Combined Beta Metric for Unsupervised Clustering of Microbiome Data

CURRENT STATUS: UNDER REVIEW

Yushu Shi
University of Texas MD Anderson Cancer Center

Liangliang Zhang
University of Texas MD Anderson Cancer Center

Christine Peterson
University of Texas MD Anderson Cancer Center

Kim-Anh Do
University of Texas MD Anderson Cancer Center

Robert Jenq
University of Texas MD Anderson Cancer Center

rrjenq@mdanderson.org Corresponding Author
ORCiD: https://orcid.org/0000-0002-5434-439X

DOI: 10.21203/rs.3.rs-17968/v1

SUBJECT AREAS
General Microbiology

KEYWORDS
Unsupervised Clustering; Bray Curtis distance; Unweighted UniFrac distance; Beta diversity
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. Results Although no single method outperformed the others consistently, we did identify key scenarios where certain methods can underperform. Specifically, the Bray Curtis metric resulted in poor clustering in a dataset where high-abundance OTUs were relatively rare. In contrast, the unweighted UniFrac metric clustered poorly when used on a dataset with a high prevalence of low-abundance OTUs. To test our proposition, we systematically modified properties of the poorly performing datasets and found that this approach resulted in improved Bray Curtis and unweighted UniFrac performance. Conclusions Based on these observations, we rationally combined the Bray Curtis metric and the unweighted UniFrac metrics and found that this new beta diversity metric showed high performance across all datasets.

Full Text
Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Figures
**Figure 1**

Heterogeneous Rand indices with PCoA plots of four example datasets using different methods
The Rand indices of PAM and hierarchical clustering correlate for the same dataset using the same beta diversity metric.
Illustration of Trimming Process, Summed Average Abundance of High - abundance OTUs Plot, Rand Index Plot and PCoA Plots of the Trimmed Schnorr Dataset (A) A schematic of a phylogenetic tree of OTUs, where the number in each node is the sum of the average abundance (left). After trimming the furthest branches, the average abundance of the new tree tip is greater than before (right). (B) The summed abundance of the high abundance
OTUs grows with increased trimming. (C) The Rand Index of Bray Curtis - PAM initially increases and then decreases with continued trimming of the phylogenetic tree towards the root. (D) The total Shannon diversity of the dataset decreases with the trimming process. (E) Bray-Curtis Beta diversity PCoA plots show the separation of two natural clusters with no trimming, 19 levels of trimming, and 29 levels of trimming.
Illustration of Branching Process, Abundance Sum of OTUs with High Abundance Plot, Rand Indexes Plot, and PCoA Plot of Martnez Dataset with Descendants (A) A schematic of the branching process, where sequences of each OTU were randomly assigned to either of the two daughter branches. (B) Abundance sum of OTUs with high mean abundance plot of the Martnez Dataset with rst, second, and third generation descendants. Lines in the bar plot indicate 95th percentile intervals of the Rand indices from 200 repeated simulations. (C) Rand indices of the Martnez Dataset with the original dataset, and simulations of the rst, second, and third generation descendants. Lines in the bar plot indicate 95th percentile intervals of the Rand indices from 200 repeated simulations. (D) Bray-Curtis beta diversity PCoA of the original Martinez dataset, as well as examples of datasets with the addition of the rst, second, and third generation novel OTUs.
Figure 5

An Illustration of Replacing Low Abundance OTU with 0 value, Rand Index Plot, PCoA Plots and the Total Shannon Diversity Plot of the Modified Smits Dataset (A) A schematic plot illustrates how we replace low abundance values with a 0 value when the threshold is set to be 1. (B) Clustering performance of unweighted Unifrac improves by increasing the threshold used to change non-zero entries to 0 in the Smits dataset, while the performance of Bray Curtis remains the same. (C) Unweighted UniFrac beta diversity PCoA of the original data, the dataset where entries less than 30 are converted to 0, and the dataset where entries less than 60 are converted to 0. (D) The total Shannon diversity of versions of the Smits dataset decreases as we raise the threshold, below which counts are converted to 0.
Figure 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 Martinez 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 Martinez dataset with an increasing number of 0 entries replaced with 1. (D) Unweighted UniFrac beta diversity PCoA of Martinez dataset with 0; 40; 60 percent of 0 entries replaced with 1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

S3.pdf
S2.pdf
S4.pdf
S1.pdf