frontiers in Genetics

METHODS
published: 14 December 2020
doi: 10.3389/fgene.2020.620143

A t-SNE Based Classification Approach to Compositional Microbiome Data

Xueli Xu[1], Zhongming Xie[2], Zhenyu Yang[1], Dongfang Li[3] and Ximing Xu[1,4]**

[1] School of Statistics and Data Science, Nankai University, Tianjin, China, [2] School of Mathematical Sciences, Nankai University, Tianjin, China, [3] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China, [4] Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin, Tianjin, China

As a data-driven dimensionality reduction and visualization tool, t-distributed stochastic neighborhood embedding (t-SNE) has been successfully applied to a variety of fields. In recent years, it has also received increasing attention for classification and regression analysis. This study presented a t-SNE based classification approach for compositional microbiome data, which enabled us to build classifiers and classify new samples in the reduced dimensional space produced by t-SNE. The Aitchison distance was employed to modify the conditional probabilities in t-SNE to account for the compositionality of microbiome data. To classify a new sample, its low-dimensional features were obtained as the weighted mean vector of its nearest neighbors in the training set. Using the low-dimensional features as input, three commonly used machine learning algorithms, logistic regression (LR), support vector machine (SVM), and decision tree (DT) were considered for classification tasks in this study. The proposed approach was applied to two disease-associated microbiome datasets, achieving better classification performance compared with the classifiers built in the original high-dimensional space. The analytic results also showed that t-SNE with Aitchison distance led to improvement of classification accuracy in both datasets. In conclusion, we have developed a t-SNE based classification approach that is suitable for compositional microbiome data and may also serve as a baseline for more complex classification models.

Keywords: microbiome data, dimension reduction, t-SNE, Aitchison distance, classification

OPEN ACCESS

Edited by:
Lixin Cheng,
Jinan University, China

Reviewed by:
Guosheng Han,
Xiangtan University, China
Weiwen Xue,
The Chinese University of Hong Kong, China

****Correspondence:***
Dongfang Li
loveli_biocc@163.com
Ximing Xu
ximing@nankai.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *22 October 2020*
Accepted: *25 November 2020*
Published: *14 December 2020*

Citation:
Xu X, Xie Z, Yang Z, Li D and Xu X (2020) A t-SNE Based Classification Approach to Compositional Microbiome Data. Front. Genet. 11:620143. doi: 10.3389/fgene.2020.620143

INTRODUCTION

The microbiome in human is involved in a large number of human essential functions, such as metabolism, nutrient intake and energy generation. In recent years, the microbiome has been found to be associated with numerous diseases, and the alterations in that by diet, disease, or environmental factors may impact on human health (Turnbaugh et al., 2006, 2009; Qin et al., 2012; Koeth et al., 2013). The next-generation sequencing technologies make it possible to study the microbiota composition through direct DNA sequencing, replacing classical microorganism

Abbreviations: t-SNE, t-distributed stochastic neighborhood embedding; LR, logistic regression; SVM, support vector machine; DT, decision tree; ACC, the classification accuracy; nMCC, the normalized Matthews correlation coefficient; AUC, the area under the receiver operating characteristic curve; AUPR, the area under the precision-recall curve.

study based on isolation and cultivation of specific species. Since the number of the sequence reads is difficult to generate equally for each sample in an experiment, the microbiome data is often required to be converted to the relative abundance for deeper analysis, resulting in compositional microbiome data (McMurdie and Holmes, 2014; Weiss et al., 2017). A single sample can often yield hundreds of millions of short sequencing reads, but for many species they are only observed in a small number of samples, so the microbiome data are typically characterized by high-dimensionality and multivariate sparsity (Li, 2015; Calle, 2019).

To gain a better understanding on the high-dimensional microbiome data, it is essential to reduce the data dimension in such a way that increases interpretability and minimizes information loss simultaneously. Traditional linear dimensionality reduction techniques, such as principal component analysis (PCA) (Hotelling, 1933; Abdi and Williams, 2010), nonnegative matrix factorization (NMF) (Lee and Seung, 1999; Jiang et al., 2012), and classical multidimensional scaling (MDS, also called principle coordinate analysis, PCoA) (Torgerson, 1952; Mugavin, 2008; Gonzalez and Knight, 2012), have difficulty capturing the nonlinear relationships in microbiome data due to their linear nature (Xu et al., 2016; Calle, 2019). In contrast, the nonlinear techniques have advantages in dealing with complex nonlinear datasets (Maaten et al., 2009). Among nonlinear dimension reduction algorithms, t-distributed stochastic neighbor embedding (t-SNE), developed by van der Maaten and Hinton (2008), has recently received increasing attention and has been applied to dimension reduction and visualization of microbiome data (Kostic et al., 2015), single-cell RNA-sequencing data (Linderman et al., 2019), bird songs (Deny et al., 2016), computational fluid dynamics (Wu et al., 2017), genomic data (Li et al., 2017), remote sensing images (Song et al., 2018) and many other application fields.

The t-SNE algorithm could efficiently project complex data sets onto a 2D or 3D plane, while the local structure of the data in the original high-dimensional space is preserved as much as possible. However, the t-SNE method does not provide a built-in way to map new data points to the corresponding low-dimensional representation, and hence it is hardly utilized for classification or regression tasks (Maaten, 2009). Some studies have attempted to cope with this out-of-sample extension problem by using neural networks for feature extraction and then perform classification on the mapped low-dimensional space from t-SNE (Maaten, 2009; Oliveira et al., 2018). However, there is little research on the application of t-SNE to the classification of microbiome data, which may be due to the unique characteristics of microbiome data, such as compositionality and relatively small sample size in many cases, limiting the performance of existing methods on such type of data.

In this article, we explore the potential of t-SNE for the classification of microbiome data, and propose a t-SNE based classification approach, which enables us to build classifiers and classify new samples in the reduced dimensional space. In our t-SNE algorithm, Aitchison distance, introduced by Aitchison (1986), is used to calculate the conditional probabilities for compositional microbiome data. To classify a new sample, its low-dimensional features are first obtained as the weighted mean vector of its nearest neighbors. Using the low-dimensional features as input, three commonly used methods-logistic regression (LR), support vector machine (SVM), and decision tree (DT) are then applied for classification in this study.

METHODS

t-Distributed Stochastic Neighbor Embedding

For a given set of p-dimensional samples $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N$ the similarity between sample $\boldsymbol{s}_j$ and sample $\boldsymbol{s}_i$ is represented by the conditional probability $p_{j|i}$, $i,\ j = 1, 2, \cdots, N$. For nearby samples, $p_{j|i}$ is relatively high, whereas for widely separated samples, $p_{j|i}$ will be almost zero. The conditional probability $p_{j|i}$ is given as

$$p_{j|i} = \frac{\exp\left(-\frac{d^2(\boldsymbol{s}_i,\boldsymbol{s}_j)}{2\sigma_i^2}\right)}{\sum_{k\neq i}\exp\left(-\frac{d^2(\boldsymbol{s}_i,\boldsymbol{s}_k)}{2\sigma_i^2}\right)} \quad \text{for } i \neq j, \text{ and } p_{i|i} = 0,$$

Where $d^2\left(\boldsymbol{s}_i, \boldsymbol{s}_j\right)$ is the square of the Euclidean distance between sample $\boldsymbol{s}_i$ and sample $\boldsymbol{s}_j$ and σ_i^2 is the variance of the Gaussian distribution that is centered on sample $\boldsymbol{s}_i$.

To circumvent the outlier problem, the symmetrized conditional probability between sample $\boldsymbol{s}_i$ and sample $\boldsymbol{s}_j$ is recommended,

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \quad \text{for } i \neq j, \text{ and } p_{ii} = 0.$$

the next step, t-SNE attempts to learn a d-dimensional map $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N$ $(d < p)$ that reflects the similarities p_{ij} between two samples $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ in the reduced dimensional space. The measure of pairwise similarities in the reduced dimensional space uses a student t-distribution rather than a Gaussian distribution to alleviate crowding problem, defined as

$$q_{ij} = \frac{\left(1 + d^2\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right)\right)^{-1}}{\sum_{k\neq l}\left(1 + d^2\left(\boldsymbol{z}_k, \boldsymbol{z}_l\right)\right)^{-1}} \quad \text{for } i \neq j, \text{ and } q_{ii} = 0,$$

where q_{ij} represents the local structure of the data points in the reduced dimensional space. To select the map points so that the two similarity matrices, P and Q, are as similar as possible, the location of the sample $\boldsymbol{z}_i$ is determined by minimizing the Kullback–Leibler divergence (Kullback and Leibler, 1951) between the low-dimensional and high-dimensional similarity distributions Q and P,

$$KL\left(P \parallel Q\right) = \sum_{i\neq j} p_{ij} \log\frac{p_{ij}}{q_{ij}},$$

using a gradient-descent method. It is worth noting that the gradient descent is an iterative optimization algorithm, and thereby the iterative number (*iter*) should be optimized. The t-SNE method has another main parameter, the perplexity (*per*), which governs the variance of the Gaussian σ_i^2 appearing in the

conditional probability $p_{j|i}$. For a detailed introduction of the t-SNE algorithm, see the literature (Maaten and Hinton, 2008).

t-SNE Based Classification for Compositional Data

The procedure of the proposed approach is shown in **Figure 1**, including two main parts, the implementation of t-SNE with Aitchison distance and the out-of-sample extension.

t-SNE With Aitchison Distance

The sample space of compositional data is simplex (Calle, 2019). Three important conditions should be fulfilled for a proper analysis of compositions: permutation invariance, scale invariance and sub-compositional coherence (Aitchison, 1986). In fact, Euclidean distance cannot meet the principles of scale invariance and sub-compositional coherence, which may lead to spurious correlations among the abundances of the different taxa (Aitchison, 1986; Li, 2015; Calle, 2019). On the other hand, Aitchison distance, which has been proved to satisfy all these criteria, was often deemed to be a solution to most problems related to compositional data (Aitchison, 1992; Calle, 2019).

The Aitchison distance d_a between two p-dimensional vectors $\mathbf{x} = (x_1, x_2, \cdots, x_p)$ and $\mathbf{y} = (y_1, y_2, \cdots, y_p)$, is defined as,

$$d_a(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{p} \left(\ln \frac{x_i}{g(\mathbf{x})} - \ln \frac{y_i}{g(\mathbf{y})} \right)^2 \right]^{1/2},$$

where g(·) denotes the geometric mean. The conditional probability $p_{j|i}$ in t-SNE is substituted by $p_{j|i}(i, j = 1, 2, \cdots, N, j = 1, 2, \cdots, N)$ calculated as,

$$p_{j|i,a} = \frac{\exp\left(-\frac{d_a^2(s_i, s_j)}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{d_a^2(s_i, s_k)}{2\sigma_i^2}\right)}, \quad \text{for } i \neq j, \text{ and } p_{i|i,a} = 0.$$

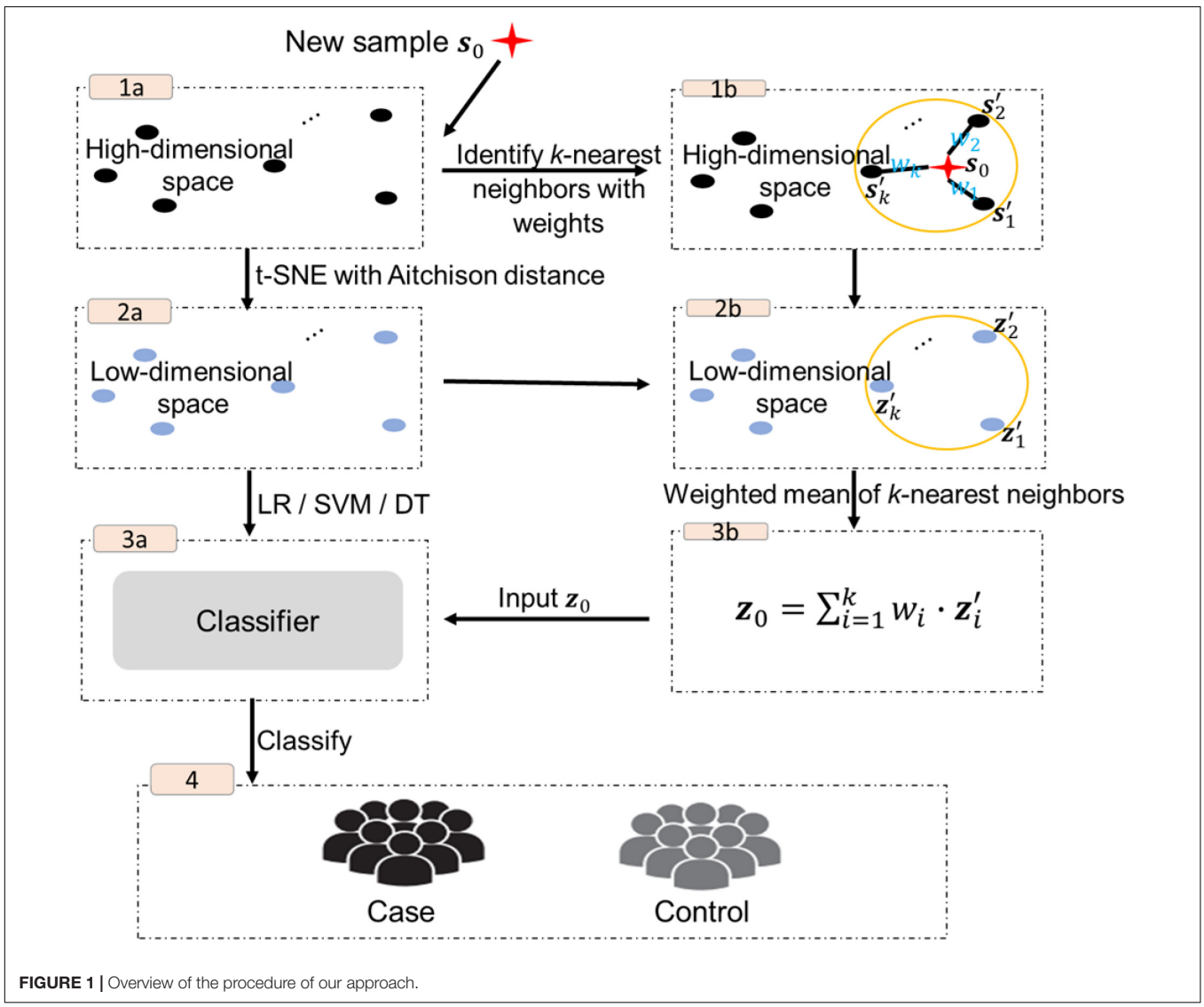

FIGURE 1 | Overview of the procedure of our approach.

The high-dimensional feature vectors of the original set are submitted to t-SNE with Aitchison distance for data dimensionality reduction (step 1a–2a in **Figure 1**), and the corresponding reduced dimensional data $\boldsymbol{z}_1, \ \boldsymbol{z}_2, \ \cdots, \boldsymbol{z}_N$ in $\mathbb{R}^d$ are used to build classifiers. In this study, we consider three widely used classification algorithms: logistic regression (LR), support vector machine (SVM) or decision tree (DT) (step 3a in **Figure 1**).

Out-of-Sample Extension

Given a new p-dimensional sample $\boldsymbol{s}_0$ in S^P, its k nearest neighbors in the original data are first identified based on the conditional probabilities between $\boldsymbol{s}_0$ and the original set of samples (step 1b in **Figure 1**),

$$sp_{0,i} = \frac{\exp(-d_a^2(\boldsymbol{s}_0, \boldsymbol{s}_i)/2\sigma_i^2)}{\sum_{h\neq i}\exp(-d_a^2(\boldsymbol{s}_i, \boldsymbol{s}_h)/2\sigma_i^2)}, \quad i = 1, 2, \cdots, N.$$

Let $sp_{0,(1)} > sp_{0,(2)} > \cdots > sp_{0,(k)} > \cdots > sp_{0,(N)}$ denote the set of N ordered probabilities, and $\boldsymbol{z}_1^{'}, \ \boldsymbol{z}_2^{'}, \ \cdots, \ \boldsymbol{z}_k^{'}$ denote the low-dimensional representations of the original data with the largest k conditional probabilities, which will be labeled as the k nearest neighbors of $\boldsymbol{s}_0$ step 2b in **Figure 1**). Then the low-dimensional representation of $\boldsymbol{s}_0$ step 3b in **Figure 1**) is given by

$$\boldsymbol{z}_0 = \sum_{i=1}^{k} w_i \cdot \boldsymbol{z}_i',$$

where $w_i = \frac{sp_{0,(i)}}{\sum_{i=1}^{k} sp_{0,(i)}}$ denotes the weight of $\boldsymbol{z}_i'$.

In the final step $\boldsymbol{z}_0$, used as an input to the classifier for prediction (step 4 in **Figure 1**).

Selection of Optimal Parameters

In t-SNE, there are several parameters that need to be tuned for good performance, such as the perplexity *per*, which is defined as a smooth measure of the effective number of neighbors. It has been suggested that a typical value for this parameter is between 5 and 50 (Maaten and Hinton, 2008). In practice, proper tuning of *per* requires users to understand the inner working of the t-SNE method as well as to have hands-on experience. In our study, the tuning of parameters can be achieved based on the performance of classifications. In particular, a grid search with fivefold cross-validation is used to tune the parameters, including the perplexity *per*, maximum number of iterations *iter*, output dimensionality *dim*, and number of neighbors *k*. The optimal combination of parameters is selected via maximizing mean cross-validation accuracy.

In addition, for the parameters of the three classification algorithms applied in the reduced dimensional space, LR is trained by tuning the lambda based on minimum mean cross-validated error. In the SVM model, the radial basis kernel is used. The two tuning parameters (gamma and cost) are chosen by minimizing the mean cross-validation error as the best combination for seven values from 10^{-6} to 10^{1} for the gamma and five values from 1 to 5 for the cost. For DT, the optimized decision tree is obtained by evaluating the cross-validated error using a grid search method and then determining the best set of hyperparameters, including min split, min bucket, max depth, and complexity. The rest of the unmentioned parameters uses the default setting in the R package.

Model Evaluation

To compare the performance of different model settings, we use the area under the receiver operating characteristic curve (AUC) which represents the trade-off between the true positive rate (specificity) vs. the false positive rate (1-sensitivity), the commonly used classification metric-accuracy (ACC) which were defined by the ratio of the samples correctly classified to the total samples, as well as two other criteria capable of overcoming the class imbalance issue, the area under the precision-recall curve (AUPR) which represents the trade-off between the precision vs. the recall, and the normalized Matthews correlation coefficient (nMCC) which projects the original range of MCC [-1, 1] into the interval [0, 1] (Matthews, 1975; Chicco and Jurman, 2020),

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}},$$

$$\mathrm{nMCC} = \frac{\mathrm{MCC}+1}{2},$$

where $\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$,

TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively. For the given data, we randomly split the data to the training and test sets in 80/20 ratio. All of the training samples were randomly divided into five sets, four of which were employed for constructing the classification model, and the remaining one as the validation set was used to validate the model for obtaining the optimal parameters. The generalization performance results were reported by ACC, nMCC, AUC, and AUPR, which were measured on test data that was held out during the training of t-SNE with Aitchison distance or t-SNE with Euclidean distance.

Implementation of the Proposed Approach

The proposed approach was implemented on R software (version 4.0.2), where t-SNE was performed using the R package tsne, LR was implemented using the R package glmnet, SVM was executed using the R package e1071, and DT was implemented using the R package rpart. In addition, both AUC and AUPR were calculated using the R package PRROC, and MCC was calculated using the R package mltools. The R code could be found at https://github.com/Xuxl2020/t-SNE-classifier.

Application to Microbiome Data

The proposed approach was performed on two microbiome datasets from diverse body sites: (1) the Mycoplasma pneumoniae (MP) infection data (Zhou et al., 2020) and (2) the idiopathic central precocious puberty (ICPP) data (Dong et al., 2020). The MP infection data was oropharyngeal (OP) microbiota derived from the MP infection study on 99 Chinese children, including 40 patients (diagnosed as MP infection, Case group) and 59

age-matched healthy children (Control group). The ICPP data was fecal microbiota from 25 girls (Case group) with idiopathic central precocious puberty and 23 healthy girls (Control group) in China. All microbiota data were generated from the Miseq platform by sequencing the V3-V4 hypervariable region of microbial 16S rDNA and were annotated with the RDP database and then calculating relative abundance for each sample in the genus taxonomic level. Both data were the compositional data. The MP infection and ICPP data contained 728 and 146 features (genus), respectively.

RESULTS

Data Visualizations With t-SNE

The results of t-SNE 2D map for MP infection data (*per* = 30, *iter* = 2,000) and ICPP data (*per* = 15, *iter* = 2,000) are illustrated in **Figure 2**. For MP infection data (**Figure 2A**), t-SNE with Aitchison distance constructs a map in which the separation between the case and control groups is almost perfect. In contrast, t-SNE with Euclidean distance produces a map in which there is no clear boundary between different groups. For ICPP data (**Figure 2B**), the map produced by t-SNE with Aitchison distance contains a few points that are clustered with the wrong group, probably due to more complex composition and more distinct individual differences in gut microbiota. Again, none of the groups are clearly separated in the t-SNE with Euclidean distance map. The computation time (seconds) of t-SNE with Aitchison distance and t-SNE with Euclidean distance for both microbiome datasets was also provided in **Supplementary Table 1**.

Impact of Output Dimensionality on Classification Performance

For MP infection data, the optimal parameters selected are: the perplexity *per* = 30, maximum number of iterations *iter* = 2,000, and number of neighbors k = 7. To examine the impact of the output dimensionality on the classification performance, the results of the proposed approach (using Aitchison distance) for *dim* = 2, 3, 5, and 7 were presented in **Figure 3A**. ACCs, AUCs, nMCCs, and AUPRs were relatively small at *dim* = 2. When *dim* = 3, the ACCs were increasing to 0.95, 0.96 and 0.92 for LR, SVM, and DT, respectively, the AUCs were increasing to 0.99, 1.00, and 0.93 for LR, SVM, and DT, respectively, the nMCCs

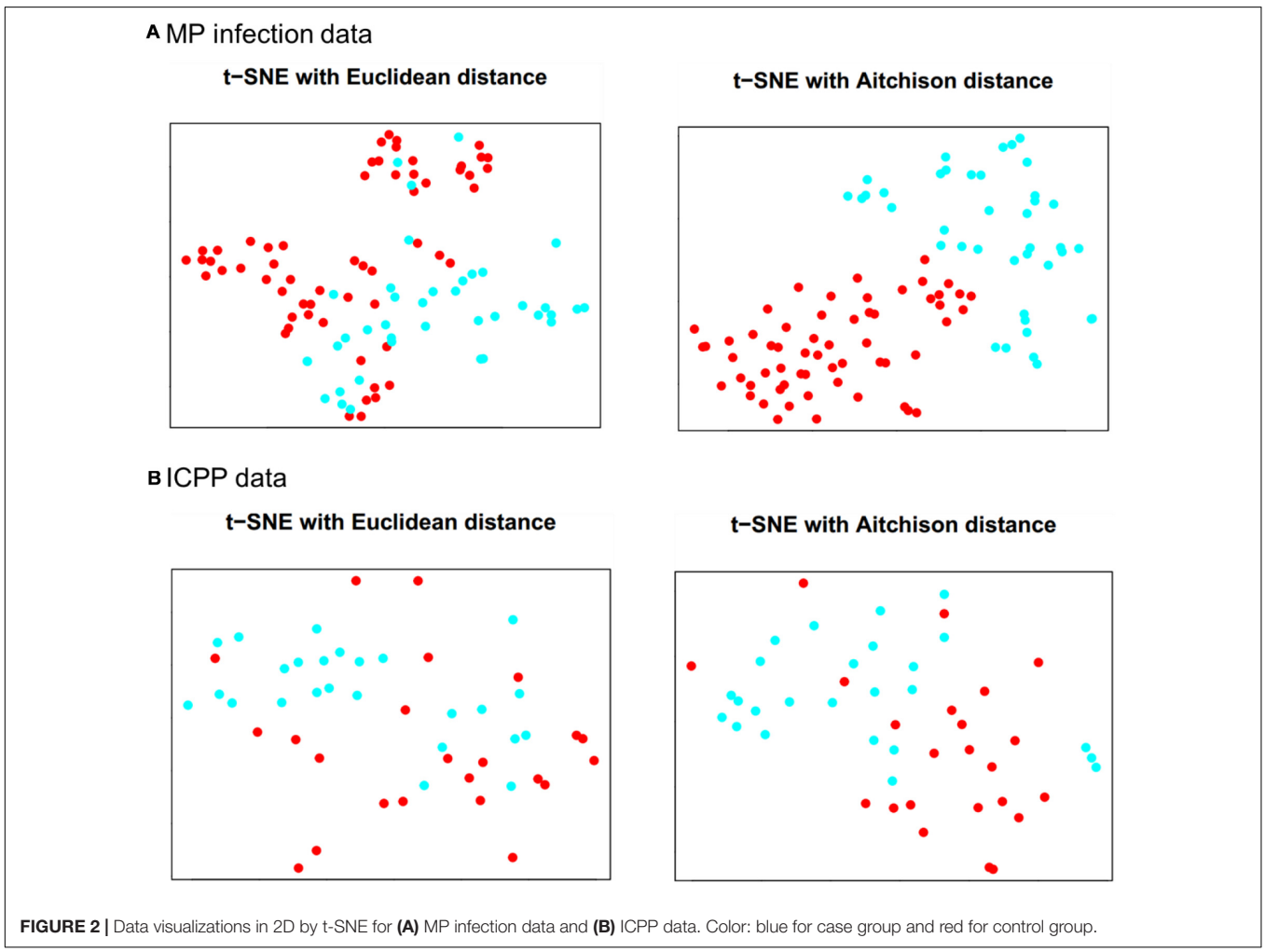

FIGURE 2 | Data visualizations in 2D by t-SNE for **(A)** MP infection data and **(B)** ICPP data. Color: blue for case group and red for control group.

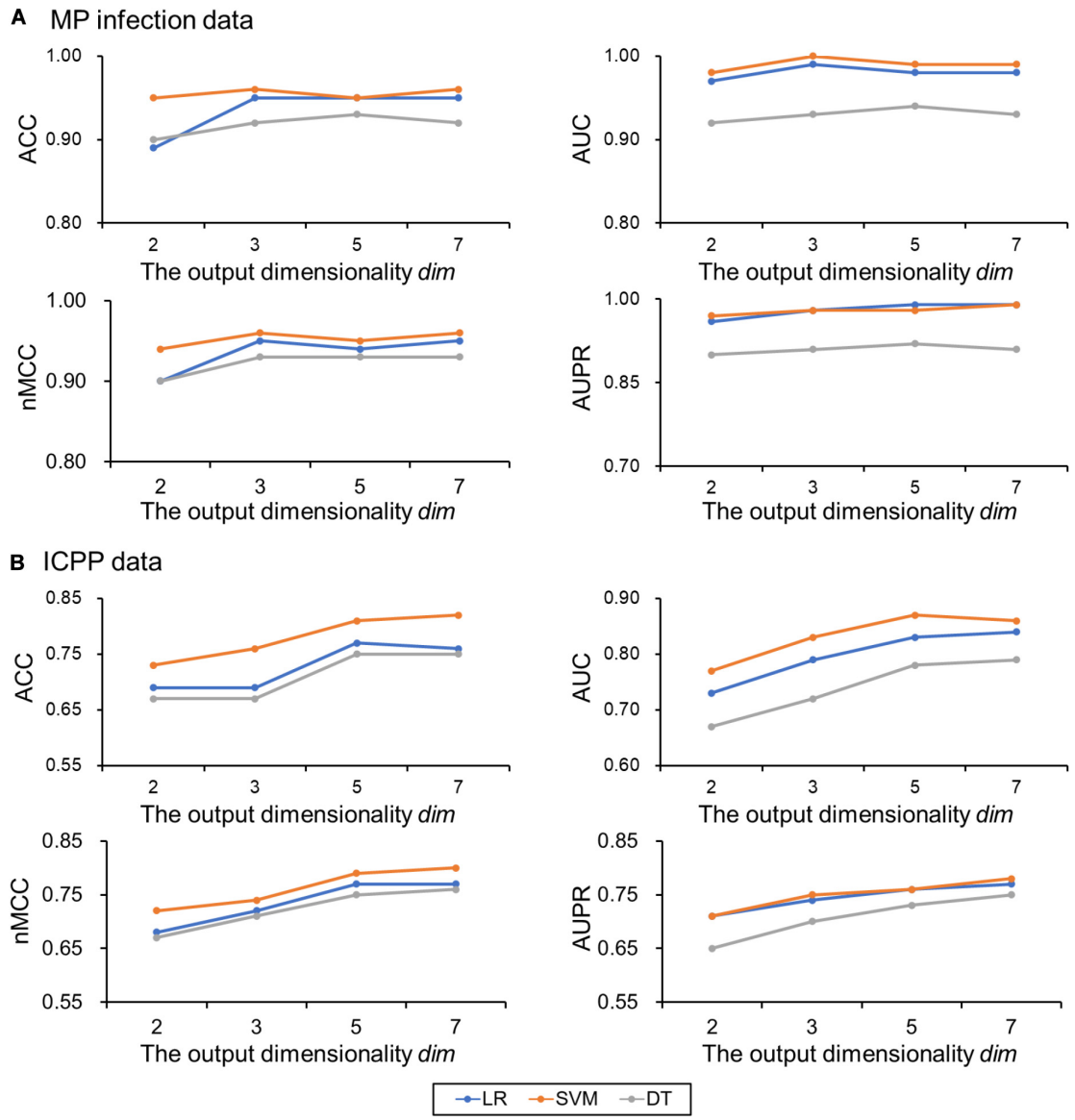

FIGURE 3 | Classification performances on the test data change with the output dimensionality *dim*. **(A)** MP infection data and **(B)** ICPP data. Color: blue for logistic regression (LR), orange for support vector machine (SVM), and gray for decision tree (DT). ACC, the classification accuracy; AUC, the area under the receiver operating characteristic curve; nMCC, the normalized Matthews correlation coefficient; AUPR, the area under the precision-recall curve.

were increasing to 0.95, 0.96, and 0.93 for LR, SVM, and DT, respectively, and the AUPRs were increasing to 0.98, 0.98, and 0.91 for LR, SVM, and DT, respectively, which were similar to the values at *dim* = 5 and 7.

For ICPP data, the optimal parameters selected are: the perplexity *per* = 15, maximum number of iterations *iter* = 2,000, and number of neighbors k = 7. The results of the proposed approach (using Aitchison distance) for *dim* = 2, 3, 5, and 7 were presented in **Figure 3B**. ACCs, AUCs, MCCs, and AUPRs were relatively small at *dim* = 2 and 3. When *dim* = 5, the ACCs were increasing to 0.77, 0.81, and 0.75 for LR, SVM, and DT, respectively, the AUCs were increasing to 0.83, 0.87, and 0.78 for LR, SVM, and DT, respectively, the nMCCs were increasing to 0.77, 0.79, and 0.75 for LR, SVM, and DT, respectively, and the AUPRs were increasing to 0.76, 0.76, and 0.73 for LR, SVM, and DT, respectively, which were similar to the values at *dim* = 7. The detailed results for different output dimensions were summarized in **Table 1**.

Impact of Different Distance Measures on Classification Performance

For MP infection data with *dim* = 3, compared to the results using Euclidean distance, the proposed approach using Aitchison distance increased the ACC by 9% for LR, 10% for SVM, and 10% for DT, respectively, increased the nMCC by 9% for LR,

TABLE 1 | Performance of classification models on the test set.

			MP infection				ICPP			
		dim =	2	3	5	7	2	3	5	7
ACC	ED	LR	0.72	0.87	0.88	0.87	0.55	0.59	0.68	0.67
		SVM	0.81	0.87	0.85	0.85	0.61	0.63	0.64	0.64
		DT	0.75	0.84	0.83	0.84	0.60	0.57	0.63	0.64
	AD	LR	0.89	0.95	0.95	0.95	0.69	0.69	0.77	0.76
		SVM	0.95	0.96	0.95	0.96	0.73	0.76	0.81	0.82
		DT	0.90	0.92	0.93	0.92	0.67	0.67	0.75	0.75
nMCC	ED	LR	0.72	0.87	0.87	0.87	0.60	0.61	0.71	0.72
		SVM	0.78	0.87	0.86	0.86	0.63	0.64	0.72	0.72
		DT	0.73	0.82	0.82	0.82	0.62	0.60	0.72	0.71
	AD	LR	0.90	0.95	0.94	0.95	0.68	0.72	0.77	0.77
		SVM	0.94	0.96	0.95	0.96	0.72	0.74	0.79	0.80
		DT	0.90	0.93	0.93	0.93	0.67	0.71	0.75	0.76
AUC	ED	LR	0.77	0.91	0.91	0.90	0.63	0.72	0.77	0.75
		SVM	0.85	0.90	0.89	0.90	0.70	0.74	0.75	0.75
		DT	0.75	0.85	0.82	0.81	0.66	0.68	0.74	0.74
	AD	LR	0.97	0.99	0.98	0.98	0.73	0.79	0.83	0.84
		SVM	0.98	1.00	0.99	0.99	0.77	0.83	0.87	0.86
		DT	0.92	0.93	0.94	0.93	0.67	0.72	0.78	0.79
AUPR	ED	LR	0.72	0.86	0.88	0.90	0.66	0.71	0.74	0.74
		SVM	0.72	0.85	0.86	0.87	0.66	0.71	0.74	0.75
		DT	0.69	0.81	0.82	0.81	0.64	0.69	0.71	0.74
	AD	LR	0.96	0.98	0.99	0.99	0.71	0.74	0.76	0.77
		SVM	0.97	0.98	0.98	0.99	0.71	0.75	0.76	0.78
		DT	0.90	0.91	0.92	0.91	0.65	0.70	0.73	0.75

ACC, the classification accuracy; nMCC, the normalized Matthews correlation coefficient; AUC, the area under the receiver operating characteristic curve; AUPR, the area under the precision-recall curve; AD, the models using Aitchison distance; ED, the models using Euclidean distance; LR, logistic regression; SVM, support vector machine; DT, decision tree.

10% for SVM, and 13% for DT, respectively, increased the AUC by 9% for LR, 11% for SVM, and 9% for DT, respectively, and increased the AUPR by 14% for LR, 16% for SVM, and 14% for DT, respectively. For MP infection data with *dim* = 5, compared to the results using Euclidean distance, the proposed approach using Aitchison distance increased the ACC by 13% for LR, 27% for SVM, and 19% for DT, respectively, increased the nMCC by 8% for LR, 10% for SVM, and 4% for DT, respectively, increased the AUC by 8% for LR, 16% for SVM, and 5% for DT, respectively, and increased the AUPR by 3% for LR, 3% for SVM, and 3% for DT, respectively. The detailed results for the comparisons were summarized in **Supplementary Table 2**.

Classification Performance in Original and Reduced Dimension Space

To compare the classification performances of the classifiers built in original and reduced dimensional space, we also used the three algorithms (LR, SVM, and DT) to build classifiers in the original dimensional space. For MP infection data (99 samples with 728 features), the ACCs were 0.95, 0.94, and 0.92 for LR, SVM, and DT, respectively, the nMCCs were 0.95, 0.95, and 0.94 for LR, SVM, and DT, respectively, the AUCs were 0.99, 0.98, and 0.94 for LR, SVM, and DT, respectively, and the AUPRs were 0.98, 0.97, and 0.94 for LR, SVM, and DT, respectively (**Figure 4A**). For ICPP data (48 samples with 146 features), the ACCs were 0.75, 0.76, and 0.73 for LR, SVM, and DT, respectively, the nMCCs were 0.77, 0.76, and 0.75 for LR, SVM, and DT, respectively, the AUCs were 0.80, 0.85, and 0.76 for LR, SVM, and DT, respectively, and the AUPRs were 0.75, 0.75, and 0.76 for LR, SVM, and DT (**Figure 4B**). In comparison with the results of the proposed approach (**Figure 4** and **Table 1**), we found that the application of dimensionality reduction technique, t-SNE with Aitchison distance, resulted in no reduction in classification accuracy.

DISCUSSION

In this work, we proposed a classification approach based on t-SNE, taking into account the compositional characteristic of microbiome data. The application of the proposed approach was illustrated on two disease-associated microbiome datasets, and demonstrated good classification performance on both datasets. Although we focused on the classification tasks, the proposed approach could be also used for regression analysis in the reduced dimensional space by t-SNE.

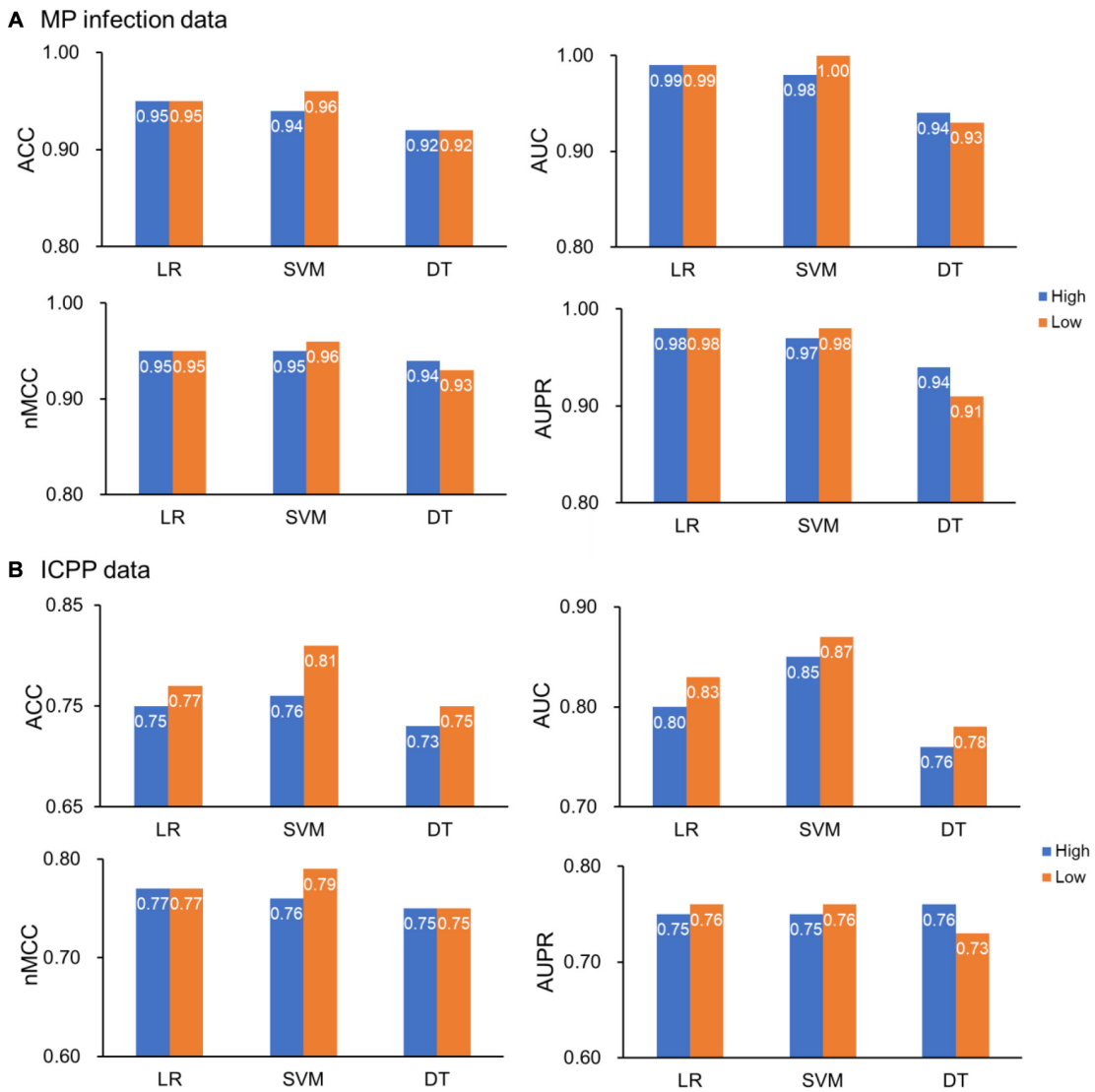

FIGURE 4 | Classification performances in original and reduced dimensional space. **(A)** MP infection data and **(B)** ICPP data. Color: blue (High) for original high-dimensional data and orange (Low) for low-dimensional data. ACC, the classification accuracy; AUC, the area under the receiver operating characteristic curve; nMCC, the normalized Matthews correlation coefficient; AUPR, the area under the precision-recall curve. Note: in the reduced dimensional space, *dim* = 3 and *dim* = 5 for MP infection data and ICPP data, respectively.

In both microbiome datasets, using the Aitchison distance to calculate the conditional probabilities in t-SNE made the case and control groups appear more clearly separated, compared to the t-SNE map with Euclidean distance (**Figure 2**), whereas the use of Aitchison distance did not increase the computation time of t-SNE, compared to the use of Euclidean distance (**Supplementary Table 1**). The classification performance was also improved for the proposed approach by using Aitchison distance (**Table 1** and **Supplementary Table 2**). This was probably because Aitchison distance satisfies the principles of scale invariance and sub-compositional coherence, and hence is more suitable for compositional data analysis, as stated by Aitchison (1986). In future work, the impact of Aitchison distance on t-SNE and the classification performance of compositional data should be studied with more rigorous mathematical theories.

In our data analysis we found that SVM outperformed LR and DT in many settings, which may be related to the fact that the classifier generalization ability varies for different types of data (Hastie et al., 2009; Oliveira et al., 2018), impacting the prediction accuracy. On the other hand, the optimization parameters often have a large impact on classification performance, and a reasonable and feasible tuning method is necessary. In our approach, a grid search with fivefold cross-validation is employed for tuning parameters, iterating over many possible parameter combinations to maximize the classification accuracy (ACC). Grid search is one of the most widely used techniques and allows us to have a transparent parameter selection (Huang et al., 2012).

Moreover, considering comparable sequencing depth (about 30,000 tags) and sequencing quality (Q20 > 95%) in each sample, the difference in model performances for different datasets may be attributed to the following factors: (1) the ICPP data is intestinal microbiome which has a higher microbial load and more complex microbial composition relative to that in oropharynx (mean of Shannon index 2.078 in ICPP data vs. 1.831 in MP data); (2) the smaller sample size in ICPP data probably limit the performance of model.

In this study, we only considered LR, SVM, and DT, whereas other models such as neural networks may have better classification performance depending on the datasets. In addition, the same idea in our approach may be also applied to other manifold learning dimension reduction techniques, such as the unified manifold approximation and projection (UMAP) developed by McInnes and Healy (2018). As a preliminary study, we compared the model performances using UMAP based on Euclidean distance and Aitchison distance. The results were presented in **Supplementary Table 3** for both microbiome datasets, showing that the use of Aitchison distance led to more accurate classification, similar to the proposed t-SNE based approach. More comprehensive and theoretical investigations on different dimension reduction methods will be conducted in future work.

The visualizations by t-SNE may be helpful for our understanding on the performance of the proposed approach. As shown in **Figure 2A** for the MP infection data, the control and case groups were well separated, and a satisfactory classification performance was expected at *dim* = 2 or 3. On the other hand, the t-SNE map at 2D for the ICPP data (**Figure 2B**) did not show a clear clustering pattern for the case and control groups which suggested that a relatively large value of *dim* would be needed to achieve a satisfactory classification performance.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

XLX, DL, and XMX designed the research. XLX and ZX performed data analysis. XLX and ZY created the figures. XLX, ZX, and ZY wrote the manuscript. XMX and DL advised and were the senior supervisor of the project. All authors read and approved the submitted manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (Grant No. 11701294) and Fundamental Research Funds for the Central Universities of China.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.620143/full#supplementary-material

REFERENCES

Abdi, H., and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisc. Rev. Comput. Stats* 2, 433–459. doi: 10.1002/wics.101

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Aitchison, J. (1992). On criteria for measures of compositional difference. *Mathe. Geol.* 24, 365–379. doi: 10.1007/BF00891269

Calle, M. L. (2019). Statistical analysis of metagenomics data. *Genom. Inform* 17:e6. doi: 10.5808/GI.2019.17.1.e6

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* 21:6. doi: 10.1186/s12864-019-6413-7

Deny, S., Mackevicius, E., Okubo, T., Berman, G., Shaevitz, J., and Fee, M. (2016). Learning stable representations in a changing world with on-line t-SNE: Proof of concept in the songbird. *Proc. Int. Conf. Learn. Represent.* 4, 1–4. doi: 10.5709/acp-0038-8

Dong, G., Zhang, J., Yang, Z., Feng, X., Li, J., Li, D., et al. (2020). The association of gut microbiota with idiopathic central precocious puberty in girls. *Front. Endocrinol.* 10:941. doi: 10.3389/fendo.2019.00941

Gonzalez, A., and Knight, R. (2012). Advancing analytical algorithms and pipelines for billions of microbial sequences. *Curr. Opin. Biotechnol.* 23, 64–71. doi: 10.1016/j.copbio.2011.11.028

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educat. Psychol.* 24, 417–441. doi: 10.1037/h0071325

Huang, Q., Mao, J., and Liu, Y. (2012). "An improved grid search algorithm of svr parameters optimization," in *Proceedings of the 14th IEEE International Conference on Communication Technology* (New York: IEEE), 1022–1026. doi: 10.1109/ICCT.2012.6511415

Jiang, X., Langille, M. G. I., Neches, R. Y., Elliot, M., Levin, S. A., Eisen, J. A., et al. (2012). Functional biogeography of ocean microbes revealed through non-negative matrix factorization. *PLoS One* 7:e43866. doi: 10.1371/journal.pone.0043866

Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., et al. (2013). Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* 19, 576–585. doi: 10.1038/nm.3145

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, H., Hyötyläinen, T., Hämäläinen, A., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* 17, 260–273. doi: 10.1016/j.chom.2015.01.001

Kullback, S., and Leibler, R. A. (1951). On Information and Sufficiency. *Anna. Math. Stat.* 22, 79–86. doi: 10.1214/aoms/1177729694

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Li, H. Z. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stats Appl.* 2, 73–94. doi: 10.1146/annurev-statistics-010814-020351

Li, W., Cerise, J., Yang, Y., and Han, H. (2017). Application of t-SNE to human genetic data. *J. Bioinform. Comput. Biol.* 15:1750017. doi: 10.1142/S0219720017500172

Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., and Kluger, Y. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* 16, 243–245. doi: 10.1038/s41592-018-0308-4

Maaten, L. V. D. (2009). Learning a parametric embedding by preserving local structure. *J. Mach. Learn. Res.* 5, 384–391.

Maaten, L. V. D., and Hinton, G. E. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2431–2456.

Maaten, L. V. D., Postma, E., and Herik, J. V. D. (2009). Dimensionality reduction: A comparative review. *Rev. Literat. Arts Am.* 10, 66–71.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

McInnes, L., and Healy, J. (2018). *UMAP: uniform manifold approximation and projection for dimension reduction*. Preprint at https://arxiv.org/abs/1802.03426.

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Mugavin, M. E. (2008). Multidimensional scaling: A brief overview. *Nurs. Res.* 57, 64–68. doi: 10.1097/01.NNR.0000280659.88760.7c

Oliveira, F., Machad, A., and Andrade, A. O. (2018). On the Use of t-Distributed Stochastic Neighbor Embedding for Data Visualization and Classification of Individuals with Parkinson's Disease. *Comput. Math. Methods Med.* 2018:8019232. doi: 10.1155/2018/8019232

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60. doi: 10.1038/nature11450

Song, W., Wang, L., Liu, P., and Choo, R. K. (2018). Improved t-SNE based manifold dimensional reduction for remote sensing data processing. *Mult. Tools Appl.* 78, 1–16. doi: 10.1007/s11042-018-5715-0

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika* 17, 401–419. doi: 10.1007/BF02288916

Turnbaugh, P. J., Hamady, M., Yatsunenk, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480–484. doi: 10.1038/nature07540

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031. doi: 10.1038/nature05414

Weiss, S., Xu, Z. Z., Peddada, S., Amnon, A., Kyle, B., Antonio, G., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Wu, J., Wang, J., Xiao, H., and Ling, J. (2017). "Visualization of high dimensional turbulence simulation data using t-SNE," in *Proceedings of the 19th AIAA Non-Deterministic Approaches Conference* (Reston: AIAA), doi: 10.2514/6.2017-1770

Xu, W., Schultz, T., and Xie, R. (2016). An efficient visualisation method for exploring latent patterns in large microbiome expression data sets. *Int. J. Data Min. Bioinform.* 15:47. doi: 10.1504/IJDMB.2016.076016

Zhou, Q., Xie, G., Liu, Y., Wang, H., Yang, Y., Shen, K., et al. (2020). Different nasopharynx and oropharynx microbiota imbalance in children with Mycoplasma pneumoniae or influenza virus infection. *Microbial. Pathogen.* 144:104189. doi: 10.1016/j.micpath.2020.104189

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.