Genomics Proteomics Bioinformatics 14 (2016) 314–324

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Construction and Analysis of Functional Networks in the Gut Microbiome of Type 2 Diabetes Patients

Lianshuo Li [1,a], Zicheng Wang [1,b], Peng He [1,c], Shining Ma [1,d], Jie Du [2,e], Rui Jiang [1,*,f]

[1] *MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China*
[2] *Beijing Anzhen Hospital, Capital Medical University, Beijing Institute of Heart, Lung, and Blood Vessel Diseases, Beijing 100029, China*

Received 5 February 2016; accepted 24 February 2016
Available online 14 October 2016

Handled by Andreas Keller

KEYWORDS

Functional network;
Microbiome;
Type 2 diabetes;
Metagenomics;
Network motif

Abstract Although networks of microbial species have been widely used in the analysis of 16S rRNA sequencing data of a **microbiome**, the construction and analysis of a complete microbial gene network are in general problematic because of the large number of microbial genes in **metagenomics** studies. To overcome this limitation, we propose to map microbial genes to functional units, including KEGG orthologous groups and the evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) orthologous groups, to enable the construction and analysis of a microbial **functional network**. We devised two statistical methods to infer pairwise relationships between microbial functional units based on a deep sequencing dataset of gut **microbiome** from **type 2 diabetes** (T2D) patients as well as healthy controls. Networks containing such functional units and their significant interactions were constructed subsequently. We conducted a variety of analyses of global properties, local properties, and functional modules in the resulting functional networks. Our data indicate that besides the observations consistent with the current knowledge, this study provides novel biological insights into the gut **microbiome** associated with T2D.

Introduction

Advancement of the next-generation sequencing technology has made it possible to sequence all genetic materials of a microbiome and to assemble millions of microbial genes from the data, resulting in the recent explosion of large-scale metagenomic studies in soil [1–3], air [4,5], marine environments [6–8], and humans [9–11], as well as in many other fields [12]. As a simple and low-cost approach, sequencing of the 16S

[*] Corresponding author.
E-mail: ruijiang@tsinghua.edu.cn (Jiang R).
[a] ORCID: 0000-0003-4548-6895.
[b] ORCID: 0000-0002-2625-3781.
[c] ORCID: 0000-0002-7035-8281.
[d] ORCID: 0000-0003-4277-1308.
[e] ORCID: 0000-0003-0447-736X.
[f] ORCID: 0000-0002-7533-3753.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2016.02.005

rRNA gene in whole or its hypervariable regions selectively can be employed to profile the taxonomic composition of a microbiome. This approach, together with the powerful network analysis methodology [13], has a variety of applications, such as identification of co-occurrence networks of microbial species in soil, marine environments [14], and, more recently, of humans [15–17]. The recently-proposed Boolean implication networks involving marine microbial species and environmental factors have also been reported [13], further enhancing the explanatory power of co-occurrence networks. Nevertheless, 16S rRNA sequencing can hardly be used to assess functions of a microbiome. In order to study functions of a microbial community, it is necessary to know which genes are in a community and how these genes interact with one another to support a complicated biological function.

Analogous to the available studies involving 16S rRNA analysis, construction of a co-occurrence network of all microbial genes is desirable. Nevertheless, a microbial community typically includes millions of microbial genes from tens of thousands of species. The construction of a network from such a large number of genes is nearly impossible in most metagenomic studies. On the other hand, many microbial genes are similar in sequence, suggesting that they may actually perform similar functions in a community. Therefore, it is possible to first define a set of functional units that can be mapped to known gene functions and then to study the relations between these functional units in a microbiome. Since the mapping of microbial genes to functional units is many-to-one, the number of functional units is expected to be much smaller than that of microbial genes. Moreover, interactions between such functional units can also reflect the interaction between the microbial genes, which helps to address the question how microbial genes work together to support particular functions of a microbiome.

We thus propose in this paper a framework for constructing a network of functional units in a microbiome. We define a node in such a microbial functional network as a KEGG orthologous group (KO) [18] or an evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) orthologous group (OG) [19], and introduce an edge connecting a pair of nodes if the nodes show a certain correlation according to their abundance levels. Specifically, using a deep sequencing dataset of the gut microbiome from type 2 diabetes (T2D) patients [10] as a demonstration, we first map raw reads to known microbial genes and obtain abundance of a functional unit as the summation of abundance of all genes belonging to the unit. Then, we design two statistical methods to infer pairwise relationships between microbial species. Using Pearson's correlation coefficient (PCC), we characterize the strength of an association between two functional units as the correlation of the relative abundance of the units across a number of individuals. Using a machine learning method called random forest (RF), we predict the abundance of a functional unit based on that of all the other units. Then the relative importance of certain units in this learning procedure is employed to measure the strength of an association between the response unit and the predictor unit. After filtering out weak associations between nodes, we obtain a network of functional units. Finally, we analyze global properties, local properties, and functional modules in the constructed functional networks.

Results

Construction of functional networks in the human gut microbiome

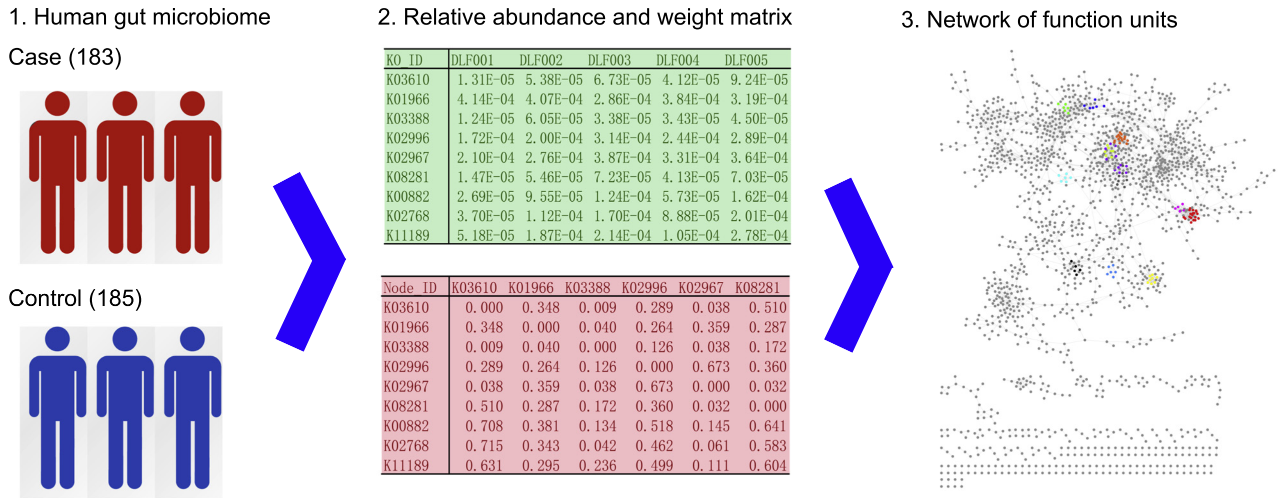
The workflow of our method is illustrated in **Figure 1**. Taking raw sequencing reads of the human gut microbiome of 183 T2D patients and 185 healthy individuals [10] as input, we adopted two statistical methods, *i.e.*, PCC and RF, to construct networks of functional units with either KO or OG, as output. To accomplish this objective in a flexible way, we divided the workflow into three independent modules: (1) an input module for calculating relative abundance levels of the units, (2) a network construction module for inferring pairwise relationships between functional units, and (3) an output module for saving and visualizing the constructed networks.

In the input module, we first mapped raw sequencing reads to known microbial genes (or contigs) and obtained raw read counts of the genes. We then normalized the read count of a particular gene to its length to obtain the abundance level of the gene. Using the total abundance normalization method, we divided the abundance of a particular gene by the summation of abundance levels of all genes to determine relative abundance of the gene. We finally mapped microbial genes to functional units and determined the relative abundance of a unit as the summation of abundance levels of all the genes belonging to the unit. In the network construction module, we designed two statistical methods to infer pairwise relationships between microbial species. We first quantified relationships between two functional units by means of PCC between the two vectors corresponding to the relative abundance levels of these units across a number of individuals. We then employed RF to predict the abundance of a functional unit using those of all the remaining units. Afterward, we used the relative importance of certain units in this learning procedure to measure the strength of an association between the response unit and the predictor unit. Finally, we enhanced the output module with basic statistics of the constructed network in addition to a plain text file that describes connections between functional units. Details of these methods are given in the Methods section.

Topological properties of functional networks

We conducted a comparative analysis of the constructed functional networks of the human gut microbiome. Using KO and OG as functional units, we obtained four functional networks: KO network based on PCC, KO network based on RF, OG network based on PCC, and OG network based on RF.

As shown in **Figure 2** and **Table 1**, we found that the KO and OG networks based on PCC share similar topological structure (Figure 2A and 2C), and so do the two networks based on RF (Figure 2B and 2D), while networks constructed using different methods vary significantly. At the default cutoff value of 1.5 for filtering weak associations between nodes in a network (see Methods), both KO networks contain 3880 edges. Nevertheless, the PCC-based KO network has only 724 nodes (Figure 2A), whereas RF-based KO network has 1687 nodes (Figure 2B), indicating that edges are distributed to nodes more uniformly when using RF in comparison with PCC. For the same reason, the PCC-based KO network has


1. Human gut microbiome
Case (183)
Control (185)
2. Relative abundance and weight matrix
3. Network of function units

KO_ID	DLF001	DLF002	DLF003	DLF004	DLF005
K03610	1.31E-05	5.38E-05	6.73E-05	4.12E-05	9.24E-05
K01966	4.14E-04	4.07E-04	2.86E-04	3.84E-04	3.19E-04
K03388	1.24E-05	6.05E-05	3.38E-05	3.43E-05	4.50E-05
K02996	1.72E-04	2.00E-04	3.14E-04	2.44E-04	2.89E-04
K02967	2.10E-04	2.76E-04	3.87E-04	3.31E-04	3.64E-04
K08281	1.47E-05	5.46E-05	7.23E-05	4.13E-05	7.03E-05
K00882	2.69E-05	9.55E-05	1.24E-04	5.73E-05	1.62E-04
K02768	3.70E-05	1.12E-04	1.70E-04	8.88E-05	2.01E-04
K11189	5.18E-05	1.87E-04	2.14E-04	1.05E-04	2.78E-04

Node_ID	K03610	K01966	K03388	K02996	K02967	K08281
K03610	0.000	0.348	0.009	0.289	0.038	0.510
K01966	0.348	0.000	0.040	0.264	0.359	0.287
K03388	0.009	0.040	0.000	0.126	0.038	0.172
K02996	0.289	0.264	0.126	0.000	0.673	0.360
K02967	0.038	0.359	0.038	0.673	0.000	0.032
K08281	0.510	0.287	0.172	0.360	0.032	0.000
K00882	0.708	0.381	0.134	0.518	0.145	0.641
K02768	0.715	0.343	0.042	0.462	0.061	0.583
K11189	0.631	0.295	0.236	0.499	0.111	0.604

Figure 1 Workflow of the proposed study

Taking raw sequencing reads of the human gut microbiome of 183 T2D patients and 185 healthy individuals as input, we adopted two statistical methods, *i.e.*, PCC and RF, to construct networks of functional units with either KO or OG, as output. To accomplish this objective in a flexible way, we divide the workflow into three independent modules: an input module for calculating relative abundance levels of the units, a network construction module for inferring pairwise relationships between functional units, and an output module for saving and visualizing the constructed networks. PCC, Pearson's correlation coefficient; RF, random forest; KO, KEGG orthologous group; OG, evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) orthologous group.

higher density and a smaller diameter than does RF-based KO network (Figure 2A and 2B). Therefore, the PCC-based KO network shows higher modularity, *i.e.*, nodes in the network tend to form denser modules; this property is also reflected by higher centralization and a greater clustering coefficient. The scale-free property is an important feature of complex networks, which requires that the degree distribution of a network fit a power law model [20]. The R^2 fitting of the two KO networks is 0.830 and 0.925, respectively. Thus both networks have good scale-free properties, although the RF-based KO network performs better.

The topological properties of the OG networks (Figure 2C and 2D) are very similar to those of the KO networks based on the same method.

By comparing the performance of these two network construction methods, it is found that the PCC method tends to build a network with fewer nodes but more modules, whereas the RF method tends to build a sparser network with more nodes, a greater diameter, and good scale-free property. The two methods exhibit different characteristics and advantages, and provide useful information from different aspects. In this study, we analyze the networks based on both methods and focus on their common features.

Highly-connected modules in functional networks

Next, to detect dense modules in the KO networks and OG networks and to explore the biological patterns in them, we identified highly-connected modules in the functional networks. A highly-connected module in a network is defined as a set of nodes where most nodes are connected with one another, *i.e.*, cliques or near-cliques. Nodes in such a module tend to be functionally similar or to interact closely thus forming a unit of a certain biological function. We applied the tool MCODE [21], a plugin for Cytoscape [22], to analyze the networks. As a result, we uncovered 12 modules present in both KO networks (Table S1) and 10 modules present in both OG networks.

Figure 3 shows four highly-connected modules present in both KO networks constructed by the two different methods. The largest module in KO networks (Module 1) is depicted in Figure 3A, which contains 19 nodes connected by 163 edges. Among them, 18 nodes belong to a KEGG pathway called "flagella assembly" as shown in Figure 3E. Nodes in this module represent KOs corresponding to different flagella biosynthetic proteins, and the interactions among them in the environment of the human intestines are quite strong. Consequently, this module can be detected by using both network construction methods. Figure 3B depicts the module ranked in the second place (Module 2), which contains 15 nodes and 102 edges. All KOs in this module represent proteins related to sporulation. The smaller module in Figure 3C (Module 5) is a 6-node clique, in which each KO corresponds to a protein as a subunit of the enzyme H^+-transporting two-sector ATPase. Figure 3D shows a 6-node near-clique (Module 7) located outside the large connected components of the KO networks, and KOs in this module represent different subunits of another enzyme, NADH: ubiquinone reductase (H^+-translocating).

Highly-connected modules detected in OG networks are similar to those in KO networks to some extent (data not shown). For example, the largest module present in both OG networks comprises 18 nodes and 134 edges. It is of note that most of the nodes correspond to functions associated with flagella, just as in KO networks.

These highly-connected modules in KO and OG networks and their biological annotations reveal that the edges in the KO and OG networks can represent interactions between proteins corresponding to the nodes, and modules can represent

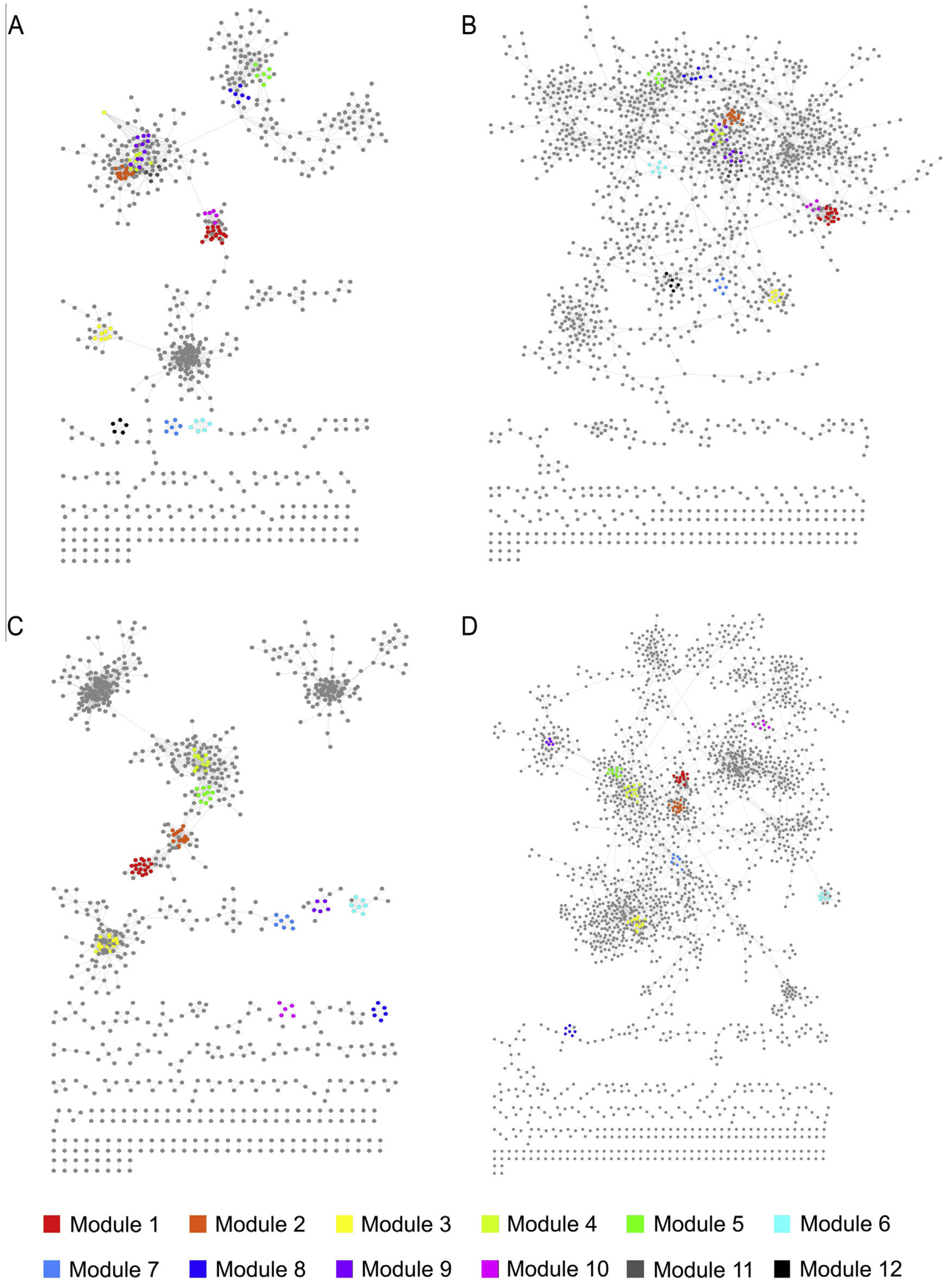


Figure 2 Functional networks in human gut microbiome

A. PCC-based KO network. **B.** RF-based KO network. Modules were detected by using MCODE in cytoscape. Each color represents a module with the color code as follows: Module 1 (red), Module 2 (orange), Module 3 (yellow), Module 4 (light green), Module 5 (green), Module 6 (cyan), Module 7 (blue), Module 8 (dark blue), Module 9 (purple), Module 10 (pink), Module 11 (dark gray), and Module 12 (black). Common modules shared in both KO networks are indicated with same color. **C.** PCC-based OG network. **D.** RF-based OG network. Modules were detected by using MCODE in cytoscape. Each color represents a module with the color code as follows: Module 1 (red), Module 2 (orange), Module 3 (yellow), Module 4 (light green), Module 5 (green), Module 6 (cyan), Module 7 (blue), Module 8 (dark blue), Module 9 (purple), and Module 10 (pink). Common modules shared in both RF networks are indicated with same color.

Table 1 Topological properties of functional networks

Network property	KO network		OG network	
	PCC-based	RF-based	PCC-based	RF-based
Nodes	724	1687	906	2118
Edges	3880	3880	4981	4981
Density	0.015	0.003	0.012	0.002
Centralization	0.103	0.016	0.075	0.013
Diameter	20	33	15	27
Clustering coefficient	0.518	0.296	0.507	0.276
Degree distribution fitting slope	−1.147	−1.813	−1.240	−1.922
Degree distribution fitting R^2	0.830	0.925	0.793	0.882

Note: KO, KEGG orthologous group; OG, eggNOG orthologous group; PCC, Pearson's correlation coefficient; RF, random forest.

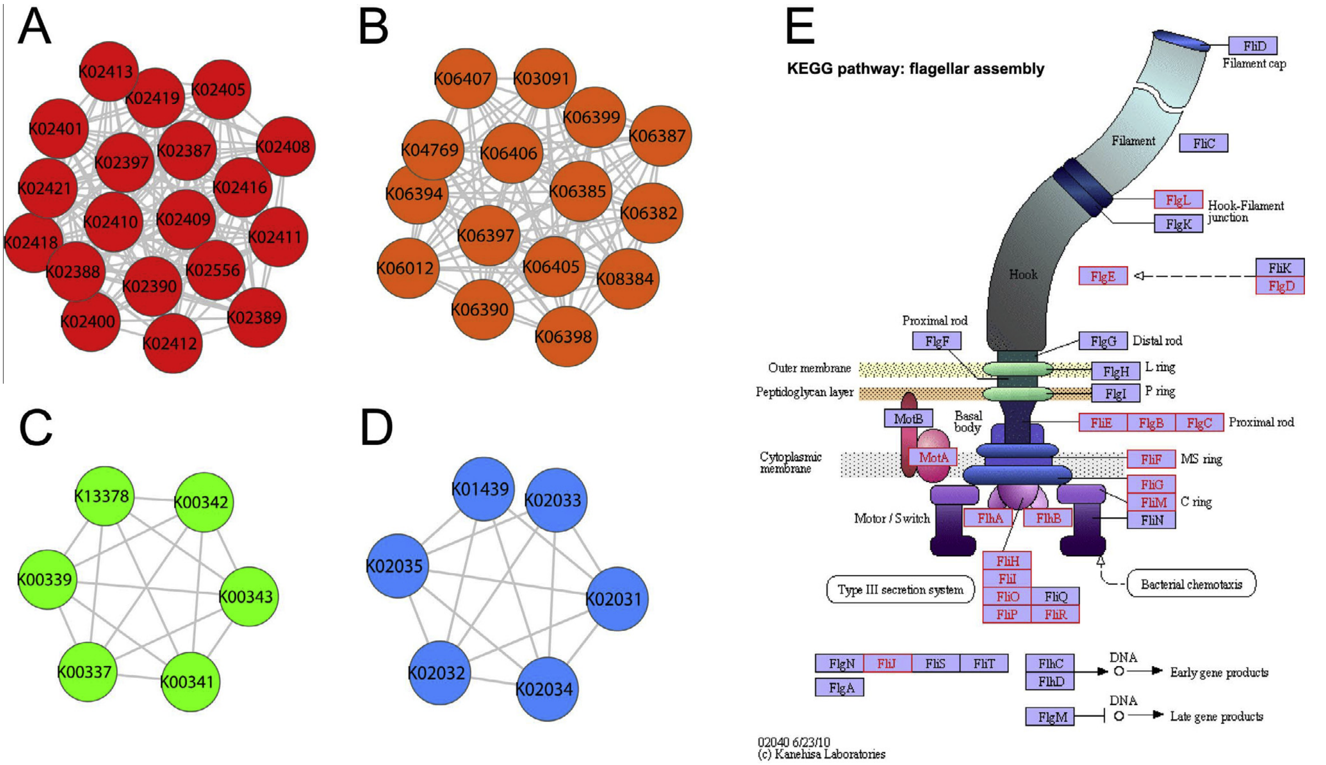


Figure 3 Highly-connected clusters in functional KO networks

The two largest highly-connected clusters present in both KO networks (Module 1 and Module 2) are shown in red (**A**) and orange (**B**), respectively, whereas two smaller highly-connected clusters present in both KO networks (Module 5 and Module 7) are shown in green (**C**) and blue (**D**), respectively. Nodes of Module 1 are enriched in the KEGG pathway "Flagella assembly" (**E**).

protein complexes with certain functions. On the other hand, these biological explanations support that the KO and OG networks are meaningful and can reflect actual interactions occurring in the human gut microbiome.

T2D-associated markers in functional networks

Disease-associated biomarkers are of great value for the diagnosis of various human diseases and for the understanding of their pathogenesis. Via a metagenome-wide association study of the human gut microbiome, we retrieved 1345 KO and 5612 OG markers associated with T2D. Among them, 876 KO markers are T2D-enriched markers, which show significantly higher abundance in T2D samples than in control samples (one-sided Wilcoxon rank sum test, $P < 10^{-4}$). On the other hand, 469 KO markers are T2D-depleted markers, which are enriched in control samples (one-sided Wilcoxon rank sum test, $P < 10^{-4}$). The numbers of T2D-enriched OG markers and T2D-depleted OG markers are 3841 and 1771, respectively.

We mapped these KO markers and OG markers to the KO networks and OG networks, respectively to determine how these markers interact with one another (Figure S1) and to illustrate the marker distribution in the two KO networks (Figure 4). As for the node sets of T2D-enriched and T2D-depleted markers in each network, we found that they are significantly densely connected but not randomly distributed ($P < 10^{-4}$ for both permutation tests, see Methods). Interestingly, some subsets of particularly highly-connected markers contain the modules described in the previous section. For example, a subset of T2D-depleted KO markers (blue nodes in the red circle in Figure S1) contains Module 1, the largest module that we identified above, which is part of microbial flagellar assembly. Similar results were obtained for the distribution of the T2D-associated OG markers in the OG networks (data not shown).

To explore the relationships between the aforementioned highly-connected markers and network modules, we tested enrichment of these markers in the modules. For the 12 modules commonly found in both KO networks, we counted the number of enriched and depleted markers in the networks based on PCC and RF, and calculated their expected values. Our results showed that actual counts in these two networks were equal and the expected counts were very close too. By Pearson's χ^2 test, we found that T2D-enriched markers were statistically significantly enriched in Modules 6 ($P < 10^{-4}$) and 7 ($P < 10^{-4}$). T2D-depleted markers were significantly enriched in Modules 1 ($P < 10^{-8}$), 5 ($P < 0.003$), 10 ($P < 10^{-5}$), and 12 ($P < 10^{-5}$). Modules 6, 7, and 12 were found to be parts of the microbial transport system. T2D-enriched and T2D-depleted markers that were enriched in the first two and last one module, respectively, might show an association with unbalanced substance transport between microbes and the human intestines. Modules 1, 5, and 10 are parts of the flagellar assembly, chemotaxis proteins [23], and NADH-quinone oxidoreductase [24], respectively. We conjecture that T2D-depleted markers enriched in these three modules may indicate a relationship between diseases and microbial movement or activity.

In further topological analysis, we found that T2D markers that were not present in modules tended to have low degrees. For example, for the PCC-based KO network, we identified two groups of nodes corresponding to T2D-enriched and -depleted markers, respectively, that were not found in the 12 aforementioned modules. For these two groups, we tested whether their average degrees were low by permuting node labels 10,000 times. As a result, the P values for T2D-depleted markers and T2D-enriched markers were 0.0006 and 0.0145, respectively, indicating that our finding was statistically significant. Moreover, node stress centrality, which was evaluated by the number of shortest paths passing through a node, was also significantly low ($P < 10^{-4}$) for T2D-enriched markers but significantly high ($P = 0.0227$) for T2D-depleted markers. These data partially indicate that the

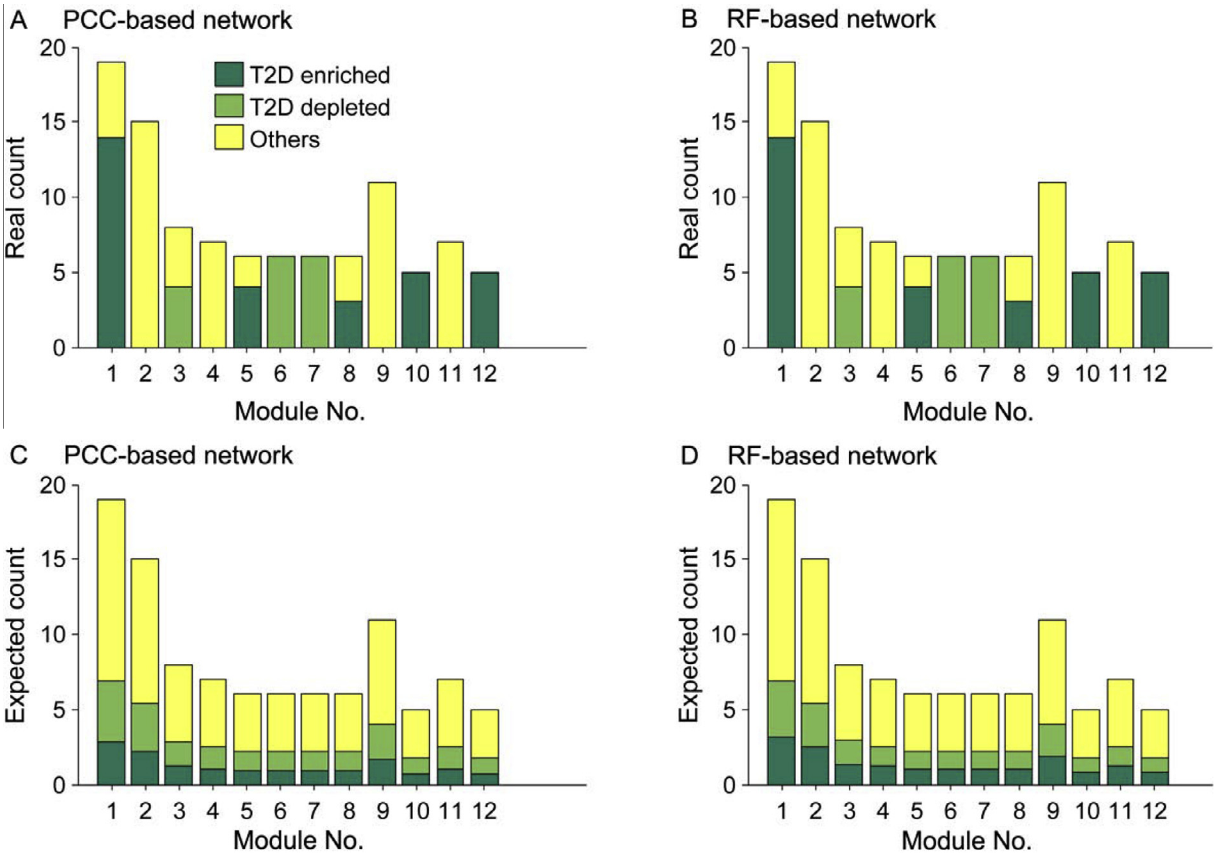


Figure 4 T2D markers in the common modules of functional networks
Real counts of modules enriched (dark green) and depleted (light green) in T2D markers, together with other functional units (yellow) in the 12 modules in both PCC (**A**) and RF (**B**) networks are shown in the upper row, whereas the corresponding expected counts are presented in panels (**C**) and (**D**), respectively, in the bottom row.

functional T2D markers that were not in modules tended to appear on the network border, which is consistent with previous studies [25]. The aforementioned procedure was also repeated for the RF-based KO network and similar results were obtained (data now shown).

A classification method based on these network markers can be implemented by treating the modules enriched with T2D markers as features. To achieve this goal, we first extracted sample data corresponding to the aforementioned modules. Then, average abundance levels of KOs in each module were calculated as features. For the 183 patient and 185 control samples, we created a classification based on logistic regression with 10-fold cross-validation. Receiver operating characteristic (ROC) and precision–recall (PR) curves are shown in Figure S2. The area under ROC curve (AUC) and area under PR curve (AUPR) are 0.695 and 0.674, respectively, suggesting that the network modules partially reflected the difference between patient and control samples.

T2D-specific functional networks

As the annotated 6313 KOs of the gut microbiome came from two groups of samples, we further divided it into two separate profiles, one for 183 T2D patients and the other for the 185 healthy individuals, thereby introducing a T2D case–control comparison. Networks were constructed similarly as described above for the general networks. We first removed the uncommon KOs that had zero abundance in some samples and got 2668 and 2870 KOs in case and control profiles of T2D, respectively. Next, we constructed four networks for these two profiles separately by means of PCC and RF.

General topologies of these four networks are shown in **Table 2**. Obvious topological difference was noticed when using different network construction methods (Figure S3). On the other hand, there is little difference between the networks from case and control samples, indicating that T2D did not affect microbial functional network topology. For further analysis, we compared local structures of these four networks. Similarly, local structures were very similar between case and control networks but different when comparing networks constructed using different methods. Taken all together, we did not observe a significant difference in both general and local structures between the case and control networks.

We found 8 modules shared by both PCC-based and RF-based case networks, and 7 modules shared in both PCC-based and RF-based control networks (Table S2). We then evaluated the abundance difference by the log ratio of the abundance in case network to that in control network. For the case networks, median abundance difference was significantly different from zero in Modules 1, 2, 3, and 5 (all $P < 0.05$). For the control networks, the modules with highly differential abundance were 1, 2, 4, and 6 (all $P < 0.05$). Interestingly, although there is only one T2D marker in Module 6 of the control networks, Module 6 is the only module that shows a significantly high abundance in T2D patients. Thus, we introduced a classifier that uses all the 7 modules in the control networks as features and obtained an AUC value of 0.87, which is higher than that reported previously [10]. The ROC and PR curves of the 10-fold cross-validation experiment are shown in Figure S4.

Network motifs in functional networks

A network motif is a small subgraph of a network that can reflect the local structure of the network. We further analyzed the four networks that were constructed above for motifs. For each network, we counted the numbers of all 6 types of tetrads and compared them to those of random networks by means of mfinder1.2 [26]. Relative significance of the six comparisons was calculated (**Figure 5**) in order to reduce the influence of the network size. We found that all the 4 networks show a similar trend of motif significance. Networks constructed using the same method had very similar relative motif significance, whereas a large difference in relative motif significance could be detected when comparing networks constructed using different methods. Therefore, difference in relative motif significance between networks is sensitive to methods used for network construction but not to data annotation. This observation indicates that microbial functional networks in the human gut may have the same local structures, while biases could be introduced when employing different methods for network construction. An obvious bias was that PCC method tends to introduce triangles into the network. Accordingly, compared to RF-based networks, we observed fewer Subgraphs 1–5, especially Subgraph 4, in PCC-based networks. Further analysis indicated that local structures of the RF-based networks (data now shown) were more similar to those of the protein structure network constructed based on physical distance [26].

Table 2 Topological properties of T2D-specific functional KO networks

Network property	PCC-based KO network		RF-based KO network		Intersection	
	Case	Control	Case	Control	Case	Control
Nodes	648	629	1788	1856	606	606
Edges	4001	4304	4001	4304	999	1247
Density	0.019	0.022	0.003	0.003	0.005	0.007
Diameter	18	12	29	24	20	10
Centralization	0.127	0.125	0.016	0.020	0.031	0.040
Clustering coefficient	0.486	0.497	0.235	0.236	0.302	0.310
Characteristic path length	3.100	3.587	4.475	4.638	4.787	4.116
Degree distribution fitting slope	−0.987	−0.968	−1.910	−1.861	−1.575	−1.536

Note: KO, KEGG orthologous group; PCC, Pearson's correlation coefficient; RF, random forest.

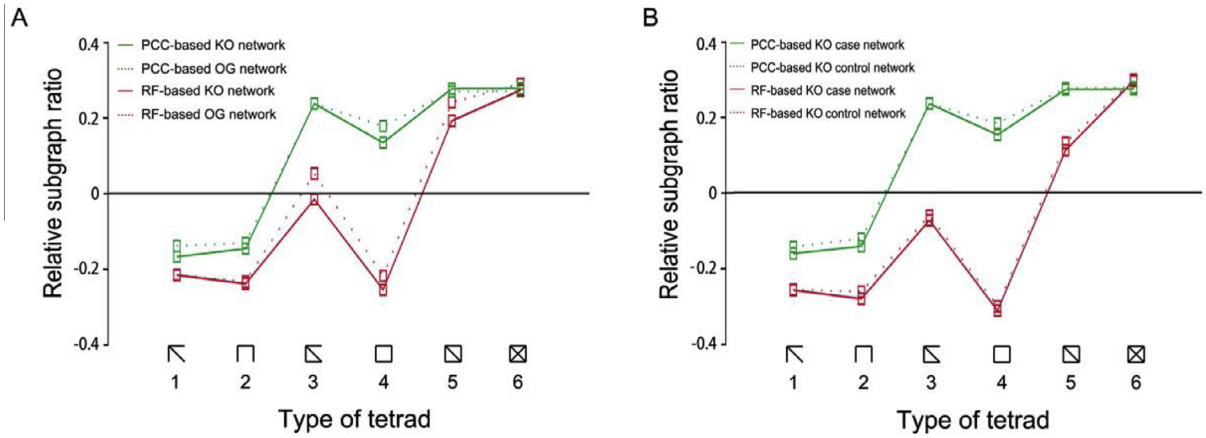


Figure 5 Network motifs in the functional networks

A. The subgraph ratio profile for the functional networks constructed based on PCC (red) and RF (green). Solid line corresponds to KO networks and dash line corresponds to OG networks. **B.** The subgraph ratio profile for the functional networks constructed based on PCC (red) and RF (green) for T2D case and control KO networks. Solid line corresponds to KO case networks and dash line corresponds to KO control networks. For convenience of description, the tetrad types are numbered 1–6 from left to right.

Conclusions and discussion

In this paper, we propose a framework for constructing and analyzing functional networks of the human gut microbiome. We apply this framework to analysis of a large-scale metagenome sequencing dataset of T2D. We find that the networks constructed using different methods and from different samples may capture different aspects of biological meanings. Our results indicate that PCC method preferentially yields dense modules and triangular local structures when constructing a network, whereas the RF method generates a more scale-free network and detects more protein complexes, suggesting that combinatorial application of both methods would be more efficient.

Our framework can be further extended from the following aspects. First, we normalized the abundance of functional units to the total level of abundance in this study. Although such normalization is widely used in metagenomics studies, it is certainly not the only choice. Methods currently used in the analysis of microarray data, for instance, quantile normalization, should be considered in our future work. Second, we adopted two statistical methods to infer relationships between functional units based on their abundance levels. Besides PCC, there are certainly a lot of alternatives such as Pearson's correlation combined with mutual information, which is capable of assessing pairwise relationships between two variables. In addition, besides RF, there are also quite a few state-of-the-art machine learning methods ready for use. How to incorporate these methods into our framework is a primary question in our upcoming studies. Finally, although our method is currently focused on the construction and analysis of functional networks in a microbiome, it is also feasible to apply our framework to analysis of 16S rRNA sequencing data, metagenome sequencing data, and metatranscriptome sequencing data. A software package that integrates these extensions into a single program is under development, aiming to facilitate the research in the field of metagenomics.

Methods

Metagenome-wide profiling of functional units in the human gut microbiome

We collected high-quality paired-end reads of 368 samples and a human gut microbial gene catalog of 4,267,985 genes annotated with 6313 KOs and 45,684 OGs. KO profiles and OG profiles were obtained using a procedure similar to that described previously [10]. We first mapped these high-quality reads of each sample to the gene catalog by means of a short oligonucleotide alignment tool SOAP2 [27] (options: -m 300 -x 400 -v 2 -n 5). The number of the reads mapped to each gene was counted to calculate the abundance of each gene. We applied the following criterion to implement an alignment as a read count of a gene: both ends of a paired-end read were aligned to the same gene with the proper insert size, or one end of a read was mapped to a gene while the other end of this read was not mapped to the genic region. Let x_i denote the number of reads aligned to gene i in a sample, then the relative abundance of gene i, denoted as a_i, is calculated as

$$a_i = \frac{x_i/l_i}{\sum_{j=1}^{g} x_j/l_j} \tag{1}$$

where l_i is the length of gene i, and g represents the total number of genes. The relative abundance was normalized to the read counts of each gene to eliminate the bias caused by different gene lengths and different sequencing depths across these samples.

With the KEGG and eggNOG annotation of the gene catalog, we calculated the abundance of each KO (or OG) in each sample by summing up the relative abundance levels of all genes annotated with this KO (or OG). Let b_i denote the abundance of the ith KO (or OG) in a sample, and the abundance is calculated as

$$b_i = \sum_{j \in S_i} a_j \tag{2}$$

where S_i is the set of genes annotated by the ith KO (or OG), and a_j represents the relative abundance of gene j. In this way, we calculated the abundance of 6313 KOs and 45,684 OGs in each of the 368 samples. Finally, we obtained a KO profile and an OG profile by reserving the 2586 KOs and 3320 OGs with nonzero abundance across all samples, which represent the abundance distribution of the common KOs and OGs in the human gut metagenomes. The KO and OG networks can thus be constructed accordingly.

Construction functional networks for the human gut microbiome

Suppose a profile of the relative abundance contains N nodes, the key step for construction of a network is to create an $N \times N$ weight matrix, in which the elements represent the strength of the pairwise association of these N nodes. One method for creating the weight matrix is simply to calculate the pairwise correlation coefficients between the abundance distributions of each of the two nodes across all samples. Here, we used PCC between two nodes to weight their association. Let X_i and Y_i denote the abundance of node X and Y in the ith sample; then, PCC is calculated as

$$r = \frac{\sum_{i=1}^{n}(Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(Xi - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Yi - \bar{Y})^2}} \tag{3}$$

where n is the number of the samples. Pairwise PCCs of all N nodes form the $N \times N$ weight matrix. We next assess the statistical significance of the correlation coefficients by a permutation test. Briefly, the significance of association between two nodes is tested by randomly permuting the abundance in all samples, recalculating PCC 10,000 times, and recording the frequency of getting a higher correlation coefficient after permutation than the real one as the P value. Then a network can be constructed by setting a P value cutoff such as 0.01.

Another way to calculate the weight matrix is based on a machine learning model. With the assumption that the abundance of a node in a sample can be predicted by the abundance of its associated nodes in the same sample, the network construction problem of N nodes can be regarded as N feature selection subproblems. In each subproblem, the abundance of one node is selected to be the learning target, and the abundance of the remaining $N - 1$ nodes serves as the feature to predict the learning target. Feature selection algorithm is applied to each subproblem to measure the importance of each feature, which is then used as the weight of the association between the target node and the feature node. In this study, we applied the RF algorithm to solve the feature selection subproblems. RF is similar to GENIE3, a state-of-the-art method used to infer gene regulatory networks [28]. In particular, in each subproblem, 1000 decision trees are grown by recursively splitting the samples based on binary tests of the features. A tree node is chosen from randomly-selected $\sqrt{N-1}$ features by optimizing the split to obtain the greatest variance reduction in the learning target. The reduction averaged across all trees can serve as the measure of feature importance. With N learning samples, we obtained the importance of all features to all targets, *i.e.*, the pairwise association of all nodes, to form the $N \times N$ weight matrix. We used the R package randomForest to apply this method based on the RF algorithm.

The $N \times N$ weight matrix can be regarded as a fully connected weighted network, and the network construction is accomplished by setting a threshold to filter out the edges with lower weights but keep the edges with higher weights. In most of this study, we completed the networks by keeping $1.5N$ edges with the highest weights, and we call 1.5 as a relative threshold because this threshold yields networks with a better scale-free topology. Different threshold values have also been tested but failed to find significant differences in the identified modules.

Analysis of functional networks

We used Cytoscape [22] to visualize biological networks, draw the networks, and calculate their topological characteristics such as densities, centralization, diameters, clustering coefficients, and scale-free properties. MCODE [21], a plugin for Cytoscape, was then used to detect modules in biological networks and find the highly-connected modules in the KO and OG networks. To do this, we first intersected the two KO (or OG) networks constructed based on different methods to obtain an intersection KO (or OG) network, and then applied MCODE (with default parameters) to the intersection network to detect modules that would be present in both KO (or OG) networks. Only modules with an MCODE score higher than 2.0 were retained for further analysis.

For a specific node group in the network, we used two hypothesis tests to detect its specific properties. The first one is a permutation test based on shuffling edges in the network. For example, to test whether a node group is densely connected, we first calculate the density of this node group in the network. Then, we switch the edges of this network $10 \times y$ times (where y is the total number of edges) while keeping the degree of each node unchanged, and thus obtain a randomized network. By calculating densities of the above-mentioned node group in 10,000 such random networks, we obtain an empirical distribution. The empirical P value, which means the frequency of greater densities in random networks than in the real network, is finally calculated and used to show the significance of the edge-shuffling permutation test. We used this method to test whether T2D markers are densely connected in functional networks, and whether nodes from a network module are highly connected in other related networks.

The second hypothesis test is also a permutation test which permutes node labels. To determine whether a node group is randomly distributed in a network, we first calculate a topological property such as the density of this node group, and then randomly select the same number of nodes from the network 10,000 times to calculate the empirical P value, which is transformed from the frequency of greater values observed. The density, average degree, and average stress centralization are all topological properties of a node group and can be used in this kind of test.

We analyzed the local structures of a network by finding network motifs using mfinder1.2 [26]. For non-directed networks, we used mfinder1.2 to find the numbers of all the 6 tetrads present in the real network and in 100 random networks created by switching edges. For each kind of tetrad, we used its number in the real network, *Nreal*, and its average number in 100 random networks, $<Nrand>$, to calculate the score [26] below:

$$\Delta_i = \frac{Nreal_i - <Nrand_i>}{Nreal_i + <Nrand_i> + 4} \tag{4}$$

After that, the subgraph ratio profile, which is a vector of Δ_i that is normalized to length 1, was calculated, representing a measure of a network's local structure of a network independent of its size.

Authors' contributions

DJ and RJ designed the project. LL and HP performed the experiments. LL, WZ, and HP analyzed the data. LL, MS, and RJ wrote the paper. All authors read and approved the final manuscript

Competing interests

The authors have declared no competing interests.

Acknowledgments

This research project was partially supported by the National Basic Research Program of China (973 Program; Grant No. 2012CB316504), the National High-tech R&D Program of China (863 Program; Grant No. 2012AA020401), the National Natural Science Foundation of China (Grant Nos. 61175002 and 61573207), and Beijing Collaborative Innovation Centre for Cardiovascular Disorders.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2016.02.005.

References

[1] King AJ, Farrer EC, Suding KN, Schmidt SK. Co-occurrence patterns of plants and soil bacteria in the high-alpine subnival zone track environmental harshness. Front Microbiol 2012;3:347.

[2] Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, et al. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol 2000;66:2541–7.

[3] Voget S, Leggewie C, Uesbeck A, Raasch C, Jaeger KE, Streit WR. Prospecting for novel biocatalysts in a soil metagenome. Appl Environ Microbiol 2003;69:6235–42.

[4] Nautiyal CS, Chauhan PS, Nene YL. Medicinal smoke reduces airborne bacteria. J Ethnopharmacol 2007;114:446–51.

[5] Ortiz G, Yague G, Segovia M, Catalan V. A study of air microbe levels in different areas of a hospital. Curr Microbiol 2009;59:53–8.

[6] Martinez A, Tyson GW, Delong EF. Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. Environ Microbiol 2010;12:222–38.

[7] Steele JA, Countway PD, Xia L, Vigil PD, Beman JM, Kim DY, et al. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. ISME J 2011;5:1414–25.

[8] Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science 2015;348:1261359.

[9] Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 2006;55:205–11.

[10] Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012;490:55–60.

[11] Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464:59–65.

[12] Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet 2012;13:260–70.

[13] Zhu C, Jiang R, Chen T. Constructing a Boolean implication network to study the interactions between environmental factors and OTUs. Quant Biol 2015;2:127–41.

[14] Hurwitz BL, Westveld AH, Brum JR, Sullivan MB. Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. Proc Natl Acad Sci U S A 2014;111:10714–9.

[15] Chaffron S, Rehrauer H, Pernthaler J, von Mering C. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res 2010;20:947–59.

[16] Levy R, Borenstein E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. Proc Natl Acad Sci U S A 2013;110:12804–9.

[17] Li B, Yang Y, Ma L, Ju F, Guo F, Tiedje JM, et al. Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. ISME J 2015;9:2490–502.

[18] Du J, Yuan Z, Ma Z, Song J, Xie X, Chen Y. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. Mol Biosyst 2014;10:2441–7.

[19] Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 2016;44: D286–93.

[20] Albert R, Barabási AL. Statistical mechanics of complex networks. Rev Mod Phys 2002;74:47.

[21] Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 2003;4:2.

[22] Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 2011;27:431–2.

[23] Djordjevic S, Stock AM. Structural analysis of bacterial chemotaxis proteins: components of a dynamic signaling system. J Struct Biol 1998;124:189–200.

[24] Brandt U. Energy converting NADH:quinone oxidoreductase (complex I). Annu Rev Biochem 2006;75:69–92.

[25] Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc Natl Acad Sci U S A 2012;109:594–9.

[26] Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, et al. Superfamilies of evolved and designed networks. Science 2004;303:1538–42.

[27] Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 2009;25:1966–7.

[28] Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods 2012;9:796–804.