Inferring Multiple Metagenomic Association Networks based on Variation of Environmental Factors

Yuqing Yang^{1,#,a}, Xin Wang^{3,#,b}, Kaikun Xie^{1,c}, Congmin Zhu^{1,d}, Ning Chen^{1,*,e}, Ting Chen^{1,2,4,*,f}

¹*Department of Computer Science and Technology and Institute of Artificial Intelligence, Tsinghua University, Beijing 100084, China*

²*Department of Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, School of Basic Medical Sciences, Fujian Medical University, Fuzhou 350122, China*

³*Department of Ultrasound, Peking Union Medical College Hospital, Beijing 100005, China*

⁴*Beijing National Research Center for Information Science and Technology, Beijing 100084, China.*

Equal contribution

* Corresponding authors :

E-mail: ningchen@tsinghua.edu.cn (Chen N), tingchen@tsinghua.edu.cn (Chen T).

Running title: *Yang Y et al / Inference of Multiple Association Networks*

26

27 **Abstract**

28 Identifying significant biological relationships or patterns is central to many
 29 metagenomic studies. Methods that estimate association networks have been proposed
 30 for this purpose, but they assume that associations are static, neglecting the fact that
 31 relationships in a microbial ecosystem may vary with changes in environmental
 32 factors, which can result in inaccurate estimations. We propose a computational
 33 model, k-Lognormal-Dirichlet-Multinomial model (kLDM), which estimates multiple
 34 association networks that correspond to specific environmental conditions according
 35 to values of environmental factors (EFs), and simultaneously infers microbe-microbe
 36 and environmental factor-microbe associations for each network. We showed the
 37 effectiveness of kLDM on synthetic data, a colorectal cancer dataset, as well as the
 38 TARA Oceans and American Gut project datasets. The results showed that the widely
 39 used Spearman's rank correlation coefficient (SCC) performed much worse than other
 40 methods, indicating the importance of separating samples by environmental
 41 conditions. We compared cancer fecal samples with cancer-free samples, and our
 42 estimation showed fewer associations among microbes but stronger associations
 43 between specific bacteria such as five colorectal cancer (CRC)-associated OTUs,
 44 indicating gut microbe translocation in cancer patients. Some
 45 environmental-factor-dependent associations were found within marine eukaryotic
 46 community, and gut microbial heterogeneity of irritable bowel disease (IBD) patients
 47 was detected. Results demonstrated that kLDM could successfully unravel the
 48 underlying biological associations. In summary, our study presents a computational
 49 framework that can elucidate the complex associations within microbial ecosystems.
 50 The kLDM program, R, and python scripts, together with all experimental datasets are
 51 all accessible at Github (<https://github.com/tinglab/kLDM.git>).

52 **KEYWORDS:** Metagenomics; Association inference; Environmental condition;
 53 Bayesian model; Clustering

54

55

56 **Introduction**

57 Microbes interact closely with the human body and the environment human lives,
 58 which affect our health and ecosystem by changing their microbiome and metabolome
 59 [1–3]. Metagenomic high-throughput sequencing technology plays a vital role in the
 60 study of microbes inhabiting various human body sites and different natural
 61 ecological environments. Such tool allows us to analyze microbial constituents of the
 62 microbiota and microbiome and interactions within a particular microbial community
 63 among microbes and between microbes and environmental factors (EF). The
 64 environmental factors, such as states of human diseases, genotypes of hosts, values of
 65 some physiological and biochemical indicators, quantization of people’s lifestyle
 66 factors, and concentrations of nutrients, are known to be associated with variation of
 67 abundance of microbes [4–8]. Recently, we have observed rapid increase of large
 68 scale metagenomic studies aiming to discover biological interactions [9,10],
 69 specifically, how microbes interact with other microbes and EFs.

70 Biological interactions can be considered as positive and negative relationships
 71 among microbes and between EFs and microbes, which can be inferred indirectly by
 72 predicting patterns of variation of microbial counts and values of EFs using
 73 association inference methods. A positive relationship may imply commensalism
 74 among microbes or microbial dependency on some EFs, whereas a negative
 75 relationship can be parasitism, competition among microbes, or inhibition of EFs on
 76 microbes [11]. Furthermore, interactions among microbes and between EFs and
 77 microbes depend on *environmental conditions* (**EF conditions**), which, in this article,
 78 are defined as some specific ranges of values of EFs. Under similar EF conditions,
 79 interactions in the microbial community can be stable, while they can change
 80 dynamically with the alteration of EF conditions (**Figure 1A**). For example, marine
 81 microbial interactions may vary by season and depth [12], and interactions among gut
 82 microbes can change according to the onset of diseases and disease states [4,13]. In

these two examples, the environmental conditions are specific seasons and ocean depths, and specific diseases and disease states, respectively. Overall, biological interactions are dynamic, a phenomenon that should be considered in any approach to association inference.

However, environmental conditions are sometimes neither obvious, nor are they known in advance in some studies. For example, the American Gut project[14] collected information on dozens of human lifestyle factors related to individuals' diet and living habits, as well as metagenomic sequencing data for thousands of individuals. However, it is unclear which subset of these individuals belong to the same environmental condition, which, by definition, not only have similar values of some, or all, of these lifestyle factors but also share identical microbial interactions. Besides, even if we have obtained hosts' disease states, grouping samples into the cases and controls while ignoring other environmental factors may be rough, because the underlying patterns of microbial compositions and associations may be very complex, consist of many subgroups, and can be aligned with neither the cases nor the controls. These problems call for novel computational methods to discover potential environmental conditions and infer association networks together.

Due to a limited number of sequencing samples in studies, previous association inference methods estimate static association network using all samples [15–20]. These methods can be classified into two categories: those that compute pairwise associations independently, and those that estimate multiple associations simultaneously. The widely used Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SCC) belong to the former, as well as the local similarity association (LSA) [19]. In contrast, CCREPE[16], SparCC[15], SPIEC-EASI[20], CCLasso[18] and mLDM[17] belong to the latter. The latter methods consider compositional bias [21] caused by the normalization process of microbial read counts, which divides microbial read counts over total sum of read counts to obtain relative abundance. Denote the relative abundance of the j^{th} OTU as $r_j = \frac{x_j}{\sum_k x_k}$ where the x_j is its read count and we get $\sum_{k \neq j} Cov(r_k, r_j) =$

112 $-\text{Var}(r_j)$. It can be seen that the commonly proposed normalization process [22]
 113 introduces dependency into microbial relative abundance, and as such, it reduces the
 114 efficiency of association inference (Figure 1D). CCREPE, SparCC and CCLasso
 115 estimate OTU-OTU (microbe-microbe) correlations by computing covariance among
 116 microbes, while SPIEC-EASI and mLDM infer conditionally dependent OTU-OTU
 117 associations by estimating the precision matrix among microbes. With the exception
 118 of mLDM, none of these methods considers associations between EFs and microbes.
 119 Taking both compositional bias and the large variance of read counts into account, our
 120 previous work, mLDM [17] estimates both OTU-OTU and EF-OTU associations
 121 more accurately by removing indirect associations among microbes induced by
 122 common EFs (Figure 1E).

123 However, all methods noted above assume only one biological network,
 124 neglecting that biological interactions can be different with the change of EFs. In this
 125 case, if samples from two different environmental conditions are combined to infer
 126 associations, the results will reflect the change of environmental conditions, instead of
 127 real associations under these two environmental conditions, which may lead to false
 128 associations and conclusions. For example, association networks of gut microbes in
 129 patients with liver cirrhosis and healthy individuals are distinctly different [23], and to
 130 predict gut microbial interactions of disease patients, samples from the healthy
 131 population should be excluded.

132 In order to estimate OTU-OTU associations and EF-OTU associations conditional
 133 on environmental conditions, we propose a new hierarchical Bayesian model,
 134 k-Lognormal-Dirichlet-Multinomial model, called kLDM (Figure 1), which can
 135 determine the number of environmental conditions automatically and infer complex
 136 associations simultaneously. Under every EF condition, kLDM considers the
 137 compositional bias and large variance of read counts, and it estimates both
 138 conditionally dependent OTU-OTU associations and direct EF-OTU associations.
 139 Most significantly, associations that vary according to environmental conditions can
 140 be elaborated by kLDM. It should be noted that the direct or conditionally dependent

141 associations are mathematical concepts that approximate biological interactions,
 142 rather than indicating direct biological interactions or causal relationships. In addition,
 143 considering sequencing of marker genes such as 16S rRNA/18S rRNA genes is
 144 widely used to investigate microbial compositions of samples, kLDM is designed for
 145 data produced by such marker gene sequencing technology. To our best knowledge,
 146 kLDM is the first method to infer multiple association networks based on variation in
 147 the values of environmental factors. Here, we validate the efficiency and robustness
 148 for sparse and compositional data of kLDM on synthetic datasets via comparisons
 149 with state-of-art approaches, then apply it to colorectal cancer (CRC) data to show its
 150 ability to find well-defined EF conditions, and explore its application on TARA
 151 Oceans data and American Gut project data to discover potential EF conditions and
 152 novel biological relationships.

153 **Materials and methods**

154 **Materials used for kLDM evaluation**

155 *Gut microbial samples from Colorectal Cancer study*

156 OTU table and meta data of metagenomic gut 16S rRNA sequencing data of healthy
 157 individuals and patients with Colorectal Cancer (CRC) were obtained from Baxter et
 158 al. [13]. We constructed a dataset which consisted of 5 known CRC-associated OTUs
 159 (*Peptostreptococcus*, *Parvimonas*, *Fusobacterium*, *Porphyromonas*, and *Prevotella*)
 160 reported in the original article, 112 common OTUs, as observed in more than half of
 161 all 490 samples, and 4 EFs, namely FIT result, age and gender, and diagnostic states
 162 of donors. The 'FIT' result refers to the positive or negative result of fecal
 163 immunochemical test. The recorded diagnostic states include 'Normal', 'High Risk
 164 Normal', 'Adenoma', 'Advanced Adenoma', and 'Cancer', all of which are determined
 165 by colonoscopy examination and review of biopsies, were also recorded. The
 166 diagnostic states were modeled as five numerical values from 1 to 5, with a higher
 167 value representing a more serious disease state.

168 *TARA Oceans eukaryotic dataset*

169 In the study of **TARA Oceans** project about marine eukaryotes [24], read counts of
 170 OTUs were obtained by sequencing the V9 region of 18S rRNA genes of eukaryotes,
 171 and 91 genus-level matched eukaryotic symbiotic interactions were summarized.
 172 Because of the limited number of samples, after filtering OTUs that exist in < 40% of
 173 samples and getting rid of samples with missing EFs or abnormal counts, a subset
 174 including 67 OTUs related to genus-level symbiotic interactions and 17 EFs from 221
 175 samples was constructed. The 17 EFs are the depth of water, chlorophyll maximum,
 176 maximum of Brünt-Väisälä frequency, maximum dissolved oxygen, and minimum
 177 dissolved oxygen; concentrations of salinity, oxygen, phosphate, silica, and
 178 chlorophyll; temperature, sunshine duration, moon phase, maximum Lyapunov
 179 exponent, residence time, latitude and longitude.

180 *16S rRNA sequencing samples from American Gut project*

181 OTU table and meta data from the American Gut project[14] were downloaded from
 182 the ftp site (<ftp://ftp.microbio.me/AmericanGut>), and 22 meta data about individuals'
 183 diet and living habits were selected. Among 22 meta data, three factors were
 184 associated with living habits (alcohol, exercise, and smoking frequencies), the
 185 remaining 19 factors were connected with diet (frequencies of fermented plant, frozen
 186 dessert, fruit, high fat red meat, home-cooked meals, meat eggs, milk cheese, milk
 187 substitute, olive oil, probiotic, red meat, salted snacks, seafood, vegetable, vitamin D
 188 supplement, vitamin B supplement, whole grain, and whole eggs). The value of meta
 189 data can be categorized as one of 'Never', 'Rarely (less than once/week)', 'Occasionally
 190 (1-2 times/week)', 'Regularly (3-5 times/week)' and 'Daily'. For convenience, we
 191 recode these categories into integers from 1 to 5 according to their frequencies. OTUs
 192 are filtered based on their average size and non-zero times in all samples. Samples
 193 with huge or small numbers of reads, as well as those with evenness < 2, are removed.
 194 Finally, 11,946 samples with 216 OTUs and 22 EFs are obtained. The python script to
 195 process the OTU .biom file can be found at the Github
 196 (<https://github.com/tinglab/kLDM.git>). Information on individuals' disease status was
 197 also extracted, and samples with diseases were labeled, including cardiovascular

disease, small intestinal bacterial overgrowth, mental illness, lactose intolerance, diabetes, inflammatory bowel disease (IBD), irritable bowel syndrome, *C. difficile* infection, cancer, and obesity.

Assumptions of kLDM hierarchical Bayesian model

The kLDM model assumes that interactions among microbes are regulated by environmental factors and that they tend to be constant when environmental conditions are identical, but may change as environmental conditions change (Figure 1A). Or in other words, kLDM accounts for variation in the values of EFs. In the same environmental condition, EF values fluctuate, but only within a small range, and core microbes stay the same, their associations remaining stable. However, when environmental conditions vary, values of EFs, microbial species, and their associations can all change. In addition, considering that distribution of environmental conditions may be continuous and complex [25], kLDM uses mixtures of multiple clusters with known distributions to approximate and capture main patterns, and each cluster can represent an environmental condition.

The hierarchical structure of kLDM model.

Let $X = \{x_i\}_{i=1}^N$ be N sequencing samples where $x_i \in \mathbb{N}^P$ is the i^{th} sequencing sample with P microbes, and x_{ij} corresponds to the read count of the j^{th} microbe or Operational Taxonomic Unit (OTU). Values of environmental factors are represented as $M = \{m_i\}_{i=1}^N$, where $m_i \in \mathbb{R}^Q$ is a Q -dimensional vector and m_{ij} records the value of environmental factor j of the i^{th} sample. kLDM models the EF vector m_i as a multivariate Gaussian distribution, and assumes that EFs and whole data consist of K mixtures of Gaussian components. Thus, samples in which EF vectors belong to the same Gaussian component can be viewed as drawn from the same environmental condition and share identical OTU-OTU and EF-OTU association networks. K environmental conditions are denoted as

224 $C = \{B^{(k)}, \Theta^{(k)}, B_0^{(k)}, \mu^{(k)}, \Sigma^{(k)}\}_{k=1}^K$, where matrices $B^{(k)}$ and $\Theta^{(k)}$ are the
 225 associations between EFs and OTUs and among OTUs for the k^{th} environmental
 226 condition, respectively, $B_0^{(k)} \in \mathbb{R}^Q$ is a basis vector, and $\mu^{(k)}$ and $\Sigma^{(k)}$ are the mean
 227 vector and covariance matrix of EFs for the k^{th} environmental condition. In addition,
 228 weights of K environmental conditions are $\pi = \{\pi_i\}_{i=1}^K$, where π_k is the weight of
 229 the k^{th} EF condition, and $\sum_{k=1}^K \pi_k = 1$.

230 Figure 1C explains the structure of the kLDM model. x_i and m_i represent the
 231 vectors of read counts and values of EFs of the i^{th} sample, respectively. The
 232 vector h_i is a P-dimensional latent variable of which the value h_{ij} is the latent
 233 relative abundance of the j^{th} microbe in sample i , and α_i can be regarded as the
 234 absolute abundance vector corresponding to h_i . kLDM assumes that the absolute
 235 abundance α_i determines the relative abundance h_i of P microbes during the
 236 sample preparation process, and that the microbial read counts vector x_i is generated
 237 based on relative microbial abundance during the sequencing process. Sample i can
 238 be seen as being collected from a specific environmental condition denoted by c_i
 239 with mixture weight π_{c_i} . The absolute abundance α_i is decided by two sources of
 240 associations: 1) associations between environmental factors and microbes under the
 241 c_i^{th} environmental condition, which can be parameterized by a linear term $B^{(c_i)^T} m_i$;
 242 and 2) associations among microbes, the effects of which are included in the EF
 243 condition-specific latent variable $z_i^{(c_i)}$. The vector $z_i^{(c_i)}$ follows a multivariate
 244 Gaussian distribution, the parameters of which are determined by the basis vector

245 $B_0^{(c_i)}$ and the precision matrix $\Theta^{(c_i)}$, which records OTU-OTU associations at the EF
246 condition c_i .

247 The generative process of the hierarchical Bayesian model, kLDM, is shown,

$$\begin{aligned} c_i &\sim \text{Categorical}(\pi) \\ m_i | c_i &\sim N(\mu_{c_i}, \Sigma_{c_i}) \\ z_i^{(c_i)} | c_i &\sim N(B_0^{(c_i)}, \Theta^{(c_i)^{-1}}) \\ \alpha_i | c_i &= \exp(B^{(c_i)^T} m_i + z_i^{(c_i)}) \\ h_i &\sim \text{Dirichlet}(\alpha_i) \\ x_i &\sim \text{Multinomial}(h_i) \end{aligned}$$

249 where $B^{(c_i)}$ is a $Q \times P$ matrix where $B_{qp}^{(c_i)}$ is the association between the q^{th}
250 environmental factor and the p^{th} microbe, and $\Theta^{(c_i)} \in \mathbb{R}^{P \times P}$ is a $P \times P$ inverse
251 covariance matrix. During the sequencing process, millions of DNA molecules are
252 selected randomly from the DNA library; therefore, we let the sequencing counts x_i
253 be a multinomial distribution,

$$P(x_i | h_i) = \binom{\sum_{j=1}^P x_{ij}}{x_{i1}, \dots, x_{iP}} \prod_{j=1}^P h_{ij}^{x_{ij}}$$

254 where the relative abundance h_i represents the microbial relative ratios within the
255 DNA library and $\sum_{j=1}^P h_{ij} = 1$. Considering the conjugacy of Dirichlet and
256 multinomial distribution, h_i follows the Dirichlet distribution,

$$P(h_i | \alpha_i) = \frac{\Gamma(\sum_{j=1}^P \alpha_{ij})}{\prod_{j=1}^P \Gamma(\alpha_{ij})} \prod_{j=1}^P h_{ij}^{\alpha_{ij}-1}$$

257 where α_i is the vector for absolute microbial abundance in the i^{th} sample, which, in
258 turn, determines the relative ratios in the library. In Dirichlet-multinomial distribution,

the covariance of two OTUs' counts x_{ij} and x_{ik} in i^{th} sample is $\text{Cov}(x_{ij}, x_{ik}) \propto$
 $-\sum_{j=1}^P x_{ij} \frac{\alpha_{ij}}{\sum_{j=1}^P \alpha_{ij}} \frac{\alpha_{ik}}{\sum_{j=1}^P \alpha_{ij}}$, which corresponds to a negative bias, and is regulated by
both the sequencing depth $\sum_{j=1}^P x_{ij}$ and OTUs' relative abundance $\frac{\alpha_{ij}}{\sum_{j=1}^P \alpha_{ij}}$. The
Dirichlet-multinomial distribution is integrated into the hierarchical structure of
kLDM to model the compositional characteristics of sequencing count data.

Assuming that microbes in the i^{th} sample come from environmental condition c_i ,
their absolute abundance α_i is affected by both environmental factors m_i and
interactions within the community. We model this into a lognormal distribution which
is suitable for most microbial abundance [26,27]:

$$P(\alpha_i | c_i, B^{(c_i)}, \Theta^{(c_i)}, B_0^{(c_i)}) \\ = \frac{1}{(2\pi)^{\frac{P}{2}} \Theta^{\frac{1}{2}}} \times \exp(-\frac{1}{2}(\ln \alpha_i - \mu_i^{(c_i)})^T \Theta (\ln \alpha_i - \mu_i^{(c_i)})) \prod_{j=1}^P \frac{1}{\alpha_{ij}}$$

where $\mu_i^{(c_i)} = B^{(c_i)T} m_i + B_0^{(c_i)}$. This can be simplified into the following formula
using the relationship between the lognormal and normal distributions,

$$\alpha_i | c_i = \exp(B^{(c_i)T} m_i + z_i^{(c_i)})$$

where $z_i^{(c_i)} | c_i \sim N(B_0^{(c_i)}, \Theta^{(c_i)^{-1}})$.

Under the c_i^{th} environmental condition, m_i follows the multivariate Gaussian
 $P(m_i | c_i) = N(\mu_{c_i}, \Sigma_{c_i})$ and the weight of the c_i^{th} EF condition is set to the
categorical distribution with the parameter π as,

$$P(c_i = k) = \prod_{c=1}^K \pi_c^{I(c_i=k)}.$$

We are interested in the parameters $B^{(k)}, \Theta^{(k)} (k = 1, \dots, K)$, where $B_{ji}^{(k)}$ infers
direct association between the i^{th} microbe and j^{th} environmental factor at the k^{th}

277 environmental condition, and $-\frac{\theta_{ij}^{(k)}}{\sqrt{\theta_{ii}^{(k)}\theta_{jj}^{(k)}}}$ is the conditionally dependent association

278 between the i^{th} and j^{th} microbes (Figure 1G).

279 **Sparse association inference of kLDM in theory**

280 The generative model can be solved theoretically via Expectation-Maximization (EM)

281 algorithm and maximum a posteriori (MAP) estimation for the latent variable $z_i^{(c_i)}$.

282 We assume that associations among microbes and between environmental factors and

283 microbes are sparse and can be inferred by kLDM with sparsity constraints. Detailed

284 formulas of inference and optimization can be found in the **Supplementary**

285 **Information.**

286 However, two potential problems confront this theoretical sparse association

287 inference, thereby limiting the practicality of the model. First, results are very

288 sensitive to the initialization of parameters because the EM algorithm can converge to

289 a local minimum. Second, estimating the number of EF conditions is very

290 time-consuming. Therefore we explored more effective approaches and finally

291 adopted an efficient split-merge algorithm as detailed below.

292 **Implementation of split and merge Algorithm for kLDM**

293 kLDM adopts a split-merge algorithm to estimate the number of environmental

294 conditions and sparse OTU-OTU and EF-OTU associations under each environmental

295 condition [28]. First, we partition samples into fine-grained clusters using values of

296 environmental factors, such that samples within a cluster can be regarded as

297 belonging to the same environmental condition. Second, since this partition is not

298 perfect, we merge multiple clusters into one if they share similar environmental

299 conditions and association networks. The final output is a set of sample groups, each
300 with distinct predicted association networks.

301 More specifically, the split process starts from a single cluster with all samples,
302 and then iteratively selects a cluster and partitions it into two clusters until the number
303 of samples in each cluster is smaller than a pre-defined threshold N_{min} . This process
304 results in the construction of a binary tree for environmental factors, each node of
305 which corresponds to one cluster of the samples, and each leaf node of which
306 corresponds to a set of samples with similar values of environmental factors. We
307 assume that EF vector m_i follows the multivariate Gaussian distribution, and when a
308 cluster is split in two, two Gaussian mixtures are used to model the distribution of EFs
309 of the cluster as,

$$P(m_i) = \pi_1 N(m_i | \mu_1, \Sigma_1) + \pi_2 N(m_i | \mu_2, \Sigma_2)$$

310 where $\pi_1 + \pi_2 = 1$ and $\pi_j (j = 1, 2)$ is the weight of the j^{th} component. After
311 estimating these parameters via the EM algorithm, we re-assign each sample of the
312 original cluster to one of the two new clusters with the larger posterior probability.

313 Then association networks are estimated for the leaves via mLDM [17], which is
314 similar to kLDM when the number of the environmental condition is set to one
315 ($K = 1$). We re-implemented the mLDM algorithm with C++ and OpenMP to
316 improve its stability and efficiency, and a comparison of running time and memory
317 usage by kLDM and mLDM on single association network inference was shown in
318 Table S1. Inferred associations will be used for the merge process.

319 The merge process aims to recover clusters being partitioned into multiple leaf
320 nodes caused by the greedy approach of the split process. The merge process adopts a
321 bottom-to-top strategy, starting from internal nodes at the lowest level and traversing
322 up to the root, to identify leaf nodes for merging. For each internal node, the
323 algorithm traverses down to its left and right branches to search for leaf nodes. The
324 two leaf nodes with the smallest inter-cluster distance, as measured by the Euclidean
325 distance between the mean values of EFs, are merged. kLDM will estimate the
326 associations and the EBIC score for the merged cluster. If its EBIC score is smaller
327 than the sum of the EBIC scores of the two leaves, then the merged cluster is kept and
328 substitutes the one that has closer mean values of EFs, while the other is discarded.
329 Otherwise, the merged cluster is discarded. This split-merge process reduces the time
330 complexity by limiting the operation at each step to partition one large cluster into
331 two, or merge two nearby clusters into one. The algorithm can run in parallel, making
332 kLDM very efficient in practice.

333 **Synthetic data generation process**

334 We generate synthetic datasets by specifying the number of microbes, environmental
335 factors, and clusters, as well as the range of the number of samples per cluster. We
336 then construct samples for every cluster via a generative process of the kLDM model
337 separately. For the i^{th} cluster ($i = 1, 2, \dots, K$), the EF-OTU association matrix B_i is
338 produced by sampling from the interval $[-0.5, 0.5]$ uniformly, with only 15 percent
339 of the elements set to non-zero. The OTU-OTU association matrix Θ_i is generated
340 using the R package ‘huge’ [29], which outputs a precision matrix for which the

adjacency matrix can be random, cluster, scale-free, hub or band graph. Every graph corresponds to a specific association structure among microbes. Values of the mean vector of the environmental factor of the cluster can be obtained by sampling uniformly from the interval $[i, (i + 0.5) * i]$. Then the Dirichlet-multinomial samplers are produced using the R package 'HMP'. For parameters of every setting, ten repetitions are conducted to generate synthetic data and then the mean and standard deviation of evaluation metrics are calculated for comparison. Public R language codes of CCLasso and SPIEC-EASI are used. The implementation of SparCC is provided by SPIEC-EASI. SCC and SCC(all) are implemented directly in built-in functions of R language. The 'mb' (meinshausen-buhlmann) estimation method is set for SPIEC-EASI, and default parameters of CCLasso, SPIEC-EASI, and SparCC are utilized. In addition, the p-value is set to 0.05 for SCC and PCC to save significant associations.

Evaluation metrics on synthetic data

ROC curves and AUC scores are utilized to evaluate the performance of association inference. Every cluster estimated by kLDM is shown by two ROC curves, the ROC curve of the OTU-OTU associations and that of the EF-OTU associations. AUC scores are the area under ROC curves. When we compute AUC scores, signs of estimated associations are neglected. As for thresholds to compute AUCs, the absolute values of calculated associations are selected for kLDM, SparCC, CCLasso, and SPIEC-EASI, and p-values are used by SCC and PCC. When plotting ROC curves for SCC and CCLasso, calculated coefficients are compared with the true

correlation matrix, as deduced by the inverse of the precision matrix Θ_i . One estimated nonzero association is regarded as a true positive if its value is also nonzero in ground truth. If values of one association are both zero in inferred and real results, the association is a true negative.

Results

Comparing kLDM with other methods on synthetic datasets

In the following, we first assess the performance of kLDM using synthetic datasets, and then we evaluate how similarity of underlying networks and missing information, i.e., EFs, affect the performance of kLDM.

Detailed experiments are conducted to show the effectiveness of kLDM by comparing it with Spearman's Correlation Coefficient (SCC), CCLasso, and SPIEC-EASI. These three models are included because they were tested in our previous studies [17] and showed advantages over other methods. SCC estimates both EF-OTU and OTU-OTU associations, and CCLasso solves the covariance matrix among microbes, while SPIEC-EASI performs well in inferring the precision matrix among OTUs. For fair comparison, since none of them considers more than one network, we partition data into clusters according to the results reported by kLDM and apply these programs on each cluster to infer associations. It should be noted that the results of SCC using all samples, denoted as SCC(all), are plotted as the baselines. The ROC curve and AUC score on every cluster are used to compare the performance.

First, we examine the relationship between the number of samples per cluster N , and the efficiency of kLDM by setting the following parameters: $K = 2$ clusters, $P = 50$ microbes, $Q = 5$ environmental factors, and N samples with two ranges denoted as $[N_{min}, N_{max}] = [100, 200]$ and $[200, 400]$. These two settings of N have identical associations of microbes and between environmental factors and microbes. **Figure 2**

shows that the ROC curves and AUC scores of kLDM are consistently better than those of the other methods, and the power of kLDM on recovering OTU-OTU and EF-OTU associations increases with increased N. However, we observe that the ROC curves of SCC (all) in most situations are lower, showing the importance of separating samples by environmental conditions. It should be noted that SPIEC-EASI does not estimate association efficiently due to its strict model selection. Samples from different EF conditions may introduce many noises and disturb results of SPIEC-EASI. Besides, sensitivities and specificities of top associations estimated by the five methods were compared using the synthetic dataset with $K=2$, $P=50$, $Q=5$ and $N \in [100, 200]$ corresponding to Figure 2A. From Table S2, we observe that kLDM infers OTU-OTU or EF-OTU associations with higher sensitivity and superior specificity.

Then, the scalability of kLDM was investigated with $P=100$ and 200 microbes. The results are plotted in Figure S1A and S1B. In both situations, the AUC scores of kLDM are all higher than those of the other approaches, which verifies its fine scalability, owing to the re-implementation of mLDM with C++ language and utilizing parallel computing. Next, we test kLDM by increasing the number of clusters to 3 and 4 (Figure S1C and S1D), and again kLDM achieves the best results.

Because the split process of kLDM is affected by similarity of environmental factors between clusters, we only change the distance between mean values of environmental factors of two clusters and keep other parameters unaltered to examine the performance of kLDM. From Table S3, as expected, when the distance between two EF mean vectors is large enough, such as 1.5 or 2.0 denoted as '**1.5 baseline**' and '**2.0 baseline**', kLDM infers associations accurately. However, the effectiveness of kLDM reduces when the distance becomes smaller, such as 1.0 ('**1.0 baseline**'), especially on the inference of OTU-OTU associations. This can be attributed to the tendency of kLDM to group samples together to infer common associations when two clusters have similar values of environmental factors, but different associations.

417 To test the effect of similarity of EF-OTU or OTU-OTU associations among
418 clusters on the performance of kLDM, we generate two new datasets for each distance
419 ('1.0, 1.5 and 2.0 baseline') by setting the B_i or θ_i of two clusters ($i=1,2$) equal
420 separately ('same EF-OTU' or 'same OTU-OTU' in Table S3). Compared with results
421 of the corresponding 'baseline', AUC scores of 'same EF-OTU' and 'same OTU-OTU'
422 do not change much. From this, we concluded that the similarity of EFs influences
423 kLDM more than the similarity of associations among environmental conditions.

424 In simulated experiments, all EFs are used to estimate association networks,
425 which may not be feasible in reality because some EFs can be missing. Therefore, we
426 simulate new datasets with only partial EFs to test kLDM's effectiveness. In Figure
427 S2, AUC scores of kLDM on two clusters are shown with different proportions (20%,
428 40%, 60% and 80%) of EFs retained to infer association networks. With subsets of
429 EFs, the AUC scores of kLDM decline. This indicates the importance of including as
430 many EFs as possible. Furthermore, with only 60% EFs, kLDM achieves results
431 comparable to those of CCLasso and SCC with the whole dataset, which again shows
432 the superiority of kLDM (Table S4).

433 **kLDM captures variation in associations of gut microbiota of colorectal cancer** 434 **patients**

435 The relationship between human gut microbiota and colorectal cancer has been
436 explored in previous studies [30–32]. We choose the metagenomic 16S rRNA
437 sequencing dataset from Baxter et al. [13], consisting of 117 OTUs and 3EFs, and
438 validate the efficiency of kLDM on capturing the variation of associations in the
439 microbial community. kLDM reported two clusters: Cluster 1 annotated as 'Cancer',
440 which includes 90% of 'Cancer' samples with significantly higher positive FIT values
441 (P-value 7.18×10^{-22}) (Table S5), and Cluster 2 denoted as 'Healthy', which contains
442 83.7% of healthy samples ('Normal' and 'High Risk Normal') (Table S6). 'Adenoma'
443 and 'Advanced Adenoma' samples exist in both Cluster 1 and Cluster 2. The
444 clustering result of kLDM does not split samples simply according to the diagnostic

state; instead, it takes all EF values, abundances and associations of microbes into account.

Different patterns of microbial abundance and associations are found between these two clusters. *Prevotella* is abundant in samples from both clusters, while the other four CRC-associated OTUs are significantly over-represented (P-value < 1e-3) in samples of Cluster 1 (Figure S3). Comparing OTU-OTU association networks between Cluster1 and Cluster2, as shown in Figure S4, most gut microbes of ‘Healthy’ patients (Figure S4B) connect quite closely, and the distribution of associations among microbes is balanced, while few associations are observed between the aforementioned 5 known CRC-associated OTUs and other microbes. In contrast, associations among gut microbes of ‘Cancer’ patients (Figure S4A) are relatively sparse, and strong correlations are found between specific bacteria (**Table 1**). *Peptostreptococcus*, *Parvimonas*, *Fusobacterium*, and *Porphyromonas* show strong correlations with each other, but don’t connect with other microbes, while *Prevotella* is more associated with other common OTUs, such as *Phascolarctobacterium* and *Clostridium_XIVa*. Based on the difference of microbial abundance and associations’ distribution within two clusters, the gut microbiota translocation in cancer samples can be found, and *Prevotella* may play a specific role in tumorigenesis.

Previous studies have shown that *Peptostreptococcus* and *Fusobacterium* are associated with inflammation [33,34]. Our results also confirm that *Peptostreptococcus* is positively correlated with *Fusobacterium* in ‘Cancer’ patients and that it is over-represented in CRC fecal samples. From Table 1, *Porphyromonas* and *Peptostreptococcus* show positive correlations with diagnostic states in Cluster 1, suggesting the significance of these two bacteria in CRC diagnosis. On the other hand, *Prevotella* is positively associated with the diagnostic states in Cluster 2, indicating a high predictive value, which could be a useful indicator for CRC diagnosis. Research has shown that *Fusobacterium* may act as a passenger microbe to perpetuate tumorigenesis since a higher load of *Fusobacterium* is related to disease severity [35].

473 From our results, patients with cancer showing a positive FIT result are more
474 susceptible to carry *Fusobacterium*.

475 **Elaborating relationships between marine eukaryotes' associations and EFs**

476 Next, we apply kLDM on the TARA Oceans eukaryotic dataset[24] to explore
477 associations in natural environments and compare predicted associations by kLDM
478 with known genus-level interactions. kLDM found two clusters corresponding to two
479 environmental conditions, Cluster 1, consisting of 168 samples with 67 OTUs, and
480 Cluster 2, consisting of 53 samples with 26 OTUs. 41 OTUs in Cluster 2 were filtered
481 out because the number of samples was smaller than the number of OTUs, and kLDM
482 removed small OTUs to infer associations efficiently. EFs of the two clusters are
483 listed in Table S7. Cluster 1 has significantly higher salinity and temperature, but
484 lower oxygen, phosphate and silicate concentration than Cluster 2. Top 1%
485 associations found in these two clusters, including 23 OTU-OTU and 12 EF-OTU
486 associations in Cluster 1, and 4 OTU-OTU and 5 EF-OTU associations in Cluster 2
487 were analyzed, and associations with literature support are listed in **Table 2**.

488 We then matched associations with known genus-level interactions among the
489 top-N associations within each of the two clusters discovered by kLDM. Since known
490 interactions are at the genus-level, we regard one association between two OTUs as
491 matched if the OTUs' genera are identical to two corresponding genera of the known
492 interaction. Simultaneously the results are compared with those that do not consider
493 the environmental condition and regard all samples as one cluster, denoted as 'Static',
494 by limiting kLDM to predict one set of EF-OTU and OTU-OTU association networks.
495 From Table S8, we observe that the combined predicted associations from the two
496 clusters are similar to that of the 'Static', with each cluster consisting of both common
497 and specific OTU-OTU associations.

498 Different types of known associations detected by kLDM are listed in Table S9.
499 We observe that four known associations, '*Phaeocystis* - *Amphibelone*',
500 '*Vampyrophrya* - *Copepoda*', '*Amoebophrya* - *Acanthometra*' and '*Blastodiniaceae* -
501 *Copepoda*', may be dominant in the ocean because they are inferred from both whole

dataset, and the bigger cluster (Cluster 1), and the association between *Amoebophrya* and *Protoperidiniaceae* may be strong in a specific EF condition related to Cluster 2. More specifically, the association between OTU-38 and OTU-25, belonging to the genera *Amoebophrya* and 'Protoperidiniaceae' respectively, only exists in Cluster 2, in which the mean values of oxygen concentration, phosphate, and silica are higher than those in Cluster 1, and the abundance levels of two OTUs are significantly higher than those in Cluster 1 (P-value < 0.05). The EF condition in Cluster 2 could be more suitable for growth of the genus *Protoperidiniaceae*, and then *Amoebophrya* would benefit from parasitism with the former. Based on results from the TARA Oceans dataset, we confirmed the effectiveness of kLDM to elaborate the relationship between OTU-OTU associations and values of EFs.

Characterizing changes in human gut microbes' association networks with different lifestyle factors

kLDM has the advantage of analyzing complex datasets with large samples to infer multiple EF conditions and the corresponding association networks, and we verified this capability on the American Gut project dataset[14]. We applied kLDM to cluster samples according to lifestyle-related factors and association networks, and obtained two big clusters, 'C1' with 6,831 samples and 'C2' with 5,003 samples, and one small cluster 'C3' with 112 samples.

Compared with 'C1' and 'C2', 'C3' exhibits different lifestyle patterns and microbiota distributions (Table S10, S11, and S12). 'C1' and 'C2' together contain almost all disease and healthy people, while 'C3' is mainly composed of IBD patients (94.64%), who make up about a quarter of all IBD patients (26.77%). Individuals in 'C3' have significantly higher alcohol frequency, high fat red meat frequency, and red meat frequency than those in clusters 'C1' and 'C2', but they also eat more vegetables and take vitamin D and B supplements and probiotics more frequently. However, their exercise frequency, milk substitute frequency, and milk cheese frequency are distinctly lower (Table S10). For genus-level microbial abundance, genera '*Prevotella*', '*Ruminococcus*', and '*Sutterella*' are more abundant in 'C3', but

531 ‘*Bifidobacterium*’ and ‘*Bacteroides*’ are less abundant (Table S13). If only
532 considering IBD samples in each cluster, we observe that the IBD samples in ‘C1’
533 and ‘C2’ are substantially different from those in ‘C3’ based on the aforementioned
534 diet frequencies and genus abundance (Table S14). Comparing IBD samples in ‘C1’
535 and ‘C2’, we observe similar lifestyle quantification values and genus abundance
536 levels except vegetable, smoking, fruit, and home-cooked meal frequencies and the
537 abundance of the genus ‘*Lachnospira*’.

538 Then the top 1% EF-OTU associations predicted by kLDM were analyzed (Table
539 S12), and various associations among three clusters matched with published literature
540 are shown. The positive correlation between probiotic frequency and ‘*Bifidobacteria*’
541 is observed in both ‘C1’ and ‘C2’, which was previously reported by Rajkumar et al.
542 [36]. In high animal-protein diet, we observed positive correlations between red meat
543 frequency and ‘*Bacteroides*’, and between poultry frequency and ‘*Ruminococcus*’ in
544 ‘C1’, and between seafood frequency and ‘*Clostridiales*’ in ‘C3’, all consistent with
545 other research [37]. Associations between high-fat red meat frequency and
546 ‘*Bacteroides*’ in ‘C1’ and ‘*Clostridiales*’ in ‘C2’ are matched with rat study showing that
547 intake of a high-fat diet results in disproportionately increases in propionate- and
548 acetate- producing species such as ‘*Clostridiales*’ and ‘*Bacteroides*’[38].

549 Discussion

550 Considering the dynamic nature of microbial interactions, we proposed a new
551 hierarchical Bayesian model, kLDM, to infer associations among microbes and
552 between environmental factors and microbes under different environmental conditions.
553 We then developed two algorithms, a theoretical EM algorithm, and a more practical
554 and efficient split-merge algorithm, to estimate both the number of environmental
555 conditions and associations for each environmental condition simultaneously. The
556 effectiveness of kLDM was verified on simulated datasets, as well as real TARA
557 Oceans, CRC, and American Gut project datasets. Although we implemented kLDM
558 with OpenMP, it can also be accelerated in MPI.

For the synthetic experiment, when we tested kLDM's scalability (Figure S1A and S1B), AUC scores of CCLasso were the second-best in most cases, but CCLasso predicted more associations than kLDM, and its ROC curves increased slowly at the beginning and climbed higher with increase of the false positive rate. However, we think the initial high true positive rate is more crucial for scientists wanting to select candidate interactions for validation; thus kLDM has its advantage here. In addition, influence of the similarity level of EF values between different EF conditions on the performance of kLDM was further explored in detail. We plotted the relationship between AUC scores and the absolute distance (see the formula in Table S3) between EF values of two clusters (Figure S5) based on the 'baseline' dataset used in Table S1. When the distance is small (distance < 2), we observed that AUC scores of OTU-OTU and EF-OTU associations would increase rapidly with the distance, and when distance is larger than two, the power of kLDM tends to be stable. To maintain all AUC scores >0.7, the distance of EF values of two clusters should be larger than 1.0. We have also computed mean values of the inverse Simpson index n_{eff} of the synthetic datasets, and the results are listed in Table S16. Values of n_{eff} of two synthetic datasets with P=100 and 200 are ≥ 13 and have high compositionality [39]. Combining Table S16 with AUC scores of kLDM in Figure S1, we can see kLDM can handle high compositional effect.

On the CRC dataset, the diagnostic state of the donors is included as one EF by kLDM. The diagnostic state was classified according to colonoscopy examination and review of biopsies, instead of the microbial compositions and associations. Therefore, patients with the same diagnostic state may not have the same underlying microbial associations. In addition, cancer may have different subtypes, and each subtype may have a different association network. A patient who will potentially develop to cancer may also have a different association network from those who will not. We also test kLDM on the CRC dataset using three environmental factors excluding the diagnostic state. We found that the two clusters stayed the same, and the compositions of the diagnostic state in the two clusters didn't change. The reason may be that the

underlying microbial networks are distinct between these two clusters. In addition, we also compare kLDM with a probabilistic clustering model, called MicrobeDMM[40], which clusters sequencing samples according to the microbial composition only. From Table S17, we observe that MicrobeDMM doesn't distinguish the healthy from the CRC samples well, and the gut samples of these two groups are mixed evenly into two clusters. In comparison, in the results of kLDM, most of the 'Normal' samples are clustered into one group, while the 'Cancer' samples to the other, which is consistent with our prior knowledge. This shows that using only microbial composition may not be sufficient, and this result, together with the result of kLDM on the synthetic dataset with partially observable EFs 1(Figure S2), proves the importance of collecting of all EFs.

On the TARA Oceans dataset, top associations found by kLDM were associated with previous studies (Table 2). For Cluster 1, the four listed OTU-OTU associations belong to two kinds of known microbial interactions. More specifically, the symbiosis between *Phaeocystis* and *Amphibelone anomala*, and the parasitism between *Amoebophrya ceratii* and *Cochlodinium_01 fulvescens* were captured by kLDM [41,42]. Associations related to values of EFs were also found. *Blastodinium mangini* tended to live in seawater with high salinity, and the concentration of PO₄ affected the growth of *Phaeocystis* [43]. In addition, *Amphibelone anomala* is associated with PO₄ because it has a close phylogenetic relationship with *Pfiesteria piscicida* [44], which is regulated by phosphate [45]. For Cluster 2, the parasite *Blastodinium_06* is linked to temperature and oxygen, and Skovgaard et al. [46] reported that some *Blastodinium spp.* living in warm temperatures can perform photosynthesis. The concentration of oxygen is associated with some *Phaeocystis*, and it was reported that the bloom of *Phaeocystis globosa* caused oxygen depletion [47]. Because most interactions among marine microorganisms are unknown, our explanation is limited.

For the American Gut project dataset, kLDM analyzed the relationship between lifestyle factors and microbiota, as well as associations within the microbial community. The resultant three clusters show differential frequencies of lifestyle

617 factors, the composition of microbes and OTU-OTU, and EF-OTU associations. Diet
618 preference is among the most influential environmental factors on gut microbiome,
619 and it can even determine microbial composition in the mammalian evolution process
620 [48]. These are reflected again in our results. IBD patients in three clusters were
621 analyzed, and the results indicated gut microbial heterogeneity of IBD patients.
622 Potentially, subgroups could be classified by individuals' lifestyle and microbial
623 community. The concept of enterotype in gut microbial community has been
624 discussed widely [49,50], and kLDM may be a useful tool to discover special
625 subgroups of IBD patients.

626 We were also interested in studying whether IBD patients in 'C3' differed from
627 other IBD patients in 'C1' and 'C2' in pathogenesis, clinical characteristics, treatment,
628 and prognosis. Notably, although patients in 'C3' consume more probiotics, they still
629 show low abundance of 'Bifidobacterium'. Probiotic diets can induce the
630 anti-inflammatory factor, IL-10, which improve gut microenvironment and reduce
631 IBD symptoms [51,52]. Individuals in 'C1' and 'C2' showed low probiotics diet
632 frequency, which agrees with studies in the literature, but IBD patients from 'C3' show
633 high probiotics diet frequency (Table S8). Philpott and Girardin [53] reported that
634 IBD patients carrying NOD2 mutations decreased transcription of IL-10. Whether
635 patients in 'C3' are more likely to carry NOD2 mutations to cause the disease is
636 suspected.

637 By further comparing the IBD diagnosis among three clusters in Table S15, all
638 IBD patients in 'C3' are those with colonic Crohn's disease, but such information of
639 IBD samples in the other two clusters is missing. Patients with Crohn's disease (CD)
640 may lack vitamin B and D [54]; therefore, the use of vitamin B and D supplements in
641 'C3' indicates the treatment for such patients. Recently, the 'Anti-Inflammatory' diet,
642 which avoids high-fat meats, has been reported to reduce symptoms in patients with
643 IBD [55]. However, high fat red meat frequency is observed for colonic Crohn's
644 disease patients in 'C3', and adjustment in this category may help them in the future.

645 Whether the lifestyle of patients with colonic Crohn's disease is affected by a doctor's
646 advice is not included in the dataset, which limits further interpretation.

647 Taking into account healthy cohorts in 'C1' and 'C2', we observed that genus
648 abundance of the IBD samples in 'C1' or 'C2' shared more similarity with that of the
649 healthy samples when compared to the IBD samples in 'C3' (Table S18). Recently
650 many microbiota-assisted models have been proposed to detect gut lesion but with
651 mixed results [50,56,57]. Adding microbial abundance improves sensitivity but with
652 the sacrifice of specificity. Our results indicate that such decreased specificity may
653 result from the close microbial composition between some patients and that of healthy
654 people. Therefore when we design microbiota-based prediction models, heterogeneity
655 within patient samples should be considered.

656 While we demonstrated that kLDM is an excellent tool for biologists to interpret
657 microbial associations under multiple environmental conditions, there are
658 implementation challenges, as well as several possible improvements. For example,
659 the mathematical score (e.g., EBIC score) may not be sensitive enough to separate
660 two environmental conditions. Studies about differential gene co-expression networks
661 may be useful for kLDM. For example, characteristics of nodes in networks, such as
662 degree, clustering coefficient, and other mathematical measures summarizing the
663 change in associations, may be included to build a more suitable approach to merging
664 sub-clusters [58]. In addition, the assumption of the Gaussian distributions for the EFs
665 may not be very suitable in the case of categorical data types, and other probabilistic
666 distributions can be considered to model categorical meta data. The scalability of
667 kLDM needs to be expanded further so that a large number of rare OTUs with
668 low-occurrence can be included in the model. Prior knowledge of microbes and their
669 interactions from known studies may also be very helpful for kLDM to reduce the
670 complexity of association inference and improve model's efficiency further [59]. The
671 sample sizes of the real TARA Oceans and CRC datasets are still relatively small,
672 which is a common phenomenon in current metagenomic studies, but we believe that
673 more large scale datasets such as the American Gut Project and Human Microbiome

674 Project, will be collected in the future. For these large-scale datasets, kLDM could be
675 the perfect tool to analyze complex associations.

676 **Conclusions**

677 We have proposed kLDM to infer EF conditions existing in the microbial community
678 and predict direct EF-OTU and conditionally dependent OTU-OTU associations
679 under every EF condition, considering compositional bias simultaneously. Compared
680 with traditional methods that estimate static association networks, kLDM has the
681 advantage of illuminating the influence of EFs on associations in the microbial
682 community. The EF condition estimated by kLDM is the result of taking EFs, OTUs'
683 abundance, and associations in the community into account comprehensively, which
684 can provide biologists with more insight into the heterogeneity of the microbial
685 community and identify microbes and interactions regulated by EFs, such as nutrients,
686 lifestyle, and health status. kLDM's superiority is validated on both synthetic data and
687 real datasets related to human gut and marine ecosystem. With the deepening of
688 research on the relationship between microorganisms and human diseases, it is
689 expected that kLDM will enable new discoveries of variation of microbes and
690 OTU-OTU and EF-OTU associations with human health and diet.

691

692 **Authors' contributions**

693 YY, XW, NC and TC designed the method. YY and XW developed the algorithm and
694 performed data analysis. YY, XW, KX, CZ, NC and TC drafted the manuscript. All
695 authors read and approved the final manuscript.

696 **Competing interests**

697 All authors declare no conflicts of interest.

698 **Acknowledgments**

699 This work is supported by the National Natural Science Foundation of China (Nos:
700 61872218, 61673241, 61721003), and Tsinghua-Fuzhou Institute Research Program,
701 Beijing National Research Center for Information Science and Technology (BNRist).

702

703 **References**

- 704 [1] Consortium IHN. The Integrative Human Microbiome Project: dynamic analysis
705 of microbiome-host omics profiles during periods of human health and disease. *Cell*
706 *Host & Microbe* 2014;16:276.
- 707 [2] Kirchman DL. Growth rates of microbes in the oceans. *Ann Rev Mar Sci*
708 2015;8:285.
- 709 [3] Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current
710 understanding of the human microbiome. *Nat Med* 2018;24:392-400.
- 711 [4] Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association
712 study of gut microbiota in type 2 diabetes. *Nature* 2013;490:55.
- 713 [5] Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, et al.
714 Genetic determinants of the gut microbiome in UK Twins. *Cell Host & Microbe*
715 2016;19:731.
- 716 [6] Scheperjans F, Aho V, Pereira PA, Koskinen K, Paulin L, Pekkonen E, et al. Gut
717 microbiota are related to Parkinson's disease and clinical phenotype. *Movement*
718 *Disorders* 2015;30:350-8.
- 719 [7] Shoaie, Saeed, Ghaffari, Pouyan, KovatchevaDatchary, Petia, et al. Quantifying
720 diet-induced metabolic changes of the human gut microbiome. *Cell Metabolism*
721 2015;22:320-31.
- 722 [8] Schwab C, Berry D, Rauch I, Rennisch I, Ramesmayer J, Hainzl E, et al.
723 Longitudinal study of murine microbiota activity and interactions with the host during
724 acute inflammation and recovery2014;8:1101-14.
- 725 [9] Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al.
726 Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*
727 2017;550:61.
- 728 [10] Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen
729 T, et al. Population-based metagenomics analysis reveals markers for gut microbiome
730 composition and diversity. *Science* 2016;352:565.

731 [11] Little AEF, Robinson CJ, Peterson SB, Raffa KF, Handelsman J. Rules of
732 engagement: interspecies interactions that regulate microbial communities. Annual
733 Review of Microbiology 2008;62:375-401.

734 [12] Fuhrman JA, Cram JA, Needham DM. Marine microbial community dynamics
735 and their ecological interpretation. Nature Reviews Microbiology 2015;13:133-46.

736 [13] Baxter NT, Th RM, Rogers MA, Schloss PD. Microbiota-based model improves
737 the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome
738 Medicine 2016;8:37.

739 [14] Mcdonald D, Hyde E, Debelius JW. American Gut: an open platform for citizen
740 science microbiome research. Msystems 2018;3:e00031-18.

741 [15] Jonathan Friedman EJA. Inferring Correlation Networks from Genomic Survey
742 Data. Plos Computational Biology 2012;8:e1002687.

743 [16] Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al.
744 Microbial co-occurrence relationships in the human microbiome. Plos Computational
745 Biology 2012;8:e1002606.

746 [17] Yang Y, Chen N, Chen T. Inference of environmental factor-microbe and
747 microbe-microbe associations from metagenomic data using a hierarchical Bayesian
748 statistical model. Cell Systems 2017;4:129.

749 [18] Fang H, Huang C, Zhao H, Deng M. CCLasso: correlation inference for
750 compositional data through Lasso. Bioinformatics 2015;31:3172.

751 [19] Ruan Q, Dutta D, Schwalbach MS, Steele JA, Fuhrman JA, Sun F. Local
752 similarity analysis reveals unique associations among marine bacterioplankton species
753 and environmental factors. Bioinformatics 2006;22:2532-8.

754 [20] Kurtz ZD, Müller CL, Miraldi ER, Dan RL, Blaser MJ, Bonneau RA. Sparse and
755 compositionally robust inference of microbial ecological networks. Plos
756 Computational Biology 2014;11:e1004226.

757 [21] Piepel G. The statistical analysis of compositional data. Technometrics
758 1982;30:120-1.

759 [22] Schloss PD, Westcott SL, Thomas R, Hall JR, Martin H, Hollister EB, et al.
760 Introducing mothur: open-source, platform-independent, community-supported
761 software for describing and comparing microbial communities. *Applied &*
762 *Environmental Microbiology* 2009.

763 [23] Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human
764 gut microbiome in liver cirrhosis. *Nature* 2014;513:59-64.

765 [24] Gipsi LM, Karoline F, Nicolas H, Johan D, Sébastien C, Fabrizio C, et al. Ocean
766 plankton. Determinants of community structure in the global plankton interactome.
767 *Science* 2015;348:192–8.

768 [25] Costea PI, Hildebrand F, Manimozhiyan A, Backhed F, Blaser MJ, Bushman FD,
769 et al. Enterotypes in the landscape of gut microbial community composition. *Nature*
770 *microbiology* 2018;3:8-16.

771 [26] Unterseher M, Jumpponen A, Ouml, Pik M, Tedersoo L, Moora M, et al. Species
772 abundance distributions and richness estimations in fungal metagenomics –
773 lessons learned from community ecology. *Molecular Ecology* 2011;20:275.

774 [27] Hong SH, Bunge J, Jeon SO, Epstein SS. Predicting microbial species richness.
775 *Proc Natl Acad Sci U S A* 2006;103:117-22.

776 [28] Zhong C, Miao D, Franti P. Minimum spanning tree based split-and-merge: A
777 hierarchical clustering method. *Information Sciences* 2011;181:3397-410.

778 [29] Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L. The huge Package for
779 high-dimensional undirected graph estimation in R. *Journal of Machine Learning*
780 *Research* 2012;13:1059-62.

781 [30] Stadlmayr A, Feng Q, Huber-Schönauer U, Niederseer D, Aigner E, Wang J, et
782 al. Gut microbiome development along the colorectal adenoma-carcinoma sequence.
783 *Zeitschrift für Gastroenterologie* 2015.

784 [31] Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, et al. Metagenomic
785 analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for
786 colorectal cancer. *Gut* 2017;66:70-8.

787 [32] Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential
788 of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems*
789 *Biology* 2015;10:766.

790 [33] Colucci F. An oral commensal associates with disease: chicken, egg, or red
791 herring? *Immunity* 2015;42:208.

792 [34] Flynn KJ, Baxter NT, Schloss PD. Metabolic and community synergy of oral
793 bacteria in colorectal cancer. *Msphere* 2016;1:e00102-16.

794 [35] Van RT, Allenvercoe E. Microbial interactions and interventions in colorectal
795 cancer. *Microbiology Spectrum* 2017;5.

796 [36] Rajkumar H, Mahmood N, Kumar M, Varikuti SR, Challa HR, Myakala SP.
797 Effect of probiotic (VSL#3) and omega-3 on lipid profile, insulin sensitivity,
798 inflammatory markers, and gut colonization in overweight adults: a randomized,
799 controlled trial. *Mediators of Inflammation* 2014;2014:348959.

800 [37] Hentges DJ, Maier BR, Burton GC, Flynn MA, Tsutakawa RK. Effect of a
801 high-beef diet on the fecal bacterial flora of humans. *Cancer Research* 1977;37:568.

802 [38] Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, et al.
803 Linking long-term dietary patterns with gut microbial enterotypes. *Science*
804 2011;334:105-8.

805 [39] Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al.
806 Correlation detection strategies in microbial data sets vary widely in sensitivity and
807 precision. *The ISME Journal* 2016;10:1669-81.

808 [40] Ian H, Keith H, Christopher Q. Dirichlet multinomial mixtures: generative
809 models for microbial metagenomics. *Plos One* 2012;7:e30126.

810 [41] Decelle J, Probert I, Bittner L, Desdevises Y, Colin S, De VC, et al. An original
811 mode of symbiosis in open ocean plankton. *Proceedings of the National Academy of*
812 *Sciences of the United States of America* 2012;109:18000-5.

813 [42] Park MG, Yih W, Coats DW. Parasites and phytoplankton, with special emphasis
814 on dinoflagellate infections. *Journal of Eukaryotic Microbiology* 2010;51:145-55.

815 [43] Lancelot C, Rousseau V, Becquevort S, Parent JY, Breton E, Leblanc C, et al.
816 Study and modeling of *Phaeocystis* in the nutrient-enriched Southern Bight of the
817 North Sea: a contribution of the Belgian AMORE (Advanced Modeling and Research
818 on Eutrophication) project. *Salud Publica De Mexico* 2003;48 suppl 2:56.

819 [44] Litaker RW, Tester PA, Colorni A, Levy MG, Noga EJ. The phylogenetic
820 relationship of *pfiesteria piscicida*, *cryptoperidiniopsis* sp. *Amyloodinium ocellatum*
821 and a *pfiesteria*-like dinoflagellate to other dinoflagellates and apicomplexans. *Journal*
822 *of Phycology* 1999;35:1379–89.

823 [45] Burkholder JAM, Jr HBG. *Pfiesteria piscicida* and other *Pfiesreria*-like
824 dinoflagellates: Behavior, impacts, and environmental controls. *Limnology &*
825 *Oceanography* 1997;42:1052-75.

826 [46] Skovgaard A, Karpov SA, Guillou L. The parasitic dinoflagellates *blastodinium*
827 spp. inhabiting the gut of marine, planktonic copepods: morphology, ecology, and
828 unrecognized species diversity. *Frontiers in Microbiology* 2012;3:305.

829 [47] Zhang B, Cai G, Wang H, Li D, Yang X, An X, et al. *Streptomyces alboflavus*
830 RPS and its novel and high algicidal activity against harmful algal bloom species
831 *Phaeocystis globosa*. *Plos One* 2014;9:e92907.

832 [48] Ley RE, Hamady M, Lozupone C, Turnbaugh P, Ramey RR, Bircher JS, et al.
833 Evolution of mammals and their gut microbes. *Science (New York, N.Y.)*
834 2008;320:1647-51.

835 [49] Costea PI, Hildebrand F, Arumugam M, Bäckhed F, Blaser MJ, Bushman FD, et
836 al. Enterotypes in the landscape of gut microbial community composition. *Nat*
837 *Microbiol* 2018;3:8-16.

838 [50] Arumugam M, Raes J, Pelletier E, Le PD, Yamada T, Mende DR, et al.
839 Enterotypes of the human gut microbiome. *Nature* 2011;473:174-80.

840 [51] Singh RK, Chang HW, Yan D, Lee KM, Ucmak D, Wong K, et al. Influence of
841 diet on the gut microbiome and implications for human health. *Journal of*
842 *Translational Medicine* 2017;15:73.

- 843 [52] Shen J, Zuo Z-X, Mao A-P. Effect of probiotics on inducing remission and
844 maintaining therapy in ulcerative colitis, crohn's disease, and pouchitis: meta-analysis
845 of randomized controlled trials. *Inflammatory Bowel Diseases* 2014;20:21-35.
- 846 [53] Philpott DJ, Girardin SE. Crohn's disease-associated Nod2 mutants reduce IL10
847 transcription. *Nature Immunology* 2009;10:455.
- 848 [54] Donnellan CF, Yann LH, Lal S. Nutritional management of Crohn's disease.
849 *Therapeutic Advances in Gastroenterology* 2013;6:231.
- 850 [55] Knightsepulveda K, Kais S, Santaolalla R, Abreu MT. Diet and Inflammatory
851 Bowel Disease. *Gastroenterology & Hepatology* 2015;11:511.
- 852 [56] Baxter NT, Ruffin MT, Rogers MAM, Schloss PD. Microbiota-based model
853 improves the sensitivity of fecal immunochemical test for detecting colonic lesions.
854 *Genome Med* 2016;8:37.
- 855 [57] Wong SH, Kwong T, Chow TC, Luk A, Dai R, Nakatsu G, et al. Original article:
856 Quantitation of faecal *Fusobacterium* improves faecal immunochemical test in
857 detecting advanced colorectal neoplasia. *Gut* 2017;66:1441-8.
- 858 [58] La Fuente AD. From 'differential expression' to 'differential networking' -
859 identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*
860 2010;26:326-33.
- 861 [59] Lo C, Marculescu R. MPLasso: Inferring microbial association networks using
862 prior microbial knowledge. *PLOS Computational Biology* 2017;13.

863

864

865 **Figure Legends**

866 **Figure 1 Schema of the kLDM model**

867 **A.** Multiple environmental conditions are assumed to exist in real environment, and
868 the EF condition can change with time. An EF-condition refers to (1) a group of
869 samples where values of EFs fall into a small and defined range, and (2) under this EF
870 condition, interactions within the microbial community are stable. **B.** Sequencing
871 samples with related meta data, possibly from multiple environmental conditions (EF

conditions), are collected. After data preprocessing, clustering, and annotating, OTUs' counts are obtained for each sample. The information about which two samples belong to the same EF-condition is unknown beforehand. **C.** The kLDM graphical model assumes K EF conditions within N samples, and infers the number of EF-conditions and associations among OTUs and associations between EFs and OTUs under every EF-condition. Two matrixes, B^k and Θ^k , record direct EF-OTU associations and OTU-OTU associations on the k^{th} EF condition, respectively. Vectors x_i and m_i are OTUs' counts and values of EFs in the i^{th} sample, respectively. At every EF-condition, we assume that values of EFs follow a multivariate Gaussian distribution that is, parameterized by μ_k and Σ_k . The rest of the variables include h_i , which represents the latent relative ratios of microbes in i^{th} sample; α_i , which is the absolute abundance of microbes; $B_0^{(k)}$, which represents the impact of unknown factors affecting OTUs' abundance; c_i , indicating that the i^{th} sample belongs to EF condition c_i ; and π_k , which is the mixture weight of the k^{th} EF condition. **D.** Compositional bias is caused by the normalization process on OTUs' counts. After normalization, microbial relative abundance sums to one. **E.** The indirect association between OTU-1 and OTU-2 induced by the common environmental factor EF-1 can be recognized by kLDM, which takes the environmental factor-microbe (EF-OTU) association into account. **F.** For kLDM, the number of EF conditions and association networks of every EF condition are estimated by a split and merge optimization algorithm. Both values of EFs and associations of microbes are taken into account to determine the EF condition. **G.** Parameters estimated by kLDM can be visualized into environmental conditions and association networks. The blue and yellow edges correspond to negative and positive associations, respectively, and the thickness of an edge is proportional to the association values.

Figure 2 Comparing ROC curves and AUC scores of kLDM with other methods on synthetic data.

‘P’, ‘Q’ and ‘N’ represent the numbers of OTUs, EFs and samples per cluster, respectively. **A.** Receiver operating characteristic (ROC) curves with 50 OTUs, 5 EFs and 2 clusters. The number of samples per cluster is set to [100, 200]. Two clusters' ROC curves on OTU-OTU and EF-OTU associations are plotted orderly. The red line corresponds to results of kLDM. **B.** ROC curves with the same parameters as **A** except that the size of clusters is set to [200, 400]. **C.** Area under the curve (AUC) scores for **A**. Scores of OTU-OTU associations and EF-OTU associations for two clusters are shown. **D.** AUC scores for **B**.

909

910 **Tables**

911 **Table 1 Top associations of CRC-associated microbes for two clusters of CRC**
912 **dataset found by kLDM**

913 **Table 2 Top 1% associations of two clusters on TARA Oceans dataset with**
914 **literature support**

915

916 **Supplementary material**

917 **Figure S1 AUC scores of kLDM with increasing numbers of microbes and EF**
918 **conditions**

919 The ‘K’ is the number of environmental conditions or clusters. **A.** AUC scores with
920 100 microbes, 8 environmental factors and 2 clusters. Size of samples of clusters
921 belongs to [400, 800]. **B.** AUC scores with 200 microbes, 10 environmental factors
922 and 2 clusters. Every cluster's number of samples range from 800 to 1600. **C.** AUC
923 scores with 50 microbes, 5 environmental factors and 3 clusters. Number of each
924 cluster's samples ranges from 200 to 400. **D.** AUC scores with 50 microbes, 5
925 environmental factors and 4 clusters.

926 **Figure S2 AUC scores of kLDM when EFs with different ratios are utilized**

927 The ‘baseline’ dataset in **Table S3** is used and we compared the results when only
928 20%, 40%, 60% and 80% EFs are found in the dataset. EFs are selected orderly and
929 set to zero because they contribute EF-OTU associations equally in the simulation

experiment. Change of AUCs of OTU-OTU and EF-OTU associations of two clusters are shown.

Figure S3 CRC-associated microbial relative abundance on two clusters

Figure S4 OTU-OTU association networks for cancer (Cluster 1) and healthy (Cluster 2) clusters

Edges represent associations with absolute weight among top 1%. Red and green edges represent positive and negative associations, respectively. The width of an edge is proportional to the association's absolute weight. Only microbes with associations in either Cluster 1 or Cluster 2 are shown. Positions of OTUs in the A and B are exactly identical and sizes of nodes are correlated with their average abundance in all samples. 5 CRC-associated microbes (Peptostreptococcus, Parvimonas, Fusobacterium, Porphyromonas and Prevotella), are arranged closely and colored with blue-green, and their associated microbes are labeled with red.

Figure S5 The relationship between the effectiveness of kLDM and the similarity of EFs of EF conditions

The 'baseline' dataset in **Table S3** is used again and we changed the distance between two clusters' mean values of EFs from 0.1 to 10.0. For every distance, AUC scores of estimated OTU-OTU associations ('Cluster 1 OTU-OTU' and 'Cluster 2 OTU-OTU') and EF-OTU associations ('Cluster 1 EF-OTU' and 'Cluster 2 EF-OTU') of two clusters are calculated.

950

Table S1 Comparison of running time and memory usage by kLDM and mLDM on single association network inference

Table S2 Mean values of sensitivities and specificities of top N associations in five different methods

Table S3 AUC scores of kLDM considering the influence of similarity of environmental conditions, EF-OTU associations and OTU-OTU associations respectively

Table S4 AUC scores of kLDM when EFs with different ratios are utilized and

959 **other methods with full dataset**

960 **Table S5 Mean values of meta data of two clusters on colorectal cancer data**

961 **Table S6 Composition of the diagnostic state in two clusters on colorectal**
 962 **cancer data**

963

964 **Table S7 Mean values of EFs of two clusters in TARA Oceans dataset**

965 **Table S8 Matched genus-level interactions on TARA Oceans dataset of kLDM**
 966 **and the 'Static'**

967 **Table S9 Types of matched genus-level interactions in results of kLDM and the**
 968 **'Static'**

969 **Table S10 20 EFs of three clusters predicted by kLDM on American Gut**
 970 **project dataset**

971 **Table S11 Health status of people in three clusters on American Gut project**
 972 **dataset**

973 **Table S12 OTU abundance, EF-OTU and OTU-OTU associations of three**
 974 **clusters on the American Gut project dataset**

975 (1) 216 OTUs' relative abundance of three clusters. (2) Top 1% OTU-OTU
 976 associations of 'C1'. (3) Top 1% OTU-OTU associations of 'C2'. (4) Top 1%
 977 OTU-OTU associations of 'C3'. (5) Top 1% EF-OTU associations of 'C1'. (6) Top 1%
 978 EF-OTU associations of 'C2' (7) Top 1% EF-OTU associations of 'C3'.

979 **Table S13 Genus-level relative abundance of three clusters on the American**
 980 **Gut project dataset**

981 44 genres' relative abundance is calculated and P-value of the abundance difference
 982 of each genus between any two clusters is computed by the Wilcoxon rank-sum test.

983 **Table S14 EF values and genus-level relative abundance of IBD samples in**
 984 **three clusters**

985 (1) EF values of IBD samples in 'C1', 'C2' and 'C3' and the P-value of each EF
 986 between any two clusters is calculated by the t-test. (2) Genus-level relative
 987 abundance of IBD samples and the abundance difference is done by the Wilcoxon

rank-sum test.

Table S15 IBD diagnosis of IBD patients in three clusters on American Gut project dataset

Table S16. Values of inverse Simpson index of synthetic datasets corresponding to Figure 2 and Figure S1.

Table S17 Composition of the diagnostic state in two clusters estimated by MicrobeDMM on colorectal cancer data

Table S18 EF values and genus-level abundance of healthy and IBD samples in 'C1' and 'C2'

(1) EF values of healthy and IBD samples in 'C1' and 'C2' and the P-value is computed by t-test. (2) Genus-level relative abundance of healthy and IBD samples in 'C1' and 'C2'.

1000

Table 1 Top associations of CRC-associated microbes for two clusters of CRC dataset found by kLDM

Cluster 1 (Cancer)		
OTU	EF or OTU	Association
<i>Parvimonas</i>	<i>Porphyromonas</i>	+0.238
<i>Prevotella</i>	<i>Phascolarctobacterium</i>	-0.234
<i>Peptostreptococcus</i>	<i>Porphyromonas</i>	+0.220
<i>Peptostreptococcus</i>	<i>Parvimonas</i>	+0.191
<i>Parvimonas</i>	<i>Fusobacterium</i>	+0.180
<i>Porphyromonas</i>	<i>Prevotella</i>	+0.173
<i>Peptostreptococcus</i>	<i>Fusobacterium</i>	+0.155
<i>Fusobacterium</i>	<i>Porphyromonas</i>	+0.155
<i>Porphyromonas</i>	<i>Akkermansia</i>	+0.121
<i>Prevotella</i>	<i>Clostridium_XIVa</i>	-0.095
<i>Prevotella</i>	<i>Clostridium_sensu_stricto</i>	+0.091
<i>Prevotella</i>	<i>Clostridium_IV</i>	-0.090

<i>Parvimonas</i>	<i>Prevotella</i>	+0.085
<i>Prevotella</i>	<i>Clostridium_XIVa</i>	-0.082
<i>Porphyromonas</i>	<u>Diagnostic State</u> (EF)	+0.294
<i>Fusobacterium</i>	<u>FIT</u> (EF)	+0.245
<i>Peptostreptococcus</i>	<u>Diagnostic State</u> (EF)	+0.281
Cluster 2 (Healthy)		
OTU	EF or OTU	Association
<i>Prevotella</i>	<u>Diagnostic State</u> (EF)	+0.274

1003 *Note: 'CRC' is the abbreviation for colorectal cancer. Results of CRC-related*
1004 *OTU-OTU and EF-OTU associations, selected from top 1% weighted associations*
1005 *from the two clusters. Five microbes related to CRC, as reported in previous study,*
1006 *are labeled in bold and the environmental factors are underlined. The 'Diagnostic*
1007 *State' consists of 'normal', 'high-risk normal', 'advanced adenoma', 'adenoma' and*
1008 *'cancer' states. 'FIT' stands for fecal immunochemical test results.*

1009

1010 **Table 2 Top 1% associations of two clusters on TARA Oceans dataset with**
1011 **literature support**

Cluster 1			
OTU	EF or OTU	Association	Article
<i>Amphibelone</i>	<i>Phaeocystis</i>	+0.172	Decelle et al. (2012) [41]
<i>Phaeocystis</i>	<i>Amphibelone anomala</i>	+0.161	Decelle et al. (2012) [41]
<i>Amoebophrya</i>	<i>Cochlodinium_01</i>	+0.120	Park et al. (2010) [42]
<i>ceratii</i>	<i>fulvescens</i>		
<i>Amoebophrya</i>	<i>Cochlodinium_01</i>	+0.118	Park et al. (2010) [42]
<i>ceratii</i>	<i>fulvescens</i>		
<i>Blastodinium</i>	<u>Salinity</u> (EF)	+0.536	Skovgaard et al. (2012) [46]
<i>mangini</i>			
<i>Phaeocystis</i>	<u>PO4</u> (EF)	+0.508	Lancelot et al. (2003) [43]
<i>Amphibelone</i>	<u>PO4</u> (EF)	+0.439	Burkholder and Jr (1997) [40]

anomala

Cluster 2

OTU	EF or OTU	Association	Article
<i>Blastodinium_06</i>	<u>Temperature</u> (EF)	+0.549	Skovgaard et al. (2012) [46]
<i>Blastodinium_06</i>	<u>Oxygen</u> (EF)	+0.410	Skovgaard et al. (2012) [46]
<i>Phaeocystis</i>	<u>Oxygen</u> (EF)	+0.389	Zhang et al. (2014) [47]

1012 *Note:* Top 1% weighted OTU-OTU and EF-OTU associations are chosen from two
 1013 clusters, and associations with related literature support are listed in the table. OTUs
 1014 are labeled in bold and EFs are underlined.

1015

1016

1017

A Assumption of the environmental condition

B OTU and meta data

OTU data

Sample	Sample-1	Sample-2	Sample-3	...	Sample-4	Sample-5	Sample-6	...
OTU								
OTU-1	x_{11}	x_{21}	x_{31}		x_{41}	x_{51}	x_{61}	
OTU-2	x_{12}	x_{22}	x_{32}		x_{42}	x_{52}	x_{62}	
OTU-3	x_{13}	x_{23}	x_{33}		x_{43}	x_{53}	x_{63}	
...								

Meta data

Sample	Sample-1	Sample-2	Sample-3	...	Sample-4	Sample-5	Sample-6	...
Meta								
EF-1	m_{11}	m_{21}	m_{31}		m_{41}	m_{51}	m_{61}	
EF-2	m_{12}	m_{22}	m_{32}		m_{42}	m_{52}	m_{62}	
...								

C kLDM graphical model

D Compositional bias

E Indirect association

F Split and merge optimization

G Inferred EF conditions and association networks

A ROC curves with $P=50, Q=5, K=2, N \in [100, 200]$

B ROC curves with $P=50, Q=5, K=2, N \in [200, 400]$

C AUC scores with $N \in [100, 200]$

D AUC scores with $N \in [200, 400]$