Absence of enterotypes in the human gut microbiomes reanalyzed with non-linear dimensionality reduction methods

Ivan Bulygin¹, Vladislav Shatov², Anton Rykachevsky¹, Arseny Rayko¹, Alexander Bernstein¹, Evgeny Burnaev^{1,4}, and Mikhail S. Gelfand^{1,3}

¹Skolkovo Institute of Science and Technology, Moscow, 121205, Russia

²School of Biology, Moscow State University, Moscow, 119991, Russia

³Institute for Information Transmission Problems (the Kharkevich Institute, RAS), Moscow, 127051, Russia

⁴Artificial Intelligence Research Institute (AIRI), Moscow, Russia

Abstract

Enterotypes of the human gut microbiome have been proposed to be a powerful prognostic tool to evaluate the correlation between lifestyle, nutrition, and disease. However, the number of enterotypes suggested in the literature ranged from two to four. The growth of available metagenome data and the use of exact, non-linear methods of data analysis challenges the very concept of clusters in the multidimensional space of bacterial microbiomes. Using several published human gut microbiome datasets, we demonstrate the presence of a lower-dimensional structure in the microbiome space, with high-dimensional data concentrated near a low-dimensional non-linear submanifold, but the absence of distinct and stable clusters that could represent enterotypes. This observation is robust with regard to diverse combinations of dimensionality reduction techniques and clustering algorithms.

1 Introduction

The human gut is populated by a diverse community of microorganisms. The microbiome of an individual gut settles in several years after birth and, by rough estimates, contains more than a thousand genera of bacteria [28]. Gut microbiota forms a dynamic ecosystem whose composition tends to be constant during the life of an individual but varies between individuals and may significantly depend on external and internal factors. Recently, much effort has been applied to correlate the results of deep sequencing of gut microbiota with diet, hereditary diseases, antibiotic consumption, etc. The initial sequencing of gut biota revealed that its composition tends to form discrete groups (enterotypes) consisting predominantly of taxa *Bacteroides*, *Prevotella*, and *Ruminococcaceae* [5]. Enterotypes have been reported in [5] as “densely populated areas in a multidimensional space of community composition” which “are not as sharply delimited as, for example, human blood groups”.

We consider enterotypes as clusters in the relative taxonomic abundance space. From a geometric point of view, clusters are defined as dense areas separated by sparse re-

gions. Clustering is a process of assigning a finite set of objects to separate groups and identifying the natural structure of the data when the relationship between objects is represented as some metric like the Euclidean distance [66]. As mentioned in [41], a major challenge is to determine the presence of such clusters through “a thorough quantitative investigation of established clustering methods and tests for microbiome data”. The published studies rely on similar approaches yet differ in the exact number of enterotypes ranging from two [48], [89], [11], [55], [39], [54], [84], [58], [85] to three [86], [19], [80], [23], [63], and four [29]. In most papers the presence of enterotypes in the microbiome data is determined by clustering of the data into K groups, with K selected by optimization of some metric of the clustering partition quality. The resulting segregation is then visualized in the projection of the first two or three principal components from the Principal Components Analysis (PCA). This approach is quite general and may involve various intermediate steps, different clustering algorithms, and metrics; however, it has certain flaws. For example, the Partitioning Around Medoids (PAM) method used in many studies may yield erroneous results for density-based clusters. Also, widely used in the studies partition quality metrics such as the Calinski–Harabasz index [9] and the Silhouette score [64] are naturally higher for convex clusters and may fail to detect density-based partition since they rely on the estimation of inter-cluster variance and cluster centers.

Another common problem is the small size of datasets in comparison to their dimensionality. The direct application of standard clustering methods for small and high-dimensional datasets, performed in most of the works, may lead to unreliable results due to the curse of dimensionality. Clustering implies the notion of dissimilarity between data samples. When the data dimensionality increases, the concepts of proximity, distance, or nearest neighbor become less qualitatively meaningful, especially for the commonly used Euclidean or Manhattan distances [1]. For example, the distance to the nearest data point approaches the distance to the farthest data point [7]. The lack of data further amplifies this problem since the data point cloud in a high-dimensional space becomes sparse, yielding unreliable esti-

mates of the probability density due to the non-asymptotic lower bound for the regression error [43], [36]. One straightforward way to overcome it is to reduce the data dimensionality by PCA. However, this would allow one only to find an affine subspace containing most of the data variance. It may not be sufficient to effectively decrease the data dimensionality without significant loss of information when the data lies near a non-linear low-dimensional manifold.

Inconsistency in the number of enterotypes found in different works and aggravating factors described above undermine the very notion of enterotypes. Several studies already showed the possibility of a gradient distribution and the absence of well defined clusters [37], [88], [14]. Different structures of enterotypes between males and females were claimed by [57]. In several works, the very concept of enterotypes was described as inconsistent and uncertain. Several factors such as the variation in microbial load between samples [79], the robustness of enterotypes clusters [34], and the microbiome variation during short periods of time [41] were considered limiting to the use of the enterotype concept. However, enterotyping of the human gut has been applied in clinical research. Numerous studies claimed correlations between enterotypes and diet [40], [11], [55], [67], [49], [39], [54], [87], [85], inflammatory gut diseases [81], [10], [52], [22], [13], [31], [20], mental health [46], acne [21], stool composition [76], [78], colorectal cancer [74], circulatory diseases [38], [26], [23], psoriasis [15], and infections such as AIDS [56] and influenza [61], [70]. The idea that the information about the enterotype of an individual may be a helpful biomarker not only to correct gut diseases but also to aid other medical interventions [28] relies on the assumption that enterotypes are discontinuous clusters that are stable in time at least on the short scale; this has been challenged recently [12].

Here, we look for stable and distinct clusters in large stool microbiome datasets. The stability of clusters implies that they should not depend on the dataset, data bootstrapping, and transformations that preserve the general form of the data point cloud. Meaningful clusters should not disappear if the dataset is changed in a non-essential way. Partition of the data into distinctive clusters should correspond to high values of an appropriate clustering validity index that is applicable both for convex and density-based clusters. In other words, there should be gaps around the clusters seen as space regions with a lower concentration of points, whereas clusters correspond to areas of the highest concentration. We introduce two new intermediate steps into the common pipeline for the microbiome clustering analysis: estimation of the intrinsic dimension and manifold learning. These steps allowed us to significantly reduce the data dimensionality while preserving most of the information. Using such low-dimensional representation of the data we demonstrate the absence of stable and distinct clusters in several large datasets of 16S rRNA genotyping of stool samples.

Firstly, using Principal Component Analysis (PCA) we find a medium-dimensional linear subspace retaining almost all data cloud variation so that the dataset projected on this subspace would not significantly differ from the original dataset. Removing dimensions with low variance also

acts as a filter that provides a more robust clustering [6]. However, instead of limiting the dimensionality reduction process solely to PCA, we then determine the intrinsic dimension of the projected data via Maximum Likelihood Estimation (MLE) [47]. This step allows for capturing a minimal but sufficient number of coordinates representing the most important features of the dataset. Following the manifold hypothesis [24], we suppose that microbial data cloud lies near some lower-dimensional manifold embedded in the high-dimensional abundance space. The goal of non-linear manifold learning is to obtain a low-dimensional representation of the data, supposedly lying on such a manifold while preserving most of the information. This information may be expressed as similarities or dissimilarities between data points, e.g. as a matrix of pairwise distances. non-linear projections per se are not interesting, since any data cloud could perfectly lie on a 1-dimensional submanifold.

Despite this 1-dimensional submanifold manifold yield perfect alignment in terms of minimization of the distance between the original data point and its projection, it does not preserve information in terms of pairwise distances. Therefore it is important to assess the quality of embeddings provided by manifold learning algorithms. Proper embedding should preserve the local and global structure e.g. points that are close in the original space should remain close in the embedding space. Given the intrinsic dimension, we further reduce the data dimensionality, using several manifold learning algorithms described in [Materials and Methods](#).

For each embedding, we apply different clustering methods, namely, Spectral Clustering [69], [82], PAM and HDBSCAN [50]. Thus we obtain robust results not biased by peculiarities of the methods. HDBSCAN and Spectral Clustering are useful when the structure of the clusters is arbitrarily shaped and non-convex. We use PAM as a baseline and for comparison with the related works. For each clustering method, we iterate over a set of hyperparameter combinations to find a partition that yields the best clustering validity metrics.

Clustering metrics based on the ratio of within-cluster compactness to between-cluster separation, like the Calinski-Harabasz index (Calinski and Harabasz 1974), the Silhouette score [64], and the Davies-Bouldin index [18] cannot handle properly arbitrarily shaped clusters and noise in form of low-density points scattered around dense clusters areas. Thus, as the main metric for clustering validity, we consider Density-Based Clustering Validation (DBCV) [53], which considers both density and shape properties of clusters, takes noise into account, and is appropriate for detecting density-based clusters.

We assess the clustering partition stability using Prediction Strength [75], initially proposed for estimating the number of clusters, which tells us how well the clustering partition decision boundaries calculated on a data subset generalize the data distribution.

In addition, to be considered as a stable enterotype, a cluster should contain a sufficiently large number of data points. Hence, we do not consider spurious clusters that contain less than 5% of data assuming that such clusters are outliers or artifacts of dimensionality reduction algorithms. Indeed, they are small, depend on manifold learning algorithms, and

are significantly separated from the main data point cloud. Nevertheless, they significantly impact the clustering quality metrics. To distinguish such ineligible partitions, we use the Shannon Entropy [68] of the probability distribution of data points to be in a certain cluster. All these metrics are described in detail in Supplementary Notes 6.1

To identify the optimal clustering, we compare the DBCV, Prediction Strength, and Entropy for each partition corresponding to different clustering hyperparameters. The balanced clustering partition that corresponds to a separation of the data cloud into distinct and stable clusters should produce a salient local maximum of the DBCV score, Entropy value, and Prediction Strength. Also, for each clustering partition, we calculate the Davies-Bouldin index and the Silhouette score. Lower Davies-Bouldin index and higher Silhouette score correspond to a better partition, where clusters are better in terms of compactness and separation. To demonstrate the continuous nature of the stool microbial data distribution, we construct 2d and 3d coordinate projections of the data using t-distributed Stochastic Neighbor Embedding (t-SNE) [77] and Uniform Manifold Approximation and Projection (UMAP) algorithms [51]. Hence, the workflow may be summarized as follows:

1. Use the standard PCA algorithm to obtain the initial affine subspace containing most of the data variance
2. Estimate the intrinsic dimension of the projected data using the MLE algorithm [47]
3. For each manifold learning method, find a low-dimensional representation of the data with the estimated intrinsic dimension, using the method's hyperparameters that provide the lowest error of the learnable inverse transformation
4. Compare the clustering quality of the data for different numbers of clusters and method's hyperparameters using the Density-Based Clustering Validation index (DBCV) [53], the Davies-Bouldin Index [18], and the Silhouette score [64]
5. Assess the clustering partition stability using the Prediction Strength [75], and evaluate partition balance using the Shannon entropy of the data distribution over clusters
6. Visualize the low-dimensional representations using t-SNE [77] and UMAP [51] algorithms

2 Materials and Methods

As the main source of the human gut microbiome data, we used the 16S rRNA genotype data from the NIH Human Microbiome Project (HMP) (Human Microbiome Project Consortium 2012) [33], [35] and American Gut Project (AGP) [3]. These largest available datasets provide sufficient number of data points for correct estimation of the clustering partition and constructing a manifold [60]. We used 3457 HMP samples from stool and rectum body sites downloaded from <https://portal.hmpdacc.org/> and 9511 samples from

Dataset	Size	Order d	Family d	Genus d
AGP	9511	53	95	180
HMP	3457	51	96	172

Table 1: Size and initial dimensionality d of the different datasets.

AGP downloaded from <https://figshare.com/> as abundance matrices. All datasets were normalized by dividing the OTUs values by the total sum of abundances for a given data sample. The numbers of objects and original dimensionality d in the relative taxon abundance space are presented in Tab. 1. The data were analyzed at the Order, Family, and Genus taxonomic levels (O, F, and G, respectively).

The data distribution is inherently sparse due to the insufficient number of samples and noisy due to the presence of possible outliers. Noise may correspond to specific patients with extremely exotic microbial communities or be caused by sample collection or data processing flaws. Moreover, the procedures may vary between laboratories, leading to considerable batch effects. Example of a dataset-specific pre-processing is a correction for microbial blooms for the AGP dataset [2]. Therefore we can not merge several individual datasets into one. Indeed, visualization using the t-SNE and UMAP dimensionality reduction algorithms illustrates that point Fig. 1. Hence, to avoid the batch effect, we considered these datasets separately. Spurious OTU that were found in less than 1% of the samples and with a standard deviation less than 0.001 were removed.

Figure 1: 3D t-SNE and UMAP visualizations of joined datasets demonstrate batch-effect at O - Order taxonomy level (tax). Red - AGP, green - HMP dataset.

For most dimensionality reduction methods, validity indices, and metrics, implementations from the 'scikit-learn' library [59] were used. For the HDBSCAN clustering method and the Density-Based Clustering Validation metric the 'hdbscan' package [50] from the 'scikit-learn-contrib' was used. For the UMAP algorithm we applied implementation from the 'umap-learn' library [51]. Code reproducing the paper results is available at <https://github.com/blufzzz/Human-Gut-Microbiome-Analysis>.

3 Results

3.1 Principal Component Analysis

We obtained significant dimensionality reduction with minuscule information loss by using projection on relatively many (16 through 52, dependent on the taxonomy level) principal components. The dimensionalities d_{PCA} of the PCA projections, are reported in Tab. 2 and defined as the number of first principal components that meet the selected threshold of 99% of the cumulative explained variance. To assess the information loss during the projection, we calculated the error of reconstruction from the projected data to the original one. The Median Absolute Error and Q_{loc} and Q_{glob} metrics (see Supplementary Notes 6.1 for details) of the reconstruction obtained using an inverse linear transformation from the projected data are presented in Tab. 3.

3.2 Estimation of the intrinsic dimension

We calculated the intrinsic dimensions d_{MLE} for each dataset at each taxonomic level by applying the Maximum Likelihood Estimation principle to the distances between close neighbors [47]. The intrinsic dimensions and the dimensionality after projection on principal components are presented in Tab. 2.

	Tax O		Tax F		Tax G	
Dataset	d_{PCA}	d_{MLE}	d_{PCA}	d_{MLE}	d_{PCA}	d_{MLE}
AGP	16	5	35	8	52	9
HMP	18	5	36	6	52	7

Table 2: Dimensionalities of two datasets (AGP and HMP) and three taxonomy levels (O - Order, F - Family, G - Genus). d_{PCA} - number of the first principal components explaining 99% variance, d_{MLE} - estimated intrinsic dimension.

Dataset	Tax	MAE	Q_{loc}	Q_{glob}
AGP	O	0.061	0.90	0.99
	F	0.042	0.96	0.99
	G	0.051	0.95	0.99
HMP	O	0.036	0.91	0.98
	F	0.041	0.92	0.98
	G	0.057	0.92	0.98

Table 3: Reconstruction MAE of the linear inverse transformation from data projected on the principal components to the original space of taxon abundances and the Q_{loc} and Q_{glob} metrics (see the text for details). Notation as in Tab. 2

3.3 Manifold learning

For non-linear dimensionality reduction from d_{PCA} to d_{MLE} , we use manifold learning algorithms, Isomap [73], Modified Locally Linear Embedding (LLE) [90], Denoising AutoEncoder (AE) [17], Spectral Embedding [69], t-distributed Stochastic Neighbor Embedding (t-SNE) [77], and Uniform

Manifold Approximation and Projection (UMAP) [51]. For each manifold learning algorithm, dataset, and taxonomic level, we obtain a low-dimensional embedding. To find an optimal algorithm for a specific dataset and taxonomic level, we iterate over various combinations of manifold learning algorithm hyperparameters, for each combination assessing how well the produced embedding represents the original data. To compare the dimensionality reduction algorithms with regard to the loss of information, we construct an inverse mapping from the obtained low-dimensional manifold back to the original space using the k-Nearest Neighbours Regression [25] and estimate the reconstruction error using Leave-One-Out procedure [65] for the Median Absolute Error (MAE). This technique assesses how well points in the original space can be reconstructed given their neighbors from the embedding space.

Inverse mapping from the embedding to the original space is usually performed by a supervised learning algorithm that minimizes the reconstruction error. Selecting an algorithm, its hyperparameters, and different ways to train it would bring ambiguity into our method. Moreover, the MAE is intractable metric which does not exhibit, what aspect of data cloud has been misrepresented. Therefore we apply two additional quality criteria of dimensionality reduction [45]. These criteria, Q_{loc} and Q_{glob} , represent the preservation of the "local" and the "global" structure, respectively and described in Supplementary Notes 6.1. It should be noted that Q_{glob} is a more important metric for the studied problem since local distortion of the data should not significantly impact the clustering partition that may be implicitly present in the data. Anyhow, for each hyperparameter combination of a dimensionality-reduction algorithm, we assess MAE, Q_{loc} , and Q_{glob} . Given hyperparameters that deliver the lowest MAE, we discard 10% of data points with the highest reconstruction MAE, which serves as denoising for a more stable clustering in the embedding space. It does not affect the clustering partition results, since these data points fall in the outliers category provided by the Local Outlier Factor algorithm [8]. The latter allows for detecting outliers by deviation of their local density with respect to their neighbors, from which it follows they are not related to any cluster as a densely populated area in a multidimensional space. Q_{loc} , and Q_{glob} for all manifold learning methods, applied to all datasets at different taxonomic levels, are shown in Fig. 2. The MAE values are listed in Tab. 4. One may see from Fig. 2 and Tab. 4, higher taxonomy levels yield more intricate data representation with higher MAE and lower Q_{loc} , Q_{glob} . This is due to data becoming more sparse and dissipated in the high-dimensional space, which hinders data representation by fitting a non-linear manifold.

3.4 Clustering

As described in the Introduction, we applied several clustering methods (Spectral Clustering [82], Partition Around Medoids, HDBSCAN [50]) for each embedding provided by dimensionality reduction algorithms described in Results 3.3. We use only the Euclidean metric for the manifold learning algorithms. For comparison with related works on the identification of enterotypes, we applied clustering

Figure 2: Horizontal axis - Q_{loc} metric (local information preservation), vertical axis - Q_{glob} metric (global information preservation) of the non-linear dimensionality reduction methods. Datasets (AGP and HMP) and taxonomy levels (O - Order, F - Family, G - Genus) are shown in the inset. SE - Spectral Embedding, LLE - Locally Linear Embedding, AE - AutoEncoder.

Dataset	Method	Tax O	Tax F	Tax G
AGP	AutoEncoder	0.05	0.19	0.21
	t-SNE	0.05	0.19	0.22
	UMAP	0.06	0.21	0.24
	Isomap	0.06	0.22	0.25
	LLE	0.06	0.22	0.24
	Spectral	0.06	0.21	0.24
HMP	AutoEncoder	0.09	0.24	0.22
	t-SNE	0.09	0.23	0.21
	UMAP	0.11	0.27	0.25
	Isomap	0.11	0.29	0.26
	LLE	0.11	0.28	0.26
	Spectral	0.12	0.34	0.28

Table 4: The reconstruction MAE, assessed using Leave-One-Out procedure. The reconstruction is done by the independent K-Nearest Neighbour Regression of the original space of relative taxon abundances data coordinates from the non-linear embedding. Notation as in Tab. 2

to the original data in the high-dimensional space of taxonomic abundances with different distance metrics: Jensen-Shannon, Manhattan, Euclidean, and Bray-Curtis as in [44].

These results are included in Fig. 4 and Fig. 3. To distinguish between the presence and absence of clusters in the data we consider the following thresholds. For the Prediction Strength, we consider a score of 0.8 for moderate support as suggested in [44] and [75]. We consider all positive values of the DBCV metric. For the Silhouette score, we consider a score of 0.5 for moderate clustering as suggested in [85], [44], [27], [4]. As a threshold value for the Davies-Bouldin index, we used 0.6 for moderate clustering [18]. Under the assumption that our data capture all microbiome variations that may be possibly related to enterotypes, we do not consider small clusters that contain less than 5% of

the data as natural clusters related to enterotypes.

To validate the ability of our methods to provide accurate clustering of data that lies in a low-dimensional non-linear manifold, embedded in a high dimensional space, we used several synthetic datasets embedded in high-dimensional space (see Supplementary Notes 6.2 for details). Since the clustering results directly depend on the algorithm hyperparameters, for each clustering method we have iterated over relevant sets of hyperparameter values. At that, for the Spectral Clustering algorithm, we considered the range from 2 to 10 as the possible number of clusters in the data, and 3, 5, and 15 as sizes of the neighborhood for computing the affinity matrix. For the PAM algorithm, we used the same range of possible clusters. For the HDBSCAN hyperparameter, we considered 10, 25, 50 as the minimal cluster size. We did not set large values for the minimal cluster size to avoid conservative clustering when more points will be considered as noise, and clusters will be restricted to more dense areas. If there is more than one partition into the same number of clusters provided by HDBSCAN, we selected the one that yields a higher DBCV index. The results of clustering demonstrated the absence of stable and distinct clusters in the data (Fig. 3 and Fig. 4).

Figure 3: Clustering results for the AGP and HMP datasets including all three taxonomy levels (O - Order, F - Family, G - Genus) - with the corresponding Silhouette score and Davies-Bouldin index. The point color represents the Entropy of the respective partition.

Figure 4: Clustering results for the AGP and HMP datasets with the corresponding DBCV index and the Prediction Strength. Notation as in Fig. 3.

In Fig. 3 and Fig. 4, we show the distribution of metrics over all clustering results, calculated for each taxonomic level, manifold learning method, clustering algorithm, and its hyperparameters. Clustering partitions with moderate

or strong support for both metrics correspond to points lying at the intersection of the blue and orange areas at Fig. 3 and Fig. 4. All partitions that correspond to points in this area were found to consist of two highly imbalanced clusters when more than 95% of the data points are concentrated in one of them. These partitions are also inconsistent in terms of clustering validity metrics when either DBCV and Prediction Score or Davies-Bouldin and Silhouette pairs pass the thresholds but not all of them. These small clusters are unstable among different partitions of the same dataset - the mean intersection between them is lower than 30%, which asserts for high dependence on manifold learning and clustering algorithms.

Three partitions that satisfy the Davies-Bouldin and Silhouette score thresholds for the AGP dataset were found in Fig. 3. One partition originates from the Locally Linear Embedding of the Order taxonomic level, where two clusters were found by the Spectral clustering algorithm. For this partition, the Davies-Bouldin index is 0.54 and the Silhouette score is 0.65, which is inconsistent with low value -0.89 of DBCV, and 0.31 for the Prediction Score. The entropy of the data mass distribution over two found clusters is equal to 0.095 which corresponds to a highly imbalanced partition where 98% of data are concentrated in one cluster.

Another two partitions of the AGP dataset into two clusters were found in the original data in high-dimensional spaces of normalized taxon-relative abundances in Order level, endowed with Jensen-Shannon and Euclidean metrics. For Jensen-Shannon the Davies-Bouldin index is 0.32, the Silhouette score is 0.81, the DBCV is -0.81 and the Prediction Score is 0.25. The entropy of the data mass distribution over two found clusters is equal to 0.099. For Euclidean distance the Davies-Bouldin index is 0.58, the Silhouette score is 0.58, the DBCV is -0.69 and the Prediction Score is 0.72. The entropy is 0.098. Another partition with DBCV and the Prediction Score above the threshold was found in the UMAP embedding of the Genus level by the HDBSCAN algorithm, presented in Fig. 4, lying at the intersection of the threshold areas. The partition consists of two clusters with the Davies-Bouldin index of 0.99, the Silhouette score of 0.26 which is incoherent with high DBCV of 0.52 and the Prediction Score of 0.92. Entropy is equal to 0.06 which corresponds to an imbalanced partition where 98% of data belong to one cluster.

For the HMP dataset, three highly imbalanced partitions into two clusters that are satisfying the metrics thresholds were found. The first one in Fig. 3 was generated by the HDBSCAN clustering algorithm in the Spectral embedding at the Family taxonomic level. This partition yields the Davies-Bouldin index of 0.54, the Silhouette score of 0.66, DBCV of -0.64, the Prediction Score of 0.21, and the Entropy of 0.091. The second one in Fig. 3 was found in the Locally Linear embedding of the Order level by the Spectral Clustering algorithm with the Davies-Bouldin index of 0.50, the Silhouette score of 0.54, DBCV of -0.84, the Prediction Score of 0.45, and the Entropy of 0.1. The only clustering partition that satisfies the Prediction Score and DBCV thresholds for the HMP dataset at Fig. 4 was found in the t-SNE embedding at the Order level, by the HDBSCAN al-

gorithm. For this clustering, the Davies-Bouldin index is 1.01, the Silhouette score is 0.11, DBCV 0.01, the Prediction Score is 0.87, and the Entropy is 0.05, which indicates 99% of data concentrated in one cluster.

Together, these results imply that various clustering methods along with different manifold learning algorithms yield highly imbalanced partitions. We do not consider clusters that contain less than 5% of the data as enterotypes. We attribute these small clusters to artifacts of manifold learning algorithms since there are evidences [16], [42] that common dimensionality reduction techniques may fail to faithfully represent the original point cloud distribution, introducing substantial distortion into the data. We observe that these small clusters are not stable, depending on the clustering method and the manifold learning algorithm. Hence, the stool metagenomes can hardly be divided into stable and distinct clusters that could be referred to as enterotypes. Our simulation on a synthetic dataset in Supplementary Notes 6.2, reaffirms that distinct and stable clusters related to enterotypes were not found because of their absence in the data rather than methodology flaws.

Despite the lack of distinct and stable clusters in the data, we demonstrate that human gut microbial communities vary continuously along a low-dimensional manifold. We observe the structure of such a manifold by mapping the point cloud, projected on the principal components, on a 2-dimensional plane, using t-SNE in Fig. 5 and UMAP in Fig. 6.

Figure 5: 2D t-SNE visualization of AGP and HMP datasets for the Order taxonomy level. Colors reflect the relative abundance of specific taxa, see the headers.

To demonstrate the continuous specificity of the data points distribution, we colored points corresponding to specific taxon relative abundances. Salient parts of the manifold correspond to higher concentrations of specific OTUs. Small clusters observed in Fig. 5 and Fig. 6 are not related to enterotypes being the direct result of specific methods hyperparameters, that may lead to tearing off the salient part of data manifold.

Figure 6: 2D UMAP visualization of AGP and HMP datasets for the Order taxonomy level. Colors reflect the relative abundance of specific taxa, see the headers.

4 Discussion

Our results support that the metagenome distribution is continuous rather than discrete and lies on a low-dimensional non-linear manifold embedded in the original high-dimensional space of relative taxon abundances. We posit that most of the previous observations may have been artifacts caused by limitations of linear methods that have been applied for the analysis of non-linear and high-dimensional microbial data. Also, overfitting in the data density estimation may occur due to insufficient numbers of data points. Small sizes of datasets lead to unstable clustering, especially if it is performed in a high-dimensional space. One may suggest an intuitive explanation of why positive clustering results were widespread in previous works. In most of them, small datasets were used, which makes the total number of abnormal microbial patterns negligible. We suppose that datasets demonstrating moderate clustering in related works have been sampled from high-density areas in the general taxonomic abundance space. Such areas correspond to prevalent types of microbial patterns, whereas areas with low density correspond to abnormal and unusual ones. Still, successful attempts to correct the human gut microbiota were made, e.g. fecal microbiota transplantation to treat *Clostridium difficile* infection (CDI) [71], inflammatory bowel disease (IBD) [72], and obesity [83]. Connecting the distribution of microbiome abundances and features of the microbial manifold with lifestyle, nutrition, or disease may be a promising direction for further research.

5 Funding

Sample preparation, data analysis, and biological interpretation of the results were supported by the Russian Science Foundation (grant 18-14-00358 to M.S.G.). Algorithm development and data processing were supported by the Ministry of Science and Higher Education (grant No. 075-10-2021-068).

6 Supplementary Notes

6.1 Metrics

The Silhouette score S_i [64] for a data point i is defined as:

$$S_i = \frac{(b_i - a_i)}{\max(b_i, a_i)} \quad (1)$$

In Eq. 1, a_i is the mean intra-cluster distance and b_i is the mean distance between the sample point i and the data points from the nearest cluster C_k that sample i is not a part of.

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in |C_i|, i \neq j} d(i, j) \quad (2)$$

$$b_i = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in |C_k|} d(i, j) \quad (3)$$

Here in Eq. 2, 3, $|C_i|$ is the number of points in a cluster C_i and $d(i, j)$ is the distance between data points i and j . The total Silhouette score is measured as the mean of all S_i and bounded between -1 for incorrect clustering and +1 for dense and well-separated clusters. The Davies-Bouldin index DB [18] in Eq. 4 measures the average similarity of the distance between clusters with the size of the clusters themselves. It estimates the cohesion based on the distance from the points in a cluster to its centroid and the separation based on the distance between centroids. Values closer to zero indicate better partitions.

$$DB = \frac{1}{k} \sum_{i=1}^K \max_{i \neq j} R_{ij} \quad R_{ij} = \frac{s_i + s_j}{d(\mu_i, \mu_j)} \quad (4)$$

In Eq. 4, s_i is the average distance between each point of cluster i and a centroid μ_i of that cluster, K is the total number of clusters, and $d(\mu_i, \mu_j)$ is the distance between centroids μ_i and μ_j of the clusters i and j .

Density-Based Clustering Validation was proposed as a metric for assessing the clustering partition quality for non-convex clusters. The metric is based on the Hartigan model of Density Contour Trees [30] and provides values between -1 and +1, with greater values indicating a better density-based clustering solution. See [53] for details.

To assess disbalance in a clustering partition, we use the Shannon Entropy [68] E of the distribution that indicates the probability p_i of one of n data points to fall into a cluster i with $|C_i|$ data points in it, see Eq. 5. Higher E values indicate more balanced partition.

$$E = - \sum_{i=1}^k p_i \log(p_i) \quad p_i = \frac{|C_i|}{n} \quad (5)$$

We estimate the reconstruction error of inverse mapping $\hat{x}_i = g(z_i)$ from the manifold learning embedding $z \in Z$ to the original space of relative taxon abundances as the normalized Median Absolute Error (Eq. 6):

$$\text{MAE}(X, \hat{X}) = \text{median} \left\{ \frac{|x_i - \hat{x}_i|_1}{|x_i|_1}, i \in \overline{1, N} \right\} \quad (6)$$

Selecting and supervised training of the inverse mapping algorithm g brings ambiguity into the method. To mitigate it, we use additional scale-independent rank-based quality criteria for dimensionality reduction proposed in [45]. Given a distance metric $d(i, j)$, two rank matrices are constructed $\rho_{ij} = \{k : d(x_k, x_j) < d(x_i, x_j)\}$ for the original data points $x_i \in X$ and $r_{ij} = \{k : d(z_k, z_j) < d(z_i, z_j)\}$ for their low-dimensional embeddings $z_i \in Z$. Then, a co-ranking matrix Q is built, where $Q_{kl} = \{(i, j) : \rho_{ij} = k, r_{ij} = l\}$ indicates the total number of instances when the k -th neighbor for some point became l -th. Thus, all non-diagonal elements of this matrix correspond to changes in arrangement of the embedded points compared to the original data. The co-ranking matrix contains all information on how the data structure is distorted in a low-dimensional representation. This information is represented further as two scalars Q_{loc} and Q_{glob} that range from 0 (bad) to 1 (good) and reflect the preservation of the "local" and "global" structure of the data cloud. See [45] for details.

For similarity measure between the predicted clustering partition X and given ground-truth labels Y we utilize Adjusted Rand Index [32], based on the Rand Index (RI) [62]. Given the predicted $X = \{X_i\}_{i=1}^r$ and the true $Y = \{Y_j\}_{j=1}^s$ clustering partitions into r and s clusters respectively, the Rand Index is non-zero for random independent partitions X and Y and 1 for the same partitions up to labeling permutations. The Adjusted Rand Index is "adjusted" version of RI such its values are close to 0 for random labeling. In Tab. 7 a contingency table is presented, where n_{ij} denotes the common number of points in clusters X_i and Y_j : $n_{ij} = |X_i \cap Y_j|$. Given a_i, b_j, n_{ij} from Tab. 7, the Adjusted Rand Index (ARI) is defined as Eq. 8. The $C_n^k = \frac{n!}{k!(n-k)!}$ denotes a k -combination from a set of n elements.

X	Y_1	Y_2	$\dots$	Y_s	$a_i = \sum_k n_{ik}$
X_1	n_{11}	n_{12}	$\dots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\dots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\dots$	n_{rs}	a_r
$b_j = \sum_k n_{kj}$	b_1	b_2	$\dots$	b_s	

(7)

$$\text{ARI} = \frac{\sum_{ij} C_{n_{ij}}^2 - \left[\sum_i C_{a_i}^2 \sum_j C_{b_j}^2 \right] / C_n^2}{\frac{1}{2} \left[\sum_i C_{a_i}^2 + \sum_j C_{b_j}^2 \right] - \left[\sum_i C_{a_i}^2 \sum_j C_{b_j}^2 \right] / C_n^2} \quad (8)$$

Synthetic dataset

In this section, we provide the results of our method on synthetic data. The positive control datasets with clusters were created similarly to [44]. We generated nine synthetic datasets with the number of clusters k ranging from two to four, with different dimensionality d equal to the O, F and G taxonomy levels (53, 96 and 180 correspondingly). Each dataset consist of 3000 data points that represent vectors of un-normalized abundances of mock OTUs. 90% of data points were sampled from one of k multivariate gaussian distributions, each reflecting a cluster, and other 10% are sampled from a distribution with large variance, representing

noise. To compare estimated clustering partitions with the ground-truth labels, we used the Adjusted Rand Index [32] (see Supplementary Notes 6.1 for details). The dimensionality that retains 99% of the cumulative explained variance after the PCA projection, estimated intrinsic dimension, the Median Absolute Error and Q_{loc} , Q_{glob} metrics are reported in Tab. 5.

d	k	d_{PCA}	d_{MLE}	MAE	Q_{loc}	Q_{glob}
53	2	2	2	0.045	0.89	0.98
53	3	2	2	0.074	0.85	0.98
53	4	4	3	0.068	0.86	0.98
96	2	2	2	0.080	0.85	0.97
96	3	6	4	0.086	0.86	0.97
96	4	22	8	0.082	0.87	0.98
180	2	109	21	0.091	0.86	0.97
180	3	58	13	0.086	0.86	0.98
180	4	21	7	0.095	0.85	0.98

Table 5: Original dimensionality d , dimensionality d_{PCA} after PCA projection and estimated intrinsic dimensionality d_{MLE} of synthetic datasets. For each dataset MAE, Q_{loc} and Q_{glob} metrics were computed for the PCA projection with respect to original data.

In Fig. 7 we display the clustering metrics distribution over all partitions that have been calculated for each dataset with a given number of clusters k and dimensionality d , each manifold learning method, and each clustering algorithm with a different combination of its hyperparameters. Clustering partitions with moderate or strong support for both metrics correspond to points lying at the intersection of the blue and orange areas in Fig. 7. For each dataset the most accurate clustering partition was found in the UMAP embedding by the HDBSCAN clustering algorithm. Such partitions have the Adjusted Rand Index and the Prediction score equal to 1, the Davies-Bouldin index lower than 0.3, the Silhouette score higher than 0.5 and DBCV higher than 0.8. Together, these metrics assert presence of stable and distinct clusters in the data for every number of clusters k and dimensionality d .

Figure 7: Silhouette score and Davies-Bouldin index (left), DBCV index and Prediction Strength (right) of clustering partitions for all nine synthetic datasets with the number of clusters k equal to 2, 3 and 4, and the dimensionality d of O, F and G taxonomy levels, equal to 53, 96 and 180.

References

- [1] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 1973, pages 420–434. Springer Berlin Heidelberg, 2001.
- [2] Amnon Amir, Daniel McDonald, Jose A. Navas-Molina, Justine Debelius, James T. Morton, Embriette Hyde, Adam Robbins-Pianka, and Rob Knight. Correcting for Microbial Blooms in Fecal Samples during Room-Temperature Shipping. *mSystems*, 2(2), 2017.
- [3] Daniel and McDonald, Embriette Hyde, Justine W. Debelius, James T. Morton, Antonio Gonzalez, Gail Ackermann, Alexander A. Aksenov, Bahar Behsaz, Caitri-ona Brennan, Yingfeng Chen, Lindsay DeRight Goldsich, Pieter C. Dorrestein, Robert R. Dunn, Ashkaan K. Fahimipour, James Gaffney, Jack A. Gilbert, Grant Gogul, Jessica L. Green, Philip Hugenholtz, Greg Humphrey, Curtis Huttenhower, Matthew A. Jackson, Stefan Janssen, Dilip V. Jeste, Lingjing Jiang, Scott T. Kelley, Dan Knights, Tomasz Kosciolk, Joshua Ladau, Jeff Leach, Clarisse Marotz, Dmitry Meleshko, Alexey V. Melnik, Jessica L. Metcalf, Hosein Mohimani, Emmanuel Montassier, Jose Navas-Molina, Tanya T. Nguyen, Shyamal Peddada, Pavel Pevzner, Katherine S. Pollard, Gholamali Rahnavard, Adam Robbins-Pianka, Naseer Sangwan, Joshua Shorenstein, Larry Smarr, Se Jin Song, Timothy Spector, Austin D. Swafford, Varykina G. Thackray, Luke R. Thompson, Anupriya Tripathi, Yoshiki Vázquez-Baeza, Alison Vrbanc, Paul Wischmeyer, Elaine Wolfe, Qiyun Zhu, and Rob Knight. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*, 3(3), 2018.
- [4] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M. Pérez, and I simnigo Perona. An extensive comparative study of cluster validity indices, 2013.
- [5] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R. Mende, Gabriel R. Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, Marcelo Bertalan, Natalia Borruel, Francesc Casellas, Leyden Fernandez, Laurent Gautier, Torben Hansen, Masahira Hattori, Tetsuya Hayashi, Michiel Kleerebezem, Ken Kurokawa, Marion Leclerc, Florence Levenez, Chaysavanh Manichanh, H. Bjørn Nielsen, Trine Nielsen, Nicolas Pons, Julie Poulain, Junjie Qin, Thomas Sicheritz-Ponten, Sebastian Tims, David Torrents, Edgardo Ugarte, Erwin G. Zoetendal, Jun Wang, Francisco Guarner, Oluf Pedersen, Willem M. de Vos, Søren Brunak, Joel Doré, Jean Weissenbach, S. Dusko Ehrlich, and Peer Bork. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, may 2011.
- [6] Asa Ben-Hur and Isabelle Guyon. Detecting Stable Clusters Using Principal Component Analysis. *Functional Genomics*, 224:159–182, 2003.
- [7] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When Is “Nearest Neighbor” Meaningful? In *Database Theory — ICDT’99*, pages 217–235. Springer Berlin Heidelberg, 1999.
- [8] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. LOF. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD ’00*, pages 93–104, New York, New York, USA, 2000. ACM Press.
- [9] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27, 1974.
- [10] Natalia Casta simno-Rodríguez, Alexander P. Underwood, Juan Merif, Stephen M. Riordan, William D. Rawlinson, Hazel M. Mitchell, and Nadeem O. Kaakoush. Gut microbiome analysis identifies potential etiological factors in acute gastroenteritis. *Infection and Immunity*, 86(7), 2018.
- [11] Tingting Chen, Wenmin Long, Chenhong Zhang, Shuang Liu, Liping Zhao, and Bruce R. Hamaker. Fiber-utilizing capacity varies in Prevotella- versus Bacteroides-dominated gut microbiota. *Scientific Reports*, 7(1):2594, 2017.
- [12] Mingyue Cheng and Kang Ning. Stereotypes About Enterotype: the Old and New Ideas. *Genomics, Proteomics and Bioinformatics*, 17(1):4–12, 2019.
- [13] Chih Min Chiu, Wei Chih Huang, Shun Long Weng, Han Chi Tseng, Chao Liang, Wei Chi Wang, Ting Yang, Tzu Ling Yang, Chen Tsung Weng, Tzu Hao Chang, and Hsien Da Huang. Systematic analysis of the association between gut flora and obesity through high-throughput sequencing and bioinformatics approaches. *BioMed Research International*, 2014:906168, 2014.
- [14] Marcus J. Claesson, Ian B. Jeffery, Susana Conde, Susan E. Power, Eibhlís M. O’connor, Siobhán Cusack, Hugh M.B. Harris, Mairead Coakley, Bhuvaneshwari Lakshminarayanan, Orla O’sullivan, Gerald F. Fitzgerald, Jennifer Deane, Michael O’connor, Norma Harnedy, Kieran O’connor, Denis O’mahony, Douwe Van Sinderen, Martina Wallace, Lorraine Brennan, Catherine Stanton, Julian R. Marchesi, Anthony P. Fitzgerald, Fergus Shanahan, Colin Hill, R. Paul Ross, and Paul W. O’toole. Gut microbiota composition correlates with diet and health in the elderly. *Nature*, 488(7410):178–184, 2012.
- [15] Francisco M. Codo sinmer, Ana Ramírez-Bosca, Eric Climent, Miguel Carrión-Gutierrez, Mariano Guerrero, Jose Manuel Pérez-Orquín, José Horga De La Parte, Salvador Genovés, Daniel Ramón, Vicente Navarro-López, and

- Empar Chenoll. Gut microbial composition in patients with psoriasis. *Scientific Reports*, 8(1):3812, 2018.
- [16] Shamus M. Cooley, Timothy Hamilton, Eric J. Deeds, and J. Christian J. Ray. A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data. *bioRxiv*, 2019.
- [17] Ian Goodfellow Courville, Yoshua Bengio, and Aaron. *Deep learning Learning*, volume 29 of *Adaptive Computation and Machine Learning series*. MIT Press, London, England, 2016.
- [18] David L. Davies and Donald W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.
- [19] Ana C. F. de Moraes, Gabriel R. Fernandes, Isis T. da Silva, Bianca Almeida-Pititto, Everton P. Gomes, Alexandre da Costa Pereira, and Sandra R. G. Ferreira. Enterotype May Drive the Dietary-Associated Cardiometabolic Risk Factors. *Frontiers in Cellular and Infection Microbiology*, 7(FEB):47, feb 2017.
- [20] Tomas De Wouters, Joël Doré, and Patricia Lepage. Does our food (environment) change our gut microbiome ('In-Vironment'): A potential role for inflammatory bowel disease? *Digestive Diseases*, 30(SUPPL. 3):33–39, 2013.
- [21] Yongqiong Deng, Hong Wang, Jiyuan Zhou, Yunzhu Mou, Guiqiang Wang, and Xia Xiong. Patients with acne vulgaris have a distinct gut microbiota in comparison with healthy controls. *Acta Dermato-Venereologica*, 98(8):783–790, 2018.
- [22] Aldona Dlugosz, Björn Winckler, Elin Lundin, Katharina Zakikhany, Gunnar Sandström, Weimin Ye, Lars Engstrand, and Greger Lindberg. No difference in small bowel microbiota between patients with irritable bowel syndrome and healthy controls. *Scientific Reports*, 5:8508, 2014.
- [23] Takuo Emoto, Tomoya Yamashita, Naoto Sasaki, Yushi Hirota, Tomohiro Hayashi, Anna So, Kazuyuki Kasahara, Keiko Yodoi, Takuya Matsumoto, Taiji Mizoguchi, Wataru Ogawa, and Ken Ichi Hirata. Analysis of gut microbiota in coronary artery disease patients: A possible link between gut microbiota and coronary artery disease. *Journal of Atherosclerosis and Thrombosis*, 23(8):908–921, 2016.
- [24] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.
- [25] Evelyn Fix and J. L. Hodges. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238, dec 1989.
- [26] Ana Carolina Franco-De-Moraes, Bianca De Almeida-Pititto, Gabriel Da Rocha Fernandes, Everton Padilha Gomes, Alexandre Da Costa Pereira, and Sandra Roberta G. Ferreira. Worse inflammatory profile in omnivores than in vegetarians associates with the gut microbiota composition. *Diabetology and Metabolic Syndrome*, 9(1):62, 2017.
- [27] J. E. Gentle, L. Kaufman, and P. J. Rousseuw. *Finding Groups in Data: An Introduction to Cluster Analysis.*, volume 47. John Wiley & Sons, jun 1991.
- [28] Jack A. Gilbert, Martin J. Blaser, J. Gregory Caporaso, Janet K. Jansson, Susan V. Lynch, and Rob Knight. Current understanding of the human microbiome. *Nature Medicine*, 24(4):392–400, apr 2018.
- [29] Takuji Gotoda. Recent topics of gut microbiota gut microbiota composition and activity in relation to host metabolic phenotype and disease risk. *Journal of Tokyo Medical University*, 73(1):16–17, 2015.
- [30] John A. Hartigan, Helmut Spath, and J. Van Ryzin. *Clustering Algorithms*, nov 1981.
- [31] Hu Huang, Pajau Vangay, Christopher E. McKinlay, and Dan Knights. Multi-omics analysis of inflammatory bowel disease. *Immunology Letters*, 162(2):62–68, 2014.
- [32] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, dec 1985.
- [33] The Human and Microbiome Project. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, jun 2012.
- [34] John Huss. Methodology and Ontology in Microbiome Research. *Biological Theory*, 9(4):392–400, 2014.
- [35] C Huttenhower, D Gevers, R Knight, B a Methé, K E Nelson, M Pop, H H Creasy, M G Giglio, J F Petrosino, S Abubucker, J H Badger, and Others. A framework for human microbiome research. *Nature*, 486(7402):215–221, jun 2012.
- [36] I A Ibragimov and R. Z. Has'minskii. *Statistical Estimation*. Springer New York, New York, NY, 1981.
- [37] Ian B. Jeffery, Marcus J. Claesson, Paul W. O'Toole, and Fergus Shanahan. Categorization of the gut microbiota: Enterotypes or gradients? *Nature Reviews Microbiology*, 10(9):591–592, 2012.
- [38] Zhuye Jie, Huihua Xia, Shi Long Zhong, Qiang Feng, Shenghui Li, Suisha Liang, Huanzi Zhong, Zhipeng Liu, Yuan Gao, Hui Zhao, Dongya Zhang, Zheng Su, Zhiwei Fang, Zhou Lan, Junhua Li, Liang Xiao, Jun Li, Ruijun Li, Xiaoping Li, Fei Li, Huahui Ren, Yan Huang, Yangqing Peng, Guanglei Li, Bo Wen, Bo Dong, Ji Yan Chen, Qing Shan Geng, Zhi Wei Zhang, Huanming Yang, Jian Wang, Jun Wang, Xuan Zhang, Lise Madsen, Susanne Brix, Guang Ning, Xun Xu, Xin Liu, Yong Hou, Huijue Jia, Kunlun He, and Karsten Kristiansen. The gut microbiome in atherosclerotic cardiovascular disease. *Nature Communications*, 8(1), 2017.

- [39] Chao Kang, Yong Zhang, Xiaohui Zhu, Kai Liu, Xiaolan Wang, Mengting Chen, Jian Wang, Hua Chen, Suocheng Hui, Li Huang, Qianrong Zhang, Jundong Zhu, Bin Wang, and Mantian Mi. Healthy subjects differentially respond to dietary capsaicin correlating with specific gut enterotypes. *Journal of Clinical Endocrinology and Metabolism*, 101(12):4681–4689, 2016.
- [40] Natalia Klimenko, Alexander Tyakht, Anna Popenko, Anatoly Vasiliev, Ilya Altukhov, Dmitry Ischenko, Tatiana Shashkova, Daria Efimova, Dmitri Nikogosov, Dmitrii Osipenko, Sergey Musienko, Kseniya Selezneva, Ancha Baranova, Alexander Kurilshikov, Stepan Toshchakov, Aleksei Korzhenkov, Nazar Samarov, Margarita Shevchenko, Alina Tepliuk, and Dmitry Alexeev. Microbiome Responses to an Uncontrolled Short-Term Diet Intervention in the Frame of the Citizen Science Project. *Nutrients*, 10(5):576, may 2018.
- [41] Dan Knights, Tonya L Ward, Christopher E McKinlay, Hannah Miller, Antonio Gonzalez, Daniel McDonald, and Rob Knight. Rethinking “Enterotypes”. *Cell Host & Microbe*, 16(4):433–437, oct 2014.
- [42] Dmitry Kobak and Philipp Berens. The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10(1):5416, dec 2019.
- [43] Michael Kohler, Adam Krzyzak, and Harro Walk. Optimal global rates of convergence for nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference*, 139(4):1286–1296, 2009.
- [44] Omry Koren, Dan Knights, Antonio Gonzalez, Levi Waldron, Nicola Segata, Rob Knight, Curtis Huttenhower, and Ruth E. Ley. A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. *PLoS Computational Biology*, 9(1):e1002863, jan 2013.
- [45] John A. Lee and Michel Verleysen. Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters*, 31(14):2248–2257, oct 2010.
- [46] Sung Ha Lee, Seok Hwan Yoon, Yeonjae Jung, Namil Kim, Uigi Min, Jongsik Chun, and Incheol Choi. Emotional well-being and gut microbiome profiles by enterotype. *Scientific Reports*, 10(1):20736, 2020.
- [47] Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in Neural Information Processing Systems*, 17:777–784, 2005.
- [48] Jing Li, Ruiqing Fu, Yajun Yang, Hans Peter Horz, Yaquun Guan, Yan Lu, Haiyi Lou, Lei Tian, Shijie Zheng, Hongjiao Liu, Meng Shi, Kun Tang, Sijia Wang, and Shuhua Xu. A metagenomic approach to dissect the genetic composition of enterotypes in Han Chinese and two Muslim groups. *Systematic and Applied Microbiology*, 41(1):1–12, 2018.
- [49] Chao Liang, Han-Chi Tseng, Hui-Mei Chen, Wei-Chi Wang, Chih-Min Chiu, Jen-Yun Chang, Kuan-Yi Lu, Shun-Long Weng, Tzu-Hao Chang, Chao-Hsiang Chang, Chen-Tsung Weng, Hwei-Ming Wang, and Hsien-Da Huang. Diversity and enterotype in gut bacterial community of adults in Taiwan. *BMC Genomics*, 18(S1):932, jan 2017.
- [50] Leland McInnes and John Healy. Accelerated Hierarchical Density Based Clustering. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2017–November(11):33–42, 2017.
- [51] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.
- [52] Gabriele Moser, Camille Fournier, and Johannes Peter. Intestinal microbiome-gut-brain axis and irritable bowel syndrome. *Wiener Medizinische Wochenschrift*, 168(3-4):62–66, mar 2018.
- [53] Davoud Moulavi, Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. Density-Based Clustering Validation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, volume 2 of *Proceedings*, pages 839–847. Society for Industrial and Applied Mathematics, Philadelphia, PA, apr 2014.
- [54] Jiro Nakayama, Koichi Watanabe, Jiahui Jiang, Kazunori Matsuda, Shiou-Huei Chao, Pri Haryono, Orawan La-ongkham, Martinus-Agus Sarwoko, I. Nengah Sujaya, Liang Zhao, Kang-Ting Chen, Yen-Po Chen, Hsueh-Hui Chiu, Tomoko Hidaka, Ning Xin Huang, Chikako Kiyohara, Takashi Kurakawa, Naoshige Sakamoto, Kenji Sonomoto, Kousuke Tashiro, Hirokazu Tsuji, Ming-Ju Chen, Vichai Leelavatcharamas, Chii-Cherng Liao, Sunee Nititsinprasert, Endang S. Rahayu, Fa-Zheng Ren, Ying-Chieh Tsai, and Yuan-Kun Lee. Diversity in gut bacterial community of school-age children in Asia. *Scientific Reports*, 5(1):8397, jul 2015.
- [55] Jiro Nakayama, Azusa Yamamoto, Ladie A. Palermo-Conde, Kanako Higashi, Kenji Sonomoto, Julie Tan, and Yuan Kun Lee. Impact of westernized diet on gut microbiota in children on Leyte island. *Frontiers in Microbiology*, 8(FEB):197, 2017.
- [56] Marc Noguera-Julian, Muntsa Rocafort, Yolanda Guillén, Javier Rivera, Maria Casadellà, Piotr Nowak, Falk Hildebrand, Georg Zeller, Mariona Parera, Rocío Bellido, Cristina Rodríguez, Jorge Carrillo, Beatriz Mothe, Josep Coll, Isabel Bravo, Carla Estany, Cristina Herrero, Jorge Saz, Guillem Sirera, Ariadna Torrela, Jordi Navarro, Manel Crespo, Christian Brander, Eugènia Negredo, Julià Blanco, Francisco Guarner, Maria Luz Calle, Peer Bork, Anders Sönnnerborg, Bonaventura Clotet, and Roger Paredes. Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine*, 5:135–146, 2016.

- [57] Kaihei Oki, Mutsumi Toyama, Taihei Banno, Osamu Chonan, Yoshimi Benno, and Koichi Watanabe. Comprehensive analysis of the fecal microbiota of healthy Japanese adults reveals a new bacterial lineage associated with a phenotype characterized by a high frequency of bowel movements and a lean body type. *BMC Microbiology*, 16(1):1–13, 2016.
- [58] Junhai Ou, Franck Carbonero, Erwin G. Zoetendal, James P. DeLany, Mei Wang, Keith Newton, H. Rex Gaskins, and Stephen J.D. O’Keefe. Diet, microbiota, and microbial metabolites in colon cancer risk in rural Africans and African Americans. *American Journal of Clinical Nutrition*, 98(1):111–120, 2013.
- [59] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [60] Josef V. Psutka and Josef Psutka. Sample size for maximum likelihood estimates of gaussian model. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9257 of *Lecture notes in computer science*, pages 462–469. Springer International Publishing, Cham, 2015.
- [61] Nan Qin, Beiwen Zheng, Jian Yao, Lihua Guo, Jian Zuo, Lingjiao Wu, Jiawei Zhou, Lin Liu, Jing Guo, Shujun Ni, Ang Li, Yixin Zhu, Weifeng Liang, Yonghong Xiao, S. Dusko Ehrlich, and Lanjuan Li. Influence of H7N9 virus infection and associated treatment on human gut microbiota. *Scientific Reports*, 5:14771, 2015.
- [62] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.
- [63] Virginia Robles-Alonso and Francisco Guarner. Progress in the knowledge of the intestinal human microbiota. *Nutrición Hospitalaria*, 28(3):553–557, 2013.
- [64] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(C):53–65, nov 1987.
- [65] David Ruppert. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 99. Springer Science & Business Media, jun 2004.
- [66] Warren S. Sarle, Anil K. Jain, and Richard C. Dubes. *Algorithms for Clustering Data*, volume 32. Prentice-Hall, Inc., USA, may 1990.
- [67] V. Shankar, M. Gouda, J. Moncivaiz, A. Gordon, N. V. Reo, L. Hussein, and O. Paliy. Differences in Gut Metabolites and Microbial Composition and Functions between Egyptian and U.S. Children Are Consistent with Their Diets. *mSystems*, 2(1), 2017.
- [68] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, jul 1948.
- [69] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [70] Nick Shortt, Hazel Poyntz, Wayne Young, Angela Jones, Aurélie Gestin, Anna Mooney, Darmiga Thayabaran, Jenny Sparks, Tess Ostapowicz, Audrey Tay, Sally Poppitt, Sarah Elliott, Georgia Wakefield, Amber Parry-Strong, Jacqui Ralston, Olivier Gasser, Richard Beasley, Mark Weatherall, Irene Braithwaite, and Elizabeth Forbes-Blom. A feasibility study: association between gut microbiota enterotype and antibody response to seasonal trivalent influenza vaccine in adults. *Clinical and Translational Immunology*, 7(3):e1013, 2018.
- [71] Loek P. Smits, Kristien E.C. Bouter, Willem M. De Vos, Thomas J. Borody, and Max Nieuwdorp. Therapeutic potential of fecal microbiota transplantation. *Gastroenterology*, 145(5):946–953, 2013.
- [72] David L. Suskind, Mitchell J. Brittnacher, Ghassan Wahbeh, Michele L. Shaffer, Hillary S. Hayden, Xuan Qin, Namita Singh, Christopher J. Damman, Kyle R. Hager, Heather Nielson, and Samuel I. Miller. Fecal microbial transplant effect on clinical outcomes and fecal microbiome in active Crohn’s disease. *Inflammatory Bowel Diseases*, 21(3):556–563, 2015.
- [73] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction, 2000.
- [74] Andrew M. Thomas, Eliane C. Jesus, Ademar Lopes, Samuel Aguiar, Maria D. Begnami, Rafael M. Rocha, Paola Avelar Carpinetti, Anamaria A. Camargo, Christian Hoffmann, Helano C. Freitas, Israel T. Silva, Diana N. Nunes, Jo simao C. Setubal, and Emmanuel Dias-Neto. Tissue-Associated Bacterial Alterations in Rectal Carcinoma Patients Revealed by 16S rRNA Community Profiling. *Frontiers in Cellular and Infection Microbiology*, 6(DEC):179, dec 2016.
- [75] Robert Tibshirani and Guenther Walther. Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, sep 2005.
- [76] E. F. Tigchelaar, M. J. Bonder, S. A. Jankipersadsing, J. Fu, C. Wijmenga, and A. Zhernakova. Gut microbiota composition associated with stool consistency. *Gut*, 65(3):540–542, mar 2016.
- [77] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2625, 2008.

- [78] Doris Vandeputte, Gwen Falony, Sara Vieira-Silva, Raul Y. Tito, Marie Joossens, and Jeroen Raes. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut*, 65(1):57–62, jan 2016.
- [79] Doris Vandeputte, Gunter Kathagen, Kevin D’Hoe, Sara Vieira-Silva, Mireia Valles-Colomer, Joa simo Sabino, Jun Wang, Raul Y. Tito, Lindsey De Commer, Youssef Darzi, Séverine Vermeire, Gwen Falony, and Jeroen Raes. Quantitative microbiome profiling links gut community variation to microbial load. *Nature*, 551(7681):507–511, 2017.
- [80] Sara Vieira-Silva, Gwen Falony, Youssef Darzi, Gipsi Lima-Mendez, Roberto Garcia Yunta, Shujiro Okuda, Doris Vandeputte, Mireia Valles-Colomer, Falk Hildebrand, Samuel Chaffron, and Jeroen Raes. Species-function relationships shape ecological properties of the human gut microbiome. *Nature Microbiology*, 1(8):16088, 2016.
- [81] Sara Vieira-Silva, João Sabino, Mireia Valles-Colomer, Gwen Falony, Gunter Kathagen, Clara Caenepeel, Isabelle Cleynen, Schalk van der Merwe, Séverine Vermeire, and Jeroen Raes. Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nature Microbiology*, 4(11):1826–1831, nov 2019.
- [82] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
- [83] Anne Vrieze, Els Van Nood, Frits Holleman, Jarkko Salojärvi, Ruud S. Kootte, Joep F.W.M. Bartelsman, Geesje M. Dallinga-Thie, Mariette T. Ackermans, Mireille J. Serlie, Raish Oozeer, Muriel Derrien, Anne Druesne, Johan E.T. Van Hylckama Vlieg, Vincent W. Bloks, Albert K. Groen, Hans G.H.J. Heilig, Erwin G. Zoetendal, Erik S. Stroes, Willem M. de Vos, Joost B.L. Hoekstra, and Max Nieuwdorp. Transfer of Intestinal Microbiota From Lean Donors Increases Insulin Sensitivity in Individuals With Metabolic Syndrome. *Gastroenterology*, 143(4):913–916.e7, oct 2012.
- [84] Jun Wang, Miriam Linnenbrink, Sven Künzel, Ricardo Fernandes, Marie-Josée Nadeau, Philip Rosenstiel, and John F. Baines. Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice. *Proceedings of the National Academy of Sciences*, 111(26):E2703–E2710, jul 2014.
- [85] Gary D. Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying Yu Chen, Sue A. Keilbaugh, Meenakshi Bewtra, Dan Knights, William A. Walters, Rob Knight, Rohini Sinha, Erin Gilroy, Kernika Gupta, Robert Baldassano, Lisa Nessel, Hongzhe Li, Frederic D. Bushman, and James D. Lewis. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.
- [86] Qinqin Wu, Xiong’e Pi, Wei Liu, Huahai Chen, Yeshe Yin, Hongwei D. Yu, Xin Wang, and Liying Zhu. Fermentation properties of isomaltooligosaccharides are affected by human fecal enterotypes. *Anaerobe*, 48:206–214, dec 2017.
- [87] Fan Xia, Jun Chen, Wing Kam Fung, and Hongzhe Li. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063, 2013.
- [88] Tanya Yatsunencko, Federico E. Rey, Mark J. Marnary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N. Baldassano, Andrey P. Anokhin, Andrew C. Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J. Gregory Caporaso, Catherine A. Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight, and Jeffrey I. Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, 2012.
- [89] Yeshe Yin, Bin Fan, Wei Liu, Rongrong Ren, Huahai Chen, Shaofeng Bai, Liying Zhu, Gang Sun, Yunsheng Yang, and Xin Wang. Investigation into the stability and culturability of Chinese enterotypes. *Scientific Reports*, 7(1):7947, 2017.
- [90] Zhenyue Zhang and Jing Wang. MLLE: Modified locally linear embedding using multiple weights. In *Advances in Neural Information Processing Systems*, volume 19, pages 1593–1600, 2007.