[Frontiers in Bioscience S4, 1333-1343, June 1, 2012]

Computational methods for the analysis of tag sequences in metagenomics studies

Qin Chang[1], Yihui Luan[1], Ting Chen[2,3], Jed A. Fuhrman[4], Fengzhu Sun[2,3]

[1]School of Mathematics, Shandong University, Jinan, Shandong 250100, PR China, [2]Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA 90089-2910, USA, [3]TNLIST/Department of Automation, Tsinghua University, Beijing 100084, PR China, [4]Department of Biological Sciences and Wrigley Institute for Environmental Studies, University of Southern California, Los Angeles, CA 90089-2910, USA

TABLE OF CONTENTS

1. ABSTRACT

Metagenomics commonly refers to the study of genetic materials directly derived from environments without culturing. Several ongoing large-scale metagenomics projects related to human and marine life, as well as pedology studies, have generated enormous amounts of data, posing a key challenge for efficient analysis, as we try to 1) understand microbial organism assemblage under different conditions, 2) compare different communities, and 3) understand how microbial organisms associate with each other and the environment. To address such questions, investigators are using new sequencing technologies, including Sanger, Illumina Solexa, and Roche 454, to sequence either particular genes, called tag sequences, mostly 16S or 18S ribosomal RNA sequences or other conserved genes, or whole metagenome shotgun sequences of all the genetic materials in a given community. In this paper, we review computational methods used for the analysis of tag sequences.

2. INTRODUCTION

Metagenomics is the study of genetic materials derived directly from a microbial community without culturing. The term "metagenome" was first used by Handelsman *et al*. (1) in 1998. Since then, many large-scale metagenomics projects have been undertaken, including, for example, the human microbial project (HMP, http://commonfund.nih.gov/hmp/), the MetaHIT project (http://www.metahit.eu/), the Global Ocean Survey (http://www.jcvi.org/cms/research/projects/gos/), ICOMM (http://icomm.mbl.edu/) and the Earth Microbiome Project (http://www.earthmicrobiome.org/). Enormous amounts of metagenomics data have resulted using sequencing technologies such as Sanger, Illumina Solexa, Roche 454 and others. The challenges for metagenomics studies include the direct sampling of genetic materials from the microbial communities, data storage and data analysis (2, 3). Here we review computational methods for analyzing particular genes, called tag sequences, which are mostly

16S or 18S rRNA sequences. It was found that 16S/18S rRNA genes can be horizontally transferred between different species and one species can contain multiple copies of 16S/18S genes. Thus, the use of 16S/18S rRNA genes may not be optimal for community comparison. Other single copy house-keeping genes such as rpoB (4) and other conserved genes (5) were also used for comparing microbial communities. The methods described in this review can equally be applicable to such sequence data.

Microbes are key players and ubiquitous in almost all natural and man-made habitats. It is estimated that the population of bacteria may be as high as 10^{30} on earth, far outnumbering all other life forms. They exist in all environments, every tissue of the human body, salt and fresh waters, within polar ice and boiling hot springs, in surface soils and deep bedrock, in acidic mine drainage and alkaline lakes. They can be found coating both the skin and gut of any animal, as well as the surface of any plant or fungus. The number of microbial cells in an average healthy human adult is about 10 times higher than human cells.

Traditional microbiological studies heavily depend on *in vitro* studies. However, only a very small fraction of microbial organisms can be cultured, and the use of culturing techniques alone significantly limits our understanding of the microbial world. Various culture-independent methodologies able to retrieve genetic materials directly from natural environments have been developed. These culture-independent techniques have revealed high microbial diversity present in many different environments. Among these techniques, profiling, or fingerprinting, methods provide information on the whole community at once, usually in the form of a list of gene fragments representing different operational taxonomic units (OTUs), and the OTUs are supposed to represent closely related organisms. These methods include T-RFLP (6, 7), DGGE (8, 9), and ARISA (10, 11). By allowing relatively easy analysis and simultaneous comparison of many samples, fingerprinting studies have revealed spatial and temporal patterns among the OTUs and environmental factors (12-14). However, profiling-based methods do not yield detailed information about the microbial organism composition within communities. Fortunately, with the rapid development of sequencing technologies and significant drop of sequencing cost, it is now possible to use new sequencing technologies to study community diversity. These studies have shown that changes of the microbial community structure within the human body are associated with human health, such as obesity (15-17) and clinically defined bacterial vaginosis (18). The human gut microbial community can significantly change after treatment with antibiotics (19, 20).

Modern molecular techniques have revealed high microbial diversity in various human tissues. In the first phase of the HMP, investigators studied the composition of the microbial communities in various human tissues. In the second phase, the microbial communities associated with human diseases will be identified. Several studies have been carried out, including the study of microbial communities in the gut (21-23), saliva (24), skin (25, 26), vagina (18) and stomach (27). In addition to HMP, many other metagenomics projects, including marine and pedology studies (28-36), are underway or are in the planning stages. For instance, Dinsdale *et al.* (37) compared the metagenomic communities of 45 distinct microbiomes and 42 distinct viromes and found that they have distinct metabolic profiles. To date, most studies of microbial diversity have used ribosomal RNA (rRNA) sequences, in particular 16S and 18S, because they are ubiquitous and largely well conserved during evolution (38). Other types of gene sequences have also been used (4, 5). These sequences are sometimes called tag sequences. Because tag sequences are generally short, very deep sequencing is possible. Tag sequences are generally highly conserved and they can be used to study the microbial organism compositions in communities. However, it is impossible to study the functions of individual genes based on tag sequences. To study functions of genes and pathways, whole genome shotgun sequences are needed.

With the accumulation of enormous amounts of sequence data, there is an urgent need for novel computational tools able to analyze them and link the results to knowledge databases, such as Greengenes (39) and SILVA (40), to learn how different organisms interact with each other and with the environment. In this paper, we review computational methods for the analysis of tag sequences, including how to 1) classify the tag sequences into OTUs, 2) compare different communities, and 3) study the association of OTUs and environmental factors.

3. OPERATIONAL TAXONOMIC UNIT (OTU)-BASED ANALYSIS OF METAGENOMICS COMMUNITIES

The comparison of different communities is an important problem in many different fields, including ecology and microbiology. Many different measures, termed beta diversity, have been proposed to compare communities, and many of the methods were reviewed in (41). Some studies comparing communities based on gene contents and their metabolic functions (32, 42) depend heavily on the accuracy of the functional annotation process. Here, we concentrate on operational taxonomic unit (OTU)-based methods using tag sequences. In this section, we review computational methods for defining OTUs and for comparing communities based on OTUs.

3.1. Computational methods for the identification of OTUs

Tag sequences can be grouped into different clusters such that the sequences in each cluster are similar, but sequences in different clusters have relatively large differences. The sequences in each cluster form an operational taxonomic unit (OTU). The motivations of using OTUs are as follows. Microbial communities are usually highly diverse, containing hundreds to thousands of microbial organisms. However, tag sequences of only a small fraction of these organisms are known and well studied. Thus, studies based on the relationships of known

tag sequences maybe biased and do not present full understanding of the microbial diversity in communities. On the other hand, OTUs do not depend on the available information about the known tag sequences and thus present an unbiased view of the microbial diversity. However, the OTUs do depend on the algorithms used and it is still an active area of research for the optimal definition of OTUs. Although the definition of OTUs is conceptually simple, the computational implementation for the identification of biologically meaningful OTUs has turned out to be a very challenging problem, and optimal methods are still being debated and developed. The difficulties in defining biologically meaningful OTUs can be attributed to several factors. First, there is a large quantity of tag sequence data from metagenomics projects, which mandates that the computational algorithm be both storage and computationally efficient. Second, errors are present in sequence reads which can make the number of predicted OTUs much higher than the true number of OTUs present in microbial communities. Third, many different clustering approaches are available, and it is not clear which clustering methods give the most biologically meaningful results. Recent active studies on this topic have begun to answer some of these questions.

Two steps are frequently used in algorithms for defining OTUs. The first step is the calculation of distances between any pair of sequences. Some programs used multiple sequence alignment (MSA) to first align the sequences. Afterwards, the distance between any pair of sequences is calculated on the basis of this alignment (43, 44). Schloss (45) and White *et al.* (46) studied factors that can significantly affect estimating the diversity of individual communities, termed alpha diversity, and comparing multiple communities, termed beta diversity. These factors include the quality of the MSA, the inclusion/exclusion of variable regions along the tag sequences, and the distance calculation methods between groups of tag sequences. It is well known that MSA for a very large number of sequences, e.g., on the order of 10^6, is computationally challenging, and no efficient algorithms are available to align such a large number of sequences. In addition, there is no guarantee that MSA of the sequences will give better results than distances calculated based on pairwise sequence alignment (PSA) of tag sequences. Therefore, as an alternative to MSA, Sun *et al.* (47) proposed using PSA to calculate the distances between any pair of sequences. Surprisingly, they showed that pairwise distances calculated from PSA yield more biologically meaningful OTUs than those based on MSA. The use of PSA also reduces the computation time significantly compared with the use of MSA.

The second step in defining OTUs is to cluster the sequences based on the pairwise distances (43, 44, 47, 48). Some programs used complete linkage in hierarchical clustering (43, 44) where the distance between two groups of sequences is defined as the maximum distance between sequence pairs from the two groups. However, recent studies showed that average linkage, where the distance between two groups of sequences is defined as the average distance between sequence pairs from the two groups, may give more biological meaningful OTUs than using complete linkage (47, 49, 50). Previously, defining OTUs has required a threshold value so that the distances between any two clusters would be above the threshold. However, if OTUs correspond to actual species, then no such threshold values exist as confirmed by recent studies (51). In addition, experimental errors such as the PCR errors and the sequencing errors are unavoidable, and as a result, the hierarchical clustering over-estimate the number of OTUs. To avoid using a threshold value in clustering, a probabilistic Bayesian clustering method, termed Clustering 16S rRNA for OTU Prediction (CROP), was recently proposed to cluster sequencing data and define OTUs (48). CROP models the sequencing data with a Gaussian mixture model, and uses a soft threshold for clustering. It was shown to accurately estimate the number of OTUs when applied to a sequence dataset of mixtures of cultures (48). Ye (52) proposed AbundantOTU to group sequences from closely related abundant species. The algorithm does not depend on pairwise or multiple sequence alignments, but is based on a consensus alignment algorithm that defines abundant OTUs. This algorithm can avoid the problem of relatively high error rate as in next generation sequence technologies. However, it cannot align sequences belonging to rare OTUs and these sequences can be analyzed using other approaches described above.

When comparing different methods for defining OTUs, investigators designed some benchmark data by either experimentally sequencing a community with known microbial species (49, 53) or computationally selecting a set of species and introducing errors in these sequences according to the sequence error models of sequencing technologies (46, 50, 51). One criterion used to evaluate algorithms for defining OTUs is comparing the number of OTUs given by the algorithms with the known number of species in the simulated community. Recently, Sun *et al.* (51) proposed using normalized mutual information (NMI) (54) and F-score (55) to evaluate algorithms for defining OTUs. The NMI score evaluates overall clustering by penalizing two types of errors: assignment of sequences from different species into the same OTU and assignement of sequences from the same species into different OTUs. The NMI score is 1 if the clustering completely agrees with the species, and it is close to 0 if the clustering of sequences is not related to the species where the tag sequences come from. As a complementary evaluative method, F-score was proposed to compare clustering of sequences from an algorithm with true underlying species classification of sequences (51). Given N sequences from m species (S_1, S_2, ..., S_m), we suppose that an algorithm clusters the sequences into n clusters (C_1, C_2, ..., C_n) . Let a_{ij} be the number of sequences from species S_i that are clustered into cluster C_j, i = 1, 2, ..., m; j = 1,2, ..., n. For species i and cluster j, the precision and recall are defined as

$$p_{ij} = \frac{a_{ij}}{|C_j|}, \qquad r_{ij} = \frac{a_{ij}}{|S_i|},$$

and the F-score is defined by

$$F_{ij} = \frac{2p_{ij}r_{ij}}{p_{ij} + r_{ij}}.$$

Finally, the F-score for the clustering from the algorithm is defined as

$$F = \sum_{i=1}^{m} \frac{|S_i|}{N} max_{j=1}^{n} F_{ij}.$$

Thus, the F-score will be one if the clusters from the algorithm are the same as the species classification of the sequences. Otherwise, the F-score is small. Both NMI and F-score have been used to evaluate the quality of OTU classifications using known sequences from a mixture of species.

The number of tag sequences in metagenomics studies is generally huge, usually on the order of 10^6. Thus, the storage of pairwise distances, as discussed above, between any pair of sequences is challenging and cannot be done in most situations. Similarly, the clustering step is computationally time-consuming. To address the storage issue, some algorithms used sparse matrix techniques, where only distances between closely related sequences are kept to represent the distance matrix. This step can overcome the storage problem because only a relatively small fraction of sequence pairwise distances are stored (44, 47). To overcome the computational issues related to clustering, several other algorithms, including CD-HIT (56), UCLUST (57) and ESPRIT-Tree (58), do not even need to generate pairwise distances. Instead of clustering all the sequences simultaneously, sequences are simply clustered as added, which reduces the computational burden significantly. It should be noted that CD-HIT and UCLUST were developed to cluster protein sequences and they were adapted to the identification of OTUs. Sun *et al.* (51) compared the three approaches and showed that ESPRIT-Tree generally gives the highest accuracy, followed by UCLUST, with CD-HIT having the lowest accuracy. On the other hand, UCLUST takes the least amount of time and ESPRIT-Tree is slightly faster than CD-HIT. These developments significantly speed up algorithms for defining OTUs and have made large-scale analysis of OTUs possible. This is an area of continuing active research.

3.2. Comparison of communities based on OTUs

Suppose that there are multiple microbial communities and that tag sequences from each of the communities are obtained. How can we compare the communities based on the tag sequences? One commonly used approach is to cluster all the tag sequences from all the communities into OTUs and then measure the differences among the communities using some distance measures, termed beta diversity, based on the distribution of OTUs in the different communities. Various beta diversity measures (41, 59) can be used to compare the communities. Specifically, beta diversity measures can be grouped into qualitative or quantitative measures. Qualitative measures, such as classic versions of Jaccard, Lennon, and Dice, consider the presence/absence of OTUs within communities without considering their abundance. On the other hand, quantitative measures, such as classic versions of Bray-Curtis, Canberra, Euclidean, and Chao's statistic (60), take the abundance of OTUs into consideration. Recently, Kuczynski *et al.* (61) studied 14 quantitative and 9 qualitative beta diversity measures based on OTUs. They showed that these measures have varied abilities to identify the relationships between community microbial composition and 1) environmental changes, or 2) community clusters. For example, Chi-square and Pearson correlation distances perform extremely well in identifying environmental gradients of the communities, while Gower and Canberra distances perform well in identifying community clusters. These beta diversity measures have been incorporated into several metagenomics analysis pipelines, including QIIME (62) and SONS (63), which is currently incorporated into MOTHUR (44). Another novel network-based community comparison method was reported in (22), where OTUs and communities were abstracted to nodes in a bipartite graph. In this scheme, an OTU is connected to a community if the OTU is present in the community. The weight of the edge is the number of sequences in the OTU belonging to the community. Network analysis tools such as Cytoscape (http://www.cytoscape.org/) can then be used to analyze the network.

4. PHYLOGENY-BASED METHODS FOR COMPARING METAGENOMICS COMMUNITIES

Phylogenetic methods are those that take evolutionary relationships of the sequences into consideration in the comparison of communities. Here, we briefly review some of the approaches, while a more complete comparison of such methods is given in (64).

4.1. The F_{ST} test and the phylogenetic (P) test

OTU-based beta diversity measures have two main disadvantages. The first, as we have seen in subsection 3.1, involves the difficult problem of accurately defining OTUs. Mistakes in defining the OTUs may lead to misleading results about community relationships. Secondly, OTUs are treated as equal in terms of presence/absence or abundance levels, even though some of them may be closely related and some may not. To overcome these shortcomings, Martin (65) introduced two statistics borrowed from population genetics and systematics for comparing samples, F_{ST} and phylogenetic (P) test, which take evolutionary relationships of sequences into consideration.

The F_{ST} statistic assesses the difference between communities by comparing the genetic diversity within each single community to the total genetic diversity in the combined community consisting of all sequences from both communities. The F_{ST} is defined as:

$$F_{ST} = \frac{\theta_T - \theta_W}{\theta_T}$$

where θ_T is the genetic diversity of the combined sample, and θ_W is the average genetic diversity within each sample (65).

There are various statistics for estimating genetic diversity in a sample. One that is commonly used takes the average nucleotide differences between two randomly chosen sequences from the sample, as calculated by

$$\theta = \sum_{i=1}^{k} \sum_{j<i} p_i p_j d_{ij}$$

where k is the number of distinct sequences, p_i and p_j are the frequencies of the ith and jth sequences, respectively, and d_{ij} is the number of differences between the ith and jth sequences (65).

The phylogenetic (P) test, also known as the parsimony test (66), can be described as follows. First, a phylogenetic tree, including all the sequences in the samples, is generated using a phylogenetic analysis tool such as PHYLIP (http://evolution.gs.washington.edu/phylip.html). Each sequence is labeled according to the community the sequence comes from. Based on this observed tree, the minimum number of changes needed to explain the labels, termed parsimony score, is calculated. If the two communities are the same, the labels of the sequences and the phylogenetic tree should be unrelated. In the literature, two randomization methods were proposed to test the hypothesis that the two communities are the same. The first approach is to randomize the tree for the sequences and keep their labels. The second approach is to randomize the labels of the sequences without changing the phylogenetic tree. For each approach, the p-value is approximated by the fraction of times that the resulting parsimony score for the randomized sample is equal to, or smaller than, the parsimony score for the original data. The p-values obtained from the two randomization approaches can be different because of the different randomization processes. Actually, the two randomization approaches test for different specific hypotheses. By randomizing the tree, the P-test tests the hypothesis that sequences from the two communities associate with each other through a random phylogenetic tree. By randomizing the labels of the sequences, the P-test tests the hypothesis that the sequences from the two communities are randomly distributed along the leaves of the observed phylogenetic tree. Both approaches have been used to compare communities. As a test strategy, the phylogenetic (P) test cannot be used as a measure of beta diversity because the p-value depends on the number of sequences in each individual community in addition to differences among all communities. The phylogenetic (P) test has been implemented in TreeClimber (63), which is now included in MOTHUR (44).

4.2. UniFrac, weighted UniFrac and variance adjusted weighted UniFrac

Two other widely used phylogenetic methods for comparing communities are UniFrac and weighted UniFrac (W-UniFrac), both proposed by Lozupone *et al.* (67, 68), and they have been widely used in many studies, e.g. (22, 69, 70). Similar to the phylogenetic (P) test, a phylogenetic tree composed of sequences from all the communities is needed, and each sequence is labeled according to the sample it comes from. UniFrac measures the distance between communities by the fraction of branch length of the tree that leads to descendants from each of two single communities, but not from both communities (67), whereas weighted UniFrac takes abundance information into consideration and weights each branch length by the difference of the fractions of sequences from the two communities belonging to the branch (68). UniFrac and W-UniFrac are calculated using the following equations:

$$\text{UniFrac} = \frac{\sum_i^n b_i |A_i - B_i|}{\sum_i^n b_i}$$

$$\text{W-UniFrac} = \frac{\sum_i^n b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|}{\sum_j^m d_j \times \left| \frac{\alpha_j}{A_T} + \frac{\beta_j}{B_T} \right|}$$

where n is the number of branches in the tree, and b_i is the length of branch i. $A_i = 1$, if there are sequences that descend from branch i in community A, and $A_i = 0$ otherwise. The case is similar for B_i. A_i and B_i are the numbers of sequences that descend from branch i in communities A and B, respectively, and A_T and B_T are the total numbers of sequences in communities A and B, respectively. m is the number of different sequences in the two communities, d_j is the distance from the root to sequence j, while α_j and β_j are the number of times sequence j is observed in communities A and B, respectively. All the above numbers of sequences should be counted with multiplicity, except m. These two statistics are implemented in the ``Fast UniFrac'' software package (71).

Based on UniFrac and weighted UniFrac, we recently proposed a new quantitative measure (72), termed variance adjusted weighted UniFrac (VAW-UniFrac). Compared to weighted UniFrac, this new statistic adjusts the weights of branch lengths according to the variance of $\frac{A_i}{A_T} - \frac{B_i}{B_T}$ under the null model that the labels of the sequences are randomly distributed on the leaves of the tree. VAW-UniFrac is calculated as follows:

$$\text{VAW-UniFrac} = \frac{\sum_{i=1}^{n} b_i \times \frac{\left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|}{\sqrt{m_i(M-m_i)}}}{\sum_{i=1}^{n} b_i \times \frac{\left| \frac{A_i}{A_T} + \frac{B_i}{B_T} \right|}{\sqrt{m_i(M-m_i)}}}$$

where $m_i = A_i + B_i$ is the number of sequences belonging to the i-th branch, $M = A_T + B_T$ is the total number of sequences in the tree, while the other annotations remain the same as in UniFrac and W-UniFrac. It was shown in (72) that VAW-UniFrac is always more powerful than W-UniFrac and is more powerful than UniFrac when the sequences from each community are not uniformly distributed along the tree, meaning that VAW-UniFrac is more likely to detect differences and find meaningful gradient among various communities.

Despite the wide applications of UniFrac and W-UniFrac, some potential problems have been observed (64) when they are used to cluster communities based on the observation that their mean values decrease with the

number of sequences from the two communities. The simulations used by Lozupone and colleagues agreed with this observation; that is, when the number of sequences is relatively small, e.g., less than 1000, then the mean values of UniFrac and weighted UniFrac decrease with the number of sequences from the communities, but their mean values become stable when the number of sequences is greater than 1000 (73). Thus, UniFrac and W-UniFrac depend on the number of sequences from the communities. To overcome this potential problem, Lozupone *et al.* (73) suggested using bootstrap to sample the same number of sequences from the communities, thus providing a method of comparison when the number of sequences from some communities are relatively small. As an extension of W-UniFrac, VAW-UniFrac experiences this same problem; hence, the bootstrap strategy should be employed when the concern warrants it. Another more philosophical issue about UniFrac is that it assumes that "differences" between communities are proportional to the phylogenetic distances of their constituent members. This may be true for some questions, but not all. It depends on how the distance is interpreted, as factors like ecological roles do not uniformly follow phylogeny. So the "ecological scale" of phylogenetically close and far distances is inherently not predictable.

5. ASSOCIATION NETWORKS OF OTUS AND ENVIRONMENTAL FACTORS

Microbial organisms do not function independently in communities. Instead, they interact with each other and with environmental factors (ENV). Without precise knowledge about organisms within communities, we can study the association of OTUs and, as a consequence, form OTU networks. Given the distribution of OTUs in a community under multiple time points, locations, or environmental conditions, Pearson correlation or Spearman correlation can be used to study the association of OTUs and ENVs. An OTU/ENV network can then be constructed, assuming that two OTUs are connected if their abundance levels are significantly associated. For presence/absence of OTUs across many different time points, locations, or environmental conditions, an OTU co-occurrence network can also be obtained as in (74). Two OTUs are connected if they are more likely to co-occur than expected, for example if they prefer similar environmental conditions or if they facilitate each other's survival, as in cooperative relationships like symbioses. Network analysis tools, such as Pajek (75) and Cytoscape (76), can be used to analyze microbial OTU/ENV association networks. Note there are also significant negative associations that may imply interactions like competition or predation, or preference for opposite seasons.

With metagenomics data from a series of time points, i.e., time series data, it is possible to define time-delayed-local association between OTU/ENVs, as defined in (77). Standard statistical approaches, such as Pearson or Spearman correlation, may not be able to capture such complex interactions in reality. For example, it was found that two OTUs may only associate within a subset of the time interval of interest. Moreover, it is possible that one OTU, OTU1, may have a time-delayed response to the abundance changes of another OTU, OTU2, thus creating a time-delayed association, as might, for example, be the case in the administration of antibiotics or host immune response to pathologic overload. As suggested, linear regression and Pearson or Spearman correlation will most likely fail to detect the relationship between OTU1 and OTU2 in such situations in that these statistics can only detect global linear relationships between OTU/OTU and OTU/ENV pairs. Obviously, these problems call for the exploration of alternate analytical methods, and in order to identify such complicated relationships between OTU/OTU and OTU/ENV pairs, we developed local similarity analysis (LSA) with time delays to study the relationship between OTU/OTU and OTU/ENV pairs (77). The following procedures were used to identify potentially time-delayed-local associations. First, the abundance levels of each OTU across the time series are normalized so that they can be considered samples from the standard normal distribution. Second, a dynamic program algorithm is then used to find potentially time-delayed-local intervals with highest absolute correlation. Third, a p-value is then calculated by randomization of the normalized abundance levels of the OTUs. Fourth, the p-values are then transformed to q-values for each pair of sequences, and an OTU network can be constructed by thresholding on the p-values or q-values. In most biological studies, both technical and biological noises are unavoidable. Here technical noise indicates errors introduced by the experiments and biological noise indicates randomness introduced during the sampling process. To study the effects of these noises on the local similarity score, biological/technical replicate experiments are usually carried out. We recently extended the original LSA to the situation with replicates termed extended LSA (eLSA) (78). With replicates, we are able to obtain the boostrap confidence interval for the LS score. The LSA software can be downloaded from http://meta.usc.edu/softs/lsa. The local association network approach has been applied to several environmental biological studies, and interesting results about the association of OTUs and environmental factors were obtained and discussed, e.g., findings reported in (79-83). For example, Steele *et al.* (81) built the largest most comprehensive ecological network using LSA in the ocean. We expect that network-based analysis of OTU/ENVs will play more important roles as more time series data are available.

6. DISCUSSION

Metagenomics is a rapidly developing field, and both tag and whole-genome shotgun sequence data are available. However, because of the large amounts of data, there is an urgent need for efficient computational tools to analyze these large datasets in order to understand microbial organism assemblage under different conditions, compare different communities, and understand how microbial organisms associate with each other and the environment. In this paper, we reviewed computational approaches for tag sequence analysis, including the definition of OTUs, the use of OTU- and phylogeny-based

methods to compare metagenomics communities, and the construction of OTU/ENV networks to study how OTUs associate with each other and with the environments.

We have seen that classifying sequences into different OTUs is an extremely difficult problem. Many shortcomings of the available methods for defining OTUs have been identified, but problems associated with new algorithms have not yet come to light. Clustering itself is an exploratory tool and can give deep insight into the microbial diversity of communities at various levels of phylognetic resolution. Due to the highly complex nature of the evolution of genomes, we recognize that OTUs based on one or a few tag sequences cannot perfectly correspond to microbial species (the characterization or even formal existence of which is still frequently debated), however the distribution of OTUs defined by clustering still has interesting and valuable ecological interpretations, e.g. (84). Although average linkage in hierarchical clustering tends to yield more stable and biologically meaningful OTUs than complete linkage, we doubt that hierarchical clustering is the optimal strategy for defining OTUs. Importantly, for short sequences in particular (as currently determined by the next-generation sequencing approaches like Illumina), the information in the distance matrix between the sequences may not be enough to cluster the sequences into clusters with certainty. Instead, probabilistic clustering may be a more reasonable alternative to hierarchical clustering of tag sequences. Specifically, it is not possible to determine if two specific sequences are definitely in the same cluster. Instead, we only know the probability that they are in the same cluster. To accommodate this idea, we developed a new method, termed CROP (48), which does not force a sequence into one cluster, but rather into different clusters based on probabilities for them to be in each cluster. At the same time, however, it has to be acknowledged that probabilistic clustering is computationally demanding and difficult to explain to non-statisticians. Despite the shortcomings of probabilistic clustering, we expect that further improvement in the computational speed of CROP will, in turn, improve OTU definition.

Once the OTUs are defined, many beta diversity measures can be used to compare communities. For instance, the study of Kuczynski *et al.* (61) highlighted the differences among a variety of beta diversity measures in recovering environmental gradients and clustering communities. However, it is not clear how the mis-specification of OTUs affects the results from different beta diversity measures.

We also reviewed phylogeny-based methods for comparing communities, including the parsimony test, UniFrac, weighted UniFrac, and our newly developed variance adjusted weighted UniFrac. UniFrac has been used in over 150 metagenomics studies, and important biological insights have been gained. On the other hand, all the methods we reviewed assume that the tree is given and is correct, that the tag sequences correctly place the sequence in a single place on the tree, and that the distances of interest between communities are reflected by phylogenetic distances. Placement on robust trees are most accurate when we match longer tag sequences unambiguously to existing RNA classification schemes, such as RDP of 16S RNA sequences, since the 16S RNA sequences are well studied, and detailed phylogenetic relationships among them are known. On the other hand, for short sequences that match multiple sequences in differnt parts of the tree nearly equally well, and for other types of tag sequences (non-16S) where the phylogenetic relationships are not clear, problems may emerge based on potential errors in properly placing the sequences on the phylogenetic tree, thus suggesting the need to further study the effects of such errors on phylogeny-based beta diversity measures.

Understanding how OTUs associate with each other and with the environment is another very important problem. Initial efforts to establish OTU co-occurrence networks have highlighted the importance of such an approach (74). Our previous analysis of marine time series data using ARISA showed interesting association patterns among OTUs and environmental factors (77, 81). Time series tag sequence data are now available (84), and local similarity analysis of such data, as described above, is giving us more detailed information on the association of microbial organisms. Nonetheless, more sophisticated local similarity analysis approaches are needed to identify other association patterns that cannot be discovered by the current version of LSA.

7. ACKNOWLEDGMENTS

This research was partially supported by NSFC grants 11071146, 60928007, and 60805010, and the National Basic Research Program of China (973 Program, No. 2007CB814901). QC is supported by Graduate Independent Innovation Foundation of Shandong University (GIIFSDU). TC, JF and FS are partially supported by US NSF DMS1043075 and OCE 1136818.

8. REFERENCES

1. J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy and R. M. Goodman: Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 5(10), R245-249 (1998)

2. K. Chen and L. Pachter: Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol* 1(2), 106-112 (2005)

3. J. C. Wooley and Y. Ye: Metagenomics: Facts and Artifacts, and Computational Challenges. *J Comput Sci Technol* 25(1), 71-81 (2009)

4. D. A. Walsh, E. Bapteste, M. Kamekura and W. F. Doolittle: Evolution of the RNA polymerase B' subunit gene (rpoB') in Halobacteriales: a complementary molecular marker to the SSU rRNA gene. *Mol Biol Evol* 21(12), 2340-2351 (2004)

5. C. von Mering, P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward and P. Bork: Quantitative

phylogenetic assessment of microbial communities in diverse environments. *Science* 315(5815), 1126-1130 (2007)

6. E. Avaniss-Aghajani, K. Jones, D. Chapman and C. Brunk: A molecular technique for identification of bacteria using small subunit ribosomal RNA sequences. *Biotechniques* 17(1), 144-149 (1994)

7. W. T. Liu, T. L. Marsh, H. Cheng and L. J. Forney: Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl Environ Microbiol* 63(11), 4516-4522 (1997)

8. G. Muyzer, E. C. de Waal and A. G. Uitterlinden: Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol* 59(3), 695-700 (1993)

9. M. Troussellier, H. Schafer, N. Batailler, L. Bernard, C. Courties, P. Lebaron, G. Muyzer, P. Servais and J. Vives-Rego: Bacterial activity and genetic richness along an estuarine gradient (Rhone River plume, France). *Aquatic Microbial Ecology* 28(1), 13-24 (2002)

10. M. V. Brown and J. A. Fuhrman: Marine bacterial microdiversity as revealed by internal transcribed spacer analysis. *Aquatic Microbial Ecology* 41(1), 15-23 (2005)

11. M. M. Fisher and E. W. Triplett: Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl Environ Microbiol* 65(10), 4630-4636 (1999)

12. J. A. Fuhrman, I. Hewson, M. S. Schwalbach, J. A. Steele, M. V. Brown and S. Naeem: Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc Natl Acad Sci U S A* 103(35), 13104-13109 (2006)

13. J. A. Fuhrman, J. A. Steele, I. Hewson, M. S. Schwalbach, M. V. Brown, J. L. Green and J. H. Brown: A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci U S A* 105(22), 7774-7778 (2008)

14. J. B. H. Martiny, B. J. M. Bohannan, J. H. Brown, R. K. Colwell, J. A. Fuhrman, J. L. Green, M. C. Horner-Devine, M. Kane, J. A. Krumins, C. R. Kuske, P. J. Morin, S. Naeem, L. Ovreas, A. L. Reysenbach, V. H. Smith and J. T. Staley: Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* 4(2), 102-112 (2006)

15. P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight and J. I. Gordon: A core gut microbiome in obese and lean twins. *Nature* 457(7228), 480-484 (2009)

16. P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis and J. I. Gordon: An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444(7122), 1027-1031 (2006)

17. Y. Sun, Y. Cai, V. Mai, W. Farmerie, F. Yu, J. Li and S. Goodison: Advanced computational algorithms for microbial community analysis using massive 16S rRNA sequence data. *Nucleic Acids Res* 38(22), e205 (2010)

18. B. B. Oakley, T. L. Fiedler, J. M. Marrazzo and D. N. Fredricks: Diversity of human vaginal bacterial communities and associations with clinically defined bacterial vaginosis. *Appl Environ Microbiol* 74(15), 4898-4909 (2008)

19. L. Dethlefsen, S. Huse, M. L. Sogin and D. A. Relman: The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* 6(11), e280 (2008)

20. C. Jernberg, S. Lofmark, C. Edlund and J. K. Jansson: Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *ISME J* 1(1), 56-66 (2007)

21. S. R. Gill, M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett and K. E. Nelson: Metagenomic analysis of the human distal gut microbiome. *Science* 312(5778), 1355-1359 (2006)

22. R. E. Ley, M. Hamady, C. Lozupone, P. J. Turnbaugh, R. R. Ramey, J. S. Bircher, M. L. Schlegel, T. A. Tucker, M. D. Schrenzel, R. Knight and J. I. Gordon: Evolution of mammals and their gut microbes. *Science* 320(5883), 1647-1651 (2008)

23. R. E. Ley, C. A. Lozupone, M. Hamady, R. Knight and J. I. Gordon: Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* 6(10), 776-788 (2008)

24. I. Nasidze, J. Li, D. Quinque, K. Tang and M. Stoneking: Global diversity in the human salivary microbiome. *Genome Res* 19(4), 636-643 (2009)

25. Z. Gao, C. H. Tseng, Z. Pei and M. J. Blaser: Molecular analysis of human forearm superficial skin bacterial biota. *Proc Natl Acad Sci U S A* 104(8), 2927-2932 (2007)

26. E. A. Grice, H. H. Kong, G. Renaud, A. C. Young, G. G. Bouffard, R. W. Blakesley, T. G. Wolfsberg, M. L. Turner and J. A. Segre: A diversity profile of the human skin microbiota. *Genome Res* 18(7), 1043-1050 (2008)

27. E. M. Bik, P. B. Eckburg, S. R. Gill, K. E. Nelson, E. A. Purdom, F. Francois, G. Perez-Perez, M. J. Blaser and D. A. Relman: Molecular analysis of the bacterial microbiota in the human stomach. *Proc Natl Acad Sci U S A* 103(3), 732-737 (2006)

28. N. Fierer, M. Breitbart, J. Nulton, P. Salamon, C. Lozupone, R. Jones, M. Robeson, R. A. Edwards, B. Felts, S. Rayhawk, R. Knight, F. Rohwer and R. B. Jackson: Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl Environ Microbiol* 73(21), 7059-7066 (2007)

29. D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier and J. C. Venter: The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5(3), e77 (2007)

30. M. L. Sogin, H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta and G. J. Herndl: Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* 103(32), 12115-12120 (2006)

31. S. G. Tringe and E. M. Rubin: Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* 6(11), 805-814 (2005)

32. S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz and E. M. Rubin: Comparative metagenomics of microbial communities. *Science* 308(5721), 554-557 (2005)

33. G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar and J. F. Banfield: Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428(6978), 37-43 (2004)

34. J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers and H. O. Smith: Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304(5667), 66-74 (2004)

35. S. Yooseph, K. H. Nealson, D. B. Rusch, J. P. McCrow, C. L. Dupont, M. Kim, J. Johnson, R. Montgomery, S. Ferriera, K. Beeson, S. J. Williamson, A. Tovchigrechko, A. E. Allen, L. A. Zeigler, G. Sutton, E. Eisenstadt, Y. H. Rogers, R. Friedman, M. Frazier and J. C. Venter: Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* 468(7320), 60-66 (2010)

36. S. Yooseph, G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J. M. Chandonia, D. A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier and J. C. Venter: The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* 5(3), e16 (2007)

37. E. A. Dinsdale, R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. Li, L. McDaniel, M. A. Moran, K. E. Nelson, C. Nilsson, R. Olson, J. Paul, B. R. Brito, Y. Ruan, B. K. Swan, R. Stevens, D. L. Valentine, R. V. Thurber, L. Wegley, B. A. White and F. Rohwer: Functional metagenomic profiling of nine biomes. *Nature* 452(7187), 629-632 (2008)

38. N. R. Pace: A molecular view of microbial diversity and the biosphere. *Science* 276(5313), 734-740 (1997)

39. T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu and G. L. Andersen: Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72(7), 5069-5072 (2006)

40. E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies and F. O. Glockner: SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35(21), 7188-7196 (2007)

41. A. E. Magurran: Measuring biological diversity. Blackwell Pub., Malden, Ma. (2004)

42. F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening and R. A. Edwards: The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9, 386 (2008)

43. P. D. Schloss and J. Handelsman: Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* 71(3), 1501-1506 (2005)

44. P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn and C. F. Weber: Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75(23), 7537-7541 (2009)

45. P. D. Schloss: The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol* 6(7), e1000844 (2010)

46. J. R. White, S. Navlakha, N. Nagarajan, M. R. Ghodsi, C. Kingsford and M. Pop: Alignment and clustering of phylogenetic markers--implications for microbial diversity studies. *BMC Bioinformatics* 11, 152 (2010)

47. Y. Sun, Y. Cai, L. Liu, F. Yu, M. L. Farrell, W. McKendree and W. Farmerie: ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res* 37(10), e76 (2009)

48. X. Hao, R. Jiang and T. Chen: Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* 27(5), 611-618 (2011)

49. S. M. Huse, D. M. Welch, H. G. Morrison and M. L. Sogin: Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* 12(7), 1889-1898 (2010)

50. P. D. Schloss and S. L. Westcott: Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *Appl Environ Microbiol* 77(10), 3219-3226 (2011)

51. Y. Sun, Y. Cai, S. M. Huse, R. Knight, W. G. Farmerie, X. Wang and V. Mai: A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief Bioinform, publshed online April 27* (2011)

52. Y. Ye: Identification and quantification of abundant species from pyrosequences of 16S rRNA by consensus alignment. *BIBM*, 153-157 (2010)

53. C. Quince, A. Lanzen, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read and W. T. Sloan: Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 6(9), 639-641 (2009)

54. C. Studholme, D. L. G. Hill and D. J. Hawkes: An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition* 32(1), 71-86 (1999)

55. C. J. Van Rijsbergen: Information retrieval. Butterworths, London ; Boston (1979)

56. W. Li, L. Jaroszewski and A. Godzik: Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17(3), 282-283 (2001)

57. R. C. Edgar: Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19), 2460-2461 (2010)

58. Y. Cai and Y. Sun: ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasi-linear computational time. *Nucleic Acids Res Published online May 19* (2011)

59. P. Legendre and L. Legendre: Numerical ecology. Elsevier, Amsterdam ; New York (1998)

60. A. Chao, R. L. Chazdon, R. K. Colwell and T. J. Shen: A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters* 8(2), 148-159 (2005)

61. J. Kuczynski, Z. Z. Liu, C. Lozupone, D. McDonald, N. Fierer and R. Knight: Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Methods* 7(10), 813-819 (2010)

62. J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Tumbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld and R. Knight: QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5), 335-336 (2010)

63. P. D. Schloss and J. Handelsman: Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl Environ Microbiol* 72(10), 6773-6779 (2006)

64. P. D. Schloss: Evaluating different approaches that test whether microbial communities have the same structure. *ISME J* 2(3), 265-275 (2008)

65. A. P. Martin: Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol* 68(8), 3673-3682 (2002)

66. P. D. Schloss and J. Handelsman: Introducing TreeClimber, a test to compare microbial community structures. *Appl Environ Microbiol* 72(4), 2379-2384 (2006)

67. C. Lozupone and R. Knight: UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71(12), 8228-8235 (2005)

68. C. A. Lozupone, M. Hamady, S. T. Kelley and R. Knight: Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* 73(5), 1576-1585 (2007)

69. E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon and R. Knight: Bacterial community variation in human body habitats across space and time. *Science* 326(5960), 1694-1697 (2009)

70. R. E. Ley, F. Backhed, P. Turnbaugh, C. A. Lozupone, R. D. Knight and J. I. Gordon: Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A* 102(31), 11070-11075 (2005)

71. M. Hamady, C. Lozupone and R. Knight: Fast UniFrac: facilitating high-throughput phylogenetic analyses of

microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* 4(1), 17-27 (2010)
72. Q. Chang, Y. Luan and F. Sun: Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics* 12(1), 118 (2011)

73. C. Lozupone, M. E. Lladser, D. Knights, J. Stombaugh and R. Knight: UniFrac: an effective distance metric for microbial community comparison. *ISME J* 5(2), 169-172 (2011)

74. S. Chaffron, H. Rehrauer, J. Pernthaler and C. von Mering: A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res* 20(7), 947-959 (2010)

75. V. Batagelj and A. Mrvar: Pajek - Analysis and visualization of large networks. In: *Graph Drawing Software*. Ed M. Junger, P. Mutzel. (2003)

76. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker: Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 13(11), 2498-2504 (2003)

77. Q. S. Ruan, D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman and F. Z. Sun: Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* 22(20), 2532-2538 (2006)

78. L. C. Xia, J. A. Steele, J. Cram, Z. G. Cardon, S. L. Simmons, J. J. Vallino, J. A. Fuhrman and F. Sun: Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Systems Biology (GIW2011)* (2011)

79. S. F. Paver and A. D. Kent: Temporal Patterns in Glycolate-Utilizing Bacterial Community Composition Correlate with Phytoplankton Population Dynamics in Humic Lakes. *Microbial Ecology* 60(2), 406-418 (2010)

80. A. Shade, C. Y. Chiu and K. D. McMahon: Differential bacterial dynamics promote emergent community robustness to lake mixing: an epilimnion to hypolimnion transplant experiment. *Environ Microbiol* 12(2), 455-466 (2010)

81. J. A. Steele, P. D. Countway, L. Xia, P. D. Vigil, J. M. Beman, D. Y. Kim, C. E. Chow, R. Sachdeva, A. C. Jones, M. S. Schwalbach, J. M. Rose, I. Hewson, A. Patel, F. Sun, D. A. Caron and J. A. Fuhrman: Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J* 5,1414-1425 (2011)

82. J. A. Fuhrman and J. A. Steele: Community structure of marine bacterioplankton: patterns, networks, and relationships to function. *Aquatic Microbial Ecology* 53(1), 69-81 (2008)

83. J. A. Fuhrman: Microbial community structure and its functional implications. *Nature* 459(7244), 193-199 (2009)
84. J. A. Gilbert, D. Field, P. Swift, L. Newbold, A. Oliver, T. Smyth, P. J. Somerfield, S. Huse and I. Joint: The seasonal structure of microbial communities in the Western English Channel. *Environ Microbiol* 11(12), 3132-3139 (2009)

Abbreviations: HMP: human microbial project, OTUs: operational taxonomic units, rRNA: ribosomal RNA, NMI: normalized mutual information, MSA: multiple sequence alignment, PSA: pairwise sequence alignment, P-test: phylogenetic test, W-UniFrac: weighted UniFrac, VAW-UniFrac: variance adjusted weighted UniFrac, ENV: environment factors

Key Words: Tag sequences, Metagenomics, Operational taxonomic units (OTUs), Community comparison, 16S rRNA, Phylogeny, Environment factors, Review

Send correspondence to: Fengzhu Sun, Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA 90089-2910, USA, Tel: 1-213-740-2413, Fax: 1-213-740-8631, E-mail: fsun@usc.edu