Computational and Structural Biotechnology Journal 19 (2021) 3930–3937

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Beyond samples: A metric revealing more connections of gut microbiota between individuals

Zhen Yang [a], Feng Xu [a], Hongdou Li [b], Yungang He [a,*]

[a] Shanghai Fifth People's Hospital, and Shanghai Key Laboratory of Medical Epigenetics, International Co-laboratory of Medical Epigenetics and Metabolism (Ministry of Science and Technology), Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China
[b] Obstetrics Gynecology Hospital, The Institute of Reproduction and Developmental Biology, Fudan University, Shanghai, China

ARTICLE INFO

Article history:
Received 5 March 2021
Received in revised form 3 July 2021
Accepted 7 July 2021
Available online 10 July 2021

Keywords:
Gut microbiota
Metric
Beta diversity
Microbiome
Difference of finite sets

ABSTRACT

Studies of gut microbiota explore their complicated connections between individuals of different characteristics by applying different metrics to abundance data obtained from fecal samples. Although classic metrics are capable to quantify differences between samples, the microbiome of fecal sample is not a good surrogate for the gut microbiome of individuals because the microbial populations of the distal colon does not adequately represent that of the entire gastrointestinal tract. To overcome the deficiency of classic metrics in which the differences can be measured between the samples analyzed, but not the corresponding populations, we propose a metric for representing composition differences in the gut microbiota of individuals. Our investigation shows this metric outperforms traditional measures for multiple scenarios. For gut microbiota in diverse geographic populations, this metric presents more explainable data variance than others, not only in regular variance analysis but also in principle component analysis and partition analysis of biologic characteristics. With time-series data, the metric further presents a strong correlation with the time interval of serial sampling. Our findings suggest that the metric is robust and powerfully detects the intrinsic variations in gut microbiota. The metric holds promise for revealing more relations between gut microbiota and human health.

1. Introduction

The human intestinal tract harbors a huge and complex microbial ecosystem, termed the gut microbiota, which has gained considerable attention. Although in controversial, the number of microbes in the human gastrointestinal tract is estimated to be greater than the number of cells in the human body [1,2]. Microbial species in the gastrointestinal tract have mutualistic interactions involving the exchange of metabolic products to influence the body's metabolic phenotype [3]. Many environmental factors including diet [4], birth delivery method [5], breastfeeding [6], and antibiotic usage [7], play important roles in determining the composition of the gut microbiome. Alterations of the gut microbiota in the human gastrointestinal tract are linked to the risk for many diseases, such as diabetes mellitus [8,9], Crohn's disease [10,11], and colorectal cancer [12–14]. Therefore, it is important to study and develop methods for comparing microbial structures between individuals and evaluate their associations with different factors.

Advances in high-throughput sequencing have made it possible to capture the microbiome composition in biologic samples. Due to the lower cost and less computational intensive analysis, amplicon sequencing is still the routine method used for studying the microbiome [15]. The amplicon approach proceeds by amplifying taxonomic marker genes of the microbiota, mainly ribosomal DNAs including prokaryotic 16s, eukaryotic 18s, and fungal ITS genes [16]. The amplicon sequences are then clustered into operational taxonomic units (OTUs) according to their sequence similarity [17]. The OTU profiles or higher-level taxa with respect to different biological or clinical factors are compared to illustrate the relationship between phenotypes and microbial dysbiosis [18,19]. Many metrics, such as Bray-Curtis, Jaccard or phylogenetic based UniFrac distances have been applied to quantify variations of the composition between microbiome samples, which is also refer to as beta-diversity [20,21]. Metrics are usually employed in biological studies under the assumption that the samples well represent their populations. This assumption holds for many applications, and the metrics work well in these situations. For studies of the gut

* Corresponding author.
E-mail address: heyungang@fudan.edu.cn (Y. He).

https://doi.org/10.1016/j.csbj.2021.07.009

microbiome, however, this assumption does not necessarily hold true [22–24]. The gut microbiota has been presented differences between gut segments, and also the stool samples from the same individual. Many factors such as acidity, transit time and microbial biomass vary greatly along the entire intestinal tract, and these differences have great effects on the microbial population [25,26]. Therefore, it is important to develop a novel metric that reliably measures differences in the gut microbiomes of individuals even when poorly representative fecal samples are obtained.

In this report, we introduce a metric that measures the differences of the gut microbiota between individuals. This novel metric is developed to measure the difference of paired original populations but not only the difference of their sampling subsets. We collected multiple fecal microbiome data sets, which include those from different geographic regions [27], or with different demographic factors, such as age and breast feeding [28]. We demonstrate the new metric outperforms other classic measures in presenting more connections between individual gut microbiomes, it could better capture the data variation by using dimension reduction, and also present better partition features for individuals from different groups. Our results indicate that this metric is a good alternative to the classic choices for investigating the microbial diversity of the gastrointestinal tract.

2. Materials & methods

2.1. Review of distances for similarities in the microbial community

Several distances of beta diversity measures were recently introduced, and are mainly classified as phylogenetic-based and non-phylogenetic-based approaches. Among the non-phylogenetic-based distances, the Jaccard distance is a qualitative measure that utilizes the presence-absence data of the species. Let S_i, S_j, and S_{ij} denote the number of species (OTUs) present in Sample i, j, and in both i and j samples, respectively, the Jaccard distance between Sample i and j is represented as:

$$Jac_{(i,j)} = \frac{S_i + S_j}{S_i + S_j + S_{ij}}$$

In contrast, the Bray-Curtis distance is a quantitative measure that uses the species abundance information for each sample. Let $S_{(A,i)}$ and $S_{(B,i)}$ be the counts of the i_{th} species (OTUs) in Sample A and B, respectively, and the Bray-Curtis distance between Sample i and j is represented as:

$$BC_{(A,B)} = 1 - 2\frac{\sum \min\left(S_{A,i}, S_{B,i}\right)}{\sum S_{A,i} + \sum S_{B,i}}$$

It should be noted that the Jaccard distance is equivalent to the Bray-Curtis distance when only presence-absence is considered [29].

The phylogenetic-based distances use the evolutionary information of representative sequences to compare whether the samples exhibit significant differences in the microbial community in a particular evolutionary lineage. The unweighted and weighted UniFrac are two representative phylogenetic distance measures. UniFrac measures the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from one sample or the other. Consider a rooted phylogenetic tree with n branches and 2 microbiome communities A and B. Let b_i be the length of the branch i and p_i^A, p_i^B the taxa proportions descending from the branch i for community A and B, respectively. The unweighted UniFrac is defined as

$$D(u) = \sum_{i=1}^{n} \frac{b_i |I(p_i^A > 0) - I(p_i^B > 0)|}{\sum_{i=1}^{n} b_i}$$

where I(.) is the indicator function and only the presence/absence of species of branch i is used. The distance ignores the taxa abundance information [30]. In contrast, the (normalized) weighted UniFrac distance weights the branch length with the species abundance difference and is defined as

$$D(w) = \frac{\sum_{i=0}^{n} b_i |p_i^A - p_i^B|}{\sum_{i=0}^{n} b_i (p_i^A + p_i^B)}$$

The consequence of using the absolute difference is that the value of $D_{(w)}$ is mainly determined by branches with large proportions and is less sensitive to the abundance changes on the branches with small proportions [31]. Both of these measures assign too much weight to either rare lineages or highly abundant lineages, which can lead to loss of power when an important composition change occurs in moderately abundant lineages. Thus, some revised measures, such as the generalized version of the UniFrac distance to address the limitations of the traditional UniFrac distance, have been proposed [32]. Using different distance measures to summarize the overall microbiota variability provides more insight into the source of microbiota variability.

2.2. Quantifying differences in finite sets

To overcome the deficiency of traditional metrics in which the differences can be measured between the samples analyzed, but not the populations, we propose a metric to quantify the differences between the populations. Given 2 finite sets, A and B, subsets A' and B' are independently generated from sets A and B by sampling n times with replacement. The sampling probabilities of the elements are unknown and further differ within and between the sets. The intersection of A' and B' is defined as C'.

Let us set O as the union of A and B. We assume that elements of O appear in A' or B' with an overall probability p. Then, the size of C' is $S_{C'} = (S_{A'} + S_{B'})p^2/2$, where $S_{A'}$ and $S_{B'}$ are the sizes of A' and B', respectively. We estimate p as $\hat{p} = \sqrt{2S_{C'}/(S_{A'} + S_{B'})}$ (Fig. 1).

Let q be the probability that an element of O does not appear in A' or B' in each sampling. The probability q increases as the size of O increases. In other words, q is larger if the intersection of A and B is smaller. Therefore, q represents the scale of difference between A and B in (0,1). We defined a distance $d = q/(1-q)$ to represent the scale of difference in $(0, \infty)$.

After n times sampling, we have the probability $q^n = 1 - p$ (Eq. (1)). When n is large, Eq. (1) can be rewritten as $d = -n/\log(1-p) - 1$. To quantify the difference between sets A and B, we propose a new distance of finite sets (DFS) as $DFS = -1/\log[1 - \sqrt{2S_{C'}/(S_{A'} + S_{B'})}]$ (Eq. (2)). It is trivial to show that DFS satisfies the triangle inequality. We note that the Jaccard distance (Tanimoto distance) $J = 1 - S_{C'}/(S_{A'} + S_{B'} - S_{C'})$ measures the difference between subsets A' and B', but not the distance between finite sets A and B.

In the present study, we apply the DFS on OTU data of the gut microbiota to explore the similarity of gut microbiota between individuals. For OTU sets A and B of the gut microbiota of paired individuals, $S_{A'}$ and $S_{B'}$ are the numbers of OTUs identified from normalized sequencing data of fecal samples, respectively. $S_{C'}$ is the number of shared OTUs between the normalized data obtained from fecal samples.

2.3. Real data analysis

2.3.1. Intestinal microbiome dataset from different geographic regions

We used the amplicon sequencing data from Yatsunenko et al., in which microbiomes were characterized from fecal samples of 531 individuals, including 100 from Venezuela, 111 from Malawi, and 316 from US metropolitan areas in different age groups [27].

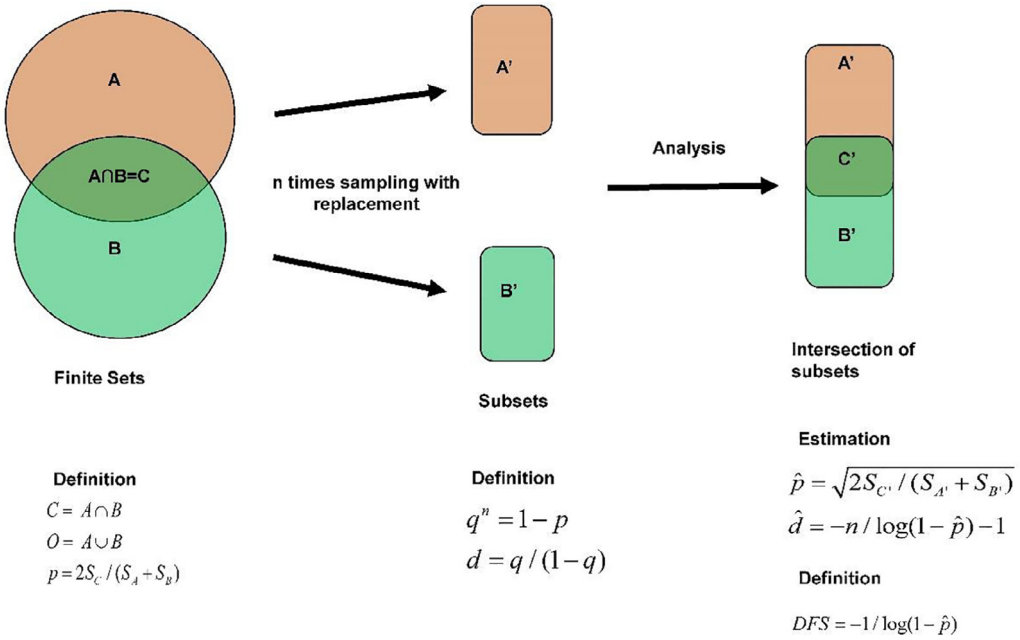

Fig. 1. The DSF shows the differences in the original finite sets, but not in their subsets. The OTUs are recognized as members of finite sets A and B for gut microbiota of two individuals, respectively. Sequence reads were obtained from each fecal sample. Then the subsets of OTUs are discovered in the fecal sample A (subset A′) and in the sample B (subset B′). The probability that the members of original OTUs appeared in the two subsets was estimated as $\hat{p}$, then the distance of the two finite OTU sets of the individuals can be obtained as $DFS = -1/\log(1-\hat{p})$.

This study was performed using variable region 4 (V4) of bacterial 16s ribosomal RNA genes, and the amplicons were sequenced based on Illumina HiSeq 2000 instrument. We obtained the dataset from the NCBI-SRA (Accession: PRJEB3079) and data processing was performed using the software tools of QIIME suite (version 1.9.1) [33]. Briefly, the closed-reference OTU picking method was used to obtain OTUs at the threshold of 97% similarity to the reference sequence. Taxonomy assignments to OTUs were designated using the Greengenes database (version 13.8) [34]. OTUs that present low frequencies (<10,000 across all the samples) were filtered out. Then, the OTU table was rarefied to the minimum value across all the samples. Finally, we obtained an OTU table with 2301 OTUs and 526 samples, and all communities were rarefied to 306,155 sequence reads per sample. Next, we calculated the beta-diversity matrices using the distance we designed, and for comparison purposes, other distances of the Euclidean, Bray-Curtis, and Jaccard measurements were also calculated. We further obtained the weighted and unweighted UniFrac distances calculated by the QIIME suite. To avoid potential bias due to the relative short reads of the HiSeq sequencing platform when performing multiple sequence alignment, we used full-length sequences of the 16s V4 region from the reference database as the representative sequence of each OTU. We also calculated the taxonomic relative abundances from phylum to genus levels based on the rarefied OTU tables.

2.3.2. Intestinal microbiome dataset from multiple time points

To test the performance of our metric on gut microbiome data collected from multiple time points within a short period, we obtained the 16s rRNA dataset from The Integrative Human Microbiome Project (HMP2) [35,36]. This dataset was processed through the HMP DACC QIIME pipeline. The OTU table was downloaded from the Human Microbiome Bioactives Resource Portal (https://portal.microbiome-bioactives.org/). This datasets consists of replicate samples collected at different time points for individuals. Only the participants with at least 4 replicate samples were used for our study. In this case, a total of 728 samples from 35 individuals were included. Then the OTU table was rarefied to the minimum value across all samples for further analysis.

2.3.3. Intestinal microbiome dataset of daily sampling for months

The proportions of the microbial taxa generally remain stable for around short time, but the microbiome composition can be altered over time by dietary changes as well as by antibiotics. To test the association performance of the distance matrices with the time interval, we used the dataset of Caporaso *et al*., which includes two healthy individuals who were sampled daily at three body sites (gut, mouth, and skin) for months [28]. Variable region 4 (V4) of 16s rRNA genes present in each community was sequenced using the Illumina Genome Analyzer IIx (SRA Accession: ERP021896). The same procedures were used for OTU picking and distance calculation. Here we only used the fecal sample data obtained from the male individual across 442 days. An OTU table with 5071 OTUs and 332 samples was finally obtained.

3. Results

3.1. More explainable variance in the top ranked dimensions

To demonstrate the utility of our *DFS*, we applied it to the amplicon sequencing data of the 16s rRNA genes of the gut microbiota published by Yatsunenko *et al*., which comprised 305 adults and 117 infants from Malawi, Venezuela, and the United States [27]. Principle Component Analysis (PCA) showed that *DFS* represents more variance in the 1st and 2nd principle component (PC) than the other common choices, such as weighted or unweighted UniFrac distance, Bray-Curtis dissimilarity, Jaccard distance, and Euclidean distance (Fig. 2). The 1st PC accounts for 86.76% of the total variance by *DFS*, while the 2nd PC accounts for another 8.8%. More than 95.58% of the total variance is represented in the 2-dimensional space. This suggests that the dimension reduction causes only a minor loss of information. A matrix of Euclidean distance can also be presented well on the 1st PC (85.05% of total variance), but with only 5.16% of the total variance on the 2nd PC. More

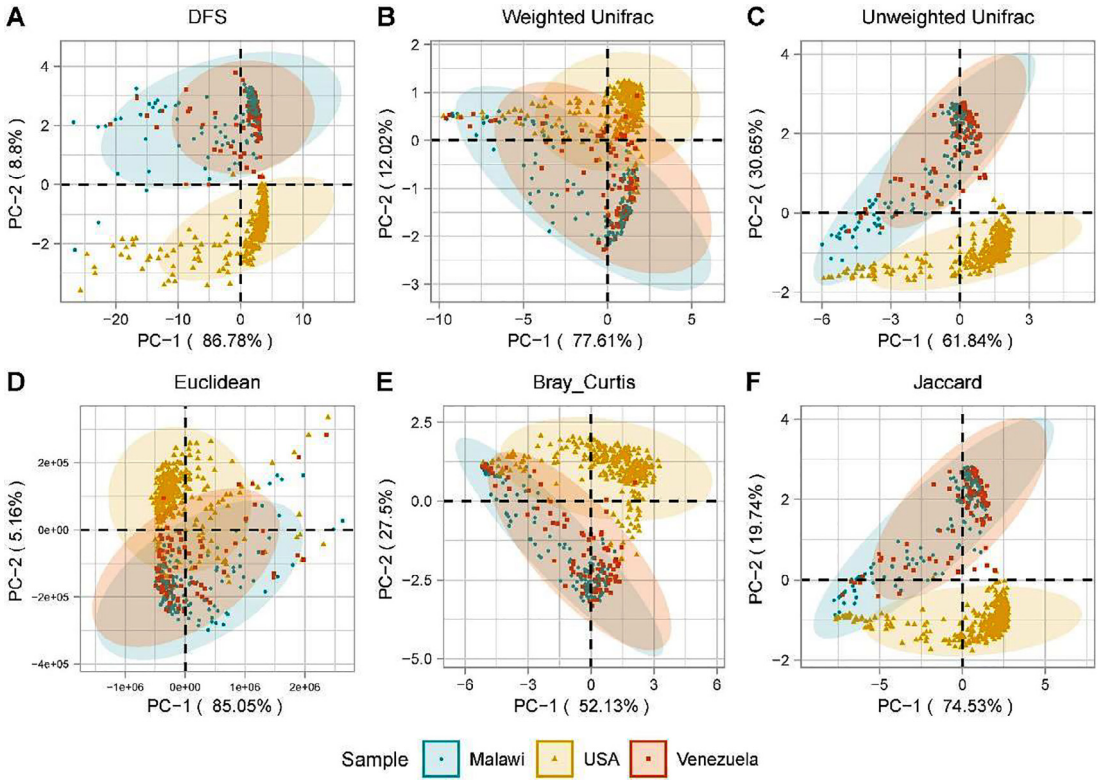

Fig. 2. Comparison of different metrics for clustering individuals from different geographic regions. Principle component analysis was performed on the matrices of A) *DFS*; B) Weighted UniFrac; C) Unweighted UniFrac; D) Euclidean; E) Bray-Curtis; and F) Jaccard. The first 2 principal components (PC-1 and PC-2) are represented by the x and y axes, respectively. Each data point is a microbial community sample colored by its geographic origin.

importantly, in the PC plots, *DFS* separates the Malawi and Venezuela clusters from that of the United States better than the other 5 distance matrices. There are some significant mix-ups between the 2 clusters in the PC plot of Euclidean distance. These mix-ups are highly likely to occur because the 2nd PC of the Euclidean distance accounts for less variance in the data. We performed the same analysis on separate data of the adult and infant groups to further demonstrate the robustness of statistic *D*. The results demonstrated that the *DFS* performs the best of the distance measures (Supplementary Figs. 1-2).

We further validate the excellent performance of *DFS* by employing the same analysis on the additional HMP2 data (see Method for details). The PCA analysis showed that the 1st PC accounts for 64.47% of the total variance by the *DFS*, while the 2nd PC accounts for another 11.89%. The *DFS* has a comparable performance with the Euclidean metric in dimension reduction, and they represent more variance in the 1st and 2nd principle component than the other metrics (Supplementary Fig. 3). Further analysis suggest that Euclidean metric is poor in representing variation of gut microbiota among individuals (see results below). Therefore, the *DFS* is the best choice among the different metrics in dimension reduction.

3.2. Better reflection of the data variance between the groups

As the PCA analysis indicated that top components have a good reflection of samples by their geographic sources, we then used a permutational multivariate analysis of variance (PERMANOVA) to evaluate the geographic factor and all the distance matrixes by using the data from Yatsunenko *et al.*, [37]. Our analysis showed that the geographic factor significantly contributes to all the different metrics (p-value < 0.001). The *DFS* provides the best separation on the basis of the geographic information, while the Unweighted UniFrac distance ranked second. The variance of the Unweighted UniFrac distance that is explained by the geographic factor was only 0.79, 0.47, and 0.82 that of the *DFS* for adults, infants, and all of them together, respectively. Further, the performance of the Weighted UniFrac distance was worse than the Unweighted UniFrac distance. The explainable variance of the weighted UniFrac distances of infants was only 0.17 that of the *DFS* (Table 1, data group 1–3). The poorly explained variance suggests that the abundance of OTUs is less informative than the diversity in studies of gut microbiota. The results of the PERMANOVA show that compared with the other distance measures, the *DFS* more effectively reflects the information of the environment.

To further demonstrate the capability of identifying small differences of gut microbiota between individuals, we employed PERMANOVA on 16s rRNA data of replicate samples of individuals from HMP2 project (see Method for details). Our results confirmed that the *DFS* is the best one among the involving metrics with a F statistic 24.51 (Table 1, data group 4). Bray-Curtis dissimilarity index was ranked as the NO.2 metric which had explainable variance only 0.85 that of the *DFS*. In this case, Weighted UniFrac distance and Jaccard distance accounted explainable variance as low as 0.49 that of the *DFS*. The *DFS* is the best choice to discover composition difference of the microbial populations among individuals.

3.3. Better partitioning to groups based on demographic factors

Compared with adult, the gut microbiome is more dynamic and unstable for neonates, one of the primary source that driving microbiota fluctuations is breast and/or formula feeding [38]. We applied our *DFS* to investigate the effects of breastfeeding on infant gut microbiota. This investigation involved sequencing data of 16s

Table 1
PERMANOVA indicates that DFS better reflects the geographic factors and individual differences.

	All sample		Adult		Infant		HMP2	
	F statistic	Explained	F statistic	Explained	F statistic	Explained	F statistic	Explained
DFS	92.99	1	81.67	1	29.35	1	24.51	1
UW UniFrac	76.39	0.82	64.65	0.79	13.81	0.47	18.76	0.77
W Unifrac	47.34	0.51	46.93	0.57	4.92	0.17	12.13	0.49
Euclidean	25.5	0.27	21.84	0.27	4.03	0.14	17.12	0.7
Bray-Curtis	45.32	0.49	32.24	0.39	6.48	0.22	20.79	0.85
Jaccard	57.01	0.61	53.77	0.66	9.42	0.32	12.13	0.49

rDNA of 39 US twins under the age of 1 year whose feeding information was available. A parsimony score was employed on neighbor joining (NJ) trees to evaluate the partitions between the infants who were breast-fed or formula-fed (Fig. 3A). The smaller the score, the better partition achieved. Our investigation shows a parsimony score of only 2 for the NJ tree of *DFS*. On the other hand, it was as high as 5 and 13 for the NJ trees of the unweighted UniFrac distance and weighted UniFrac distance measures, respectively (Supplementary Fig. 4). The result suggests that the *DFS* was superior to others for identifying different feeding types. We further identified that *Acinetobacter*, *Bifidobacterium*, and *Klebsiella* were among the top-ranked genera related to breast/formula feeding (Supplementary Table 1). This observation was supported by several other studies. For instance, *Bifidobacterium* was found to be abundant in the gut of breast milk-fed full-term infants than that of formula fed infants without probiotic supplements [39], whereas *Klebsiella* was more abundant in the formula-fed group [40].

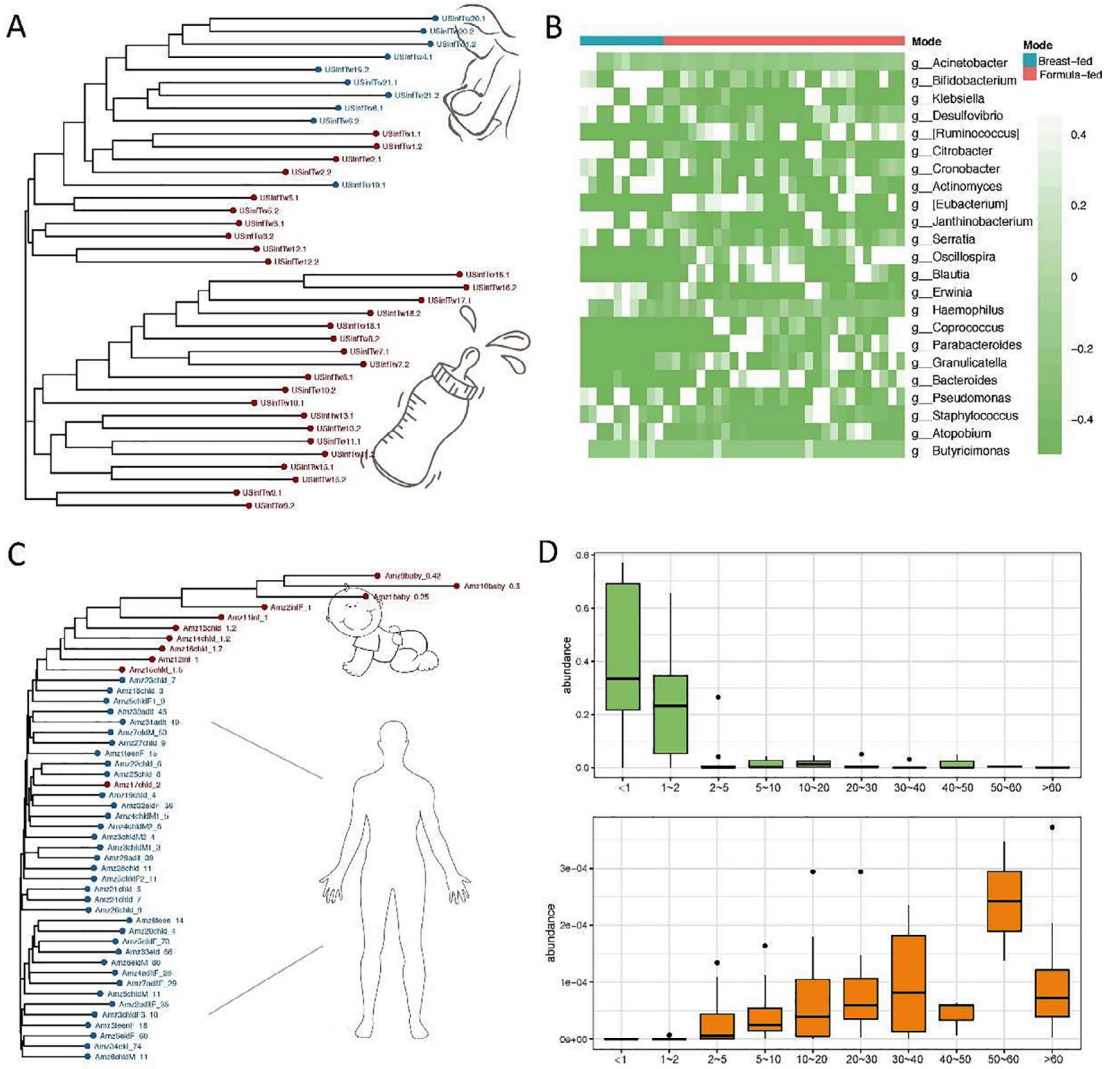

Fig. 3. DFS provides better partitioning for phylogeny reconstruction. A) A neighbor-joining tree clearly separates the breast-feeding and formula-feeding samples. B) Microbiota have major genera differences between the breast-feeding and formula-feeding samples. C) A neighbor-joining tree separates samples from different ages. D). *Dehalobacterium* (Upper green plot, FDR: 6.67E-13) and *Bifidobacterium* are major microorganisms differentially present across different ages (Lower orange plot, FDR: 3.35E-13).

Studies have indicated that the composition and diversity of gut microbiota presents age-related change pattern [41]. Here we employed the *DFS* to investigate differences in the microbial composition of Venezuelans from newborn to 82 years of age. We constructed NJ trees for the samples using the matrixes of different distance measures. The NJ tree of *DFS* clearly separated individuals under two years of age from older individuals, except for one of them (Fig. 3C). The result was supported by multiple studies that observed major shifts in the microbiota population between children two years of age and older individuals [42]. Our result is comparable with those using the unweighted UniFrac distance and Jaccard distance measures, and superior to the other measures (Supplementary Fig. 5). We further identified *Dehalobacterium*, *Ruminococcus*, and *Methanobrevibacter* are among the top-ranked genera using the Spearman rank correlation for individual age and abundance of genera with a false discovery rate <0.05 (Fig. 3D).

3.4. Strong correlation with the time interval of serial samplings

To evaluate the ability of *DFS* to discriminate fluctuations in the microbiota over period of time, we calculated the distance measures on serial data of 332 fecal samples that were collected from the same individual across 442 days [28]. Pearson's correlation coefficient showed a strong correlation between the *DFS* values and time intervals of serial samplings for any given sample and all of the others (Fig. 4). Most of the coefficients (97.3%) were larger than or equal to 0.3, and 37.9% of them were larger than or equal to 0.6. The performance was comparable to that of the other frequency-independent distances, such as the Unweighted UniFrac distance and Jaccard distance measures. In contrast, for frequency-dependent distances (i.e., Weighted UniFrac distance, Euclidean distance, and Bray-Curtis distance), the majority of the coefficients were <0.3 (95.5%, 95.1%, and 71.1%, respectively). Further, PCA analysis presents 2 potential clusters of serial samples for the frequency-independent distances but not frequency-dependent metrics (Supplementary Fig. 6). The *DFS*, as well as the other frequency-independent distance measures, may be less sensitive to the noise fluctuations of microbiota abundance. This may explain why the *DFS* outperform the frequency-dependent metrics with the variability of an individual's gut microbiota across years, weeks, and even days.

4. Discussion

Microbiome data are essentially count based abundance data in their original form. Due to their high dimensionality, phylogenetic constraints among species/OTUs and excessive zeros, it is statistically challenging to analyze their data structures. To address the problem, microbiome data are often measured as pairwise dissimilarity matrix, which used to test the association of microbiome composition with environmental factors [43–45]. For studies of the gut microbiome, fecal sample was usually used as surrogate due to the ease of obtaining. Unfortunately, it has increasing recognized that fecal microbial populations may not be fully representative of those in gastrointestinal tract. Stool is useful in analyzing the microbial populations of the distal colon, it does not adequately represent the entire gastrointestinal tract. On one hand, regional differences in gut microbial populations was observed along the rostral to caudal due to the functional heterogeneity of gastrointestinal tract segments [46,47], on another hand, storage, transportation and handling methods of the fecal samples may greatly influence the microbiome composition [48–50]. Due to the inherent instability variable feature of the microbiome data obtained by using fecal samples, it is often difficult to capture reliable information about the ecological structures of the gut microbiota for a given population by using traditional beta-diversity metrics [51]. Here we propose a novel beta-diversity measures of gut microbiota. By applying it to microbiome data from fecal samples, we show that our method is robust and powerful in detecting the ecological structure from populations. Our method is general and can be applied to microbiome data from different scenarios, such as detecting microbial composition structures driven by

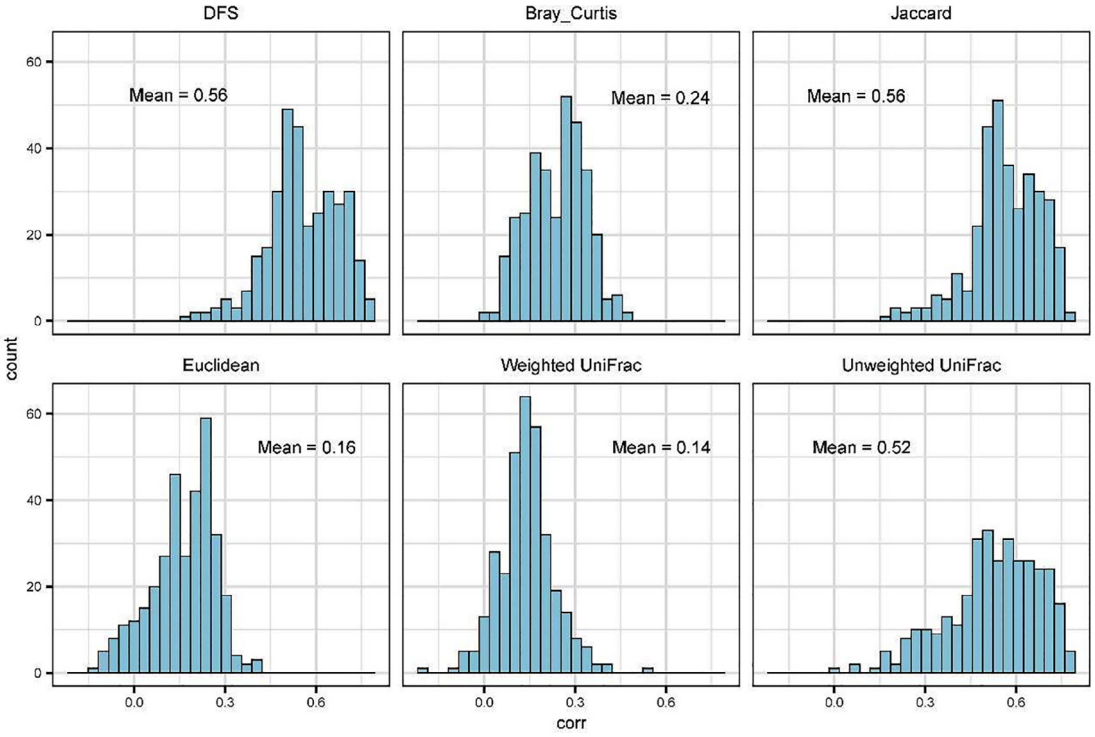

Fig. 4. Histogram of Pearson correlation coefficients for paired samples with different distances and time intervals.

geographic or aging factors. Our method is also superior to other metrics for more explainable data variance by using dimensionality reduction methods.

We used PCA analysis for dimension reduction in the present study. The *DFS* presents more explainable variance on the 1st and 2nd dimension in the PCA analysis, which means that less information is lost in the dimension reduction. In addition, *DFS* has another significant advantage in evolutionary studies. Given a distance matrix of 4 taxa, the phylogeny can be shown without information loss in a quartet network with 6 length parameters. Although the phylogenetic relationship is indicated in a quartet tree where only 5 length parameters are allowed, information will be lost due to the transformation. There is, however, an exception. Information loss is not going to occur if the distances between 4 taxa can be fully presented in a 2-dimensional linear space. In other words, we expect to lose less information in a *DFS* matrix in the phylogeny reconstruction than in the other distance matrixes. Our results further show excellent performance of *DFS* in phylogeny reconstruction. Other ordination methods (PCoA, NMDS) can also be employed on *DFS* to rearrange the samples in a low-dimensional space (Supplementary Fig. 7), while the methods aim to present the relation, but not to maximize the variance represented by the PCs.

It is crucial to thoroughly present the connections of gut microbiota between individuals. For the numerous distance measures now available to address this issue, one major type of measure depends only on the variation of the components between microbiome samples. Qualitative measures like Sorensen and Jaccard distances use the presence/absence (binary) data to compare the community. In contrast, quantitative measures such as Bray-Curtis distance and Kullbach-Leibler divergence take the relative abundance of each type of organism into account [29,52]. The limited range of the classic distances (i.e., 0 to 1) is a significant drawback in presenting the connection of the gut microbiota of many individuals. Studies of gut microbiota require a good measurement that is additive with a range from zero to infinity. The additive feature allows for less information loss and better reconstruction when the relation tree of the samples is reconstructed with the measurements [53]. *DFS* has a range from zero to infinity. Our results show the *DFS* leads to less information loss and performs well in tree reconstruction (Figs. 2, 3). We therefore propose *DFS* as a good replacement for the aforementioned classic measures in studies of gut microbiota.

The phylogeny based UniFrac distances are other major choices for the quantification of distances of fecal samples. The UniFrac distances, however, are easily influenced by the tree topology of OTUs [54]. Moreover, the unweighted UniFrac distance is highly sensitive to sequencing depth and rarefaction instances with no clear structure or separation between groups [55]. Further, the calculation of UniFrac distances is far more complicated than the other distances, and calls for the development of a novel dissimilarity measure to overcome the drawbacks. Our *DFS* distance is not influenced by the tree topology of the OTUs, and it is easy to obtain using straightforward algebra (Eq. (2)). Furthermore, our results show the *DFS* performs better than UniFrac distances in multiple datasets. *DFS* is an excellent alternative to represent the complicated connections of gut microbiota between individuals.

5. Conclusions

The *DFS* is a unique metric. It was developed to measure the differences in finite sets or individuals, but not subsets or biologic samples. In studies of gut microbiota, this unique metric can lead us to inspect more connections between individuals, and not only those of fecal samples. Our study demonstrated many advantages of this unique metric on multiple datasets using different scenarios. In future, it is important to compare the different metrics in studies of fungal diversity using 18s rRNA and ITS sequencing data of replicate samples from different sources. The usefulness of other metrics should be evaluated in future studies. It is highly unlikely that a single metric can perform well for all research purposes in all scenarios. Further, we strongly suggest caution when interpreting *DFS* analysis results, and the properties of *DFS* metric must be thoroughly considered.

Funding

This work was supported by grants from the National Natural Science Foundation of China (Grant No. 31871255 and 91731310 to Y.H.; No. 91959106 to Y.Z.). Y.H. was also supported by Shanghai Municipal Science and Technology (grant 2017SHZDZX01).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We thank the innovative research team of the High-level Local University in Shanghai for their support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.07.009.

References

[1] Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, et al. An estimation of the number of cells in the human body. Ann Hum Biol 2013;40(6):463–71.
[2] Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol 2016;14(8):e1002533.
[3] Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. Host-bacterial mutualism in the human intestine. Science 2005;307(5717):1915–20.
[4] De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci U S A 2010;107(33):14691–6.
[5] Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci U S A 2010;107(26):11971–5.
[6] Harmsen HJM, Wildeboer–Veloo ACM, Raangs GC, Wagendorp AA, Klijn N, Bindels JG, et al. Analysis of intestinal flora development in breast-fed and formula-fed infants by using molecular identification and detection methods. J Pediatr Gastroenterol Nutr 2000;30(1):61–7.
[7] Gasparrini AJ, Crofts TS, Gibson MK, Tarr PI, Warner BB, Dantas G. Antibiotic perturbation of the preterm infant gut microbiome and resistome. Gut Microbes 2016;7(5):443–9.
[8] Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012;490(7418):55–60.
[9] Zhou W, Sailani MR, Contrepois K, Zhou Y, Ahadi S, Leopold SR, et al. Longitudinal multi-omics of host-microbe dynamics in prediabetes. Nature 2019;569(7758):663–71.
[10] Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 2006;55(2):205–11.
[11] Dicksved J, Halfvarson J, Rosenquist M, Järnerot G, Tysk C, Apajalahti J, et al. Molecular analysis of the gut microbiota of identical twins with Crohn's disease. ISME J 2008;2(7):716–27.
[12] Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 2014;10(11):766.

[13] Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun 2015;6(1):6528.
[14] Yu J, Feng Q, Wong SH, Zhang D, Liang Qy, Qin Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut 2017;66(1):70–8.
[15] Wooley JC, Godzik A, Friedberg I, Bourne PE. A primer on metagenomics. PLoS Comput Biol 2010;6(2):e1000667.
[16] Tkacz A, Hortala M, Poole PS. Absolute quantitation of microbiota abundance in environmental samples. Microbiome 2018;6(1):110.
[17] Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009;75(23):7537–41.
[18] Cho I, Yamanishi S, Cox L, Methé BA, Zavadil J, Li K, et al. Antibiotics in early life alter the murine colonic microbiome and adiposity. Nature 2012;488(7413):621–6.
[19] Alekseyenko AV, Perez-Perez GI, De Souza A, Strober B, Gao Z, Bihan M, et al. Community differentiation of the cutaneous microbiota in psoriasis. Microbiome 2013;1(1):31.
[20] Barwell LJ, Isaac NJ, Kunin WE. Measuring beta-diversity with species abundance data. J Anim Ecol 2015;84(4):1112–22.
[21] Swenson NG, Hector A. Phylogenetic beta diversity metrics, trait evolution and inferring the functional beta diversity of communities. PLoS ONE 2011;6(6):e21264.
[22] Altomare A, Putignani L, Del Chierico F, Cocca S, Angeletti S, Ciccozzi M, et al. Gut mucosal-associated microbiota better discloses inflammatory bowel disease differential patterns than faecal microbiota. Dig Liver Dis 2019;51(5):648–56.
[23] Rangel I, Sundin J, Fuentes S, Repsilber D, de Vos WM, Brummer RJ. The relationship between faecal-associated and mucosal-associated microbiota in irritable bowel syndrome patients and healthy subjects. Aliment Pharmacol Ther 2015;42(10):1211–21.
[24] Ouwehand AC, Salminen S, Arvola T, Ruuska T, Isolauri E. Microbiota composition of the intestinal mucosa: association with fecal microbiota? Microbiol Immunol 2004;48(7):497–500.
[25] Leite GGS, Weitsman S, Parodi G, Celly S, Sedighi R, Sanchez M, et al. Mapping the segmental microbiomes in the human small bowel in comparison with stool: a REIMAGINE study. Dig Dis Sci 2020;65(9):2595–604.
[26] Hillman ET, Lu H, Yao T, Nakatsu CH. Microbial ecology along the gastrointestinal tract. Microbes Environ 2017;32(4):300–13.
[27] Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature 2012;486(7402):222–7.
[28] Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. Genome Biol 2011;12(5):R50.
[29] Anderson MJ, Crist TO, Chase JM, Vellend M, Inouye BD, Freestone AL, et al. Navigating the multiple meanings of beta diversity: a roadmap for the practicing ecologist. Ecol Lett 2011;14(1):19–28.
[30] Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 2005;71(12):8228–35.
[31] Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol 2007;73(5):1576–85.
[32] Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 2012;28(16):2106–13.
[33] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010;7(5):335–6.
[34] DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 2006;72(7):5069–72.
[35] Schiffer L, Azhar R, Shepherd L, Ramos M, Geistlinger L, Huttenhower C, et al. HMP16SData: efficient access to the human microbiome project through bioconductor. Am J Epidemiol 2019;188(6):1023–6.
[36] Integrative HMPRNC. The integrative human microbiome project. Nature. 2019;569(7758):641–8.
[37] Tang Z-Z, Chen G, Alekseyenko AV. PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances. Bioinformatics 2016;32(17):2618–25.
[38] Wang Z, Neupane A, Vo R, White J, Wang X, Marzano SL. Comparing Gut microbiome in mothers' own breast milk- and formula-fed moderate-late preterm infants. Front Microbiol 2020;11:891.
[39] Lee SA, Lim JY, Kim B-S, Cho SJ, Kim NY, Kim OB, et al. Comparison of the gut microbiota profile in breast-fed and formula-fed Korean infants using pyrosequencing. Nutr Res Pract 2015;9(3):242–8.
[40] Bezirtzoglou E, Tsiotsias A, Welling GW. Microbiota profile in feces of breast- and formula-fed newborns by using fluorescence in situ hybridization (FISH). Anaerobe 2011;17(6):478–82.
[41] Kim S, Jazwinski SM. The gut microbiota and healthy aging: a mini-review. Gerontology 2018;64(6):513–20.
[42] Avershina E, Storrø O, Øien T, Johnsen R, Pope P, Rudi K. Major faecal microbiota shifts in composition and diversity with age in a geographically restricted cohort of mothers and their children. FEMS Microbiol Ecol 2014;87(1):280–90.
[43] Zhang J, Wei Z, Chen J. A distance-based approach for testing the mediation effect of the human microbiome. Bioinformatics 2018;34(11):1875–83.
[44] Plantinga AM, Chen J, Jenq RR, Wu MC. pldist: ecological dissimilarities for paired and longitudinal microbiome association analysis. Bioinformatics 2019;35(19):3567–75.
[45] Zhao Ni, Chen J, Carroll I, Ringel-Kulka T, Epstein M, Zhou H, et al. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. Am J Hum Genet 2015;96(5):797–807.
[46] Martinez-Guryn K, Leone V, Chang EB. Regional diversity of the gastrointestinal microbiome. Cell Host Microbe 2019;26(3):314–24.
[47] Stanley D, Geier MS, Chen H, Hughes RJ, Moore RJ. Comparison of fecal and cecal microbiotas reveals qualitative similarities but quantitative differences. BMC Microbiol 2015;15:51.
[48] Ezzy AC, Hagstrom AD, George C, Hamlin AS, Pereg L, Murphy AJ, et al. Storage and handling of human faecal samples affect the gut microbiome composition: a feasibility study. J Microbiol Methods 2019;164:105668.
[49] Ott SJ, Musfeldt M, Timmis KN, Hampe J, Wenderoth DF, Schreiber S. In vitro alterations of intestinal bacterial microbiota in fecal samples during storage. Diagn Microbiol Infect Dis 2004;50(4):237–45.
[50] Hill CJ, Brown JRM, Lynch DB, Jeffery IB, Ryan CA, Ross RP, et al. Effect of room temperature transport vials on DNA quality and phylogenetic composition of faecal microbiota of elderly adults and infants. Microbiome 2016;4(1):19.
[51] Risely A, Gillingham MAF, Béchet A, Brändel S, Heni AC, Heurich M, et al. Phylogeny- and abundance-based metrics allow for the consistent comparison of core gut microbiome diversity indices across host species. Front Microbiol 2021;12:659918.
[52] Roden VJ, Kocsis ÁT, Zuschin M, Kiessling W. Reliable estimates of beta diversity with incomplete sampling. Ecology 2018;99(5):1051–62.
[53] Retzlaff N, Stadler PF. Phylogenetics beyond biology. Theory Biosci 2018;137(2):133–43.
[54] Lozupone CA, Knight R. The UniFrac significance test is sensitive to tree topology. BMC Bioinf 2015;16:211.
[55] Wong RG, Wu JR, Gloor GB, Moreno-Hagelsieb G. Expanding the UniFrac Toolbox. PLoS ONE 2016;11(9):e0161196.