Binning application in low-dimensional metagenomic sequences: performance of Barnes-Hut t-Stochastic Neighbor Embeddings, assessment of internal cluster validity indices

J. Ceballos[1], L. Ariza-Jiménez[2], O. L. Quintero[2], and N. Pinel[1]

[1]Biodiversity, Evolution, & Conservation Res. Group, Universidad EAFIT, Medellín, Colombia
[2]Mathematical Modelling Research Group, Universidad EAFIT, Medellín, Colombia

Abstract

Metagenomic studies aim to reconstruct the structure of microbial communities through the use of DNA sequence data of complex composition. To this end, they generally embed multidimensional data into low dimensional spaces followed by a binning process. The performance of the dimensionality reduction techniques, the clustering methods, and the internal cluster validity indices vary depending on the biological, statistical and computational features that are part of the metagenomic analysis, yet it is seldom evaluated systematically. The explained problematic was explored through an unsupervised binning of metagenomic DNA sequences, based on the Subtractive and Fuzzy c-means algorithms applied to the two- and three-dimensional metagenomic sequences obtained via the Barnes-Hut t-Stochastic Neighbor Embedding (BH-SNE) algorithm in conjunction with Principal Component Analysis (PCA), with the aim of assessing the performance of the BH-SNE including and not including a preliminary PCA, besides the assessment of four Internal Cluster Validity Indices (ICVI) that conditioned the clustering procedure. In addition, the assessment of the ICVIs demonstrated that the Silhouette index had the best performances based on the median values of the F measure. Moreover, Silhouette index was also the most consistent index obtaining the highest values of F median in two- and three-dimensional treatments. In the case of high AAI ranges, the Silhouette index had equal results compared with Calinski-Harabasz index in terms of highest values of F median in three-dimensional treatment, although there were differences between their performance in two-dimensional treatments. In particular, Dunn index generated the worst performances in the low AAI percentages, while the Davies-Bouldin index was the worst in high AAI percentages. Additionally, the Dunn and Davies-Bouldin indices were the most consistent generating the lowest F median values. Moreover, the results of this research suggest that the biology of the metagenomic sequences could have an incidence over the best ICVIs performances. Finally, it was possible to determine that the highest F median values were obtained by the ICVIs in 3D embeddings, with equal results for BH-SNE including and not including preliminary PCA. Furthermore, it was also demonstrated that there was no significance between the results that included or not included a preliminary PCA. In terms of consistency, it was not possible to determine which was the most consistent treatment (2D or 3D embedding with BH-SNE including and not including preliminary PCA) that led the ICVIs to obtaining the best and worst F median results.

Keywords: metagenomics, clustering, cluster validity index, fuzzy clustering, embedding, BH-SNE

1 Introduction

Metagenomic studies aim to reconstruct the ecological structure of microbial communities, based on complex samples of DNA, without relying on the cultivation of the constituent populations. In general, there are two metagenomic sequencing approaches: targeted, which sequences an amplified gene that serves as a taxonomic marker; and shotgun sequencing, which sequences mixed DNA fragments without specific amplification [1] aiming for Whole Metagenome Sequencing (WMS). The targeted approach is limited to the analysis of groups for which taxonomically informative genetic markers are known and can be amplified [2]; the WMS approach reduces this bias by considering all regions of the genome [3]. WMS is often used to analyze unknown mixed microbial communities without a previous knowledge of its genomes, gaining insight into the functional and evolutionary processes shape a community.

Metagenome reconstruction from sequences obtained with high throughput technologies such as Illumina requires assembly and binning, in order to identify the constituent populations. The complexity of microbial communities interferes with the quality of genome assemblies [4]. The lack of prior knowledge about the number of microbial groups and their relative abundance makes the binning procedures a very challenging task [5]. The different levels of relatedness among populations complicate the assembly process [6].

In WMS studies, it is common to embed sequence data into a low-dimensional space, and then perform a binning process [7, 8, 9]. Barnes-Hut t-Stochastic Neighbor Embedding (BH-SNE) is a non-linear dimensionality reduction method introduced as an approach to enable the visualization and subsequent binning of genomic fragments [7]. BH-SNE applies an initial Principal Component Analysis, retaining by default the 50 most informative dimensions for embedding into a two or three dimensional space [9, 10], even though most studies default to the former. Data loss through dimensionality reduction algorithms such as BH-SNE or PCA may compromise the interpretability of the results [11], and yet their use has been justified on the improvement of genome reconstruction originated from the reduction of the data complexity [4].

Generally speaking, there are two main binning strategies for metagenomic shotgun DNA sequences: supervised, or taxonomy dependent; and unsupervised, or taxonomy independent, where taxonomy refers to the group of close organisms that are included in the same biological classification [12]. Supervised methods require prior knowledge of the analyzed species and a reference data set; unsupervised methods focus on the extraction of taxon parameters using machine learning techniques such as clustering to analyze higher dimensionality data sets, without prior knowledge about the species [3]. Methods used in unsupervised approaches to cluster re-projected, low-dimensional metagenomic sequence data include clustering K-means and Adaptive Resonance Theory (ART) [13]; Mcluster (improving K-means) [14]; Average Linkage Hierarchical clustering with Euclidean metric [15]; Cluster size insensitive FCM method (csiFCM), AbundaceBin and MetaCluster3.0 [16]; improved version of Fuzzy C-Means method (IFCM) [17]; assembly-assisted method MetaProb [18].

Evaluating the quality of the groups resulting from clustering algorithms is a subject that has been implemented inconsistently in metagenomic studies [7, 8, 10, 19, 20]. Various Internal Cluster Validity Indeces (ICVI), such as the Calinski-Harabasz (CH) index, the Dunn index (DI), the Davies-Bouldin index (DB), The Silhouette index (SI), and others have been employed [21], albeit in a mostly idiosyncratic manner. Gisbrecht et al [19] proposed the Dunn Index as an appropriate metric of quality, while Lux et al [8] recommended the use of the Davies-Boulding index, but resulting from limited comparisons.

A difficulty with evaluating in a comparative manner the performance of the metagenomic sequence binning approaches, is the lack of consistent metrics for genome relatedness in the construction of the synthetic test communities [7, 8, 10, 19, 20]. The amino acid identity (AAI) among orthologous genes shared by two genomes provides a measure of genome relatedness [22], an by extension a taxonomic proxy. While not a perfect measure, given its inability to account for unique genes, the use of AAI as a criterion in the construction of synthetic test communities could aid in assessing the real capacity of the embedding and unsupervised methods in metagenomic analysis.

The stochasticity in many of the algorithms employed in metagenomic binning pipelines has also not been considered in performance assessments, and consequently statistics on the consistency of the methods are not available. In particular, BH-SNE is inherently an stochastic algorithm, since its core objective function is minimized using a gradient descent optimization which is initiated randomly [23], then different runs of the BH-SNE algorithm provides different embeddings on metagenomic data.

The present work has as a starting point the research [9], expanding the exploration of BH-SNE into three dimensional embeddings and the impact of the preliminary PCA step. We adopt a larger and more diverse data set than previously used [9], structuring it based on AAI relationships among genomes. We develop a statistically rigorous evaluation scheme, implementing four different validity indeces and multiple replicates. In summary, this work aims to address the following questions:

- When applied to metagenomic data embedding, is there any effect in the performance of BH-SNE, positive or negative, from the use of the preliminary PCA dimensionality reduction?
- Are 3D BH-SNE embeddings equally fitted for unsupervised binning of metagenomic sequence data as the traditional 2D embeddings?
- Among the following internal validity indices: Calinski-Harabasz, Dunn, Davies-Bouldin, and Silhouette; which index presents the best performance and consistency?
- In the case of the best dimensionality reduction strategy in conjunction with the most outstanding internal validity index, can the index performance be affected by the biology of the metagenomic sequences?

This article consists of the following sections: *theoretical framework*, where the theory on embedding techniques, clustering methods and internal cluster validity indices and others are explained; *materials and methods*, which includes the methodology implemented to achieve the aims of the research; *results and discussion*, where the results of all the

procedures are presented and discussed; and *conclusions*, where are summarized the main achievements of the investigation.

2 Theoretical framework

2.1 Dimensionality reduction techniques

Two-dimensionality reduction methods are considered in this study, namely BH-SNE and PCA. BH-SNE is a non-linear dimensionality reduction technique designed to embed high-dimensional data into a space of two or three dimensions [10]. In the case of PCA, it reduces the dimensionality of multivariate data preserving the major quantity of relevant information through principal components, which are a linear combination of the original data [24, 25]. In addition, PCA can be used to pre-process high-dimensional data in order to speed up the computation of the BH-SNE method and suppress some noise [10].

The BH-SNE method has been proposed as a variant of another dimensionality reduction technique named t-Distributed Stochastic Neighbor Embedding (t-SNE), due to the optimization problem associated with the latter technique requires a time $\mathcal{O}\left(N^2\right)$, while the BH-SNE overcomes this issue by requiring a time $\mathcal{O}\left(N \log N\right)$ [23].

Given a set of objects $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ in a high-dimensional space $\mathbb{R}^r$, the goal of the t-SNE [23] is to learn an s-dimensional embedding in which each object is represented by a point of a set $\mathcal{E} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, with $\mathbf{y}_i \in \mathbb{R}^s$. In particular, this algorithm looks for an embedding such that nearby and distant points correspond to similar and dissimilar objects in the original space, respectively. To this end, the joint probabilities p_{ij} that measure the pairwise similarity between objects $\mathbf{x}_i$ and $\mathbf{x}_j$ by symmetrizing two conditional probabilities are defined as follows:

$$p_{i|j} = \frac{\exp\left(-d\left(\mathbf{x}_i, \mathbf{x}_j\right)^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-d\left(\mathbf{x}_i, \mathbf{x}_k\right)^2 / 2\sigma_i^2\right)} \,, \qquad p_{i|i} = 0 \tag{1}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}, \tag{2}$$

Where $d\left(\mathbf{x}_i, \mathbf{x}_j\right)$ is a function that computes a distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ (usually the Euclidean distance), and σ_i is the bandwidth of the Gaussian kernel whose value varies per object [26]. Likewise, joint probabilities q_{ij} that measure the pairwise similarity between the two corresponding points $\mathbf{y}_i$ and $\mathbf{y}_j$ are defined as follows:

$$q_{ij} = \frac{\left(1 + d\left(\mathbf{y}_i, \mathbf{y}_j\right)^2\right)^{-1}}{\sum_{k \neq i} \left(1 + d\left(\mathbf{y}_i, \mathbf{y}_k\right)^2\right)^{-1}} \,, \qquad q_{ii} = 0 \,. \tag{3}$$

However, unlike the normalized Gaussian kernel used to measure similarities in the original space, a normalized Student-t kernel is used in the embedding in order to account for the difference in volume between both spaces.

To learn the aforementioned embedding, i.e. the location of points y_i in $\mathbb{R}^s$, the t-SNE technique tries to match these two distributions as well as possible by minimizing a cost function

$$C\left(\mathcal{E}\right) = KL\left(P||Q\right) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) \,, \tag{4}$$

Which is a sum of Kullback-Leibler divergences between the original (p_{ij}) and induced (q_{ij}) joint distributions for each object. In particular, this cost function can be minimized by descending along the gradient $\frac{\partial C}{\partial \mathbf{y}_i}$, which it is initialized randomly, as follows:

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} \left(p_{ij} - q_{ij}\right) q_{ij} Z \left(\mathbf{y}_i - \mathbf{y}_j\right) \,, \tag{5}$$

Where

$$Z = \sum_{k \neq i} \left(1 + d\left(\mathbf{y}_i, \mathbf{y}_j\right)^2\right)^{-1} \,. \tag{6}$$

The Barnes-Hut-based variant of the t-SNE implements a sparse approximation to the probabilities p_{ij}. This approximation is viable since probabilities p_{ij} corresponding to dissimilar objects $\mathbf{x}_i$ and $\mathbf{x}_j$ are nearly infinitesimal

when similarities are computed using a Gaussian kernel. Then, if $\mathcal{N}_i$ is the subset of nearest neighbors of $\mathbf{x}_i$, the BH-SNE algorithm redefines the pairwise similarities p_{ij} as follows:

$$p_{i|j} = \begin{cases} \frac{\exp\left(-d(\mathbf{x}_i,\mathbf{x}_j)^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-d(\mathbf{x}_i,\mathbf{x}_k)^2/2\sigma_i^2\right)}\,, & \text{if } j \in \mathcal{N}_i \\ 0\,, & \text{otherwise} \end{cases} \tag{7}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}\,, \tag{8}$$

Where each subset $\mathcal{N}_i$ is found in $\mathcal{O}\left(N\log N\right)$ time by building a vantage-point tree on the data set [27].

Afterwards, the BH-SNE address the problem of the t-SNE cost function (Equations 7 and 8) which can be rewritten as:

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4\sum_{j\neq i} p_{ij}\left(1 + d\left(\mathbf{y}_i,\mathbf{y}_j\right)^2\right)^{-1} - 4\sum_{j\neq i} q_{ij}^2 Z\left(\mathbf{y}_i - \mathbf{y}_j\right) \tag{9}$$

Notice that the first term of the right in Equation (9) is computationally efficient in comparison with the second one. In this regard, the BH-SNE algorithm constructs a quadtree on the current embedding and the quadtree is traversed using a deep-first search, then at every node in the quadtree, the algorithm decides whether the corresponding cell can be used as a "summary" for the gradient contributions of all points in that cell. In particular, for a given point $\mathbf{y}_i$ and any two points $\mathbf{y}_j$ and $\mathbf{y}_k$ inside a cell with N_{cell} points that is sufficiently small and sufficiently far away from the first point, we have that $d\left(\mathbf{y}_i,\mathbf{y}_j\right) \approx d\left(\mathbf{y}_i,\mathbf{y}_k\right) \gg d\left(\mathbf{y}_j,\mathbf{y}_k\right)$. Therefore, the gradient contributions of all points $\mathbf{y}_j$ inside the aforementioned cell is given by $\sum_{j=1}^{N_{cell}} q_{ij}^2 Z\left(\mathbf{y}_i - \mathbf{y}_j\right)$, and can be approximated by $N_{cell} q_{i,cell}^2 Z\left(\mathbf{y}_i - \mathbf{y}_{cell}\right)$, where $\mathbf{y}_{cell}$ is the center-of-mass of the embedding points that are located inside the cell. Once the above equations are applied, the second term of the right in Equation (9) is approximated in $\mathcal{O}\left(N\log N\right)$ time instead of $\mathcal{O}\left(N^2\right)$ time.

2.2 Clustering methods

Generally speaking, clustering aims to divide a given set of n objects $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ into clusters such that objects in the same cluster are similar and objects in different clusters are dissimilar to each other [28]. If the similarity is evaluated in terms of distances, then clustering aims to minimize intra-cluster distances while inter-cluster distances are maximized. The clustering methods known as Subtractive clustering and FCM have been selected based on the applications presented by [9].

2.2.1 Subtractive clustering

Subtractive clustering [29] is a method used to estimate the centers of clusters of a data set. The algorithm can be used as a preliminary method for more complex clustering algorithms that need to know in advance the initial clusters centers of a data set. However, the algorithm computational cost grows as the square of the number of the objects in the data set.

Estimated centers by this algorithm are objects of the data set itself and they are identified based on a measure of potential, iteratively. In the first iteration ($i = 1$), the potential $D_i(x_j)$ for an object $\mathbf{x}_j$ is a function of its distance to the remaining objects in the data set:

$$D_i\left(\mathbf{x}_j\right) = \sum_{l=1}^{n} \exp\left(-\frac{\left\|\mathbf{x}_j - \mathbf{x}_l\right\|^2}{\left(r_a/2\right)^2}\right)\,, \tag{10}$$

Where r_a is a positive radius used to define a neighborhood around each object in order to measure its potential value. Thus, an object with many nearby objects (i.e. with a high density of surrounding objects) will have a high potential value. In particular, the first estimated cluster center $\mathbf{c}_i$ is selected as the object having the highest potential $D_i(x_j)$. Then, a value of potential is subtracted from each object as a function of its distance from the first center c_i:

$$D_{i+1}\left(\mathbf{x}_j\right) = D_i\left(\mathbf{x}_j\right) - D_i\left(\mathbf{c}_i\right)\exp\left(-\frac{\left\|\mathbf{x}_j - \mathbf{c}_i\right\|^2}{\left(r_b/2\right)^2}\right)\,. \tag{11}$$

In particular, r_b is a second positive radius used to define the neighborhood of a found cluster that will experiment a subtraction (reduction) of its potential. In practice, it is suggested to set r_b larger than r_a (e.g. $r_b = 1.25r_a$) in order to ensure the identification of cluster centers that are sufficiently separated [29].

Once the potential of the objects near the first cluster is reduced, the object with the highest remaining potential D_i is selected as the next estimated cluster center c_i in a second iteration ($i = i + 1$). Henceforward, the method iterates k steps (until $i \leq k$) between the procedure of reducing the potential of objects around found cluster centers and finding new potential cluster centers based on proposed criteria. A strategy to improve the process of finding new clusters by considering the possibility of rejecting some of them and automatically establish the number of clusters is also developed in [29].

2.2.2 Fuzzy C-Means

FCM is a clustering algorithm that considers that each point of the data set can belong to more than one cluster with a certain degree of membership [11]. Indeed, FCM aims to partition a set of n data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ into C fuzzy clusters such that the following cost function J is minimized:

$$J = \sum_{i=1}^{C} \sum_{j=1}^{n} (u_{ij})^m \left\| \mathbf{x}_j - \mathbf{c}_i \right\|^2 \tag{12}$$

Where $0 \leq u_{ij} \leq 1$ and $\sum_{i=1}^{C} u_{ij} = 1$; $m > 1$ is a fuzzification parameter; and $\mathbf{c}_i$ is the center of i-th fuzzy cluster.

The cost function J can be minimized with an iterative procedure that updates the following equations:

$$\mathbf{c}_i = \frac{\sum_{j=1}^{n} (u_{ij})^m \mathbf{x}_j}{\sum_{j=1}^{n} (u_{ij})^m} , \tag{13}$$

$$u_{ij} = \left[\sum_{l=1}^{C} \left(\frac{\left\| \mathbf{c}_i - \mathbf{x}_j \right\|}{\left\| \mathbf{c}_l - \mathbf{x}_j \right\|} \right)^{2/(m-1)} \right]^{-1} . \tag{14}$$

The u_{ij} are the entries of the fuzzy membership or partition matrix U, and each of them describes the degree of membership of the data point $\mathbf{x}_j$ in the i-th fuzzy cluster with a value (between 0 and 1) that is inversely proportional to the distance of this data point to the cluster center $\mathbf{c}_i$. The degree of membership of a data point to every fuzzy cluster is fixed, so the sum of elements in each column of the matrix U is equal to 1.

The fuzzification parameter m affects the entries of the fuzzy membership matrix U and determines the level of cluster fuzziness. As m increases, the entries of U decreases, and thus clusters become fuzzier. If m is close to 1, the entries of U converge to 0 or 1, so clusters become crispier and FCM behaves like the conventional K-means algorithm. By default, the value assigned to m is 2 [28].

2.3 Cluster validity indices

Cluster validation is the process of estimating how well the partitions of an input data set fits its underlying structure [21]. Cluster validation can be performed by computing external and internal indices, where the external indices measure the agreement between the results from the first clustering procedure and ground truth clustering, whereas internal indices measure the goodness of the performed clustering without external information by examining just the partitioned data [11].

Based on the best performing indices in [21], the following internal indices are taken in account: the Calinski-Harabasz (CH) index, the Dunn index (D), Davies-Bouldin index (DB) and The Silhouette index (S).

The notation for the explanation of the internal indices is given by: Given a set X of N objects represented as vectors in an S-dimensional Euclidean space $X = \{x_1, x_2, ..., x_N\} \epsilon \mathcal{R}^{\mathcal{S}}$. The clustering or partition of X, is noted as $\mathcal{U} = \{U_1, U_2, \ldots, U_K\}$ with K clusters. The distance between the i-th and j-th binned objects is represented by $d(x_i, x_j)$; d also refers to a distance between other objects. $\overline{u_k} == \frac{1}{|u_k|} \sum_{x_i \in u_k} x_i$ is the mean vector that specifies the center of the k-th cluster.

2.3.1 Calinski-Harabasz index

The Calinski-Harabasz index (CH) identifies the clusters of points in a multidimensional Euclidean space based on the density of clusters [30]. The cohesion is calculated taking into account the distances from points to the centroid of the cluster where they belong, whilst the inter-cluster distance is obtained from the distance between centroids and the global centroids [21].

The CH index is given by the Equation (15) that relates the inter-cluster and intra-cluster sums of squares [21, 31, 32].

$$CH(U) = \frac{N-k}{k-1} \frac{\sum_{u_k \in U} |u_k| \, d(\overline{u_k}, \overline{X})}{\sum_{x_i \in u_k} d(x_i, \overline{u_k})} \tag{15}$$

Where $\overline{X}$ Equation (16), is the mean vector of the whole data set.

$$\overline{X} = \frac{1}{N \sum_{x_i \in X} x_i} \tag{16}$$

2.3.2 Dunn index

The Dunn index (D) estimates the quality of the binning methods, it is a ratio-type index between the clustering separation (inter-cluster distance) and the clustering cohesion (intra-cluster distance), which are estimated by the nearest neighbor distance and the maximum cluster diameter, respectively. Therefore, large values of D corresponds to good clusters; The D of the clustering $\mathcal{U}$ is given by Equation (17) [21, 33].

$$\mathrm{D}\left(\mathcal{U}\right) = \frac{\max\limits_{1 \leq i,j \leq K, i \neq j} \left\{\delta\left(U_i, U_j\right)\right\}}{\max\limits_{1 \leq k \leq K} \left\{\triangle\left(U_k\right)\right\}} \, , \tag{17}$$

Where:

$$\triangle\left(U_k\right) = \max_{\mathbf{x_i}, \mathbf{x_j} \in U_k} \left\{d\left(\mathbf{x_i}, \mathbf{x_j}\right)\right\} \tag{18}$$

$$\delta\left(U_i, U_j\right) = \min_{\mathbf{x_i} \in U_i, \mathbf{x_j} \in U_j} \left\{d\left(\mathbf{x_i}, \mathbf{x_j}\right)\right\} \tag{19}$$

$\triangle\left(U_k\right)$ represents the diameter of the k-th cluster of $\mathcal{U}$, defined as the maximum distance between any two objects in the cluster. On the other hand, $\delta\left(U_i, U_j\right)$ is the distance between clusters U_i and U_j, defined as the minimum distance between any two objects in different clusters [34].

2.3.3 Davies-Bouldin index

The Davies-Bouldin index (DB) is a measure of the appropriateness of data partitions, it indicates the similarity of clusters, bringing the opportunity to compare the relative appropriateness of the overall number of clusters; the measure is independent of the number of clusters and does not vary depending on the clustering method [35].

The Davies-Bouldin index (DB) of the object $\mathbf{x}_i$ in X and assigned to cluster U_k is defined in Equation (20) [21, 35].

$$DB(U) = \frac{1}{k} \sum_{u_k \in U} \max_{u_l \in U \setminus U_k} \left\{ \frac{S(u_k) + S(u_l)}{d(\overline{u_k}, \overline{u_l})} \right\} \tag{20}$$

Where:

$$S(u_k) = \frac{1}{|u_k|} \sum_{x_i \in u_k} d(x_i, \overline{u_k}) \tag{21}$$

Understanding $\overline{u_l}$ as the mean vector that specifies the center of the l-th cluster, calculated like $\overline{u_k}$ see notation paragraph. Due to that, the DB index is the mean value among all the clusters [31].

2.3.4 Silhouette index

The Silhouette index (S) is a confidence indicator on the membership of an object to the cluster to which it is assigned in comparison with the remaining clusters [33], it indicates the goodness of the clustering over a data set. The S of the object $\mathbf{x}_i$ in X and assigned to cluster U_k is defined as $S(x_i)$ in Equation (22) [21].

$$\mathrm{S}\left(\mathbf{x}_i\right) = \frac{\mathrm{b}\left(\mathbf{x}_i, U_k\right) - \mathrm{a}\left(\mathbf{x}_i, U_k\right)}{\max\left\{\mathrm{a}\left(\mathbf{x}_i, U_k\right), \mathrm{b}\left(\mathbf{x}_i, U_k\right)\right\}} \tag{22}$$

Where:

$$\mathrm{a}\left(\mathbf{x}_i, U_k\right) = \frac{1}{|U_k|} \sum_{\mathbf{x}_j \in U_k} d\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{23}$$

$$\mathrm{b}\left(\mathbf{x}_i, U_k\right) = \min_{U_l \in U \setminus U_k} \left\{ \frac{1}{|U_l|} \sum_{\mathbf{x}_j \in U_l} d\left(\mathbf{x}_i, \mathbf{x}_j\right) \right\} \tag{24}$$

For a clustering $\mathcal{U}$ of X, its (overall average) silhouette index $S(\mathcal{U})$ is defined in Equation (25).

$$\mathrm{S}\left(\mathcal{U}\right) = \frac{1}{N} \sum_{U_k \in \mathcal{U}} \sum_{\mathbf{x}_i \in U_k} \frac{\mathrm{b}\left(\mathbf{x}_i, U_k\right) - \mathrm{a}\left(\mathbf{x}_i, U_k\right)}{\max\left\{\mathrm{a}\left(\mathbf{x}_i, U_k\right), \mathrm{b}\left(\mathbf{x}_i, U_k\right)\right\}} \tag{25}$$

$\mathrm{a}\left(\mathbf{x}_i, U_k\right)$ represents the average dissimilarity of object $\mathbf{x}_i$ in U_k with the remaining members of the same cluster. Since its value indicates how well an object is assigned to a certain cluster, a small value indicates a better assignment.

On the other hand, $\mathrm{b}\left(\mathbf{x}_i, U_k\right)$ represents the minimum average dissimilarity of object $\mathbf{x}_i$ with any other cluster different from U_k. The cluster with the minimum average dissimilarity is considered the "neighboring cluster" of object $\mathbf{x}_i$ representing the second-best cluster choice for $\mathbf{x}_i$.

The Silhouette index for an object ranges from -1 to 1. A value close to 1 indicates that the object is "well-clustered" because any neighboring cluster is on average not as close to the object as its own cluster. A value near 0 indicates that the object is presumably on the border of two clusters because it lies on average equally far away from two clusters. A value close to -1 indicates that the object has been erroneously assigned because the object is on average much closer to a neighboring cluster than to its own cluster. Thus, it would be more natural to assign this object to a neighboring cluster [36].

In the case of most objects have a high silhouette index, then the overall clustering configuration is presumably appropriate. On the other hand, if most objects have low or negative values, then the clustering configuration is maybe underestimating or overestimating the number of clusters.

3 Materials and methods

This section presents the methodology implemented to achieve the aims of the research. The data set construction is explained in Subsection 3.1; with respect to the data set embedding procedures and parameters, these are presented in Subsection 3.2; the proposed unsupervised binning method based on Subtractive and FCM clustering, and ICVIs is detailed in Subsection 3.3; finally, the Subsection 3.4 shows the measures used to assess the performance of the binning method. The scheme that summarizes all these procedures is presented in the Figure 1.

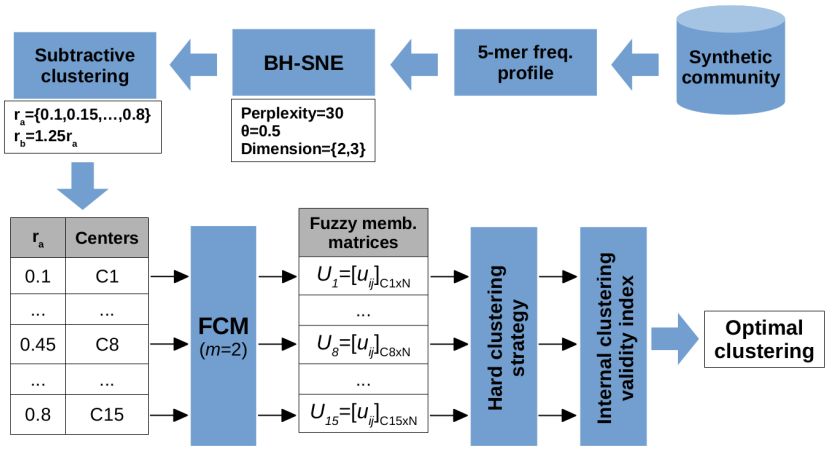

Figure 1: **Scheme of fuzzy binning application in low-dimensional metagenomic DNA sequences, including ICVI.**

3.1 Data set construction

Data were obtained from The Human Microbiome Project (HMP) [37] data portal (`http://hmpdacc.org/HMRGD/`). Forty-six complete or nearly-complete genome sequence assemblies of intestinal microorganisms were retrieved in Fasta format. The average Amino Acid Identity (AAI), an index of genome sequence similarity (and an approximate proxy to phylogenetic relatedness) [22], was calculated for all possible genome sequence pairs, using The Kostas Laboratory online AAI calculator (`http://enve-omics.ce.gatech.edu/aai/`). A total of 1035 AAI pairwise were obtained.

Synthetic Communities (SC) were constructed by applying a graph theory-based approach. To begin with, each genome was considered as the vertex of a graph and undirected edges between every pair of genomes were created; then, a weight was given to each edge connecting a pair of genomes and its value was determined by the AAI value between them. Next, four subgraphs were derived from the above AAI graph such that edges in each graph have AAI weights in ranges of $[25, 40)\%$, $[40, 55)\%$, $[55, 70)\%$, and $[70, 85)\%$, respectively. Then, k-vertex cliques, i.e. a subset of k vertices such that every two distinct vertices in the subset are adjacent, were identified in each AAI subgraph in order to build candidate SC of k genomes having pairwise AAI value in the same range. In particular, k was set to 5 and 10. Sets of 11 pseudo-biological replicates were assembled by filtering the candidate SC such that no two communities within each set shared more than 2 or 6 genomes in the 5-genome SC or 10-genome SC, respectively. Graph construction and clique selection were performed using a custom R script, and candidate SC were filtered using a custom Python script.

Contigs of 1000 bp in length, and of random starting points were generated from each genome to achieve a 1x genome sequence coverage. The number of contigs in each community varied according to the constituent genomes' sizes. This approach was seen as more biologically realistic than constructing data sets with an equal number of sequence fragments from each genome, unequal population abundances not-withstanding. Pentamer (5-mer) count profiles were constructed from all sequence fragments, combining each pentamer sequence with its reverse-complement (simplified *5-mer* dictionary). Counts were initiated with 0.1 pseudo-counts to avoid zero values in the table of frequencies. *5-mer* frequencies were obtained as specific counts divided by the total number of *5-mer*.

3.2 BH-SNE-based data set embedding

As mentioned before, the BH-SNE algorithm was used to embed the high-dimensional metagenomic data into low-dimensional spaces wherein a binning algorithm can be applied.

In the same way, as in [7], some of the default values of the BH-SNE algorithm's input parameters were used. The perplexity, a parameter controlling the effective number of local neighbors based on which neighborhood structure was captured, was set to 30. The parameter θ, a trade-off parameter ranging from 0 to 1, which controls the speed and the accuracy of the approximations provided by the Barnes-Hut implementation of the t-SNE algorithm, was set to 0.5, since preliminary tests with different values for θ (0.25, 0.50, 0.75) did not show significant differences in binning performance.

Two and three dimensional embeddings were obtained, in order to investigate the effects of increasing the dimensionality of the output space over the subsequent binning of genomic fragments. In addition, the low-dimensional mappings were generated with or without the application of a prior PCA-based dimensionality reduction step before the BH-SNE was performed. In the case of the application of PCA, data were mapped on the first 50 principal components and then the reduced data were used as an input to BH-SNE.

For the purposes of this work, the term "treatment" is used to describe the way in which mappings were obtained regarding the dimensionality of the output embedding and the application of PCA as a pre-processing step. Accordingly, below are listed the treatments considered in this study and their corresponding code-names:

- 2D: two-dimensional map, embedding obtained using the BH-SNE.
- 2DPCA: two-dimensional map, embedding includes preliminary PCA followed by the BH-SNE.
- 3D: three-dimensional map, embedding obtained using the BH-SNE.
- 3DPCA: three-dimensional map, embedding includes preliminary PCA followed by the BH-SNE.

Finally, because of the non-deterministic nature of the BH-SNE algorithm, a set of 11 technical replicates of the embeddings were generated for treatment applied to each SC.

3.3 Unsupervised method for metagenomic binning

As in [9], an unsupervised group identification method that uses the Subtractive and FCM clustering algorithms, in conjunction with internal clustering validity indices, was used for binning either two- and three-dimensional mappings of genomic fragments obtained via the BH-SNE algorithm. In particular, the aforementioned method has two stages each having particular objectives.

In the first stage, Subtractive clustering algorithm [29] was used to address the problem of estimate the centers of clusters in the embedded metagenomic data which correspond to different genomic populations in a SC. As mentioned before, Subtractive assumes each data object has the potential of being a cluster center based on the density of its surrounding data objects, and it uses a particular pair of positive radii r_a and r_b, with values between 0 and 1, to condition the effect of data objects around potential cluster centers in the density measure. As in [9], second radius was set as $r_b = 1.25r_a$, and the Subtractive algorithm was run for values of r_a from 0.1 to 0.8 with increments of 0.05, to obtain estimates of the number of clusters and their corresponding locations in the embedding space as a function of the radius r_a.

Next, the second stage deals with the problem of binning genomic fragments in an unsupervised way. To this end, the FCM algorithm (with a fuzzification parameter m = 2) was used to cluster the embedded metagenomic data based on each of the cluster center estimates obtained in the first stage. Then, based on the resulting fuzzy partition matrix in each instance, a hard clustering was obtained by assigned each data object to the cluster with the largest measure of membership. Since in practice there is no ground-truth information for this kind of data, the goodness of each hard clustering was measured using internal validity indices to determine which of these clustering results better fitted the underlying structure of the data. For this last task, several indices were evaluated, namely Silhouette (S), Davies-Bouldin (DB), Calinski-Harabasz (CH), and Dunn (D). In particular, this selection was based on the results of an extensive comparative study performed in [21].

3.4 Performance evaluation

The F-measure, a commonly used performance metric in metagenomic binning studies [38, 39, 40] when the true number of microbial populations is known. Let N be the number of microbial populations in a metagenomic data set, and C the number of clusters estimated by the binning algorithm on the same data set. Let A_{ij} be the number of DNA fragments from the j-th population assigned by the algorithm to the i-th cluster. Then, according to [11], the Equation (26) presents the accuracy of the resulting clustering in terms of F-measure.

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

Precision and Recall are defined in Equations (27,28) according to [38].

$$\text{Precision} = \frac{\sum_{i=1}^{K} \max_j A_{ij}}{\sum_{i=1}^{K} \sum_{j=1}^{n} A_{ij}} \tag{27}$$

$$\text{Recall} = \frac{\sum_{j=1}^{n} \max_i A_{ij}}{\sum_{i=1}^{K} \sum_{j=1}^{n} A_{ij} + \#\text{ unassigned reads}} \tag{28}$$

The accuracy in terms of F-measure was calculated for the binning results derived from each ICVI (CH, D, DB, and S) in each one of the eleven technical replicates, per treatment. This allows the calculus of the F median values for each index in each treatment. The Wilcoxon signed rank test [41] applied to the Coefficient of Variation (CV) [42] of their F median values contributed to explore the differences between the ICVI results per treatment.

4 Results and discussion

This section presents the results of our study and their corresponding discussion. Section 4.1 presents the results regarding the data set construction and the subsequent embedding. Then, the assessment of the binning results based on the Subtractive and FCM clustering in conjunction with the ICVIs are presented in Section 4.2. There, Section 4.2.1 lets to identify the index with the best performance by means of a global analysis, and Section 4.2.2 shows the index with the highest consistency in its results by means of a local analysis. Additionally, observations such as the impact of the biology of the metagenomic sequences over the ICVI results, the incidence of the AAI range over the ICVI results, are also presented in this section. Finally, Section 4.3 presents the results related to the performance of BH-SNE algorithm with and without a PCA preliminary dimensionality reduction step.

4.1 Data set construction and embedding

Table (1) summarizes the data sets obtained for this study. These data sets consist of 41 Synthetic Communities (SCs), which are composed of ten, five, or three distinct genomes. The low number of distinct genomes in the Subset 5 and the Subset 6 was due to the nature of the original data set which only allowed to identify 3-vertex cliques at most in the subgraphs with AAI within $[55, 70)\%$ and $[70, 85)\%$. For this AAI range, the number of generated SCs was the smallest, especially in the last range, and this number increases as the AAI percentage decreases. The communities used in this paper ranged from 6443 to 33199 contigs.

Table 1: **Amino Acid Identity communities per range**

AAI range	Subset	Genomes	Shared genomes	# SC	SC index
$[25, 40)\%$	1	5	2	10	$1 - 10$
	2	10	6	10	$11 - 20$
$[40, 55)\%$	3	5	2	10	$21 - 30$
	4	10	6	7	$31 - 37$
$[55, 70)\%$	5	3	2	3	$38 - 40$
$[70, 85)\%$	6	3	0	1	41

The table presents the constraints related to the generation of SC, its columns presents: AAI range, the range of percentages that were taken into account to create the subsets of communities; Subset, indicates the number of the subset assigned to the grouped SC; Genomes, the number of genomes that were included in each obtained SC; Shared genomes, the amount of maximum genomes allowed to be shared between all the grouped SC; # SC, the number of SC in the corresponding subset of data; SC index, index assigned to each genome per SC.

Figures (2,3) illustrate the two-dimensional BH-SNE-based embedding of the sixth technical replicate of the SC 22, which is composed by the following genomes: *Butyrivibrio fibrisolvens* 16/4, *Faecalibacterium prausnitzii* SL3/3, *Ruminiclostridium* [Eubacterium] *siraeum* V10Sc8a, Lachnospiraceae bacterium 2_1_46FAA, and *Acidaminococcus* sp. HPA0509. The embeddings were obtained without performing the default preliminary PCA step.

As indicated in Table (1), 41 SCs were obtained for this study. Then, given that there were four possible embedding treatments for each SC and eleven technical replicates of each treatment, a total of 1804 BH-SNE-based embeddings were generated and processed by our unsupervised method for metagenomic binning afterward.

4.2 Performance evaluation

The performance evaluation was executed based on two approaches: a global analysis and a local analysis. In both approaches, median values of the accuracy of the binning results in terms of F-measure (F) and the estimated number of genomic populations (C) were computed to perform the analysis. It is worth to mention that in the global approach were presented the indices with the highest and lowest F median and C median values, whereas in the local approach, it was analyzed the consistency of the indices generating the highest and lowest F median and C median results.

4.2.1 Global analysis

Table (2) presents the indices with the best and worst performances in terms of an overall F median and C median among all the treatments per AAI range. This table was constructed by firstly calculate the F median and C median values generated by each ICVI taking into account their results in the eleven technical replicates, for each treatment, per SC. Later, an overall F median and C median values were calculated per treatment for each AAI range, based on the F and C median results obtained by the ICVIs, per SC. Then, the ICVIs with the highest and lowest F median values (for each treatment) were included in the table. Finally, for the ICVIs added in the third step, their corresponding C median values were included in the table. It is important to mention that the most relevant results in the Table were highlighted in bold.

Based on Table (2), two trends in relation with the binning results on the two- and three-dimensional embeddings can be observed. In the majority of cases, selection based on the S index led to binning results in the treatments 2D and 2DPCA with the highest F measure. In the cases of subsets with 10 number of genomes, even the best performing indices, the S index and the CH index, tended to identify C median numbers below the expected number. However, in the cases of subsets with five number of genomes, the S and CH indices were able to calculate the right number of C medians in the AAI range of $[25, 40)\%$.

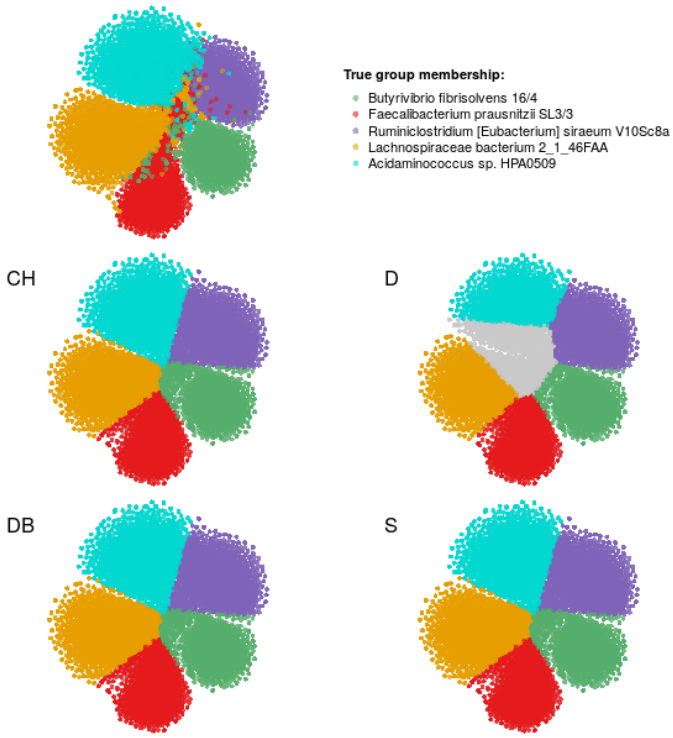

Figure 2: **Two-dimensional BH-SNE-based embedding of the sixth technical replicate of the SC 22**. The upper left panel represents with colors the organismal origin of the genomic fragments. The remaining panels present the group assignment suggested by our unsupervised method for metagenomic binning as a function of distinct cluster validity indices.

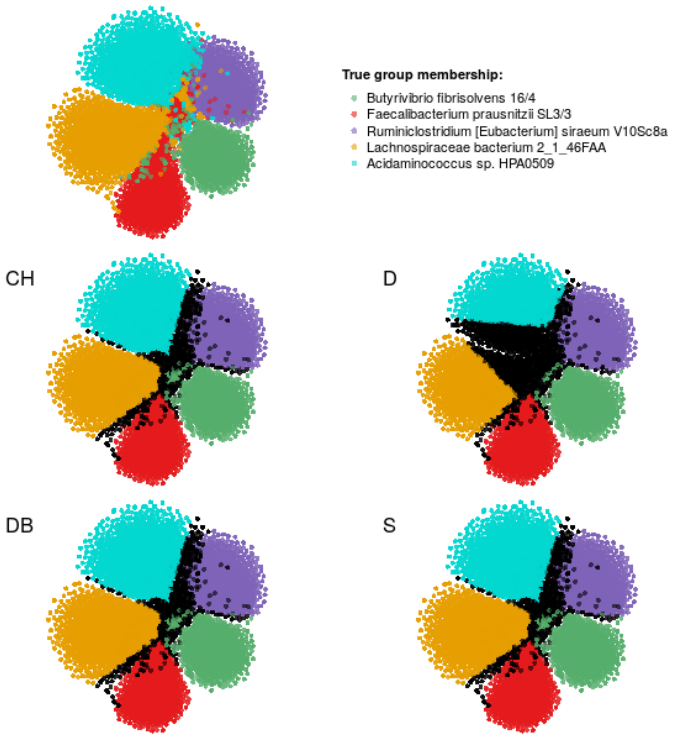

Figure 3: **Two-dimensional BH-SNE-based embedding of the sixth technical replicate of the SC 22**. Same as Figure 2, except that sequences assigned to the wrong genome by the distinct cluster validity indices are colored in black.

On the contrary, the worst binning results for the 2D treatments were generated by the D index in the first four subsets. However, D index was also able to calculate the correct number of C median in the cases of 5 genomes. Additionally, for the last two subsets wherein AAI percentages are higher, the DB index presented the worst performances. Indeed, D index was not able to calculate the proper number of C median.

Metagenomic binning based on three-dimensional embeddings show better performance for all treatments. In particular, the S index led to higher F median values in five of six subsets independently of the applied treatment. In the remaining subset (subset 3, AAI range of $[40, 55)\%$ with 10 genomes), the DB index generated the higher values of F median. Additionally, CH and S indices presented equal F median results in both three-dimensional treatments.

With regard to the worst binning results in the 3D treatments, the D index presented the worst performance in low AAI ranges and DB in high AAI ranges. However, the D index was able to calculate the right number of C median when the subsets consisted of 5 genomes, whereas the DB index led to over-clustering results in high AAI ranges. In particular, results based on D and DB indices, under 3D treatments, indicated that the number of genomes and their biology could affect the performance of the ICVIs.

Finally, the way in which performance of the binning method based on the S index was compromised when it comes to determine the correct C median value in a SC with different number of genomes, and how the F median value decreased as the AAI percentage of the subsets increases could also support the idea that the binning results associated to the ICVIs could be affected by the biology of the genomes.

4.2.2 Local analysis

This approach was used to analyze the consistency of the indices generating the highest and lowest F median and C median results. The analysis was done by exploring the SC located in each subset per AAI range, where the subsets were composed by the SC listed in Table 1. The ICVI with the best and worst values of the F median and their corresponding C median values for each SC per AAI range were presented in Tables 5 to 10 (in the annexes). In particular, Table 5 corresponds to the results of the Subset 1, Table 6 to the results of Subset 2, and so on. Based on the mentioned tables, the indices with the highest frequency having the best and worst performances of F median and their corresponding

Table 2: **Overall median results of F and C**

Subset	AAI	Genomes	Best F measure				Worst F measure			
			Treatment	Index	F	C	Treatment	Index	F	C
1	[25, 40)%	5	2D	S	0.862	**5**	2D	D	0.765	**5**
			2DPCA	DB	0.869	**5**	**2DPCA**	**D**	**0.760**	**5**
				S	0.869	**5**				
			3D	**S**	**0.884**	**5**	3D	D	0.792	**5**
			3DPCA	**S**	**0.884**	**5**	3DPCA	D	0.792	**5**
2	[25, 40)%	10	2D	S	0.872	8	**2D**	**D**	**0.745**	6
			2DPCA	S	0.887	8	2DPCA	D	0.768	6
			3D	**S**	**0.894**	8	3D	D	0.803	6
			3DPCA	**S**	**0.894**	8	3DPCA	D	0.803	6
3	[40, 55)%	5	2D	S	0.797	4	**2D**	**D**	**0.716**	**5**
			2DPCA	S	0.807	4	2DPCA	D	0.723	**5**
			3D	**DB**	**0.817**	**5**	3D	D	0.750	**5**
			3DPCA	**DB**	**0.817**	**5**	3DPCA	D	0.750	**5**
4	[40, 55)%	10	2D	CH	0.759	**5**	2D	D	0.654	7
			2DPCA	S	0.778	7	**2DPCA**	**D**	**0.620**	7
			3D	**S**	**0.786**	7	3D	D	0.672	7
			3DPCA	**S**	**0.786**	7	3DPCA	D	0.672	7
5	[55, 70)%	3	2D	S	0.632	5	2D	DB	0.115	12
			2DPCA	S	0.681	**3**	2DPCA	DB	0.192	12
			3D	**CH**	**0.741**	**3**	**3D**	**DB**	**0.0352**	12
				S	**0.741**	**3**				
			3DPCA	**CH**	**0.741**	**3**	**3DPCA**	**DB**	**0.0352**	12
				S	**0.741**	**3**				
6	[70, 85)%	3	2D	S	0.365	4	**2D**	**DB**	**0.073**	15
			2DPCA	S	0.454	4	2DPCA	DB	0.134	15
			3D	**CH**	**0.529**	5	3D	DB	0.207	15
				S	**0.529**	4				
			3DPCA	**CH**	**0.529**	5	3DPCA	DB	0.207	15
				S	**0.529**	4				

This table presents the indices with the best and worst performances per subset, AAI range, and number of genomes. Multi-column of best F measure presents the treatment, F and C values of the indices with the highest F median and their C median; multi-column of worst F measure shows the treatment, F and C values of the indices with the worst F median and their C median.

treatments were summarized in Table (3).

Table 3: **Indices with the highest frequencies obtaining the best and worst F median results**

Performance in Frequency						
Subset	AAI range	Genomes	Best		Worst	
			Index	Treatment	Index	Treatment
1	[25, 40)%	5	S	3DPCA	D	3DPCA
2	[25, 40)%	10	S	2DPCA	D	2D
3	[40, 55)%	5	DB	2DPCA	D	3DPCA
			S	2DPCA		
			DB	3D		
4	[40, 55)%	10	CH	2D	D	ALL
			S	3D		
5	[55, 70)%	3	S	ALL	DB	2D
					DB	2DPCA
					DB	3D
					D	3DPCA
6	[70, 85)%	3	S	ALL	DB	2D
					DB	2DPCA
					DB	3D
					D	3DPCA

The Table presents the indices with the highest frequency having the best and worst performances of F and their corresponding treatments. Multi-columns named best presents the indices and the treatments with the highest frequency being the best ICVI; multi-columns named worst shows the indices and the treatments with the highest frequency being the worst ICVI.

From Table (3), it can be inferred that the S index presented equal results for all of the treatments in high AAI range. However, in low AAI ranges, it was not possible to determine in which treatment the S index had more consistency obtaining the highest F media values. In particular, results of low AAI range include: in the case of subset 1, the most

consistent combination was the S index with treatment 3DPCA; for subset 2 and 3, the combination of S index and treatment 2D had equal results than the combinations of DB index with treatment 2D and the DB index with treatment 3D; in the case of subset 4, the S index with treatment 3D had equal results than the CH index with treatment 2D.

Hence, based on the obtained results for the S index in low AAI ranges, it is suggested to use the S index in conjunction with the cited treatment in the best column (see Table (3)) with the purpose to be able to perform accurately a metagenomic binning in low AAI ranges. However, it is worth mentioning that the S index presented the highest F median results in low AAI ranges when it was used in 3D treatments (not the most consistent with that treatment), as can be seen in the analysis of the Table (2).

In summary, in the case of the best index, the analysis of the Table (3) led to identify that the S index had more consistency generating the highest values of F median per subset than the rest of ICVIs, since its F median results were high in all the 6 subsets. Thus, the S index was not only the best index (such as it was determined in Subsubsection 4.2.1) but also was the most consistent index generating the best results. In particular, the impacts of the AAI changes over the ICVIs performances were specified in Subsubsection 4.2.1.

When it comes to analyze the ICVI with the highest consistency having the lowest values of F median, it can be inferred, from Table (3), that D index was more consistent generating the worst performances in low AAI ranges, while DB index was more consistent in high AAI ranges. Thus, D and DB indices (in their mentioned AAI ranges) were not only the most consistent generating the worst results, but also generated the lowest F median values, such as was determined in Subsubsection 4.2.1.

In order to analyze the consistency of the ICVIs results, Table (4) included the frequencies of all the indices having the best and worst F median performances. In particular, Table (4) was constructed based on the Tables 11 to 16 (in the annexes), where it can be found the number of times that each ICVI obtained the best and worst performances based on their F median values, discriminating between treatments in each subset. It is important to highlight that Tables 11 to 16 (in the annexes) were created using the information of the Tables 5 to 10 (in the annexes), which in turn were constructed applying the embedding and binning procedure in each one of the eleven technical replicates per SC.

Table 4: **Frequencies of all the indices having the best and worst F median performances**

Best performance						
Index	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6
CH	15	22	12	11	5	2
D	2	0	0	0	0	0
DB	24	21	25	8	0	1
S	27	32	26	14	12	4
Worst performance						
Index	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6
CH	3	3	6	2	0	0
D	35	35	33	26	3	1
DB	3	2	2	0	10	3
S	2	0	1	0	0	0

The Table presents all the ICVI with the number of times that each index has obtained the best and worst performances of F median per subset, independent of the treatments.

The Table (4) presents: the index of S, previously determined as the best and most consistent index, in the case of an AAI of $[25, 40)\%$ was more consistent when there were 10 genomes than with 5 genomes, instead of an AAI of $[40, 55)\%$ where it presented the inverse results, S index was more consistent with 5 genomes than with 10 genomes. This situation did not allow to determine if there were a proper number of genomes that let to obtain a better binning and a better validity process. Therefore, it supports the idea that the biology of the metagenomic sequences have an incidence over the best ICVI performances, such as was determined in Subsubsection 4.2.1.

Lastly, in the case of the ICVI with the worst performances, D index was the worst ICVI almost the same number of times in the subsets 1, 2, 3 and 4; particularly, it seems that the biology of the metagenomic sequences did not affect the D index results.

4.3 Performance of Barnes-Hut *t*-SNE

This subsection presents the results obtained to define the relevance of the performance of BH-SNE algorithm with and without a PCA preliminary dimensionality reduction step.

In the case of the three-dimensional mappings, Table (2) indicates that there were no differences between the performances of the best and worst ICVI in the treatments 3D and 3DPCA, since the indices led to the same numbers of F median and C median. Although the F median measures decreased as the AAI percentage increased, the indices in both treatments (3D and 3DPCA) found the proper number of C median in the subsets with 5 and 3 genomes. However, an exception was evident in the subset 6, where a higher number of C median than the expected (i.e. an over-clustering) was obtained.

Regarding the two-dimensional-based results, Table (2) indicates for the F median value that there was a better performance of every ICVI when a PCA-based preliminary step was applied (2DPCA). In particular, the differences of the F median values between 2D and 2DPCA treatments vary between 0.007 and 0.089. With reference to C median, in general, the ICVIs in 2D and 2DPCA treatments found C median equal values in four subsets. In the particular case of low AAI range of $[25, 40)\%$ with 5 genomes, the ICVIs were available to find the proper number of C median, whilst in the particular cases of higher AAI ranges, the ICVIs were not available to find the proper number of C median.

These 2D and 3D results led to conclude that the highest F median values were obtained in 3D and 3DPCA treatments (equal results in both 3D treatments). The results in Table (3) indicated that there were no consistent treatments where the ICVIs had obtained the highest F median values.

From Tables (2) and (3) it can be concluded that it was not possible to identify the treatment that led the indices to generate the worse binning results. Neither was it possible to identify the most consistent treatment where the indices used to obtain the worse binning results.

To evaluate the relevance of applying a preliminary PCA step before running the BH-SNE algorithm, the Coefficient of Variation (CV) of the F median values was calcuted in order to analyze the relationship between the F median and the variability of the F value. The Wilcoxon Signed-Rank Test was applied to the F median CV to determine whether the differences between the dispersions of each index in all treatments were significant without assuming them to follow the normal distribution. Figures (4,5,6,7,8,9) indicated that there was no significance between all the CV of F results of each index, comparing their results obtained in each treatment and performing the analysis per subset. This analysis indicated that there were no significant changes regarding the metagenomic binning performance on embeddings that included or not preliminary PCA.

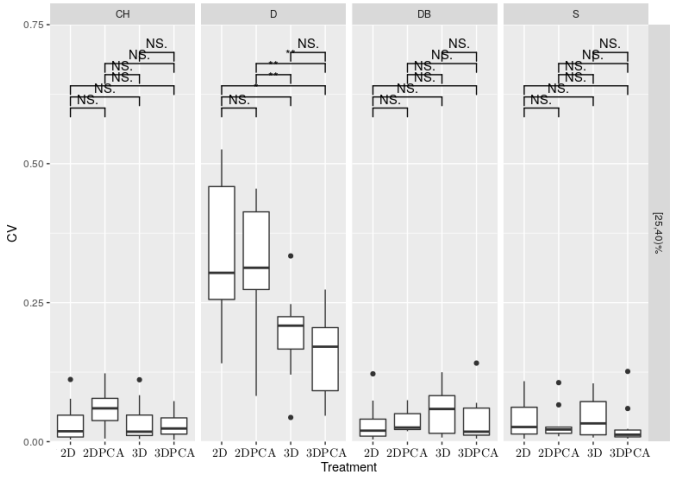

Figure 4: **Wilcoxon Signed-Rank Test of the Coefficient of Variation (CV) of the F measure median results in each treatment**. Each panel represents the Wilcoxon test between the CV results obtained by the ICVI of CH, D, DB, S in all the treatments (2D, 2DPCA, 3D, 3DPCA) with SC of 5 genomes included in AAI range $[25, 40)\%$. Horizontal lines at top indicate group-wise comparisons with the level of significance (***=0.001, **=0.01, *=0.05, NS.= Not significant).

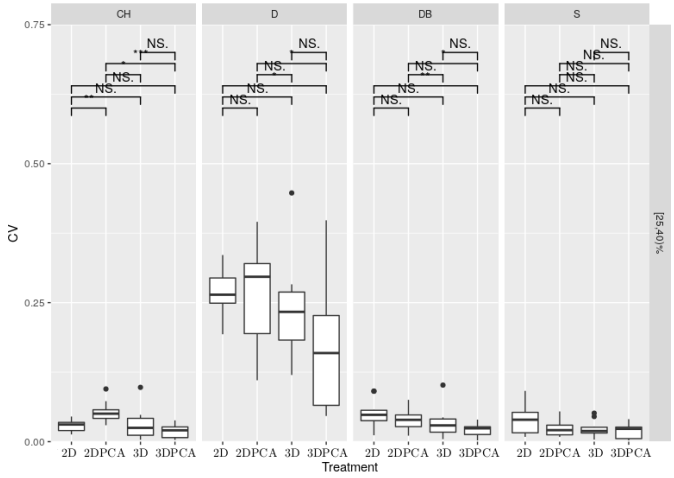

Figure 5: **Wilcoxon Signed-Rank Test of the Coefficient of Variation (CV) of the F measure median results in each treatment**. Each panel represents the Wilcoxon test between the CV results obtained by the ICVI of CH, D, DB, S in all the treatments (2D, 2DPCA, 3D, 3DPCA) with SC of 10 genomes included in AAI range $[25, 40)\%$. Horizontal lines at top indicate group-wise comparisons with the level of significance (***=0.001, **=0.01, *=0.05, NS.= Not significant).

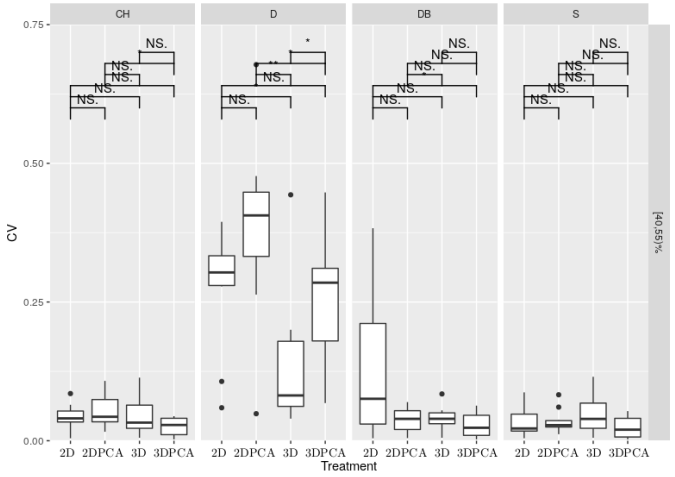

Figure 6: **Wilcoxon Signed-Rank Test of the Coefficient of Variation (CV) of the F measure median results in each treatment**. Each panel represents the Wilcoxon test between the CV results obtained by the ICVI of CH, D, DB, S in all the treatments (2D, 2DPCA, 3D, 3DPCA) with SC of 5 genomes included in AAI range $[40, 55)\%$. Horizontal lines at top indicate group-wise comparisons with the level of significance (***=0.001, **=0.01, *=0.05, NS.= Not significant).

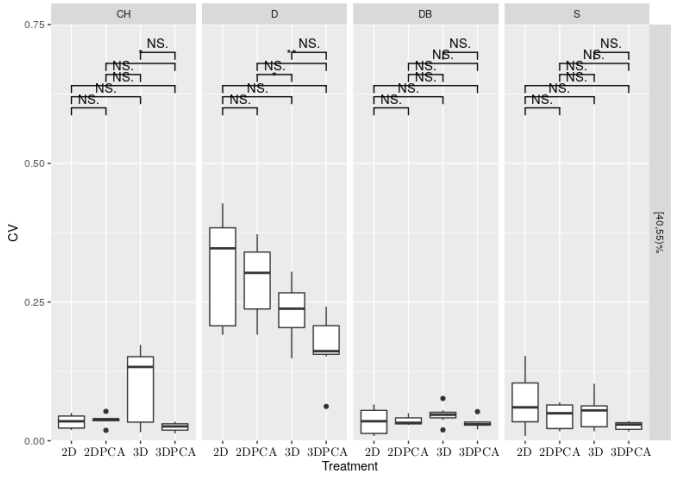

Figure 7: **Wilcoxon Signed-Rank Test of the Coefficient of Variation (CV) of the F measure median results in each treatment**. Each panel represents the Wilcoxon test between the CV results obtained by the ICVI of CH, D, DB, S in all the treatments (2D, 2DPCA, 3D, 3DPCA) with SC of 10 genomes included in AAI range $[40, 55)\%$. Horizontal lines at top indicate group-wise comparisons with the level of significance (***=0.001, **=0.01, *=0.05, NS.= Not significant).

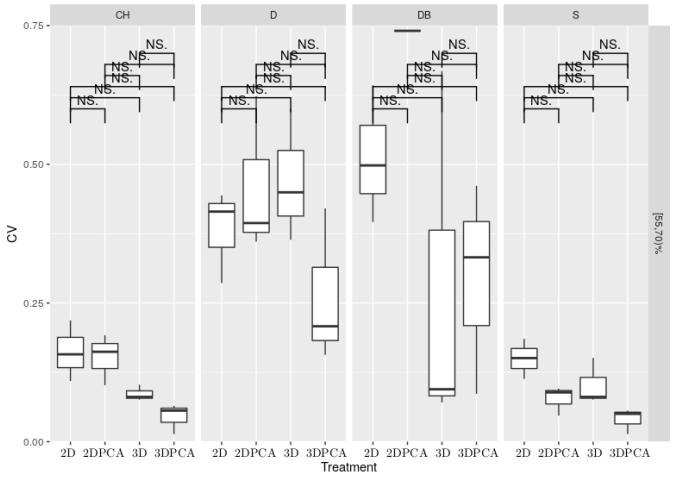

Figure 8: **Wilcoxon Signed-Rank Test of the Coefficient of Variation (CV) of the F measure median results in each treatment**. Each panel represents the Wilcoxon test between the CV results obtained by the ICVI of CH, D, DB, S in all the treatments (2D, 2DPCA, 3D, 3DPCA) with SC of 3 genomes included in AAI range $[55, 70)\%$. Horizontal lines at top indicate group-wise comparisons with the level of significance (***=0.001, **=0.01, *=0.05, NS.= Not significant).

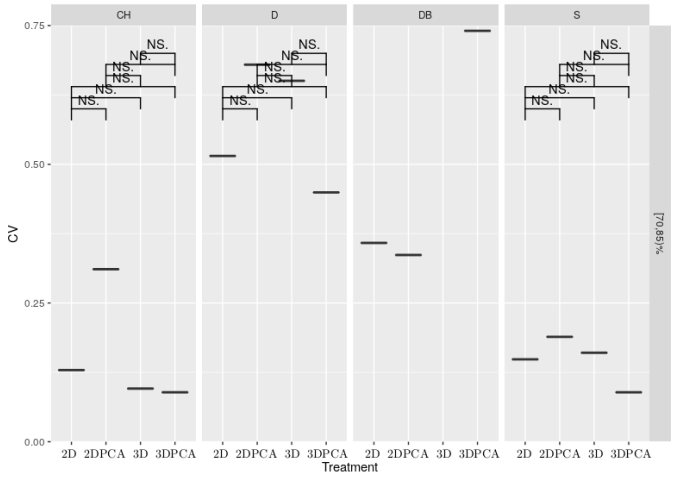

Figure 9: **Wilcoxon Signed-Rank Test of the Coefficient of Variation (CV) of the F measure median results in each treatment**. Each panel represents the Wilcoxon test between the CV results obtained by the ICVI of CH, D, DB, S in all the treatments (2D, 2DPCA, 3D, 3DPCA) with SC of 3 genomes included in AAI range $[70, 85)\%$. Horizontal lines at top indicate group-wise comparisons with the level of significance (***=0.001, **=0.01, *=0.05, NS.= Not significant).

5 Conclusions and future work

This work presented an unsupervised binning method of metagenomic DNA sequences, through the use of Subtractive and FCM algorithms applied to the two- and three-dimensional metagenomic sequences obtained with BH-SNE including and no including a preliminary PCA.

By comparing the performance of BH-SNE with and without PCA, it was possible to determine that the highest F median values were obtained by the ICVIs in 3D embeddings, with equal results for BH-SNE with or without the preliminary PCA step. In terms of consistency, it was not possible to determine which was the most consistent treatment

(2D or 3D embedding with BH-SNE including and not including preliminary PCA) that led the ICVIs to obtaining the best and worst F median results. Furthermore, it was demonstrated that there was no significant difference between the results that included or not included a preliminary PCA.

In addition, the assessment of the ICVIs demonstrated that the Silhouette index had the best performances based on the median values of the F measure. Moreover, Silhouette index was also the most consistent index obtaining the highest values of F median in two- and three-dimensional treatments. In the case of high AAI ranges, the Silhouette index had equal results compared with the Calinski-Harabasz index in terms of highest values of F median in three-dimensional treatment, despite differences between their performance in two-dimensional treatments. In particular, Dunn index generated the worst performances in the low AAI percentages, while the Davies-Bouldin index was the worst in high AAI percentages. Additionally, the Dunn and Davies-Bouldin indices were the most consistent generating the lowest F median values.

Results of this research suggest that the biology of the metagenomic sequences could have an incidence over the best ICVIs performances. As a future work, with the aim to contribute to complement the results presented in this research, it is suggested to perform experiments with a larger number of genomes per comunnity, specially in the case of high percentages of AAI, in order to achieve a wide analysis of the biological impact in a fuzzy binning and its validation procedures.

References

[1] S. Nayfach and K. S. Pollard, "HHS Public Access," *Cell*, vol. 166, no. 5, pp. 1103–1116, 2016.

[2] T. J. Sharpton, "An introduction to the analysis of shotgun metagenomic data," *Frontiers in Plant Science*, vol. 5, no. June, pp. 1–14, 2014. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fpls.2014.00209/abstract

[3] K. Sedlar, K. Kupkova, and I. Provaznik, "Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 48–55, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.csbj.2016.11.005

[4] D. R. Mende, A. S. Waller, S. Sunagawa, A. I. Järvelin, M. M. Chan, M. Arumugam, J. Raes, and P. Bork, "Assessment of metagenomic assembly using simulated next generation sequencing data," *PLoS ONE*, vol. 7, no. 2, 2012.

[5] S. S. Mande, M. H. Mohammed, and T. S. Ghosh, "Classification of metagenomic sequences: Methods and challenges," *Briefings in Bioinformatics*, vol. 13, no. 6, pp. 669–681, 2012.

[6] J. S. Ghurye, V. Cepeda-Espinoza, and M. Pop, "Metagenomic assembly: Overview, challenges and applications," *Yale Journal of Biology and Medicine*, vol. 89, no. 3, pp. 353–362, 2016.

[7] C. C. Laczny, N. Pinel, N. Vlassis, and P. Wilmes, "Alignment-free visualization of metagenomic data by nonlinear dimension reduction." *Scientific Reports*, vol. 4, p. 4516, 2014.

[8] M. Lux, A. Sczyrba, and B. Hammer, "Automatic discovery of metagenomic structure," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.

[9] L. Ariza-Jiménez, O. L. Quintero, and N. Pinel, "Unsupervised fuzzy binning of metagenomic sequence fragments on three-dimensional barnes-hut t-stochastic neighbor embeddings," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2018, pp. 1315–1318.

[10] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *Journal of Machine Learning Research*, vol. 15, pp. 3221–3245, 2014.

[11] R. Xu and D. C. Wunsch, "Clustering algorithms in biomedical research: a review." *IEEE reviews in biomedical engineering*, vol. 3, pp. 120–54, 2010.

[12] S. A. Thomson, R. L. Pyle, S. T. Ahyong, M. Alonso-Zarazaga, J. Ammirati, J. F. Araya, and A. et al, "Taxonomy based on science is necessary for global conservation," *PLOS Biology*, vol. 16, no. 3, p. e2005075, 2018. [Online]. Available: http://dx.plos.org/10.1371/journal.pbio.2005075

[13] S. D. Essinger, R. Polikar, and G. L. Rosen, "Neural network-based taxonomic clustering for metagenomics," *Proceedings of the International Joint Conference on Neural Networks*, vol. 19104, 2010.

[14] Ruiqi Liao, Ruichang Zhang, Jihong Guan, and Shuigeng Zhou, "A New Unsupervised Binning Approach for Metagenomic Sequences Based on N-grams and Automatic Feature Weighting," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 42–54, 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6654133/

[15] T.-F. Chen, R.-M. Chen, J. J. P. Tsai, and R.-M. Hu, "Fine Classification of Human Gut Microbiota by Using Hierarchical Clustering Approach," *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 109–112, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7789967/

[16] Y. Liu, F. Liu, T. Hou, and K. Wang, "Unsupervised Binning of Metagenomic Datasets Using Cluster Size Insensitive Fuzzy c-means Method," no. 1, pp. 3936–3939, 2016.

[17] Y. Liu, T. Hou, B. Kang, and F. Liu, "Unsupervised Binning of Metagenomic Assembled Contigs Using Improved Fuzzy C-Means Method," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 6, pp. 1459–1467, 2017.

[18] S. Girotto, M. Comin, and C. Pizzi, "Binning metagenomic reads with probabilistic sequence signatures based on spaced seeds," *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–8, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8058538/

[19] A. Gisbrecht, B. Hammer, B. Mokbel, and A. Sczyrba, "Nonlinear dimensionality reduction for cluster identification in metagenomic samples," *Proceedings of the International Conference on Information Visualisation*, pp. 174–179, 2013.

[20] M. Hauser, C. E. Mayer, and J. Söding, "KClust: Fast and sensitive clustering of large protein sequence databases," *BMC Bioinformatics*, vol. 14, no. 1, 2013.

[21] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.

[22] K. T. Konstantinidis and J. M. Tiedje, "Towards a genome-based taxonomy for prokaryotes," *Journal of Bacteriology*, vol. 187, no. 18, pp. 6258–6264, 2005. [Online]. Available: https://jb.asm.org/content/187/18/6258

[23] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[24] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, vol. 2, no. 11, pp. 559–572, 1901. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/14786440109462720

[25] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[26] G. Hinton and S. Roweis, "Stochastic neighbor embedding," pp. 857–864, 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=2968618.2968725

[27] P. N. Yianilos, "Data structures and algorithms for nearest neighbor search in general metric spaces," pp. 311–321, 1993. [Online]. Available: http://dl.acm.org/citation.cfm?id=313559.313789

[28] G. Dougherty, *Pattern Recognition and Classification*. New York, NY: Springer New York, 2013.

[29] S. L. Chiu, "Fuzzy model identification based on cluster estimation," *Journal of Intelligent and Fuzzy Systems*, vol. 2, no. 3, pp. 267–278, 1994.

[30] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[31] B. Desgraupes, "Clustering Indices," *CRAN Package*, no. April, pp. 1–10, 2013. [Online]. Available: cran.r-project.org/web/packages/clusterCrit

[32] C. Cengizler and M. Kerem-Un, "Evaluation of Calinski-Harabasz Criterion as Fitness Measure for Genetic Algorithm Based Segmentation of Cervical Cell Nuclei," *British Journal of Mathematics & Computer Science*, vol. 22, no. 6, pp. 1–13, 2017. [Online]. Available: http://www.sciencedomain.org/abstract/19643

[33] N. Bolshakova and F. Azuaje, "Cluster validation techniques for genome expression data," *Signal Processing*, vol. 83, no. 4, pp. 825–833, 2003.

[34] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recognition*, vol. 37, no. 3, pp. 487–501, 2004.

[35] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[36] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[37] The Human Microbiome Project Consortium, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[38] Y. Wang, H. C. Leung, S. M. Yiu, and F. Y. Chin, "Metacluster 5.0: A two-round binning approach for metagenomic data for low-abundance species in a noisy sample," *Bioinformatics*, vol. 28, no. 18, pp. 356–362, 2012.

[39] L. V. Vinh, T. V. Lang, L. T. Binh, and T. V. Hoai, "A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads," *Algorithms for Molecular Biology*, vol. 10, no. 1, pp. 1–12, 2015.

[40] S. Girotto, M. Comin, and C. Pizzi, "Metagenomic reads binning with spaced seeds," *Theoretical Computer Science*, vol. 698, pp. 88–99, 2017.

[41] E. Whitley and J. Ball, "Statistics review 6: Nonparametric methods," *Critical Care*, vol. 6, no. 6, pp. 509–513, 2002.

[42] G. L. Kesteven, "The coefficient of variation," *Nature*, vol. 158, pp. 520–521, 1946.

Annexes

Table 5: **Best and worst F median per SC in AAI range $[25, 40)\%$ with 5 genomes**

SC index	Best F measure				Worst F measure			
	Treatment	Index	F	C	Treatment	Index	F	C
1	2D	CH	0.842	5	2D	D	0.675	9
		DB	0.842	5				
		S	0.842	5				
	2DPCA	DB	0.848	5	2DPCA	D	0.685	6
		S	0.848	5				
	3D	CH	0.853	5	3D	D	0.724	5
		DB	0.853	5				
		S	0.853	5				
	3DPCA	DB	0.848	5	3DPCA	D	0.797	5
2	2D	CH	0.780	4	2D	D	0.714	4
	2DPCA	CH	0.786	6	2DPCA	D	0.649	3
	3D	CH	0.768	3	3D	D	0.744	4
		DB	0.768	3				
	3DPCA	CH	0.816	4	3DPCA	D	0.799	5
		S	0.816	4				
3	2D	DB	0.874	5	2D	D	0.744	5
		S	0.874	5				
	2DPCA	DB	0.884	5	2DPCA	D	0.758	9
	3D	DB	0.905	5	3D	D	0.797	5
		S	0.905	5				
	3DPCA	DB	0.922	5	3DPCA	D	0.714	4
		S	0.922	5				
4	2D	S	0.870	5	2D	D	0.747	4
	2DPCA	S	0.885	5	2DPCA	D	0.811	6
	3D	S	0.859	6	3D	D	0.761	4
	3DPCA	DB	0.897	5	3DPCA	D	0.591	4
		S	0.897	5				
5	2D	CH	0.941	5	2D	D	0.724	6
		S	0.941	5				
	2DPCA	DB	0.941	5	2DPCA	D	0.866	6
		S	0.941	5				
	3D	CH	0.948	5	3D	D	0.898	5
		DB	0.948	5				
		S	0.948	5				
	3DPCA	S	0.962	5	3DPCA	D	0.960	5
6	2D	DB	0.883	4	2D	D	0.841	6
	2DPCA	DB	0.901	5	2DPCA	CH	0.699	9
		S	0.901	5				
	3D	DB	0.916	5	3D	D	0.850	4
		S	0.916	5				
	3DPCA	DB	0.941	5	3DPCA	D	0.861	5
		S	0.941	5				
7	2D	CH	0.811	5	2D	D	0.804	4
		DB	0.811	5				
	2DPCA	DB	0.827	5	2DPCA	D	0.660	6
		S	0.827	5				
	3D	DB	0.819	5	3D	CH	0.669	3
	3DPCA	CH	0.848	5	3DPCA	D	0.665	5
		S	0.848	5				
8	2D	D	0.795	5	2D	DB	0.765	3
	2DPCA	D	0.833	4		S	0.765	3
	3D	CH	0.890	4	2DPCA	CH	0.793	7
	3DPCA	DB	0.924	5	3D	DB	0.765	3
		S	0.924	5		S	0.765	3
					3DPCA	D	0.859	5
9	2D	DB	0.867	5	2D	D	0.709	6
		S	0.867	6				
	2DPCA	DB	0.870	6	2DPCA	D	0.757	5
	3D	DB	0.940	5	3D	D	0.829	7
		S	0.940	5				
	3DPCA	DB	0.935	5	3DPCA	D	0.871	5
		S	0.935	5				
10	2D	CH	0.917	4	2D	D	0.853	5
	2DPCA	CH	0.929	4	2DPCA	D	0.608	4
		S	0.929	4				
	3D	CH	0.928	4	3D	D	0.728	4
		S	0.928	4	3D	DB	0.728	3
	3DPCA	CH	0.942	4	3DPCA	D	0.940	4
		DB	0.942	4				
		S	0.942	4				

Best and worst F median per SC in subset 1, includes C median obtained with Treatment-Index combination.

Table 6: **Best and worst F median per SC in AAI range $[25, 40)\%$ with 10 genomes**

SC index	Best F measure				Worst F measure			
	Treatment	Index	F	C	Treatment	Index	F	C
11	2D	CH	0.828	8	2D	D	0.750	7
	2DPCA	CH	0.827	10	2DPCA	D	0.668	7
	3D	CH	0.900	8	3D	D	0.875	8
		DB	0.900	8				
		S	0.900	8				
	3DPCA	S	0.907	8	3DPCA	D	0.861	7
12	2D	S	0.870	8	2D	D	0.757	7
	2DPCA	DB	0.914	9	2DPCA	D	0.824	9
		S	0.914	9				
	3D	CH	0.895	8	3D	D	0.817	7
		S	0.895	8				
	3DPCA	DB	0.941	9	3DPCA	D	0.925	9
		S	0.941	9				
13	2D	S	0.882	9	2D	D	0.802	10
	2DPCA	DB	0.891	9	2DPCA	D	0.722	9
		S	0.891	9				
	3D	S	0.891	9	3D	D	0.796	7
	3DPCA	DB	0.880	8	3DPCA	CH	0.871	11
		S	0.880	8				
14	2D	CH	0.887	8	2D	D	0.743	6
		DB	0.887	8				
		S	0.887	8				
	2DPCA	S	0.890	8	2DPCA	D	0.786	8
	3D	CH	0.892	8	3D	D	0.767	8
		DB	0.892	8				
	3DPCA	CH	0.903	8	3DPCA	D	0.870	8
15	2D	S	0.874	8	2D	D	0.740	8
	2DPCA	DB	0.915	8	2DPCA	D	0.716	9
		S	0.915	8				
	3D	DB	0.896	7	3D	CH	0.880	8
		S	0.896	7				
	3DPCA	S	0.938	8	3DPCA	D	0.901	8
16	2D	DB	0.847	7	2D	D	0.747	8
		S	0.847	7				
	2DPCA	DB	0.894	8	2DPCA	CH	0.722	14
		S	0.894	8				
	3D	CH	0.849	7	3D	D	0.779	8
	3DPCA	CH	0.905	8	3DPCA	D	0.803	7
		DB	0.905	8				
		S	0.905	8				
17	2D	CH	0.907	8	2D	D	0.784	7
		DB	0.907	8				
		S	0.907	8				
	2DPCA	CH	0.916	8	2DPCA	D	0.861	8
		DB	0.916	8				
		S	0.916	8				
	3D	CH	0.928	8	3D	D	0.785	8
		DB	0.928	8				
		S	0.928	8				
	3DPCA	CH	0.936	8	3DPCA	D	0.933	8
		DB	0.936	8				
		S	0.936	8				
18	2D	CH	0.839	8	2D	D	0.774	7
		S	0.839	7				
	2DPCA	CH	0.851	8	2DPCA	D	0.720	6
		DB	0.851	8				
		S	0.851	8				
	3D	CH	0.842	8	3D	D	0.816	6
	3DPCA	CH	0.908	9	3DPCA	D	0.831	7
		DB	0.908	9				
		S	0.908	9				
19	2D	CH	0.824	8	2D	D	0.700	9
	2DPCA	S	0.847	7	2DPCA	DB	0.815	6
	3D	CH	0.817	6	3D	D	0.758	4
		S	0.817	6				
	3DPCA	CH	0.883	7	3DPCA	DB	0.851	6
20	2D	S	0.878	10	2D	D	0.675	8
	2DPCA	DB	0.884	10	2DPCA	D	0.767	13
		S	0.884	11				
	3D	CH	0.911	10	3D	D	0.661	10
		DB	0.911	10				
		S	0.911	10				
	3DPCA	CH	0.919	10	3DPCA	D	0.848	10
		DB	0.919	10				
		S	0.919	10				

Best and worst F median per SC in subset 2, includes C median obtained with Treatment-Index combination.

Table 7: **Best and worst F median per SC in AAI range $[40, 55)\%$ with 5 genomes**

SC index	Best F measure				Worst F measure			
	Treatment	Index	F	C	Treatment	Index	F	C
21	2D	DB	0.797	4	2D	D	0.786	4
		S	0.797	4				
	2DPCA	DB	0.848	4	2DPCA	D	0.689	9
		S	0.848	4				
	3D	CH	0.880	4	3D	D	0.747	4
		DB	0.880	4				
		S	0.880	4				
	3DPCA	S	0.886	4	3DPCA	D	0.748	4
22	2D	CH	0.913	5	2D	D	0.859	6
		DB	0.913	5				
		S	0.913	5				
	2DPCA	DB	0.918	5	2DPCA	D	0.318	29
		S	0.918	5				
	3D	CH	0.926	5	3D	D	0.886	6
		DB	0.926	5				
		S	0.926	5				
	3DPCA	CH	0.931	5	3DPCA	D	0.895	5
		DB	0.931	5				
		S	0.931	5				
23	2D	CH	0.689	3	2D	DB	0.671	9
	2DPCA	DB	0.781	4	2DPCA	CH	0.661	10
		S	0.781	4				
	3D	DB	0.774	5	3D	CH	0.687	2
	3DPCA	S	0.832	5	3DPCA	D	0.798	6
24	2D	CH	0.775	5	2D	D	0.704	5
	2DPCA	DB	0.758	7	2DPCA	D	0.721	6
	3D	DB	0.814	5	3D	CH	0.706	2
	3DPCA	DB	0.870	5		S	0.706	2
		S	0.870	5	3DPCA	D	0.731	5
25	2D	DB	0.834	5	2D	D	0.695	6
	2D	S	0.8345	5				
	2DPCA	DB	0.862	5	2DPCA	D	0.710	5
		S	0.862	5				
	3D	S	0.858	5	3D	D	0.597	3
	3DPCA	S	0.824	6	3DPCA	D	0.7363	6
26	2D	S	0.799	6	2D	D	0.765	6
	2DPCA	S	0.798	6	2DPCA	D	0.733	8
	3D	CH	0.801	6	3D	D	0.729	4
		S	0.801	6				
	3DPCA	DB	0.813	6	3DPCA	D	0.757	5
27	2D	DB	0.806	4	2D	D	0.686	7
	2DPCA	DB	0.817	5	2DPCA	CH	0.772	7
	3D	CH	0.824	4	3D	D	0.7913	4
		DB	0.824	5				
		S	0.824	4	3DPCA	CH	0.829	6
	3DPCA	S	0.898	5	3DPCA	D	0.829	5
28	2D	CH	0.806	2	2D	DB	0.694	7
	2DPCA	S	0.851	3	2DPCA	D	0.749	5
	3D	DB	0.809	3	3D	CH	0.804	2
	3DPCA	S	0.879	3	3DPCA	D	0.854	4
29	2D	DB	0.729	4	2D	D	0.586	3
		S	0.729	4				
	2DPCA	CH	0.799	4	2DPCA	D	0.670	5
		DB	0.799	4				
		S	0.799	4				
	3D	DB	0.745	4	3D	D	0.590	4
	3DPCA	CH	0.806	4	3DPCA	D	0.738	5
30	2D	S	0.793	2	2D	D	0.712	6
	2DPCA	CH	0.823	3	2DPCA	D	0.740	4
		DB	0.823	3				
		S	0.823	3				
	3D	DB	0.828	3	3D	D	0.789	2
	3DPCA	DB	0.857	3	3DPCA	D	0.795	3

Best and worst F median per SC in subset 3, includes C median obtained with Treatment-Index combination.

Table 8: **Best and worst F median per SC in AAI range $[40, 55)\%$ with 10 genomes**

SC index	Best F measure				Worst F measure			
	Treatment	Index	F	C	Treatment	Index	F	C
31	2D	CH	0.755	10	2D	D	0.562	25
	2DPCA	S	0.773	10	2DPCA	D	0.646	15
	3D	S	0.765	10	3D	CH	0.651	4
	3DPCA	DB	0.826	8	3DPCA	D	0.708	6
32	2D	CH	0.727	14	2D	D	0.700	15
	2DPCA	S	0.757	8	2DPCA	D	0.690	7
	3D	S	0.772	9	3D	D	0.687	6
	3DPCA	DB	0.847	9	3DPCA	D	0.720	5
		S	0.847	9				
33	2D	CH	0.779	11	2D	D	0.569	8
	2DPCA	CH	0.776	11	2DPCA	D	0.685	11
	3D	S	0.779	8	3D	D	0.702	7
	3DPCA	S	0.808	10	3DPCA	D	0.717	9
34	2D	CH	0.816	10	2D	D	0.741	7
	2DPCA	CH	0.826	10	2DPCA	D	0.649	14
	3D	CH	0.847	9	3D	D	0.666	9
		S	0.847	9				
	3DPCA	DB	0.839	9	3DPCA	D	0.790	8
35	2D	DB	0.742	10	2D	D	0.652	5
	2DPCA	S	0.775	8	2DPCA	D	0.608	4
	3D	S	0.777	8	3D	CH	0.668	4
	3DPCA	S	0.813	9	3DPCA	D	0.701	6
36	2D	CH	0.794	9	2D	D	0.677	16
	2DPCA	DB	0.784	9	2DPCA	D	0.593	5
		S	0.784	9				
	3D	DB	0.798	10	3D	D	0.530	4
		S	0.798	9				
	3DPCA	DB	0.795	9	3DPCA	D	0.755	8
37	2D	CH	0.808	10	2D	D	0.687	6
	2DPCA	CH	0.80	11	2DPCA	D	0.677	8
	3D	CH	0.839	9	3D	D	0.692	6
		DB	0.839	9				
	3DPCA	S	0.841	10	3DPCA	D	0.763	7

Best and worst F median per SC in subset 4, includes C median obtained with Treatment-Index combination.

Table 9: **Best and worst F median per SC in AAI range $[55, 70)\%$ with 3 genomes**

SC index	Best F measure				Worst F measure			
	Treatment	Index	F	C	Treatment	Index	F	C
38	2D	S	0.782	2	2D	DB	0.106	48
	2DPCA	S	0.680	3	2DPCA	DB	0.266	14
	3D	CH	0.786	2	3D	DB	0.035	163
		S	0.786	2				
	3DPCA	CH	0.692	3	3DPCA	D	0.691	3
		S	0.692	3	3DPCA	DB	0.691	3
39	2D	S	0.613	3	2D	DB	0.188	23
	2DPCA	S	0.798	2	2DPCA	DB	0.117	41
	3D	CH	0.637	3	3D	DB	0.037	155
		S	0.637	3				
	3DPCA	S	0.795	2	3DPCA	D	0.579	4
40	2D	S	0.631	3	2D	DB	0.197	22
	2DPCA	S	0.673	3	2DPCA	DB	0.178	22
	3D	CH	0.741	2	3D	DB	0.032	185
		S	0.741	2				
	3DPCA	CH	0.773	2	3DPCA	D	0.676	3
		S	0.773	2				

Best and worst F median per SC in subset 5, includes C median obtained with Treatment-Index combination.

Table 10: **Best and worst F median per SC in AAI range $[70, 85)\%$ with 3 genomes**

SC index	Best F measure				Worst F measure			
	Treatment	Index	F	C	Treatment	Index	F	C
41	2D	S	0.365	4	2D	DB	0.073	15
	2DPCA	S	0.454	3	2DPCA	DB	0.134	15
	3D	CH	0.529	2	3D	DB	0.207	9
		S	0.529	2				
	3DPCA	CH	0.533	2	3DPCA	D	0.445	2.5
		DB	0.533	2				
		S	0.533	2				

Best and worst F median per SC in subset 6, includes C median obtained with Treatment-Index combination.

Table 11: **F median frequencies in subset 1**

AAI 2540-5				
Treatment	Best performance		Worst performance	
	Index	Quantity	Index	Quantity
2D	CH	5	CH	0
	D	1	D	9
	DB	4	DB	1
	S	5	S	1
2DPCA	CH	2	CH	2
	D	1	D	8
	DB	6	DB	0
	S	6	S	0
3D	CH	5	CH	1
	D	0	D	8
	DB	7	DB	2
	S	7	S	1
3DPCA	CH	3	CH	0
	D	0	D	10
	DB	7	DB	0
	S	9	S	0

Number of times that each ICVI obtained the best and worst performances based on the F median values.

Table 12: **F median frequencies in subset 2**

AAI 2540-10				
Treatment	Best performance		Worst performance	
	Index	Quantity	Index	Quantity
2D	CH	5	CH	0
	D	0	D	10
	DB	3	DB	0
	S	8	S	0
2DPCA	CH	3	CH	1
	D	0	D	8
	DB	7	DB	1
	S	9	S	0
3D	CH	8	CH	1
	D	0	D	9
	DB	5	DB	0
	S	7	S	0
3DPCA	CH	6	CH	1
	D	0	D	8
	DB	6	DB	1
	S	8	S	0

Number of times that each ICVI obtained the best and worst performances based on the F median values.

Table 13: **F median frequencies in subset 3**

AAI 4055-5				
Treatment	Best performance		Worst performance	
	Index	Quantity	Index	Quantity
2D	CH	4	CH	0
	D	0	D	8
	DB	5	DB	2
	S	6	S	0
2DPCA	CH	2	CH	2
	D	0	D	8
	DB	8	DB	0
	S	8	S	0
3D	CH	4	CH	3
	D	0	D	7
	DB	8	DB	0
	S	5	S	1
3DPCA	CH	2	CH	1
	D	0	D	10
	DB	4	DB	0
	S	7	S	0

Number of times that each ICVI obtained the best and worst performances based on the F median values.

Table 14: **F median frequencies in subset 4**

AAI 4055-10				
Treatment	Best performance		Worst performance	
	Index	Quantity	Index	Quantity
2D	CH	6	CH	0
	D	0	D	7
	DB	1	DB	0
	S	0	S	0
2DPCA	CH	3	CH	0
	D	0	D	7
	DB	1	DB	0
	S	4	S	0
3D	CH	2	CH	2
	D	0	D	5
	DB	2	DB	0
	S	6	S	0
3DPCA	CH	0	CH	0
	D	0	D	7
	DB	4	DB	0
	S	4	S	0

Number of times that each ICVI obtained the best and worst performances based on the F median values.

Table 15: **F median frequencies in subset 5**

AAI 5570-3				
Treatment	Best performance		Worst performance	
	Index	Quantity	Index	Quantity
2D	CH	0	CH	0
	D	0	D	0
	DB	0	DB	3
	S	3	S	0
2DPCA	CH	0	CH	0
	D	0	D	0
	DB	0	DB	3
	S	3	S	0
3D	CH	3	CH	0
	D	0	D	0
	DB	0	DB	3
	S	3	S	0
3DPCA	CH	2	CH	0
	D	0	D	3
	DB	0	DB	1
	S	3	S	0

Number of times that each ICVI obtained the best and worst performances based on the F median values.

Table 16: **F median frequencies in subset 6**

AAI 7085-3				
Treatment	Best performance		Worst performance	
	Index	Quantity	Index	Quantity
2D	CH	0	CH	0
	D	0	D	0
	DB	0	DB	1
	S	1	S	0
2DPCA	CH	0	CH	0
	D	0	D	0
	DB	0	DB	1
	S	1	S	0
3D	CH	1	CH	0
	D	0	D	0
	DB	0	DB	1
	S	1	S	0
3DPCA	CH	1	CH	0
	D	0	D	1
	DB	1	DB	0
	S	1	S	0

Number of times that each ICVI obtained the best and worst performances based on the F median values.