Molecular ecological network analyses

Ye Deng1,2, Yi-Huei Jiang1, Yunfeng Yang3, Zhili He1, Feng Luo5 and Jizhong Zhou1,3,4*

Abstract

Background: Understanding the interaction among different species within a community and their responses to environmental changes is a central goal in ecology. However, defining the network structure in a microbial community is very challenging due to their extremely high diversity and as-yet uncultivated status. Although recent advance of metagenomic technologies, such as high throughout sequencing and functional gene arrays, provide revolutionary tools for analyzing microbial community structure, it is still difficult to examine network interactions in a microbial community based on high-throughput metagenomics data.

Results: Here, we describe a novel mathematical and bioinformatics framework to construct ecological association networks named molecular ecological networks (MENs) through Random Matrix Theory (RMT)-based methods. Compared to other network construction methods, this approach is remarkable in that the network is automatically defined and robust to noise, thus providing excellent solutions to several common issues associated with high-throughput metagenomics data. We applied it to determine the network structure of microbial communities subjected to long-term experimental warming based on pyrosequencing data of 16S rRNA genes. We showed that the constructed MENs under both warming and unwarming conditions exhibited topological features of scale free, small world and modularity, which were consistent with previously described molecular ecological networks. Eigengene analysis indicated that the eigengenes represented the module profiles relatively well. In consistency with many other studies, several major environmental traits including temperature and soil pH were found to be important in determining network interactions in the microbial communities examined. To facilitate its application by the scientific community, all these methods and statistical tools have been integrated into a comprehensive Molecular Ecological Network Analysis Pipeline (MENA), which is open-accessible now (http://ieg2.ou.edu/MENA).

Conclusions: The RMT-based molecular ecological network analysis provides powerful tools to elucidate network interactions in microbial communities and their responses to environmental changes, which are fundamentally important for research in microbial ecology and environmental microbiology.

Keywords: Ecological network, Random Matrix Theory, Microbial community, Microbiological ecology, Network interaction, Environmental changes

Background

In an ecosystem, different species/populations interact with each other to form complicated networks through various types of interactions such as predation, competition and mutualism. On the basis of ecological interactions, ecological networks can be grouped as antagonistic, competitive and mutualistic networks [1]. Traditionally, food webs have been intensively studied in ecological research because they are critical to study the complexity and stability of ecological communities [2,3]. Recent studies showed that food webs possessed typical properties of network topology (e.g. degree distribution, small world effect) [1,4,5]. Within the last decade, mutualistic networks have also been intensively studied [6]. But, it appears that no studies have been performed to examine competitive networks. This is most likely because the network structure is not available based on competitive interactions. Unlike food webs and plant-animal mutualistic networks where the structure is already known, quantifying competitive interactions among different species/populations within a given habitat is difficult so that the network structure for
competitive interactions is unknown. This is also true for network studies in microbial ecology. Because of their vast diversity, as-yet uncultivated status [7] and of the lack of appropriate theoretical frameworks and experimental data, very few community-scale network studies have been performed in microbial ecology.

Various network approaches have been developed and widely applied in genomic biology [8]. To reveal the interactions among biological molecules including genes and proteins, differential equation-based network methods [9–12], Bayesian network methods [13,14], and relevance/co-expression network methods [15–20], have been used to infer cellular networks based on gene expression data. Among them, the correlation-based relevance network method is most commonly used largely due to its simple calculation procedure and noise tolerance [21]. However, most studies involving relevance network analysis use arbitrary thresholds, and thus the constructed networks are subjective rather than objective [8]. This problem has been solved by our recent development of a random matrix theory (RMT)-based approach, which is able to automatically identify a threshold for cellular network construction from microarray data [22–24]. Our results showed that the developed novel RMT-based approach can automatically identify cellular networks based on microarray data. Our results also indicated that this approach is a reliable, sensitive and robust tool for identifying transcriptional networks for analyzing high-throughput genomics data for modular network identification and gene function prediction [22,23].

High-throughput technologies such as microarrays and high throughput sequencing have generated massive amounts of data on microbial community diversity and dynamics across various spatial and temporal scales [25,26]. These data offer an unprecedented opportunity to examine interactions among different microbial populations [7]. Recently, a novel conceptual framework, termed molecular ecological networks (MENs), has been proposed and applied to characterize microbial communities in response to elevated CO$_2$ [27,28]. Here, we provide detailed mathematical and bioinformatic foundation of this novel approach, and further applications to characterize microbial community network interactions in response to long-term experimental warming. Additionally, we provide an online tool, named the Molecular Ecological Network Analyses Pipeline (MENA pipeline), which is freely accessible to the scientific community.

**Results**

**Overview of MENA**

An ecological network is a representation of various biological interactions (e.g., predation, competition, mutualism) in an ecosystem, in which species (nodes) are connected by pairwise interactions (links) [1,29–32]. As previously described, we refer such molecule-based ecological networks in microbial communities as molecular ecological networks (MENs) [27,28], in which different nodes (molecular markers, e.g., OTUs, functional genes, intergenic regions) are linked by edges (i.e., interactions). The MENs derived from functional gene markers are referred as functional molecular ecological networks (fMENs) [27] and those based on phylogenetic gene markers as phylogenetic molecular ecological networks (pMENs) [28].

The whole process of MENA can be divided into two phases and each phase is comprised of several major steps (Figure 1). The first phase is network construction, which includes four major steps: data collection, data transformation/standardization, pair-wise similarity matrix calculation, and the adjacent matrix determination by RMT-based approach. Among them, the last step is the key to RMT-based network construction (Figure 2), which has been well established in biological systems [22,23]. Once the adjacency matrix is defined, an undirected network graph can be drawn. The second phase of MENA is network analyses, which is composed of network topology characterization (Table 1, 2), module detection, module-based eigengene analysis and identification of modular roles. These methods are important for revealing the networks' overall and modular organizations and identifying key populations at OTU level. In addition, eigengene network analysis can be performed to reveal higher order organization of MENs, and the associations of network properties to environmental characteristics can be established. Finally, the network differences can be compared under different conditions to analyze how environments affect network structure and interactions.

**Molecular network under experimental warming**

Here we used 16 S rRNA gene-based pyrosequencing data from a long-term experimental warming site [48] to construct pMENs and demonstrate the whole process of MENA. The experimental site was established in grassland with two atmospheric temperature treatments, ambient (unwarming) and +2 °C warming. Six replicate plots were set up for each treatment. The environmental DNA of microbial community was extracted from the soil samples of those 12 plots and 2 or 3 unique tags with 16 S rRNA gene conserved primers were used to amplify the V4-V5 hypervariable regions of the 16 S rRNA genes. Altogether, there were 14 replicate datasets for each treatment of warming or unwarming. After preprocessing all raw sequences, the numbers of sequences for all 28 samples ranged from 1,033 to 5,498. After defining OTUs within 0.03 sequence difference, an OTU distribution table with 1,417 distinct
OTUs across 28 samples was obtained. Since the numbers of sequences of all samples were diverse, the abundance data were transformed into relative abundance by dividing the sum of each sample as described previously [48]. The relative abundance table was split into two datasets: warming and unwarming. For each of the datasets, only the OTUs appeared in 7 or more replicates were used for correlation calculations, resulting in 228 and 197 OTUs for warming and unwarming datasets, respectively. After threshold scanning through RMT-based approach, the phylogenetic molecular ecological networks (pMENs) under warming and unwarming conditions were constructed with an identical similarity threshold 0.76 (Table 3). The final warming and unwarming pMENs included 177 and 152 nodes which had at least one edge, and 279 and 263 total edges, respectively.

The robustness of MENs to noise
In order to examine the robustness of MEN approach to noise, different levels (1 to 100 % of original standard deviation) of Gaussian noise were added to the warming dataset. Once various levels of noise were added, new correlation matrices based on these noise-added datasets were calculated. The same similarity threshold used for the original datasets was used for defining adjacency matrices in the new datasets. When less than 40 % noise was added, roughly 90 % of the original OTUs were still detected in the perturbed networks (Figure 3). With 100 % Gaussian noise, more than 85 % nodes from

---

**Figure 1 Overview of the Random Matrix Theory (RMT)-based molecular ecological network analysis.** Two major parts are included, network construction and network analyses. In each of them, several key steps are outlined.
original network were still preserved and accounted for 75% nodes of perturbed network. These results indicate that the RMT-based MEN construction approach is robust to noise.

**The overall MENs topology**

Scale-free, small-world, modularity and hierarchy are common network properties in many complex systems (Table 2) [8,53,54]. The overall topology indices (Table 3) revealed that all curves of network connectivity distribution were fitted well with the power-law model ($R^2$ values from 0.74 to 0.92), indicative of scale-free networks. Also, the average path lengths ($GD$) were 3.09 to 5.08, which were close to logarithms of the total number of network nodes and comparable to those in other networks displaying small-world behavior, suggesting that the MENs in these microbial communities had the typical property of small world. For modularity, all modularity values ($M$) were from 0.44 to 0.86, which were significantly higher than the $M$ values from their corresponding randomized networks. Therefore, all constructed MENs appeared to be modular. Finally, the hierarchy property was examined by the scaling of clustering coefficient. $R^2$ values of the linear relationship between logarithms of clustering coefficients and the logarithms of connectivity ranged from 0.10 to 0.73, indicating the hierarchical behavior was quite variable. MENs from certain habitats may have highly hierarchical structures like sediment samples from Lake DePue (0.73), but others may not (Table 3). Overall, our constructed MENs from different habitats clearly exhibit scale free, small world and modularity properties, but hierarchy property is displayed on certain networks.

**Modular structure**

Modularity is a very important concept in ecology. It could originate from specificity of interactions (e.g.
Table 1 The network topological indexes used in this study

| Indexes                   | Formula                        | Explanation                                                                 | Note                                                                 | Ref  |
|---------------------------|--------------------------------|-----------------------------------------------------------------------------|                                                                      |      |
| **Part I: network indexes for individual nodes** |                                |                                                                            |                                                                      |      |
| Connectivity              | $k_i = \sum_{j} a_{ij}$        | $a_{ij}$ is the connection strength between nodes i and j.                 | It is also called node degree. It is the most commonly used concept for describing the topological property of a node in a network. | [33] |
| Stress centrality         | $SC_i = \sum_{j} \sigma(j, i, k)$ | $\sigma(j, i, k)$ is the number of shortest paths between nodes j and k that pass through node i. | It is used to describe the number of geodesic paths that pass through the ith node. High Stress node can serve as a broker. | [34] |
| Betweenness               | $B_i = \sum_{j} \frac{\sigma(j, k)}{\sigma(j, k) - 1}$ | $\sigma(j, k)$ is the total number of shortest paths between j and k.       | It is used to describe the ratio of paths that pass through the ith node. High Betweenness node can serve as a broker similar to stress centrality. | [34] |
| Eigenvector centrality    | $EC_i = \frac{1}{2} \sum_{j} M(\{i\}) \cdot EC_j$ | $M(\{i\})$ is the set of nodes that are connected to the ith node and $\lambda$ is a constant eigenvalue. | It is used to describe the degree of a critical node that is connected to other central nodes. | [35] |
| Clustering coefficient    | $CC_i = \frac{2l_i}{k_i (k_i - 1)}$ | $l_i$ is the number of links between neighbors of node i and $k_i$ is the number of neighbors of node i. | It describes how well a node is connected with its neighbors. If it is fully connected to its neighbors, the clustering coefficient is 1. A value close to 0 means that there are hardly any connections with its neighbors. It was used to describe hierarchical properties of networks. | [36,37] |
| Vulnerability             | $Vi = \frac{E_i}{E}$          | $E$ is the global efficiency and $E_i$ is the global efficiency after the removal of the node i and its entire links. | It measures the disease of node i on the system performance if node i and all associated links are removed. | [38] |
| **Part II: The overall network topological indexes** |                                |                                                                            |                                                                      |      |
| Average connectivity      | $avgK = \frac{1}{n} \sum_{i} k_i$ | $k_i$ is degree of node i and n is the number of nodes.                    | Higher $avgK$ means a more complex network.                          | [39] |
| Average geodesic distance | $GD = \frac{1}{(n-1)} \sum_{i \neq j} d_{ij}$ | $d_{ij}$ is the shortest path between node i and j.                      | A smaller $GD$ means all the nodes in the network are closer.         | [39] |
| Geodesic efficiency       | $E = \frac{1}{(n-1)} \sum_{i \neq j} \frac{1}{d_{ij}}$ | All parameters shown above.                                              | It is the opposite of $GD$. A higher $E$ means that the nodes are closer. | [40] |
| Harmonic geodesic distance| $HD = \frac{1}{E}$            | $E$ is geodesic efficiency.                                               | The reciprocal of $E$, which is similar to $GD$ but more appropriate for disjoint graph. | [40] |
| Centralization of degree  | $CD = \sum_{i=1}^{n} (\max(k) - k_i)$ | max(k) is the maximal value of all connectivity values and $k_i$ represents the connectivity of $i^{th}$ node. Finally this value is normalized by the theoretical maximum centralization score. | It is close to 1 for a network with star topology and in contrast close to 0 for a network where each node has the same connectivity. | [41] |
| Centralization of betweenness | $CB = \sum_{i=1}^{n} (\max(B) - B_i)$ | max(B) is the maximal value of all betweenness values and $B_i$ represents the betweenness of $i^{th}$ node. Finally this value is normalized by the theoretical maximum centralization score. | It is close to 0 for a network where each node has the same betweenness, and the bigger the more difference among all betweenness values. | [41] |
| Centralization of stress centrality | $CS = \sum_{i=1}^{n} (\max(SC) - SC_i)$ | max(SC) is the maximal value of all stress centrality values and $SC_i$ represents the stress centrality of $i^{th}$ node. Finally this value is normalized by the theoretical maximum centralization score. | It is close to 0 for a network where each node has the same stress centrality, and the bigger the more difference among all stress centrality values. | [41] |
predation, pollination), habitat heterogeneity, resource partition, ecological niche overlap, natural selection, convergent evolution, and phylogenetic relatedness, and it could be important for system stability and resilience [55]. In MENs, a module in the network is a group of OTUs that are highly connected among themselves, but had much fewer connections with OTUs outside the group. Random matrix theory-based approach is able to delineate separate modules, but some modules could still be very big.

We used several methods, including short random walks [56], leading eigenvector of the community matrix [57], simulated annealing approach [58,59] and the greedy modularity optimization [57], to define modules and submodules within a large module. From the evaluation of warming and unwarming pMENs, short random walks generated 27 and 31 modules with average values 0.61 and 0.56, respectively; the leading eigenvector of the matrix generated 22 and 28 modules with average values 0.61 and 0.54, respectively; the greedy modularity optimization had 18 and 20 modules with average values 0.67 and 0.61, respectively; the simulated annealing approach had average 18 and 19 modules with average values 0.67 and 0.61, respectively. From these results, the greedy modularity optimization and simulated annealing approach had higher module values than two other approaches, indicating they are more effective in separating the complex networks into submodules. Notably, since the simulated annealing approach was stochastic [59], the submodules of pMENs generated by this approach were slightly different with different runs. Therefore, the greedy modularity optimization approach was preferred to identify the submodular structure of MENs. The modular pMEN of warming pyrosequencing dataset was shown in Figure 4A. A total of 10 joint submodules with ≥8 nodes were isolated from a single large module and all the other isolated modules were relatively small (2 to 4 nodes). The size of modules or submodules varied with 2 to 24 nodes.

### Eigengene network analysis and the modular topological roles

After modules and submodules are determined, the eigengene analysis is used to reveal higher order organizations in the network structure [60–62]. In the eigengene analysis, each module is represented by its singular value decomposition (SVD) of abundance profile called module eigengene [62]. In the warming pMEN, the module eigengenes from top 10 large submodules (≥8 nodes) explained 30 - 68 % variations of relative abundance across different replicates, suggesting that these eigengenes represented the module profiles relatively well. The correlations among module eigengenes were used to define the eigengene network. Eigengene analysis is important for revealing higher order organization and identifying key populations based on network topology [62]. In warming pMEN, these correlations of 10 largest submodules were visualized as a heat-map and hierarchical clustering diagram (Figure 4B). The eigengenes within several groups of submodules showed significant
correlations and clustered together as super-groups, such as #6 and #8, #2, #5 and #3, and #1, and #7 and #9, which were referred as meta-modules among submodules. Besides, within each module, eigengene analysis approach was able to show the representative abundance profile and identify key members as shown in our previous paper [28].

Different nodes play distinct topological roles in the network [33]. The analysis of modular topological roles is important to identify key populations or functional networks based on the nodes’ roles in their own modules. Their topological roles can be defined by two parameters, within-module connectivity (zi) and among-module connectivity (Pi). The topological roles of nodes in warming and unwarming pMENs were illustrated in ZP-plot (Figure 4C). According to values of zi and Pi, the roles of nodes were classified into four categories: peripherals, connectors, module hubs and network hubs. From ecological perspectives, peripherals might represent specialists whereas module hubs and connectors were close to generalists and network hubs as super-generalists [55]. Here, the majority of OTUs (90.9 %) under warming and unwarming conditions were peripherals with most of their links inside their own modules. A total of 26 nodes (7.9 %) were connectors and only four nodes (1.2 %) were module hubs. Those four OTUs as module hubs were derived from Planctomyces (Planctomycetes), Nocardioides (Actinobacteria) under warming condition, and Thermoleophilum (Actinobacteria) and GP4 (Acidobacteria) under unwarming condition, indicating that the hubs of pMENs were substantially different under different conditions.

The correlations between network topologies with environmental traits

The relationships between microbial network topology and environmental characteristics can be examined in both direct and indirect ways. Indirectly, as a first step, the OTU significance (GS) is calculated and defined as the square of Pearson correlation coefficient (r²) of OTU abundance profile with environmental traits. Then the correlation between GS and nodes’ topological indices (e.g., connectivity) was used to measure the relationship of network topology with traits. For instance, in warming pMEN, the GS of average soil temperature was significantly correlated with the nodes’ connectivity (r = 0.30, p = 4.7 × 10⁻⁵), indicating that the nodes with higher connectivity were inclined to have closer relationships with temperature. If multiple GS was involved, Mantel and partial Mantel tests could be implemented to calculate correlations between the connectivity and multiple GS of environmental traits to reveal the internal associations between network topology and environmental changes. In warming pMEN, the nodes’ connectivity was significantly associated with the GS of pH values, soil NO₃-nitrogen and soil carbon contents when the effect of temperature was controlled (rₖₐ₉ = 0.104, P = 0.013). Meanwhile, the GS of temperature was also significantly associated with both temperature and environmental factors such as soil geochemistry factors were controlled (rₖₐ₉ = 0.159, P = 0.003) (Table 4). Moreover, the OTUs of β-Proteobacteria and Verrucomicrobia were highly associated with the changes of soil geochemistry (rₖₐ₉ = 0.59 and 0.926 respectively, both P = 0.013). These results suggested that the OTUs topology in warming pMEN was significantly associated with both temperature and the selected soil variables. In addition, OTUs from β-Proteobacteria and Verrucomicrobia were most sensitive to the changes of soil variables.

The correlations between module-based eigengenes and environmental factors can be used to detect the modules’ response to environmental changes. In warming pMEN, the coefficients (r values) and significances (p values) were shown in a heatmap (Figure 5). Submodules #1 and #9 were positively correlated with the average soil temperature significantly (p < 0.01) but
Table 3 Topological properties of the empirical molecular ecological networks (MENs) of additional microbial communities and their associated random MENs*

| Habitats of communitiesb | Similarity threshold (st) | Network size (n) | $R^2$ of power law | $R^2$ of scaling law | Average path (GD) | Average Clustering coefficient (avgCC) | Modularity & (the number of modules) | Average path (GD) | Average clustering coefficient (avgCC) | Modularity (M) |
|-------------------------|--------------------------|-----------------|--------------------|---------------------|------------------|----------------------------------------|-----------------------------------|------------------|---------------------------------------|----------------|
| **Functional MENs**     |                          |                 |                    |                     |                  |                                        |                                    |                  |                                       |                |
| Grassland soils under elevated CO$_2$, MN (i) | 0.80                     | 254             | 0.79               | 0.25                | 3.09             | 0.22                                   | 0.44 (18)                       | 3.00 ± 0.03   | 0.099 ± 0.009                         | 0.31 ± 0.01 |
| Grassland soils under ambient CO$_2$, MN (i) | 0.80                     | 184             | 0.88               | 0.11                | 4.21             | 0.10                                   | 0.65 (16)                       | 3.84 ± 0.06   | 0.028 ± 0.007                         | 0.52 ± 0.01 |
| Lake sediment, Lake DePue, WI (ii) | 0.92                     | 151             | 0.85               | 0.73                | 3.47             | 0.09                                   | 0.48 (8)                         | 3.46 ± 0.05   | 0.046 ± 0.010                         | 0.45 ± 0.01 |
| Groundwater, Well 101–2, Oak Ridge, TN (iii) | 0.95                     | 107             | 0.74               | 0.44                | 3.12             | 0.29                                   | 0.52 (11)                       | 3.13 ± 0.07   | 0.081 ± 0.017                         | 0.40 ± 0.01 |
| Groundwater Well 102–2, Oak Ridge, TN (iii) | 0.89                     | 140             | 0.79               | 0.21                | 4.22             | 0.17                                   | 0.67 (12)                       | 3.89 ± 0.08   | 0.033 ± 0.012                         | 0.53 ± 0.01 |
| Groundwater Well 102–3, Oak Ridge, TN (iii) | 0.87                     | 117             | 0.85               | 0.19                | 3.57             | 0.25                                   | 0.64 (13)                       | 3.54 ± 0.09   | 0.049 ± 0.013                         | 0.48 ± 0.01 |
| **Phylogenetic MENs (454 pyrosequencing)** |                          |                 |                    |                     |                  |                                        |                                    |                  |                                       |                |
| Grassland soils under warming, Norman, OK (iv) | 0.76                     | 177             | 0.83               | 0.48                | 3.91             | 0.13                                   | 0.67 (18)                       | 3.94 ± 0.20   | 0.020 ± 0.008                         | 0.44 ± 0.01 |
| Grassland soils under unwarming, Norman, OK (iv) | 0.76                     | 152             | 0.88               | 0.10                | 2.71             | 0.09                                   | 0.61 (20)                       | 3.39 ± 0.23   | 0.038 ± 0.010                         | 0.47 ± 0.01 |
| Grassland soils under elevated CO$_2$, MN (i) | 0.78                     | 263             | 0.89               | 0.26                | 3.95             | 0.25                                   | 0.81 (34)                       | 3.98 ± 0.22   | 0.015 ± 0.006                         | 0.61 ± 0.02 |
| Grassland soils under ambient CO$_2$, MN (i) | 0.77                     | 292             | 0.87               | 0.22                | 4.26             | 0.27                                   | 0.85 (36)                       | 4.10 ± 0.20   | 0.017 ± 0.005                         | 0.59 ± 0.01 |
| Agricultural soil, Africa (v) | 0.77                     | 384             | 0.86               | 0.20                | 4.99             | 0.34                                   | 0.86 (32)                       | 3.99 ± 0.04   | 0.020 ± 0.004                         | 0.48 ± 0.01 |
| Human intestine, Stanford, CA (vi) | 0.86                     | 215             | 0.92               | 0.18                | 3.55             | 0.13                                   | 0.69 (27)                       | 4.23 ± 0.10   | 0.025 ± 0.009                         | 0.58 ± 0.01 |

*Various parameters of the empirical networks and generation of random networks are explained in the Table 1.

bSample sources: (i) The grassland soils under elevated and ambient CO$_2$ were collected from a free-air CO$_2$ enrichment field in Minnesota which were analyzed with both GeoChip3.0 and 16 S pyrosequencing [49]. The MENs analysis was designed in Zhou et al. [27] and pMENs analysis was designed in Zhou et al. [28]. (ii) The lake sediment samples from Lake DePue were analyzed with GeoChip2.0. (iii) The groundwater samples from three different Wells in Oak Ridge, Tennessee were analyzed with GeoChip 2.0 [50]. (iv) The grassland samples under warming and unwarming were collected from the long term warming experiment at Oklahoma [51] and analyzed with 16 S pyrosequencing [48]. (v) The pyrosequencing data of agricultural soils from Africa and the groundwater samples from Oak Ridge was provided by Dr. Tiedje and his colleagues at Michigan State University. (vi) The human intestine sample from Stanford was described elsewhere [52].
negatively \((p < 0.01)\) with soil pH values and soil carbon contents, indicating that the members in these two sub-modules might be stimulated by temperature but inhibited by soil pH and carbon. Also, submodules #6 and #8 were positively correlated with soil pH \((p < 0.01)\), #4 was positively correlated with NO\(_3\)- concentration \((p = 0.001)\) and soil carbon content \((p = 0.013)\). While #3 was positively correlated with soil carbon content \((p = 0.016)\), #7 was negatively correlated with soil carbon content \((p = 0.025)\). In addition, #2 and #6 were negatively correlated with temperature \((p < 0.05)\). All above results demonstrated that different submodules in warming pMEN responded to the environmental changes differently and the changes of temperature could have significant impacts on members of some submodules (e.g., #1, #2, #6 and #9).

### Open-access pipeline

To facilitate the application of MENA in the scientific community, an open-access pipeline for MEN construction and related analyses (Figure 6). RMT-based threshold searching is performed using a Java script [22] and some network analyses are called in the programs of sna [63], igraph [64] and WGCNA [65] packages in the R project. The MENAP includes the following components: (i) user registration and login, (ii) data upload, (iii) network construction by the RMT-based method (perhaps other methods as well), (iv) network analysis, and (v) dataset and network management (Figure 6).

The network analysis component is further divided into three major parts:

(a) **Network characterization.** Various network properties are calculated and evaluated, such as connectivity, betweenness, clustering coefficient, and geodesic distance. The module/submodule detection and modularity analyses is performed using fast greedy modularity optimization [66]. Eigengene network analysis is performed to understand network characteristics at higher organization levels and to identify key microbial populations or key functional genes in terms of network topology.

(b) **Network visualization.** An automatic pipeline is constructed to visualize the constructed network. Moreover, the file format for software Cytoscape 2.6.0 [67] is prepared to visualize more complex and delicate network graphs. Other data associated with OTUs, such as taxonomy, relative abundance, edge information, and positive and negative correlations is imported and visualized in network figures.

(c) **Network comparison.** Various randomization methods like the Maslov-Sneppen method [68] are used obtain random networks for network comparison. Various indices are evaluated for comparing the differences of networks among different communities in terms of sensitivity and robustness. In addition, OTU significances are calculated to reveal associations of the network structure to the ecological functional traits [27].

### Discussion and conclusions

Most previous studies on the biodiversity of microbial communities have been focused on the number of species and the abundance of species, but not interactions among species. However, species interactions could be more important to ecosystem functioning than species richness and abundance, especially in complex ecosystems [1,27–29]. Several recent analyses show that the ecological networks of ecosystems are highly structured [1,69,70], thus ignoring the structure of network and the interactions among network components precludes further assessment of biodiversity and its dynamics. Several recent breakthroughs have been made to analyze species...
interactions of animals and plants [1,4,31,55,70,71], but it is difficult to detect network interactions of a microbial community [72–74]. Therefore, in this study, we systematically described a mathematical and bioinformatic framework of MENA based on RMT, a powerful method well established in quantitative physics [23,75,76]. Our results demonstrate that the RMT-based approach is powerful in discerning network interactions in microbial communities.

The network approach described is based on the transition of two universal distributions from the random matrix theory. A major advantage of RMT method is that the threshold to construct network is automatically determined. In contrast, most other methods studies use arbitrary thresholds, which are usually based on limited knowledge of biological information [8,72–74,77]. RMT-based approach selects an optimal threshold without ambiguity, which ensures its construction of optimal

---

**Figure 4 The submodules of the warming pMEN.** (A) The network graph with submodule structure by the fast greedy modularity optimization method. Each node signifies an OTU, which could correspond to a microbial population. Colors of the nodes indicate different major phyla. A blue edge indicates a positive interaction between two individual nodes, while a red edge indicates a negative interaction. (B) The correlations and heatmap to show module eigengenes of warming pMEN. The upper part is the hierarchical clustering based on the Pearson correlations among module eigengenes and the below heatmap shows the coefficient values (r). Red color means higher correlation whereas green color signified lower correlation. (C) ZP-plot showing distribution of OTUs based on their module-based topological roles. Each dot represents an OTU in the dataset of warming (red), or unwarming (green). The topological role of each OTU was determined according to the scatter plot of within-module connectivity (z) and among-module connectivity (P) [55,60].
networks. Another advantage of RMT-based approach is its remarkable capacity in tolerating noise, resulting in reliable, robust networks. Our results show even with 100% Gaussian noise, more than 85% nodes from original network are still preserved. This characteristic could be very important for dealing with the large-scale data, such as metagenomics and microarrays, which are generally inherent with high noise.

Nevertheless, characterizing ecological network of microbial communities poses major challenges. MENs are constructed by the adjacency matrix originated from the pair-wise correlations of relative OTU abundance across different samples. Therefore, a network interaction between two OTUs or genes describes the co-occurrence of these two OTUs or genes across different samples. The co-occurrence might be caused by species or genes performing similar or complementary functions, or shared environmental conditions that microbial species coexist in [28]. However, the former possibility can be complicated by the observations that functionally redundant genes are not necessary co-regulated, but instead co-regulated through other genes, which is coined as transitive co-regulation [78]. The latter possibility can be complicated by the distinctiveness of individuals in microbial niches observed in their behaviors and responses to environmental perturbation [79]. Therefore, caution must be taken for the interpretation of underlying mechanisms that shape microbial communities.

A long-held tenet is that the structure of ecological networks has significant influence on the dynamics [1,80]. Most complex systems have common characteristics such as small world, scale-free, modularity and hierarchy [8,53,54]. Consistently, MENs were found to be scale-free, small world and modular, in addition to hierarchical property in some MENs. These network properties are important for the robustness and stability of complex systems [8,27,28,81]. For example, our results showed that any two microbial species in the community can be linked by just a few other neighbor species, showing small-world property. This may imply that the energy, materials and information can be easily transported through entire systems. In microbial communities, this characteristic drives efficient communications among different members so that relevant responses can be taken rapidly to environmental changes. Meanwhile, it is intriguing to note that modularity is prevailing in MENs, while hierarchy is present only in some MENs. Research on a wide range of architectural patterns in mutualistic (pollination) and trophic (predation) networks showed that hierarchy, also called nestedness, was strong in mutualistic networks, but that modularity was strong in trophic networks [82]. Although ecological networks of microbial communities are very complicated and cannot be classified into simple mutualistic or trophic networks, it would be interesting to compare a number of ecological networks of microbial communities to catalog different architectural patterns and to explore the mechanisms underlying the stability and resilience of communities.

In addition to interactions among microbes within a community, MENs allow for analyses of interactions with their environment through correlations with abiotic environmental measurements, which might provide insights on the conditions that have significant impact on the co-occurring organisms. It is also possible to link groups of organisms with biogeochemical measurements

| Phylogeny          | # nodes | GS of soil geochemistry | partial $r_M$ | partial $P$ | GS of temperature | partial $r_M$ | partial $P$ |
|--------------------|---------|-------------------------|---------------|-------------|-------------------|---------------|-------------|
| All detected OTUs  | 177     | 0.104                   | 0.018         |             | 0.159             | 0.003         |             |
| Acidobacteria      | 35      | 0.059                   | 0.234         |             |                  | -0.054        | 0.800       |
| Actinobacteria     | 63      | -0.033                  | 0.650         | 0.077       | 0.367             | 0.108         |             |
| Chloroflexi        | 5       | -0.339                  | 0.663         | 0.367       | -0.202            | 0.788         |             |
| Planctomycetacia   | 6       | -0.082                  | 0.521         | 0.096       | 0.155             | 0.155         |             |
| α-Proteobacteria   | 26      | -0.057                  | 0.721         | 0.096       | -0.001            | 0.430         |             |
| β-Proteobacteria   | 12      | 0.590                   | 0.013         |             |                  | -0.013        | 0.430       |
| δ-Proteobacteria   | 6       | 0.338                   | 0.088         | -0.298      | 0.877             | 0.877         |             |
| γ-Proteobacteria   | 4       | 0.030                   | 0.772         | 0.796       | 0.243             | 0.243         |             |
| Verrucomiobia      | 5       | 0.926                   | 0.013         | -0.755      | 1.000             |               |             |

*a* Soil variables used for OTU significance calculations: pH values, NO$_3$-Nitrogen and soil carbon contents.

*b* Correlation coefficient based on Mantel test.

*c* The significance (probability) of Mantel test.
to reveal the functional role of organism in biogeochemical processes. These kinds of data are important for generating hypotheses to help explain natural environments that microbial communities reside, which might lead to forecasting responses of microbial communities when environment changes [73].

In summary, our study provides a mathematical/bioinformatic framework for network construction based on metagenomics data such as sequencing [28] and microarray hybridization data [27]. It is useful, as demonstrated with the microbial communities under experimental warming, for dissecting interactions within a microbial community as well as with environment, thus allowing microbial ecologists to address a variety of ecological questions at the community-wide scale [83,84]. It is also possible to extend MENA to emerging fields of microbial ecology such as high-throughput proteomics, since RMT is not stringent on data types. In addition, broad application of MENA will generate a number of ecological networks that allow for exploration of architectural patterns of microbial communities [1]. This RMT-based molecular ecological network analysis provides powerful tools to elucidate network interactions in microbial communities and their responses to environmental changes, which are fundamentally important for research in microbial ecology, systems microbiology, and global change.

**Methods**

**Data standardization**

The network construction begins with a data table with \(n\) distinct operational taxonomic units (OTUs) based on 16 S rRNA genes or functional genes observed across \(m\) replicates or samples. Typically OTUs are used to refer taxonomic classification based on ribosomal RNA genes. For convenience, in the following sections, we use OTUs to refer the classifications derived from both 16 S rRNA genes and/or functional genes. Let \(y_{ik}\) represent the abundance or relative abundance of the \(i\)-th OTU in the \(k\)-th sample (\(i \in \{1, \ldots, n\}\); \(k \in \{1, \ldots, m\}\)) and \(Y_{nxm} = [y_{ik}]\) is the abundance matrix. Usually, the abundance profile of \(i\)-th OTU is standardized as below. If the mean and standard deviation of \(y_i\) across all samples are \(\bar{y}_i\) and \(\sigma_i\), the standardized abundance of the \(i\)-th OTU in the \(k\)-th sample is

\[
x_{ik} = \frac{y_{ik}}{\bar{y}_i} \frac{1}{\sigma_i},
\]

where \(x_{ik}\) has mean value of 0 and variance value of 1. \(X_{nxm}\) is the standardized data matrix and used for subsequent correlation analysis.

**Defining adjacency matrix**

Molecular ecological networks can be built on the basis of the measurements of relative OTU abundance in microbial communities. In MENs, each OTU corresponds to a node. Each network corresponds to an adjacency matrix (or interaction matrix), \(A_{nxn} = [a_{ij}]\), which encodes the connection strength between each pair of nodes [20]. In an unweighted network, the adjacency \(a_{ij} = 1\) if nodes \(i\) and \(j\) are connected, and \(a_{ij} = 0\) otherwise [20]. For an undirected network, the adjacency matrix is symmetric. In weighted network, the pairwise adjacency has values between 0 and 1, i.e., \(0 \leq a_{ij} \leq 1\). The adjacency matrix is the foundation of all subsequent steps in network analysis.

To define the adjacency matrix, the similarity of OTU abundance across all samples should be measured first. Such similarity measures the degree of concordance between the abundance profiles of OTUs across different

![Figure 5](https://example.com/figure5.png)

**Figure 5** The correlations between module eigengenes and environmental traits in the warming pMEN. The color of each plot indicates the correlation between corresponding module eigengene and environmental trait. Red color means highly positive correlation and green color means highly negative correlation. The numbers in each plot are the correlation coefficient (\(r\)) and significance (\(p\)) in parentheses. The environmental traits include soil pH value (pH), NO3-nitrogen content (NO3N), soil carbon content (SC) and average soil temperature (avgT).
Similar to widely used gene co-expression analyses \[20,61,85,86\], pairwise Pearson correlation coefficients \((r_{ij})\) are used to measure the similarity between \(i\)-th and \(j\)-th OTU across different samples. Let \(R_{n \times n} = [r_{ij}]\) be the Pearson correlation matrix, then

\[
r_{ij} = \frac{\text{cov}(x_i, x_j)}{\sqrt{\text{var}(x_i) \text{var}(x_j)}} = \frac{\sum_{k=1}^{m} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{m} (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{m} (x_{jk} - \bar{x}_j)^2}}
\]

where \(x_{ik}\) and \(x_{jk}\) are the standardized abundance of the \(i\)-th and \(j\)-th OTUs in the \(k\)-th sample. \(\bar{x}_i\) and \(\bar{x}_j\) are the mean values of the \(i\)-th and \(j\)-th OTUs over samples. In general, the absolute value of the correlation coefficient \((r_{ij})\) is used to define the abundance similarity between \(i\)-th and \(j\)-th OTU \((s_{ij})\), that is

\[
s_{ij} = |r_{ij}|, \text{ where } i, j \in \{1, \ldots, n\}
\]

Between the \(i\)-th and \(j\)-th OTU is defined by thresholding the OTU abundance similarity \[33\]:

\[
a_{ij} = \begin{cases} s_{ij} & \text{if } s_{ij} \geq s_{th} \\ 0 & \text{if } s_{ij} < s_{th} \end{cases}
\]

where \(s_{th}\) is the threshold parameter. The resulting adjacency matrix, \(A^{pp}\), is generally smaller than the similarity matrix because the rows or columns are removed if all of their elements are less than the threshold value.

**Determining the threshold by random matrix theory-based approach**

The structure of relevance network strongly depends on the threshold value, \(s_t\). In some network analysis, the threshold value is chosen arbitrarily based on known biological information or set by the empirical study \[8\]. Thus, the resulting network is more or less subjective \[19,20,85,87\]. However, it is difficult to select appropriate thresholds, especially for poorly studied organisms/communities. In MENA, we use the random matrix theory (RMT)-based approach, which is able to identify the threshold automatically based on the data structure itself \[22,46\] to select the final threshold parameter, \(s_t\).

**Basic concept of RMT**

Initially proposed by Wigner and Dyson in the 1960s for studying the spectrum of complex nuclei \[88\], random matrix theory (RMT) is a powerful approach for
identifying and modeling phase transitions associated with disorder and noise in statistical physics and materials science. It has been successfully used for studying the behavior of different complex systems, such as spectra of large atoms [89], metal insulator transitions in disorder systems, spectra of quasiperiodic systems [90], chaotic systems [91], the stock market [76], brain response [92], gene co-expression networks [22] and protein interaction networks [46]. However, its suitability for complex biological systems, especially microbial communities, remains largely unexplored.

RMT predicts two universal extreme distributions of the nearest neighbor spacing distribution (NNSD) of eigenvalues: Gaussian orthogonal ensemble (GOE) statistics, which corresponds to random properties of complex system, and Poisson distribution, which corresponds to system-specific, nonrandom properties of complex systems [89]. These two different universal laws depend on the properties of the matrix. On one hand, if consecutive eigenvalues are completely uncorrelated, the NNSD follows Poisson statistics. Considering a series of eigenvalues, the probability of an eigenvalue falling in a scale \([D, D + s]\) is independent of the start point \(D\), where \(s\) can be any positive values. It means the probability of an eigenvalue falling in any scales with certain length \(s\) will be identical, no matter where the scales begin. The NNSD under such assumption follows a Poisson random process, so-called exponential distribution of Poisson process [89]. On the other hand, for correlated eigenvalues, the NNSD has Gaussian orthogonal ensemble (GOE) statistics. Given a series of correlated eigenvalues, the probability of one eigenvalue falling into a scale \([D, D + s]\) is proportional to \(s\). Wigner illustrated that the NNSD under this assumption was closely to Gaussian orthogonal ensemble so-called Wigner surmise [89].

The key concept of RMT is to mainly concern with the local property between eigenvalues rather than the global property of a series of eigenvalues. Here, the local property between eigenvalues means the eigenvalue fluctuations and the global property is the average eigenvalue density. In order to reveal the fluctuations of eigenvalues, the average eigenvalue density has to be removed from system so that the average eigenspacing is constant. Also, this procedure to generate a uniform eigenvalues distribution is called unfolding. The unfolded eigenvalues will fall between 0 and 1, and its density does not depend on the overall level distribution. Consider a sequence of eigenvalues \(\lambda_1, \lambda_2, \ldots, \lambda_n\) from adjacency matrix, and those eigenvalues have been ordered as \(\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n\). In practice, we replace eigenvalues \(\lambda_i\) with \(e_i = N_{av}(\lambda_i)\) where \(N_{av}\) is the continuous density of eigenvalues obtained by fitting and smoothing the original integrated density of eigenvalues to a cubic spline or by local density average.

After unfolding the eigenvalues, three statistical quantities can be used to extract information from a sequence of eigenvalues, namely, eigenvalue spacing distribution \(P(d)\), number variance of eigenvalues \(\sum\), and spectral rigidity \(\Delta\). \(P(d)\) is the probability density function for unfolded eigenvalue spacing, \(d_i = |e_{i+1} - e_i|\), which is the NNSD for eigenvalues. For the completely uncorrelated eigenvalues, \(P(d)\) follows Poisson statistic and it can be expressed by

\[
P(d) = \exp(-d).
\]

On the other hand, for the correlated eigenvalues, \(P(d)\) closely follows Wigner-Dyson distribution of the GOE statistics and it can be expressed by

\[
P(d) \approx \frac{\pi d}{2} \exp\left(-\frac{\pi d^2}{2}\right).
\]

We use the \(\chi^2\) goodness-of-fit test to assess whether NNSD follows Wigner-Dyson distribution or Poisson distribution. We assume that the NNSD of any biological system obeys these two extreme distributions [22,23,27,28], and that there is a transition point from GOE to Poisson distribution, and this transition point can be used as the threshold for defining adjacency matrix.

**Algorithms of detecting the threshold value**

The following major steps are used to define the threshold \(s_t\) based on the standardized relative abundance of OTUs across different samples (Figure 2).

(a) Calculate the Pearson correlation matrix, \(R^{nxn}\), based on the standardized relative abundance of OTUs, \(X^{nxm}\) with \(n\) distinct OTUs across \(m\) samples.

(b) Obtain similarity data, \(S^{nxn}\), by taking the absolute value of correlation matrix \(R^{nxn}\).

(c) Set an initial threshold value, \(s_{th}\) (e.g., 0.3 based on our experiences).

(d) Calculate the adjacency matrix, \(A^{exp} = [a_{ij}]\) according to \(s_{th}\) where \(p\) is the number of OTUs retained in the adjacency matrix with non-zero rows or columns.

(e) Calculate eigenvalues \(\lambda_i\) of the adjacency matrix based on the equation \((S - \lambda I)v = 0\), where \(\lambda\) is the eigenvalue, \(v\) is the corresponding eigenvector, and \(I\) is the identity matrix. Because \(S\) is the symmetric matrix and \(v\) is a non-zero vector, we can get \(p\) number of eigenvalues to solve the equation \((S - \lambda I)v = 0\). To test NNSD distribution, order the eigenvalues as \(\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p\).

(f) To get unfolded eigenvalues, replace \(\lambda_i\) with \(e_i = N_{av}(\lambda_i)\), where \(N_{av}\) is the continuous density of eigenvalues and can be obtained by fitting the
original integrated density to a cubic spline or by local average.

(g) Calculate the nearest neighbor spacing distribution of eigenvalues, \( P(d) \), which defines the probability density of unfolded eigenvalues spacing, \( d_i = |e_{i+1} - e_i| \).

(h) Using the \( \chi^2 \) goodness-of-fit test to determine whether the probability density function \( P(d) \) follows the exponential distribution of Poisson statistic, \( \exp(-d) \).

\[ H_0: P(d) \text{ follows the Poisson distribution.} \]
\[ H_1: P(d) \text{ does not follow the Poisson distribution.} \]

The \( \chi^2 \) goodness-of-fit test has the test statistics, \( \chi^2 = \sum_i \frac{(d_i - E(d_i))^2}{E(d_i)} \), where \( d_i \) is the observed nearest neighbor spacing and \( E(d_i) \) is an expected (theoretical) nearest neighbor spacing from Poisson distribution. The resulting \( \chi^2 \) value is compared to the \( \chi^2 \) distribution. Let \( \chi^2_{u}(0.01) \) be the critical value at a significant level of 0.01 based on \( \chi^2 \) distribution with \( u \) degrees of freedom.

(i) If \( \chi^2 \geq \chi^2_{u}(0.01) \), the null hypothesis \( H_0 \) is not rejected. Then go to step (j).

If \( \chi^2 < \chi^2_{u}(0.01) \), the null hypothesis \( H_0 \) is rejected. Then, increase the threshold by 0.1, \( s_{tb} + 0.1 \), and repeat the steps from (e) to (h).

(j) Find a finer scale threshold value by increasing the threshold with 0.01 within the range of \([s_{tb}, 0.1, s_{tb}]\). Then repeat the steps from (e) to (h).

(k) If \( H_0 \) is accepted, i.e., the \( P(d) \) follows Poisson distribution, the finer scale threshold identified is used as the optimal threshold for defining the adjacency matrix.

Once the final threshold value \( s_t \) is determined at a finer scale, an adjacency matrix is obtained by retaining all the OTUs whose abundance similarity values are greater than the determined threshold. Currently we have only adopted the unweighted network in the following network topological analysis. Hence, the final adjacency \( a_{ij} \) is:

\[ a_{ij} = \begin{cases} 
1 & \text{if } s_{ij} \geq s_t \\
0 & \text{if } s_{ij} < s_t 
\end{cases} \]

where \( s_t \) is the final threshold parameter. Two nodes are linked if the similarity between their abundance profiles across all samples is equal to 1.

**Calculation of MEN topological indices and general features**

Once MENs are determined, various network topology indices can be calculated based on the adjacency matrix (Table 1). The overall topological indices describe the overall network topology in different views and thus are useful in characterizing various MENs identified under different situations. The indices for describing individual nodes are useful in assessing their roles in the network.

Scale-free, small world, modularity and hierarchy are most common network characteristics of interest [8,53,93]. A scale-free network is a network whose connectivity follows a power law, at least asymptotically [94], that is, only a few nodes in the network have many connections with other nodes while most of nodes have only a few connections with other nodes. It can be expressed by \( P(k) \sim k^{-\lambda} \), where \( k \) is connectivity and \( \lambda \) is a constant. A small-world network is the network in which most nodes are not neighbors of one another, but most nodes can be reached by a few paths (typically, less than 6). Small world network has a small average shortest path (GD) typically as the logarithm of the number of nodes [43]. In addition, there is no formal definition for hierarchical topology [95]. One of the most important signatures for hierarchical, modular organizations is that the scaling of clustering coefficient follows \( C(k) \sim k^{-\gamma} \), in which \( k \) is connectivity and \( \gamma \) is a constant. By log-transformation, \( \log[C(k)] \sim -\gamma \log(k) \), the logarithms of clustering coefficients have a linear relationship with the logarithms of connectivity.

**Module detection**

Modularity is a fundamental characteristics of biological networks as well as many engineering systems [53]. In MENs, a module in the network is a group of OTUs that are highly connected within the group, but very few connections outside the group. The maximum modularity score is used to separate the graph into multiple dense sub-graphs or modules. The modularity of each network \( (M) \) is estimated using the equation [66]:

\[ M = \sum_{b=1}^{N_M} \frac{l_b}{L} - \left( \frac{K_b}{2L} \right)^2 \]

where \( N_M \) is the number of modules in the network, \( l_b \) is the number of links among all nodes within the \( b^{th} \) module, \( L \) is the number of all links in the network, and \( K_b \) is the sum of degrees (connectivity) of nodes which are in the \( b^{th} \) module. \( M \) measures the extension whose nodes have more links within their own modules than expected if linkage is random. It varies with the range of \([-1, 1]\).

Several different algorithms can be used to separate modules, including short random walks, leading eigenvector of the community matrix, simulated annealing approach, and fast greedy modularity optimization [56,57]. The algorithm of short random walks is based on the idea that all random walks tend to stay in the densely connected parts of a graph that was corresponding to the modules [56]. After calculating a distance
between two nodes or between sets of nodes by random walk algorithm, it uses a hierarchical clustering approach to present the structural similarities between all nodes. Thereafter this approach will choose the best partition automatically. The advantage of this algorithm is efficient and fast computation.

Once the network modularity value (M) was explicitly defined, theoretically the module structure can be determined by maximizing M values over all possible divisions of network. However, exhaustive maximization over all divisions is computational intractable [57]. The algorithm of leading eigenvector is one of several approximate optimization methods have been proven effectively obtained higher M values with high speed. It simplified the maximization process in terms of a modularity matrix B^{n \times n} that can be obtained by the adjacent matrix A^{n \times n} subtracting an expected edges matrix P^{n \times n} from a null model. Then the network can be split into two groups by finding the leading eigenvector that was corresponding to the largest positive eigenvalue of modularity matrix. This splitting process can be looped until any further divisions will not increase the M value [57]. This method shows more accurate separations and other algorithms in several well-studied social networks [57].

The algorithm of simulated annealing approach usually produces the best separation of the modules by direct maximization of M [58]. The simulated annealing is a stochastic optimization technique to find “low cost” configurations [96]. It carries out the exhaustive search on network structures to merge and split priori-modules and move individual nodes from one module to another. Although this is a time-consuming process, it is expected to obtain clear module separations with a higher M.

The algorithm of fast greedy modularity optimization is to isolate modules via directly optimizing the M score [66,97]. It starts with treating each node as the unique member of one module, and then repeatedly combines two modules if they generate the largest increase in modularity M. This algorithm has advantages with fast speed, accurate separations and ability to handle huge networks [66,97].

Identification of key module members
After all modules are separated, each node can be assigned a role based on its topological properties [59], and the role of node i is characterized by its within-module connectivity (\( z_i \)) and among-module connectivity (\( P_i \)) as follows

\[
 z_i = \frac{k_{ib} - k_b}{\sigma_{ki}},
\]

and

\[
P_i = 1 - \sum_{c=1}^{N_M} \left( \frac{k_{ic}}{k_i} \right)^2,
\]

where \( k_{ib} \) is the number of links of node i to other nodes in its module b, \( k_b \) and \( \sigma_{ki} \) are the average and standard deviation of within-module connectivity, respectively over all the nodes in module b, \( k_i \) is the number of links of node i in the whole network, \( k_{ic} \) is the number of links from node i to nodes in module c, and \( N_M \) is the number of modules in the network.

The within-module connectivity, \( z_i \), describes how well node i is connected to other nodes in the same module, and the participation coefficient, \( P_i \), reflects what degree that node i connects to different modules. \( P_i \) is also referred as the among-module connectivity [98]. If all links of node i only belong to its own module, \( P_i = 0 \). If the links of node i are distributed evenly among modules, \( P_i = 1 \). The topological roles of individual nodes can be assigned by their position in the z-parameter space. Originally, Guimera et al. [33,59] divided the topological roles of individual nodes into seven categories. Olesen et al. [98] simplified this classification into four categories for pollination networks. In this study, we use the simplified classification as follows: (i) Peripheral nodes (\( z_i < 2.5, P_i \leq 0.62 \)), which have only a few links and almost always to the nodes within their modules, (ii) Connectors (\( z_i < 2.5, P_i > 0.62 \)), which are highly linked to several modules, (iii) Module hubs (\( z_i > 2.5, P_i \leq 0.62 \)), which are highly connected to many nodes in their own modules, and (iv) Network hubs (\( z_i > 2.5, P_i > 0.62 \)), which act as both module hubs and connectors. From ecological perspective, peripheral nodes represent specialists whereas the other three are generalists.

Eigen-gene analysis
One of the grand challenges in dealing with high throughput metagenomics data is the high dimensionality. Various statistical approaches are used to reduce dimensions and extract major features, including principal component analysis (PCA), detrended correspondence analysis (DCA), and singular value decomposition (SVD). SVD is an orthogonal linear transformation of data (e.g., microbial data) from the complexity to the comprehensibility [99]. Based on SVD analysis, the Eigengene is a linear combination of genes and eigenvalues. In the diagonalized data, each eigengene is just expressed in the corresponding eigen arrays. Langfelder and Horvath [61] proposed eigengene network analysis to summarize the gene expression data from each module as a centroid. Eigengene network analysis is powerful to reveal higher order organization among gene co-
expression modules [33,61,62]. Here, we have adopt this method to analyze modules in MENs.

**SVD analysis to define module eigen-gene**

Suppose there are \( n^b \) OTUs in the \( b \)-th module. Let \( X^b = [x^b_{i,q}] \) represent the relative abundance matrix of the \( b \)-th module, where \( x^b_{i,q} \) is the relative abundance of the \( i \)-th OTU in the \( q \)-the sample \( (i \in \{1, \ldots, n^b\}, q \in \{1, \ldots, m\}) \). In SVD analysis, \( X^b \) can be decomposed as follows:

\[
X^b = U^b D^b (V^b)^T,
\]

where both \( U^b_{(n^b \times n^b)} \) and \( V^b_{(m \times m)} \) are column-orthogonal matrices, and \( D^b_{(m \times m)} \) is a diagonal matrix of the singular values \( \{d^b_{q,j}\} \). The matrices \( V^b \) and \( D^b \) are denoted as

\[
V^b = (v^b_1, v^b_2, \ldots, v^b_m) \quad \text{and} \quad D^b = \text{diag}(d^b_1, d^b_2, \ldots, d^b_m).
\]

Assuming that the singular values are arranged in decreasing order, the first column of \( V^b \) is referred as the Module Eigen-gene, \( E^b \), for the \( b \)-th module. That is, \( E^b = v^b_1 \).

The relative abundance profile of the OTUs within a module is represented by the eigen-gene. In addition, the sum of variance of OTU abundances equals to the sum of the diagonal matrix in SVD. Therefore, the percentage of the variance explained by the eigen-gene is given by \( \Phi^b \) as

\[
\Phi^b = \frac{|d^b_1|^2}{\sum_{j=1}^m |d^b_j|^2}.
\]

Generally, the module eigen-gene can explain approximately 50% or more of the variance of the OTU abundances in the module [61]. Since PCA and SVD are identical if each OTU relative abundance has been standardized to mean 0 and variance 1, \( E^b \) is the first principal component based on PCA analysis [61].

**Module membership**

Module eigen-gene provides the best summary of variation in relative abundance of OTUs within a module, but it is a centroid of a module rather than a real OTU. In practice, it is always important to understand how close it is between a given actual OTU and its eigen-gene. The correlation of the eigen-gene in module \( b \) to the \( i \)-th actual OTU across all experimental samples is defined as

\[
MM^b_{i} = \text{cor}(x_i, E^b)
\]

If \( MM^b_{i} \) is close to 1 or -1, it is evident that the \( i \)-th OTU is close to the centroid of module \( b \).

**Random network construction and network comparison**

Since only a single data point is available for each network parameter, we are not able to perform standard statistical analyses to assess statistical significances. Similar to the concept of hypothesis testing, the null model is generated to assess the performance of the alternative model. Thus, the random networks are generated to compare different complex networks using the Maslov-Sneppen procedure [68]. The Maslov-Sneppen method keeps the numbers of nodes and links unchanged but rewire the positions of all links in the MENs so that the sizes of networks are the same and the random rewired networks are comparable with original ones. This method has been typically used for ecological network analyses [4]. For each network identified, a total of 100 randomly rewired networks are usually generated by the Maslov-Sneppen procedure [68] and all network indices are calculated individually for each randomized network. Then the average and standard deviation for each index of all random networks are obtained. The statistical Z-test is able to test the differences of the indices between the MEN and random networks. Meanwhile, for the comparisons between the network indices under different conditions, the Student t-test can be employed by the standard deviations derived from corresponding random networks.

**Trait-based gene significance measure**

In gene expression network analyses, the gene significance \( (G_{S_{ih}}) \) is the correlation between the expression profile of the \( i \)-th gene and the \( h \)-th sample trait, \( T_h \) [33]. The higher \( G_{S_{ih}} \), the more biologically significant gene \( i \) is related to the sample trait \( h \). Similarly, in this study, the trait-based OTU significance is defined as:

\[
G_{S_{ih}} = [\text{cor}(x_i, T_h)]^2
\]

where \( x_i \) is the relative abundance of the \( i \)-th OTU \( i \in \{1, \ldots, n\} \) and \( T_h \) is the \( h \)-th sample trait (e.g. soil pH, N content, total plant biomass) \( (h \in \{1, \ldots, g\}) \). Since the measurement units for different traits vary, all trait data should be standardized prior to statistical analysis. Finally, an OTU significance matrix, \( G^{\text{avg}}_{S} \), is obtained.

**Relationships of microbial interaction networks with soil variables**

To discern the relationships between molecular ecological networks and soil properties, Mantel tests can be performed [100]. The relationships between the MENs and environmental variables were determined as follows: First, the significances of variables are calculated with the above equation (Eq 13) and the OTU significance matrix is generated. Then the Euclidean distance matrix \( D^{x \times n}_{GS} \) is generated by calculating the Euclidean distance
between every two OTUs. The distance matrix among all OTUs’ connectivity ($D^x_{ij}$) was calculated as well. In addition, Mantel tests are performed between the distance matrices of the connectivity ($D^x_{GS}$) and OTU significance ($D^s_{GS}$) to examine the relationships between network structure (i.e., connectivity) and soil variables. The Mantel tests were performed using the programs available in R vegan package [101].

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work has been supported, through contract DE-SC0004613 and contract DE-AC02-05CH11231 (as part of ENIGMA, a Scientific Focus Area) and contract DE-SC0004601, by the US Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics: GTL Foundation Science, the United States Department of Agriculture (Project 2007-35319-18305) through NSF-USDA Microbial Observatory Programs, the Oklahoma Bioenergy Center (OBC) of State of Oklahoma, and State Key Joint Laboratory of Environment Simulation and Pollution Control (Grant 11203ESPCT) at Tsinghua University.

Author details
1 Institute for Environmental Genomics and Department of Botany and Microbiology, University of Oklahoma, Norman OK 73019, USA. 2 Glomics Inc, Norman OK 73072, USA. 3 State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China. 4 Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley CA 94720, USA. 5 School of Computing, Clemson University, Clemson SC 29634, USA.

Authors’ contributions
YD carried out the analysis, constructed the pipeline and wrote the article. YJ carried out part of method development and method writing. YY provided the RMT method and contributed part of the discussion. FL provided oversight of the work and helped finalize the article. All authors read and approved the final manuscript.

Received: 29 December 2011 Accepted: 30 May 2012
Published: 30 May 2012

References
1. Montoya JM, Pimm SL, Sole RV: Ecological networks and their fragility. Nature 2006, 442(7100):259–264.
2. May RM: Stability and complexity in model ecosystems. Princeton, New Jersey: Princeton University Press; 1973.
3. Pim S: Food webs. London: Chapman & Hall; 1982.
4. Dunne JA, Williams RJ, Martinez ND: Food web structure and network theory: The role of connectance and size. Proc Natl Acad Sci USA 2002, 99(20):12917–12922.
5. Montoya JM, Sole RV: Small world patterns in food webs. J Theor Biol 2002, 214(3):405–412.
6. Rezende EL, Lavabre JE, Guimaraes PR, Jordan P, Bascompte J: Non-random coextinctions in phylogenetically structured mutualistic networks. Nature 2007, 448(7156):925–926.
7. Raes J, Bork P: Molecular eco-systems biology: towards an understanding of community function. Nat Rev Microbiol 2008, 6(6):693–699.
8. Barabasi AL, Oltvai ZN: Network biology: Understanding the cell’s functional organization. Nat Rev Genet 2004, 5(2):101–115.
9. Akutsu T, Miyano S, Kuhara S: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pac Symp Biocomput 1999, r7–r28.
10. Gardner TS, di Bernardo D, Lorenz D, Collins JJ: Inferring genetic networks and identifying compound mode of action via expression profiling. Science 2003, 301(5629):102–105.
11. Liang S, Fuhrman S, Somogyi R: Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pac Symp Biocomput 1998, 3:18–29.
12. Yeung MKS, Tegner J, Collins JJ: Reverse engineering gene networks using singular value decomposition and robust regression. Proc Natl Acad Sci USA 2002, 99(9):6163–6168.
13. Friedman N, Linial M, Nachman I: Pe'er D: Using Bayesian networks to analyze expression data. Journal of Computational Biology 2000, 7(3–4):601–620.
14. Gentzur M, Baudis M, Moch H, Beerenwinkel N: Quantifying cancer progression with conjunctive Bayesian networks. Bioinformatics 2009, 25(21):2809–2815.
15. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS: Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci USA 2000, 97(22):12182–12186.
16. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, et al: Variations in DNA elucidate molecular networks that cause disease. Nature 2008, 452(7186):429–435.
17. Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurnce MF, Zhao W, Qi S, Chen Z, et al: Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proc Natl Acad Sci USA 2006, 103(46):17402–17407.
18. Oldham MC, Horvath S, Geschwind DH: Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci USA 2006, 103(47):17973–17978.
19. Schmitt WA Jr, Raab RM, Stephanopoulos G: Elucidation of gene interaction networks through time-lagged correlation analysis of transpositional data. Genome Res 2004, 14(8):1654–1663.
20. Zhang S, Horvath S: A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Bio 2005, 4:1.
21. Gardner TS, Faith JJ: Reverse-engineering transcription control networks. Phys Life Rev 2005, 2(1):65–88.
22. Luo F, Yang Y, Zong J, Gao H, Khan L, Thompson DK, Zhou J: Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinformatics 2007, 8:299.
23. Luo F, Zhong JK, Yang YS, Scheuermann RH, Zhou J: Application of random matrix theory to biological networks. Phys Lett A 2006, 357(6):420–423.
24. Yang Y, Harris DP, Luo F, Xiong W, Joachimiak M, Wu L, Dehail P, Jacobsen J, Yang Z, Palumbo AV, et al: Snapshot of iron response in Shewanella oneidensis by gene network reconstruction. BMC Genomics 2009, 10(1):131.
25. Handelmann J, Tiedje JM, Alvarez-Cohen L, Ashburner M, Cann IKO, DeLong E, Erlebach T, Dunne JA: The network structure of food webs. In: Theory and practice of ecological network analysis: methodological foundations. Washington: National Academy of Sciences; 2007:1–158.
26. He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, Huang Z, Wu W, Gu B, Jardine P, et al: GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. GBE J 2007, 1(1):67–77.
27. Zhou J, Deng Y, Luo F, He Z, Tu Q, Zhi X: Functional molecular ecological networks. mBio 2010, 1(4):e00169–00110.
28. Zhou J, Deng Y, Luo F, He Z, Yang Y: Phylogenetic molecular ecological network of soil microbial communities in response to elevated CO2. mBio 2011, 2(4):e00122–00111.
29. Bascompte J: Networks in ecology. Basic Appl Ecol 2007, 8(6):485–490.
30. Dunne JA, Williams RJ, Martinez ND, Wood RA, Erwin DH: Compilations and network analyses of cambrian food webs. Plos Biol 2008, 6(4):102.
31. Dunne JA: The network structure of food webs. In: Ecological Networks: Linking Structure to Dynamics in Food Webs. Edited by M P, Dunne JA: Oxford: Oxford University Press; 2006. 227–286.
32. Chaffon H, Rehauern H, Barenmaller V, J von Mering C: A global network of coexisting miomes from environmental and whole-genome sequence data. Genome Res 2010, 20(7):947–959.
33. Guimera R, Sales-Pardo M, Amaral LA: Classes of complex networks defined by role-to-role connectivity profiles. Nat Phys 2007, 3(1):163–169.
34. Brandes U, Erlebach T: Network analysis: methodological foundations. Berlin: Springer; 2005.
35. Bonacich P: Power and Centrality - a Family of Measures. Am J Social 1987, 92(5):1170–1182.
36. Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks. Nature 1998, 393(6684):440–442.

37. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of metabolic networks. Science 2002, 297(5586):1551–1555.

38. Costa LD, Rodrigues FA, Travieso G, Boas PPR. Characterization of complex networks: A survey of measurements. Adv Phys 2007, 56(1):167–242.

39. West DB. Introduction to Graph Theory. Upper Saddle River, NJ: Prentice Hall; 1996.

40. Latora V, Marchioni M. Efficient behavior of small-world networks. Phys Rev Lett 2001, 87(19):198701.

41. Wasserman S, Faust K. Social Network Analysis: Methods and applications. Cambridge: Cambridge University Press; 1994.

42. Krackhardt D. Graph Theoretical Dimensions of Informal Organizations. Hillsdale, NJ: Lawrence Erlbaum and Associates; 1994.

43. Amaral LA, Scala A, Barthelmy M, Stanley HE. Classes of small-world networks. Proc Natl Acad Sci USA 2000, 97(21):11149–11152.

44. Girvan M, Newman ME. Community structure in social and biological networks. Proc Natl Acad Sci USA 2002, 99(12):7821–7826.

45. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci USA 2006, 103(23):8577–8582.

46. Luo F, Zhong J, Yang Z, Zhou J. Application of random matrix theory to miRNA data for discovering functional gene modules. Phys Rev E 2005, 73(3 Pt 1):031924.

47. Ravasz E, Barabasi AL. Hierarchical organization in complex networks. Phys Rev E 2003, 67(2):026112.

48. Zhou J, Zhong J, Yang Z, Zhou J. Reproducibility and quantitativeness of amplicon sequence-based detection. J Mol Biol 2011, 313:1–1313.

49. He Z, Xu M, Deng Y, Kang S, Zhou J. Metagenomic analysis reveals a marked divergence in the structure of the belowground microbial communities at elevated CO2. Ecol Lett 2010, 13(5):564–575.

50. Pons P, Latapy M. Computing communities in large networks using random walks. Computer and Information Sciences - Iscis 2005, Proceedings. 2005, 3733:284–293.

51. Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 2006, 74(3):036104.

52. Guimerà R, Amaral LAN. Characterization of complex networks: modules and universal roles. J Stat Mech-Theory Exp 2005, 2005:02034.

53. Guimerà R, Amaral LAN. Functional cartography of complex metabolic networks. Nature 2005, 433(7028):859–900.

54. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Nelson CE, Schadt CW, Garten CT Jr, et al. A human gut microbial catalog established by metagenomic sequencing. Nature 2010, 464(7285):59–65.

55. Zhou J, Xiao MC, Wong WH. Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci USA 2002, 99(20):12783–12788.

56. Epstein SS. Miobial awakenings. Nature 2009, 457(7233):1083.

57. May RM. Stability and complexity in model ecosystems. Princeton, NJ: Princeton University Press; 2001.

58. Shubin H. Biological robustness. Nat Rev Genet 2004, 5(11):826–837.

59. Theilbaut E, Fontaine C. Stability of ecological communities and the architecture of mutualistic and trophic networks. Science 2010, 329(5993):853–856.

60. Wang F, Zhou H, Meng J, Peng X, Jiang L, Sun P, Zhang C, Van Nostrand JD, Deng Y, He Z, et al. GeoChip-based analysis of metabolic diversity of miobial communities at the Juan de Fuca Ridge hydrothermal vent. Proc Natl Acad Sci USA 2009, 106(46):19480–19485.

61. Zhou J, Kang S, Schadt CW, Garten CT Jr. Spatial scaling of functional gene diversity across various microbial taxa. Proc Natl Acad Sci USA 2008, 105(22):7768–7773.

62. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 2000, 5:418–429.

63. Carter DA. Comprehensive strategies to study neuronal function in transgenic animal models. Biol Psychiatry 2004, 55(8):785–788.

64. Carlson MR, Zhang B, Fang Z, Mischel PS, Horvath S, Nelson SF. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics 2006, 7:40.

65. Wagner EP, Random Matrices in Physics. San Francisco: World Scientific; 2000.

66. Mehta ML. Random Matrices, 2nd edition, Academic Press 1990.

67. Clime MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christians R, Avila-Campillo J, Ieich M, Gross B, et al. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2007, 2(10):2366–2382.

68. Maslov S, Sneppen K. Specificity and stability in topology of protein networks. Science 2002, 296(5569):910–913.

69. Bascompte J, Jordano P, Melian CJ, Olesen JM. The nested assembly of plant-plant mutualistic networks. Proc Natl Acad Sci USA 2003, 100(16):9383–9387.

70. Bastolla U, Fortuna MA, Pascual-Garcia A, Ferrera A, Luque B, Bascompte J. The architecture of mutualistic networks minimizes competition and ineanes biodiversity. Nature 2009, 458(7241):1018–1020.

71. Bascompte J, Jordano P. Plant-animal mutualistic networks: The architecture of biodiversity. Annu Rev Ecol Evol Syst 2007, 38:567–593.

72. Horvath S, Zhang B, Mayr C, Grootseker T, Schindelbeck T, Rahman MM, Mischel PS, Horvath S, Nelson SF, et al. Hierarchical clustering using pairwise entropy measurements. PLoS Comput Biol 2007, 3(4):e17000.
97. Newman ME: Fast algorithm for detecting community structure in networks. Phys Rev E 2004, 69(6 Pt 2):066133.
98. Olesen JM, Bascompte J, Dupont YL, Jordano P: The smallest of all worlds: pollination networks. J Theor Biol 2006, 240(2):270–276.
99. Alter O, Brown PO, Botstein D: Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 2000, 97(18):10101–10106.
100. Mantel N: Detection of Disease Clustering and a Generalized Regression Approach. Cancer Research 1967, 27(2):209.
101. Dixon P: VEGAN, a package of R functions for community ecology. J Veg Sci 2003, 14(6):927–930.

doi:10.1186/1471-2105-13-113
Cite this article as: Deng et al.: Molecular ecological network analyses. BMC Bioinformatics 2012 13:113.