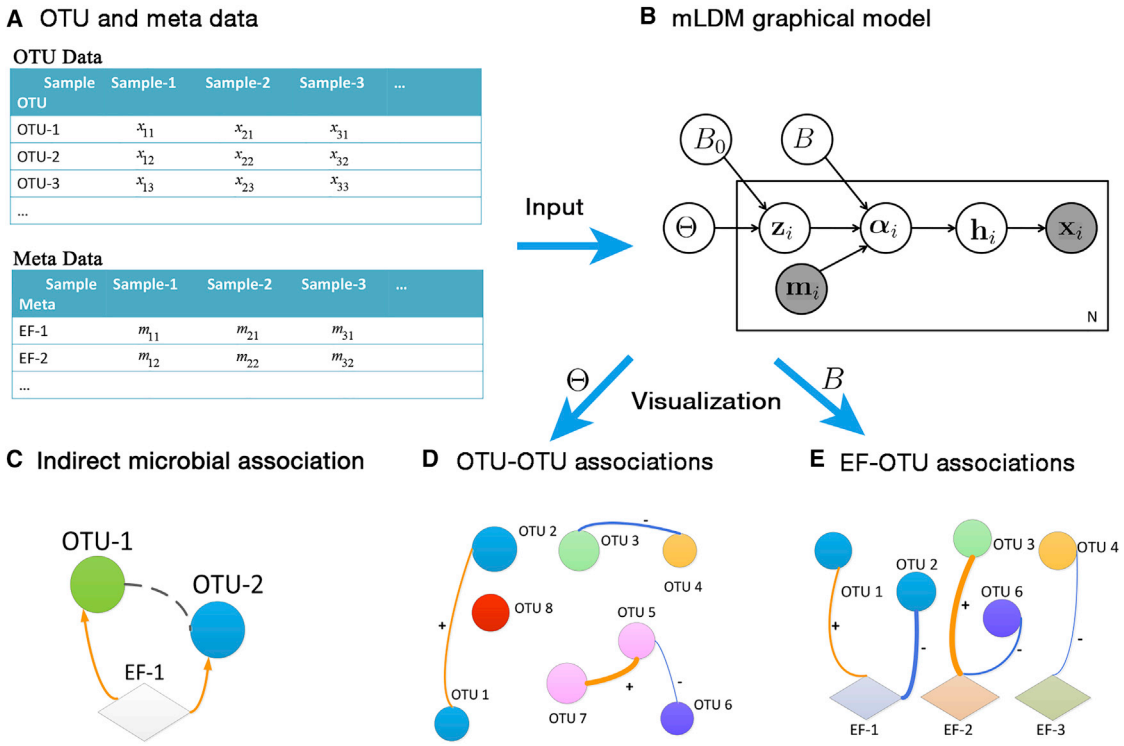


Figure 1. Schematic of mLDM

(A) OTU and meta data. OTU data consist of OTU read counts, and meta data record values of environmental factors. The data were preprocessed by omitting the missing values, filtering out OTUs with low frequency of occurrence, and removing abnormal samples.

(B) The mLDM graphical model accepts input from the OTU and meta data and estimates associations among microbes and between microbes and environmental factors. Matrix Θ represents conditionally dependent associations among microbes, and matrix B expresses direct associations between microbes and environmental factors. These two matrices can be respectively visualized by two networks (D) and (E). Gray indicates experimentally measured variables, read counts (x_i), and environmental factors (m_i), for sample i. The other variables in the model include B_0, which represents the average effect of all factors that affect microbial abundance but are not explicitly modeled; z_i, the latent variable that includes the influence on microbial abundance from the OTU-OTU associations; α_i, the absolute abundance of microbes; and h_i, the relative abundance levels of microbes in the sample. N metagenomic samples are generated according to variables within the box, which model the sequencing process (see STAR Methods for a full description of the model).

(C) Indirect microbial association between OTU-1 and OTU-2. mLDM could detect and remove the indirect association between OTU-1 and OTU-2 and identify common environmental factor EF-1 that actually affects their abundance.

(D) Microbial association (OTU-OTU) network. "+" and "−" correspond to the positive (orange edges) and negative (blue edges) associations, respectively. Colors of OTUs represent different taxa, and the same colors belong to identical taxa. The thickness of edges is correlated to the strength of associations.

(E) Environmental factor-microbe (EF-OTU) association network.

associations without considering the variance of metagenomic data, shows that mLDM achieves the highest AUC scores across all five graphical structures. We also observe that mLDM has smaller Δ_1 distances than most of the other methods, suggesting that mLDM is able to accurately estimate the weight and sign of conditionally dependent associations. On the cluster graph, the ROC curves of SparCC and CCLasso increase more slowly than those of mLDM at the beginning, but climb higher as the false-positive rates become larger. This can be explained by the local density of each standalone cluster in the graph. Under these conditions, mLDM tends to shrink edges with low weights, finally retaining fewer edges than either SparCC or CCLasso. However, we argue that an initial high true-positive rate, when the false-positive rate is small, is very significant in biological applications, essentially because a higher ratio of predicted associations will be true.

Figure 2B shows the ROC curves for the estimated associations between EFs and OTUs (EF-OTU), where simulation parameters are set the same as those shown in Figure 2A. The corresponding AUC scores and $\Delta_1^{(2)}$ distances are shown in Table S2. CCREPE, SparCC, CCLasso, and SPIEC do not estimate EF-OTU associations; therefore, we compared mLDM with PCC, SCC and glasso only. From the ROC curves, we observe that mLDM has higher true-positive rates and lower false-positive rates than the other four methods. From the AUC scores, we observe that mLDM has better performance than the other methods. For $\Delta_1^{(2)}$ distances, mLDM also performs better than the other methods, with the exception of SCC, which does slightly better in the Band graph. More comparisons (with DirMulti, Dirichlet-multinomial regression [Chen and Li, 2013]) can be found in Figure S4.

Next, to show the sensitivity of the computational models with respect to different sample sizes, we fixed the number of microbes as p = 50 and the number of EFs as Q = 5 and simulated metagenomic sequencing datasets with various sample sizes, including n = 25, 50, 200, and 500. The AUC scores of the

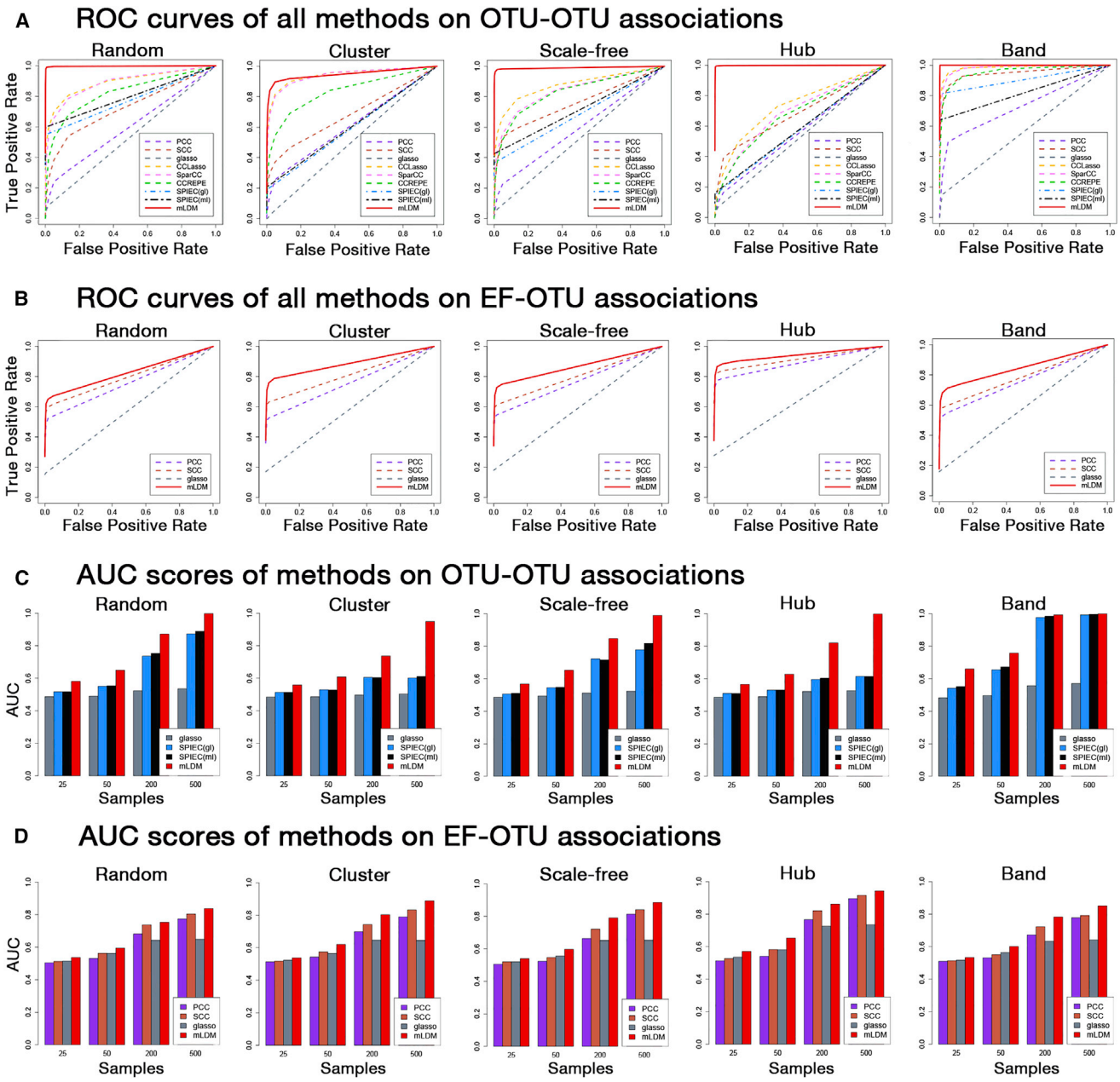


Figure 2. Performance of Association Inference of Nine Methods on Synthetic Experiment

(A) Comparisons of ROC curves of computational methods for predicting OTU-OTU associations on five different graphical structures (random, cluster, scale-free, hub and band). Average results of 20 simulations with the same parameters are displayed. Each simulated dataset consists of 500 samples (n = 500) with 50 OTUs (p = 50) and five environmental factors (Q = 5).

(B) Comparisons of ROC curves of computational methods for predicting EF-OTU associations on five graphical structures. Average results of 20 simulations with the same parameters are displayed.

(C) Comparisons of AUC scores of computational methods for predicting OTU-OTU associations using different numbers of samples (n = 25, 50, 200, and 500) with the numbers of microbes and environmental factors fixed (p = 50 and Q = 5).

(D) Comparisons of AUC scores of computational methods for predicting EF-OTU associations using p = 50, Q = 5, n = 25, 50, 200, and 500.

estimated OTU-OTU associations by glasso, SPIEC (gl), SPIEC (ml), and mLDM are plotted in Figure 2C. As expected, the AUC scores of all five methods increase when the sample size increases. Among these methods, mLDM gives the highest AUC scores on all five graphical structures, which again proves that mLDM can accurately estimate conditionally dependent associations. The AUC scores of the estimated EF-OTU associations by PCC, SCC, glasso, and mLDM are shown in Figure 2D, and, again, the AUC scores of mLDM are higher than those of PCC, SCC, or glasso.

Performance on TARA Oceans Eukaryotic Data

To validate the performance of mLDM on discovering OTU-OTU associations from real metagenomic sequencing data, we show the results of mLDM, as well as eight other methods, on TARA Oceans eukaryotic data (Lima-Mendez et al., 2015). The eukaryotic abundance profiles were estimated by sequencing and clustering the V9 region of eukaryotic 18 s rRNA genes. A subset for evaluations was extracted from datasets established by the original authors. This subset consists of 67 OTUs with 28 known genus-level interactions and 17 EFs from 221 samples.

It should be noted that the known interactions are at genus level, and thus we evaluated the results at genus level. Since the exact OTU-OTU associations at species level are unidentified, we further specified that a predicted association between two OTUs would match a known genus-level interaction if the two OTUs belonged to two interacting genera. Since this is not a ground truth dataset because of its incompleteness, we reported the numbers of matched genus-level associations among the top-N predicted associations (with the highest weights) of all methods, as listed in Table S3. It can be seen that mLDM is superior to other programs in terms of the number of matched associations for six cases, demonstrating its power of association inference. SCC is competitive with mLDM when $N \leq 40$, but its performance decreases as N increases. Both CCLasso and SparCC tend to report a dense association network, which includes a large number of false-positive associations, as shown in Figures 3E and 3F. In contrast, mLDM assumes network sparsity and therefore selects associations with higher weights, as shown in Figure 3D.

The ground truth, consisting of 28 genus-level symbiotic interactions, as listed in Table S4, and the top 40 highest valued genus-level associations discovered by mLDM, are plotted in Figures 3A and 3B, respectively. The strong negative association between the genus *Amoebophrya* and the genus *Alexandrium*, as given by mLDM, implies a parasitic interaction, which matches known parasitic interactions (Chambouvet et al., 2011). The known parasitic interactions between *Amoebophrya* and *Peridiniaceae*, and between *Amoebophrya* and *Acanthometra* were also detected by mLDM as having negative associations (Gunderson et al., 2002). We also list the top 10 predicted OTU-OTU associations, i.e., those with largest weights, together with relevant citations in Table S5.

Figure 3C shows EF-OTU associations estimated by mLDM. Compared to OTU-OTU associations estimated by mLDM, as shown in Figure 3D, fewer EF-OTU associations are found, indicating that EFs have direct effect on only some OTUs, while OTU-OTU associations comprise the greater share of forces that drive the changes of microbial community. Similarly, we show the top 10 estimated EF-OTU associations in Table S5. Some of these predictions were consistent with findings reported in the literature. For example, Cope-1 (*Corycaeus sp.*) is positively associated with the depth of maximum Brunt-Väisälä frequency, which is a measure of the stability of ocean's stratification. This discovery is consistent with a previous work about the predictability of the depth of maximum Brunt-Väisälä frequency to Cope-1 (Irigoien et al. (2011). The relationships between the depth of maximum chlorophyll and Cope-2 (*Oithona sp.*) and between moon phase and Cope-7 (*Centropages fu.*) were also studied in other projects (Munk, 1993; Osore et al., 2004). Figure 3G shows a non-linear association found by mLDM, where estimated absolute abundances of Corycaeus sp vary with concentrations of oxygen.

EF-OTU Associations on Human Gut Microbes from Colorectal Cancer Dataset

We next evaluated EF-OTU associations estimated by mLDM on human gut microbes from a colorectal cancer dataset (Baxter et al., 2016). The composition of gut microbes of patients with colorectal cancer (CRC) has been found to differ from that of the normal gut microbial community, and some microbes, such as *Fusobacterium*, *Peptostreptococcus*, *Parvimonas*, and *Porphyromonas*, have been reported to be enriched in the patients' gut (Feng et al., 2015; Yu et al., 2015; Zeller et al., 2014). A total of 117 OTUs and five meta data (FIT results, site, Dx_Bin, age, and gender) out of 490 samples were selected from the dataset provided by the original authors to construct association networks. Among the meta data, "site" contains four cities with three cities in the US and one in Canada, and "Dx_Bin" comprises five diagnostic states, "normal," "high-risk normal," "adv adenoma," "Adenoma," and "cancer." Results of mLDM and previous studies were compared to verify our model's effectiveness. Table 1 lists top 12 EF-OTU associations estimated by mLDM.

We observed that four OTUs, *Peptostreptococcus* (OTU310), *Porphyromonas* (OTU105), *Parvimonas* (OTU281), and *Fusobacterium* (OTU264), appear among the top 12 EF-OTU associations and are positively associated with CRC. The unclassified *Prevotella* (OTU57) reported by the original authors is also found to be positively associated with CRC, and it is the 25th largest EF-OTU association (+0.185). These results are consistent with previous studies. We believe that this is convincing validation for the accuracy of EF-OTU associations estimated by mLDM.

In addition, for eight out of the 12 EF-OTU associations, p values via the Wilcoxon rank-sum test are shown in Table 1. All these associations are statistically significant, which again shows the efficiency of mLDM as a predictor of EF-OTU associations. Interestingly, among the top 12 EF-OTU associations, we also discover that two microbes are associated with the "age," *Veillonella* (OTU66) (+0.364), and Parasutterella (OTU82) (−0.275), and that a special species, *Pasteurellaceae* (OTU58), is negatively associated (−0.298) with CRC. More studies are needed to explain these associations.

Association Inference on West English Channel Data

Finally, we applied mLDM to other marine metagenomic sequencing data to infer the underlying OTU-OTU associations and EF-OTU associations. In the marine community, huge numbers of marine microbes play important roles in ocean food chains. However, very little is known about how marine microbes interact with each other or how they are affected by environmental factors. Gilbert et al. (2012) studied the dynamics of the marine microbial community in the West English Channel by analyzing high-throughput 16S rRNA data sampled from 2003 to 2008. From these data, we extracted 48 OTUs and eight EFs that appear in 46 samples and employed mLDM to infer associations.

The OTU-OTU association network for the 48 OTUs is shown in Figure 3H. In general, the number of positive associations (brown edges) among OTUs is more than that of the negative

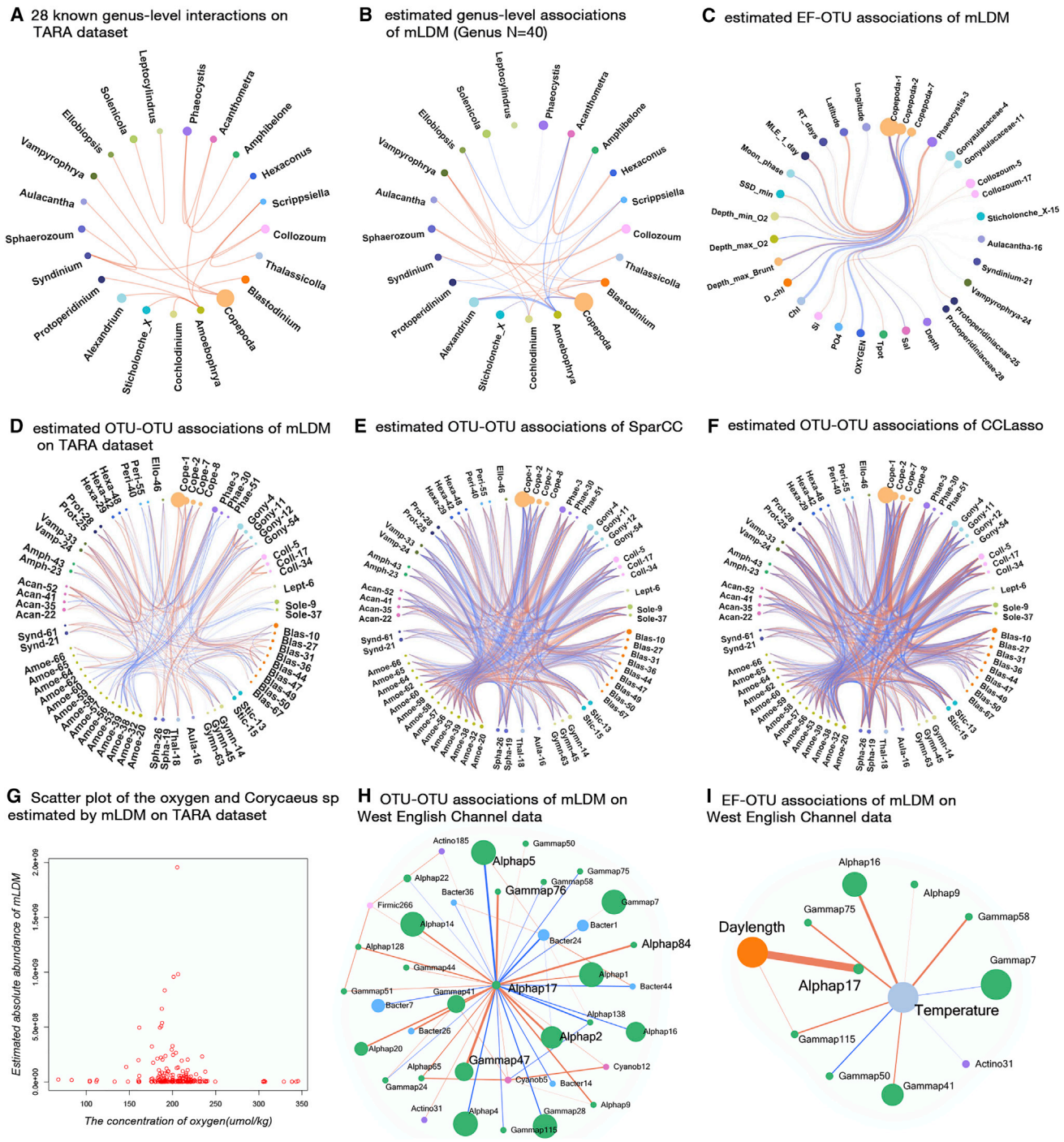


Figure 3. Results of Experiments on TARA Oceans Eukaryotic Dataset and West English Channel Data

(A) A network for 28 known genus-level symbiotic interactions from the TARA Oceans Eukaryotic dataset. Since the signs of the interactions are unknown, we show them in brown for convenience. Sizes of nodes are proportional to their relative abundance.

(B) The genus-level association network (TARA Oceans Eukaryotic dataset) discovered by mLDM where only the top n = 40 genus-level associations are plotted. OTUs that belong to the same genera are labeled with the same colors. The brown and blue edges represent positive and negative associations, respectively. Thickness of an edge is proportional to its absolute edge weight.

(C) Predicted EF-OTU association network by mLDM (TARA Oceans Eukaryotic dataset).

(D) Predicted OTU-OTU association network by mLDM (TARA Oceans Eukaryotic dataset).

(E) Predicted OTU-OTU association network by SparCC (TARA Oceans Eukaryotic dataset).

(F) Predicted OTU-OTU association network by CCLasso (TARA Oceans eukaryotic dataset).

(legend continued on next page)

Table 1. Top 12 EF-OTU Associations Estimated by mLDM on Colorectal Cancer Data Dataset

OTU	EF	Association	p value (Wilcoxon rank-sum test)
Peptostreptococcus (OTU310)	cancer	+0.865	2.00×10^{-15}
Porphyromonas (OTU105)	cancer	+0.617	2.08×10^{-14}
Fusobacterium (OTU264)	normal	−0.463	1.74×10^{-5}
Fusobacterium (OTU264)	FIT positive	+0.442	N/A
Parvimonas (OTU281)	cancer	+0.378	3.50×10^{-12}
Porphyromonas (OTU105)	normal	−0.372	7.34×10^{-6}
Veillonella (OTU66)	age	+0.364	N/A
Parvimonas (OTU281)	normal	−0.307	7.94×10^{-5}
Pasteurellaceae (OTU58)	cancer	−0.298	2.95×10^{-5}
Porphyromonas (OTU105)	FIT positive	+0.288	N/A
Parasutterella (OTU82)	age	−0.275	N/A
Fusobacterium (OTU264)	cancer	+0.272	2.68×10^{-7}

All predicted EF-OTU associations are sorted in descending order according to their absolute values. Wilcoxon rank-sum test is performed on eight out of 12 EF-OTU associations where the EFs include five diagnostic states, "normal," "high-risk normal," adv Adenoma," "adenoma," and "cancer," "FIT" (fecal immunochemical test), and "age." The content in the "OTU" column consists of annotated OTUs and OTU numbers from the original article. Wilcoxon rank-sum test was performed on the diagnostic state-OTU associations but is not applicable for other types of EF-OTU associations.

associations (blue edges). The network is clearly dominated by OTUs from *Proteobacteria*, which are colored green. This result is consistent with the original discovery by Gilbert et al. (2012). The OTU Alphap17, which belongs to the family *Rhodospirillaceae*, plays an important role in the network, as it is a hub connecting most OTUs. *Rhodospirillaceae* is known to produce energy through photosynthesis, which is critical to the marine microbial community on the surface of the ocean. Gilbert et al. (2012) also found that a single *Rhodobactereaceae* OTU acts as a hub and is correlated with different groups. Although the OTU Alphap5 from the genus *Thalassobacter*, the OTU Alphap2 from the family SAR11, and the OTU Alphap17 are from the same class, *Alphaproteobacteria*, their associations are different. Alphap5 and Alphap17 have a strong negative association while Alphap2 and Alphap17 have a positive association. The OTUs Gammap47 and Gammap76 are from the same family, SAR86, and both have a positive association with the OTU Alphap17. It is remarkable that the relative abundance of Alphap17 is so low, while still connecting many big OTUs with high relative abundance levels, such as Alphap1, Alphap2, Gammap76, and Gammap7, implying that we should pay more attention to rare OTUs with low abundance in future research.

Figure 3I shows the EF-OTU association network between eight EFs and 48 OTUs. We observe that temperature has the most significant impact on OTUs, especially on the phylum *Proteobacteria*. This is consistent with previous observations. Furthermore, the OTU Alphap17, which connects many other OTUs, is very strongly and positively associated with day length. This is consistent with the photosynthesis function of OTU Alphap17 and further confirms that the photosynthesis of Alphap17 is critical to the whole marine microbial community. Gilbert et al. (2012) also associated day length with the variance of microbial community via discriminant function analysis. In addition, the OTU Alphap16 from the family *Rhodobacteraceae* has a positive association with temperature. The top ten OTU-OTU and EF-OTU associations are shown in Table S6. The positive associations between temperature and both Alphap16 and Gammap58 were previously reported by Lefort and Gasol (2013).

DISCUSSION

To discover the underlying associations among microbes from metagenomic samples, we propose mLDM, a hierarchical Bayesian model with sparsity constraints to discover associations among microbes and between microbes and the environmental factors that affect them. mLDM can infer both conditionally dependent associations among microbes and direct associations between microbes and environmental factors, by taking into account both compositional bias and variance of metagenomic data, an approach not previously studied. This newly discovered conditionally dependent association provides insight into the mechanisms underlying a microbial community by capturing the direct relationship underlying each microbial pair and removing the indirect connection induced from other common factors. The effectiveness of mLDM was verified on the basis of experiments involving both synthetic and real datasets.

To address the question whether environmental factors are important for the inference of OTU-OTU associations, we applied mLDM on synthetic datasets, when only one type of associations, either OTU-OTU or EF-OTU associations, was estimated, similar to the approach of Ovaskainen et al. (2010). The results are shown in Figure S1. Compared to the methods considering both types of associations, we observe lower ROC curves for those estimating only one type of associations. Therefore, it can be concluded that environmental factors would affect the estimation of OTU-OTU associations, that OTU-OTU associations would affect EF-OTU associations, and that both types of associations should be considered in association estimation.

Since mLDM assumes sparsity of true association, we also test whether the sparsity pattern of OTU-EF interactions (matrix B) would affect association estimation by this method. In matrix B, coefficients of a row correspond to the impact of an environmental factor, and coefficients of a column correspond to the impact of multiple environmental factors to an out. Therefore,

(G) Scatterplot of the concentrations of oxygen and estimated absolute abundances of *Corycaeus sp.* by mLDM (TARA Oceans Eukaryotic dataset).
(H) Estimated OTU-OTU associations by mLDM on the West English Channel data. Nodes in the same color belong to the same phylum, and the diameter of each node is proportional to the relative abundance of the OTU. Edges in brown and blue colors denote positive and negative associations, respectively.
(I) Estimated EF-OTU associations by mLDM on West English Channel data.

we assumed some sparsity patterns of B by assigning some fractions of rows or columns as nonzero and plotted the ROC curves of estimated EF-OTU associations by mLDM, as shown in Figure S2. Overall, mLDM is insensitive to these patterns and works well in all cases. These results indicate that mLDM can be applied to estimate various types of sparse associations.

mLDM assumes that microbes respond linearly to environmental factors. However, the abundance of microbes may reach optima under certain environmental conditions, such as some range of temperature and depth. While this appears to be a limitation, we argue that some nonlinear associations can, to some extent, be captured by our model, when data points are distributed askew, which is typical in real datasets. For example, among EF-OTU associations estimated by mLDM on the TARA Oceans Eukaryotic Data, we observed that the OTU Cope-1 annotated with the strain *Corycaeus sp.* is negatively associated with oxygen concentration. Almost 95% of all 221 samples of the TARA Oceans dataset are either from the surface waters or from the deep chlorophyll maximum subsurface, whose depths range from 5.374 to 183.31 m. From the samples near the ocean surface, the abundance of *Corycaeus sp.* does not increase linearly with the increase of oxygen but rather tends to be more abundant when the concentration of oxygen is within a certain range, as plotted in Figure 3G. This example demonstrates that mLDM is capable of capturing some of these nonlinear associations.

Consistency is an important property that shows the robustness of methods against noise. Accordingly, we tested mLDM on the Human Microbiome Project dataset (HMP) and constructed two datasets to evaluate consistency. Since some subjects had two gut samples from two time points, we constructed the first dataset using the samples from the first time point and the second dataset using those from the second time point. Consistency was measured by Jaccard similarity, i.e., the fraction of the number of intersections among the top 200 largest OTU-OTU associations, as estimated from the two datasets, over the number of the union of these two sets of associations. The consistency of nine methods was then plotted, and it is shown in Figure S3. We observe that the consistency of mLDM is ranked fifth among all, and among the four methods that estimate conditionally dependent associations, including glasso, SPIEC (gl), SPIEC (ml), and mLDM, mLDM is the second best. Of the nine methods compared, CCLasso had the highest consistency, while glasso had the lowest consistency. Overall, methods that estimate direct correlations had higher consistency than those that estimate conditionally dependent associations. However, approaches that estimate conditionally dependent associations are sensitive to heterogeneity or noise within the dataset, particularly in model selection. However, consistency needs not to be the best standard to assess methods because, to some extent, consistency may reflect this method is misled by systematic bias.

For future work, we will develop a more scalable mLDM model to analyze large microbial network structures with tens of thousands of microbes by using stochastic gradient descent and parallel computing techniques. For rare OTUs, which only exist in a small fraction of the samples, the lognormal distribution may be not suitable, and other appropriate distributions need to be explored. We will also develop dynamic mLDM models to analyze time series data and learning time-varying network structures.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHODS DETAILS
  - The metagenomic Lognormal-Dirichlet-Multinomial Model
  - Sparse association estimation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data Generation and Evaluation Metrics in Synthetic Experiment
  - Preprocessing of TARA Oceans Eukaryotic Data
  - Preprocessing of Colorectal Cancer Data
  - Preprocessing of West English Channel Data
- DATA AND SOFTWARE AVAILAVILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.12.012.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.Y., N.C., and T.C.; Methodology, Y.Y. and N.C.; Software, Y.Y.; Format Analysis, Y.Y.; Investigation, Y.Y., N.C., and T.C.; Writing – Original Draft, Y.Y., N.C., and T.C.; Writing – Review & Editing, Y.Y., N.C., and T.C.; Visualization, Y.Y.; Supervision, N.C. and T.C.

ACKNOWLEDGMENTS

We are thankful for the assistance of Jun Zhu on methodology. This work is supported by the National Natural Science Foundation of China (nos: 61305066, 61561146396, 61322308, 61673241). An early version of this paper was submitted to and peer reviewed at the 2016 Annual International Conference on Research in Computational Molecular Biology (RECOMB). The manuscript was revised and then independently further reviewed at *Cell Systems*.

Received: March 26, 2016
Revised: August 2, 2016
Accepted: December 20, 2016
Published: January 25, 2017

REFERENCES

Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., and Polz, M.F. (2005). PCR-induced sequence artifacts and bias: Insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl. Environ. Microbiol. *71*, 8966–8969.

Agresti, A., and Hitchcock, D.B. (2005). Bayesian inference for categorical data analysis. Stat. Methods Appl. *14*, 297–330.

Aitchison, J. (1982). The statistical analysis of compositional data. J. R. Stat. Soc. B *44*, 139–177.

Albert, R., and Barabasi, A.L. (2001). Statistical mechanics of complex networks. Rev. Mod. Phys. *74*, xii.

Andrew, G., and Gao, J. (2007). Scalable training of L1-regularized log-linear models. Proceedings of the 24th International Conference on Machine Learning. 33–40.

Barberán, A., Bates, S.T., Casamayor, E.O., and Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. ISME J. *6*, 343–351.

Baxter, N.T., Ruffin, M.T., 4th, Rogers, M.A.M., and Schloss, P.D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Med. *8*, 37.

Caporaso, J.G., Paszkiewicz, K., Field, D., Knight, R., and Gilbert, J.A. (2012). The Western English Channel contains a persistent microbial seed bank. ISME J. *6*, 1089–1093.

Chambouvet, A., Laabir, M., Sengco, M., Vaquer, A., and Guillou, L. (2011). Genetic diversity of Amoebophryidae (Syndiniales) during Alexandrium catenella/tamarense (Dinophyceae) blooms in the Thau lagoon (Mediterranean Sea, France). Res. Microbiol. *162*, 959–968.

Chen, J., and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. Biometrika *95*, 759–771.

Chen, J., and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. Ann. Appl. Stat. *7*, 418–442.

Chow, C.-E.T., Sachdeva, R., Cram, J.A., Steele, J.A., Needham, D.M., Patel, A., Parada, A.E., and Fuhrman, J.A. (2013). Temporal variability and coherence of euphotic zone bacterial communities over a decade in the Southern California Bight. ISME J. *7*, 2259–2273.

Chow, C.-E.T., Kim, D.Y., Sachdeva, R., Caron, D.A., and Fuhrman, J.A. (2014). Top-down controls on bacterial community structure: Microbial network analysis of bacteria, T4-like viruses and protists. ISME J. *8*, 816–829.

Eiler, A., Heinrich, F., and Bertilsson, S. (2012). Coherent dynamics and association networks among lake bacterioplankton taxa. ISME J. *6*, 330–342.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015). CCLasso: Correlation inference for compositional data through Lasso. Bioinformatics *31*, 3172–3180.

Faust, K., and Raes, J. (2012). Microbial interactions: From networks to models. Nat. Rev. Microbiol. *10*, 538–550.

Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. PLoS Comput. Biol. *8*, e1002606.

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat. Commun. *6*, 6528.

Friedman, J., and Alm, E.J. (2012). Inferring correlation networks from genomic survey data. PLoS Comput. Biol. *8*, e1002687.

Friedman, J.H., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso *9*, 432–441.

Gilbert, J.A., Steele, J.A., Caporaso, J.G., Steinbrück, L., Reeder, J., Temperton, B., Huse, S., McHardy, A.C., Knight, R., Joint, I., et al. (2012). Defining seasonal marine microbial community dynamics. ISME J. *6*, 298–308.

Gunderson, J.H., John, S.A., Boman, W.C., and Coats, D.W. (2002). Multiple strains of the parasitic dinoflagellate Amoebophrya exist in Chesapeake Bay. J. Eukaryot. Microbiol. *49*, 469–474.

Hong, S.H., Bunge, J., Jeon, S.O., and Epstein, S.S. (2006). Predicting microbial species richness. Proc. Natl. Acad. Sci. USA *103*, 117–122.

Irigoien, X., Chust, G., Fernandes, J.A., Albaina, A., and Zarauz, L. (2011). Factors determining the distribution and betadiversity of mesozooplankton species in shelf and coastal waters of the Bay of Biscay. J. Plankton Res. *33*, 1182–1192.

Konopka, A. (2009). What is microbial community ecology? ISME J. *3*, 1223–1230.

Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., and Bonneau, R.A. (2015). Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput. Biol. *11*, e1004226.

Lefort, T., and Gasol, J.M. (2013). Global-scale distributions of marine surface bacterioplankton groups along gradients of salinity, temperature, and chlorophyll: A meta-analysis of fluorescence in situ hybridization studies. Aquat. Microb. Ecol. *70*, 111–130.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J.C., Roux, S., Vincent, F., et al.; Tara Oceans coordinators (2015). Ocean plankton. Determinants of community structure in the global plankton interactome. Science *348*, 1262073.

Liu, D.C., and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. Mathprogram *45*, 503–528.

Metzker, M.L. (2010). Sequencing technologies—the next generation. Nat. Rev. Genet. *11*, 31–46.

Munk, P. (1993). Differential growth of larval sprat Sprattus sprattus across a tidal front in the eastern North Sea. Mar. Ecol. Prog. Ser. *99*, 17–27.

Murphy, K.P. (2012). Machine learning: A probabilistic perspective. Mathematics Education Library *58*, 27–71.

Osore, M., Mwaluma, J.M., Fiers, F., and Daro, M.H. (2004). Zooplankton composition and abundance in Mida Creek, Kenya. Zool. Stud. *43*, 415–424.

Ovaskainen, O., Hottola, J., and Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology *91*, 2514–2521.

Pascual-García, A., Tamames, J., and Bastolla, U. (2014). Bacteria dialog with Santa Rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions? BMC Microbiol. *14*, 284.

Proctor, L.M. (2015). Overview of the Phase One (2007-2012) of the NIH Human Microbiome Project. Encyclopedia of Metagenomics, 2015, 488–494.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60.

Ruan, Q., Dutta, D., Schwalbach, M.S., Steele, J.A., Fuhrman, J.A., and Sun, F. (2006). Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. Bioinformatics *22*, 2532–2538.

Schwab, C., Berry, D., Rauch, I., Rennisch, I., Ramesmayer, J., Hainzl, E., Heider, S., Decker, T., Kenner, L., Müller, M., et al. (2014). Longitudinal study of murine microbiota activity and interactions with the host during acute inflammation and recovery. ISME J. *8*, 1101–1114.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., Arrieta, J.M., and Herndl, G.J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc. Natl. Acad. Sci. USA *103*, 12115–12120.

Steele, J.A., Countway, P.D., Xia, L., Vigil, P.D., Beman, J.M., Kim, D.Y., Chow, C.-E.T., Sachdeva, R., Jones, A.C., Schwalbach, M.S., et al. (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. ISME J. *5*, 1414–1425.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J.R. Stat. Soc. *58*, 267–288.

Ulrich, W., and Ollik, M. (2004). Frequent and occasional species and the shape of relative-abundance distributions. Divers. Distrib. *10*, 263–269.

Unterseher, M., Jumpponen, A., Opik, M., Tedersoo, L., Moora, M., Dormann, C.F., and Schnittler, M. (2011). Species abundance distributions and richness estimations in fungal metagenomics—lessons learned from community ecology. Mol. Ecol. *20*, 275–285.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., and Hui, F.K. (2015). So many variables: Joint modeling in community ecology. Trends Ecol. Evol. *30*, 766–779.

Wooley, J.C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. PLoS Comput. Biol. *6*, e1000667.

Yu, J., Feng, Q., Wong, S., Zhang, D., Liang, Q.Y., Qin, Y., Tang, L., Zhao, H., Stenvang, J., and Li, Y. (2015). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut.

Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol. Syst. Biol. *10*, 766.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. J. Mach. Learn. Res. *13*, 1059–1062.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
TARA Oceans eukaryotic data	(Lima-Mendez et al., 2015)	http://doi.pangaea.de/10.1594/PANGAEA.843018
TARA Oceans environmental data	(Lima-Mendez et al., 2015)	http://www.raeslab.org/companion/ocean-interactome.html
Colorectal Cancer data	(Baxter et al., 2016)	https://github.com/SchlossLab/Baxter_glne007Modeling_GenomeMed_2015
West English Channel data	(Gilbert et al., 2012)	https://vamps.mbl.edu/
Software and Algorithms		
huge	(Zhao et al., 2012)	https://cran.r-project.org/web/packages/huge/index.html
HMP	R package	https://cran.r-project.org/web/packages/HMP/index.html
CCREPE	(Faust et al., 2012)	http://bioconductor.org/packages/release/bioc/html/ccrepe.html
SPIEC-EASI	(Kurtz et al., 2015)	https://github.com/zdk123/SpiecEasi
CCLasso	(Fang et al., 2015)	https://github.com/huayingfang/CCLasso
lbfgs	R package	https://cran.r-project.org/web/packages/lbfgs/index.html
mLDM	This paper	https://github.com/tinglab/mLDM/

CONTACT FOR REAGENT AND RESOURCE SHARING

Please contact the corresponding author Dr. Ting Chen (tingchen@tsinghua.edu.cn) for further information and requests about codes and datasets.

METHODS DETAILS

The metagenomic Lognormal-Dirichlet-Multinomial Model

Suppose there are N samples $\mathbf{X} = \{\mathbf{x_i}\}_{i=1}^{N}$. Each $\mathbf{x_i} \in N^P$ is a P-dimensional vector that contains P microbes (or Operational Taxonomic Units (OTUs)), where $\mathbf{x}_{ij}$ represents the sequence/read count of the j-th microbes in the i-th sample. Let $\mathbf{M} = \{\mathbf{m_i}\}_{i=1}^{N}$ represent the environmental factors, where each $\mathbf{m_i} \in R^Q$ is a Q-dimensional vector and $\mathbf{m}_{ij}$ represents the value of the j-th environmental factor associated with the i-th sample.

Figure 1B illustrates the mLDM model for metagenomic sequencing, where $\mathbf{x_i}$ is the read count vector of the i-th sample and $\mathbf{m_i}$ records values of the environmental factors corresponding to the i-th sample. The latent variable $\mathbf{h_i}$ is the vector of the relative abundance levels of P microbes in the extracted sample, and $\boldsymbol{\alpha_i}$ represents the absolute abundance levels of the microbes in the original community. We assume that the counts $\mathbf{x_i}$ are proportional to the latent microbial ratios $\mathbf{h_i}$ which are determined by their absolute abundance $\boldsymbol{\alpha_i}$. Microbial absolute abundance $\boldsymbol{\alpha_i}$ can be influenced by two factors: **1)** environmental factors $\mathbf{m_i}$, whose effects on the microbes are denoted by a linear regression model $\mathbf{B}^T\mathbf{m_i}$, and **2)** the associations among microbes encoded by a latent vector $\mathbf{z_i}$, which is determined by the matrix Θ that records microbial associations and the mean vector $\mathbf{B}_0$ that affects the basic absolute abundance of microbes. The microbial basic absolute abundance can be regarded as the average result of effects of all other factors that have an effect on microbial abundance, but are not included in the mLDM. More specifically, the generative process of the metagenomic Lognormal-Dirichlet-Multinomial hierarchical model is defined as:

$$z_i \sim Gaussian\left(\mathbf{B}_0, \Theta^{-1}\right)$$

$$\boldsymbol{\alpha_i} = \exp\left(\mathbf{B}^T\mathbf{m_i} + \mathbf{z_i}\right)$$

$$\mathbf{h_i} \sim Dirichlet\left(\boldsymbol{\alpha_i}\right)$$

$$\mathbf{x_i} \sim Multinomial\left(\mathbf{h_i}\right)$$

where $\mathbf{B}$ is a $Q \times P$ parameter matrix, $\mathbf{B}_0$ is a P-dimensional vector, and Θ is the inverse covariance matrix (i.e., precision matrix) of a multivariate Gaussian distribution. With this model, our goal is to infer both $\mathbf{B}$, the environmental factor-microbe (or EF-OTU)

associations, and Θ, the microbe-microbe (or OTU-OTU) associations, under some sparsity regularization which will be made clear in next section. We now explain the design of each component in mLDM.

We assume that read count data $\mathbf{x_i}$ follows a multinomial distribution with the microbial ratio parameter $\mathbf{h_i}$:

$$\mathrm{P}(\mathbf{x_i} \mid \mathbf{h_i}) = \binom{s(\mathbf{x_i})}{x_{i1}, \cdots, x_{iP}} \prod_{j=1}^{P} h_{ij}^{x_{ij}} \tag{1}$$

where $s(\mathbf{x_i}) = \sum_{j=1}^{P} x_{ij}$ is the total read count of the i-th sample. Since the multinomial parameter $\mathbf{h_i}$ is subject to the constraint that $\sum_{j=1}^{P} h_{ij} = 1$, we assume it follows a Dirichlet distribution

$$\mathrm{P}(\mathbf{h_i} \mid \boldsymbol{\alpha_i}) = \frac{1}{T(\boldsymbol{\alpha_i})} \prod_{j=1}^{P} h_{ij}^{\alpha_{ij}-1} \tag{2}$$

where $T(\boldsymbol{\alpha_i}) = (\prod_{j=1}^{P} \Gamma(\alpha_{ij}) / \Gamma(s(\boldsymbol{\alpha_i})))$, $\Gamma(\cdot)$ is the Gamma function and $s(\alpha_i) = \sum_{j=1}^{P} \alpha_{ij}$. Based on the conjugacy of Dirichlet and multinomial distribution, we can obtain the following Dirichlet-Multinomial distribution via integrating h_{ij} out

$$P(\mathbf{x_i} \mid \boldsymbol{\alpha_i}) = \int P(\mathbf{x_i} \mid \mathbf{h_i}) P(\mathbf{h_i} \mid \boldsymbol{\alpha_i}) d\mathbf{h_i} = \binom{s(\mathbf{x_i})}{x_{i1}, \cdots, x_{ip}} \frac{T(\boldsymbol{\alpha_i} + \mathbf{x_i})}{T(\boldsymbol{\alpha_i})} \tag{3}$$

The flexible variance-covariance property of the Dirichlet-multinomial distribution is suitable for modeling the sequencing data. A simple explanation is as follow. We calculate the variance of the read count x_{ij}, $\mathrm{Var}(x_{ij}) = s(x_i) \cdot C \cdot r_{ij} \cdot (1 - r_{ij})$, and the covariance of two read counts x_{ij} and x_{ik}, $Cov(x_{ij}, x_{ik}) = -s(x_i) \cdot C \cdot r_{ij} \cdot r_{ik}$, where $C = (s(\mathbf{x_i}) + s(\boldsymbol{\alpha_i}) / 1 + s(\boldsymbol{\alpha_i}))$ and $r_{ij} = \alpha_{ij} / s(\boldsymbol{\alpha_i})$, $r_{ik} = \alpha_{ik} / s(\boldsymbol{\alpha_i})$ are true relative abundance levels. We can see that both the variance and covariance of microbial counts are regulated by the sequencing depth $s(\mathbf{x_i})$ and the true relative abundance r_{ij} of the microbes. Moreover, the coefficient between x_{ij} and x_{ik} is negative, which models the compositional negative bias.

We further assume that the absolute abundance $\boldsymbol{\alpha_i}$ for all microbes in the i-th sample follows the multivariate lognormal distribution with mean $\boldsymbol{\mu_i}$ and covariance Θ^{-1} which is commonly used to model most microbial abundance except for some occasional species (Hong et al., 2006; Ulrich and Ollik, 2004; Unterseher et al., 2011). Microbes survive in a community through conditionally dependent associations. However, at the same time, microbes are also subjected to unpredictable fluctuations impacted by their microenvironment. Therefore, we record associations among microbes in the matrix Θ and let the mean $\boldsymbol{\mu_i}$ vary with the environmental data vector $\mathbf{m_i}$ by a linear regression model. Then the prior distribution is defined as

$$P(\boldsymbol{\alpha_i} \mid \mathbf{B}, \mathbf{B}_0, \Theta, \mathbf{m_i}) = \frac{1}{(2\pi)^{\frac{P}{2}} |\Theta|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2} (\log \boldsymbol{\alpha_i} - \boldsymbol{\mu_i})^T \Theta (\log \boldsymbol{\alpha_i} - \boldsymbol{\mu_i}) \right) \prod_{j=1}^{P} \frac{1}{\alpha_{ij}} \tag{4}$$

where $\boldsymbol{\mu_i} = B^T \mathbf{m_i} + B_0$. Using the relationship between the lognormal and Gaussian distributions, it is also equivalent to the following form:

$$\boldsymbol{\alpha_i} = \exp(B^T \mathbf{m_i} + \mathbf{z_i}) \tag{5}$$

where $\mathbf{z_i} \sim N(B_0, \Theta^{-1})$. This formulation avoids positivity constraint in the lognormal distribution. This is beneficial for finding the estimates, e.g., by using some unconstrained optimization algorithms, as explained in the next section.

With the above model, we capture both the conditionally dependent associations among microbes and the direct associations between microbes and environmental factors. More specifically, the conditionally dependent associations among microbes are encoded in the precision matrix Θ. To visualize the microbial association network, we use an undirected graph denoted as $G^{(1)} = (V^{(1)}, E^{(1)})$ employed in the Gaussian Markov random field (Murphy, 2012) to represent Θ, where $V^{(1)}$ represents the set of nodes denoting P microbes and $E^{(1)}$ is the set of conditionally dependent associations with each element $e_{ij}^{(1)}$ representing the association between the i-th and j-th microbes. If $\Theta_{ij} = 0$, then the i-th and the j-th microbes are conditionally independent, and hence, no edge exists between the two microbes in graph $G^{(1)}$. The weight of edge $e_{ij}^{(1)}$, $w_{ij}^{(1)} = -(\Theta_{ij} / \sqrt{\Theta_{ii}\Theta_{jj}})$, is the strength of the association between the two microbes.

The direct associations between microbes and environmental factors are encoded in weight matrix B. The association between the i-th microbe and the j-th environmental factor is B_{ij}, and we can plot them in another bipartite graph $G^{(2)} = (V^{(2)}, E^{(2)})$, where the set of nodes $V^{(2)}$ represents both P microbes and Q environmental factors, and the edge $e_{ij}^{(2)}$ in $E^{(2)}$ represents the direct association between the j-th environmental factor and the i-th microbe. The weight of edge $e_{ij}^{(2)}$ equals $w_{ij}^{(2)} = B_{ji}$.

Overall, our metagenomic association network consists of these two graphs $G^{(1)}$ and $G^{(2)}$, as illustrated in Figures 1D and 1E.

Sparse association estimation

We now explain how to estimate the metagenomic association network by using sparsity regularization. Given metagenomic data $\mathbf{X}$ and environmental factors $\mathbf{M}$, the posterior distribution of the latent factors $\mathbf{Z}$ is

$$\mathrm{P}(\mathbf{Z} \mid \mathbf{X}, \mathbf{M}, B, B_0, \Theta) \propto \mathrm{P}(\mathbf{X}, \mathbf{Z} \mid B, B_0, \Theta, \mathbf{M}) \propto \mathrm{P}(\mathbf{X} \mid \boldsymbol{\alpha}) \mathrm{P}(\boldsymbol{\alpha} \mid \mathbf{Z}, B, B_0, M) \mathrm{P}(\mathbf{Z} \mid B_0, \Theta) \tag{6}$$

where $P(\mathbf{X} \mid \boldsymbol{\alpha})$ can be calculated with Equation 3, and $P(\mathbf{Z} \mid B_0, \Theta) = \prod_{i=1}^{N} P(\mathbf{z_i} \mid B_0, \Theta)$ with each factor $P(\mathbf{z_i} \mid B_0, \Theta)$ being a Gaussian distribution. As a consequence of the deterministic relationship $\boldsymbol{\alpha_i} = \exp(B^T \mathbf{m_i} + \mathbf{z_i})$, it should be noted that the distribution $P(\boldsymbol{\alpha} \mid \mathbf{Z}, B, B_0, M)$ is a Dirac delta function. In general, associations among microbes are not expected to be dense and only a few environmental factors will predominate. This motivated us to identify a sparse association network which could be effectively achieved by sparse learning techniques (Tibshirani, 1996). Also, in practice, the number of samples is usually smaller than the number of microbes, or $N<<P$. Therefore, introducing sparsity regularization helps avoid overfitting. Specifically, we estimate the sparse association network by solving the following problem:

$$\min_{B, B_0, \Theta, Z} f(B, B_0, \Theta, \mathbf{Z}) + \frac{\lambda_1}{2} \| \Theta \|_1 + \lambda_2 \| B \|_1 \tag{7}$$

where $f(B, B_0, \Theta, \mathbf{Z}) = -(1/N) \log P(\mathbf{Z} \mid \mathbf{X}, \mathbf{M}, B, B_0, \Theta) = -(1/N) \sum_{i=1}^{N} (\sum_{j=1}^{P} \tilde{\Gamma}(\alpha_{ij} + x_{ij}) - \tilde{\Gamma}(s(\alpha_i) + s(x_i)) - \sum_{j=1}^{P} \tilde{\Gamma}(\alpha_{ij}) + \tilde{\Gamma}(s(\alpha_i)))$ $-(1/2) \log |\Theta| + (1/2N) \sum_{i=1}^{N} (z_i - B_0)^T \Theta (z_i - B_0)$, $\tilde{\Gamma}(\cdot) = \log \Gamma(\cdot)$ is the log gamma function, and the positive parameters λ_1 and λ_2 are used to control the sparsity of the solution with larger values representing sparser results. Then, the model parameters can be estimated by optimizing the objective function with respect to Z,B,B_0 and Θ alternately.

1) For $\mathbf{Z}$, we minimize the objective function in Equation 7 with respect to $\mathbf{Z}$. Because of independence, we can solve for each $\mathbf{z_i}$ independently by the gradient descent methods. Here, we adopt the limited-memory quasi-Newton (L-BFGS) algorithm (Liu and Nocedal, 1989), which is a quasi-Newton method and converges fast. L-BFGS requires the derivative of z_{ij}, which is computed as follows:

$$\frac{\partial f}{\partial z_{ij}} = -\frac{1}{N} \left(\tilde{\Gamma}'(\alpha_{ij} + x_{ij}) - \tilde{\Gamma}'(s(\boldsymbol{\alpha_i}) + s(\mathbf{x_i})) - \tilde{\Gamma}'(\alpha_{ij}) - \tilde{\Gamma}'(s(\boldsymbol{\alpha_i})) \right) \alpha_{ij} + \frac{1}{N} \Theta_{j:} (\mathbf{z_i} - B_0) \tag{8}$$

where $\tilde{\Gamma}'(\alpha_{ij})$ is the digamma function and Θ_j is the j-th row of the matrix Θ.

2) For B, we minimize Equation 7 with respect to B. The objective is not differentiable by the existence of the L_1 norm regularizer. Therefore we use the orthant-wise limited-memory quasi-Newton (OWL-QN) algorithm (Andrew and Gao, 2007), which is based on L-BFGS and can minimize the log likelihood function with L_1 regularization for optimization. The derivative of B_{ij} is

$$\delta_{ij}(B) = \begin{cases} \partial_{ij}^{-} f(B) & \text{if } \partial_{ij}^{-} f(B) > 0 \\ \partial_{ij}^{+} f(B) & \text{if } \partial_{ij}^{+} f(B) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where

$$\partial_{ij}^{\pm} f(B) = \frac{\partial f(B)}{\partial B_{ij}} + \begin{cases} \lambda_2 sign(B_{ij}) & \text{if } B_{ij} \neq 0 \\ \pm \lambda_2 & \text{if } B_{ij} = 0 \end{cases},$$

and $(\partial f(B) / \partial B_{ij}) = (1/N) \sum_{k=1}^{N} (\tilde{\Gamma}'(\alpha_{ij} + x_{ij}) - \tilde{\Gamma}'(s(\boldsymbol{\alpha_i}) + s(\mathbf{x_i})) - \tilde{\Gamma}'(\alpha_{ij}) + \tilde{\Gamma}'(s(\boldsymbol{\alpha_i}))) \alpha_{kj} m_{kj}$.

3) For B_0, we have the update rule $B_0 = (1/N) \sum_{\mathbf{i}=1}^{N} \mathbf{z_i}$, which is the mean of the latent vectors $\mathbf{z_i}$.
4) For Θ, this step is equal to solving the classical problem of a graphical lasso (glasso):

$$\min_{\Theta} -\log |\Theta| + tr(S\Theta) + \lambda_1 \| \Theta \|_1, \tag{10}$$

where the empirical covariance $S = (1/N) \sum_{i=1}^{N} (\mathbf{z_i} - B_0)(\mathbf{z_i} - B_0)^T$. This problem is also termed as sparse inverse covariance estimation and can be solved with a standard graphical lasso (glasso) algorithm by (Friedman et al., 2008). However, different from the fully observed glasso, where the empirical covariance is computed once, we should note that our S depends on the inferred latent vectors $\mathbf{z}$ and needs to update at each iteration. Since $\mathbf{z_i}$ and $\mathbf{m_i}$ mutually influence each other in explaining the observed data $\mathbf{x}$ (see the Figure 1B), the learned sparse graph (i.e., Θ) is affected by environmental factors, matching our intuition in Figure 1C.

For model selection, we choose the best parameters for λ_1 and λ_2 via extended Bayesian information criteria (EBIC) (Chen and Chen, 2008). EBIC improves the original BIC by assigning larger prior to lower dimension models, a strategy more suitable for model selection in large model spaces.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data Generation and Evaluation Metrics in Synthetic Experiment

The synthetic data can be naturally produced via our generative process. First, the environmental factor matrix **M** is sampled from the multivariate normal distribution $N(\mathbf{0}, \mathbf{I})$ and then normalized with $\sum_{i=1}^{N} M_{ij} = 0$ and $(1/N-1)\sum_{i=1}^{N} M_{ij}^2 = 1$. The element B_{ij} of matrix B is sampled from the uniform distribution of $[-0.5, 0.5]$ and set to 0 with probability of 0.85. Since dominant microbes are found in some microbial communities, we produce vector B_0 by uniformly sampling from $[6, 8]$ with probability of 0.2 and $[2, 4, 5]$ with probability of 0.8 to affect the distribution of absolute abundance of microbes. To evaluate the ability of mLDM to recover network structures, we follow Kurtz et al. (2015) and use five different precision matrices Θ whose adjacency matrices are as follows:

Random Graph: Edge $e_{ij}^{(1)}$ in $E^{(1)}$ is set to nonzero with probability $(3/P)$ and about $(3/2)(P-1)$ edges are produced.
Cluster Graph: Nodes $V^{(1)}$ are randomly split into $\lfloor P/20 \rfloor$ groups and within the same group the nodes i and j are connected with probability of 0.3.
Scale-free Graph: The B-A algorithm (Albert and Barabasi, 2001) is used to produce a graph in which **a)** initially two nodes in $G^{(1)}$ are connected and **b)** every new node is added in by linking to a node in the current graph with probability proportional to the degree of the node.
Hub Graph: Nodes $V^{(1)}$ are randomly split into $\lfloor P/20 \rfloor$ groups, and within the same group, every node is connected with a center node with probability of 1. Finally, random $P - \lfloor P/20 \rfloor$ edges are included in the $E^{(1)}$.
Band Graph: Each adjacent node pair i and j in $V^{(1)}$ is connected if $|i-j| = 1$ and $P-1$ edges are generated in $E^{(1)}$.

We use the huge package (Zhao et al., 2012) to generate Θ and obtain the positive definite covariance matrix $\Sigma = \Theta^{-1}$. In order to make the covariance matrix Σ sparse, and thus beneficial to methods estimating the correlations, we set $\Sigma_{ij} = 0$ if $|\Sigma_{ij}| < 0.1$. Then, $\mathbf{z_i}$ is sampled from the normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\alpha_i}$ is calculated via Equation 5. Next, we generate the Dirichlet-multinomial samples $\mathbf{x_i}$ from Equation 3. This process relies on the R package 'HMP', which includes the generation of Dirichlet-multinomial samplers. For B, B_0 and Θ with five structures, all methods are compared with the following four experimental settings: P = 50, Q = 5 and N = 25,50,200 and 500. We use public codes glasso, CCREPE, SPIEC-EASI, CCLasso and the implementation of SparCC in SPIEC-EASI. Here PCC and SCC are implemented in R language, and the candidates of associations are selected via p value. We set p value at 0.05 for PCC, SCC and CCREPE, and the threshold of correlation for SparCC is 0.1. For each parameter setting, we randomly generate 20 sets of data for evaluation. For all experimental results, it should be noted that we show the mean and variance of evaluation results from the 20 synthetic datasets.

We use three metrics for evaluation:

ROC curve: We plot the ROC curves using two criteria. For PCC, SCC, CCREPE, SparCC and CCLasso, which estimate pairwise correlations, we compare their results with the true correlation matrix $\boldsymbol{\rho}$ with each element being $\rho_{ij} = (\Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}})(i<j)$. For glasso, SPIEC-EASI and mLDM, which estimate conditional independence, we compare their results with the true precision matrix Θ.
AUC score: We compute the area under the ROC curves directly. The AUC scores are calculated by ignoring the sign of edges.
Δ_1 **distance**: It is defined as the L_1 distance between the estimated edge weights and the true weights in the graph. A smaller Δ_1 distance indicates a higher accuracy. Let $\Delta_1^{(1)}$ and $\Delta_1^{(2)}$ denote the Δ_1 distance for the OTU-OTU and EF-OTU association graphs, respectively. For the pairwise correlation methods, $\Delta_1^{(1)} = (2/P(P-1))\sum_{i<j}|\widehat{\rho}_{ij} - \rho_{ij}|$, where $\widehat{\rho}$ is the estimated value and ρ is the true value. For the conditional independence methods, $\Delta_1^{(1)} = (2/P(P-1))\sum_{i<j}|\widehat{\Theta}_{ij} - \Theta_{ij}|$, and $\Delta_1^{(2)} = (1/QP)\sum_{i=1}^{Q}\sum_{j=1}^{P}\left|\widehat{B}_{ij} - B_{ij}\right|$.

Preprocessing of TARA Oceans Eukaryotic Data

The TARA Oceans eukaryotic OTU table and environmental data, including the known genus-level eukaryotic symbiotic interactions were downloaded from the PANGAEA website (https://doi.pangaea.de/10.1594/PANGAEA.843018) and the TARA OCEANS project website (http://www.raeslab.org/companion/ocean-interactome.html). A total of 91 genus-level mapped eukaryotic symbiotic interactions that consist of both parasitism and mutualism were collected based on the literature (Lima-Mendez et al., 2015) and were used to evaluate the effectiveness of all methods. Samples with missing environmental factor values or with too large or small read counts were removed. OTUs that appear in less than 40% of the samples were omitted. For comparison, we chose OTUs that were involved in known genus-level symbiotic interactions. Finally we constructed a dataset consisting of 67 OTUs with 28 known genus-level interactions and 17 environmental factors from 221 samples for evaluation.

Preprocessing of Colorectal Cancer Data

We adopted the dataset directly from Baxter et al. (2016) and downloaded the OTU and meta data from the github (https://github.com/SchlossLab/Baxter_glne007Modeling_GenomeMed_2015). We selected a total of 117 OTUs, including 112 that existed in at least half of all 490 samples, and 5 that were CRC-associated OTUs reported in the article, including *Prevotella* (OTU57), *Porphyromonas* (OTU105), *Fusobacterium* (OTU264), *Parvimonas* (OTU281) and *Peptostreptococcus* (OTU310).

Preprocessing of West English Channel Data

For the West English Channel data, we downloaded the OTU table from the VAMPS website (https://vamps.mbl.edu/). Forty-seven samples from position L_4 $(50°25.18'N, 4°21.89'W)$ were selected for association estimation. We extracted 48 OTUs that appeared in

at least 46 samples, and the total abundance of these OTUs exceeds 50% of the total read counts. This dataset has 8 EFs, including *temperature*, *day length*, as well as concentrations of *salinity*, *ammonia*, *chlorophyll*, *nitrate*, *phosphate* and *silicate*, which were used to infer EF-OTU associations.

DATA AND SOFTWARE AVAILAVILITY

The program of mLDM is freely available at https://github.com/tinglab/mLDM/. Now, for the synthetic dataset with 50 OTUs, 5 EFs and 500 samples, mLDM runs about 20 min on a server with an Intel Xeon v3 2.5GHz CPU and 128G RAM.