BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 24 2011, pages 3341–3347
doi:10.1093/bioinformatics/btr583

Genome analysis

Advance Access publication October 20, 2011

MetaRank: a rank conversion scheme for comparative analysis of microbial community compositions

Tse-Yi Wang[1,†], Chien-Hao Su[1,2,3,†] and Huai-Kuang Tsai[1,2,*]

[1]Institute of Information Science, [2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, 115 and [3]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 106, Taiwan

Associate editor: Martin Bishop

ABSTRACT

Motivation: Metagenomics involves sampling and studying the genetic materials in microbial communities. Several statistical methods have been proposed for comparative analysis of microbial community compositions. Most of the methods are based on the estimated abundances of taxonomic units or functional groups from metagenomic samples. However, such estimated abundances might deviate from the true abundances in habitats due to sampling biases and other systematic artifacts in metagenomic data processing.

Results: We developed the MetaRank scheme to convert abundances into ranks. MetaRank employs a series of statistical hypothesis tests to compare abundances within a microbial community and determine their ranks. We applied MetaRank to synthetic samples and real metagenomes. The results confirm that MetaRank can reduce the effects of sampling biases and clarify the characteristics of metagenomes in comparative studies of microbial communities. Therefore, MetaRank provides a useful rank-based approach to analyzing microbiomes.

Contact: hktsai@iis.sinica.edu.tw

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on August 9, 2011; revised on October 4, 2011; accepted on October 15, 2011

1 INTRODUCTION

Metagenomics is a field that involves sampling, sequencing and analyzing the genetic material of uncultured microorganisms in microbial communities. In metagenomic experiments, the nucleic acids of microorganisms are directly isolated from natural environments. This direct approach is not hampered by the limitation that most microorganisms are uncultured in the laboratory, and provides a more complete picture of a microbial community (Hugenholtz and Tyson, 2008). Recent advances in sequencing technologies have fueled an increase in both the number and scope of metagenomic projects, and large amounts of data are being accumulated for analyzing microbiomes (Biers *et al.*, 2009; Dinsdale *et al.*, 2008; Rusch *et al.*, 2007). A key question in metagenomics is whether and how changes in the microbial abundances of taxonomic units or functional groups relate to alterations of the habitats (Hamady and Knight, 2009). To characterize the relationship, it is important to compare microbial community compositions in different environments (Wooley *et al.*, 2010).

Several statistical methods developed for comparative metagenomics try to identify differentially abundant features between microbial communities. Most of these methods are designed to compare two communities, e.g. XIPE-TOTEC (Rodriguez-Brito *et al.*, 2006), IMG/M (Markowitz *et al.*, 2008), MEGAN (Mitra *et al.*, 2009), RAMMCAP (Li, 2009), Galaxy (Kosakovsky Pond *et al.*, 2009) and STAMP (Parks and Beiko, 2010). However, a few methods, like Metastats (White *et al.*, 2009), are developed specifically for comparing two sets of multiple communities, while other methods, such as ShotgunFunctionalizeR (Kristiansson *et al.*, 2009), are capable of performing both kinds of comparative analysis. All of the above methods employ statistical hypothesis tests to determine whether member (taxonomic unit or functional group) abundances are equal in distinct communities, and focus on the quantitative differences between microbial community compositions.

The drawback of the above methods is that they are highly dependent on the precision of estimated values (e.g. proportions) in member abundances. The estimated values might be noisy due to sampling biases and other systematic artifacts in metagenomic data processing, e.g. 16S rRNA chimeras and copy number variation, artificial replicates and inaccurate binning (Ashelford *et al.*, 2005; Brady and Salzberg, 2009; Gomez-Alvarez *et al.*, 2009; Mavromatis *et al.*, 2007). Although systematic artifacts can be corrected through improvements in data-processing techniques, sampling biases will remain unavoidable unless exhaustive inventories of different organisms in microbial communities become available (Wooley and Ye, 2010). According to the recent study of Gifford *et al.* (2011), an exhaustive sample for per liter of seawater requires >10 trillions 454 FLX pyrosequencing reads. Currently, only a fraction of genetic materials in habitats are collected as non-exhaustive samples in metagenomic experiments, and no dataset comprises trillions of reads. Therefore, in comparative analysis of metagenomes, sampling biases inevitably have impact on the analyzing accuracy of the above methods.

To mitigate the effects of sampling biases, we propose the MetaRank scheme, which performs a series of rank conversions for analysis of microbial communities based on the relative order of member abundances, rather than their estimated values. The rationale behind MetaRank is that, since the relative order of large

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

values is robust to small deviations, the ranks of highly abundant members are less affected by sampling biases. Based on this concept, MetaRank comprises a series of hypothesis tests to compare member abundances within communities and determine their relative order as follows. Highly abundant members are converted to large ranks. For any two members, if their estimated proportions of abundances cannot be distinguished from each other with statistical significance, they are converted to the same rank.

Empirical tests on two metagenomic datasets (Kurokawa *et al.*, 2007; Ley *et al.*, 2006; Mavromatis *et al.*, 2007) and synthetic samples confirmed that MetaRank is able to diminish the effects of sampling biases and help to clarify the characteristics of metagenomes in comparative analysis of microbiomes. The ranks converted by MetaRank have smaller normalized SDs than the estimated proportions and the ordinary ranks, which are converted by straightforward sorting abundances in numerical order. When measuring similarities of metagenomes by Pearson's correlations, we found that the ranks converted by MetaRank with small deviations clearly revealed the central tendencies of metagenomes; while detecting differences by t-test, using MetaRank capably identified the discriminative features with biological relevance. Therefore, MetaRank reduces the impact of sampling biases and provides a useful rank-based approach to analyze the microbial community compositions.

2 METHODS AND MATERIALS

The MetaRank scheme is designed to mitigate the effects of sampling biases by converting abundances into ranks. The scheme utilizes statistical hypothesis testing to iteratively select and convert highly abundant members into large ranks. To evaluate the effectiveness of MetaRank, we performed simulation experiments and analyzed the synthetic metagenomic samples under distinct conditions. In addition, we applied MetaRank to real metagenomes and assessed its usefulness in comparing microbial community compositions. All related methods and materials are described in the following sections, including binomial and multinomial tests employed by MetaRank, random sampling to generate synthetic samples, coefficient of variation for variability measurement between synthetic samples and Pearson's correlation for similarity measurement as well as t-test for difference detection between real metagenomes. Also, we introduce two public metagenomic datasets used in this study, inclusive of 16S rRNA sequences and shotgun sequences in human gut microbiomes from Ley *et al.* (2006) and Kurokawa *et al.* (2007) studies.

2.1 MetaRank

Given a metagenomic sample of a microbial community, MetaRank first employs binomial tests to iteratively select highly abundant members within the community, and then implement multinomial tests to rank the selected members in each run. We describe the steps in detail below.

2.1.1 Using binomial tests to select highly abundant members For N members in a microbial community, let X_n represent the abundance of the n-th member in the metagenomic sample, and $\widehat{p}_n$ (i.e. X_n/S) be the sample proportion of the n-th member, where $n=1,2,\ldots,N$ and $S=X_1+X_2+\ldots+X_N$. Under the assumption that all nucleic acids of microorganisms in habitats are equally likely to be sampled and sequenced in metagenomic experiments, the abundance X_n of the n-th member in the sample is modeled as a binomial random variable.

$$X_n \sim \text{Binomial}(S, p_n),$$

where p_n is the unknown population proportion of the n-th member in the habitat and estimated by the sample proportion $\widehat{p}_n$.

To select highly abundant members with proportions that are significantly higher than the average proportion ($1/N$), MetaRank applies hypothesis tests, $H_o: p_n \le 1/N$ versus $H_a: p_n > 1/N$ for all $1 \le n \le N$. Since $X_n \sim \text{Binomial}(S, p_n)$ with mean $E(X_n)=Sp_n$ and variance $\text{Var}(X_n)=Sp_n$ $(1-p_n)$, the binomial distribution of the test statistic X_n under H_o is approximated by normal distribution with z-statistic Z_n,

$$Z_n=\frac{X_n-E(X_n)}{\sqrt{\text{Var}(X_n)}}=\frac{X_n-\frac{S}{N}}{\sqrt{\frac{S}{N}\left(1-\frac{1}{N}\right)}}=\frac{\widehat{p}_n-\frac{1}{N}}{\sqrt{\frac{1}{SN}\left(1-\frac{1}{N}\right)}}\sim N(0,1),$$

when sample size S is large enough such that $0 \le E(X_n) \pm 3\sqrt{\text{Var}(X_n)} \le S$. Otherwise, exact binomial test is applied when S is small such that $E(X_n)-3\sqrt{\text{Var}(X_n)}<0$ or $S<E(X_n)+3\sqrt{\text{Var}(X_n)}$. The P-value for exact binomial test is calculated as follows,

$$P[X_n \ge x_n]=\sum_{k=x_n}^{S}\binom{S}{k}\frac{1}{N^k}\left(1-\frac{1}{N}\right)^{S-k},$$

where x_n is the observed value of the test statistic X_n.

For members that reject the null hypothesis with statistical significance, MetaRank considers them as highly abundant. For those that fail to reject the null hypothesis (assuming N' members remain), MetaRank continues to select members whose proportions are significantly larger than the average ($1/N'$) in the next iteration. When none of the remaining members rejects the null hypothesis, MetaRank terminates the selection procedure and considers all remaining members as rare members. Thus, in each iteration, the selected members (whose proportions are larger than the average) are higher than the remaining members (whose proportions are equal to or smaller than the average). Moreover, the members selected in distinct iterations are ranked in their selected order; for example, the member selected in first iteration is assigned a higher rank than the one selected in the second iteration. Eventually, the rare members are the lowest in the community.

2.1.2 Using multinomial tests to rank highly abundant members Based on the above procedure, MetaRank ranks the abundances in the target community according to the following three rules. First, all rare members are assigned the same smallest rank. Second, the members selected in distinct iterations are ranked according to the order in which they were selected; thus, the members selected in the first iteration of the procedure are assigned higher ranks than all the others. Third, if two abundances (the i-th and j-th members) are selected in the same iteration, MetaRank determines their ranks ($R_i > R_j$, $R_i < R_j$ or $R_i = R_j$) by two hypothesis tests, $H_o: p_i \le p_j$ versus $H_a: p_i > p_j$ and $H'_o: p_j \le p_i$ versus $H'_a: p_j > p_i$. If H_o is rejected, $R_i > R_j$; conversely, if H'_o is rejected, $R_i < R_j$. However, if both H_o and H'_o are accepted, we have $R_i = R_j$.

Under the same assumption that all nucleic acids are equally likely to be sampled and sequenced, each abundance X_n is modeled as a binomial random variable; any two abundances X_i and X_j are jointly modeled by the multinomial distribution (i.e. the generalization of binomial distribution in multidimension).

$$(X_i, X_j) \sim \text{Multinomial}(S, p_i, p_j),$$

where p_i and p_j are the unknown population proportions of the i-th and j-th members in habitat and estimated by the sample proportions $\widehat{p}_i$ and $\widehat{p}_j$. For large S such that

$$0 \le E(X_i) \pm 3\sqrt{\text{Var}(X_i)} \le S \text{ and } 0 \le E(X_j) \pm 3\sqrt{\text{Var}(X_j)} \le S,$$

the z-statistics of the approximate tests are

$$Z_{ij}=\frac{X_i-X_j}{\sqrt{X_i+X_j}} \text{ and } Z_{ji}=\frac{X_j-X_i}{\sqrt{X_j+X_i}}.$$

Otherwise, the exact multinomial tests are applied. The P-values are calculated as

$$P[X_i-X_j \ge x_i-x_j]=\sum_{h=x_i-x_j}^{S}\sum_{k=0}^{h-(x_i-x_j)}\frac{S!}{h!k!(S-h-k)!}\left(\frac{\widehat{p}_i+\widehat{p}_j}{2}\right)^{h+k}\left(1-\widehat{p}_i-\widehat{p}_j\right)^{S-h-k},$$

and

$$P[X_j - X_i \geq x_j - x_i] = \sum_{h=0}^{S-(x_j-x_i)} \sum_{k=h+(x_j-x_i)}^{S} \frac{S!}{h!k!(S-h-k)!} \left(\frac{\widehat{p}_i + \widehat{p}_j}{2}\right)^{h+k} \left(1-\widehat{p}_i-\widehat{p}_j\right)^{S-h-k},$$

where x_i and x_j are the observed values of X_i andX_j.

As a result, the sorted abundances $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(m)} \leq \ldots \leq X_{(M)}$ are converted into ranks $1 \leq R_{(1)} \leq R_{(2)} \leq \ldots \leq R_{(m)} \leq \ldots \leq R_{(M)} \leq M$, where the subscript in parentheses (m) denotes the m-th order in the community, and M is the total number of members. For members whose abundances cannot be distinguished from each other by hypothesis testing, MetaRank converts them into their average order; i.e. for any m', m'' such that $R_{(m')} < R_{(m'+1)} = R_{(m'+2)} = \ldots = R_{(m''-1)} < R_{(m'')}$ (given $R_{(0)} = 0$ and $R_{(M+1)} = M+1$), we have

$$R_{(m'+1)} = R_{(m'+2)} = \cdots = R_{(m''-1)} = \frac{m'+m''}{2}.$$

For example, the ranks of the rare members (assuming N'' members remain in the last iteration) are converted into $(N''+1)/2$.

2.2 Random resampling to generate synthetic metagenomic samples

To investigate the effects of sampling biases, synthetic metagenomic samples are generated by re-sampling reads from a pooled dataset of real metagenomes. Given a set of real metagenomic samples G_w, $w = 1, 2, \ldots, W$, all the reads in G_w are pooled together to construct a synthetic library D. To generate a synthetic sample, we first determine the number of reads (T) and then randomly choose the reads from D. The number T is arbitrarily assigned by the uniform distribution, UNIF(l, u), where l and u are pre-determined as follows. First, we performed simulation analysis under distinct sample-sequencing depth $r \in \{10\%, 20\%, \ldots, 90\%\}$, which is defined as the percentage of reads in a sample represented in library (Gifford *et al.*, 2011). For each r, we let $l = (r-5\%)|D|$, and $u = (r+5\%|D|$, where $|D|$ denotes the number of reads in D. Second, for simulation experiments under the condition that synthetic samples comprise similar numbers of reads as real metagenomes, we let $l = \min_{1 \leq w \leq W} |G_w|$ and $u = \max_{1 \leq w \leq W} |G_w|$, where $|G_w|$ denotes the number of reads in G_w. Further, once the number of reads is determined, we randomly choose reads from D without replacement under the assumption that each read is equally likely to be chosen.

We not only pooled all reads in Ley *et al.* (2006) study as a library for random resampling, but also built four other simulated microbial communities, including SimLC, SimMC and SimHC with low, middle and high complexity in taxonomic compositions, as well as SimCOG in COG-functional compositions. The first three were modified from the simulated metagenomes of Mavromatis *et al.* (2007) and the last was obtained by pooling all reads in Kurokawa *et al.* (2007) study. To make sure that SimLC, SimMC and SimHC contained reasonable numbers of highly abundant and rare OTUs, we modified the original compositions of Mavromatis *et al.* such that the taxonomic abundance distributions (Supplementary Fig. S1) fit the three well-known mathematical models in biodiversity: Motomura's geometric series, Fisher's log-series and Preston's log-normal series (Magurran, 1988).

2.3 Measuring variability between samples resulting from sampling biases

Given a set of synthetic metagenomic samples generated from a pooled dataset, we recruit coefficient of variation (CV), which is the normalized SD, to measure the variability between synthetic samples. Let Z_{uv} be the converted ranks (using MetaRank), the estimated proportions of raw abundances (without MetaRank) or the ordinary ranks (by straightforward sorting abundances in numerical order) of the u-th member in the v-th sample, $u = 1, 2, \ldots, M$ and $v = 1, 2, \ldots, C$. For each member (say the u-th), Z_{uv} may be different between distinct samples due to random biases in resampling. The variability of Z_{uv} in the u-th member is measured by

$$\text{CV}_u = \frac{\sigma_u}{|\mu_u|},$$

where μ_u and σ_u are the mean and SD of Z_{uv},

$$\mu_u = \frac{1}{C}\sum_{v=1}^{C} Z_{uv} \text{ and } \sigma_u = \sqrt{\frac{1}{C-1}\sum_{v=1}^{C}\left(Z_{uv}-\mu_u\right)^2}.$$

Furthermore, considering all members as a whole, the variability are measured by

$$\text{average of CV} = \frac{1}{M}\sum_{u=1}^{M} \text{CV}_u.$$

2.4 Pearson's correlation for measuring similarities between microbial communities

The Pearson's correlation is used to measure the similarity between two microbial communities as follows:

$$\rho = \frac{\sum_{k=1}^{M}(U_k - \overline{U})(V_k - \overline{V})}{\sqrt{\sum_{k=1}^{M}(U_k - \overline{U})^2 \sum_{k=1}^{M}(V_k - \overline{V})^2}},$$

where U_k and V_k, $k = 1, 2, \ldots, M$, are the ranks converted by MetaRank, the estimated proportions of raw abundances or the ordinary ranks of straightforward sorted abundances in the two communities. The distance between the two communities is defined as

$$d = 1 - \rho.$$

For a set E of microbial communities, we calculate the distance ($d_{ee'}$) between any two communities (e, $e' \in E$), and use Unweighted Pair Group Method with Arithmetic Mean (UPGMA) for hierarchical clustering. In each step, the closest two clusters are combined to form one cluster. The distance between two clusters is the unweighted arithmetic mean of all the distances between pairs of communities in distinct clusters:

$$\frac{1}{|A||B|}\sum_{a \in A}\sum_{b \in B} d_{ab},$$

where A and B are two clusters of communities; and $|A|$ and $|B|$ are the numbers of communities in A and B, respectively.

2.5 Two-sample *t*-test for detecting differences between microbial communities

Given two sets of microbial communities, we apply two-sample t-test to examine whether a member is differentially abundant between the two sets with the t-statistic:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{s_1^2}{c_1} + \frac{s_2^2}{c_2}}},$$

where $\overline{Y}_1$ and $\overline{Y}_2$ are the means of the ranks converted by MetaRank or the proportions of raw abundances, s_1 and s_2 are the SDs, and c_1 and c_2 are the numbers of communities in the two sets. In addition, we use Bonferroni correction to deal with the problem of multiple comparisons.

2.6 Metagenomic datasets

We used two real metagenomic datasets (Kurokawa *et al.*, 2007; Ley *et al.*, 2006) for empirical tests. The first one comprises human gut 16S rRNA sequences in Ley *et al.* (2006) study, and the second comprises human gut shotgun sequences in Kurokawa *et al.* (2007) study. We downloaded the first dataset from GeneBank (with accession numbers: DQ793220-DQ802819, DQ803048, DQ803139-DQ810181, DQ823640-DQ825343 and AY974810-AY986384) and assigned the 16S rRNA sequences to taxonomic units by

RDP classifier (Cole *et al.*, 2009) (version: RDP release 10; default settings: display depth 10 and confidence threshold 80%). Then, we obtained the taxonomy-based microbial community compositions of 50 samples. In the second dataset, the shotgun sequences had been assigned to clusters of orthologous groups (COGs) by Kurokawa *et al.* (2007). We directly obtained the function-based microbial community compositions of 13 samples from their Supplementary Materials.

The first dataset contained 46 obese samples and four lean controls of human gut microbiomes in a 1 year study of diet (Ley *et al.*, 2006), and the second dataset included four infant and nine adult samples of human gut microbiomes (Kurokawa *et al.*, 2007). In the first dataset, the obese samples were extracted from 12 obese individuals (I1, I2, …, I12) at four distinct time points (Weeks 0, 12, 26 and 52), but two samples were missing (I7 at Week 52 and I10 at Week 26) Ley *et al*. (2006) study; the lean controls were extracted from two lean individuals (I13 and I14) at two time points (0 and 52 weeks). In this study, we denote the obese samples and lean controls by IxWy, where x represents the x-th individual and y represents the time point. Since I7W52 and I10W26 were missing, we excluded I7W0 for concordance in comparison of the samples between Weeks 0 and 52. Also, when comparing the samples between four distinct time points, we excluded the ones of the individuals I7 and I10.

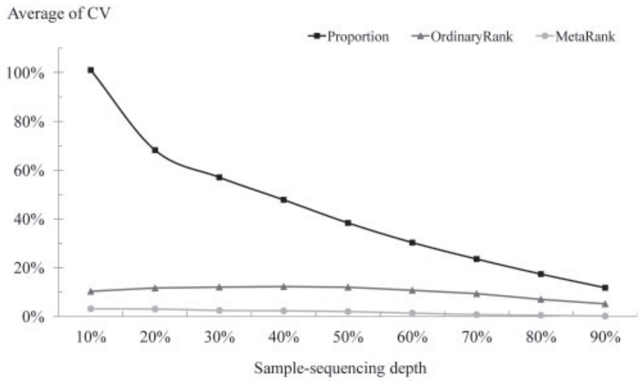

Fig. 1. The averages of CV, which is the normalized SD, in the ranks converted by MetaRank, the estimated proportions and the ordinary ranks at the phylum level of the 5000 synthetic samples for each sample-sequencing depth $r \in \{10\%, 20\%, \ldots, 90\%\}$. Under distinct sample-sequencing depth, the averages of CV in the ranks converted by MetaRank were smaller than the ones in the others.

3 RESULTS AND DISCUSSION

To evaluate the utility of MetaRank in comparative analysis of microbiomes, we applied it to synthetic samples (see Section 2 for details) and real metagenomes (Kurokawa *et al.*, 2007; Ley *et al.*, 2006). In synthetic samples generated from Ley *et al*. (2006) data and the other four simulated microbial communities (SimLC, SimMC, SimHC and SimCOG), we measured the variability between samples resulting from sampling biases and found that the ranks converted by MetaRank were with less variability than the estimated proportions of raw abundances and the ordinary ranks of straightforward sorted abundances. With less variability, the ranks converted by MetaRank are more able to clarify the central tendencies of samples than the estimated proportions and the ordinary ranks. Therefore, when measuring the similarities between real metagenomes in Ley *et al*. (2006) data, using MetaRank clearly revealed the common characteristics in the central tendencies of similar samples. While detecting the differences in Kurokawa *et al*. (2007) data, applying MetaRank capably identified the discriminative features with biological relevance. Our results suggested that MetaRank can mitigate the effects of sampling biases and provide a useful rank-based approach to analyze the microbiomes.

3.1 MetaRank reduces the variability resulting from sampling biases

We first generated synthetic samples from Ley *et al*. (2006) data and examine whether MetaRank was able to diminish the variability between synthetic samples resulting from sampling biases. All the 16S rRNA sequences were pooled together as a synthetic library, and at each taxonomic level of phylum, class, order, family and genus, 5000 synthetic samples were generated for each sample-sequencing depth $r \in \{10\%, 20\%, \ldots, 90\%\}$ (see Section 2 for details). The taxonomy-based microbial community composition in each sample was obtained by RDP classifier (Cole *et al.*, 2009). We measured the variability between the synthetic samples in the ranks converted by MetaRank, the estimated proportions of raw abundances and the ordinary ranks of straightforward sorted abundances by the normalized SD (see Section 2 for details). As shown in Figure 1, at the phylum level, the normalized SDs in the ranks converted by MetaRank were smaller than the ones in the estimated proportions and the ordinary ranks. Similar observations were found at the taxonomic levels of class, order, family and genus (Supplementary Fig. S2), as well as in SimLC, SimMC, SimHC and SimCOG (Supplementary Fig. S3). The results confirmed that MetaRank is able to reduce the variability between samples resulting from sampling biases.

With reduced variability resulting from sampling biases, we further found that MetaRank is able to clearly reveal the central tendencies of metagenomic samples. Under the condition that synthetic samples comprise similar numbers of sequences as real metagenomes, we generated 5000 synthetic lean controls from the 4 real lean controls (see Section 2 for details). As listed in Supplementary Table S1 at the phylum level, the estimated proportions and the ordinary ranks were different between distinct synthetic lean controls due to sampling biases. However in the ranks converted by MetaRank, 4983 synthetic lean controls (~99.67% of them) were identical to one another with common traits. The phyla *Firmicutes* and *Bacteroidetes* were the largest and second largest ranks; the other phyla were rare with the same smallest ranks. It is noted that, with small normalized SDs, the ranks converted by MetaRank are close to the central tendencies to reveal the common characteristics of metagenomes.

Moreover, given two sets of samples, using MetaRank is not only able to reveal the central tendency of samples in each set, but also helps to clarify whether the central tendency in one set is similar to or different from the central tendency in the other set. Among the real metagenomes in Ley *et al.* (2006) data, there are obese samples before and after a year of diet (Supplementary Tables S2–S4). The obese samples at Week 52 (after diet) were supposed to be more similar to the four lean controls than the obese samples at Week 0 (before diet). For each pair of obese sample and lean control, we adopted Pearson's correlation to measure similarity. As shown in Figure 2, in the ranks converted by MetaRank, the correlations between obese samples at Week 52 and lean controls were higher than the ones between obese samples at Week 0 and lean controls (significantly with $P = 1.97 \times 10^{-4}$ by t-test). But the

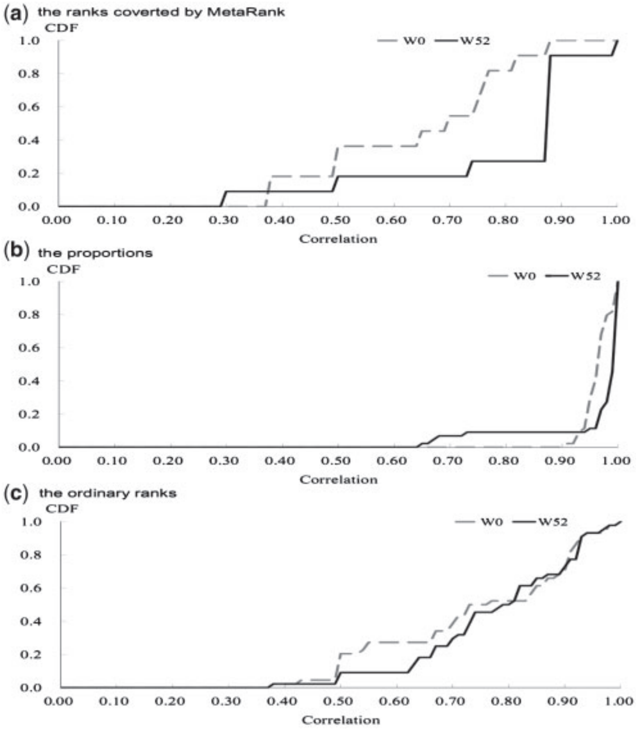

Fig. 2. The cumulative distribution functions (CDFs) of Pearson's correlations in ranks converted by MetaRank, the estimated proportions and the ordinary ranks. The gray dotted line denotes the CDF of the Pearson's correlations between the obese samples at Week 0 and the four lean controls; the black line denotes the CDF of the Pearson's correlations between the obese samples at Week 52 and the four lean controls. In the ranks converted by MetaRank, the correlations in the black line are significantly higher than the ones in the gray line with $P = 1.97 \times 10^{-4}$ by t-test. At Week 0, most obese samples were different from the four lean controls (∼80% of the correlations <0.75); at Week 52, most obese samples were similar to the lean controls (∼80% of the correlations >0.75). However, in the estimated proportions and ordinary ranks, the correlations in the black line are not higher than the ones in the gray line ($P = 0.57$ and 0.16).

estimated proportions ($P = 0.57$) and the ordinary ranks ($P = 0.16$) failed to show the similarities between obese samples after diet and lean controls.

Since MetaRank reduces the variability owing to sampling biases, the converted ranks with small deviations to central tendencies are able to reveal the common traits within a set of metagenomes and facilitate the comparative analysis of two sets of metagenomes. In the following sections, we demonstrate the usefulness of MetaRank in measuring similarities and detecting differences between microbial community compositions in Ley *et al.* (2006) and Kurokawa *et al.* (2007) data.

3.2 MetaRank reveals the central tendencies and helps to measure the similarities of metagenomes

We demonstrated the utility of MetaRank to measure similarities of microbiomes in Ley *et al.* (2006) 1 year study of diet. The taxonomic abundances in the obese samples and the lean controls were converted into ranks by MetaRank, followed by hierarchical clustering with UPGMA(Section 2). To facilitate the description of the clustering results using MetaRank, we first illustrate the simple case study that only consists of the samples at Weeks 0 and 52 (before and after diet). Then, we describe the results of all samples at Weeks 0, 12, 26 and 52. As shown in Figure 3, given a fix distance 0.2 (i.e. Pearson's correlation 0.8), there were three main clusters, where the unweighted arithmetic mean of distances within clusters were <0.2. The four lean controls were closely grouped together in one cluster that contained some obese samples at Week 0 and almost all the obese samples at Week 52 except one (I4W52). More than half of the obese samples at Week 0 were in the other two clusters. Our results showed that before diet only some but after diet almost all the obese samples clustered together with the four lean controls as the unweighted arithmetic mean of distances <0.2 (i.e. Pearson's correlation >0.8). Previous studies (Turnbaugh *et al.*, 2009; Zhang *et al.*, 2009) have also reported that the microbiomes were less diverse and more similar in lean than in obese human gut. Similar observations were found at the taxonomic levels of class, order, family and genus (Supplementary Fig. S4). However without using MetaRank, hierarchical clustering analysis of the estimated proportions or the ordinary ranks failed to show that most obese samples after diet were similar to and grouped together with the four lean controls (Supplementary Fig. S5).

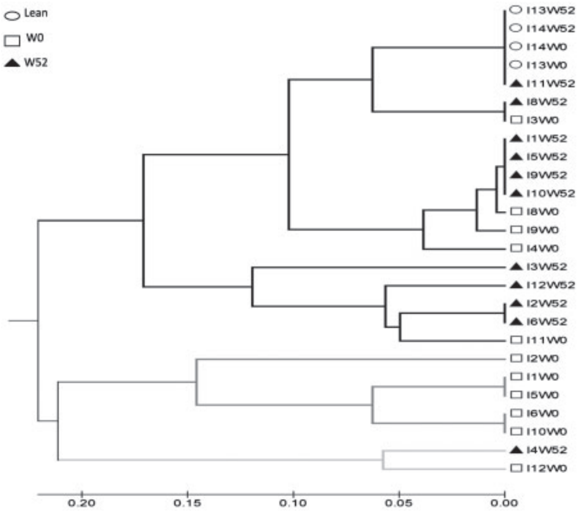

Fig. 3. The hierarchical clustering results of the ranks converted by MetaRank at the phylum level in 12 obese individuals at Week 0 (I1W0, I2W0, …, I12W0) and 52 (I1W52, I2W52, …, I12W52), including the four lean controls (I13W0, I14W0, I13W52 and I14W52), based on UPGMA. The hierarchical agglomerative clustering (bottom-up clustering) initially treated each sample as a single cluster at the bottom and then successively agglomerated pairs of nearest clusters until all clusters were merged into a single cluster at the top. Given a fix distance 0.2 (i.e. Pearson's correlation 0.8), there were three main clusters, where the unweighted arithmetic mean of distances within clusters were <0.2.

Close inspection on these obese samples at Week 52 and the lean controls (in the ranks converted by MetaRank in Supplementary Table S4) revealed their common characteristics of the central tendencies. The phyla *Firmicutes* and *Bacteroidetes* were the largest and second largest ranks; *Actinobacteria*, *Proteobacteria* or *Verrucomicrobia* was the third largest or smaller rank; the others were rare with the smallest ranks. Further comparison of these obese samples at Weeks 0 and 52 showed the differences between before and after diet. The converted rank of the phylum *Bacteroidetes* at Week 0 was smaller than the one at Week 52 (significantly with $P=0.015$ by t-test); the converted rank of *Actinobacteria* at Week 0 was larger than the ones at Week 52 ($P=0.029$). Therefore, using MetaRank not only grouped together most obese samples after diet with the lean controls in clustering analysis, but also clearly revealed their common traits as well as the differences between before and after diet.

In addition, considering these obese samples at Weeks 12 and 26 in clustering analysis of the ranks converted by MetaRank (Fig. S6), we still found the similar results that are consistent with the ones in Figure 2. The obese samples at Weeks 12 and 26 were less diverse and more similar to the lean controls than the ones at Week 0, but more diverse and less similar to the lean controls than the ones at Week 52 (Supplementary Fig. S6). As longer the individuals were treated on diet, more obese samples were grouped together with the lean controls in the cluster with unweighted arithmetic mean of distances <0.2. Using MetaRank revealed the alterations in obese human gut microbiomes with the treatment of diet. In microbial community compositions, we observed that the phylum *Bacteroides* moved up to the second largest rank; *Actinobacteria* moved down to the third largest or smaller ranks (Supplementary Table S4).

Since the ranks converted by MetaRank are close to the central tendencies with small deviations, using MetaRank in hierarchical clustering analysis of metagenomes helps to reveal the similarities and diversities of microbial community compositions.

3.3 MetaRank helps detecting differences between metagenomes

We also applied MetaRank to identify the discriminative features in the dataset of COG-functional compositions in four infant and nine adult samples (Kurokawa *et al.*, 2007). First, we used MetaRank to convert the raw abundances (Supplementary Table S5) into ranks (Supplementary Table S6) and then applied t-test to identify the differences between the infant and adult samples. We found that using MetaRank identified 41 COG-function features as rank-based differences between the two groups significantly with adjusted P-values <5% (Supplementary Table S7). Without using MetaRank, the same t-test identified 17 COG-function features as quantitative differences in proportions (see all statistics of the 3869 COG-functions in Supplementary Table S9). It is noted that only COG1808, COG4277 and COG5000 were both detected as rank-based and quantitative differences. The other 38 COG-functional features identified by MetaRank were not detectable in proportions by the same t-test. Since non-parametric and parametric methods are complementary to each other in statistics (no one can replace the other), we considered MetaRank as a useful rank-based (non-parametric) approach complementary to current quantitative (parametric) methods.

We further compared the 41 COG-functional features with 192 ones identified by Metastats (White *et al.*, 2009), a more advanced quantitative method than simple t-test in proportions. In Table 1, we list the 11 COG-function features detected by MetaRank, but not by Metastats (Supplementary Table S8). The other 30 COG-functions identified by both MetaRank and Metastats are listed in Supplementary Table S9. Moreover, the 11 COG-function features in Table 1 were consistent with Kurokawa *et al.*'s and other previous studies (Brink *et al.*, 1991; Greger *et al.*, 1989; Heijnen *et al.*, 1993; Kobayashi *et al.*, 1975; Kurokawa *et al.*, 2007; van de Graaf *et al.*, 2007; Ziegler and Fomon, 1983). Among the 11 COG-function features, COG780, COG1183, COG1468, COG-2088, COG2355, COG2361, COG5523 and COG5587 were identified more enriched in the adult samples than in the infant samples by both our approach and Kurokawa *et al.*'s; COG1135 and COG1414 were more enriched in the infant samples than in the adult samples (Kurokawa *et al.*, 2007). With regard to COG598 (Mg^{2+} and Co^{2+} transporters), other studies have verified that lactose enhanced the intestinal absorption of magnesium in infancy (Brink *et al.*, 1991; Greger *et al.*, 1989; Heijnen *et al.*, 1993; Kobayashi *et al.*, 1975; van de Graaf *et al.*, 2007; Ziegler and Fomon, 1983). Therefore, the 11 COG-function features identified by MetaRank but not by Metastats were supported by other previous studies with biological relevance.

Table 1. The 11 COG-function features identified by MetaRank with a simple t-test, but not detected by Metastats

COG	Definition	P-value[a]
598	Mg^{2+} and Co^{2+} transporters	0.0497
780	Enzyme related to GTP cyclohydrolase I	0.0033
1135	ABC-type metal ion transport system, ATPase component	0.0011
1183	Phosphatidylserine synthase	0.0043
1414	Transcriptional regulator	1.26×10^{-6}
1468	RecB family exonuclease	0.0046
2088	Uncharacterized protein, involved in the regulation of septum location	0.0015
2355	Zn-dependent dipeptidase, microsomal dipeptidase homolog	0.0025
2361	Uncharacterized conserved protein	0.0211
5523	Predicted integral membrane protein	0.0129
5587	Uncharacterized conserved protein	0.0061

[a] Bonferroni correction.

Therefore, a simple t-test with MetaRank was capable of a useful approach to detect the differences in ranks associated with environmental alterations. Complementary to the differences in proportions that are identified by quantitative methods, such as Metastats, the differences in ranks are also informative for understanding the relationship between microbial communities and habitats.

4 CONCLUSION

Most current statistical methods for comparative analysis of microbial community compositions rely on estimated abundances. However, when processing metagenomic data, sampling biases and systematic artifacts cause noisy deviations such that estimated abundances differ from true abundances (Ashelford *et al.*, 2005; Brady and Salzberg, 2009; Mavromatis *et al.*, 2007). In this study, we propose a rank conversion scheme, MetaRank, which converts highly abundant members into larger ranks. Since the ranks of highly abundant members are robust to small deviations, the ranks converted by MetaRank in samples are less affected by sampling biases. Empirical tests on synthetic samples and two real metagenomes confirmed that using MetaRank is able to reduce the variability between samples resulting from sampling biases, facilitates the comparative analysis of metagenomes, and helps to reveal the characteristics in common or the discriminative features with biological relevance. Therefore, MetaRank provides a useful rank-based alternative to analyze the microbial community compositions.

Some limitations of the proposed scheme should be mentioned. Since the proportions of rare members are too small to be distinguished from zero with statistical significance (nor from each other), MetaRank converts them into the same smallest rank. However, the same smallest rank provides no information for comparative analysis of rare members. As the rare members may play functional roles in microbial communities (Galand *et al.*, 2009), it is inappropriate to use ranks converted by MetaRank to explore rare microbiomes. However, since it is difficult to investigate rare members that are unobserved in samples but present in habitats, using general quantitative approaches to characterize rare microbiomes is also inappropriate. In addition, although MetaRank is able to mitigate the noisy effects of sampling biases, such rank-based methods have disadvantages as compared with quantitative approaches. In particular, there is a loss of information and the loss of ability to provide several valuable measures for statistical inference, e.g. estimated proportions, confidence intervals, effect sizes, etc. Hence, we consider MetaRank as a rank-based approach complementary to current quantitative methods.

In this study, the statistical techniques including binomial and multinomial tests, Pearson's correlation for measuring similarities and simple t-test to detect differences, represent the initial steps in the analysis of ranks converted by MetaRank. More advanced techniques such as the *Kendall Tau* coefficient and permutation tests can be considered for further implementation. In our future work, we will put efforts on developing more specific ways to estimate the relative orders in rare microbiomes.

ACKNOWLEDGEMENTS

We wish to thank Daryi Wang and Jen-Hao Cheng for his valuable suggestions and comments.

Funding: National Science Council of Taiwan under grant (NSC99-2627-B-001-003 to H.-K.T.).

Conflict of Interest: none declared.

REFERENCES

Ashelford,K.E. *et al.* (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.*, **71**, 7724–7736.

Biers,E.J. *et al.* (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl. Environ. Microbiol.*, **75**, 2221–2229.

Brady,A. and Salzberg,S.L. (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods*, **6**, 673–676.

Brink,E.J. *et al.* (1991) Inhibitory effect of dietary soybean protein vs. casein on magnesium absorption in rats. *J. Nutr.*, **121**, 1374–1381.

Cole,J.R. *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.

Dinsdale,E.A. *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature*, **452**, 629–632.

Galand,P.E. *et al.* (2009) Ecology of the rare microbial biosphere of the Arctic Ocean. *Proc. Natl Acad. Sci. USA*, **106**, 22427–22432.

Gifford,S.M. *et al.* (2011) Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J.*, **5**, 461–472.

Gomez-Alvarez,V. *et al.* (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J.*, **3**, 1314–1317.

Greger,J.L. *et al.* (1989) Interactions of lactose with calcium, magnesium and zinc in rats. *J. Nutr.*, **119**, 1691–1697.

Hamady,M. and Knight,R. (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res.*, **19**, 1141–1152.

Heijnen,A.M. *et al.* (1993) Ileal pH and apparent absorption of magnesium in rats fed on diets containing either lactose or lactulose. *Br. J. Nutr.*, **70**, 747–756.

Hugenholtz,P. and Tyson,G.W. (2008) Microbiology: metagenomics. *Nature*, **455**, 481–483.

Kobayashi,A. *et al.* (1975) Effects of dietary lactose and lactase preparation on the intestinal absorption of calcium and magnesium in normal infants. *Am. J. Clin. Nutr.*, **28**, 681–683.

Kosakovsky Pond,S. *et al.* (2009) Windshield splatter analysis with the Galaxy metagenomic pipeline. *Genome Res.*, **19**, 2144–2153.

Kristiansson,E. *et al.* (2009) ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics*, **25**, 2737–2738.

Kurokawa,K. *et al.* (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.*, **14**, 169–181.

Ley,R.E. *et al.* (2006) Microbial ecology: human gut microbes associated with obesity. *Nature*, **444**, 1022–1023.

Li,W. (2009) Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics*, **10**, 359.

Magurran,A.E. (1988) *Ecological Diversity and Its Measurement*. Princeton University Press, Princeton.

Markowitz,V.M. *et al.* (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.*, **36**, D534–D538.

Mavromatis,K. *et al.* (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods*, **4**, 495–500.

Mitra,S. *et al.* (2009) Visual and statistical comparison of metagenomes. *Bioinformatics*, **25**, 1849–1855.

Parks,D.H. and Beiko,R.G. (2010) Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, **26**, 715–721.

Rodriguez-Brito,B. *et al.* (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics*, **7**, 162.

Rusch,D.B. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, **5**, e77.

Turnbaugh,P.J. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

van de Graaf,S.F. *et al.* (2007) Physiology of epithelial Ca^{2+} and Mg^{2+} transport. *Rev. Physiol. Biochem. Pharmacol.*, **158**, 77–160.

White,J.R. *et al.* (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.*, **5**, e1000352.

Wooley,J.C. and Ye,Y. (2010) Metagenomics: facts and artifacts, and computational challenges. *J. Comput. Sci. Technol.*, **25**, 71–81.

Wooley,J.C. *et al.* (2010) A primer on metagenomics. *PLoS Comput. Biol.*, **6**, e1000667.

Zhang,H. *et al.* (2009) Human gut microbiota in obesity and after gastric bypass. *Proc. Natl Acad. Sci. USA*, **106**, 2365–2370.

Ziegler,E.E. and Fomon,S.J. (1983) Lactose enhances mineral absorption in infancy. *J. Pediatr. Gastroenterol. Nutr.*, **2**, 288–294.