Clustering and classification of human microbiome data: evaluating the impact of different settings in bioinformatics workflows

Debora Santo*, Tatjana Lončar-Turukalo†, Blaž Stres‡, Vladimir Crnojević* and Sanja Brdar*
*BioSense Institute, University of Novi Sad, Dr Zorana Đinđića 1, 21000 Novi Sad, Serbia
†Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
‡Biotechnical Faculty, University of Ljubljana, Jamnikarjeva 101, Sl-1000 Ljubljana, Slovenia
Email: debora.santo@biosense.rs

***Abstract*—Microbiome studies are attracting increasing interest, especially in human health applications, where their use for disease prognostics, diagnostics and treatment can have immense effects on life quality. The settings in the microbiome data preprocessing stage can lead to the great variability of the generated operational taxonomic unit (OTU) tables, reflected in the size and sparseness of this data matrix. As there are still no solid guidelines on the best practices, it is valuable to assess which machine learning algorithms provide higher stability of results under variable preprocessing settings.**

In this study, we have generated OTU tables using data from the *Moving pictures of human microbiome* study using two different reference databases (Greengenes and Silva) and four levels of the similarity threshold (ranging from 90 to 99%), processed in the QIIME bioinformatics package. The results of the two best-performing classification and clustering algorithms are presented in detail: Random Forest classifier (RF) and Spectral clustering (SC). The random forest classifier has outperformed spectral clustering in terms of accuracy. As the rate of data generation increases, while the cost of labeling remains high, further improvement of clustering performance and ensemble approaches should be explored.

***Keywords*-Bioinformatics; microbiome; machine learning; OTU table; 16S rRNA.**

I. Introduction

Microbiome studies have a great potential for resolving questions and problems in the medical, environmental and agricultural field when coupled with reliable and extensively collected multi-scale meta-data of the samples. The rising awareness of the many roles microbes take up in their hosts and habitats is increasing the interest and funding that researchers have for these studies. When talking about the human microbiome, because of the impact it has on the host body health and physiology, it's rightfully called the second genome [1].

Microbiome information is most often accessed via 16s rRNA marker sequencing [2] or DNA metabarcoding approaches that apply whole-genome sequencing (WGS) [3], [4]. While both can reveal the diversity and relative abundance of microbial communities within the researched samples, the later can also provide additional information on the span of their functions within the given habitat. The rapidly decreasing price of high-throughput sequencing methods and computing equipment is making metagenomic studies more accessible than ever [5], but the amount of data produced is enormous and extracting meaningful and reliable biological knowledge is still challenging.

A few large projects have been collecting the microbiome data and knowledge within several different fields, the two biggest being The Human Microbiome Project (HMP) geared towards studies of the microbiome's effect on the human health [6], [7], [8], [9] and the Earth Microbiome Project (EMP) that focuses on general microbial diversity across habitats [10], [11], [12], [13]. One of the remaining issues is that as many as 99,5% of the microbes, especially among those found in soil samples, might still be undetectable as they have not been successfully isolated and cultured so far, and thus are not included in the referenced databases [14], [15]. This bottleneck in which microbes we are able to detect in our samples can limit the power of the biological interpretations of the study results and therefore it is crucial to invest time and money in building more extensive microbe reference databases.

To efficiently extract biologically relevant and reliable information from microbiome data coupled with multi-scale metadata, we also need robust computational and modelling approaches. During the handling of this data, numerous dilemmas can arise in choosing the optimal preprocessing parameters (sequence quality filtering, reference database choice, software package choice, picking operational taxonomic units (OTUs), generating OTU tables, etc.), but also in choosing a clustering or classification algorithm to unveil the structure of the microbial communities. There is a need for higher stability and reliability in the derived conclusions. As there are still no clear standards for the optimal parameter setups, the assessment of the effect their variation has on the performance of the machine learning algorithms has a great value for the research community. The OTU stability is often neglected, but it's an important part within the analysis [16]. Variability can arise during *de novo* clustering, when a reference sequence is lacking or when different reference databases are used, as a result of implementing different β-diversity measures for phylogenetic differentiation [17], [18], [19], when different

2471-7819/19/$31.00 ©2019 IEEE
DOI 10.1109/BIBE.2019.00156

clustering or classification algorithms are being applied, or due to grouping into enterotypes (based on the prevalent bacterial genera in the gut microbiome) where it has been shown that the results highly depend on data source type, sequencing depth, the distance metric, similarity thresholds, OTU-picking, etc. [20].

The *Moving pictures of human microbiome* dataset [1] used in our evaluation contains temporal analyses of the microbial data. Such longitudinal studies come with their sets of challenges in correctly assessing what variation span can be attributed to normal microbial community dynamics, what leads to a pathological state, and how therapeutics can help restore a healthy microbiome [21], [22], [23]. In that study, it has been shown that only a few taxa persist over time and that the temporal variability in the microbiota of an individual is quite pronounced - which demonstrates one of the challenges in the interpretation of the results. Despite that, the differences between body sites and individuals are stable and sufficient for correct sample classification, which makes this dataset suitable for comparing the performance of different algorithms in capturing the structure of the data. We have evaluated the stability and robustness of the clustering and classification algorithms' output with respect to variation in the parameter settings within the QIIME preprocessing pipeline [24], [25]. More specifically, we have observed the influence of the reference database used, exploring the Silva and Greengenes as options, as well as the influence of the similarity threshold settings used to identify taxa.

II. Material and methods

A. Data

In this research, we used the data from the *Moving pictures of human microbiome* study [1], which was the largest human microbiome time series analysis performed to that date. The size of the dataset is $\approx$ 12GB and it encompasses millions of 16S rRNA sequences from 1967 samples, obtained from three body sites (oral, skin, gut) of one male and one female subject collected during 15 and 6 months, respectively, covering over 396 time-points (see Table I).

Table I
THE SAMPLES' GENDER AND SITE DISTRIBUTION IN THE *Moving pictures of the human microbiome* DATASET.

Gender	Site	Number of samples
Female	Oral	135
Female	Skin	268
Female	Gut	131
Male	Oral	373
Male	Skin	724
Male	Gut	336

As sufficient stability was demonstrated in the differentiation between body sites, this dataset was chosen due to its suitable characteristics for testing classification and clustering algorithms. The data was accessed via MG-RAST API [26] after quality filtering.

B. Methods

The data was processed in QIIME [24], [25]. The overall experimental workflow included the following steps:

1) Preprocessing (removing primers, demultiplexing, quality filtering);
2) Picking OTUs (clustering reads by sequence similarity to a reference sequence database to determine their taxonomic belonging, discarding the rest);
3) Building an OTU table (containing OTU counts per sample);
4) Measuring β-diversity between samples and rarefaction;
5) Training and cross-validating classification algorithm on an OTU table, resulting from a step 3 for each combination of the parameter settings;
6) Clustering samples based on β-diversity matrices, that are the result of step 4.

The first 4 steps are part of QIIME package that were run with different settings. The last 2 belong to the machine learning part of the study that was used for algorithm stability evaluation.

When picking OTUs in the second step, the variability in the results was explored in the light of two different reference databases (Greengenes [27], [28], [29] and Silva [30], [31]), as it has been shown earlier how the choice of the reference database can have an impact on the results [32]. We used the latest available releases *gg_13_5* from 2013 and *Silva_132_release* from 2017 for Greengenes and Silva, respectively. Four levels of taxonomy were induced by choosing different similarity thresholds (90 to 99%, with 97% being the usual threshold when defining the taxonomy to the species level) for clustering the sequences using the default taxonomy assigner - UCLUST [33]. This was done to see how these settings affect the number and sparseness of features within the generated OTU tables. The picked OTUs (retained and taxonomically assigned sequences) identified in the analyzed samples, form an OTU table that represents the microbial taxonomy per sample in a form of observation counts per sample. The OTU table is a data matrix, an input to the machine learning algorithms.

The Random Forest (RF) was used as a classification algorithm, with n=500 trees, while the evaluation of the classification accuracy was assessed using confusion matrices, estimated using 10-fold cross-validation.

For the clustering algorithms, the similarity measure between samples is the first to be defined. β-diversity (microbial community diversity between samples from different sites) was measured from OTU tables to generate pairwise

distance matrices. 24 non-phylogenetic diversity (distance) measures were explored: (1) binary Chi-square, (2) binary Chord, (3) binary Euclidean, (4) binary Hamming, (5) binary Jaccard, (6) binary Lennon, (7) binary Ochiai, (8) binary Pearson, (9) binary Sörensen-Dice, (10) Bray-Curtis, (11) Canberra, (12) Chi-square, (13) Chord, (14) Euclidean, (15) Gower, (16) Hellinger, (17) Jaccard, (18) Kulczynski, (19) Manhattan distance, (20) Morisita-Horn, (21) Pearson, (22) Soergel, (23) Spearman rank, and (24) Species profile distance.

To quantify how similar are the β-diversity matrices, we measured the correlation between different β-diversity measures using the Mantel test [34] formulated as:

$$rm = \frac{1}{d-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} D_{Xij} D_{Yij} \quad (1)$$

where $d = n(n-1)/2$, and D_{Xij}, D_{Yij} are standardized distances, and n is the number of samples that are taken pairwise to calculate the distance matrix.

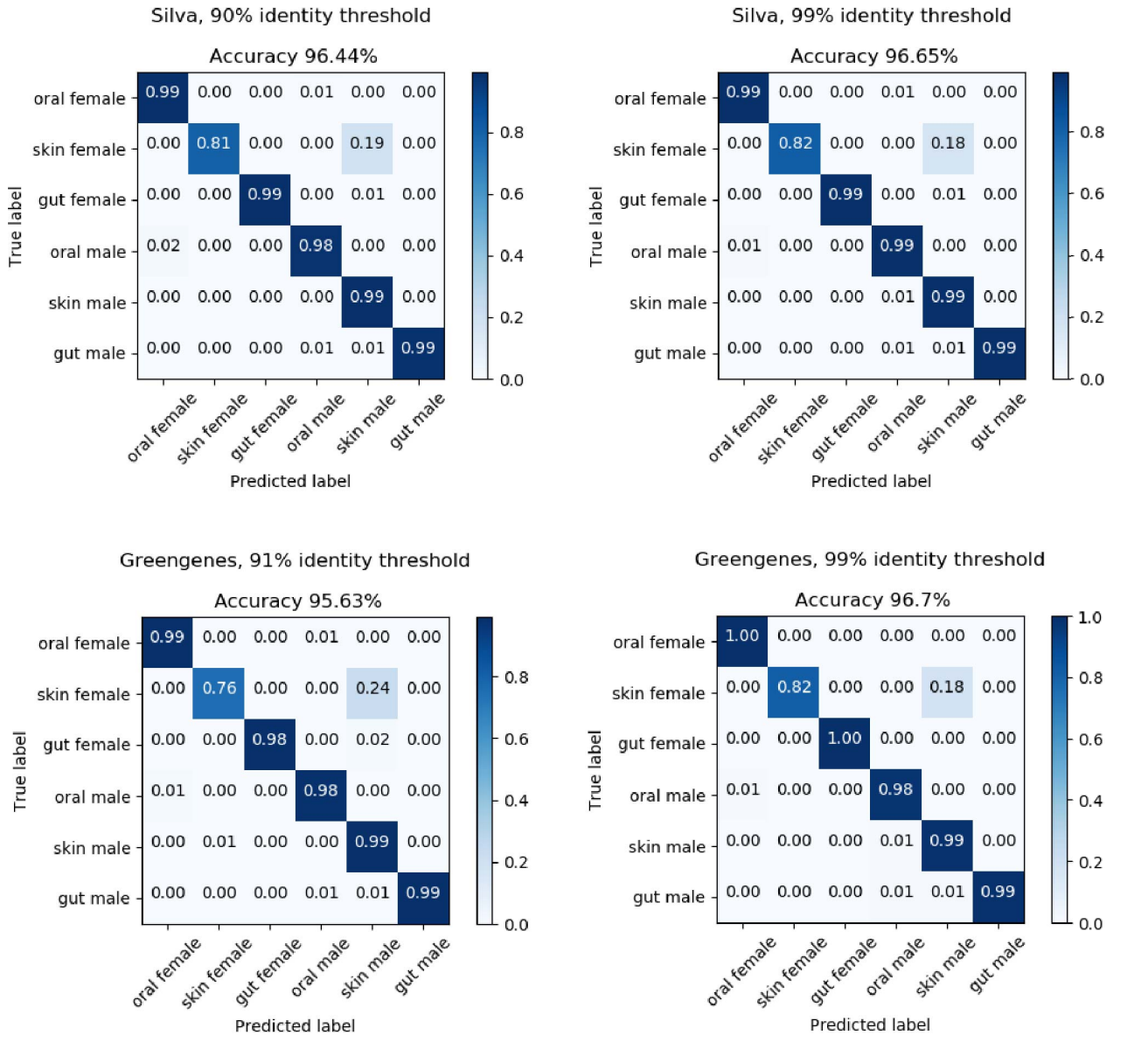

Figure 1. Evaluation of the Random Forest algorithm based on confusion matrices showing the corresponding classification accuracy for the similarity thresholds at 90 or 91% and 99%, using different databases: Greengenes and Silva.

Prior to clustering, we have performed the Kernel Principal Component Analysis (KPCA) to visually evaluate how well the samples are discriminated based on the first two principal components that rely on different β-diversity measures.

Clustering was performed using the spectral clustering algorithm [35] as it works directly on the β-diversity distance matrices. For all β-diversity metrics the pairwise distances between samples were transformed into similarities following the element-wise transformation:

$$S_{ij} = e^{-D_{ij}^2/(2\sigma^2)} \quad (2)$$

where D is a pair-wise β-diversity matrix, σ is the mean value of that matrix, and S is the final similarity matrix. The spectral clustering was performed for all 24 β-diversity metrics independently, using the corresponding similarity matrix as input, resulting in 24 partitions for each combination of the similarity threshold and the reference database evaluated.

We used the Adjusted Rand Index (ARI) [36] to evaluate and compare the clustering results. Rand index [37] is the measure of similarity between two data clustering results (expected versus observed), while ARI is corrected-for-chance version of the Rand Index [38].

III. Results

The variability in the QIIME preprocessing settings (the reference databases used and similarity thresholds set) have highly impacted the number of features (ranging from 6637 to 87069) and values in the OTU tables (i.e. matrix sparseness). This was because the possibility to detect the specific microbe sequence in a certain database and the sequence similarity cutoff determines the number of OTUs in the future analysis. Initially, several classification and clustering algorithms were tested with multiple splits into training and test samples, but the performance of the two best is presented: Random Forest classifier (RF) and Spectral clustering (SC).

A. Random Forest classifier (RF)

The RF algorithm has demonstrated a capacity to learn on the sparse data and provided very stable classification results regardless of the reference database and the identity threshold settings. The generated confusion matrices (Fig. 1) show the RF performance evaluation for the Silva database with the similarity thresholds set to 90% and 99% (upper panel), and the corresponding matrices for the Greengenes database with the similarity thresholds set to 91% and 99% (lower panel). The lower similarity threshold limit differs between the two databases because the same similarity threshold option was not available in both, so the closest was chosen (90% and 91%, respectively). The classification accuracy, estimated by the 10-fold cross-validation, was in average around 96% for different parameter settings (ranging from 95.63% to 96.7%, and increasing as the similarity thresholds have been increased).

B. β-diversity evaluation

The evaluation of the correlation between β-diversity measures was performed using the Mantel test. The test results are summarized via boxplot and presented in Fig. 2. The boxplot presents the distribution of the measured correlation between all pairs of the β-diversity matrices. With the median of ≈ 0.55, the obtained results provide insight that different β-diversities capture different relationships between microbiome samples, which has raised concerns on the diversity of the partitions in the clustering step.

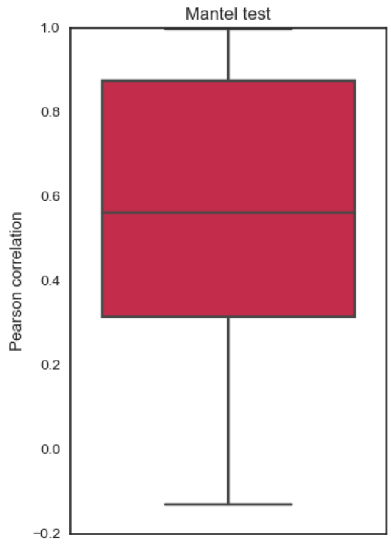

Figure 2. Evaluation of the correlation between the β-diversity measures using the Mantel test.

When it comes to the low dimensional representation of the samples in the OTU table, the KPCA was able to resolve the body sites, but without gender separation. This was achieved using the first two principal components (see Fig. 3).

C. Spectral clustering (SC)

SC has shown a large ARI variability when different β-diversity matrices were used and a high sensitivity to any change in the parameters. Agreement of the clustering results with true classes measured by ARI varied from 0.35 to 0.59, with mean value 0.51 ± 0.06 standard deviation. The ARI for partitions using all β-diversity measures is shown in Fig. 4 and across all threshold in Fig. 5.

Fig. 4 uncovers the performance of clustering on each of the used β-diversity measures, reevaluated with different thresholds for similarities with sequences in the two reference databases. While for some β-diversity measures the threshold selection does not highly impact the result (noticed as the close grouping of dots of the same color),

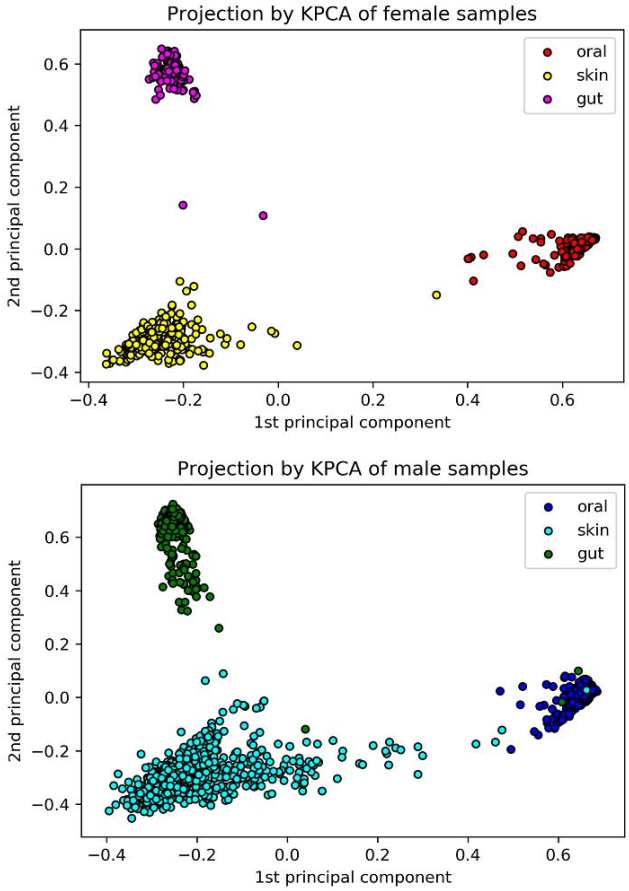

Figure 3. Kernel Principal Component Analysis (KPCA) separation of female and male study participant's samples according to body site based on PC1 and PC2 of the abundance weighted Jaccard distance as the β-diversity metric. On both figures, different colors indicate different body sites (oral, skin and gut).

others are more sensitive (e.g. specprof - Species profile distance, whose ARI for different thresholds is taking up a wider range of values). Interestingly, the clustering on the binary β-diversity measures provided results that better align with true labels, as evaluated by ARI, but at the same time, the results were more sensitive to the threshold and database selection.

Fig. 5 shows a different perspective on the variability of the clustering results. We have measured the stability of ARI for spectral clustering using β-diversity measures under a few different similarity thresholds and two different databases (Silva and Greengenes). We can notice the differences between the results generated using different reference databases and the smallest variability for the 99% threshold, especially in the case where the Greengenes database was used.

The clustering evaluation results were confirmed by an alternative external validation measure as well - Normalized Mutual Information (NMI) [39].

IV. Conclusion

The efficient classification of a healthy baseline microbiome data provides grounds for a robust modelling of microbiome data for health-related datasets. The RF classifier has better separated the samples according to their labels, outperforming the clustering with respect to both accuracy and the stability of the results. However, as clustering is prevailing in use as no labels are available in most microbiome studies (but are often yet to be discovered and assigned to samples), further improvements of clustering by dimensionality reduction, ensemble or semi-supervised approaches should be explored [40], [41]. A new release of QIIME, QIIME2 has been published [42], opening the possibility to compare the performance of the enhanced preprocessing analysis pipeline. Also, alternative preprocessing tools could be included in the analysis to comprehensively evaluate different settings in bioinformatics workflows and their performance in microbiome analyses.

Acknowledgment

This research was in part supported by the grants III43002 and III44006 of the Ministry of Education, Science and Technological Development of the Republic of Serbia. It was based upon work done within the COST Action CA18131: Statistical and machine learning techniques in human microbiome studies, that's supported by the COST Association (European Cooperation in Science and Technology).

References

[1] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer *et al.*, "Moving pictures of the human microbiome," *Genome biology*, vol. 12, no. 5, p. R50, 2011.

[2] J. M. Janda and S. L. Abbott, "16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls," *Journal of Clinical Microbiology*, vol. 45, no. 9, pp. 2761–2764, 2007. [Online]. Available: https://jcm.asm.org/content/45/9/2761

[3] M. L. Zepeda Mendoza, T. Sicheritz-Pontén, and M. T. P. Gilbert, "Environmental genes and genomes: understanding the differences and challenges in the approaches and software for their analyses," *Briefings in Bioinformatics*, vol. 16, no. 5, pp. 745–758, 02 2015. [Online]. Available: https://doi.org/10.1093/bib/bbv001

[4] E. Aylagas, A. Borja, X. Irigoien, and N. Rodríguez-Ezpeleta, "Benchmarking dna metabarcoding for biodiversity-based monitoring and assessment," *Frontiers in Marine Science*, vol. 3, p. 96, 2016. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmars.2016.00096

[5] J. Zhou, Z. He, Y. Yang, Y. Deng, S. G. Tringe, and L. Alvarez-Cohen, "High-throughput metagenomic technologies for complex microbial community analysis: Open and closed formats," *mBio*, vol. 6, no. 1, 2015. [Online]. Available: https://mbio.asm.org/content/6/1/e02288-14

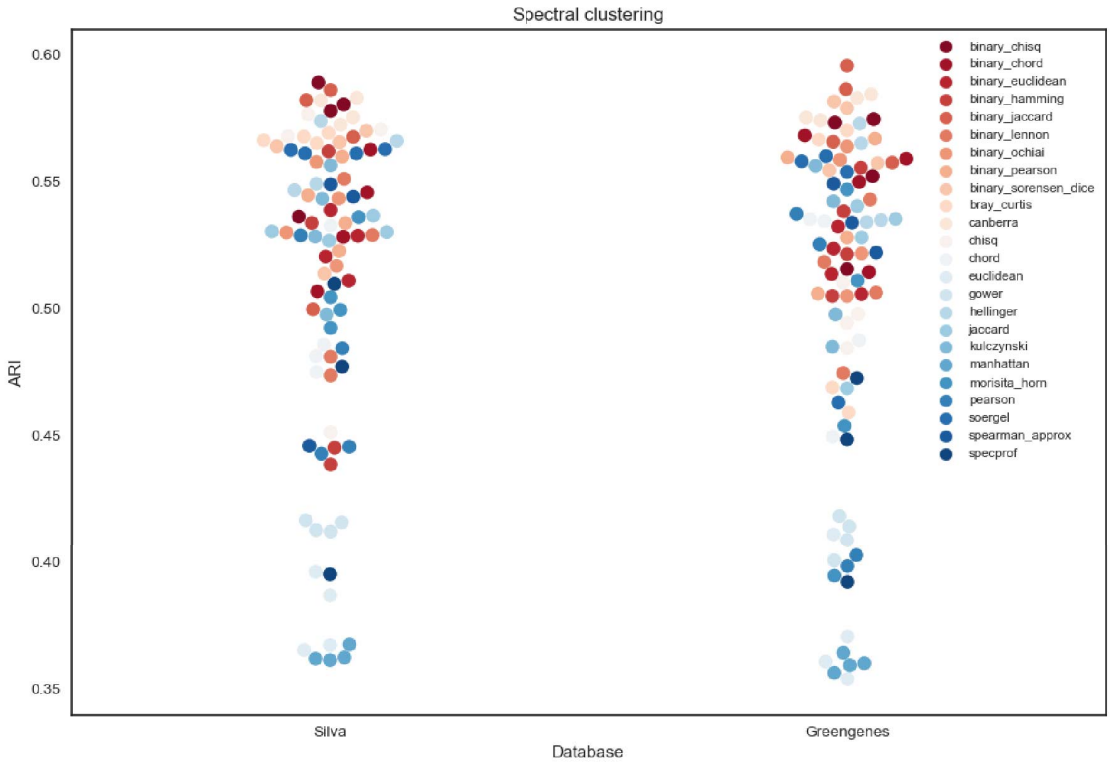

Figure 4. The ARI for the spectral clustering partitions based on different β-diversity measures (indicated by color) for four different thresholds (90-91%, 94%, 97% and 99%) separated for the Silva and Greengenes database. The aggregation of dots in the same color (i.e. the same β-diversity measure) indicates the stability of the clustering results regardless of the threshold similarity used.

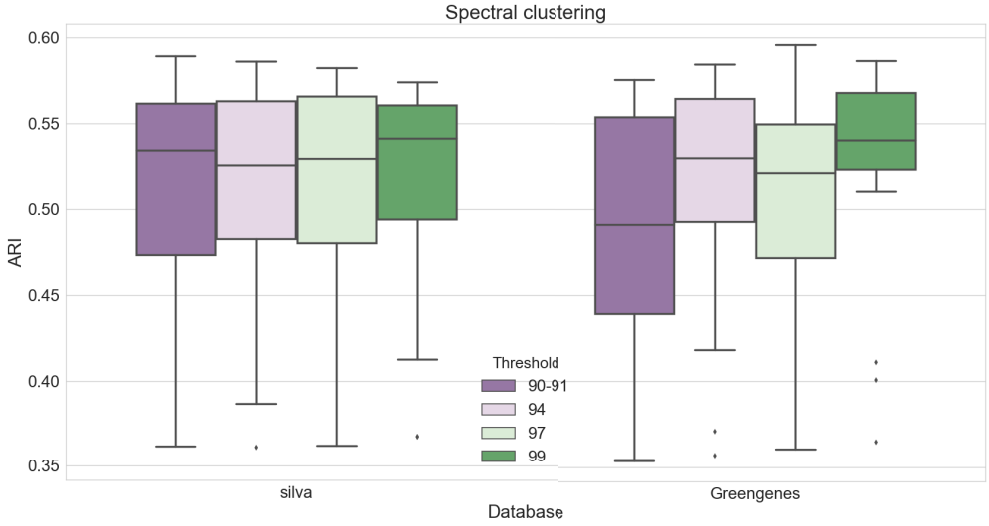

Figure 5. The boxplots of the ARI clustering evaluation for all β-diversity measures when different similarity thresholds and different databases are used.

[6] J. Peterson et al., "The NIH human microbiome project," *Genome research*, vol. 19, no. 12, pp. 2317–2323, Dec 2009, 19819907[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19819907

[7] D. Gevers, R. Knight, J. F. Petrosino, K. Huang, A. L. McGuire, B. W. Birren, K. E. Nelson, O. White, B. A. Methé, and C. Huttenhower, "The human microbiome project: a community resource for the healthy human microbiome," *PLoS biology*, vol. 10, no. 8, pp. e1 001 377–e1 001 377, 2012, 22904687[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22904687

[8] P. et al., "The integrative human microbiome project," *Nature*, vol. 569, no. 7758, pp. 641–648, 2019. [Online]. Available: https://doi.org/10.1038/s41586-019-1238-8

[9] "The human microbiome project website," https://www.hmpdacc.org/hmp/, accessed: 2019-07-14.

[10] J. A. Gilbert, J. K. Jansson, and R. Knight, "The earth microbiome project: successes and aspirations," *BMC Biology*, vol. 12, no. 1, p. 69, Aug 2014. [Online]. Available: https://doi.org/10.1186/s12915-014-0069-1

[11] L. R. Thompson et al., "A communal catalogue reveals earth's multiscale microbial diversity," *Nature*, vol. 551, pp. 457 EP –, Nov 2017, article. [Online]. Available: https://doi.org/10.1038/nature24621

[12] J. A. Gilbert, J. K. Jansson, and R. Knight, "Earth microbiome project and global systems biology," *mSystems*, vol. 3, no. 3, pp. e00 217–17, Apr 2018, 29657969[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29657969

[13] "The earth microbiome project website," http://www.earthmicrobiome.org/, accessed: 2019-07-14.

[14] T. M. Vogel, P. Simonet, J. K. Jansson, P. R. Hirsch, J. M. Tiedje, J. D. van Elsas, M. J. Bailey, R. Nalin, and L. Philippot, "Terragenome: a consortium for the sequencing of a soil metagenome," *Nature Reviews Microbiology*, vol. 7, pp. 252 EP –, Apr 2009, editorial. [Online]. Available: https://doi.org/10.1038/nrmicro2119

[15] "The terragenome: International soil metagenome sequencing consortium website," https://www.terragenome.org/, accessed: 2019-07-14.

[16] Y. He, J. G. Caporaso, X.-T. Jiang, H.-F. Sheng, S. M. Huse, J. R. Rideout, R. C. Edgar, E. Kopylova, W. A. Walters, R. Knight, and H.-W. Zhou, "Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity," *Microbiome*, vol. 3, no. 1, p. 20, May 2015. [Online]. Available: https://doi.org/10.1186/s40168-015-0081-x

[17] J. Kuczynski, Z. Liu, C. Lozupone, D. McDonald, N. Fierer, and R. Knight, "Microbial community resemblance methods differ in their ability to detect biologically relevant patterns," *Nature methods*, vol. 7, no. 10, pp. 813–819, Oct 2010, 20818378[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20818378

[18] P. Legendre and M. De Cáceres, "Beta diversity as the variance of community data: dissimilarity coefficients and partitioning," *Ecology Letters*, vol. 16, no. 8, pp. 951–963. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.12141

[19] D. H. Parks and R. G. Beiko, "Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities," *The ISME journal*, vol. 7, no. 1, pp. 173–183, Jan 2013, 22855211[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22855211

[20] O. Koren, D. Knights, A. Gonzalez, L. Waldron, N. Segata, R. Knight, C. Huttenhower, and R. E. Ley, "A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets," *PLOS Computational Biology*, vol. 9, no. 1, pp. 1–16, 01 2013. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1002863

[21] E. C. Dinleyici, D. Martínez-Martínez, A. Kara, A. Karbuz, N. Dalgic, O. Metin, A. S. Yazar, S. Guven, Z. Kurugol, O. Turel, M. Kucukkoc, O. Yasa, M. Eren, M. Ozen, J. M. Martí, C. P. Garay, Y. Vandenplas, and A. Moya, "Time series analysis of the microbiota of children suffering from acute infectious diarrhea and their recovery after treatment," *Frontiers in Microbiology*, vol. 9, p. 1230, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmicb.2018.01230

[22] B. J. Ridenhour, S. L. Brooker, J. E. Williams, J. T. Van Leuven, A. W. Miller, M. D. Dearing, and C. H. Remien, "Modeling time-series data from microbial communities," *The ISME journal*, vol. 11, no. 11, pp. 2526–2537, Nov 2017, 28786973[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28786973

[23] K. Faust, L. Lahti, D. Gonze, W. M. de Vos, and J. Raes, "Metagenomics meets time series analysis: unraveling microbial community dynamics," *Current Opinion in Microbiology*, vol. 25, pp. 56 – 66, 2015, environmental microbiology • Extremophiles. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1369527415000478

[24] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight, "Qiime allows analysis of high-throughput community sequencing data," *Nature Methods*, vol. 7, pp. 335 EP –, Apr 2010, correspondence. [Online]. Available: https://doi.org/10.1038/nmeth.f.303

[25] "Qiime™: Quantitative insights into microbial ecology," http://qiime.org/, accessed: 2019-07-10.

[26] A. Wilke, J. Bischof, T. Harrison, T. Brettin, M. D'Souza, W. Gerlach, H. Matthews, T. Paczian, J. Wilkening, E. M. Glass, N. Desai, and F. Meyer, "A restful api for accessing microbial community data for mg-rast," *PLOS Computational Biology*, vol. 11, no. 1, pp. 1–8, 01 2015. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1004008

[27] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen, "Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb," *Applied and Environmental Microbiology*, vol. 72, no. 7, pp. 5069–5072, 2006. [Online]. Available: https://aem.asm.org/content/72/7/5069

[28] "Greengenes 16s rrna gene database," https://greengenes.secondgenome.com/, accessed: 2019-07-11.

[29] D. McDonald, M. N. Price, J. Goodrich, E. P. Nawrocki, T. Z. DeSantis, A. Probst, G. L. Andersen, R. Knight, and P. Hugenholtz, "An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea," *The ISME journal*, vol. 6, no. 3, pp. 610–618, Mar 2012, 22134646[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22134646

[30] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, "The silva ribosomal rna gene database project: improved data processing and web-based tools," *Nucleic acids research*, vol. 41, no. D1, pp. D590–D596, 2012.

[31] "Silva: high quality ribosomal rna database," https://www.arb-silva.de/, accessed: 2019-07-11.

[32] M. Balvočiūtė and D. H. Huson, "Silva, rdp, greengenes, ncbi and ott — how do these taxonomies compare?" *BMC Genomics*, vol. 18, no. 2, p. 114, Mar 2017. [Online]. Available: https://doi.org/10.1186/s12864-017-3501-4

[33] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 08 2010. [Online]. Available: https://doi.org/10.1093/bioinformatics/btq461

[34] N. Mantel, "The detection of disease clustering and a generalized regression approach," *Cancer Research*, vol. 27, no. 2 Part 1, pp. 209–220, 1967. [Online]. Available: http://cancerres.aacrjournals.org/content/27/2_Part_1/209

[35] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, ser. NIPS'01. Cambridge, MA, USA: MIT Press, 2001, pp. 849–856. [Online]. Available: http://dl.acm.org/citation.cfm?id=2980539.2980649

[36] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec 1985. [Online]. Available: https://doi.org/10.1007/BF01908075

[37] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356

[38] S. Wagner and D. Wagner, "Comparing clusterings - an overview," Universität Karlsruhe, Karlsruhe, Tech. Rep. 4, 2007.

[39] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 09, p. P09008, 2005.

[40] S. Brdar and V. Crnojević, "Ensemble approaches for stable assessment of clusters in microbiome samples," in *Computational Intelligence Methods for Bioinformatics and Biostatistics. CIBB 2016. Lecture Notes in Computer Science*, A. Bracciali, G. Caravagna, D. Gilbert, and R. Tagliaferri, Eds., vol. 10477. Springer, Cham, 2017, pp. 199–208.

[41] P. Yang, X. Su, L. Ou-Yang, H.-N. Chua, X.-L. Li, and K. Ning, "Microbial community pattern detection in human body habitats via ensemble clustering framework," *BMC Systems Biology*, vol. 8, no. 4, p. S7, Dec 2014. [Online]. Available: https://doi.org/10.1186/1752-0509-8-S4-S7

[42] "Qiime 2™," https://qiime2.org/, accessed: 2019-07-10.