Extended Graphical Lasso for Multiple Interaction Networks for High Dimensional Compositional Data

Yang Xu[1,2], Hongmei Jiang[2], Wenxin Jiang[2],

1 Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan, Shandong, China
2 Department of Statistics, Northwestern University, Evanston, Illinois, United States of America

Abstract

Compositional data are quantitative descriptions of the parts of some whole, conveying relative information, which are ubiquitous in many fields. There has been a spate of interest in association networks for such data in biological and medical research, for example, microbial interaction networks. In this paper, we propose a novel method, the extended joint hub graphical lasso (EDOHA), to estimate multiple related interaction networks for high dimensional compositional data across multiple distinct classes. To be specific, we construct a convex penalized log likelihood optimization problem using log ratios of compositional data and solve it with an alternating direction method of multipliers (ADMM) algorithm. The performance of the proposed method in the simulated studies shows that EDOHA has remarkable advantages in recognizing class-specific hubs than the existing comparable methods. We also present three applications of real datasets. Biological interpretations of our results confirm those of previous studies and offer a more comprehensive understanding of the underlying mechanism in disease.

Author summary

Reconstruction of multiple association networks from high dimensional compositional data is an important topic, especially in biology. Previous studies focused on estimating different networks and detecting common hubs among all classes. Integration of information over different classes of data while allowing difference in the hub nodes is also biologically plausible. Therefore, we propose a method, EDOHA, to jointly construct multiple interaction networks with capacity in finding different hub networks for each class of data. Simulation studies show the better performance over conventional methods. The method has been demonstrated in three real world data.

Introduction

Compositional data consist of vectors of proportions or percentages normalized to a constant sum, which makes inference on correlations between unconstrained features complicated due to the information loss from normalization [1]. Compositional data are commonplace in molecular biology, especially in the 'omics' (e.g. metagenomics, proteomics and metabolomics) studies. One of the motivating examples in the real applications is the microbiome data. Numerous studies have shown strong evidence that

microbial compositions are closely related with various diseases such as diabetes [2], inflammatory bowel disease [3] and obesity [4]. Microbial count data are usually generated by sequencing variable regions of bacterial 16S rRNA gene. They are not directly comparable across samples and are usually transformed to relative abundance or proportion by dividing the total counts in the sample. Another motivating example is protein compositional data where relative quantification of proteins is achieved by comparison of mass spectrometry (MS) or tandem MS (MS/MS) signals of the differentially labeled moieties [5]. A common characteristic of these compositional omics data is the deficiency of independent samples (n) in comparison with the abundance of features (p), that is to say, $p \gg n$.

With advances in high-throughput sequencing and omics technologies, biological information is being collected at an amazing rate, which stimulates researchers to discover modular structure, relationships and regularities in complex data. A wide range of methods have been proposed to construct biologial correlation networks for composition data, such as SPIEC-EASI [6], SparCC [7], Reboot [8], REBACCA [9], CCLasso [10] and COAT [11] for microbial interaction networks. Interactions between various biological nodes (e.g. genes, proteins, metabolites) on different levels (e.g. gene regulation, cell signalling) can be represented as graphs and, thus, analysis of such networks might shed new light on the function of biological systems. Hubs, the highly connected nodes at the tail of the power law degree distribution, are known to play a crucial role in biological networks. Some studies have shown that scale-free topology exists in many different organizational levels, such as metabolic networks [12] and cellular networks [13]. Hub nodes may be the most essential elements for community stability and play an important role in the infection and pathogenesis of the virus.

The objective of our research is to estimate multiple interaction networks for high dimensional compositional data across multiple classes. Aforementioned methods are used to depict the relationship networks between features for one class only. When there are multiple classes, such as healthy and diseased conditions, a straightforward method is to construct the network for each class separately and then compare their differences. However, these procedures may sacrifice the similarity shared between multiple classes, which may be critically important to find out the principal elements related to the disease. One would expect these networks to be similar to each other, since they are from the same type of entities. The joint graphical lasso (JGL) [14] is proposed to estimate multiple models simultaneously, which ignores the scale-free network and is unable to detect hubs explicitly. In the model of JRmGRN [15], it identifies common hub elements across multiple classes by jointly using distinct datasets. In many situations, hub nodes that are specific to an individual network also exist. For example, in the tissue-specific networks associated with SARS-CoV-2, both common and class-specific key hubs are revealed in diverse tissues [16]. Common hub features are essential to all class and class-specific hubs could convey particular biology information. This inspires us to explore a new model to incorporate both common and class-specific hub nodes when jointly constructing interaction networks.

Materials and methods

Transformation of compositional data

We will first briefly discuss compositional data for one class using microbiome data as an example. The absolute abundances or counts of p microbes, $\mathbf{y} = [y_1, y_2, \ldots, y_p]$, living in an environment such as human gut are usually not directly observable. However, the relative abundances $\mathbf{x} = [\frac{y_1}{m}, \frac{y_2}{m}, \ldots, \frac{y_p}{m}]$ where $m = \sum_{i=1}^{p} y_i$, can be measured using 16S rRNA sequencing technologies. It is apparently that $\mathbf{x}$ is

compositional data due to the sum being one and its elements are conditionally dependent with each other, since they are constrained to lie in the unit simplex. Here we apply the centered log-ratio transform [1] to remove the unit-sum constraint of compositional data. For a compositional variable $\mathbf{x} = (x_1, \ldots, x_p)$, we have

$$\mathbf{r} = clr(\mathbf{x}) = [log(\frac{x_1}{g(\mathbf{x})}), \ldots, log(\frac{x_p}{g(\mathbf{x})})] = [log(\frac{y_1}{g(\mathbf{y})}), \ldots, log(\frac{y_p}{g(\mathbf{y})})],$$

where $g(\mathbf{x}) = [\prod_{i=1}^{p} x_i]^{\frac{1}{p}}$ is the geometric mean of the composition vector. It is easy to show that there is a relationship between the covariance matrix $\mathbf{\Gamma}$ of $\mathbf{r}$ and the population covariance of the log-transformed absolute abundances $\mathbf{\Sigma} = Cov[log\mathbf{Y}]$: $\mathbf{\Gamma} = \mathbf{G\Sigma G}$ [1,6], where $\mathbf{G} = \mathbf{I}_p - \frac{1}{p}\mathbf{J}$, $\mathbf{I}_p$ is the p-dimensional identity matrix, and $\mathbf{J}$ is $p * p$ matrix with each of the entries equals 1. Kurtz et al. [6] mention that the matrix $\mathbf{G}$ is close to the identity matrix for high-dimensional data, and thus a finite sample estimator $\hat{\mathbf{\Gamma}}$ of $\mathbf{\Gamma}$ may be as a good approximation of $\hat{\mathbf{\Sigma}}$. Actually, Cao et al. [11] have shown that $\mathbf{\Gamma}$ could be a proxy for $\mathbf{\Sigma}$ as long as $\mathbf{\Sigma}$ belongs to a class of large sparse covariance matrices.

Gaussian graphical models (GGMs) are now frequently used to describe biological feature association networks and to detect conditionally dependent features. Correlation networks could be expressed as an undirected, weighted graph $\mathscr{G} = (\mathbf{V}, \mathbf{E})$ where the vertex set $\mathbf{V} = \{v_1, v_2, \ldots, v_p\}$ represents the p feature nodes (e.g., microbes or proteins) and the edge set $\mathbf{E}$ contains the possible associations among nodes. Suppose the log-transformed observations $(r_1, \ldots, r_p)$ are drawn from a multivariate normal distribution with covariance $\mathbf{\Gamma}$, the non-zero elements of the off-diagonal entries of the inverse covariance matrix $\mathbf{\Theta} = \mathbf{\Gamma}^{-1}$ define the adjacency matrix of the graph $\mathscr{G}$ and thus describe the factorization of the normal distribution into conditionally dependent components [17]. Because the number of samples n is smaller than the number of features p and $\mathbf{\Theta}$ are expected to be sparse, penalized maximum likelihood approaches are proposed to estimate the precision matrix $\mathbf{\Gamma}^{-1}$, which yields an estimation of precision matrix $\hat{\mathbf{\Theta}} = \hat{\mathbf{\Gamma}}^{-1}$ where $\hat{\mathbf{\Gamma}}$ is the sample covariance matrix.

The general formulation for EDOHA

We present the extended joint hub graphical lasso (EDOHA) algorithm for constructing multiple interaction networks from multiple classes. Suppose that there are K classes of data sets, corresponding to K different levels of a phenotype variable or K different conditions, such as control group, carrier group and disease group. Let $\mathbf{X}^{(k)} \in \mathbb{R}^{n_k * p}$ be a matrix representing the compositional data of p features, and $\mathbf{R}^{(k)} \in \mathbb{R}^{n_k * p}$ be the matrix for the centered log-ratio transformed data. Assume that the transformed observations are independent, identically distributed: $\mathbf{r}_1^{(k)}, \ldots, \mathbf{r}_{n_k}^{(k)} \sim N(\mu^{(k)}, \mathbf{\Gamma}^{(k)})$, where $\mathbf{r}^{(k)}$ represents relative biological data from the kth class. The log likelihood for the data takes the form

$$l(\{\mathbf{\Theta}\}) = \frac{1}{2}\sum_{k=1}^{K} n_k\big(\log(\det\mathbf{\Theta}^{(k)}) - \mathrm{tr}(\hat{\mathbf{\Gamma}}^{(k)}\mathbf{\Theta}^{(k)})\big). \tag{1}$$

where $\hat{\mathbf{\Gamma}}^{(k)}$ is the empirical covariance estimation of $\mathbf{r}^{(k)}$ and $\hat{\mathbf{\Gamma}}^{(k)}$ is an approximation of $\hat{\mathbf{\Sigma}}^{(k)}$. The non-zero element $\theta_{ij}^{(k)}$ in $\mathbf{\Theta}^{(k)} = {\mathbf{\Gamma}^{(k)}}^{-1}$ indicates node i and j for the kth class are conditionally dependent. Most elements in $\mathbf{\Theta}^{(k)}$ are expected to be zero. In addition, a few hub nodes are expected to have a large number of interactions with many other nodes, some of which are common among all classes and others are specific to different classes. Unlike JRmGRN [15], our method aims to investigate these class-specific hub nodes explicitly. The precision matrix $\mathbf{\Theta}^{(k)}$ is decomposed into two

parts: the elementary symmetric network for the kth class $\mathbf{Z}^{(k)}$, mainly containing the non-hub node correlation information, and the network for hub nodes $\mathbf{V}^{(k)}$ for the kth class, including common and class-specific ones, where $\mathbf{V}^{(k)}$ is a matrix with entirely zero or almost completely nonzero columns. To estimate $\{\mathbf{\Theta}\} = (\mathbf{\Theta}^{(1)}, \mathbf{\Theta}^{(2)}, \ldots, \mathbf{\Theta}^{(K)})$ when $p > n_k$, we take a penalized log likelihood approach

$$\min_{\{\mathbf{\Theta}\}} - \sum_{k=1}^{K} n_k(\log(\det \mathbf{\Theta}^{(k)}) - \mathrm{tr}(\hat{\mathbf{\Gamma}}^{(k)}\mathbf{\Theta}^{(k)})) + P(\{\mathbf{\Theta}\}). \tag{2}$$

The penalty function $P(\{\mathbf{\Theta}\})$ has the following form,

$$\begin{aligned} P(\{\mathbf{\Theta}\}) = \lambda_1 &\sum_{k=1}^{K} \|\mathbf{Z}^{(k)} - \mathrm{diag}(\mathbf{Z}^{(k)})\|_1 \\ &+ \lambda_2 \sum_{k<k'} \|\mathbf{Z}^{(k)} - \mathbf{Z}^{(k')} - \mathrm{diag}(\mathbf{Z}^{(k)} - \mathbf{Z}^{(k')})\|_1 \\ &+ \lambda_3 \sum_{k} \|\mathbf{V}^{(k)} - \mathrm{diag}(\mathbf{V}^{(k)})\|_1 + \lambda_4 \sum_{k} \|\mathbf{V}^{(k)} - \mathrm{diag}(\mathbf{V}^{(k)})\|_{1,2} \\ &+ \lambda_5 \sum_{k<k'} \|\mathbf{V}^{(k)} - \mathbf{V}^{(k')} - \mathrm{diag}(\mathbf{V}^{(k)} - \mathbf{V}^{(k')})\|_1 \end{aligned}$$

where $\mathbf{Z}^{(k)} + \mathbf{V}^{(k)} + (\mathbf{V}^{(k)})^T = \mathbf{\Theta}^{(k)}$, and $\|\mathbf{V}^{(k)}\|_{1,2} = \sum_{j=1}^{p} \|\mathbf{V}_j^{(k)}\|_2$, $\mathbf{V}_j^{(k)}$ is the jth column of matrix $\mathbf{V}^{(k)}$. Here $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ are five nonnegative tuning parameters. λ_1 and λ_3 control the sparsity of elementary network $\mathbf{Z}^{(k)}$ and hub network $\mathbf{V}^{(k)}$ respectively. λ_4 allows $\mathbf{V}^{(k)}$ to have zero columns and dense non-zero columns, where the non-zero columns represent the respective hub nodes in kth class. And λ_2, λ_5 encourage the elementary networks and hub networks to have the similarity. When $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 are fixed, the expression of (2) is a convex optimization problem, which can be solved by efficient algorithms. The convexity of (2) is based on the following facts: both negative log determinant and norm functions are convex functions, so is the nonnegative combination of convex functions.

Apparently, compared with the JRmGRN model, we replace the common hub network with respective hub network for each class and thus we are able to find out the common and class-specific hub nodes simultaneously. Common hub features across multiple classes could be crucial to regulate biological interaction, while class-specific hubs may mediate specific phenotype. Our proposed method may help to explain which features play a significant part in different phenotypic traits or in different conditions.

An ADMM algorithm for EDOHA

We solve the problem using an alternating directions method of multipliers algorithm [18], which allows us to decouple some of the terms that are difficult to optimize jointly. We assume that $\mathbf{\Theta}^{(k)}$ is positive definite for $k = 1, \ldots, K$. We note that the problem can be reformulated as a consensus problem [19]:

$$\min \Phi(\mathbf{X}) + h_1(\tilde{\tilde{\mathbf{V}}}) + \Psi(\tilde{\mathbf{X}}) \quad s.t. \quad \mathbf{X} = \tilde{\mathbf{X}} \quad \mathbf{V} = \tilde{\tilde{\mathbf{V}}}, \tag{3}$$

where $\mathbf{X} = \big(\mathbf{\Theta}^{(1)}, \mathbf{Z}^{(1)}, \mathbf{V}^{(1)}, \ldots, \mathbf{\Theta}^{(K)}, \mathbf{Z}^{(K)}, \mathbf{V}^{(K)}\big)$,
$\tilde{\mathbf{X}} = \big(\tilde{\mathbf{\Theta}}^{(1)}, \tilde{\mathbf{Z}}^{(1)}, \tilde{\mathbf{V}}^{(1)}, \ldots, \tilde{\mathbf{\Theta}}^{(K)}, \tilde{\mathbf{Z}}^{(K)}, \tilde{\mathbf{V}}^{(K)}\big)$, and

$$\Phi(\mathbf{X}) = f(\mathbf{\Theta}) + g(\mathbf{Z}) + h(\mathbf{V}), \tag{4}$$

$$\Psi(\tilde{\mathbf{X}}) = \sum_{k=1}^{K} I(\tilde{\mathbf{\Theta}}^{(k)} = \tilde{\mathbf{Z}}^{(k)} + \tilde{\mathbf{V}}^{(k)} + (\tilde{\mathbf{V}}^{(k)})^T), \tag{5}$$

where

$$f(\mathbf{\Theta}) = -\sum_{k=1}^{K} n_k(\log(\det\mathbf{\Theta}^{(k)}) - \mathrm{tr}(\hat{\mathbf{\Gamma}}^{(k)}\mathbf{\Theta}^{(k)})), \tag{6}$$

$$\begin{aligned} g(\mathbf{Z}) = \lambda_1 \sum_{k=1}^{K} \|\mathbf{Z}^{(k)} - \mathrm{diag}(\mathbf{Z}^{(k)})\|_1 \\ + \lambda_2 \sum_{k<k'} \|\mathbf{Z}^{(k)} - \mathbf{Z}^{(k')} - \mathrm{diag}(\mathbf{Z}^{(k)} - \mathbf{Z}^{(k')})\|_1, \end{aligned} \tag{7}$$

$$\begin{aligned} h(\mathbf{V}) = \lambda_3 \sum_{k} \|\mathbf{V}^{(k)} - \mathrm{diag}(\mathbf{V}^{(k)})\|_1 \\ + \lambda_4 \sum_{k} \|\mathbf{V}^{(k)} - \mathrm{diag}(\mathbf{V}^{(k)})\|_{1,2}, \end{aligned} \tag{8}$$

$$h_1(\tilde{\tilde{\mathbf{V}}}) = \lambda_5 \sum_{k<k'} \|\tilde{\tilde{\mathbf{V}}}^{(k)} - \tilde{\tilde{\mathbf{V}}}^{(k')} - \mathrm{diag}(\tilde{\tilde{\mathbf{V}}}^{(k)} - \tilde{\tilde{\mathbf{V}}}^{(k')})\|_1. \tag{9}$$

And

$$I(\tilde{\mathbf{\Theta}}^{(k)} = \tilde{\mathbf{Z}}^{(k)} + \tilde{\mathbf{V}}^{(k)} + (\tilde{\mathbf{V}}^{(k)})^T) = \begin{cases} 0 & \text{if} \quad \tilde{\mathbf{\Theta}}^{(k)} = \tilde{\mathbf{Z}}^{(k)} + \tilde{\mathbf{V}}^{(k)} + (\tilde{\mathbf{V}}^{(k)})^T \\ \infty & \text{otherwise} \end{cases}$$

The scaled augmented Lagrangian is given by

$$\begin{aligned} L(\mathbf{X}, \tilde{\mathbf{X}}, \tilde{\tilde{\mathbf{V}}}, \mathbf{W}, \tilde{\mathbf{W}}) = \Phi(\mathbf{X}) + h_1(\tilde{\tilde{\mathbf{V}}}) + \Psi(\tilde{\mathbf{X}}) + \frac{\rho}{2}\|\mathbf{X} - \tilde{\mathbf{X}} + \mathbf{W}\|_F^2 - \frac{\rho}{2}\|\mathbf{W}\|_F^2 \\ + \frac{\rho}{2}\|\tilde{\tilde{\mathbf{V}}} - \mathbf{V} + \tilde{\mathbf{W}}_V\|_F^2 - \frac{\rho}{2}\|\tilde{\mathbf{W}}_V\|_F^2, \end{aligned} \tag{10}$$

where $\mathbf{X}, \tilde{\mathbf{X}}, \tilde{\tilde{\mathbf{V}}}$ are the primal variables, $\mathbf{W} = (\{\mathbf{W}_{\Theta}^{(k)}\}, \{\mathbf{W}_{Z}^{(k)}\}, \{\mathbf{W}_{V}^{(k)}\})$, $\tilde{\mathbf{W}}_V$ are the dual variables. $\|\mathbf{A}\|_F^2$ denotes the Frobenius norm of $\mathbf{A}$. Here ρ is a positive parameter for the scaled Lagrangian form. We set $\rho = 2.5$ as is used in Deng et al. [15].

The iteration of ADMM can be described as follows:

$$\begin{cases} \mathbf{X}_{t+1} = \mathrm{argmin}_{\mathbf{X}} \Big\{ \Phi(\mathbf{X}) + \frac{\rho}{2}\|\mathbf{X} - \tilde{\mathbf{X}}_t + \mathbf{W}_t\|_F^2 \\ \qquad\qquad + \frac{\rho}{2}\|\tilde{\tilde{\mathbf{V}}}_t - \mathbf{V} + \tilde{\mathbf{W}}_{V_t}\|_F^2 \Big\} \\ \tilde{\tilde{\mathbf{V}}}_{t+1} = \mathrm{argmin}_{\tilde{\tilde{\mathbf{V}}}} \Big\{ h_1(\tilde{\tilde{\mathbf{V}}}) + \frac{\rho}{2}\|\tilde{\tilde{\mathbf{V}}} - \mathbf{V}_{t+1} + \tilde{\mathbf{W}}_{V_t}\|_F^2 \Big\} \\ \tilde{\mathbf{X}}_{t+1} = \mathrm{argmin}_{\tilde{\mathbf{X}}} \Big\{ \Psi(\tilde{\mathbf{X}}) + \frac{\rho}{2}\|\mathbf{X}_{t+1} - \tilde{\mathbf{X}} + \mathbf{W}_t\|_F^2 \Big\} \\ \mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{X}_{t+1} - \tilde{\mathbf{X}}_{t+1} \\ \tilde{\mathbf{W}}_{V_{t+1}} = \tilde{\mathbf{W}}_{V_t} + \tilde{\tilde{\mathbf{V}}}_{t+1} - \mathbf{V}_{t+1} \end{cases} \tag{11}$$

Theorem 1. *There exists a solution* $(\mathbf{X}^*, \tilde{\mathbf{X}}^*, \tilde{\tilde{\mathbf{V}}}^*)$ *to the EDOHA optimization problem (3), and the ADMM iteractions via (11) approach the optimal value, i.e.* $p_t \to p^*$*, where* $p_t = \Phi(\boldsymbol{X}_t) + h_1(\tilde{\tilde{\mathbf{V}}}_t) + \Psi(\tilde{\mathbf{X}}_t)$ *and* $p^* = \Phi(\boldsymbol{X}^*) + h_1(\tilde{\tilde{\mathbf{V}}}^*) + \Psi(\tilde{\mathbf{X}}^*)$.

The theorem establishes the convergence of the ADMM algorithm to achieve the optimal solution for EDOHA. It also automatically establishes algorithmic convergence for any optimization problem that can be regarded as a sub-case of EDOHA, for example, JRmGRN (by taking $\lambda_5 = 0$), which was not established before. A general algorithm for solving the optimization problem is shown in S1 Text. And the proof of Theorem 1 is shown in S2 Text.

Faster computations for EDOHA

We now present a theorem that lead to substantial computational improvements to the EDOHA. Using the theorem, one can inspect the empirical covariance matrices $\hat{\mathbf{\Gamma}}^{(1)}, \ldots, \hat{\mathbf{\Gamma}}^{(K)}$ in order to determine whether the solution to the EDOHA optimization problem is block diagonal after some permutation of the features. Previous studies [14, 15] use uniform thresholding to decompose the precision matrices of different classes in exactly the same way. Non-uniform thresholding generates a non-uniform feasible partition by thresholding the K empirical covariance matrices separately. In a non-uniform partition, two variables of the same group in one class may belong to different groups in another class [20]. Here we recommend a novel non-uniform thresholding approach that can split precision matrices into smaller submatrices without ignoring the different sparsity patterns from different matrices. Now we provide the key result. The following theorem states the sufficient conditions for the presence of non-uniform block diagonal structure.

Theorem 2. *A sufficient condition for the solution to (2) to be block diagonal with blocks given by $C_1^k, C_2^k, \ldots, C_{T_k}^k$ is that*

$$min\left\{\frac{\lambda_1-(K-1)\lambda_2}{n_1}, \ldots, \frac{\lambda_1-(K-1)\lambda_2}{n_K}, \frac{\lambda_3-(K-1)\lambda_5}{2n_1}, \ldots, \frac{\lambda_3-(K-1)\lambda_5}{2n_1}\right\} \geq |\hat{\Gamma}_{ij}^{(k)}|$$

for $\forall k$, $i \in C_t^k, j \in C_{t'}^k, t \neq t'$.

Proof of Theorem 2 is given in S3 Text. It supplies us with a criterion if, given a partition of features $C_1^k, C_2^k, \ldots, C_{T_k}^k$, $\sum_t C_t^k = p$, the solution of the optimization problem is block diagonal with each block corresponding to features in C_t^k. In practice, for any given $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$, we can quickly perform the following two-step procedure to identify any block structure in each class in the solution.

- Create $\mathbf{B}^{(k)}$, a $p * p$ matrix with $B_{ii}^{(k)} = 1$ for $i = 1, \ldots, p$. For $i \neq j$, let $B_{ij}^{(k)} = 0$ if the conditions specified in Theorem 2 are met for that pair of variables. Otherwise, set $B_{ij}^{(k)} = 1$.
- Identify the connected components of the undirected graph whose adjacency matrix is given by $\mathbf{B}^{(k)}$.

Theorem 2 guarantees that the connected components identified correspond to distinct blocks in kth class. Therefore, one can quickly obtain these solutions based on a non-uniform feasible partition. The block diagonal condition leads to massive computational speed-ups. Instead of computing the eigen decomposition of K $p * p$ matrices, we compute the eigen decomposition of $\sum_k T_k$ matrices of dimensions $p_{C_1^k} * p_{C_1^k}, ..., p_{C_{T_k}^k} * p_{C_{T_k}^k}$. The computational complexity per-iteration is reduced from $O(p^3)$ to $\sum_k \sum_{t=1}^{T_k} O(p_{C_t^k}^3)$.

Tuning parameter selection

In this paper, we use Bayesian information criterion(BIC)-type quantity to select tuning parameters. We choose $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ to minimize the following function which

balances the model likelihood and model complexity.

$$
\begin{aligned}
BIC(\hat{\mathbf{\Theta}}, \hat{\mathbf{Z}}, \hat{\mathbf{V}}) = & \sum_{k=1}^{K} \Big[n_k \big(-\log(\det \hat{\mathbf{\Theta}}^{(k)}) + \mathrm{tr}(\hat{\mathbf{\Gamma}}^{(k)} \hat{\mathbf{\Theta}}^{(k)}) \big) \Big] \\
& + \sum_{k=1}^{K} log(n_k) |\hat{\mathbf{Z}}^{(k)}| - log(n) |\bigcap \hat{\mathbf{Z}}^{(k)}| \\
& + \sum_{k=1}^{K} log(n_k) \big(\hat{v}^{(k)} + c(|\hat{\mathbf{V}}^{(k)}| - \hat{v}^{(k)}) \big) \\
& - log(n) \big(\hat{v} + c(|\bigcap \hat{\mathbf{V}}^{(k)} - \hat{v}|) \big),
\end{aligned} \tag{12}
$$

where $\{\hat{\mathbf{\Theta}}^{(k)}, \hat{\mathbf{Z}}^{(k)}, \hat{\mathbf{V}}^{(k)}\}$ is the estimated parameters with a fixed set of tuning parameters $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$, $|*|$ is the cardinality, $\hat{v}^{(k)}$ is the number of estimated hubs for kth class and $\hat{v}$ is the number of estimated common hubs, and c is a constant between zero and one. We select the set of tuning parameters $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ which minimizes the quality BIC$(\hat{\mathbf{\Theta}}^{(k)}, \hat{\mathbf{Z}}^{(k)}, \hat{\mathbf{V}}^{(k)})$. Note that BIC will favor more hub nodes in $\hat{\mathbf{V}}^{(k)}$ when constant c is small. In this paper, we take $c = 0.3$.

We use the grid search to find the tuning parameters. However, computing BIC over a range of values for five tuning parameters $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ may be computationally intensive. In this case, we suggest a dense search over $(\lambda_1, \lambda_3, \lambda_4)$ while holding (λ_2, λ_5) at fixed low values, followed by a quick search over (λ_2, λ_5), holding $(\lambda_1, \lambda_3, \lambda_4)$ at the selected values. With the number of features involved in the analysis dramatically increasing, tuning parameter selection becomes very complicated. In this situation, we need to explore some theoretical properties of the problem that can be used to provide guidance on our search of tuning parameters. This approach follows Deng et al. [15] and we provide the following theorems that extend their theoretical results to our present case with class-specific hubs.

Theorem 3. *Let* $\left(\mathbf{\Theta}^{*(k)}, \boldsymbol{Z}^{*(k)}, \boldsymbol{V}^{*(k)}\right)$ *be a solution to (2), a sufficient condition for* $\boldsymbol{Z}^{*(k)}$ *to be a diagonal matrix is that* $\lambda_3 + \lambda_4 < 2\lambda_1$ *and* $\lambda_5 < 2\lambda_2$.

Proof of Theorem 3 is given in S4 Text.

Theorem 4. *Let* $\left(\mathbf{\Theta}^{*(k)}, \boldsymbol{Z}^{*(k)}, \boldsymbol{V}^{*(k)}\right)$ *be a solution to (2), a sufficient condition for* $\boldsymbol{V}^{*(k)}$ *to be a diagonal matrix is that* $2\lambda_1 < \lambda_3 + \frac{\lambda_4}{\sqrt{p}}$ *and* $2\lambda_2 < \lambda_5$.

Proof of Theorem 4 is given in S5 Text.

Corollary 1. *Let* $\left(\mathbf{\Theta}^{*(k)}, \boldsymbol{Z}^{*(k)}, \boldsymbol{V}^{*(k)}\right)$ *be a solution to (2), a necessary condition for both* $\boldsymbol{Z}^{*(k)}$ *and* $\boldsymbol{V}^{*(k)}$ *to be non-diagonal matrices is that*

$$
\lambda_3 + \frac{\lambda_4}{\sqrt{p}} + \lambda_5 \leq 2(\lambda_1 + \lambda_2) \leq \lambda_3 + \lambda_4 + \lambda_5. \tag{13}
$$

Specifically, we require that both $\mathbf{Z}^{(k)}$ and $\mathbf{V}^{(k)}$ are non-diagonal to produce non-trivial edges and hubs. With Corollary 1, we could reduce the search space of parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 as these five tuning parameters are related. In this paper, we use a uniformed grid of log space from 0.01 to 5 (size=20) for parameter $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 satisfying $\lambda_3 + \frac{\lambda_4}{\sqrt{p}} + \lambda_5 \leq 2(\lambda_1 + \lambda_2) \leq \lambda_3 + \lambda_4 + \lambda_5$.

Results

Simulation studies

To examine the efficiency of the proposed method for the better identification of common and class-specific hub nodes, we simulate Erdös-Rényi (ER)-based network [21] and then generate corresponding compositional data to assess and validate the method. We compare the performance of EDOHA with the existing methods, such as JGL and JRmGRN. Results show that EDOHA is more efficient than other methods when analyzing compositional data correlation networks which have both common and class-specific hub nodes.

Simulation strategy

To simulate a biological compositional data set such as microbiome count data, we consider the data are drawn with two steps. We first generate the basis abundance and proportion for each feature and then generate count data given a sequencing size (i.e. library size). The data structure characteristics are reflected in the basis covariance, which will be introduced in details later. Here we assume that basis proportions vary from sample to sample and are generated from one of three different distributions, namely, log ratio normal (LRN), Poisson log normal (LNP) and Dirichlet log normal (LND) distributions [9]. These three methods are presented in S6 Text. Then we extract count data from a multinomial distribution using the proportions, which reflects a random process that all sequences are equally likely to be selected in a biological sample.

To evaluate EDOHA comprehensively, we consider that the features are associated with ER-based network, in which each pair of nodes is selected with equal probability and connected with a predefined probability. The scale-free ER-based networks are generated by modifying the procedure used in Deng et al. [15]. Specifically, for a given number of classes (K), nodes (p), samples (n_k), we use the following procedures to simulate ER-based network and corresponding compositional data.

Step 1 We generate the base sparse matrix $\mathbf{A}$ in which A_{ij} is set as a random number in $[-0.75, -0.25] \bigcup [0.25, 0.75]$ with probability α (elementary network sparsity 1-α) and zero otherwise.

Step 2 Given the number of hubs m, we randomly choose m nodes and for each element that represents the correlation between ith hub node and other node j, $\tilde{h}_{ij}$, we set it to be a random number in $[-0.75, -0.25] \bigcup [0.25, 0.75]$ with probability β (hub sparsity 1-β) and zero otherwise.

Step 3 To construct the hub matrix $\mathbf{H}^{(k)}$, we randomly choose a fraction δ (network difference) of the hub nodes and reset them to be random numbers from $1, 2, \ldots, p$. The modified hub nodes are denoted by $h^{(k)}$. As for nonzero elements in $\mathbf{H}^{(k)}$, we first set $h_{ij}^{(k)} = \tilde{h}_{ij}$ and then randomly adjust a fraction of δ of these nonzero elements and reset their values to be random numbers in $[-0.75, -0.25] \bigcup [0.25, 0.75]$ with probability β.

Step 4 To construct the elementary network, $\mathbf{Z}^{(k)}$, we first set it equal to $\mathbf{A}$, and then randomly choose a fraction of δ of elements and reset their values to be random numbers in $[-0.75, -0.25] \bigcup [0.25, 0.75]$ with probability α and zero otherwise. We set $\mathbf{Z}^{(k)} = \mathbf{Z}^{(k)} + t(\mathbf{Z}^{(k)})$ so that $\mathbf{Z}^{(k)}$ is symmetric.

Step 5 We define the precision matrix $\mathbf{\Theta}^{(k)}$ as $\mathbf{Z}^{(k)} + \mathbf{H}^{(k)} + (\mathbf{H}^{(k)})^T$. If $\mathbf{\Theta}^{(k)}$ is not positive definite, we add the diagonal element of $\mathbf{\Theta}^{(k)}$ by $0.1 - \lambda_{\min}(\mathbf{\Theta}^{(k)})$, where $\lambda_{\min}(\mathbf{\Theta}^{(k)})$ is the minimum eigenvalue of $\mathbf{\Theta}^{(k)}$.

Step 6 We generate the compositional data of n_k samples for the kth class from a multinomial distribution using the proportion obtained from LRN with basis covariance $(\mathbf{\Theta}^{(k)})^{-1}$.

Here the simulation studies are conducted for three classes with 40 or 80 samples for each class. The elementary network sparsity, the hub sparsity and the network difference are set as 0.98, 0.7, 0.2, respectively. We simulate three networks with 80, 160, 300 nodes, respectively. As we have mentioned, we use the BIC and the grid search to find the appropriate tuning parameters and model.

Simulation results

We consider simulated network described in the previous section with 80, 160, 300 nodes and estimate corresponding system with sample size n=40, n=80, respectively. The effects of EDOHA penalties vary with the sample size. To better present the simulation study results, we multiply the tuning parameters $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ by the sample size before performing the EDOHA.

We compare the performance of EDOHA and JRmGRN of identifying non-zero edges and class-specific edges. The results are computed averaging over 100 simulated data sets. We say that an edge (i, j) in the kth network is detected if the estimated association $\hat{\Theta}_{ij}^{(k)} \neq 0$ and we say that the edge is correctly detected if $\Theta_{ij}^{(k)} \neq 0$. The number of differential edges, which differ between classes, is defined as follows [14]:

$$\sum_{k<k'} \sum_{i<j} I\Big(\Theta_{ij}^{(k)} \neq \Theta_{ij}^{(k')}\Big).$$

We record the sensitivity and specificity associated with detecting non-zero edges and detecting differential edges. The sensitivity is the proportion of the non-zero or differential edges that are correctly detected and the specificity represents the proportion of the zero or non-differential edges that are correctly detected. Hence the sensitivity and specificity of edge detection (ED) and differential edge detection (DED) are computed as

- ED Sensitivity= $\dfrac{\sum_{k=1}^{K} \sum_{i<j} I\left(\hat{\Theta}_{ij}^{(k)} \neq 0 \text{ and } \Theta_{ij}^{(k)} \neq 0\right)}{\sum_{k=1}^{K} \sum_{i<j} I\left(\Theta_{ij}^{(k)} \neq 0\right)}$
- ED Specificity= $\dfrac{\sum_{k=1}^{K} \sum_{i<j} I\left(\hat{\Theta}_{ij}^{(k)} = 0 \text{ and } \Theta_{ij}^{(k)} = 0\right)}{\sum_{k=1}^{K} \sum_{i<j} I\left(\Theta_{ij}^{(k)} = 0\right)}$
- DED Sensitivity= $\dfrac{\sum_{k<k'} \sum_{i<j} I\left(\hat{\Theta}_{ij}^{(k)} \neq \hat{\Theta}_{ij}^{(k')} \text{ and } \Theta_{ij}^{(k)} \neq \Theta_{ij}^{(k')}\right)}{\sum_{k<k'} \sum_{i<j} I\left(\Theta_{ij}^{(k)} \neq \Theta_{ij}^{(k')}\right)}$
- DED Specificity= $\dfrac{\sum_{k<k'} \sum_{i<j} I\left(\hat{\Theta}_{ij}^{(k)} = \hat{\Theta}_{ij}^{(k')} \text{ and } \Theta_{ij}^{(k)} = \Theta_{ij}^{(k')}\right)}{\sum_{k<k'} \sum_{i<j} I\left(\Theta_{ij}^{(k)} = \Theta_{ij}^{(k')}\right)}$

As shown in Table 1, if only the number of non-zero edges is considered, there is little difference between EDOHA and JRmGRN in terms of the total number of detected pairwise node-node associations. However, the sensitivity of detecting differential edges using EDOHA has more than doubled in all cases compared with JRmGRN. This is mainly because EDOHA is equiped with better ability to identify the class-specific edges. Comprehensive evaluation of EDOHA on identifying the common and

class-specific hubs are presented in Table 2. These results suggest the usefulness of EDOHA in identifying true hub nodes through network reconstruction.

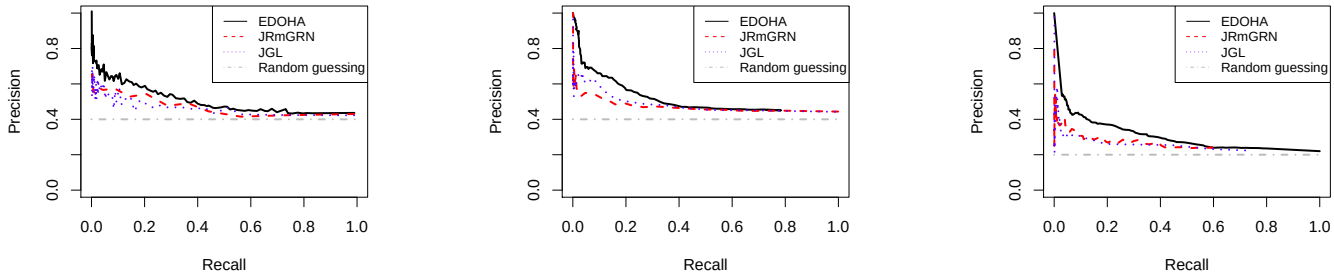
Table 1. Means (Standard deviations) over 100 replicates using EDOHA and JRmGRN are shown for sensitivity and specificity of edge detection (ED) and differential edge detection (DED)

		n=40				n=80			
		ED Sensitivity	ED Specificity	DED Sensitivity	DED Specificity	ED Sensitivity	ED Specificity	DED Sensitivity	DED Specificity
p=80	EDOHA	0.614(0.089)	0.925(0.054)	0.396(0.098)	0.959(0.049)	0.597(0.110)	0.960(0.024)	0.416(0.104)	0.968(0.019)
	JRmGRN	0.535(0.055)	0.914(0.039)	0.138(0.112)	0.986(0.011)	0.542(0.132)	0.928(0.058)	0.161(0.095)	0.989(0.004)
p=160	EDOHA	0.352(0.055)	0.974(0.011)	0.288(0.059)	0.977(0.005)	0.430(0.063)	0.977(0.010)	0.329(0.058)	0.979(0.006)
	JRmGRN	0.355(0.039)	0.955(0.012)	0.077(0.043)	0.992(0.004)	0.417(0.053)	0.989(0.003)	0.114(0.068)	0.987(0.002)
p=300	EDOHA	0.347(0.032)	0.971(0.007)	0.217(0.081)	0.974(0.016)	0.288(0.036)	0.990(0.005)	0.293(0.050)	0.978(0.011)
	JRmGRN	0.297(0.052)	0.939(0.025)	0.068(0.067)	0.988(0.018)	0.253(0.028)	0.991(0.003)	0.110(0.037)	0.991(0.002)

Table 2. The ability of EDOHA and JRmGRN to identify hub nodes are shown by True Positive (TP), False Positive (FP), sensitivity and specificity. The number of hubs is the average of different classes and sample size n=80.

	80 nodes: 5 hubs				160 nodes: 8 hubs				300 nodes: 12 hubs			
	TP	FP	Sensitivity	Specificity	TP	FP	Sensitivity	Specificity	TP	FP	Sensitivity	Specificity
EDOHA	5	2	1	0.973	5	2	0.625	0.987	7	1	0.583	0.997
JRmGRN	4	3	0.8	0.96	5.5	3.5	0.688	0.977	4	3	0.333	0.990

We also show that EDOHA has substantial improvements over several other methods. Since the hub nodes cannot be found out by JGL explictly, the precision recall curve is constructed based on the differential non-zero edges in the network, which is compared with the results from aforementioned methods intuitively. We simulate the networks with varying sparsity and similarity in two conditions and estimate corresponding networks with 160 nodes. The sample size is 80. To compare the results from different methods, we simulate each situation 100 times. As can be seen in Figure 1, the precision of EDOHA stays high through a larger range of recall, whereas for the other methods it quickly drops to the level of random guessing. This agrees with our expectation since EDOHA distinguishes the differences among elementary networks and hub networks respectively, which fits the data in the model better.



(a) E_sparsity:0.8,H_sparsity:0.4,Difference=0.4 (b) E_sparsity:0.8,H_sparsity:0.6,Difference=0.4 (c) E_sparsity:0.8,H_sparsity:0.6,Difference=0.2

Fig 1. The Precision-Recall curve of EDOHA, JRmGRN and JGL for different networks setting. 'E_sparsity' is the sparsity of elementary network; 'H_sparsity' is the sparsity of Hub network, the last parameter shown in title is the difference of two elementary networks.

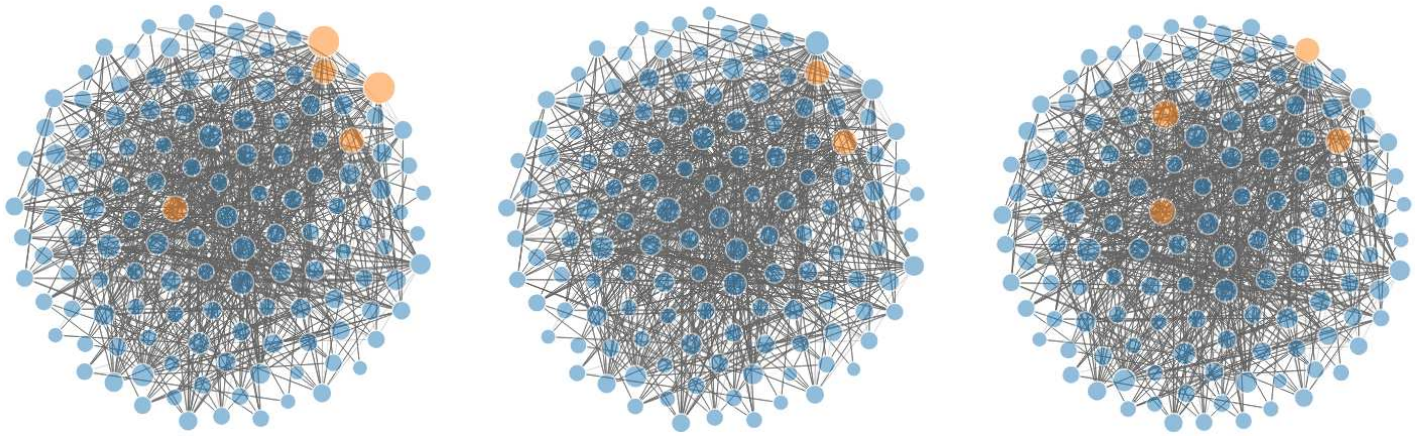
Real data analysis

We apply the proposed model on three real data sets: one is proteomic data and the other two are microbiome data. Compared with the analysis methods used in the original publications, our model demonstrates better performance in constructing multiple networks with common and class-specific hubs across multiple classes.

Application to mouse skin microbiome data

We apply EDOHA to a mouse skin microbial data set including three groups of individuals: non-immunized (Control), immunized-healthy (Healthy), and immunized-diseased (EBA). Microbial communities are measured utilizing variable regions of bacterial 16S rRNA sequencing data. These regions are amplified, sequenced, and then grouped into common Operational Taxonomic Units (OTUs) according to the similarity and quantified, with OTU counts serving as an intermediary to the underlying microbial populations abundances. The data set contains 131 core OTUs mainly coming from four prime phyla, which are Firmicutes (44 OTUs), Proteobacteria (35 OTUs), Bacteroidetes (26 OTUs), Actinobacteria (17 OTUs). We analyze their abundance data from 261 mouse skin samples. In particular, we wish to reconstruct the pair-wise conditional correlations network and identify the OTUs that are hubs. Such OTUs likely play an important role in the environment.

In Figure 2, we plot the networks for the three groups. The hub OTUs are highlighted in orange. Only OTUs from Firmicutes and Actinobacteria are identified as hub OTUs. The three networks share only one common hub while Healthy and EBA groups have another common hub. However, three hub OTUs in the Healthy group do not appear as hubs in the EBA group. Note that two OTU hubs shared by the Healthy and the Control groups are not hubs in the EBA group. Such information may be useful to understand the mechanism of protection from disease, which would not be available without our method of class-specific hub detection.



(a) Healthy (b) EBA (c) Control

Fig 2. The estimated microbial networks for the groups of non-immunized (Control), immunized-healthy (Healthy) and immunized-diseased (EBA) individuals. The hub OTUs are highlighted in orange.

In addition to comparing the hub OTUs, we also investigate whether the correlation patterns among the 131 OTUs are different for different groups of disease status. A correlated pair of OTUs is considered consistent between two groups if the correlations in both groups have the same sign. We come to the same conclusion that correlations from the non-immunized individuals are less consistent with other two immunized

groups than between the two immunized groups. There are 687 consistent pairs between the two immunized groups, while there are only 639 consistent pairs between the Control and Healthy and 632 between the Control and EBA groups. The results obtained in Ban et al. [9] were 532 consistent pairs between the two immunized groups, the other two were 236 and 212 respectively. Hence the gaps between these groups in our research are much less (See S1 Fig). The reason is mainly that we jointly model multiple networks simultaneously so that the similarity of network could be constructed more accurately by using datasets from multiple classes, which results in more accurate class-specific networks.

Application to IBD microbiome data

We preform our proposed method on the inflammatory bowel disease (IBD) multi-omics database from the Integrative Human Microbiome Project (HMP2 metadata) focusing on the functions of microbes in human health and disease. IBD further includes two main subtypes, Crohn's disease (CD) and ulcerative colitis (UC). Our samples consist of 86 CD patients, 46 UC patients and 46 healthy controls with 342 OTUs. As is known, IBD is a chronic and relapsing inflammatory condition of the gastrointestinal (GI) tract and the GI microbiome of healthy humans is dominated by four major bacterial phyla: Firmicutes, Bacteroidetes, Proteobacteria and Actinobacteria. The data set contains 225, 44, 38, 23 OTUs from these four prime phyla respectively.

We aim to reconstruct the multiple microbial networks of the human gut that represent the interactions among the OTUs, as well as to identify hub OTUs that tend to have many interactions with other ones. Identifying such regulatory OTUs will lead to a better understanding of the mechanism of IBD, and eventually may lead to new therapeutic treatments. A large-scale cross-measurement type association network for host and microbial molecular interactions has been constructed [22]. Figure 3 displays the microbial interaction networks for the three classes. More hub OTUs are identified in CD than in UC and healthy controls. And almost each hub in UC and healthy groups is covered in the hub sets of CD group. We find that species from Actinobacteria are not detected as hub OTUs in three groups. Several studies [3, 23, 24] discovered that Faecalibacterium were differentially abundant in IBD and healthy group. Subdoligranulum, Roseburia and Fusobacterium have also been identified as hubs, all of which are associated, metatranscriptionally as well as metagenomically, with taxonomic features. In our study, Rumiococcus gnavus and Roseburia are found in CD and UC groups but not in healthy group, OTUs from Alistiles are only detected as hubs in CD group, which may lead to an entirely new line of medical research into IBD.

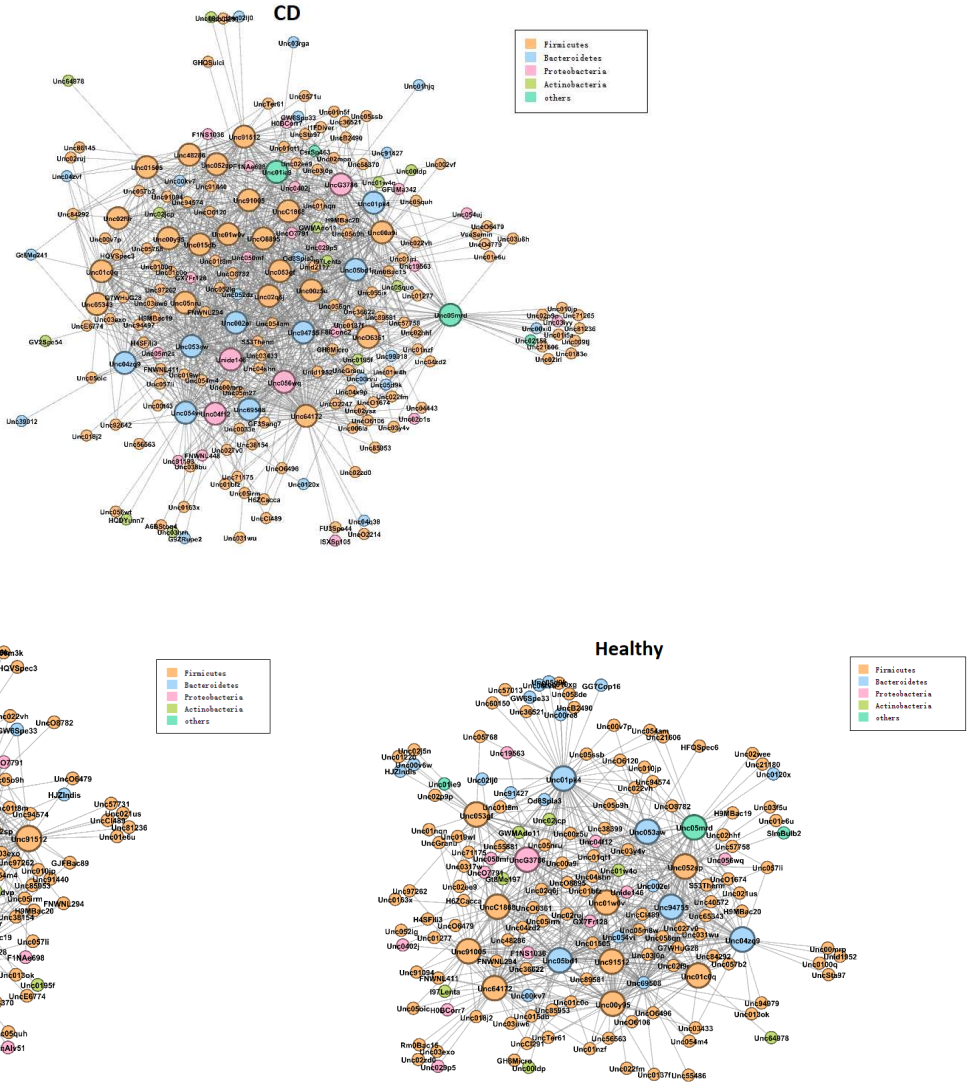
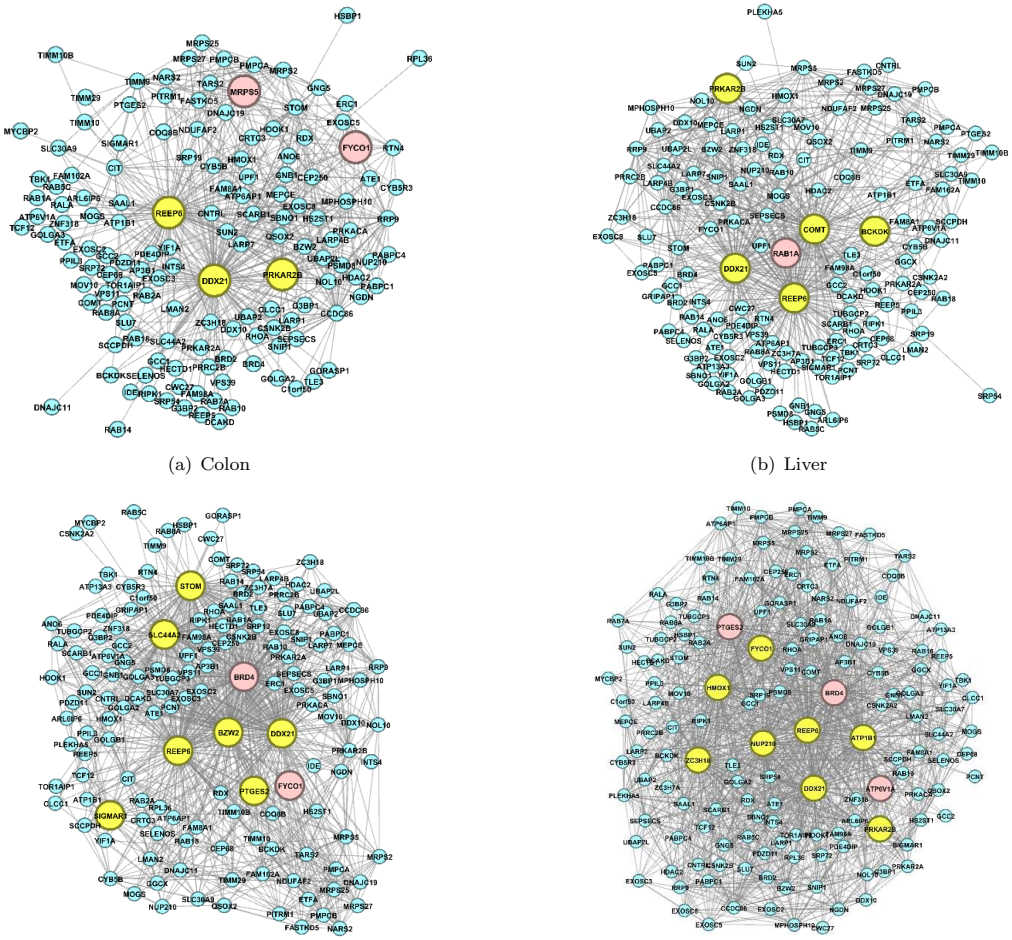


Fig 3. The estimated microbial networks for CD, UC and healthy groups. Four major bacterial phyla are marked by different colors. The larger nodes represent the hub OTUs.

Application to SARS-CoV-2 infection proteomic data

A most recent study depicted the SARS-CoV-2 protein interaction networks for 332 proteins [25] in four kinds of tissues: colon, liver, lung and kidney. The data downloaded from the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) database provide relative protein abundance. We apply the proposed method, EDOHA, to construct proteome-wide networks ans reveal common key hubs across different types of tissues and tissue-specific hubs.

As shown in Figure 4, we identify three common hub proteins BRD4, FYCO1 and PTGES2. And we identify MRPS5 as a hub only in colon, RAB1A only in liver and ATP6V1A only in kidney, which is consistent with previous studies [16]. We also detect many other common hubs, including DDX21, PRKAR2B and REEP6, which appear as hubs in two or three organs. Moreover, BCKDK and COMT are involved in hubs only in liver. BWZ2, SIGMAR1, SLC44A2 and STOM are recognized as hubs merely in lung. And ATP1B1, HMOX1, NUP210 and ZC3H18 are screened as hubs only in kidney. All



(a) Colon (b) Liver

(c) Lung (d) Kidney

Fig 4. The estimated interaction networks of proteins affected by SARS-Cov-2 in four tissues. The hub proteins are highlighted in yellow and pink. And the pink ones are involved in top 5 hub proteins detected by previous study.

of these hub proteins detected in four tissues and their functions in living organism are shown in S1 Table.

All of these hubs have a connectivity bigger than 100, which is at least 4 times larger than that of any non-hubs. Ubiquitous hubs in multiple tissues would be promising drug targets to rescue multi-organ injury and deal with inflammation. Certain tissue-specific hubs might mediate specific dysfunction. Such information is urgently needed for the identification of the therapeutic targets for intervention and vaccine development.

Discussion

Currently, there has been an increasing interest in the structure of multiple interaction networks. In most cases, people implicitly assume that each node has roughly the same number of interactions within the network when analyzing omics data, and each pair of nodes has equal probability to be an edge and all edges are independent of each other. However, this assumption is not appropriate in some real-world networks. In biological networks, scale-free properties are quite universal, which means the number of edges for each node follows a power-law distribution and a small proportion of nodes interact

with many other ones. Barabasi and Oltvai [26] has found that most networks within the cell approximate a scale-free topology, including the metabolic networks, protein-protein interactions and genetic regulatory networks. The presence of hubs seems to be a general feature of all cellular networks. For example, hub proteins play critical roles in the organization and function of cellular protein interaction networks. It has also been demonstrated that such hub proteins may constitute an important pool of attractive drug targets. One typical aim is to capture more complex interactions and identify class-specific hubs in class-specific networks. Constructing biological association networks based on data sets from the same tissue with different phenotypes or different tissue enables us to screen out the influential features contributing to life health and disease, which provides insights into understanding the essential elements in living organisms and ecosystems. As researches into biological correlation networks continue, it has become important to develop a novel model to jointly estimate the scale-free interactions networks from different classes.

In this paper we propose a new statistical procedure to construct class-specific networks and select informative hub features among multiple classes for high dimensional compositional data. Hub features, including common and specific ones, are accurately identified by decomposing the precision matrix into two parts. New penalty terms are added to single out class-specific hubs. Moreover, theoretical properties for selecting tuning parameters are investigated to improve computation efficiency. The synthetic data are generated with ER-based network to model as closely as possible the situation in experimental biological compositional data. Our simulation studies show that the proposed method achieves higher accuracy in detecting the differential edges from different classes. We show that EDOHA has the potential to recognize the class-specific hub features and gains the larger area under the Precision-Recall curves compared with other methods. We also apply the proposed method on three real omics data sets. One of them is proteomic data from different tissues, and the other two are microbial data from microbial communities with different phenotypes. Across all three data sets, EDOHA successfully builds multiple networks and the results are basically consistent with previous reports. Furthermore, EDOHA identifies some hub features, both common and class-specific ones, which provides a deeper understanding of the mechanisms involved. Overall, EDOHA could not only jointly reconstruct multiple networks but also detect class-specific hubs explicitly for compositional data with multiple distinct classes. It is promising in generating networks with such data structure.

Supporting information

S1 Text. A detailed ADMM algorithm for EDOHA.

S2 Text. The proof of the convergence of the ADMM algorithm for EDOHA.

S3 Text. The proof of the sufficient conditions for the non-uniform block diagonal structure.

S4 Text. The proof of Theorem 3.

S5 Text. The proof of Theorem 4.

S6 Text. Methods for generating the basis proportion for each feature.

S1 Fig. Venn diagram of consistent correlated OTUs from Control, Healthy and EBA groups. The figure shows the number of possible pairs within the same group and between different groups. It is suggested that the gaps between these groups in our research are much less than Ban et al. [9].

S1 Table. The hub proteins detected in four organs and their functions in living organism. The table lists hub proteins detected as common ones as well as tissue-specific ones, and introduces their functions.

References

1. Aitchison J. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological). 1982; 44(2):139–160.
2. Dunne JL, Triplett EW, Gevers D, Xavier R, Insel R, Danska J, et al. The intestinal microbiome in type 1 diabetes. Clinical & Experimental Immunology. 2014; 177(1):30–37.
3. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome biology. 2012; 13(9):R79.
4. Perry RJ, Peng L, Barry NA, Cline GW, Zhang D, Cardone, RL, et al. Acetate mediates a microbiome–brain–β-cell axis to promote metabolic syndrome. Nature. 2016; 534(7606):213–217.
5. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nature biotechnology. 1999; 17(10):994–999.
6. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLoS computational biology. 2015; 11(5):e1004226.
7. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS computational biology. 2012; 8(9):e1002687.
8. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS computational biology. 2012; 8(7):e1002606.
9. Ban Y, An L, Jiang H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. Bioinformatics. 2015; 31(20):3322–3329.
10. Fang H, Huang C, Zhao H, Deng M. CCLasso: correlation inference for compositional data through Lasso. Bioinformatics. 2015; 31(19):3172–3180.
11. Cao Y, Lin W, Li H. Large covariance estimation for compositional data via composition-adjusted thresholding. Journal of the American Statistical Association. 2019; 114(526):759–772.
12. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical organization of modularity in metabolic networks. science. 2002; 297(5586):1551–1555.

13. Ravasz E. Detecting hierarchical modularity in biological networks. Computational Systems Biology. 2009; 145–160.

14. Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2014; 76(2):373–397.

15. Deng W, Zhang K, Liu S, Zhao P, Xu S, Wei H. JRmGRN: joint reconstruction of multiple gene regulatory networks with common hub genes using data from multiple tissues or conditions. Bioinformatics. 2018; 34(20):3470–3478.

16. Feng L, Yin Y, Liu C, Xu K, Li Q, Wu J, et al. Proteome-wide Data Analysis Reveals Tissue-specific Network Associated with SARS-CoV-2 Infection. Journal of Molecular Cell Biology. 2020.

17. Lauritzen SL. Graphical models. Clarendon Press; 1996.

18. Boyd S, Parikh N, Chu E. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning. 2011; 3(1):1–122.

19. Ma S, Xue L, Zou H. Alternating direction methods for latent variable Gaussian graphical model selection. Neural computation. 2013; 25(8):2172–2198.

20. Tang Q, Yang C, Peng J, Xu J. Exact hybrid covariance thresholding for joint graphical lasso. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2015;593–607.

21. Mendes P, Sha W, Ye K. Artificial gene networks for objective comparison of analysis algorithms. Bioinformatics. 2003; 19(suppl_2):ii122–ii129.

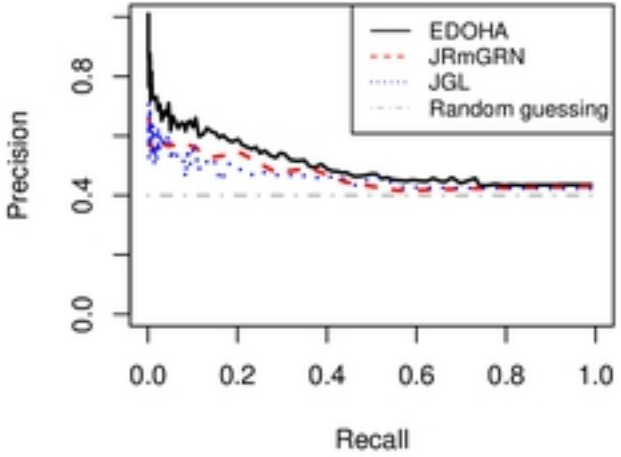
22. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019; 569(7758):655–662.

23. Kang S, Denman SE, Morrison M, Yu Z, Dore J, Leclerc M, et al. Dysbiosis of fecal microbiota in Crohn's disease patients as revealed by a custom phylogenetic microarray. Inflammatory bowel diseases. 2010; 16(12):2034–2042.

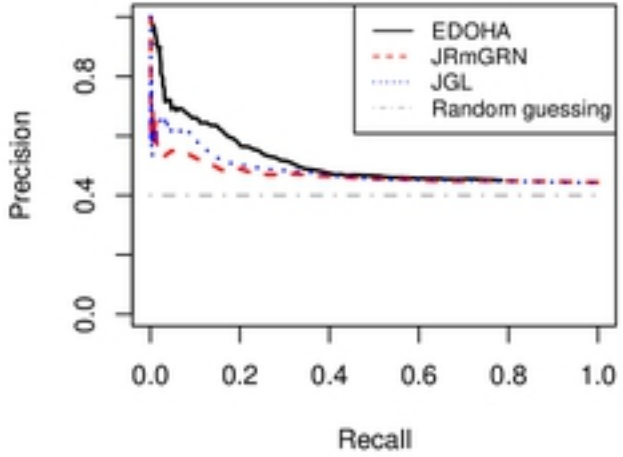
24. Mondot S, Barreau F,Al Nabhani Z, Dussaillant M, Le RK, Doré J, et al. Altered gut microbiota composition in immune-impaired Nod2-/- mice. Gut. 2012; 61(4):634–635.

25. Gordon, DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020;1–13.

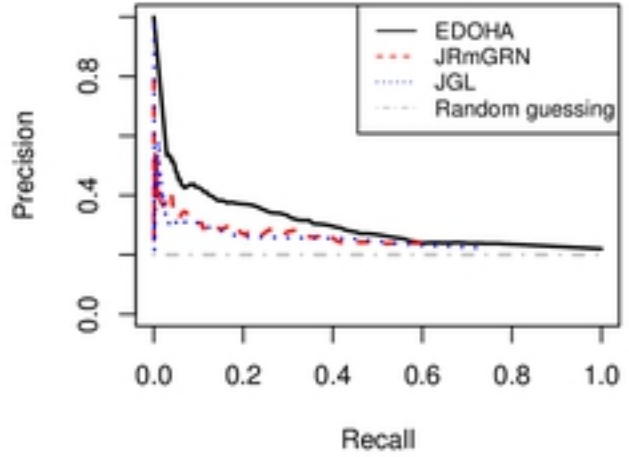
26. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nature reviews genetics. 2004; 5(2):101–113.



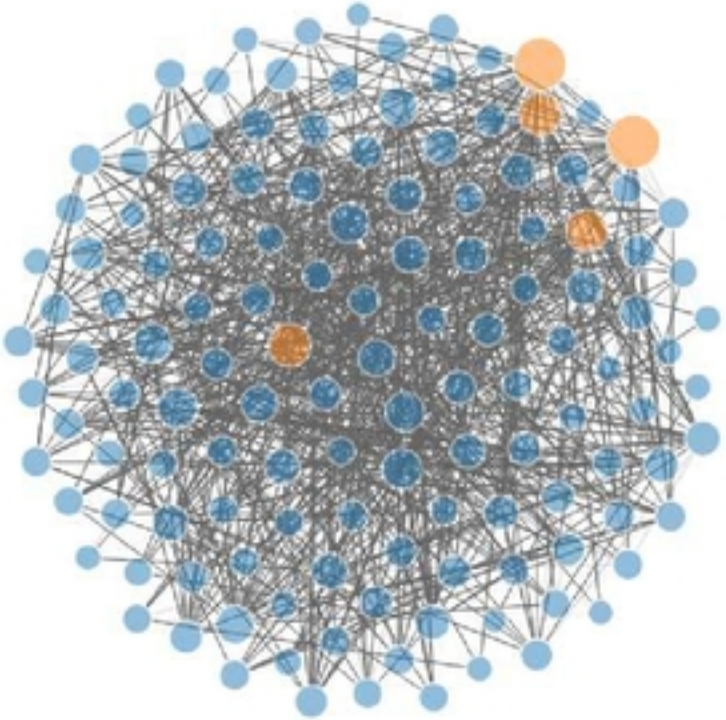
(a) E_sparsity:0.8,H_sparsity:0.4,Difference=0.4



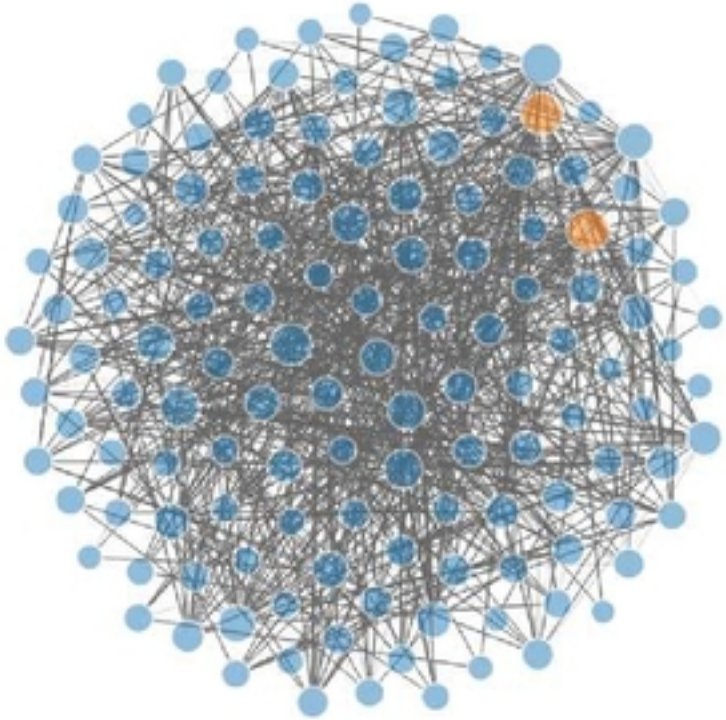
(b) E_sparsity:0.8,H_sparsity:0.6,Difference=0.4



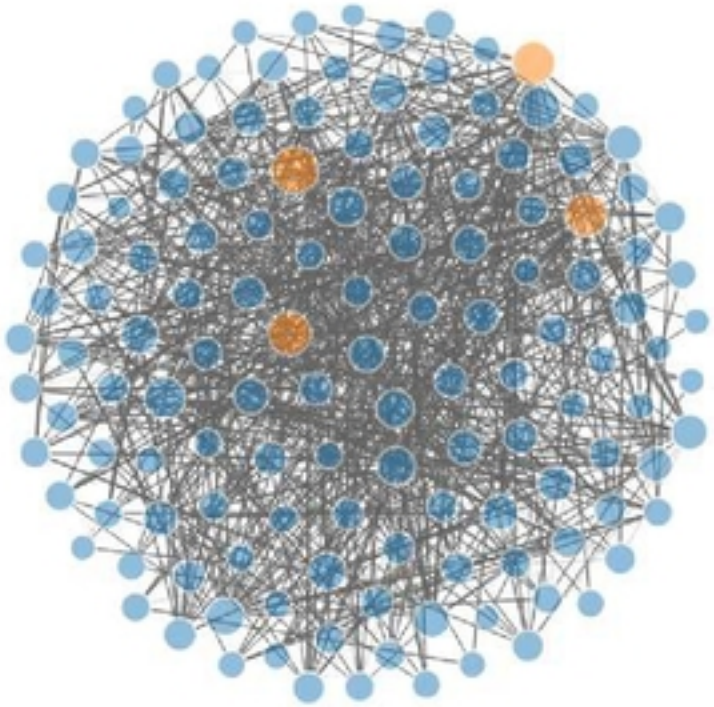
(c) E_sparsity:0.8,H_sparsity:0.6,Difference=0.2



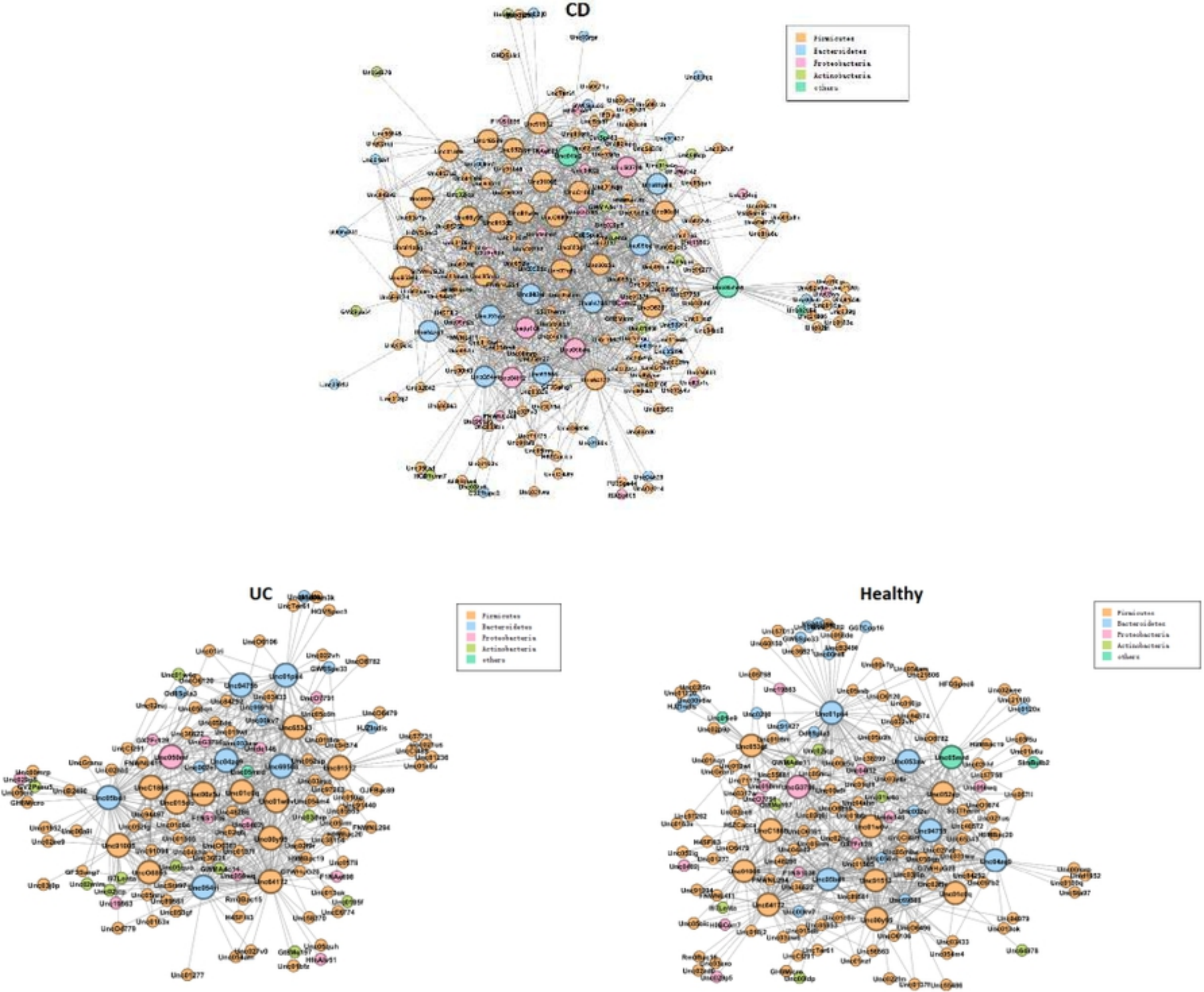
(a) Healthy

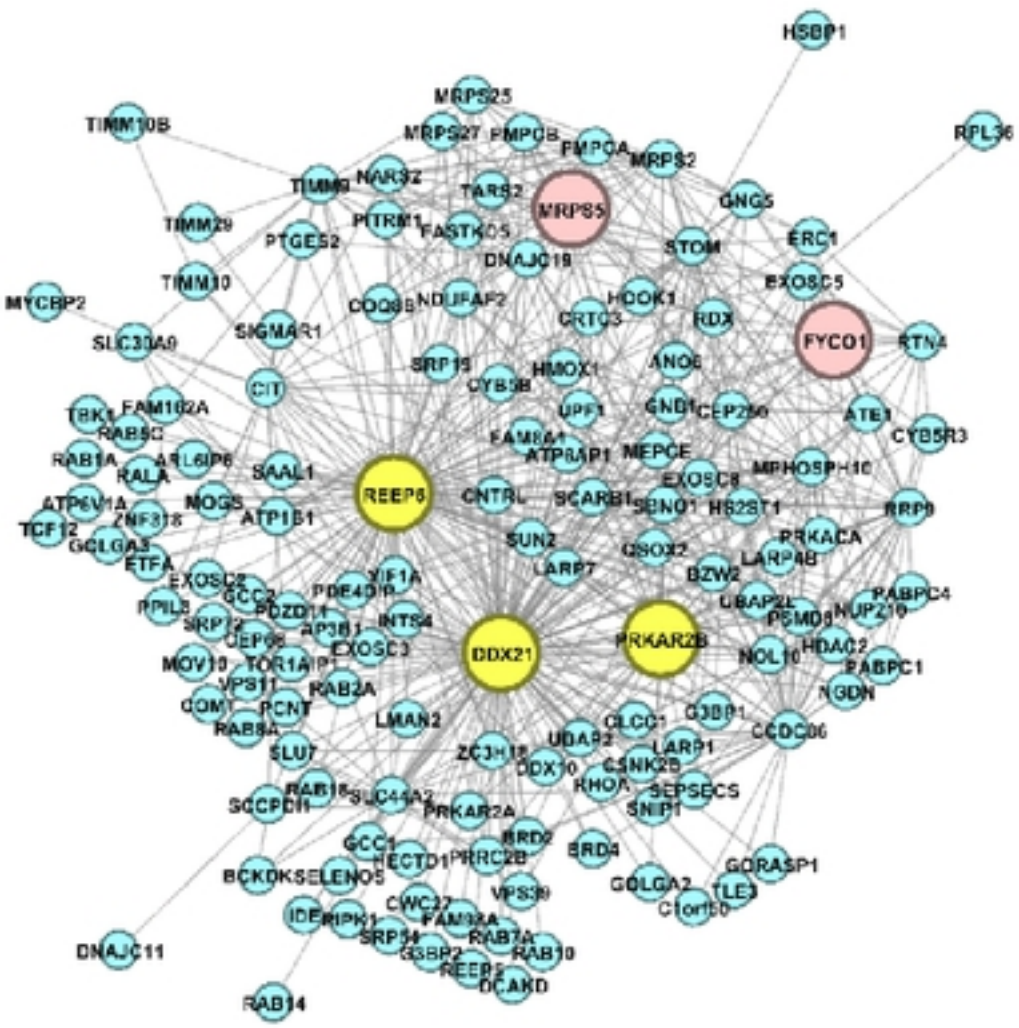


(b) EBA



(c) Control

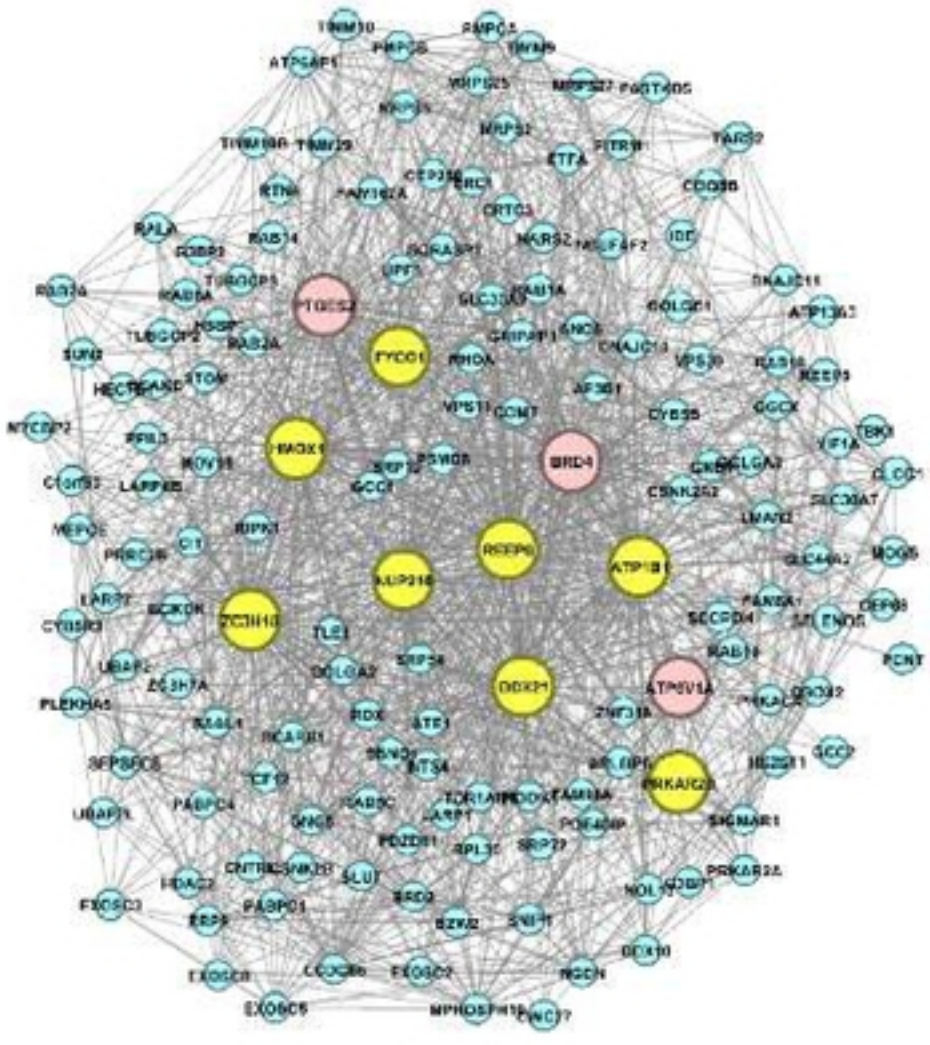

CD
Firmicutes
Bacteroidetes
Proteobacteria
Actinobacteria
others
UC
Firmicutes
Bacteroidetes
Proteobacteria
Actinobacteria
others
Healthy
Firmicutes
Bacteroidetes
Proteobacteria
Actinobacteria
others



(a) Colon

(b) Liver

(c) Lung



(d) Kidney