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

Ivan Bulygin[1], Vladislav Shatov[2], Anton Rykachevskiy[1], Arsenii Raiko[1], Alexander Bernstein[1], Evgeny Burnaev[1,3] and Mikhail S. Gelfand[1,4]

[1] Skolkovo Institute of Science and Technology, Moscow, Russia
[2] Moscow State University, Moscow, Russia
[3] Artificial Intelligence Research Institute (AIRI), Moscow, Russia
[4] Institute for Information Transmission Problems, 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 of variable 16S rRNA regions, 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.

Subjects Bioinformatics, Genomics, Microbiology, Gastroenterology and Hepatology, Data Science
Keywords Human gut microbiome, Dimensionality reduction, Clustering, Enterotypes

Submitted 2 November 2022
Accepted 12 July 2023
Published 8 September 2023

Corresponding author
Ivan Bulygin, bulygin@phystech.edu

Academic editor
Alexander Bolshoy

Additional Information and Declarations can be found on page 20

DOI **10.7717/peerj.15838**

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 (*Gilbert et al., 2018*). 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. The initial sequencing of the gut biota revealed that its composition tends to form discrete groups (enterotypes) consisting predominantly of taxa *Bacteroides*, *Prevotella*, and *Ruminococcus* (*Arumugam et al., 2011*). Enterotypes have been reported in *Arumugam et al. (2011)* as "densely populated areas in a multidimensional space of community composition" which "are not as sharply delimited as, for example, human blood groups".

In *Costea et al. (2017)* this concept was revisited, suggesting a more careful definition reflecting non-discrete structure and non-uniform density of the microbial composition.

How to cite this article Bulygin I, Shatov V, Rykachevskiy A, Raiko A, Bernstein A, Burnaev E, Gelfand MS. 2023. Absence of enterotypes in the human gut microbiomes reanalyzed with non-linear dimensionality reduction methods. *PeerJ* **11**:e15838
http://doi.org/10.7717/peerj.15838

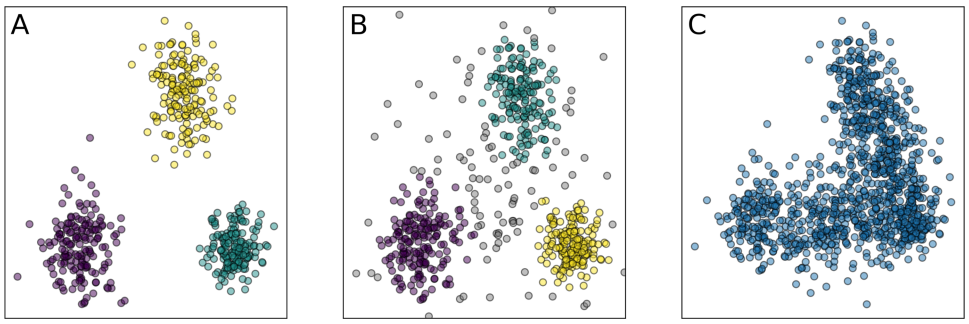

Figure 1 Examples of the clustering partitions. (A) Clusters as well-separated sets of points; (B) clusters as regions of points with a higher density than the background; (C) absence of clusters, but presence of the low-dimensional structure. Similar distinction also applies for non-convex clusters. Colors indicate different clusters.

Full-size DOI: 10.7717/peerj.15838/fig-1

Still, many recent papers consider enterotypes as discrete clusters in the relative taxonomic abundance, and here we attempt to follow this approach, but in a more rigorous way. From a geometric point of view, clusters are defined as dense areas separated by sparse regions. 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 a metric, *e.g.*, the Euclidean distance (*Maronna, Charu & Chandan, 2015*). Enterotypes may be defined either as well-separated clusters without points between them or as regions with a higher density of data points, indicating preferential clustering, as shown in Figs. 1A and 1B, respectively. We address both cases using diverse methods and metrics, sensitive for either the first or both types of clusters.

A major current challenge is to determine the existence of enterotypes *via* "a thorough quantitative investigation of established clustering methods and tests for microbiome data" (*Knights et al., 2014*). Published studies rely on similar approaches yet differ in the exact number of enterotypes ranging from two (*Li et al., 2018*; *Yin et al., 2017*; *Chen et al., 2017*; *Nakayama et al., 2017*; *Kang et al., 2016*; *Nakayama et al., 2015*; *Wang et al., 2014*; *Ou et al., 2013*; *Wu et al., 2011*) to three (*Wu et al., 2017*; *De Moraes et al., 2017*; *Vieira-Silva et al., 2016*; *Emoto et al., 2016*; *Robles-Alonso & Guarner, 2013*) and four (*Gotoda, 2015*). 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 to the first two or three principal components from the principal components analysis (PCA). This approach is general and may involve various intermediate steps, different clustering algorithms, and a variety of metrics; however, it has certain flaws. For example, the 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 (*Calinski & Harabasz, 1974*) and the Silhouette score (*Rousseeuw, 1987*) 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 (*Aggarwal, Hinneburg & Keim, 2001*). For example, the distance to the nearest data point approaches the distance to the farthest data point (*Beyer et al., 1999*). The lack of data further amplifies this problem since the data point cloud in a high-dimensional space becomes sparse, yielding unreliable estimates of the probability density due to the non-asymptotic lower bound for the regression error (*Kohler, Krzyzak & Walk, 2009*; *Ibragimov & Has'minskii, 1981*). 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 the aggravating factors described above undermine the very notion of enterotypes. Several studies demonstrated the possibility of a gradient distribution and the absence of well-defined clusters (*Jeffery et al., 2012*; *Yatsunenko et al., 2012*; *Claesson et al., 2012*; *Cheng & Ning, 2019*; *Costea et al., 2017*). Different structures of enterotypes between males and females were claimed by *Oki et al. (2016)*. In several works, the very concept of enterotypes was described as inconsistent and uncertain (*Knights et al., 2014*; *Jeffery et al., 2012*). Factors such as the variation in the microbial load between samples (*Vandeputte et al., 2017*), robustness of enterotypes clusters (*Huss, 2014*), and microbiome variation during short periods of time (*Knights et al., 2014*) were considered limiting to the use of the enterotype concept. However, enterotyping of the human gut has been applied in clinical research. Published studies claimed correlations between enterotypes and diet (*Chen et al., 2017*; *Nakayama et al., 2017*; *Kang et al., 2016*; *Nakayama et al., 2015*; *Wu et al., 2011*; *Klimenko et al., 2018*; *Shankar et al., 2017*; *Liang et al., 2017*; *Xia et al., 2013*), inflammatory gut diseases (*Vieira-Silva et al., 2019*; *Castaño Rodríguez et al., 2018*; *Moser, Fournier & Peter, 2017*; *Dlugosz et al., 2015*; *Chiu et al., 2014*; *Huang et al., 2014*; *De Wouters, Doré & Lepage, 2012*), mental health (*Lee et al., 2020*), acne (*Deng et al., 2018*), stool composition (*Tigchelaar et al., 2015*; *Vandeputte et al., 2015*), colorectal cancer (*Thomas et al., 2016*), circulatory diseases (*Emoto et al., 2016*; *Jie et al., 2017*; *Franco-de Moraes et al., 2017*), psoriasis (*Codoñer et al., 2018*), and infections such as AIDS (*Noguera-Julian et al., 2016*) and influenza (*Qin et al., 2015*; *Shortt et al., 2018*). The idea that 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 (*Gilbert et al., 2018*) 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 (*Cheng & Ning, 2019*).

Here, we look for balanced, stable, and distinct clusters in large stool microbiome datasets. Balance implies that each cluster should contain a sufficient number of points to

be considered as a potential enterotype. Stability means that clusters should not depend on 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 separating gaps seen as space regions with a lower concentration of points, whereas clusters correspond to areas of highest concentration. All these requirements are addressed by appropriate metrics and methods, described in 'Materials & Methods'. To ensure accurate analysis of the high-dimensional data, 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 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-genotyped stool samples. This absence of natural clusters is seen in the example in Fig. 1C.

MATERIALS & 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) (*The Human Microbiome Project Consortium, 2012a*; *The Human Microbiome Project Consortium, 2012b*) and American Gut Project (AGP) (*McDonald et al., 2018*). These largest available datasets provide a sufficient number of data points for correct estimation of the clustering partition and constructing a manifold (*Psutka & Psutka, 2015*). Both datasets were collected in the United States. No patterns driven by geography or lifestyle were explicitly taken into account in our framework. We did not use longitudinal sampled microbiota datasets, as we were not concentrating on the dynamics of enterotypes, but rather on their existence. Similarly, we did not use shotgun sequencing data (*Pasolli et al., 2017*), as its characterized taxonomy composition is limited to sequenced genomes. We used 4,587 HMP samples from stool and rectum body sites downloaded from the Human Microbiome Project (https://portal.hmpdacc.org) and 9,511 samples from AGP downloaded from figshare (https://figshare.com/articles/dataset/American_Gut_Project_fecal_sOTU_relative_abundance_table/6137198) as abundance matrices. For comparison with the original research (*Arumugam et al., 2011*), we analyzed Sanger (*Gill et al., 2006*), Illumina (*Qin et al., 2010*), and Pyroseq (*Turnbaugh et al., 2008*) datasets from (http://www.bork.embl.de/Docu/Arumugam_et_al_2011/). The results are presented in Text S3, Table S2, Figs. S2–S5. All datasets were normalized by dividing Operational Taxonomic Units (OTUs) values by the total sum of abundances for a given data sample. An OTU found in less than 1% of the samples or with a standard deviation less than 0.001 were removed to ease the preprocessing step. To account for outliers with microbiome dominated by single or few species, *e.g.*, in patients with extreme gut microbiota we repeated the analysis, for the HMP and AGP datasets with removed OTUs accounting for >70% abundance. The results were consistent, see Text S5, Tables S3–S4, Figs. S8–S15.

As the first step of dimensionality reduction after preprocessing, we use PCA to identify a medium-dimensional linear subspace retaining almost all data cloud variation. The dataset projected on this subspace does not significantly differ from the original dataset, while removing dimensions with low variance acts as a filter that provides a more robust clustering (*Ben-Hur & Guyon, 2003*). We preserve the variances after projection, since removing them may hinder the subsequent clustering process and lead to erroneous results. Instead of limiting the dimensionality reduction process solely to PCA, as in previous studies, we then determine the intrinsic dimension of the projected data *via* the maximum likelihood estimation (MLE) (*Levina & Bickel, 2004*). This step allows for capturing a minimal but sufficient number of coordinates representing the most important features of the dataset. Following the manifold hypothesis (*Fefferman, Mitter & Narayanan, 2016*), we suppose that a 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. At that non-linear projections per se are not interesting, since any data cloud could perfectly lie on a one-dimensional submanifold.

While this one-dimensional submanifold yields a 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 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 data dimensionality using several manifold learning algorithms, namely: Isomap (*Tenenbaum, De Silva & Langford, 2000*), locally linear embedding (LLE) (*Zhang & Wang, 2007*), denoising autoencoder (AE) (*Goodfellow, Bengio & Courville, 2016*), spectral embedding (SE) (*Shi & Malik, 2000*), t-distributed Stochastic Neighbor Embedding (t-SNE) (*Van Der Maaten & Hinton, 2008*), and uniform manifold approximation and projection (UMAP) (*McInnes et al., 2018*). A detailed description of these algorithms and their pros and cons is beyond the scope of this article. These methods are conceptually different and susceptible mostly for quantitative comparison, rather than qualitative. A short description is provided in Table S5. For each manifold learning algorithm, dataset, and taxonomic level, we obtain a low-dimensional embedding. To find the near-optimal hyperparameters of the manifold learning algorithm for a specific dataset and taxonomic level, we iterate over various combinations of hyperparameters. For each combination, we assess how well an embedding produced by an algorithm with this combination of parameters, represents the original data.

We selected a computationally feasible hyperparameters range, with reasonable values, according to our expectation of the number of clusters and common machine learning practice. To ensure reproducibility of the results, we restricted the number of hyperparameter combinations from eight to 40, depending on the algorithm. For details

on the training procedures and hyperparameters choice, refer to the software link in the 'Data Availability' section.

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 k-nearest neighbors regression (*Fix & Hodges, 1989*) with five nearest neighbors and distance weighting. Then, we estimate the reconstruction error using the Leave-One-Out procedure (*Ruppert, 2004*). We report the resulting error as median of the absolute error (MAE) across all reconstructed points. This technique assesses how well points coordinates 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 alone is an intractable metric that does not show what aspect of the data cloud has been misrepresented. Therefore, we apply two additional criteria of quality of dimensionality reduction (*Lee & Verleysen, 2010*). These criteria, Q_{loc} and Q_{glob}, represent the preservation of the "local" and the "global" structure and are described in Text S1. 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. Given hyperparameters that deliver the lowest MAE, we iteratively discard up to 10% of initial 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 are outliers as determined by the Local Outlier Factor algorithm (*Breunig et al., 2000*). The latter allows for detecting outliers by deviation of their local density with respect to their neighbors, from which it follows that they are not related to any cluster as the latter are densely populated areas in a multidimensional space.

For each low-dimensional embedding, we apply several clustering methods-Spectral Clustering (*Shi & Malik, 2000*; *Von Luxburg, 2007*), PAM, and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (*McInnes & Healy, 2017*). A detailed description of these algorithms and their pros and cons is beyond the scope of this article, while a short description is provided in Table S6. 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 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. As a result, we obtain more robust results not biased by the peculiarities of the algorithms and the choice of hyperparameters.

Clustering metrics based on the ratio of within-cluster compactness to between-cluster separation, like the Calinski–Harabasz index (*Calinski & Harabasz, 1974*), the Silhouette score (*Rousseeuw, 1987*), and the Davies–Bouldin index (*Davies & Bouldin, 1979*) cannot handle arbitrarily shaped clusters and noise in the 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) (*Moulavi et al., 2014*), which accounts for both density and shape properties of clusters, tolerates noise, and is appropriate

for detecting density-based clusters. We assess the clustering partition stability using prediction strength (*Tibshirani & Walther, 2005*) initially proposed for estimating the number of clusters, which tells us how well the decision boundaries of the clustering partition, 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 separated from the main data point cloud. Nevertheless, they significantly impact the clustering quality metrics. To detect such imbalanced partitions, we use Shannon Entropy (*Shannon, 1948*) of the probability distribution of data points to be in a certain cluster. All these metrics are described in detail in Text S1.

To identify the optimal clustering, we compare DBCV, prediction strength, and entropy for each partition respective to different clustering hyperparameters. A balanced clustering partition that corresponds to 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, following previous studies, for each clustering partition, we calculate the Davies–Bouldin Index and the Silhouette score. Lower Davies–Bouldin Index and higher Silhouette score correspond to better partitions, where clusters are better in terms of compactness and separation. We summarize all steps described above in a single framework schematically shown in Fig. 2.

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) (*Van Der Maaten & Hinton, 2008*) and UMAP algorithms (*McInnes et al., 2018*). For most dimensionality reduction methods, validity indices, and metrics, implementations from the 'scikit-learn' library (*Pedregosa et al., 2011*) were used. For the HDBSCAN clustering method and the DBCV metric the 'hdbscan' package (*McInnes & Healy, 2017*) from the 'scikit-learn-contrib' was used. For the UMAP algorithm we applied the implementation from the 'umap-learn' library (*McInnes et al., 2018*).

RESULTS

Data preprocessing

The numbers of objects for both datasets and their dimensionality d in the relative taxon abundance space, before and after preprocessing, are presented in Table 1. The datasets were analyzed at the Order, Family, and Genus taxonomic levels (denoted 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 exotic microbial communities or be caused by sample collection or data processing artifacts. Moreover, the procedures may vary between laboratories, leading to considerable batch effects; for instance, dataset-specific preprocessing has been applied to correct for microbial blooms in the AGP dataset (*Amir et al., 2017*). Therefore, we

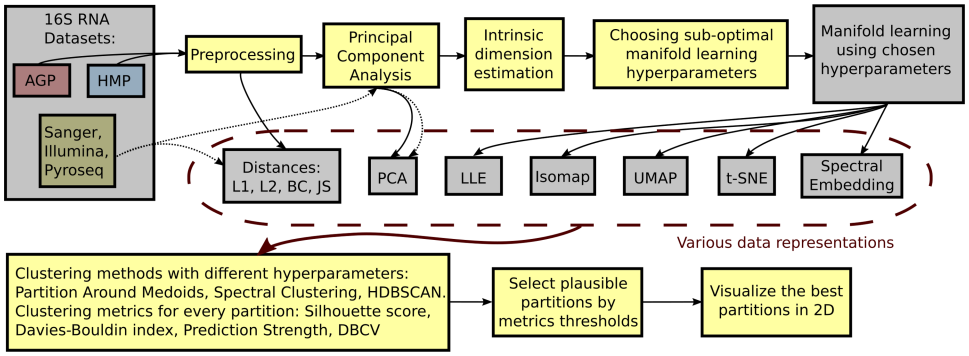

Figure 2 Schematic depiction of our framework for clustering high-dimensional data. This procedure is applied to each dataset (AGP, HMP) at every taxonomy level (O, Order; F, Family; G, Genus). The data is preprocessed by removing OTUs that are found in less than 1% of the samples or have a standard deviation less than 0.001. For the preprocessed data, pairwise distances are calculated: Manhattan (L1), Euclidean (L2), Bray-Curtis (BC), and Jensen–Shannon (JS). PCA representation is obtained by projecting the data onto principal components explaining 99% variance. After such projection, the intrinsic dimension of the data is estimated. The intrinsic dimension for every dataset at different taxonomy levels is given, and suboptimal hyperparameters for every manifold learning algorithm are found by minimizing the reconstruction median absolute error over different hyperparameter combinations. Then, the collection of data representations is extended by adding the results of nonlinear dimensionality reduction methods obtained using the found suboptimal hyperparameters. For every data representation in the collection, various clustering algorithms with different hyperparameters are applied. Partition Around Medoid (PAM), spectral clustering, and HDBSCAN. Then, for every found partition, clustering metrics are assessed. As a result, only partitions that pass certain metrics thresholds are considered plausible.
Full-size DOI: 10.7717/peerj.15838/fig-2

Table 1 Size and dimensionality d of the AGP and HMP datasets in Order, Family, and Genus taxonomic levels. Initial (init.) dimensionality corresponds to raw data and processed (proc.) corresponds to data, after removing OTU found in less than 1% of the samples or with a standard deviation less than 0.001.

Dataset	Size	Order *d*		Family *d*		Genus *d*	
		init.	proc.	init.	proc.	init.	proc.
AGP	9511	168	39	258	69	535	108
HMP	4587	179	39	267	70	574	97

cannot merge several individual datasets into one. Indeed, visualization using t-SNE (*Van Der Maaten & Hinton, 2008*) and UMAP (*McInnes et al., 2018*) dimensionality reduction algorithms illustrates that point in Fig. 3. Hence, to avoid the batch effect, we considered these datasets separately.

Principal Component Analysis (PCA)

We obtained significant dimensionality reduction with minuscule information loss by using projection on relatively many (16 through 47, dependent on the taxonomy level) principal components. Dimensionalities d_{PCA} of the PCA projections are reported in Table 2 and defined as the number of first principal components that meet the selected threshold of 99% of the cumulative explained variance. Contribution of original taxonomic coordinates to the principal components, also known as the PCA loadings, can be calculated as the Euclidean norm of the corresponding principal vectors coordinates multiplied by the square

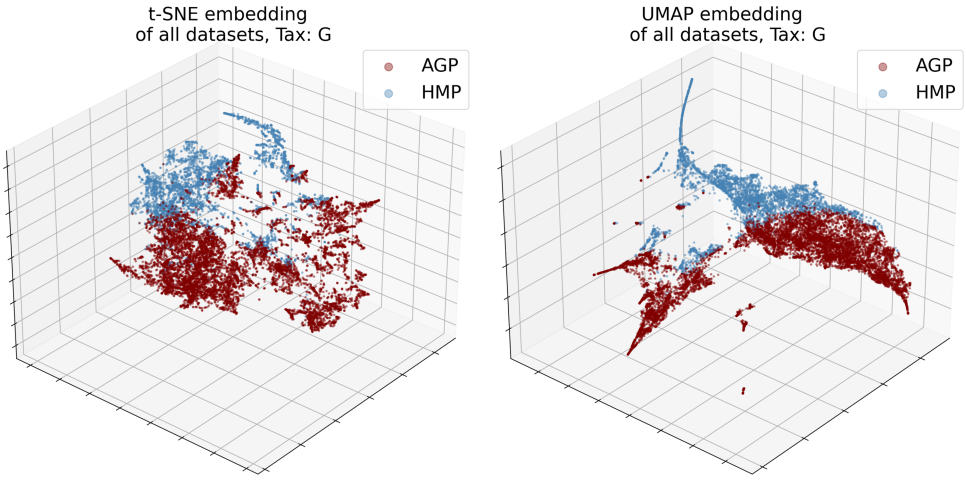

Figure 3 3D t-SNE and UMAP visualizations of joined datasets demonstrate batch-effect at Genus taxonomy level (tax). Red—AGP, green—HMP dataset.

Full-size DOI: 10.7717/peerj.15838/fig-3

Table 2 Dimensionalities d_{PCA} and d_{MLE} of two datasets (AGP and HMP) in three taxonomy levels (O, Order; F, Family; G, Genus).

Dataset	Tax O		Tax F		Tax G	
	d_{PCA}	d_{MLE}	d_{PCA}	d_{MLE}	d_{PCA}	d_{MLE}
AGP	16	6	34	8	47	8
HMP	18	5	35	6	40	6

Notes.

d_{PCA}, number of the first principal components explaining 99% variance; d_{MLE}, estimated intrinsic dimension..

root of the associated eigenvalue. The cumulative explained variance and PCA loadings are presented in Fig. 4. Evidently, for both datasets at the Genus level *Bacteroides* and *Prevotella* contribute the most to the variance and the resulting PCA components. At the Family level it is again *Bacteroidaceae* and *Prevotellaceae* for both datasets, but with *Ruminococcaceae* for AGP and *Enterobacteriaceae* for HMP as additional strong drivers of the variance. At the Order level, the variance is dominated by *Bacteroidales, Enterobacteriales*, and *Clostridiales* for both datasets. To assess the information loss during projection, we calculated the error of reconstruction from the projected data to the original one. The Median Absolute Error (MAE) and Q_{loc} and Q_{glob} metrics (for details see Text S1) are presented in Table 3. Reconstruction of the original data from the data projected on principal components was obtained using an inverse linear transformation.

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 (*Levina & Bickel, 2004*). As a dimension estimation we have selected the median of the intrinsic dimension distribution across the neighborhood cardinality, which varies from 5 to 100. The distribution is estimated using bootstrapping technique for 50 trials.

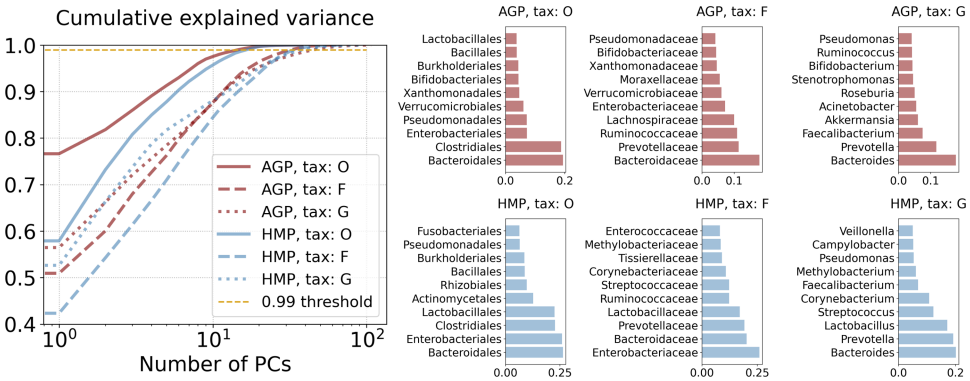

Figure 4 **Principal components analysis.** Cumulative explained variance and PCA loadings, representing contribution of the original taxonomy coordinates to the principal components.
Full-size DOI: 10.7717/peerj.15838/fig-4

Table 3 **PCA dimensionality reduction metrics.** Median absolute error (MAE) of the linear inverse transformation from the data projected on the principal components to the original space of taxon abundances. Q_{loc} and Q_{glob} metrics denote preservation of the local and global data structure (see the text for details). Notation as in Table 2.

Dataset	Tax	MAE	Q_{loc}	Q_{glob}
AGP	O	0.061	0.90	0.99
	F	0.041	0.96	0.99
	G	0.050	0.95	0.99
HMP	O	0.036	0.91	0.99
	F	0.044	0.92	0.98
	G	0.058	0.92	0.99

The resulting intrinsic dimensions and the dimensionality after projection on principal components are presented in Table 2.

Manifold learning

Subsequent non-linear dimensionality reduction from d_{PCA} to d_{MLE}, is performed by different manifold learning algorithms described in the Materials & Methods section and Table S5. In Fig. 5, we present the Q_{loc}, and Q_{glob} metrics that represent the preservation of the local and global data structure after dimensionality reduction. They were evaluated for all manifold learning methods with near-optimal hyperparameters, applied to all datasets at different taxonomic levels. Near-optimal hyperparameters were found using manifold learning algorithms with different combinations of potential hyperparameters and selecting the ones with the lowest reconstruction MAE. The corresponding MAE values of the original data reconstruction from a nonlinear embedding are listed in Table 4. It shows that higher taxonomy levels yield more intricate data representation with higher MAE and lower Q_{loc}, Q_{glob}. This is due to data becoming sparser and more dissipated in high-dimensional space, which hinders data representation by fitting a non-linear manifold.

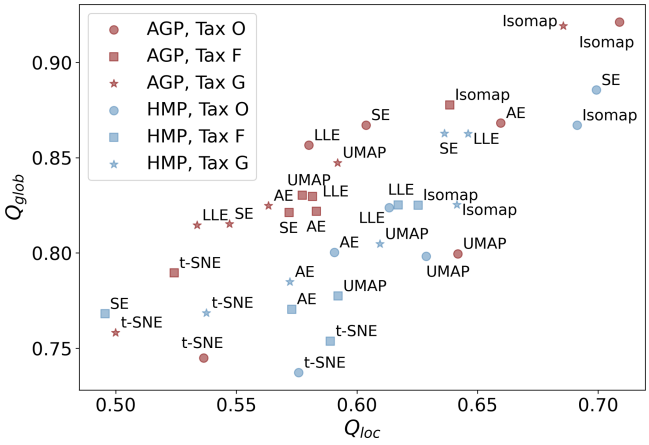

Figure 5 Metrics of the data structure preservation for different dimensionality reduction methods, taxonomy levels and datasets. Horizontal axis-Q_{loc} metric (local information preservation), vertical axis-Q_{glob} metric (global information preservation) of the non-linear dimensionality reduction 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.

Full-size DOI: 10.7717/peerj.15838/fig-5

Table 4 Median Absolute Error (MAE) of the data reconstruction from different manifold learning embeddings. MAE is assessed using Leave-One-Out procedure. The reconstruction is done by the independent k-nearest neighbors regression of the coordinates in the original space of relative taxon abundances from the non-linear embedding. Notation as in Table 2.

Dataset	Method	Tax O	Tax F	Tax G
AGP	AutoEncoder	0.06	0.19	0.22
	t-SNE	0.05	0.20	0.22
	UMAP	0.06	0.22	0.24
	Isomap	0.06	0.22	0.25
	LLE	0.06	0.21	0.24
	Spectral	0.06	0.21	0.24
HMP	AutoEncoder	0.09	0.24	0.22
	t-SNE	0.09	0.24	0.22
	UMAP	0.11	0.27	0.25
	Isomap	0.11	0.29	0.27
	LLE	0.12	0.29	0.26
	Spectral	0.13	0.34	0.29

Clustering

We applied several clustering methods using the Euclidean metric - Spectral Clustering (*Von Luxburg, 2007*), PAM and HDBSCAN (*McInnes & Healy, 2017*)-for each embedding provided by the dimensionality reduction algorithms. Following related works on the identification of enterotypes, we also applied clustering to the original data in high-dimensional space of taxonomic abundances with a variety of distance metrics: Jensen–Shannon, Manhattan, Euclidean, and Bray-Curtis as in *Koren et al. (2013)*. It should be noted that we used only the Euclidean and Manhattan metric in the manifold learning

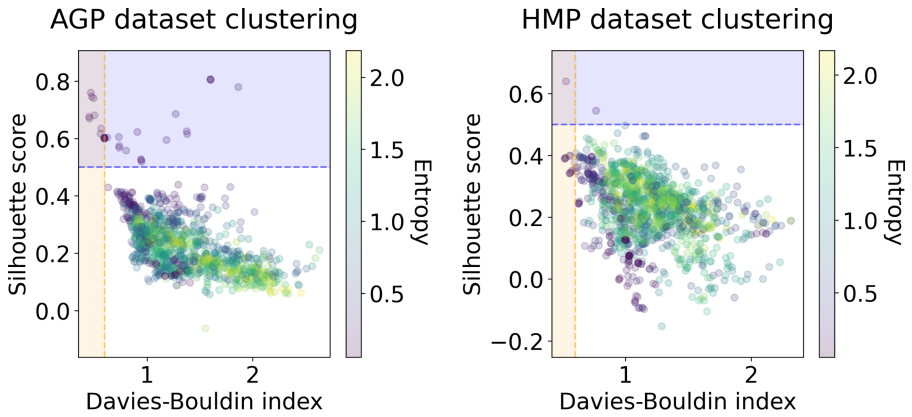

Figure 6 Silhouette score and Davies–Bouldin Index of the clustering results for the AGP and HMP datasets. All three taxonomy levels (O, Order; F, Family, G, Genus) are displayed. The point color represents the entropy of the respective partition.

Full-size DOI: 10.7717/peerj.15838/fig-6

algorithms since distribution-based distances such as Jensen–Shannon are not applicable after the PCA projection. Originally, *Bacteroides, Prevotella*, and *Ruminococcus* have been considered as the main drivers of microbial variation that contribute most to the enterotypes (*Arumugam et al., 2011*). To ensure that there is no evident clustering structure within these genera, we visualize the three-dimensional distribution of the normalized abundances of these OTUs in Fig. S6 (top). We observe a continuous distribution with no apparent natural clusters. We support this observation, by adding this three-dimensional projection of the datasets into the pool of all representations to which we apply clustering methods. Such representations are comprised by manifold-learning embeddings, original data in different metric spaces, and data projected on principal components. The resulting metrics of all clustering results in different data representations are shown in Figs. 6 and 7. 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 *Tibshirani & Walther (2005)* and *Koren et al. (2013)*. 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 *Wu et al. (2011)*; *Koren et al. (2013)*; *Gentle, Kaufman & Rousseuw (1991)*; *Arbelaitz et al. (2013)*. As a threshold value for the Davies–Bouldin index, we used 0.6 for moderate clustering (*Davies & Bouldin, 1979*). 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 high dimensional data, we also provide results on simple synthetic datasets (for details see Text S2). Since the clustering results directly depend on the algorithm hyperparameters, for each clustering method we have iterated over combinations of relevant sets of hyperparameter values. For the spectral clustering algorithm, we considered the range from two to nine as a possible number of clusters in the data, and 5, 15, 25, 50 as sizes of the neighborhood for computing

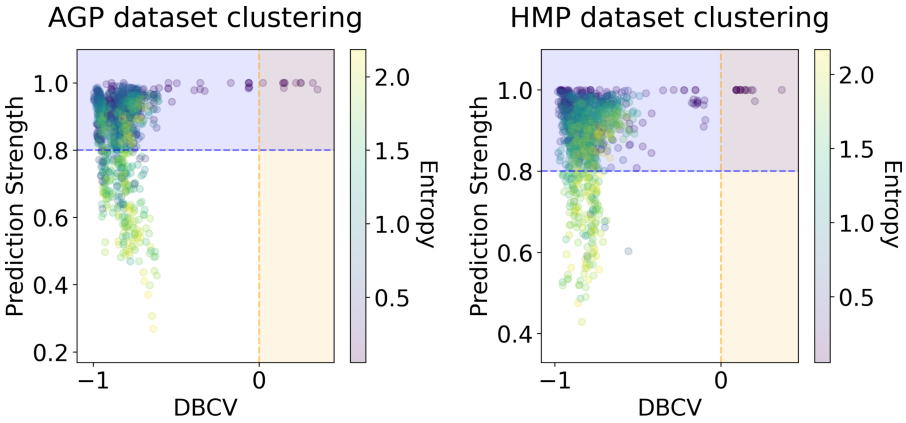

Figure 7 DBCV index and the prediction strength of the clustering results for the AGP and HMP datasets. All three taxonomy levels (O, Order; F, Family; G, Genus) are displayed. The point color represents the entropy of the respective partition.

Full-size DOI: 10.7717/peerj.15838/fig-7

the affinity matrix. For the precomputed distance matrices of the original data, we use values 1, 5, 10, 15 for the "gamma" parameter in construction of the affinity matrix using a radial basis function. For the PAM algorithm, we used the same range of possible clusters. For the HDBSCAN hyperparameters, we considered 5, 10, 25, 50 as the minimal cluster size and 5, 10, 15, 20 as the minimal number of samples in a neighborhood for a point to be considered a core point. We did not set larger 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.

In Figs. 6 and 7, we show the distribution of metrics over all clustering results. They are comprised by partitions calculated for each dataset, taxonomic level, manifold learning method, and clustering algorithm with different hyperparameters. Clustering partitions with moderate or strong support for both metrics correspond to points lying at the intersection of the blue and orange areas. All partitions respective to points in this area were found to consist of two or three highly imbalanced clusters with more than 95% of the data points concentrated in one cluster. These partitions are also inconsistent in terms of clustering validity metrics when either DBCV and Prediction Strength, or Davies–Bouldin and Silhouette pairs pass the thresholds but not all of them. Small clusters that constitute less than 5% of the data were found to be unstable among different partitions of the same dataset, demonstrating high dependence on the manifold learning and clustering algorithms. As expected, different hyperparameters combinations of the clustering algorithm can lead to the same partitions for fixed dataset, taxonomic level and embedding type. Among them, given that they satisfy the clustering metrics, we selected the one with the highest entropy, which should match the most balanced partition.

In Table 5 we present these partitions with the respective metrics. Only few partitions into two and three clusters were found at different taxonomic levels, that meet the threshold criteria of at least one of the metrics pairs: the Davies–Bouldin and Silhouette score or the DBCV and Prediction Strength. All partitions are stable, according to Prediction Strength.

Table 5 Selected clustering results with high or moderate clustering partition metrics. Clustering validity metrics and the number of clusters k for different partitions obtained from different data representations. Repr. denotes a data representation used for clustering. It is either an embedding provided by a manifold learning algorithm (SE - Spectral Embedding, LLE - Locally Linear Embedding) or pairwise distances inferred from the data (L1 - Manhattan distance in the original space of taxonomic abundances). Spectral, Spectral Clustering algorithm. D-B index, Davies-Bouldin index. Silh. score, Silhouette score. DBCV, Density-Based Clustering Validation index. Ent., Entropy. Notation as in Table 2.

	Tax	Repr.	Cluster method	k	D-B index	Silh. score	DBCV	Prediction Strength	Ent.
AGP	**O**	L1	Spectral	2	0.60	0.60	−0.63	0.98	0.06
	O	LLE	Spectral	2	0.49	0.74	−0.86	0.94	0.06
	O	LLE	Spectral	3	0.60	0.60	−0.91	0.91	0.18
	O	SE	Spectral	2	0.50	0.68	−0.91	0.96	0.09
	O	SE	Spectral	3	0.57	0.63	−0.92	0.94	0.19
HMP	**F**	t-SNE	HBDSCAN	2	1.38	0.14	0.15	1.00	0.09
	F	UMAP	HBDSCAN	2	1.02	0.17	0.22	1.00	0.06
	G	UMAP	HBDSCAN	2	1.03	0.23	0.25	1.00	0.08
	O	t-SNE	HBDSCAN	2	1.00	0.13	0.12	1.00	0.06
	O	UMAP	HBDSCAN	2	0.87	0.15	0.19	1.00	0.08
	O	UMAP	HBDSCAN	3	1.02	0.06	0.19	1.00	0.16
	F	UMAP	HBDSCAN	2	1.03	0.08	0.10	1.00	0.08
	F	SE	HBDSCAN	2	0.53	0.64	−0.63	1.00	0.09
	F	t-SNE	HBDSCAN	2	1.11	0.09	0.21	0.97	0.09
	G	UMAP	HBDSCAN	2	1.24	−0.02	0.16	1.00	0.06

Yet, plausible partitions with moderate Silhouette scores varying from 0.60 to 0.74 exhibit low DBCV index from −0.92 to −0.63. Similarly, partitions with moderate DBCV from 0.10 to 0.25 yield lower Silhouette score in the range from −0.02 to 0.23 and higher Davies–Bouldin index varying from 0.87 to 1.38. For all presented clustering partitions, the entropy of the data mass distribution over clusters is low, indicating a highly imbalanced partition. The maximal entropy over two found clusters among partitions is 0.09, indicative of a clustering where 98% of data are concentrated in one cluster. While the entropy of 0.19 for partition into three clusters is higher than for two clusters, it is still imbalanced, with 96% of the data concentrated in the one cluster. Therefore, there are no partitions found, that would satisfy all clustering metrics criteria at the same time.

Among all clustering partitions passing the metrics thresholds in Figs. 6 and 7, we visualize the ones with the highest entropy. The resulting projections for AGP and HMP datasets, for every pair of metrics, are presented in Figs. 8 and 9. Since these partitions were found in an embedding space with dimensionality larger than three, we use for visualization PCA in Fig. 8 and Large Margin Nearest Neighbor method (*Weinberger & Saul, 2009*) in Fig. 9. Different approaches were chosen for the sake of better visualization of the clustering partition. The large margin nearest neighbor method allows for dimensionality reduction *via* linear transformation. This transformation is conditioned on the clustering partition of the data so that neighbor points from the same cluster are kept close, whereas points from different clusters are separated by a large gap.

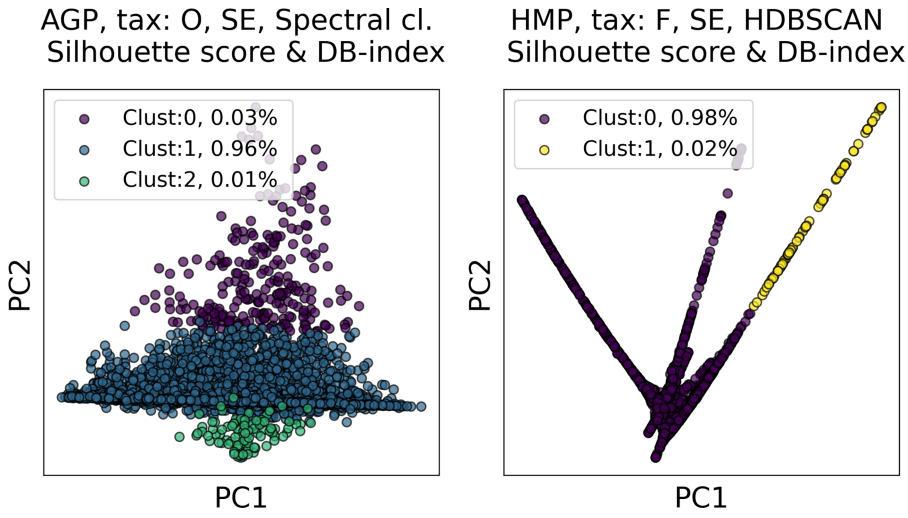

Figure 8 Visualization of the clustering results for the AGP and HMP datasets in the first two principal components. The visualized clustering partitions have the highest entropy among all other partitions satisfying Silhouette score and Davies–Bouldin (DB) Index thresholds. Dataset name, taxonomy level, representation of the data, clustering algorithm, and the pair of metrics used to select the partition are shown in the title. SE, spectral embedding. Color indicates different clusters. The percentage of the data belonging to each cluster is depicted on the legend.

Full-size DOI: 10.7717/peerj.15838/fig-8

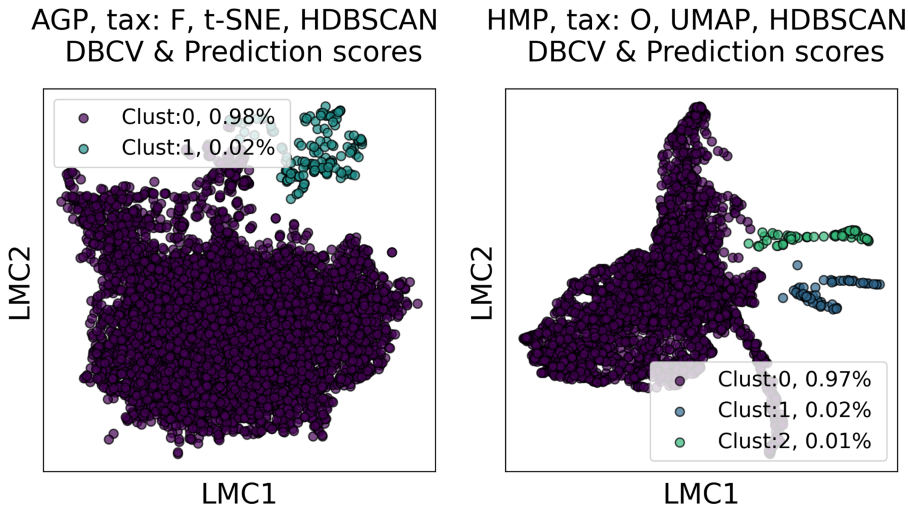

Figure 9 Visualization of the clustering results for the AGP and HMP datasets, using Large Margin Nearest Neighbor method. The visualized clustering partitions have the highest entropy among all partitions satisfying DBCV index and prediction strength thresholds. Dataset name, taxonomy level, representation of the data, clustering algorithm, and the pair of metrics used to select the partition are shown in the title. Color indicates different clusters. The percentage of the data belonging to each cluster is depicted on the legend.

Full-size DOI: 10.7717/peerj.15838/fig-9

Further, we have demonstrated that straightforward assignment of each data point to a potential enterotype, based on the originally reported distribution of *Bacteroides*, *Prevotella* and *Ruminococcus* Genera in enterotypes (*Arumugam et al., 2011*), does not reveal any natural clusters in different data representations. The corresponding distributions of clustering metrics and the visualization are presented in Text S4, Figs. S6–S7.

Together, these results imply that various clustering methods along with different manifold learning algorithms yield only highly imbalanced partitions, with more than 95% of data concentrated in one cluster. 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 is evidence (*Cooley et al., 2019*; *Kobak & Berens, 2019*) 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, presented in Text S2, Table S1, and Fig. S1, proves that distinct and stable clusters related to enterotypes have not been found because of their absence in the data rather than methodology flaws.

Visualization

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 clouds of data from the Genus taxonomic level on a two-dimensional plane using UMAP in Fig. 10 and t-SNE in Fig. 11. As explained above, prior to this step the datasets have been projected on the principal components capturing 99% of variance. To remove noise and outliers, after the dimensionality reduction small clusters of points containing less than 1% of the data were removed using the Local Outlier Factor algorithm (*Breunig et al., 2000*). To demonstrate the continuity of the data points distribution, we colored points as specific taxon relative abundances, corresponding to the genera most relevant for the definition of enterotypes, according to the initial finding (*Arumugam et al., 2011*). Salient parts of the manifold represent higher concentrations of specific OTUs. Small clusters observed in Figs. 10 and 11 are not related to enterotypes, being the direct result of specific methods hyperparameters, that may lead to tearing off the salient part of the data manifold.

To estimate the density of points in the visualization, we performed standard kernel density estimation (KDE) of this 2-dimensional data. The bandwidth parameter of the KDE is equal to the median value of pairwise distances distribution from every point to 100 closest neighbors. For the two-dimensional visualizations produced by both UMAP and t-SNE in Figs. 10 and 11, we observe that the density of data distribution is not uniform, indicating that there are regions of preferential concentrations of data, as shown in Fig. 12 for UMAP and in Fig. 13 for t-SNE algorithm. Nevertheless, this can be related only to the variations of OTU concentration (*Bacteroides* and *Prevotella*) and features of the dimensionality reduction methods rather than to formation of distinct clusters. This is

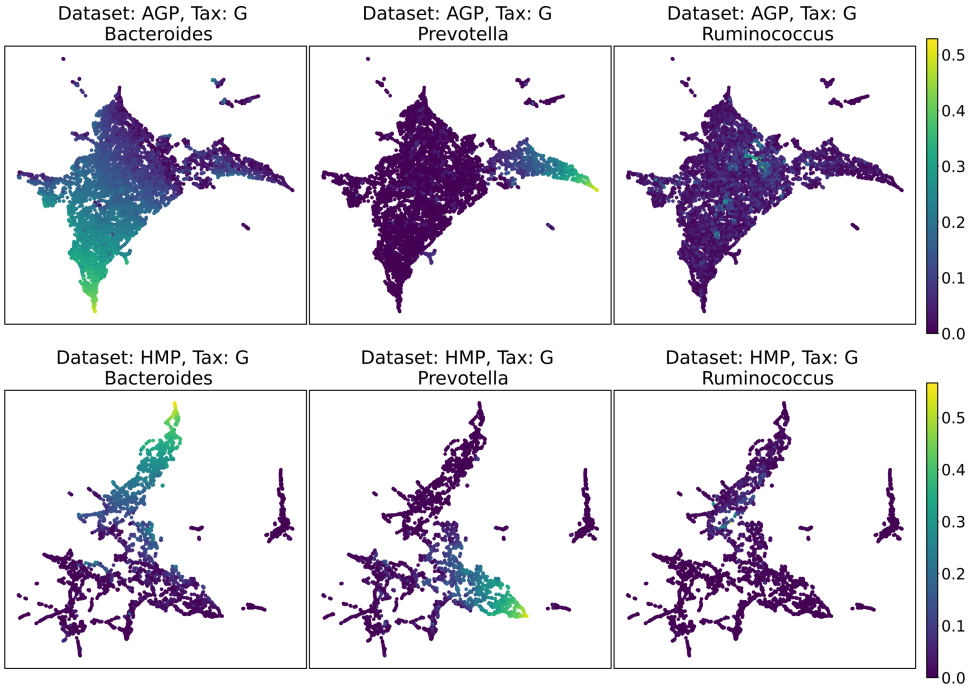

Figure 10 2D UMAP visualization of AGP and HMP datasets for the Genus taxonomy level. Colors reflect the relative abundance of specific taxa, see the headers.

Full-size DOI: 10.7717/peerj.15838/fig-10

ensured by the density-based clustering algorithm HDBSCAN, robust to noise and clusters shape, that would have indicated the existence of clusters by the high DBCV metric. Therefore, these regions are not related to the enterotypes in the original definition.

We supported this intuition by analyzing the OTU distribution in the high-density areas of visualizations in Figs. 10 and 11. Removing all regions in Figs. 12 and 13 with density less than 70% percentile of the total density distribution, we obtained separated, high-density areas. In Figs. 14 and 15, we show these regions in the two-dimensional visualization, along with the distributions of the ten most significant OTUs within the regions. The most significant OTUs are the ones with the largest mean value among all points that belong to the high-density regions. We observe that for the UMAP visualization in Fig. 14, the difference in the OTU distribution between clusters is mostly controlled by *Bacteroides* and *Prevotella* and an unclassified OTU at the Genus level, denoted as *Rest*. This observation is consistent with the abundance gradient visualization in Figs. 10 and 11, as well as previously reported results (*Costea et al., 2017*), indicating the continuous nature of the OTU distribution with preferential high-density regions. The same applies to the analysis of the t-SNE visualization in Fig. 15, with the difference, that the variation between high-density regions is also controlled by *Faecalibacterium* for AGP and *Lactobacillus and Streptococcus* for HMP.

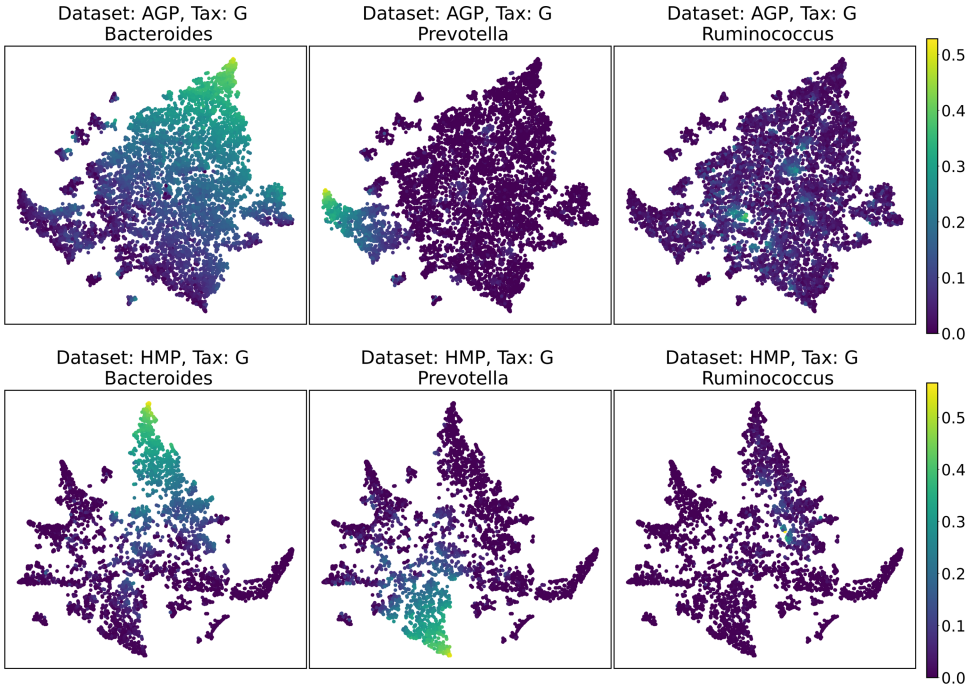

Figure 11 **2D t-SNE visualization of AGP and HMP datasets for the Genus taxonomy level.** Colors reflect the relative abundance of specific taxa, see the headers.

Full-size DOI: 10.7717/peerj.15838/fig-11

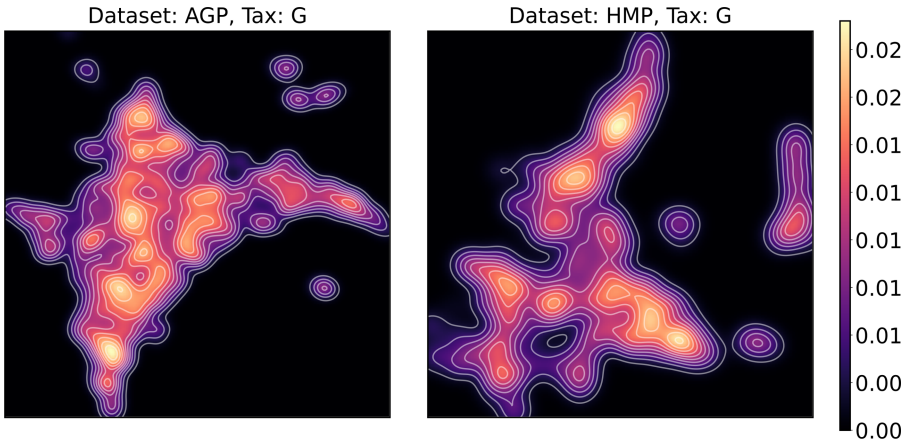

Figure 12 **Kernel density estimation (KDE) of 2D UMAP visualization.** Color indicates relative likelihood of the point to belong to the data distribution, according to KDE.

Full-size DOI: 10.7717/peerj.15838/fig-12

DISCUSSION

Our results demonstrate 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 applied for

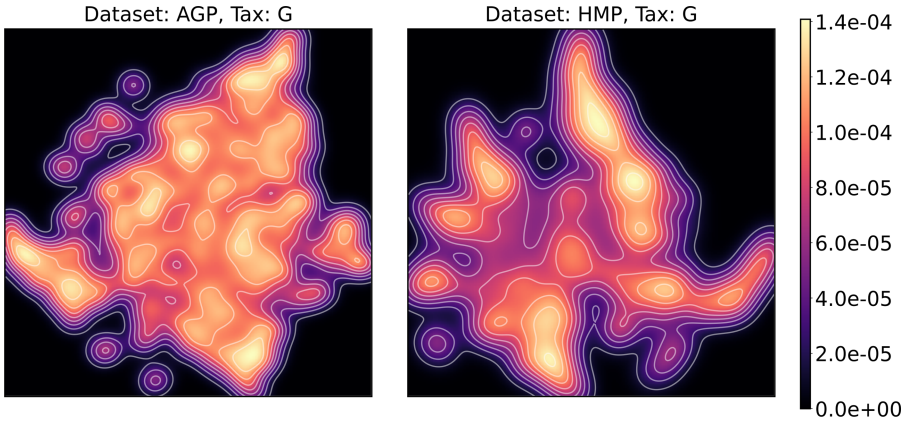

Figure 13 **Kernel density estimation (KDE) of 2D t-SNE visualization.** Color indicates relative likelihood of the point to belong to the data distribution, according to KDE.
Full-size DOI: 10.7717/peerj.15838/fig-13

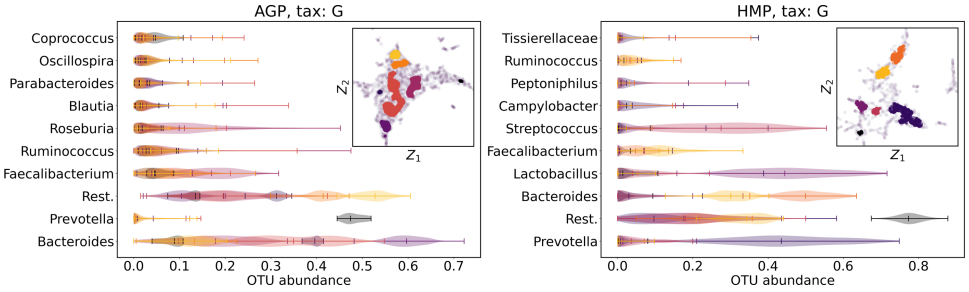

Figure 14 **Analysis of the high-density regions of 2D UMAP visualization.** The colored regions correspond to the kernel density estimation likelihood larger than 70% percentile of the total likelihood distribution. Color indicates different high-density clusters, depicted in the two-dimensional scatter plot with Z_1 and Z_2 coordinates. For the first ten selected OTU with the largest mean value among all high-density regions, the violin plots depict their distribution within each region.
Full-size DOI: 10.7717/peerj.15838/fig-14

the analysis of non-linear, high-dimensional metagenomic 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 the latter 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 intermediate 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. A more discrete structure could arise if more diverse samples are studied, including people with sharply differing lifestyles and diets. Still, successful attempts to correct the human gut microbiota were made, *e.g.*, fecal microbiota transplantation to treat the *Clostridium difficile* infection (CDI) (*Smits et al., 2013*), inflammatory bowel disease (IBD) (*Suskind et al., 2015*), and obesity (*Vrieze et al., 2012*). Connecting the distribution of microbiome abundances and structural features of

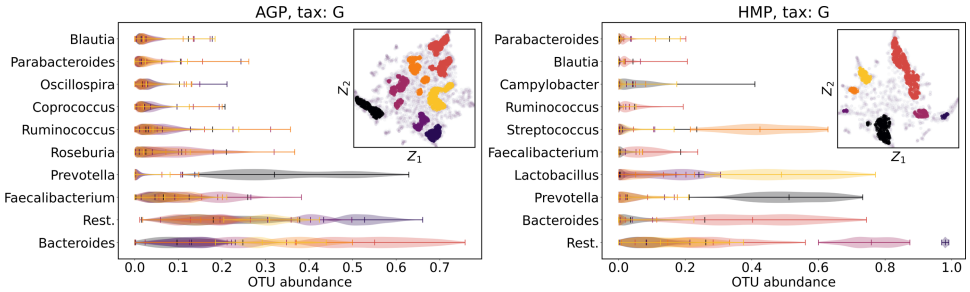

Figure 15 **Analysis of the high-density regions of 2D t-SNE visualization.** The colored regions correspond to the Kernel Density Estimation likelihood larger than 70% percentile of the total likelihood distribution. Color indicates different high-density clusters, depicted in the two-dimensional scatter plot with Z_1 and Z_2 coordinates. For the first ten selected OTU with the largest mean value among all high-density regions, the violin plots depict their distribution within each region.

Full-size DOI: 10.7717/peerj.15838/fig-15

the microbial manifold with lifestyle, nutrition, or disease may be a promising direction for further research.

CONCLUSIONS

We have shown the absence of enterotypes, defined as stable dense regions, or as stable and well separated clusters in the taxonomic abundance space, in human gut microbiomes. This challenges the current consensus, demonstrating that the metagenome distribution is continuous rather than discrete. We improved the standard methodology of microbial data analysis by applying a large variety of linear and non-linear dimensionality reduction methods to properly estimate the intrinsic dimension. We demonstrate that some of these methods do preserve the global and local data structure. This allowed us to achieve robustness of the clustering methods and compare results produced by different approaches. To the best of our knowledge, this is the first study applying a wide range of non-linear methods for validating the existence of enterotypes, and hence it may serve as a starting point for a more adequate analysis of metagenome datasets, which may reveal an intrinsic connection to nutrition, lifestyle, or disease. This study is also relevant for computational biologists seeking a general approach for clustering in high-dimensional data.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Sample preparation, data analysis, and biological interpretation of the results were supported by the Russian Foundation for Basic Research (grant 20-54-81007). Algorithms development and data processing were supported by the Russian Foundation for Basic Research (grant 21-51-12005 NNIO_a). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

Russian Foundation for Basic Research: 20-54-81007, 21-51-12005 NNIO_a.

Competing Interests

Mikhail S. Gelfand is an Academic Editor for PeerJ.

Author Contributions

- Ivan Bulygin performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Vladislav Shatov performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Anton Rykachevskiy performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Arsenii Raiko performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Alexander Bernstein conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Evgeny Burnaev conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Mikhail S. Gelfand conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Data and software are available at Figshare:

Bulygin, Ivan; Shatov, Vladislav; Rykachevskiy, Anton; Raiko, Arsenii; Bernstein, Alexander; Burnaev, Evgeny; et al. (2022). Absence of enterotypes in the human gut microbiomes reanalyzed with non-linear dimensionality reduction methods. figshare. Dataset. https://doi.org/10.6084/m9.figshare.19091423.v5

The software is also available at GitHub: https://github.com/blufzzz/Human-Gut-Microbiome-Analysis.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.15838#supplemental-information.

REFERENCES

Aggarwal CC, Hinneburg A, Keim DA. 2001. On the surprising behavior of distance metrics in high dimensional space. *Database Theory—ICDT* **2001**:420–434 DOI 10.1007/3-540-44503-x_27.

Amir A, McDonald D, Navas-Molina JA, Debelius J, Morton JT, Hyde E, Robbins-Pianka A, Knight R. 2017. Correcting for microbial blooms in fecal samples during room-temperature shipping. *mSystems* **2(2)**:e00199-16 DOI 10.1128/msystems.00199-16.

Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. 2013. An extensive comparative study of cluster validity indices. *Pattern Recognition* **46**:243–256 DOI 10.1016/j.patcog.2012.07.021.

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto J-M, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, De Vos WM, Brunak S, Doré J, Weissenbach J, Ehrlich SD, Bork P. 2011. Enterotypes of the human gut microbiome. *Nature* **473**:174–180 DOI 10.1038/nature09944.

Ben-Hur A, Guyon I. 2003. Detecting stable clusters using principal component analysis. *Functional Genomics* **224**:159–182 DOI 10.1385/1-59259-364-x:159.

Beyer K, Goldstein J, Ramakrishnan R, Shaft U. 1999. When is nearest neighbor meaningful? In: *Database theory—{ICDT'99}*. 21. Berlin, Heidelberg: Springer, 7–35 DOI 10.1007/3-540-49257-7_15.

Breunig MM, Kriegel H-P, Ng RT, Sander J. 2000. LOF. *ACM SIGMOD Record* **29**:93–104 DOI 10.1145/335191.335388.

Calinski T, Harabasz J. 1974. A dendrite method for cluster analysis. *Communications in Statistics—Theory and Methods* **3**:1–27 DOI 10.1080/03610927408827101.

Chen T, Long W, Zhang C, Liu S, Zhao L, Hamaker BR. 2017. Fiber-utilizing capacity varies in Prevotella- versus Bacteroides-dominated gut microbiota. *Scientific Reports* **7(1)**:2594 DOI 10.1038/s41598-017-02995-4.

Cheng M, Ning K. 2019. Stereotypes about enterotype: the old and new ideas. *Genomics, Proteomics and Bioinformatics* **17**:4–12 DOI 10.1016/j.gpb.2018.02.004.

Chiu C-M, Huang W-C, Weng S-L, Tseng H-C, Liang C, Wang W-C, Yang T, Yang T-L, Weng C-T, Chang T-H, Huang H-D. 2014. Systematic analysis of the association between gut flora and obesity through high-throughput sequencing and bioinformatics approaches. *BioMed Research International* **2014**:906168 DOI 10.1155/2014/906168.

Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S, Harris HMB, Coakley M, Lakshminarayanan B, O'Sullivan O, Fitzgerald GF, Deane J, O'Connor M, Harnedy N, O'Connor K, O'Mahony D, Sinderen DVan, Wallace M, Brennan L, Stanton C, Marchesi JR, Fitzgerald AP, Shanahan F, Hill C, Ross RP, O'Toole PW. 2012. Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488**:178–184 DOI 10.1038/nature11319.

Codoñer FM, Ramírez-Bosca A, Climent E, Carrión-Gutierrez M, Guerrero M, Pérez-Orquín JM, Parte JHorgadela, Genovés S, Ramón D, Navarro-López V, Chenoll E. 2018. Gut microbial composition in patients with psoriasis. *Scientific Reports* **8(7)**:e00060-18 DOI 10.1038/s41598-018-22125-y.

Cooley SM, Hamilton T, Aragones SD, Ray JCJ, Deeds EJ. 2019. A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-seq data. *bioRxiv* DOI 10.1101/689851.

Costea PI, Hildebrand F, Arumugam M, Bäckhed F, Blaser MJ, Bushman FD, De Vos WM, Ehrlich SD, Fraser CM, Hattori M, Huttenhower C, Jeffery IB, Knights D, Lewis JD, Ley RE, Ochman H, O'Toole PW, Quince C, Relman DA, Shanahan F, Sunagawa S, Wang J, Weinstock GM, Wu GD, Zeller G, Zhao L, Raes J, Knight R, Bork P. 2017. Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology* **3**:8–16 DOI 10.1038/s41564-017-0072-8.

Davies DL, Bouldin DW. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI* **1(2)**:224–227 DOI 10.1109/tpami.1979.4766909.

De Moraes ACF, Fernandes GR, Da Silva IT, Almeida-Pititto B, Gomes EP, Pereira ADC, Ferreira SRG. 2017. Enterotype may drive the dietary-associated cardiometabolic risk factors. *Frontiers in Cellular and Infection Microbiology* **7**:47 DOI 10.3389/fcimb.2017.00047.

De Wouters T, Doré J, Lepage P. 2012. Does our food (Environment) change our gut microbiome ('In-Vironment'): a potential role for inflammatory bowel disease? *Digestive Diseases* **30**:33–39 DOI 10.1159/000342595.

Deng Y, Wang H, Zhou J, Mou Y, Wang G, Xiong X. 2018. Patients with acne vulgaris have a distinct gut microbiota in comparison with healthy controls. *Acta Dermato Venereologica* **98**:783–790 DOI 10.2340/00015555-2968.

Dlugosz A, Winckler B, Lundin E, Zakikhany K, Sandström G, Ye W, Engstrand L, Lindberg G. 2015. No difference in small bowel microbiota between patients with irritable bowel syndrome and healthy controls. *Scientific Reports* **5**:8508 DOI 10.1038/srep08508.

Emoto T, Yamashita T, Sasaki N, Hirota Y, Hayashi T, So A, Kasahara K, Yodoi K, Matsumoto T, Mizoguchi T, Ogawa W, Hirata K. 2016. 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**:908–921 DOI 10.5551/jat.32672.

Fefferman C, Mitter S, Narayanan H. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society* **29**:983–1049 DOI 10.1090/jams/852.

Fix E, Hodges JL. 1989. Discriminatory analysis, nonparametric discrimination: consistency properties. *International Statistical Review / Revue Internationale de Statistique* **57(3)**:238–247 DOI 10.2307/1403797.

Gentle JE, Kaufman L, Rousseuw PJ. 1991. Finding groups in data: an introduction to cluster analysis. *Biometrics* **47**:788 DOI 10.2307/2532178.

Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. 2018. Current understanding of the human microbiome. *Nature Medicine* **24**:392–400 DOI 10.1038/nm.4517.

Gill SR, Pop M, De Boy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. 2006. Metagenomic analysis of the human distal gut microbiome. *Science* **312**:1355–1359 DOI 10.1126/science.1124234.

Goodfellow I, Bengio Y, Courville A. 2016. *Deep learning*. Cambridge, Massachusetts: MIT Press.

Gotoda T. 2015. 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**:16–7.

Huang H, Vangay P, McKinlay CE, Knights D. 2014. Multi-omics analysis of inflammatory bowel disease. *Immunology Letters* **162**:62–68 DOI 10.1016/j.imlet.2014.07.014.

Huss J. 2014. Methodology and ontology in microbiome research. *Biological Theory* **9**:392–400 DOI 10.1007/s13752-014-0187-6.

Ibragimov IA, Has'minskii RZ. 1981. *Statistical estimation*. New York: Springer DOI 10.1007/978-1-4899-0027-2.

Jeffery IB, Claesson MJ, O'Toole PW, Shanahan F. 2012. Categorization of the gut microbiota: enterotypes or gradients? *Nature Reviews Microbiology* **10(9)**:591–592 DOI 10.1038/nrmicro2859.

Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, Zhong H, Liu Z, Gao Y, Zhao H, Zhang D, Su Z, Fang Z, Lan Z, Li J, Xiao L, Li J, Li R, Li X, Li F, Ren H, Huang Y, Peng Y, Li G, Wen B, Dong B, Chen J-Y, Geng Q-S, Zhang Z-W, Yang H, Wang J, Wang J, Zhang X, Madsen L, Brix S, Ning G, Xu X, Liu X, Hou Y, Jia H, He K, Kristiansen K. 2017. The gut microbiome in atherosclerotic cardiovascular disease. *Nature Communications* **8**:845 DOI 10.1038/s41467-017-00900-1.

Kang C, Zhang Y, Zhu X, Liu K, Wang X, Chen M, Wang J, Chen H, Hui S, Huang L, Zhang Q, Zhu J, Wang B, Mi M. 2016. Healthy subjects differentially respond to dietary capsaicin correlating with specific gut enterotypes. *The Journal of Clinical Endocrinology & Metabolism* **101**:4681–4689 DOI 10.1210/jc.2016-2786.

Klimenko N, Tyakht A, Popenko A, Vasiliev A, Altukhov I, Ischenko D, Shashkova T, Efimova D, Nikogosov D, Osipenko D, Musienko S, Selezneva K, Baranova A, Kurilshikov A, Toshchakov S, Korzhenkov A, Samarov N, Shevchenko M, Tepliuk A, Alexeev D. 2018. Microbiome responses to an uncontrolled short-term diet intervention in the frame of the citizen science project. *Nutrients* **10**:576 DOI 10.3390/nu10050576.

Knights D, Ward TL, McKinlay CE, Miller H, Gonzalez A, McDonald D, Knight R. 2014. Rethinking enterotypes. *Cell Host & Microbe* **16**:433–437 DOI 10.1016/j.chom.2014.09.013.

Kobak D, Berens P. 2019. The art of using t-SNE for single-cell transcriptomics. *Nature Communications* **10(1)**:5416 DOI 10.1038/s41467-019-13056-x.

Kohler M, Krzyzak A, Walk H. 2009. Optimal global rates of convergence for nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference* **139**:1286–1296 DOI 10.1016/j.jspi.2008.07.012.

Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, Huttenhower C, Ley RE. 2013. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLOS Computational Biology* **9**:e1002863 DOI 10.1371/journal.pcbi.1002863.

Lee JA, Verleysen M. 2010. Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters* **31**:2248–2257 DOI 10.1016/j.patrec.2010.04.013.

Lee S-H, Yoon S-H, Jung Y, Kim N, Min U, Chun J, Choi I. 2020. Emotional well-being and gut microbiome profiles by enterotype. *Scientific Reports* **10**(**1**):20736 DOI 10.1038/s41598-020-77673-z.

Levina E, Bickel PJ. 2004. Maximum likelihood estimation of intrinsic dimension. NIPS..

Li J, Fu R, Yang Y, Horz H-P, Guan Y, Lu Y, Lou H, Tian L, Zheng S, Liu H, Shi M, Tang K, Wang S, Xu S. 2018. A metagenomic approach to dissect the genetic composition of enterotypes in Han Chinese and two Muslim groups. *Systematic and Applied Microbiology* **41**:1–12 DOI 10.1016/j.syapm.2017.09.006.

Liang C, Tseng H-C, Chen H-M, Wang W-C, Chiu C-M, Chang J-Y, Lu K-Y, Weng S-L, Chang T-H, Chang C-H, Weng C-T, Wang H-M, Huang H-D. 2017. Diversity and enterotype in gut bacterial community of adults in Taiwan. *BMC Genomics* **18**:932 DOI 10.1186/s12864-016-3261-6.

Maronna R, Charu A, Chandan R. 2015. Data clustering: algorithms and applications. *Statistical Papers* **57**:565–566 DOI 10.1007/s00362-015-0661-7.

McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, De Right Goldasich L, Dorrestein PC, Dunn RR, Fahimipour AK, Gaffney J, Gilbert JA, Gogul G, Green JL, Hugenholtz P, Humphrey G, Huttenhower C, Jackson MA, Janssen S, Jeste DV, Jiang L, Kelley ST, Knights D, Kosciolek T, Ladau J, Leach J, Marotz C, Meleshko D, Melnik AV, Metcalf JL, Mohimani H, Montassier E, Navas-Molina J, Nguyen TT, Peddada S, Pevzner P, Pollard KS, Rahnavard G, Robbins-Pianka A, Sangwan N, Shorenstein J, Smarr L, Song SJ, Spector T, Swafford AD, Thackray VG, Thompson LR, Tripathi A, Vázquez-Baeza Y, Vrbanac A, Wischmeyer P, Wolfe E, Zhu Q, Knight R, Mann AE, Amir A, Frazier A, Martino C, Lebrilla C, Lozupone C, Lewis Jr CM, Raison C, Zhang C, Lauber CL, Warinner C, Lowry CA, Callewaert C, Bloss C, Willner D, Galzerani DD, Gonzalez DJ, Mills DA, Chopra D, Gevers D, Berg-Lyons D, Sears DD, Wendel D, Lovelace E, Pierce E, TerAvest E, Bolyen E, Bushman FD, Wu GD, Church GM, Saxe G, Holscher HD, Ugrina I, German JB, Caporaso JG, Wozniak JM, Kerr J, Ravel J, Lewis JD, Suchodolski JS, Jansson JK, Hampton-Marcell JT, Bobe J, Raes J, Chase JH, Eisen JA, Monk J, Clemente JC, Petrosino J, Goodrich J, Gauglitz J, Jacobs J, Zengler K, Swanson KS, Lewis K, Mayer K, Bittinger K, Dillon L, Zaramela LS, Schriml LM, Dominguez-Bello MG, Jankowska MM, Blaser M, Pirrung M, Minson M, Kurisu M, Ajami N, Gottel NR, Chia N, Fierer N, White O, Cani PD, Gajer P, Strandwitz P, Kashyap P, Dutton R, Park RS, Xavier RJ, Mills RH, Krajmalnik-Brown R, Ley R, Owens SM, Klemmer S, Matamoros S, Mirarab S, Moorman S, Holmes S, Schwartz T, Eshoo-Anton TW, Vigers T, Pandey V, Treuren WV, Fang X, Zech Xu Z, Jarmusch A, Geier J, Reeve N, Silva R, Kopylova E, Nguyen D, Sanders K, Benitez RASalido, Heale AC, Abramson M, Waldispühl J, Butyaev A, Drogaris C, Nazarova E, Ball M, Gunderson B. 2018. American gut: an open platform for citizen science microbiome research. *mSystems* **3**(**3**):e00031-18 DOI 10.1128/msystems.00031-18.

McInnes L, Healy J. 2017. Accelerated hierarchical density based clustering. In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. DOI 10.1109/icdmw.2017.12.

McInnes L, Healy J, Saul N, Großberger L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**:861 DOI 10.21105/joss.00861.

Franco-de Moraes AC, De Almeida-Pititto B, Da Rocha Fernandes G, Gomes EP, Da Costa Pereira A, Ferreira SRG. 2017. Worse inflammatory profile in omnivores than in vegetarians associates with the gut microbiota composition. *Diabetology & Metabolic Syndrome* **9**:62 DOI 10.1186/s13098-017-0261-x.

Moser G, Fournier C, Peter J. 2017. Intestinal microbiome-gut-brain axis and irritable bowel syndrome. *Wiener Medizinische Wochenschrift* **168**:62–66 DOI 10.1007/s10354-017-0592-0.

Moulavi D, Jaskowiak PA, Campello RJGB, Zimek A, Sander J. 2014. Density-based clustering validation. In: *Proceedings of the 2014 SIAM international conference on data mining*. DOI 10.1137/1.9781611973440.96.

Nakayama J, Watanabe K, Jiang J, Matsuda K, Chao S-H, Haryono P, La-ongkham O, Sarwoko M-A, Sujaya IN, Zhao L, Chen K-T, Chen Y-P, Chiu H-H, Hidaka T, Huang N-X, Kiyohara C, Kurakawa T, Sakamoto N, Sonomoto K, Tashiro K, Tsuji H, Chen M-J, Leelavatcharamas V, Liao C-C, Nitisinprasert S, Rahayu ES, Ren F-Z, Tsai Y-C, Lee Y-K. 2015. Diversity in gut bacterial community of school-age children in Asia. *Scientific Reports* **5**:8397 DOI 10.1038/srep08397.

Nakayama J, Yamamoto A, Palermo-Conde LA, Higashi K, Sonomoto K, Tan J, Lee Y-K. 2017. Impact of westernized diet on gut microbiota in children on Leyte Island. *Frontiers in Microbiology* **8**:197 DOI 10.3389/fmicb.2017.00197.

Noguera-Julian M, Rocafort M, Guillén Y, Rivera J, Casadellà M, Nowak P, Hildebrand F, Zeller G, Parera M, Bellido R, Rodríguez C, Carrillo J, Mothe B, Coll J, Bravo I, Estany C, Herrero C, Saz J, Sirera G, Torrela A, Navarro J, Crespo M, Brander C, Negredo E, Blanco J, Guarner F, Calle ML, Bork P, Sönnerborg A, Clotet B, Paredes R. 2016. Gut microbiota linked to sexual preference and HIV infection. *EBioMedicine* **5**:135–146 DOI 10.1016/j.ebiom.2016.01.032.

Oki K, Toyama M, Banno T, Chonan O, Benno Y, Watanabe K. 2016. 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)**:284 DOI 10.1186/s12866-016-0898-x.

Ou J, Carbonero F, Zoetendal EG, Lany JPDe, Wang M, Newton K, Gaskins HR, O'Keefe SJ. 2013. Diet, microbiota, and microbial metabolites in colon cancer risk in rural Africans and African Americans. *The American Journal of Clinical Nutrition* **98**:111–120 DOI 10.3945/ajcn.112.056689.

Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, Beghini F, Malik F, Ramos M, Dowd JB, Huttenhower C, Morgan M, Segata N, Waldron L. 2017.

Accessible, curated metagenomic data through ExperimentHub. *Nature Methods* **14**:1023–1024 DOI 10.1038/nmeth.4468.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* **12(85)**:2825–2830 DOI 10.5555/1953048.2078195.

Psutka JV, Psutka J. 2015. Sample size for maximum likelihood estimates of gaussian model. In: *Computer analysis of images and patterns*. 462–469 DOI 10.1007/978-3-319-23117-4_40.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD, Wang J. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**:59–65 DOI 10.1038/nature08821.

Qin N, Zheng B, Yao J, Guo L, Zuo J, Wu L, Zhou J, Liu L, Guo J, Ni S, Li A, Zhu Y, Liang W, Xiao Y, Ehrlich SD, Li L. 2015. Influence of H7N9 virus infection and associated treatment on human gut microbiota. *Scientific Reports* **5**:14771 DOI 10.1038/srep14771.

Robles-Alonso V, Guarner F. 2013. Progress in the knowledge of the intestinal human microbiota. *Nutricion Hospitalaria* **28**:553–557 DOI 10.3305/nh.2013.28.3.6601.

Castaño Rodríguez N, Underwood AP, Merif J, Riordan SM, Rawlinson WD, Mitchell HM, Kaakoush NO. 2018. Gut microbiome analysis identifies potential etiological factors in acute gastroenteritis. *Infection and Immunity* **86(7)**:e00060-18 DOI 10.1128/iai.00060-18.

Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**:53–65 DOI 10.1016/0377-0427(87)90125-7.

Ruppert D. 2004. The elements of statistical learning: data mining, inference, and prediction. *Journal of the American Statistical Association* **99**:567–567 DOI 10.1198/jasa.2004.s339.

Shankar V, Gouda M, Moncivaiz J, Gordon A, Reo NV, Hussein L, Paliy O. 2017. Differences in gut metabolites and microbial composition and functions between Egyptian and U.S. Children are consistent with their diets. *mSystems* **2(1)**:e00169-16 DOI 10.1128/msystems.00169-16.

Shannon CE. 1948. A mathematical theory of communication. *Bell System Technical Journal* **27**:379–423 DOI 10.1002/j.1538-7305.1948.tb01338.x.

Shi J, Malik J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**:888–905 DOI 10.1109/34.868688.

Shortt N, Poyntz H, Young W, Jones A, Gestin A, Mooney A, Thayabaran D, Sparks J, Ostapowicz T, Tay A, Poppitt S, Elliott S, Wakefield G, Parry-Strong A, Ralston J, Gasser O, Beasley R, Weatherall M, Braithwaite I, Forbes-Blom E. 2018. A feasibility study: association between gut microbiota enterotype and antibody response to seasonal trivalent influenza vaccine in adults. *Clinical & Translational Immunology* **7**:e1013 DOI 10.1002/cti2.1013.

Smits LP, Bouter KEC, De Vos WM, Borody TJ, Nieuwdorp M. 2013. Therapeutic potential of fecal microbiota transplantation. *Gastroenterology* **145**:946–953 DOI 10.1053/j.gastro.2013.08.058.

Suskind DL, Brittnacher MJ, Wahbeh G, Shaffer ML, Hayden HS, Qin X, Singh N, Damman CJ, Hager KR, Nielson H, Miller SI. 2015. Fecal microbial transplant effect on clinical outcomes and fecal microbiome in active Crohn's disease. *Inflammatory Bowel Diseases* **21**:556–563 DOI 10.1097/mib.0000000000000307.

Tenenbaum JB, De Silva V, Langford JC. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**:2319–2323 DOI 10.1126/science.290.5500.2319.

The Human Microbiome Project Consortium. 2012a. A framework for human microbiome research. *Nature* **486**:215–221 DOI 10.1038/nature11209.

The Human Microbiome Project Consortium. 2012b. Structure, function and diversity of the healthy human microbiome. *Nature* **486(7402)**:207–214 DOI 10.1038/nature11234.

Thomas AM, Jesus EC, Lopes A, Aguiar S, Begnami MD, Rocha RM, Carpinetti PA, Camargo AA, Hoffmann C, Freitas HC, Silva IT, Nunes DN, Setubal JC, Dias-Neto E. 2016. Tissue-associated bacterial alterations in rectal carcinoma patients revealed by 16S rRNA community profiling. *Frontiers in Cellular and Infection Microbiology* **6**:179 DOI 10.3389/fcimb.2016.00179.

Tibshirani R, Walther G. 2005. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics* **14**:511–528 DOI 10.1198/106186005x59243.

Tigchelaar EF, Bonder MJ, Jankipersadsing SA, Fu J, Wijmenga C, Zhernakova A. 2015. Gut microbiota composition associated with stool consistency. *Gut* **65**:540–542 DOI 10.1136/gutjnl-2015-310328.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. 2008. A core gut microbiome in obese and lean twins. *Nature* **457**:480–484 DOI 10.1038/nature07540.

Van Der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9(86)**:2579–2605.

Vandeputte D, Falony G, Vieira-Silva S, Tito RY, Joossens M, Raes J. 2015. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* **65**:57–62 DOI 10.1136/gutjnl-2015-309618.

Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, Tito RY, De Commer L, Darzi Y, Vermeire S, Falony G, Raes J. 2017. Quantitative

microbiome profiling links gut community variation to microbial load. *Nature* **551**:507–511 DOI 10.1038/nature24460.

Vieira-Silva S, Falony G, Darzi Y, Lima-Mendez G, Garcia Yunta R, Okuda S, Vandeputte D, Valles-Colomer M, Hildebrand F, Chaffron S, Raes J. 2016. Species–function relationships shape ecological properties of the human gut microbiome. *Nature Microbiology* **1**(**8**):16088 DOI 10.1038/nmicrobiol.2016.88.

Vieira-Silva S, Sabino J, Valles-Colomer M, Falony G, Kathagen G, Caenepeel C, Cleynen I, Merwe Svander, Vermeire S, Raes J. 2019. Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nature Microbiology* **4**:1826–1831 DOI 10.1038/s41564-019-0483-9.

Von Luxburg U. 2007. A tutorial on spectral clustering. *Statistics and Computing* **17**:395–416 DOI 10.1007/s11222-007-9033-z.

Vrieze A, Van Nood E, Holleman F, Salojärvi J, Kootte RS, Bartelsman JFWM, Dallinga-Thie GM, Ackermans MT, Serlie MJ, Oozeer R, Derrien M, Druesne A, Van Hylckama Vlieg JET, Bloks VW, Groen AK, Heilig HGHJ, Zoetendal EG, Stroes ES, Vos WMDe, Hoekstra JBL, Nieuwdorp M. 2012. Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. *Gastroenterology* **143**:913–916.e7 DOI 10.1053/j.gastro.2012.06.031.

Wang J, Linnenbrink M, Künzel S, Fernandes R, Nadeau M-J, Rosenstiel P, Baines JF. 2014. 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 of the United States of America* **111**(**26**):E2703-10 DOI 10.1073/pnas.1402342111.

Weinberger K, Saul L. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* **10**(**9**):207–244 DOI 10.5555/1577069.1577078.

Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD. 2011. Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**:105–108 DOI 10.1126/science.1208344.

Wu Q, Pi X, Liu W, Chen H, Yin Y, Yu HD, Wang X, Zhu L. 2017. Fermentation properties of isomaltooligosaccharides are affected by human fecal enterotypes. *Anaerobe* **48**:206–214 DOI 10.1016/j.anaerobe.2017.08.016.

Xia F, Chen J, Fung WK, Li H. 2013. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69**:1053–1063 DOI 10.1111/biom.12079.

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. *Nature* **486**:222–227 DOI 10.1038/nature11053.

Yin Y, Fan B, Liu W, Ren R, Chen H, Bai S, Zhu L, Sun G, Yang Y, Wang X. 2017. Investigation into the stability and culturability of Chinese enterotypes. *Scientific Reports* **7**:7947 DOI 10.1038/s41598-017-08478-w.

Zhang Z, Wang J. 2007. MLLE: modified locally linear embedding using multiple weights. *Advances in Neural Information Processing Systems* **19**:1593–1600 DOI 10.7551/mitpress/7503.003.0204.