Comparison of Beta Diversity Measures in Clustering the High-dimensional Microbial Data

Biyuan Chen†, Xueyi He‡, Bangquan Pan‡, Xiaobing Zou† and Na You‡*

Abstract

Background: Heterogeneity of disease is a major concern in medical research, which is commonly recognized as subtypes with different pathogenesis, exhibiting distinct prognosis and treatment effects. The classification of population into homogenous subgroups is challenging, especially for the complex diseases. Recent studies show that gut microbiome compositions play a vital role in disease development, and it is of great interest to cluster the patients according to their microbial profiles. There are a variety of beta diversity measures to quantify the dissimilarity between the compositions of different samples for clustering. However, using different beta diversity measures results in different clusters, and it is difficult to make a choice.

Results: Considering the microbial compositions from 16S rRNA sequencing, which are presented as a high-dimensional vector with large proportion of extremely small or even zero-valued elements, we set up three simulation experiments to mimic the microbial compositional data and evaluate the performance of different beta diversity measures in clustering. Their performance in two real datasets demonstrates the validity of the simulation experiments.

Conclusion: It is shown that J-divergence/Jensen-Shannon divergence/square root of Jensen-Shannon divergence and Bhattacharyya/Hellinger can capture the compositional changes at low abundance elements more efficiently, and Bhattacharyya/Hellinger is suggested.

Keywords: Enterotype; Clustering Analysis; Beta Diversity; Dissimilarity; Microbial composition; High-dimensional Data

Background

Heterogeneity of disease is the primary concern of precision medicine, which challenges the medical research in many aspects, from the identification of risk factors to the development of specific treatments [1–3]. The patients with the perceived same disease may respond quite differently to the same treatment and show a distinct prognosis in clinical practice. Most common diseases are so complex that there exist various subtypes, and the patients vary in etiology and pathogenesis between subtypes [3–5]. Rather than treating the patients uniformly, it is more plausible to classify...
them into subgroups, and develop precise treatments specifically.

Recently, many studies indicate that the gut microbiome plays an important role in the origin and development of disease through the gut-brain axis [6–9]. With the advantage of high efficiency and low cost, the abundance of microbial genes in the gut samples are popularly measured by the 16S rRNA high-throughput sequencing [10]. The analysis pipeline [11] clusters the sequenced reads as the operational taxonomic units (OTUs) and measures their abundance by the binned reads coverage. Annotating the OTU sequences at different taxonomic levels results in the microbial compositions and their abundance at different levels. Considering the microbial evolution and taxonomy accuracy of 16S rRNA sequencing, the analysis at the genus level is of great interest, where the OTUs abundance of each sample is represented as a high-dimensional vector. Moreover, because the abundance values are non-negative, and through the normalization to make the total amount of each sample to be one, there appears large proportion of extremely small or even zero-valued compositions in the OTU table.

The clustering of microbial compositional vectors reveals the heterogeneity of patients in the gut microbiome. The clustered subgroups are characterized as enterotypes, attracting much attention once upon their appearance [12–14]. In order to classify the samples into subgroups according to their compositional profiles, the dissimilarity between samples needs to be measured, which is termed as the beta diversity in the microbial community. According to different definitions of the dissimilarity, there have been a variety of beta diversity measures being proposed, such as over 40 ones included in the R package phyloseq [15]. There consist of not only the commonly used Euclidean and Jensen-Shannon divergence, but also the diversity measures for presence-absence data [16] and the Unifrac distance utilizing phylogenetic information [17]. Besides, the R package Phlintropy [18] implements 46 distance and similarity measures in [19] for comparing probability functions. Although there are fruitful choices for beta diversity, different measures may provide significantly different clusters [20]. It is confused for the users to make one selection from those resulted clusters, even according to the indices to evaluate the clustering performance such as the Caliński-Harabasz statistic, Silhouette coefficient, and prediction strength, since they may give different suggestions [13, 20].

Numerical evaluation based on the simulations can provide an objective comparison of the performance of different beta diversity measures. However, the previous works mainly focus on the analysis of low dimensional data [21–23]. In this paper, we set up three simulation experiments to mimic the microbial compositions to investigate the performance of different beta diversity measures in clustering. By comparing with the truth, we can infer the advantages of different measures in revealing the dispersion of compositions in different manners, guiding the choice of beta diversity in practical data analysis.

Note that we focus on the dispersion in the compositional abundance between samples, rather than the presence-absence of species in the microbiome. Considering their popularity being used in the literature for microbial analysis, we include 13 beta diversity measures for comparison. The paper is structured as follows. In Methods section, we present the definition of each beta diversity under investigation. Three simulation experiments are introduced in Results section, where the performance of different beta diversity measures are explored, followed by two real datasets analysis. A short Discussion section and Conclusion section are presented at the end.
Methods

Denoted by \( x_i = (x_{i1}, \ldots, x_{im})' \) the \( m \) compositions of the \( i \)th subject in the population \( x_1, \ldots, x_N \), the compositional constraints \( x_{ik} \geq 0 \) and \( \sum_{k=1}^{m} x_{ik} = 1 \) hold for \( i = 1, 2, \ldots, N \). Given a pre-specified number of clusters \( K \), the clustering algorithm, such as the partitioning around medoids method [24], classifies the population into \( K \) groups according to the dissimilarity matrix \( D = (d_{ij})_{N \times N} \), where \( d_{ij} \), termed as the beta diversity in the microbial community, quantifies the dissimilarity between the compositional vectors of two distinct samples \( x_i \) and \( x_j \).

The most commonly used beta diversity measures, Euclidean \( \beta_1 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \) and Manhattan \( \beta_2 = \sum_{k=1}^{m} |x_{ik} - x_{jk}| \) [18], are actually \( L_2 \) or \( L_1 \) norm developed in the real space. The Bray-Curtis [15], or called as Canberra [18],

\[
\beta_3 = \frac{\sum_{k=1}^{m} |x_{ik} - x_{jk}|}{\sum_{k=1}^{m} (x_{ik} + x_{jk})}
\]
gives the equivalent dissimilarity matrix as Manhattan for clustering, due to \( \beta_3 = \beta_2/2 \) provided that \( \sum_{k=1}^{m} x_{ik} = \sum_{k=1}^{m} x_{jk} = 1 \). The Jaccard [15], or Tanimoto [18], is defined as

\[
\beta_4 = 1 - \frac{\sum_{k=1}^{m} \min(x_{ik}, x_{jk})}{\sum_{j=1}^{m} \max(x_{ik}, x_{jk})}
\]

which is a monotone function of \( \beta_2 \), i.e., \( \beta_4 = 2\beta_2/(1 + \beta_2) \). Because of the equivalence of Bray-Curtis and Manhattan, we will not include Bray-Curtis in the comparison analysis.

The Kullback-Leibler divergence [25] reflects the difference between two probability measures. Its discrete version can be directly applied to measure the dispersion between two compositional vectors, yielding the J-divergence [23]

\[
\beta_5 = \sqrt{D(x_i||x_j) + D(x_j||x_i)},
\]

and the widely used Jensen-Shannon divergence (JSD) [18]

\[
\beta_6 = \frac{1}{2} \left[ D(x_i||\frac{x_i + x_j}{2}) + D(x_j||\frac{x_i + x_j}{2}) \right],
\]

where \( D(x_i||x_j) = \sum_{k=1}^{m} x_{ik} \ln(x_{ik}/x_{jk}) \) indicates the Kullback-Leibler divergence. Since JSD does not satisfy the triangle inequality, it is not a mathematical distance, but its square root (rJSD) [26] \( \beta_7 = \sqrt{\beta_6} \) is, so that rJSD is usually alternatively employed in the literature of enterotype studies [12, 13]. Substituting \( D(x_i||x_j) \) by \( \sum_{k=1}^{m} x_{ik} \ln(x_{ik}/x_{jk}) \), we have \( \beta_5 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})(\ln x_{ik} - \ln x_{jk})} \), indicating that it not only measures the absolute difference between two compositions, but also those with log-transformations. This offers J-divergence more power in quantifying the compositional changes than the measures that only consider the difference at original scale, such as Euclidean, Manhattan, and Jaccard, especially when the compositional values are small. Meanwhile, both JSD and rJSD also acquire this advantage by utilizing the Kullback-Leibler divergence. Comparing \( x_i \) with the middle rather than themselves makes JSD and rJSD less sensitive to the dispersion in compositions than J-divergence, however, it can help them remain the signals among noisy compositional changes from lower abundance, as shown later in the simulations. It is clearly seen that J-divergence does not apply to zero compositions in \( x_{ik} \) or \( x_{jk} \), neither JSD nor rJSD when both of them are zeros. In our analysis, we use the R package \textit{philentropy} [18] for computation, where \( x/0 \) is replaced by \( x/\epsilon \), \( x \ln(0) \) by \( x \ln(\epsilon) \), and \( \epsilon = 1e-5 \).

The compositional vectors of \( m \) dimensions vary within the \( m - 1 \) dimensional simplex space [21], for instance when \( m = 3 \), it is a triangle formed by three vertexes \((1,0,0)'\), \((0,1,0)'\) and \((0,0,1)'\) with its interior. Considering the limited variation of compositional vectors in radii, the angle contained by two vectors...
with the center at \(0\), Angular [23]

\[
\beta_8 = \arccos \left( \frac{\sum_{k=1}^{m} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{m} x_{ik}^2} \sqrt{\sum_{k=1}^{m} x_{jk}^2}} \right)
\]

reflects the dispersion between their compositions in a great extent. Note that Euclidean \(\beta_1\) is the chord length between two compositional vectors corresponding to the angle \(\beta_8\). With \(\lambda_{x_i} = \sum_{k=1}^{m} x_{ik}^2 / (\sum_{k=1}^{m} x_{ik})^2 = \sum_{k=1}^{m} x_{ik}^2 / \sum_{k=1}^{m} x_{jk}^2\) for compositional vectors, Horn-Morisita [15], abbreviated to Horn,

\[
\beta_9 = 1 - \frac{2 \sum_{k=1}^{m} x_{ik} x_{jk}}{(\lambda_{x_i} + \lambda_{x_j}) \sum_{k=1}^{m} x_{ik} \sum_{k=1}^{m} x_{jk}} = 1 - \frac{2 \sum_{k=1}^{m} x_{ik} x_{jk}}{\sum_{k=1}^{m} x_{ik}^2 + \sum_{k=1}^{m} x_{jk}^2},
\]

which is also Dice of Drost [18]. It is related to Angular by \(\beta_9 = 1 - 2 \cos \beta_8 \sqrt{\sum_{k=1}^{m} x_{ik}^2 \sum_{k=1}^{m} x_{jk}^2} / (\sum_{k=1}^{m} x_{ik}^2 + \sum_{k=1}^{m} x_{jk}^2)\) and Euclidean via \(\beta_9 = \beta_1^2 / (\sum_{k=1}^{m} x_{ik}^2 + \sum_{k=1}^{m} x_{jk}^2)\). These connections cause Angular, Horn and Euclidean may have similar clustering performance, but differ from each other when the variances among compositional elements \(\sum_{k=1}^{m} x_{ik}^2\) and \(\sum_{k=1}^{m} x_{jk}^2\) matters. Note that \(\sum_{k=1}^{m} x_{ik}^2\) and \(\sum_{k=1}^{m} x_{jk}^2\) are actually the squared radii of compared compositions.

It is arbitrary to take the radii off consideration or, in certain manners as Horn does, to account for the dissimilarity in the simplex space. The mapping from \(x_i\) to \(\sqrt{x_i}\), \(i = 1, 2, \ldots, N\), yields the projection of simplex space onto the unit hypersphere with the same radii and derives the beta diversity measures defined by the angle, Bhattacharyya[23]

\[
\beta_{10} = \arccos \left( \sum_{k=1}^{m} \sqrt{x_{ik}} \sqrt{x_{jk}} \right),
\]

and the chord length, Hellinger [23]

\[
\beta_{11} = \sqrt[2]{\left( \sum_{i=1}^{m} (\sqrt{x_{ik}} - \sqrt{x_{jk}}) \right)^2} = \sqrt{2(1 - \cos \beta_{10})}.
\]

Bhattacharyya and Hellinger are more reasonable in dealing with the effect of radii than Angular and Horn. Besides, the square root mapping leads to Bhattacharyya and Hellinger in favor of the dissimilarity between small composition values.

The log-transformations proposed by Aitchison [21] set up the foundations for compositions modelling, where \(arl(x_i) = (\ln(x_{i1}/x_{im}), \ldots, \ln(x_{i,m-1}/x_{im}))\) maps the \(m\)-dimensional simplex space \(S^m\) to \((m-1)\)-dimensional real space \(R^{m-1}\), \(clr(x_i) = (\ln(x_{i1}/g(x_i)), \ldots, \ln(x_{im}/g(x_i)))\)' with \(g(x_i) = (\prod_{k=1}^{m} x_{ik})^{1/m}\) converts \(S^m\) to a hyperplane of real space \(U^m = \{(u_1, \ldots, u_m) : u_1 + \cdots + u_m = 0\}\), and \(ilr(x) = V'clr(x)\) projects \(S^m\) to \(R^{m-1}\) with \(V'\) the transport of \(m \times (m-1)\) matrix \(V\) which columns form an orthonormal basis of \(U^m\) [27]. The dissimilarity measures being developed in the real space, such as Euclidean or Manhattan, can be applied to the transformed data and used as the beta diversity for compositional vectors. We notice that none of these three transformations is compatible with zero compositions. The R package compositions [28] calculates \(clr\) and \(ilr\) by omitting zeros out for transformation and then patching them back. Due to the close relationship between \(clr\) and \(ilr\), we only include \(ilr\) in the comparison analysis.

**Results**

**Simulations**

In order to investigate the performance of different beta diversity measures in clustering the population into subgroups, we set up three simulation experiments to mimic the microbial compositional data. All through the simulations, we set \(m = 500\) and \(K = 2\) clusters, each with \(n = 100\) samples. The adjusted Rand index (ARI) [29] is calculated to evaluate the accuracy of the clustering results.
Experiment 1

In the first experiment, we generate the compositional vectors using the log-normal distribution as stated in Lu, et al. [30]. Denoted by $LN(\mu, \Sigma)$ the multivariate log-normal distribution with mean $\mu$ and covariance matrix $\Sigma$, the random vector $z = (z_1, \ldots, z_m)'$ which is generated from $LN(\mu, \Sigma)$ is converted to the compositional vector via $x = z/\sum_{i=1}^{m} z_i$. We set $\mu = \mu_k$ in cluster $k$, $k = 1, 2$, and $\Sigma = (0.5^{(i-j)})_{m \times m}$ as the same in both clusters. The elements in $\mu_1$ is randomly assigned using the normal distributions $N(\mu, \sigma)$ with mean $\mu$ and standard deviation $\sigma$, and $\mu_2$ is constructed by manipulating $\mu_1$. Specifically, the first 50 (10% of total) elements of $\mu_1$ are independently generated from $N(9, 1)$, the following 50 (10%) from $N(6, 1)$ and the rest 400 (80%) from $N(3, 1)$, resulting in the compositions of cluster 1 with three levels of abundance, that are high around $1e-2$, median around $4e-4$ and low around $2e-5$. To explore how the compositional changes affect the clustering results using different beta diversity measures, we randomly select 10% of $\mu_1$ elements at different abundance levels and add perturbations to construct $\mu_2$. Corresponding to the perturbations from $N(0, 1)$ being added onto the high level $\mu_1$ elements, $N(0, 3)$ on the median, and $N(0, 5)$ on the low level, three datasets 1.1, 1.2 and 1.3 are simulated respectively.

The implemented perturbations cause compositional differences between clusters. The significance of these differences rather than their absolute values affects the clustering. In Figure 1, we present the absolute mean differences, marked by their p-values of the Wilcoxon signed-rank test between two clusters, along with all the compositions, as well as the resulted ARIs using different beta diversity measures in three datasets. It is shown that the significant differences between clusters in datasets 1.1, 1.2, and 1.3 mostly appear at the elements with high, median, or low abundance, with the smaller and smaller absolute mean differences.

According to the obtained ARIs, the beta diversity measures compared in our analysis are divided into five categories, i.e., Manhattan/Jaccard, J-divergence/JSD/rJSD, Euclidean/Angular/Horn, Bhattacharyya/Hellinger, and ilr-transformed Manhattan/Euclidean. As presented in Section 2, the measures within each category have a close relationship, resulting in very similar clusters. The ilr-transformed Manhattan/Euclidean seems wobbly in different datasets, which are ranked as the worst in dataset 1.1 but reach the highest ARIs in datasets 1.2 and 1.3. Besides, different abilities of the other measures in capturing different levels of compositional changes determine their clustering performance. Note that due to the compositional constraints, all the compositional changes are actually with small values. The
logarithm aids J-divergence, JSD, and rJSD to capture the tiny compositional changes, so does the square root that Bhattacharyya/Hellinger utilizes. In three datasets, J-divergence/JSD/rJSD and Bhattacharyya/Hellinger consistently show higher ARIs than Manhattan/Jaccard and Euclidean/Angular/Horn, and J-divergence performs as the best among them.

Although the measures within the same category perform similarly in many situations, it is remarkable that they also possibly present quite different ARIs, for instance in dataset 1.3, J-divergence gives much higher ARI than JSD/rJSD, and Euclidean yields significantly lower ARI than Angular/Horn. Nevertheless, no matter how ARIs vary within the categories, J-divergence/JSD/rJSD and Bhattacharyya/Hellinger can always produce top ARIs among the others.

**Experiment 2**

The multivariate Dirichlet distribution is a natural choice to generate compositional vectors. In the second experiment, we simulate the clustered datasets according to the multivariate Dirichlet distribution $D(\alpha)$, where $\alpha$ is the positive parameter of length $m$, and $\alpha = \alpha_k$ in the $k$th cluster, $k = 1, 2$. The first 50 (10% of total) elements of $\alpha_1$ are independently generated from the chi-square distribution $\chi^2(10)$ with degrees of freedom 10, the following 50 (10%) from $\chi^2(1)$ and the rest 400 (80%) from $\chi^2(0.1)$, corresponding to three levels of abundance in cluster 1, which are high around $2e-2$, median around $3e-4$, or low with over 85% less than $1e-10$ even zero values. As similarly as in Experiment 1, $\alpha_2$ is set up by manipulating $\alpha_1$. The random perturbations from $\chi^2(2)$, $\chi^2(1)$ or $\chi^2(1/2)$ are superposed on 50 high, medium, or low abundance elements, resulting in the dataset 2.1, 2.2 or 2.3, respectively. The absolute mean difference between two clusters and resulted in ARIs using different beta diversity measures in three datasets are presented in Figure 2.

In datasets 2.2 and 2.3, when the significant differences are located at median or low abundance elements, J-divergence/JSD/rJSD and Bhattacharyya/Hellinger show significantly higher ARIs than Manhattan/Jaccard and Euclidean/Angular/Horn. However, in dataset 2.1, they present less ARIs. Note that the compositions with high abundance in this experiment are more compact with fewer variances than in Experiment 1, causing the ARIs particularly resistant to the choice of beta diversity when the perturbations are added onto high abundance elements. We choose the perturbation parameter in dataset 2.1 to produce different ARIs to evaluate the clustering performance of different beta diversity measures. However, at this time, the intended compositional differences between two clusters are insufficient to overcome the enlarged noises from lower levels, when J-divergence/JSD/rJSD and Bhattacharyya/Hellinger are applied. The messed up clusters in dataset 2.1 using J-divergence/JSD/rJSD or Bhattacharyya/Hellinger reveal their preference on the compositional changes at low abundance elements. The *ilr*-transformed Manhattan/Euclidean can not show competitive ARIs in any datasets.

**Experiment 3**

It is worth noting that the perturbations on the parameter $\mu$ of log-normal distribution or $\alpha$ of Dirichlet distribution have no straightforward relationship with the compositional changes in $x_i$. Due to the correlations between compositions, the parameter perturbations at one level may also bring the compositional changes at the other levels. In order to minimize this impact on the conclusions, we set up the third experiment to simulate the datasets using the multinomial distribution $Mul(N, P)$, where $N$ is the total counts, and $P = (P_1, \ldots, P_m)'$ such that $P_i \geq 0$ and $\sum_{i=1}^{m} P_i = 1$.

First, we estimate $P$ and the distribution of $N$ by the Monte Carlo method. A total of $1e+4$ composi-
Figure 2 Delta is the absolute mean difference between the compositions of two clusters at each compositional coordinate; \(\times\) indicates the coordinates with p-values of Wilcoxon signed-rank test between two clusters being smaller than 0.001; \(\circ\) shows those coordinates with p-values between 0.001 and 0.01. The ARIs obtained in Datasets 2.1, 2.2 and 2.3 are respectively presented by the solid, dashed or dotted lines.

Figure 3 shows those coordinates with p-values between 0.001 and 0.01. The ARIs obtained in Datasets 2.1, 2.2 and 2.3 are respectively presented by the solid, dashed or dotted lines.

In the simulation settings of cluster 1 in the first experiment, representing an empirical distribution \(\hat{F}_N\) of \(N\), and a Monte Carlo estimate \(\hat{P}\) for \(P\), whose first 10% elements at high abundance around \(1e-2\), followed by 10% median around \(8e-4\), and 80% low around \(3e-5\). Then we let \(P_1 = \hat{P}\), and generate the compositions in cluster 1 by three steps, first generating \(N\) from \(\hat{F}_N\), then simulating the counts’ vector from \(Mul(N, P_1)\), and finally normalizing the counts as compositions by dividing them by their summation. A subset of \(s\) elements in \(P_1\), denoted by \(Q_s\), is collected and perturbed as \(Q'_s = Q_s \oplus \epsilon\), where \(\oplus\) is the addition operator in the simplex space [21], and \(\epsilon\) is a random sample from \(D(\gamma \cdot 1)\). Finally, \(P_2\) is given by replacing \(Q_s\) in \(P_1\) by \(Q'_s\), and used to generate cluster 2. We randomly select \(s = 50\) elements from those with high abundance for the perturbation \(\gamma = 1e+4\), median for \(\gamma = 1e+3\), or low for \(\gamma = 10\), providing datasets 3.1, 3.2 and 3.3 respectively. The absolute mean difference between two clusters and the obtained ARIs using different beta diversity measures in three datasets are presented in Figure 3.

Similarly as in Experiment 2, when the compositional changes are intended at high abundance level in dataset 3.1, J-divergence/JSD/rJSD and Bhattacharyya/Hellinger may give smaller ARIs than Manhattan/Jaccard and Euclidean/Angular/Horn. While the compositional changes move to lower levels in datasets 3.2 and 3.3, the advantage of JSD/rJSD and Bhattacharyya/Hellinger shows more and more significant. Note that in datasets 3.1 and 3.2, J-divergence provides significantly smaller ARIs than JSD/rJSD. Considering the discussion in Section 2 based on their definitions, J-divergence takes more risk that the signals are covered by the variations from lower abundance levels. With the highest ARIs from J-divergence in Experiment 1, it is implied that the performance of J-divergence is more data-dependent than that of JSD/rJSD.

Real Analysis

Autism Dataset

The gut samples of 278 children from the third affiliated hospital of Sun Yat-sen University, including 209 autism patients and 69 healthy controls, were analyzed using the 16S rRNA sequencing. The microbial genome annotation at the genus level results in the compositions of 278 samples among 780 OTUs, with 50%, 75%, 90%, and 95% quantiles of the compositional values being 5.2e-7, 4.4e-6, 6.2e-5, and 8.3e-4 respectively. In particular, 87.5% elements of the OTU table are zeros, while only 1.7% is higher than 0.01, and 3.5% is greater than 0.001. We used the aforementioned beta diversity measures to cluster the population into \(K = 2\) to 10 subgroups, and calculated the Caliński-Harabasz indices, Silhouette coefficients and prediction strength.
of these clustering results. These indices do not significantly increase as $K$ changes from $2$ to $10$. Therefore we set $K = 2$ in the following analysis.

The heatmap of ARIs between the clustering results using different beta diversity measures are presented in Figure 4. It is shown they gather into four groups, J-divergence/JSD/rJSD/Bhattacharyya/Hellinger, Manhattan/Jaccard, Euclidean/Angular/Horn, and the $ilr$-transformed Euclidean/Manhattan, with J-divergence departing slightly from the others in the same group, as they perform in simulations. The difference in compositions between two clusters identified by different beta diversity measures is further investigated. The numbers of OTUs whose adjusted p-values with false discovery rate (FDR) control are smaller than 0.05, and their mean OTU abundance is listed in Table 1. It is indicated that, except $ilr$-transformed measures, JSD/rJSD/Bhattacharyya/Hellinger detects the most numbers of OTUs with the adjusted p-values less than 0.05, followed by J-divergence. They are also more than that Manhattan/Jaccard and Euclidean/Angular/Horn present, especially at the elements whose mean abundance is lower than 0.001, demonstrating the superior capability of JSD/rJSD/Bhattacharyya/Hellinger in finding out the compositional changes at low abundance level.

**Human Gut Metagenomes**

Arumugam, et al. [12] first proposed the concept of enterotype by clustering 33 faecal samples using rJSD into three subgroups, according to 249 OTUs annotated at the genus level. They defined three enterotypes, which are named as Bacteroides, Prevotella, and
Table 1 Numbers of OTUs whose adjusted p-values with FDR control are smaller than 0.05, and their frequencies at different abundance levels.

| OTU mean abundance | Total | >0.001 | 0.001~1e-5 | <1e-5 |
|--------------------|--------|--------|------------|-------|
| Manhattan           | 23     | 14     | 9          | 0     |
| Jaccard             | 23     | 14     | 9          | 0     |
| J_divergence        | 29     | 18     | 11         | 0     |
| JSD                 | 35     | 19     | 15         | 1     |
| rJSD                | 35     | 19     | 15         | 1     |
| Euclidean           | 4      | 4      | 0          | 0     |
| Angular             | 7      | 5      | 1          | 1     |
| Horn                | 3      | 3      | 0          | 0     |
| Bhattacharyya       | 33     | 18     | 15         | 0     |
| Hellinger           | 33     | 18     | 15         | 0     |
| Manhattan-ilr       | 84     | 5      | 42         | 37    |
| Euclidean-ilr       | 82     | 16     | 39         | 27    |

Ruminococcus respectively, with the sample sizes of 19, 6 or 8 correspondingly. We apply all the included beta diversity measures in this paper to reanalyze the OTU table. The JSD/Bhattacharyya/Hellinger provides exactly the same clusters as rJSD. The J-divergence yields a unique but also quite similar clustering results to rJSD, only moving one sample in Ruminococcus to Bacteroides with ARI 0.90. The Manhattan/Jaccard and Euclidean/Angular/Horn obtain the same partitions too, but have two samples being adjusted from Prevotella to Bacteroides, with ARI 0.82 comparing to rJSD. However, ilr-transformed Manhattan/Euclidean presents very distinct clusters from rJSD, whose ARIs are only 0.02 and 0.15.

The consistency between the clusters from different beta diversity measures indicates that the compositional changes in this dataset are mainly located at a high abundance level. Comparing to rJSD, the assignment of FR_AD.3 from Ruminococcus to Bacteroides by J-divergence yields more significantly different OTUs with low abundances, such as Akkermansia and Gordonibacter. Their highest compositions are 0.09 and 0.003 respectively. Moving two samples, DA.AD.4 and FR.AD.6, from Prevotella to Bacteroides by Manhattan/Jaccard/Euclidean/Angular/Horn makes the number of OTUs whose adjusted p-values with FDR control are smaller than 0.1 reduce from 4 to 2, of which the compositions of 2 vanishing OTUs, Rhodopirillum and Escherichia/Shigella, are with low abundance less than $2e^{-5}$ and 0.035, respectively. It is illustrated that rJSD/JSD/Bhattacharyya/Hellinger focuses more on smaller compositional change at lower abundance elements than Manhattan/Jaccard/Euclidean/Angular/Horn, and J-divergence does further.

**Discussion**

In this paper, we proposed three simulation experiments to mimic the high-dimensional microbial clusters and investigate the performance of different beta diversity measures in clustering the compositional data into subgroups. The findings can be used to guide the choice of beta diversity, and their correspondingly resulted in clusters. There are still many other measures not included in this comparison analysis. Their performance can also be evaluated using the proposed simulation experiments, or inferred by exploring the connections between their defined formulas and the measures discussed here.

**Conclusion**

Besides the Aitchison transformations that do not apply to the high-dimensional compositional data analysis and show wobbly results, the beta diversity measures under investigation in this paper can be partitioned into two classes, according to their clustering performance, that is Manhattan/Bray-Curtis/Jaccard/Euclidean/Angular/Horn, and J-divergence/JSD/rJSD/Bhattacharyya/Hellinger. The former emphasizes on the compositional changes at higher abundance elements, while the latter favors the changes at lower abundance elements. For the gut microbial compo-
sitions with a large number of low abundance elements, we recommend the measures in the latter class for subgroup clustering. Among them, Bhattacharyya/Hellinger is further suggested, considering the dataset sensitivity of J-divergence and arbitrary $\epsilon$ setting of J-divergence/JSD/rJSD in dealing with zeros.

**List of abbreviations**

OTU: operational taxonomic unit; JSD: Jensen-Shannon divergence; rJSD: square root of Jensen-Shannon divergence; ARI: Adjusted Rand Index; FDR: false discovery rate

**Declarations**

**Ethics approval and consent to participate**
No ethics approval was required for the study.

**Consent for publication**
Not applicable.

**Availability of data and materials**
The Autism Dataset used for illustration during the current study is available from the corresponding author on reasonable request. The Human Gut Metagenomes analyzed during the current study is available online at: http://www.bork.embl.de/Docu/Arunugam_et_al_2011/data/tables/.

**Competing interests**
The authors declare that they have no competing interests.

**Funding**
This work was supported by the National Natural Science Foundation of China (11671409, 81873801), Science and Technology Program of Guangzhou, China (20200703011, 201903010040) for the corresponding authors’ work and publication cost of this paper. The funders did not play any roles in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Author’s contributions**
NY and BC designed the study, BC and XZ provided the Autism dataset. NY, XH and BP implemented the analysis. NY and XH wrote the manuscript. All authors read and approved the final manuscript.

**Acknowledgements**
Not applicable

**Author details**
²Child Development and Behavior Center, The Third Affiliated Hospital, Sun Yat-Sen University, 510630 Guangzhou, China. ³School of Mathematics, Sun Yat-sen University, 510275 Guangzhou, China.

**References**

1. Schlicker A, Beran G, Chresta CM, McWalter G, Pritchard A, Weston S, et al. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. BMC medical genomics. 2012;5(1):66.

2. Punt CJ, Koopman M, Vermeulen L. From tumour heterogeneity to advances in precision treatment of colorectal cancer. Nature reviews Clinical oncology. 2017;14(4):235–246.

3. Ogino S, Nishihara R, VanderWeele TJ, Wang M, Nishi A, Lochhead P, et al. The role of molecular pathological epidemiology in the study of neoplastic and non-neoplastic diseases in the era of precision medicine. Epidemiology (Cambridge, Mass). 2016;27(4):602.

4. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–352.

5. Schneider LS, Frangakis C, Dye LT, Devanand D, Marano CM, Mintzer J, et al. Heterogeneity of treatment response to citalopram for patients with Alzheimer’s disease with aggression or agitation: the CitAD randomized clinical trial. American Journal of Psychiatry. 2016;173(5):465–472.

6. Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. Nature Reviews Microbiology. 2011;9(4):279–290.

7. Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proceedings of the National Academy of Sciences. 2012;109(2):594–599.

8. Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T, et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell. 2013;155(7):1451–1463.

9. Mayer EA, Knight R, Mazmanian SK, Cryan JF, Tillisch K. Gut microbes and the brain: paradigm shift in neuroscience. Journal of Neuroscience. 2014;34(46):15490–15496.

10. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, et al. The NIH human microbiome project. Genome research. 2009;19(12):2317–2323.

11. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 2010;7(5):335.

12. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. nature. 2011;473(7346):174–180.

13. Carlson AL, Xia K, Azcarate-Peril MA, Goldman BD, Ahn M, Stynier MA, et al. Infant gut microbiome associated with cognitive development. Biological psychiatry. 2018;83(2):148–159.

14. Costea PI, Hildebrand F, Arumugam M, Bäckhed F, Blaser MJ, Bushman FD, et al. Enterotypes in the landscape of gut microbial community composition. Nature microbiology. 2018;3(1):8–16.

15. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS one.
16. Koleff P, Gaston KJ, Lennon JJ. Measuring beta diversity for presence–absence data. Journal of Animal Ecology. 2003;72(3):367–382.

17. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. The ISME journal. 2011;5(2):169–172.

18. Drost HG. Philentropy: information theory and distance quantification with R. Journal of Open Source Software. 2018;3(26):765.

19. Cha SH. Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences. 2007;1(4):300–307.

20. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, et al. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS computational biology. 2013;9(1).

21. Aitchison J. A concise guide to compositional data analysis. Girona: 2nd Compositional Data Analysis Workshop; 2003.

22. Legendre P, Legendre LF. Numerical ecology. Amsterdam: Elsevier; 2012.

23. Palarea-Albaladejo J, Martín-Fernández JA, Soto JA. Dealing with distances and transformations for fuzzy C-means clustering of compositional data. Journal of classification. 2012;29(2):144–169.

24. Izenman AJ. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. New York: Springer; 2008.

25. Kullback S, Leibler RA. On information and sufficiency. The annals of mathematical statistics. 1951;22(1):79–86.

26. Endres DM, Schindelin JE. A new metric for probability distributions. IEEE Transactions on Information theory. 2003;49(7):1858–1860.

27. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C. Isometric logratio transformations for compositional data analysis. Mathematical Geology. 2003;35(3):279–300.

28. Van den Boogaart KG, Tolosana-Delgado R. Analyzing compositional data with R. Berlin: Springer; 2013.

29. Hubert L, Arabie P. Comparing partitions. Journal of classification. 1985;2(1):193–218.

30. Lu J, Shi P, Li H. Generalized linear models with linear constraints for microbiome compositional data. Biometrics. 2019;75(1):235–244.

Additional Files

Additional file 1 — R Code
Functions to generate the compositional clusters in simulation experiment 1, 2, and 3, as described in details in the Results section.