

METHODOLOGY

Open Access



Performance determinants of unsupervised clustering methods for microbiome data

Yushu Shi¹, Liangliang Zhang², Christine B. Peterson^{3†}, Kim-Anh Do^{3†} and Robert R. Jenq^{4,5*†} 

Abstract

Background: In microbiome data analysis, unsupervised clustering is often used to identify naturally occurring clusters, which can then be assessed for associations with characteristics of interest. In this work, we systematically compared beta diversity and clustering methods commonly used in microbiome analyses. We applied these to four published datasets where highly distinct microbiome profiles could be seen between sample groups, as well a clinical dataset with less clear separation between groups.

Results: Although no single method outperformed the others consistently, we did identify the key scenarios where certain methods can underperform. Specifically, the Bray Curtis (BC) metric resulted in poor clustering in a dataset where high-abundance OTUs were relatively rare. In contrast, the unweighted UniFrac (UU) metric clustered poorly on dataset with a high prevalence of low-abundance OTUs. To explore these hypotheses about BC and UU, we systematically modified the properties of the poorly performing datasets and found that this approach resulted in improved BC and UU performance. Based on these observations, we rationally combined BC and UU to generate a novel metric. We tested its performance while varying the relative contributions of each metric and also compared it with another combined metric, the generalized UniFrac distance. The proposed metric showed high performance across all datasets.

Conclusions: Our systematic evaluation of clustering performance in these five datasets demonstrates that there is no existing clustering method that universally performs best across all datasets. We propose a combined metric of BC and UU that capitalizes on the complementary strengths of the two metrics.

Keywords: Unsupervised clustering, Bray Curtis distance, Unweighted UniFrac distance, Beta diversity

Background

The development of next-generation sequencing in the past decade has increased access for researchers to analyze microbial communities [1, 2]. A common method involves deep sequencing of 16S rRNA genes and

grouping bacteria at a certain level of 16S rRNA gene similarity [3]. The highest resolution bacterial sequence is referred to as an operational taxonomic unit (OTU). A challenge with microbiome analyses is that these sequences are derived from bacteria that can be incredibly diverse and simultaneously have a relatedness structure that is internally associated, both genetically and phenotypically. Metrics to account for phylogenetic closeness have been proposed specifically for microbiome data [4–6]. Previously, Fukuyama [7] showed that deep and shallow parts of the tree contribute differently to phylogenetic distances.

*Correspondence: rrjenq@mdanderson.org

†Christine Peterson, Kim-Anh Do, and Robert Jenq contributed equally to this work.

⁴Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, 1881 East Road, 35CR5.4102, Unit 1954, Houston, TX, 77054, USA

⁵Department of Stem Cell Transplantation and Cellular Therapy, CPRIT Scholar in Cancer Research, Texas, USA

Full list of author information is available at the end of the article



© The Author(s). 2022 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Microbiome data analysis can suffer from overgeneralization (e.g., comparing observations by alpha diversity measures such as Shannon [8] or Simpson [9] diversity) or overspecification (e.g., performing association tests for each taxon, which incurs a heavy penalty for multiple testing). A useful and common strategy to address both of these limitations is to first apply an unsupervised clustering methodology. After considering the microbiome data as a whole and identifying naturally occurring clusters of samples, these clusters can then be assessed for associations with sample characteristics of interest. Previous studies, for instance, have shown that human stool microbiome samples naturally form clusters that are associated with dietary and geographic factors [10, 11]. A problem that often confuses researchers is that clustering performance results often vary depending on the algorithm or the beta diversity metrics used, observed previously by Koren et al. [12] and Claesson et al. [13].

When performing microbiome sample clustering, both model-based methods and machine learning methods have been used. Machine learning methods, which rely on defined distance metrics, are used more frequently than model-based statistical methods, due to their efficient implementation and easy interpretation. In this paper, we focused on the partition around medoids (PAM) [14] clustering method, which is related to but considered more robust than K -means. In contrast to K -means, which can be sensitive to the effects of outliers, PAM's optimization goal is to minimize the sum of distances to the medoids instead of minimizing the sum of the squared distances to the cluster centers. We also evaluated another frequently used method for non-Euclidean metrics, hierarchical clustering with complete linkage. This method initially bundles the closest observations (with distances defined by the longest distance between any two observations in two clusters) and gradually results in a binary tree that combines the two closest clusters.

[Supplemental Figure S1](#) illustrates the differences between these clustering methods. The choice of clustering algorithm and distance metric together determine the performance of a machine learning clustering method. The metrics considered in this study are commonly used and include the Bray Curtis(BC) dissimilarity metric [15], the unweighted UniFrac(UU) distance [4], the weighted UniFrac distance [5], and the Aitchison distance [16], which is a Euclidean distance quantified following centered log ratio (CLR) transformation of abundances. Previously, a parameter-dependent metric, generalized UniFrac, was proposed by Chen et al [6]. The results from the generalized UniFrac are shown in [Supplemental Figure S6](#).

In addition to PAM and hierarchical clustering, we also included a third approach, the model-based Dirichlet multinomial model (DMM) [17]. This algorithm assumes

that observations of the same cluster come from the same multinomial distribution, and that parameters of the multinomial distribution have a Dirichlet distribution. This model better captures overdispersion than simply assuming a multinomial.

In this paper, we systematically compared methods for clustering microbiome observations from four published studies with either geographical or seasonal variables as the true cluster label, which enables biological interpretation of the group separation. We first applied clustering with five methods, and assessed performance of the various methods using the adjusted Rand index with the true clustering assignment. The adjusted Rand index is based on a pairwise membership agreement and is corrected for the expected value, where a score of 0 is expected with random clustering, and a score of 1 indicates perfect clustering. We then explored the relationships between the differences in performance with properties of the datasets considered. With the findings obtained from the method comparisons, we proposed a novel combined metric that provided high performance for all four datasets. We further test our findings with a clinical dataset where clusters are less separated.

Results

We used four published stool microbiome datasets to test the performance of the clustering methods. A brief summary of these datasets is shown in [Table 1](#). Each study comprises two groups of samples with relatively distinct microbiome profiles. The Shannon diversity of the dataset is calculated as the sum of Shannon diversities over all samples in the dataset, and details related to this alpha diversity are provided in a later section. The percent of sequences in high abundance OTUs is the percent of sequences in OTUs with average abundance greater than 0.001. While the individual studies varied in their 16S rRNA gene deep sequencing methods, we processed the downloaded raw sequences using the same pipeline, including VSEARCH [18] to dereplicate and remove singletons and the UNOISE2 function in usearch [19] to identify OTUs. More information about the four example datasets is provided in [Supplemental Table S1](#). A heatmap of the high abundance OTUs can also be found in [Supplemental Figure S2](#).

The first three studies compare cohorts separated by geography. The De Filippo [20] study consists of samples from healthy children living in a rural West African village in Burkina Faso, who consume a high-fiber, largely vegetarian diet, as well as samples from healthy European children of a similar age consuming a Mediterranean diet. The Martínez et al. [11] dataset consists of samples from the Asaro and Sausi communities in Papua New Guinea, as well as those from the USA. The dataset from Schnorr et al. [10] compares fecal samples from Italian urban

Table 1 Summary of the example datasets

Dataset	Number of OTUs	Cluster	Number of samples	Average sequencing depth	Shannon Diversity of the dataset	Abundance sum of OTUs with average abundance >0.001
De Filippo	2 803		29 (total)	14,015	152.95	0.382
		Italy	15	14,378		
		Africa	14	13,626		
Martinez	10 223		62 (total)	16,035	336.14	0.484
		Papua	40	16,504		
		US	22	15,182		
Schnorr	4 707		43 (total)	6011	238.74	0.058
		Hadza	27	4521		
		Italy	16	8526		
Smits	12 002		259 (total)	21,441	1233.86	0.621
		Late Dry	197	19,587		
		Early Wet	62	27,335		

adults with samples from adult hunter-gatherers living in Hadza land in Tanzania, Africa, who belong to one of the few populations that have maintained a traditional lifestyle with limited exposure to processed foods.

The fourth study, Smits et al. [21], comprises fecal samples longitudinally collected from a cohort of Hadza. Their diets are significantly affected by seasonal food availability, with more berries and honey during the wet season and more meat in the dry season. In this study, the authors found that some taxa disappear when certain foods become scarce and reappear when the seasons change. For the purpose of unsupervised learning, we used samples from the two most distinct seasons, the late dry and the early wet seasons.

The first column in Fig. 1 illustrates the adjusted Rand indices of the four datasets, with different colors indicating the beta diversity metrics using PAM clustering or Dirichlet multinomial mixture model (DMM). Because of its limited capacity to handle high-dimensional data, when fitting the DMM model, we binned the OTUs present in less than 20% of the observations into one OTU when calculating Rand indices. Also, due to the fact that Dirichlet multinomial mixture is model-based, there is no corresponding distance matrix that can be shown in a PCoA plot. The black bars in the Rand index plots and the last column in the PCoA plots are generated from our newly proposed method introduced later in this paper. Due to the generally similar results between PAM and hierarchical clustering (see Fig. 2), we presented the results in our main text only using PAM. Among these existing methods, we found that there is no clearly superior one that universally performs well in all four datasets. Interestingly, UU performs well in three of four datasets, despite lacking information regarding bacterial abundances. Similar

observations for the presence/absence methods have been made by Martino et al. [22]. Moreover, the Schnorr dataset has a perfect Rand index with UniFrac distances, but a low Rand index with BC, Aitchison, and DMM, all of which ignore phylogenetic relationships between OTUs. We devote the next section to exploring aspects of these differences in performance. For the completeness of comparison, we also tested the performance of 22 less common metrics that are provided by QIIME2 [23] (Fig. S7).

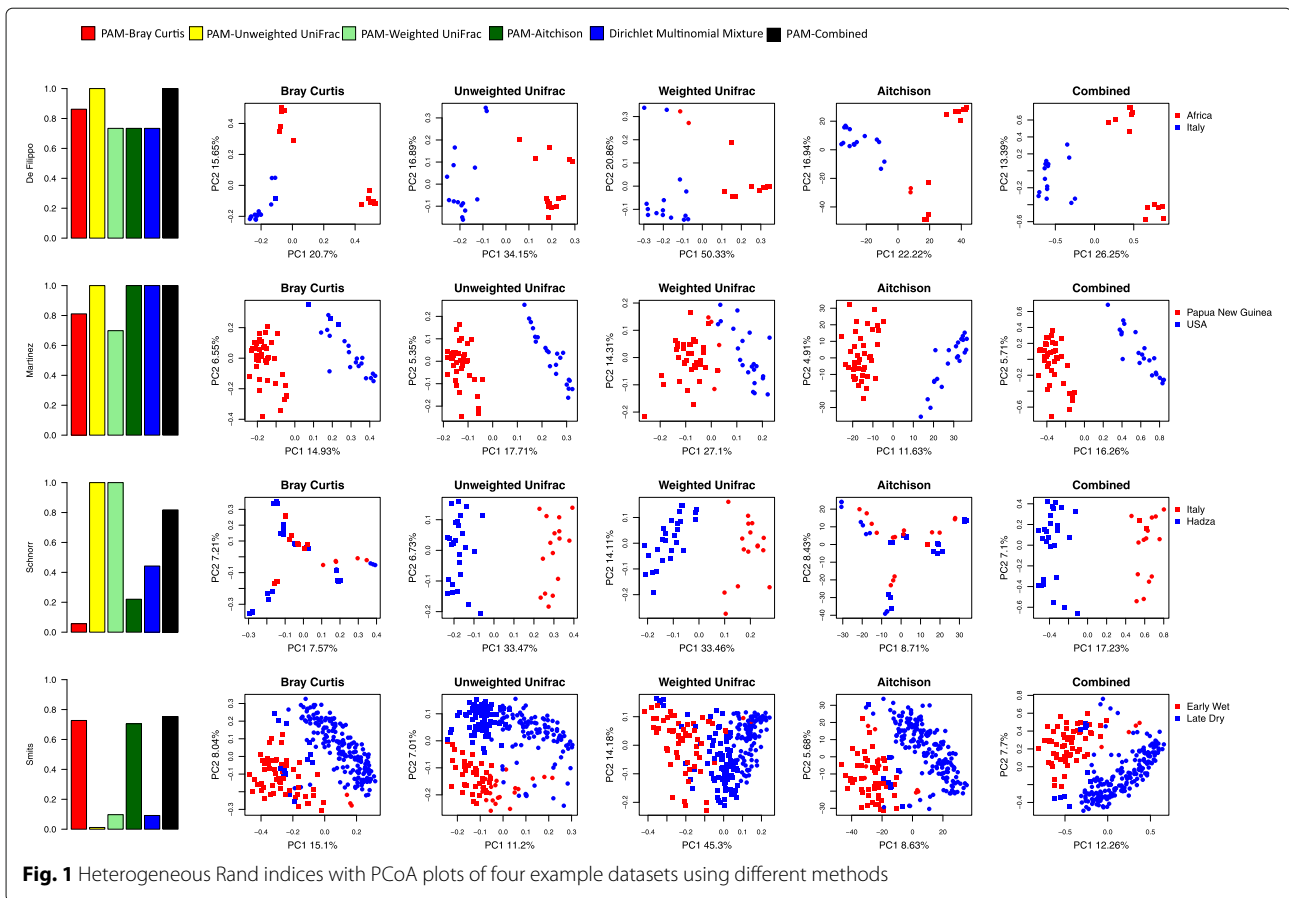
In this section, we will focus on the underperformance of BC for the Schnorr dataset, followed by the underperformance of UU for the Smits dataset. Because in reality, researchers may not know the information of the true cluster assignment, we will focus on a priori properties of the dataset, rather than the differences between the two assigned clusters.

High abundance OTUs drive the performance of the Bray-Curtis dissimilarity

For a $p \times n$ table with p OTUs and n observations, the BC distance between two observations A and B is calculated with the actual counts:

$$d^{BC} = \frac{\sum_{i=1}^p |n_{Ai} - n_{Bi}|}{\sum_{i=1}^p (n_{Ai} + n_{Bi})},$$

where n_{Ai} is the count for the i th OTU for observation A, and n_{Bi} is the count for the i th OTU for observation B. By examining the definition, we find that the dissimilarity of the BC metric is driven by differences in high-abundance rather than low-abundance OTUs. Here, we chose to define a metric, to potentially identify poor BC performance by summing the average abundance of “high abundance OTUs”, which are those with a mean



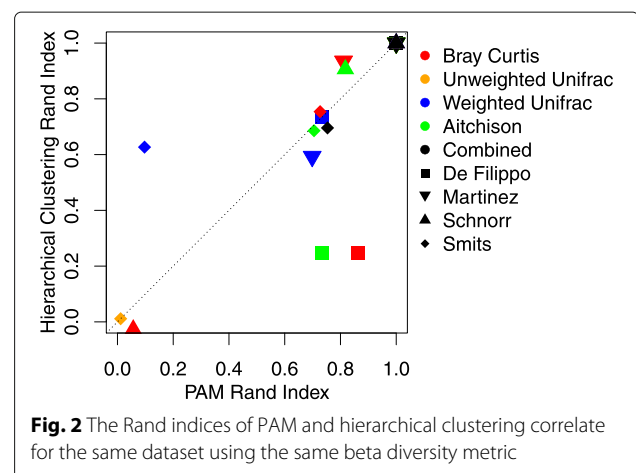
abundance greater than 0.001, though similar results can be obtained across a range of thresholds.

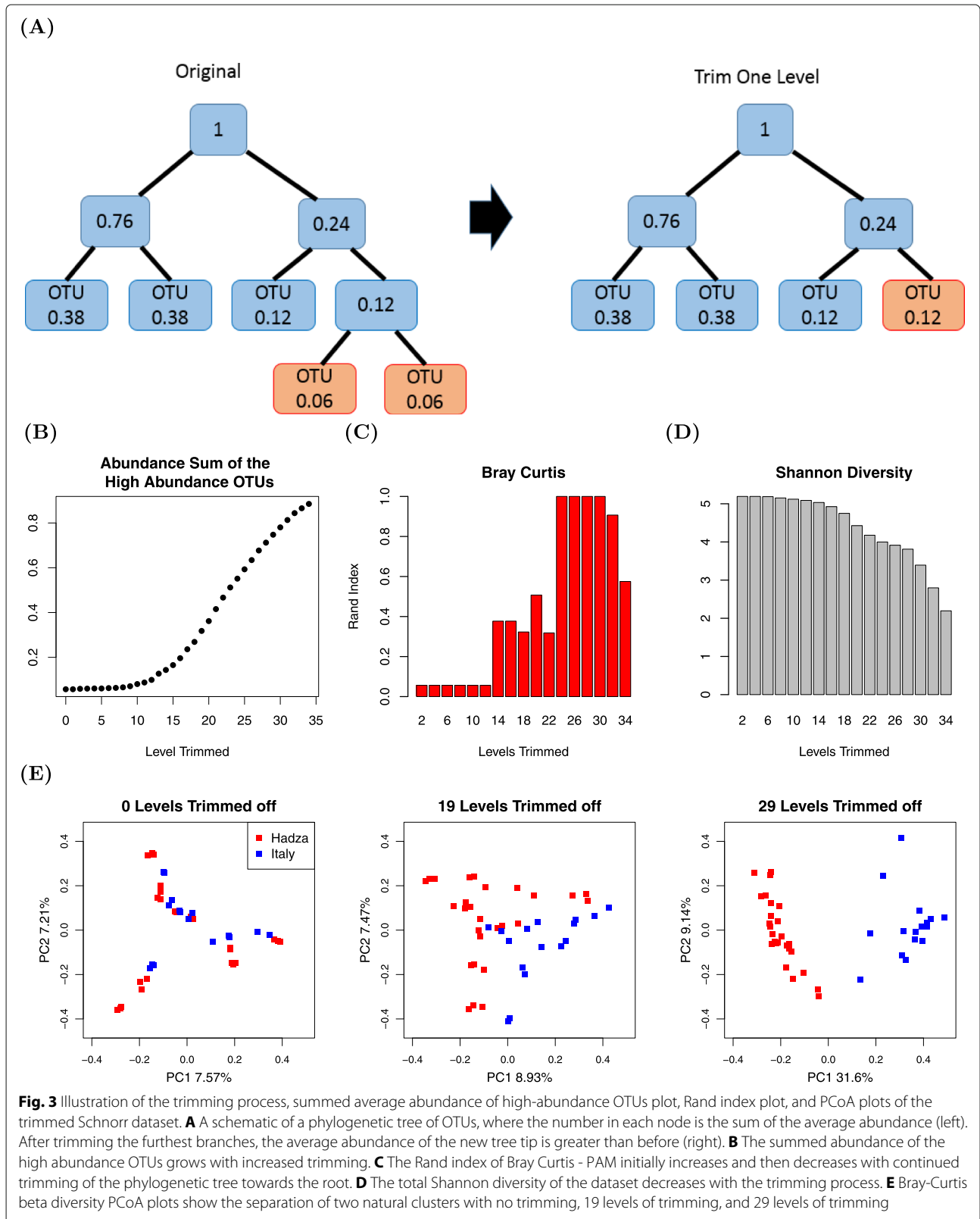
As shown in the last column of Table 1 and the last column of the Supplemental Table S1, the sum of average abundance and the high-abundance OTUs for the Schnorr dataset is far below that of the other three datasets. A visual description is also provided as a heatmap in Fig. S2. Thus, the Schnorr dataset is characterized by both unusually poor clustering performance by BC and few high-abundance OTUs.

To calculate this association, we attempted to improve the performance of BC by generating novel OTUs with higher mean abundances. We did this by first generating a phylogenetic tree of the aligned OTUs, and then systematically merging OTUs starting with those most distal to the tree root. In essence, we performed “trimming” of branches from the tree, by combining the abundances of the distal OTUs to generate new, more proximal OTUs with higher mean abundances, as shown in Fig. 3A.

In Fig. 3B, C, the *x*-axes indicate the number of levels trimmed, ranging from the first most distal branch to, at most, 34 levels proximal. Figure 3B shows how the summed abundance of high-abundance OTUs increases with more iterations of the trimming process, while Fig. 3C shows that the Rand index improves initially

as distal sparse OTUs are merged together. As trimming continues, the Rand index begins to worsen, as presumably excessive binning results in loss of distinctive OTU information. The three PCoA plots of Fig. 3E correspond to the original dataset (where no levels have been trimmed), to a modified dataset where 19 levels are trimmed off, and to a modified dataset where 29 levels are trimmed off from the most distal branch towards the root.





The two natural clusters begin to distinctly separate when 19 levels of branches are trimmed and are fully separated when 29 levels are trimmed. The corresponding heatmaps of the three datasets in Fig. S3 also demonstrate that trimming can increase the number of OTUs with high mean abundance.

Notably, as we progressively reduced the resolution of microbiome data during the trimming process, reflected by the decreasing Shannon diversity in Fig. 3D, the PAM clustering showed improved performance. Shannon diversity is a commonly used alpha diversity metric used to quantify the information within a dataset. The Shannon diversity for observation j is defined as:

$$\text{Shannon}_j = - \sum_{i=1}^p p_{ij} \log(p_{ij}), \quad p_{ij} > 0,$$

where p_{ij} is the abundance of the i th taxa for observation j . A reasonable explanation for the increase in Rand indices is that the trimming process actually utilizes the tree phylogenetic information, and similar thinking was previously mentioned in Gopalakrishnan et al. [24] and Peled et al. [25]. We encourage researchers to explore their datasets by trimming with the tools provided in our R package “MicrobiomeCluster.” Interestingly, datasets with clear clustering into two groups are likely to produce a bimodal histogram of pairwise distances between samples. One mode corresponds to small within-cluster distances, while the other mode corresponds to larger between-cluster distances.

As a counterexample, we chose to examine in more detail the Martínez dataset for its high clustering performance when using the BC metric. We reasoned that a high degree of discriminating information was being provided by OTUs with high mean abundances. To simulate OTUs with low mean abundances progressively, we took counts assigned to an OTU and assigned those to a new OTU in randomly selected samples from each tip of the original tree to form first-generation descendants, then grew two new branches from each of the newly formed tree tips to obtain the second-generation descendants, and so on. As zero counts do not contribute to the Shannon diversity, we essentially “grew” branches without increasing the information from the original dataset, as shown in Fig. 4A. Sequences of each OTU were randomly assigned to either of the two daughter branches. As shown in Fig. 4B, the summed abundance of the OTUs with high mean abundance is reduced. The performance of both BC- and Aitchison-based clustering show a decreasing trend, while tree-structure-based UniFrac methods are minimally affected in these simulations (Fig. 4C). With an increasing number of descendants, fewer OTUs have a high mean abundance when averaged over samples, corresponding with a

drop in the performance of the BC- and Aitchison-based methods.

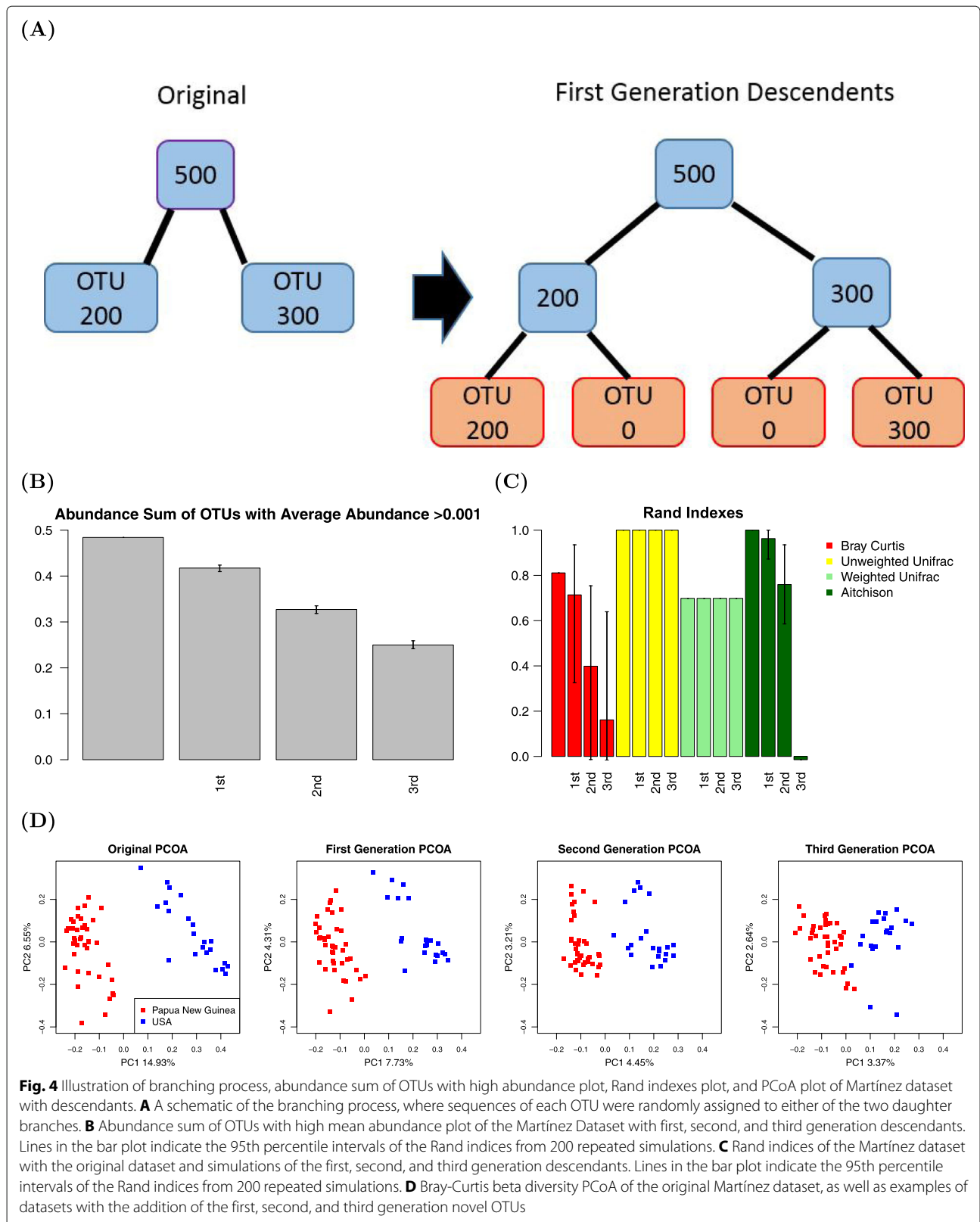
To visualize the effects of reduced OTU abundances, we plotted the PCoA of the original dataset, as well as examples of datasets with the addition of the first, the second, and the third generation simulated distal OTUs. As additional descendants are generated, the separation between two clusters becomes less apparent (Fig. 4D). We also plotted heatmaps of the original dataset, as well as examples of modified datasets with the addition of the first, and second generation descendants in Fig. S4. As tree branches progressively diverge, the most abundant OTUs have reduced mean abundances.

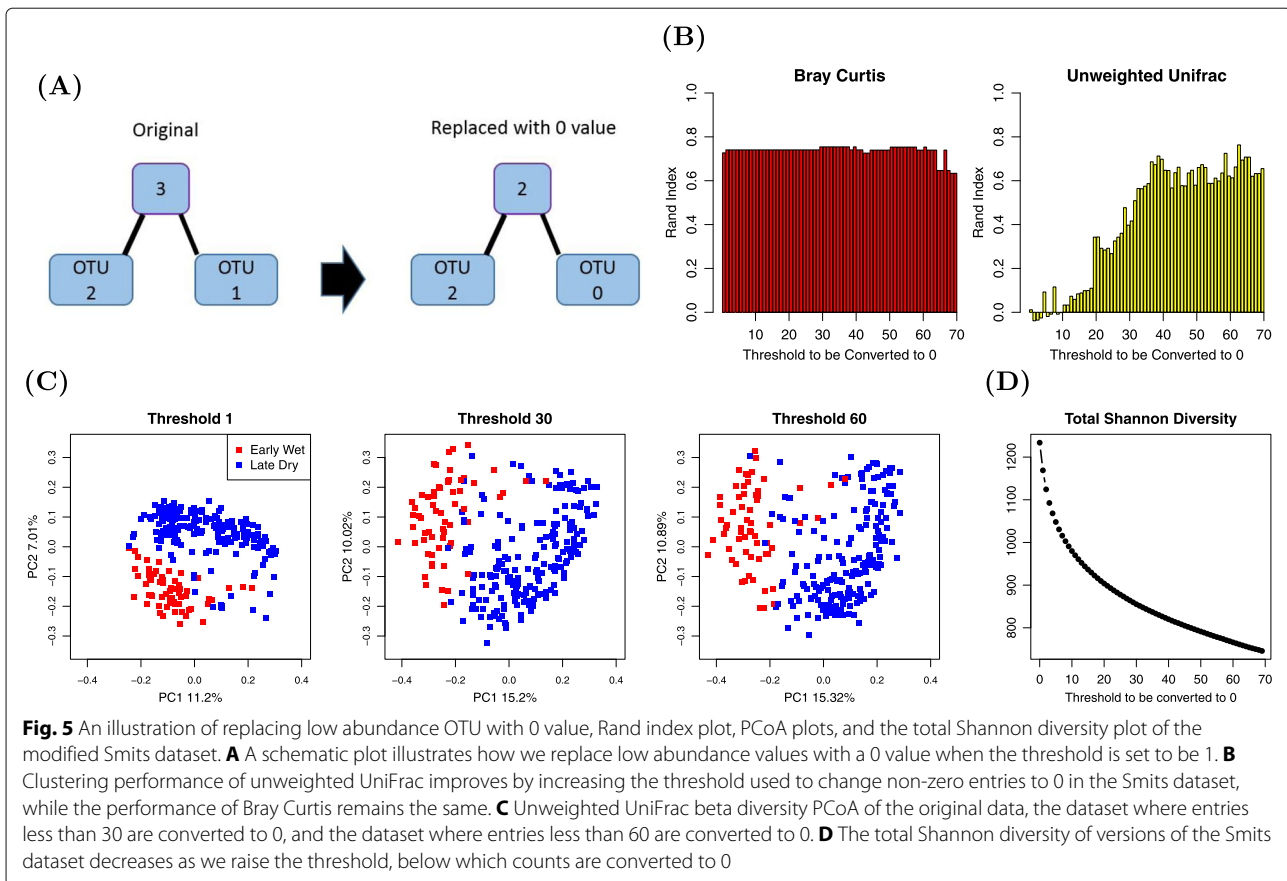
Prevalence of low-abundance OTUs inhibits clustering performance of the unweighted UniFrac distance

Another key observation from our examination of clustering performance in the four datasets is that the UU distance performs well for three of the datasets, but markedly underperforms in the Smits dataset. This dataset is distinct from the others in that samples were collected at different time points from the same group of individuals, rather than being collected from two distinct groups of subjects. Thus, the association of the microbiome is with time (distinct in two seasons), rather than the source. We would anticipate that, during the late dry season, many bacteria more suited for the early wet season will be reduced in abundance but may not be totally eliminated and can recover when the host diet eventually becomes more hospitable, and vice versa. The geographically separated clusters from the other studies are less likely to have an extensive overlap in presence of bacterial taxa, with many taxa being present only in one population and absent in the other.

To explore the determinants of when the UU metric could lead to reduced clustering performance, we focused on the unweighted nature of the measure, where only the presence or absence of an OTU is quantified rather than the abundance of an OTU. In the Smits dataset, we expected that a large number of OTUs are shared by both clusters. To identify a metric that captures the information uncaptured by UU, we propose using the total Shannon diversity, which is the sum of the Shannon diversities over all the samples of the dataset.

Because zero entries in the OTU table do not contribute to total Shannon diversity, the total Shannon diversity in a dataset is a metric that can reflect the amount of information uncaptured by UU across samples. To examine this, we simulated new repetitions of the data from the Smits dataset by gradually switching non-zero entries to 0, in essence adjusting the criteria by which an OTU would be considered “present” in a sample, as shown in Fig. 5A. In both Fig. 5B and D, the x -axis is the threshold going





from 0 to 70. During the process, entries with a number of sequences less than or equal to the threshold will be converted to 0. In Fig. 5B, the performance of UU goes up with the number of entries converted to 0, whereas the performance of BC only fluctuates slightly. In Fig. 5D, the total Shannon diversity drops down sharply with the number of entries converted to 0.

PCoA also visually depicts improved discrimination of UU with an increasing threshold. In Fig. 5C, we plotted the PCoA of the original data, the simulated dataset where entries less than 30 are converted to 0, and the simulated dataset where entries less than 60 are converted to 0. As more non-zero entries are converted to 0, the two clusters separate further in the first coordinate.

As a counter example, we chose to examine in more detail the Martínez dataset, for which the UU displays high clustering performance. To simulate poorly discriminating data, we increasingly substituted 0 entries with a single count value. Such an alteration of the dataset imposes a minimal change to the original data. As with the Smits dataset, we quantified the median with the 95% empirical confidence intervals of Rand indices and Shannon diversities over 200 simulated datasets. In the x -axes of both Fig. 6B and C, the probability of filling a 0-value entry with 1 goes from 0 to 1, with a 0.1

increment. Figure 6B demonstrates that the total Shannon diversity indeed increases as 0-value entries are increasingly replaced with a value of 1. In turn, we found that the Rand indices of other methods are almost completely unchanged with the perturbation, while the performance of the UU-based clustering is progressively reduced in Fig. 6C.

The PCoA plots of the original dataset, as well as a simulated dataset where 40% of 0-value entries in the OTU table are replaced with a value of 1, and a simulated dataset where 60% of 0-value entries are replaced with a value of 1 are shown in Fig. 6D. We found that, as percentage of replacement increased, the two groups of samples become progressively less distinct.

Combined metric shows stable good performance

With the above findings, it would be appealing to develop a metric that can use the information of both high- and low-abundance OTUs. An intuitive way of constructing such a new metric is to combine information from both BC and UU metrics, which tend to complement each other. After normalizing pairwise BC and UU metrics by their largest values the two metrics lie between 0 and 1. We define the distance between two observations in the new metric as:

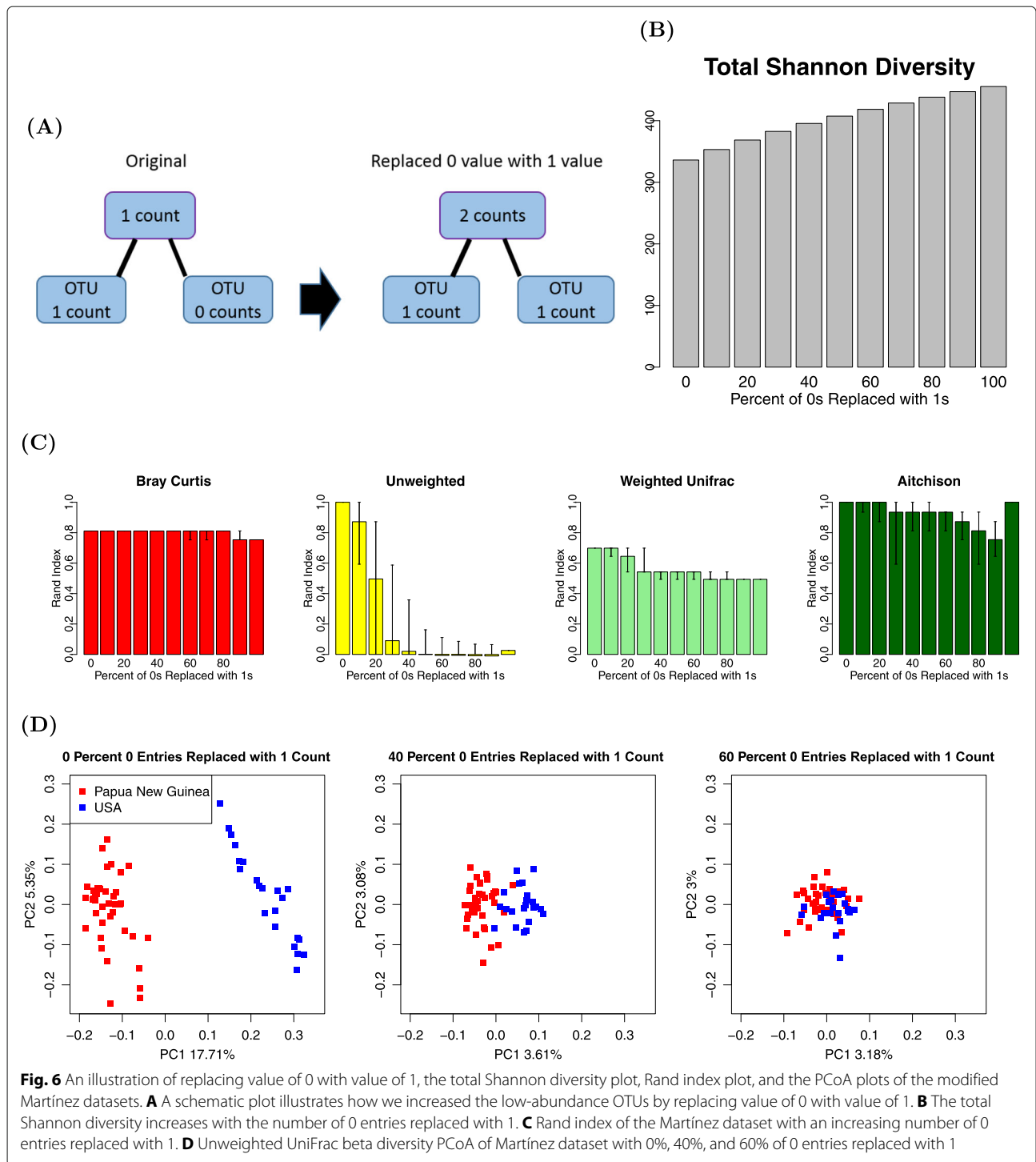
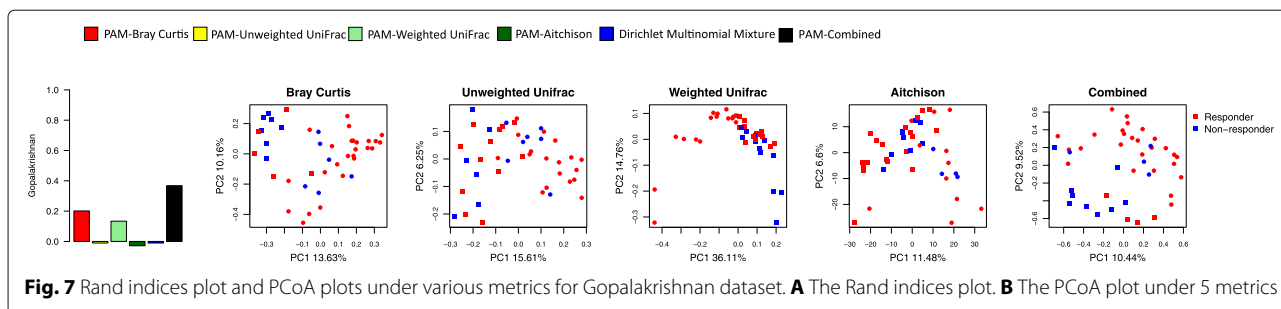


Fig. 6 An illustration of replacing value of 0 with value of 1, the total Shannon diversity plot, Rand index plot, and the PCoA plots of the modified Martínez datasets. **A** A schematic plot illustrates how we increased the low-abundance OTUs by replacing value of 0 with value of 1. **B** The total Shannon diversity increases with the number of 0 entries replaced with 1. **C** Rand index of the Martínez dataset with an increasing number of 0 entries replaced with 1. **D** Unweighted UniFrac beta diversity PCoA of Martínez dataset with 0%, 40%, and 60% of 0 entries replaced with 1

$$d^{\text{combined}} = \alpha d_{\text{normalized}}^{\text{UU}} + (1 - \alpha) d_{\text{normalized}}^{\text{BC}}$$

Different α values change the relative contributions of UU and BC distances respectively, and we suggest using $\alpha = 0.5$ as a default simple choice. The Fig. S6 shows that our combined metric tends to do better than the generalized UniFrac with various parameter values, and the

default choice gives a robust performance across example datasets. As shown in the last column of the bar plots and the last column of PCoA plots of Fig. 1, the proposed new metric (with $\alpha = 0.5$), which inherits complementary insights from the BC and the UU dissimilarities, results in high performance in Rand indices for all the datasets. As shown in Fig. S7, the combined met-



ric also outperforms other less commonly used metrics in QIIME2.

Evaluation in a setting without strong cluster separation

We also explored a clinical setting where one does not expect global separation between groups [24]. The Gopalakrishnan study examines the relationship between microbiome and response to immunotherapy drugs for melanoma patients. Here, we consider patients who responded to (30 patients) and patients who did not respond to immunotherapy drugs (13 patients) as two clusters. There are 1455 OTUs in this dataset and the average sequencing depth is 48,765. Compared with four datasets discussed in previous sections, its abundance sum of OTUs with average abundance > 0.001 is higher (0.874), while its Shannon diversity is lower (132.75). However, if we convert low abundance OTUs to 0s as we did for Smits dataset, the performance of UU improves while the Shannon diversity decreases (Fig. S5).

As shown in Fig. 7 and Fig. S6, all Rand indices are less than 0.4, with the combined metric being the highest. It is interesting to notice that the combined metric can outperform each of its individual components. The BC had the second best performance, corresponding with the presence of high-abundance OTUs. Of note, clustering performance by UU is poor, and the low Shannon diversity of the dataset would normally indicate the potential for better performance.

Discussion

A quantitative recommendation regarding when to avoid certain metrics would require a systematic examination of many more datasets. We did observe that in four of the datasets we considered, hierarchical clustering, though less frequently used with microbiome data, can provide similar and sometimes superior results compared to the more commonly used PAM method (Fig. 2). Unlike PAM clustering, the number of clusters do not have to be pre-determined prior to hierarchical clustering, which is a potential advantage. For supervised settings where the group membership is known, beta diversity metrics, including the proposed combined beta metric, can be used as input to distance-based testing approaches such

as PERMANOVA. For the data sets considered here, PERMANOVA produces significant p -values across virtually all metrics and data sets. This is because PERMANOVA takes the group labels as known, and tests the null hypothesis that both the centroids and dispersion are the same across groups. There may be evidence to reject this null even in settings where the groups overlap substantially.

Conclusions

Our systematic evaluation of clustering performance in these five datasets show that there is no existing clustering method that universally performs the best across all datasets. In general, the weighted UniFrac outperforms other methods across studies. When OTUs with high mean abundance are rare, clustering methods that do not consider phylogeny (i.e., BC, Aitchison, DMM) perform less well. In contrast, in a dataset with lots of low abundance OTUs, the UU tends to produce poor separation, while other methods can still perform well. To capitalize on the complementary strength of the two metrics, we propose a new metric that incorporates the BC metric and the UU metric, which gives good performance in a dataset representative of the common scenario where clusters are less separated.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40168-021-01199-3>.

Additional file 1: Figure S1. An illustrative plot of commonly used clustering methods.

Additional file 2: Figure S2. Heatmap of the most abundant 300 OTUs for the four example datasets. This figure shows the most abundant 300 OTUs for the four published datasets plotted in log₁₀ scale. Red color indicates high abundance, whereas white color indicates low abundance. As shown in the plot, the abundances of “high abundance” OTUs in Schnorr dataset are lower than those of the other three datasets. Gray and blue indicate different clusters in each dataset.

Additional file 3: Figure S3. Heatmap of the most abundant 300 OTUs for the Schnorr dataset with 0, 19, and 29 levels trimmed off. This figure plots the most abundant 300 OTUs of the Schnorr Dataset with 0, 19, and 29 levels trimmed off. As trimming goes along, the abundant OTUs aggregate sequences from distant OTUs. In each subpanel, the gray left part is the Italian sample set, while the blue right part is the Hadza sample set.

Additional file 4: Figure S4. Heatmap of the most abundance 300 OTUs of the Martínez dataset, with descendants. This figure plots the most

abundance 300 OTUs of the Martínez Dataset with its first, second and third generation descendants. As the tree branches diverge, fewer sequences are left in the most abundant 300 OTUs. In each subpanel, the gray left part is the Papua sample set, while the blue right part is the US sample set.

Additional file 5: Figure S5. Covering low abundance OTUs to 0s improves the performance of UU for Gopalakrishnan dataset. (A) Shannon diversity of the dataset decreases as more OTUs are converted to 0s. (B) The trend of the Rand index rises by converting OTUs to 0s, however excessive count removal affects the performance of UU eventually.

Additional file 6: Figure S6. Rand indices with different α values for the proposed metric and comparison with the generalized UniFrac metric. This figure shows the Rand indices with different values of α (0.2, 0.4, 0.6, 0.8) for the proposed metric and compares the results with the Generalized UniFrac under different parameters for it (0, 0.25, 0.5, 0.75).

Additional file 7: Figure S7. Performance of other metrics. The figure shows the performance of other less common metrics provided by QIIME2. Among them, city-block distance and species-by-species Euclidean distance are modifications of the Euclidean distance, while Jaccard distance and Canberra distances are similar to the Bray Curtis distance. As pointed out in the vegan package manual [27], Euclidean and Manhattan distances are not good in separating groups. More details of beta diversity metrics can be found in QIIME2 [23].

Additional file 8: Table S1. Additional summary of the example datasets.

Acknowledgements

Not applicable

Authors' contributions

RRJ designed the research. YS conducted the data analysis and wrote the manuscript, with input from LZ, CBP, and KAD. RRJ, CBP, and KAD revised the manuscript. All authors read and approved the final manuscript.

Funding

CBP is partially supported by NIH/NCI CCSG grant P30CA016672. RRJ is partially supported by NIH R01 HL124112 and CPRIT RR160089 grants. KAD is partially supported by a Cancer Center Support Grant NCI Grant P30 CA016672, NIH grants UL1TR003167, 5R01GM122775, the prostate cancer SPORE P50 CA140388, CPRIT grant RP160693, and the Moon Shots funding at MD Anderson Cancer Center. RRJ is also funded by Cancer Center Support Grant NCI Grant P30 526 CA016672 and Moon Shots funding at MD Anderson Cancer Center.

Availability of data and materials

We created an R package "MicrobiomeCluster," which contains methods implemented in this manuscript (<https://github.com/YushuShi/MicrobiomeCluster.git>) [26], including the proposed combined metric as well as some basic data manipulation tools.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

Dr. Robert Jenq is a consultant of Merck, Karius, and MicrobiomeDx and a SAB member of Seres and Kaleido and has patent license of Seres. The other authors declare that they have no competing interests.

Author details

¹Department of Statistics, The University of Missouri, Columbia, 209D Middlebush Hall, Columbia, MO, 65201, USA. ²Department of Population and Quantitative Health Sciences, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH, 44106, USA. ³Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler St, 4th Floor, Houston, TX, 77030, USA. ⁴Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, 1881 East Road, 35CR5.4102, Unit 1954, Houston, TX, 77054, USA. ⁵Department of Stem Cell Transplantation and Cellular Therapy, CPRIT Scholar in Cancer Research, Texas, USA.

Received: 13 September 2021 Accepted: 15 November 2021

Published online: 05 February 2022

References

- Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. *Nat Med*. 2018;24(4):392–400.
- Knight R, Callewaert C, Marotz C, Hyde ER, Debelius JW, McDonald D, Sogin ML. The microbiome and human biology. *Annu Rev Genomics Hum Genet*. 2017;18:65–86.
- Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchell T, Perry T, Kao D, Mason AL, Madsen KL, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front Microbiol*. 2016;7:459.
- Lozupone C, Knight R. Unifrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*. 2005;71(12):8228–35.
- Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol*. 2007;73(5):1576–85.
- Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*. 2012;28(16):2106–13. <https://doi.org/10.1093/bioinformatics/bts342>.
- Fukuyama J. Emphasis on the deep or shallow parts of the tree provides a new characterization of phylogenetic distances. *Genome Biol*. 2019;20(131):.
- Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27(4):623–56. <https://doi.org/10.1002/j.1538-7305.1948.tb00917.x>.
- Simpson EH. Measurement of diversity. *Nature*. 1949;163:688.
- Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, Turroni S, Biagi E, Peano C, Severgnini M, et al. Gut microbiome of the Hadza hunter-gatherers. *Nat Commun*. 2014;5:3654.
- Martínez I, Stegen JC, Maldonado-Gómez MX, Eren AM, Siba PM, Greenhill AR, Walter J. The gut microbiota of rural Papua New Guineans: composition, diversity patterns, and ecological processes. *Cell Rep*. 2015;11(4):527–38.
- Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, Huttenhower C, Ley RE. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput Biol*. 2013;9(1):1–16. <https://doi.org/10.1371/journal.pcbi.1002863>.
- Claesson M, Jeffery I, Conde S, Power S, O'Connor E, Cusack S, M B Harris H, Coakley M, Lakshminarayanan B, O'Sullivan O, F Fitzgerald G, Deane J, O'Connor M, Harnedy N, O'Connor K, O'Mahony D, Van Sinderen D, Wallace M, Brennan L, W O'Toole P. Gut microbiota composition correlates with diet and health in the elderly. *Nature*. 2012;488:178–84. <https://doi.org/10.1038/nature11319>.
- Kaufman L, Rousseeuw PJ. Partitioning around medoids (program pam). In: *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Ltd; 1990. p. 68–125. <https://doi.org/10.1002/9780470316801.ch2>. <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316801.ch2>.
- Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. *Ecol Monogr*. 1957;27(4):325–49.
- Aitchison J. Principal component analysis of compositional data. *Biometrika*. 1983;70(1):57–65. <https://doi.org/10.1093/biomet/70.1.57>.
- Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE*. 2012;7(2):1–15. <https://doi.org/10.1371/journal.pone.0030126>.
- Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*. 2016;4:2584.
- Edgar RC. UNOISE2: improved error-correction for Illumina 16S and its amplicon sequencing. *bioRxiv*. 2016. <https://doi.org/10.1101/081257>.
- De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P. 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.
- Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, Reid G, Knight R, Manjuran A, Changalucha J, Elias JE, Dominguez-Bello MG, Sonnenburg JL. Seasonal cycling in the gut microbiome of the Hadza

- hunter-gatherers of Tanzania. *Science*. 2017;357(6353):802–6. <https://doi.org/10.1126/science.aan4834>.
22. Martino C, Morton JT, Marotz CA, Thompson LR, Tripathi A, Knight R, Zengler K. A novel sparse compositional technique reveals microbial perturbations. *mSystems*. 2019;4(1):. <https://doi.org/10.1128/mSystems.00016-19>.
 23. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciulek T, Kreps J, Langille MG, Lee J, Ley R, Liu Y-X, Lofffield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, Mciver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hoft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol*. 2019;37(8):852–7. <https://doi.org/10.1038/s41587-019-0209-9>.
 24. Gopalakrishnan V, Spencer C, Nezi L, Reuben A, Andrews M, Karpnits T, Prieto P, Vicente D, Hoffman K, Wei S, et al. Gut microbiome modulates response to anti-pd-1 immunotherapy in melanoma patients. *Science*. 2018;359(6371):97–103.
 25. Peled JU, Devlin SM, Staffas A, Lumish MA, Khanin R, Littmann ER, Ling L, Kosuri S, Maloy MA, Slingerland J, Ahr KF, Rodriguez KAP, Shono Y, Slingerland AE, Docampo M, Sung AD, Weber D, Alousi AM, Gyurkocza B, Ponce DMC, Barker J, Perales M-A, Giralto SA, Taur Y, Pamer EG, Jenq RR, van den Brink MRM. Intestinal microbiota and relapse after hematopoietic-cell transplantation. *J Clin Oncol*. 2017;35(15):1650–9.
 26. Shi Y. *MicrobiomeCluster*. 2020. R package. <https://github.com/YushuShi/MicrobiomeCluster.git>. Accessed 28 Jan 2022.
 27. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. *Vegan: community ecology package*. 2020. R package version 2.5-7. <https://CRAN.R-project.org/package=vegan>. Accessed 28 Jan 2022.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

