Uniform Manifold Approximation and Projection (UMAP) Reveals Composite Patterns and Resolves Visualization Artifacts in Microbiome Data

George Armstrong,[a,b,c] Cameron Martino,[a,b,c] Gibraan Rahman,[a,c] Antonio Gonzalez,[a] Yoshiki Vázquez-Baeza,[b] Gal Mishne,[d,e] Rob Knight[a,e,f]

[a]Department of Pediatrics, School of Medicine, University of California, San Diego, California, USA
[b]Center for Microbiome Innovation, Jacobs School of Engineering, University of California San Diego, La Jolla, California, USA
[c]Bioinformatics and Systems Biology Program, University of California, San Diego, California, USA
[d]Halıcıoğlu Data Science Institute, University of California, San Diego, La Jolla, California, USA
[e]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA
[f]Department of Bioengineering, University of California, San Diego, La Jolla, California, USA

ABSTRACT Microbiome data are sparse and high dimensional, so effective visualization of these data requires dimensionality reduction. To date, the most commonly used method for dimensionality reduction in the microbiome is calculation of between-sample microbial differences (beta diversity), followed by principal-coordinate analysis (PCoA). Uniform Manifold Approximation and Projection (UMAP) is an alternative method that can reduce the dimensionality of beta diversity distance matrices. Here, we demonstrate the benefits and limitations of using UMAP for dimensionality reduction on microbiome data. Using real data, we demonstrate that UMAP can improve the representation of clusters, especially when the clusters are composed of multiple subgroups. Additionally, we show that UMAP provides improved correlation of biological variation along a gradient with a reduced number of coordinates of the resulting embedding. Finally, we provide parameter recommendations that emphasize the preservation of global geometry. We therefore conclude that UMAP should be routinely used as a complementary visualization method for microbiome beta diversity studies.

IMPORTANCE UMAP provides an additional method to visualize microbiome data. The method is extensible to any beta diversity metric used with PCoA, and our results demonstrate that UMAP can indeed improve visualization quality and correspondence with biological and technical variables of interest. The software to perform this analysis is available under an open-source license and can be obtained at https://github.com/knightlab-analyses/umap-microbiome-benchmarking; additionally, we have provided a QIIME 2 plugin for UMAP at https://github.com/biocore/q2-umap.

KEYWORDS beta diversity, dimensionality reduction

An important step in microbiome research is visualizing the relationships between samples. In the study of microbial communities through next-generation sequencing (NGS), these comparisons are typically done through the visualization of beta diversities with principal-coordinate analysis (PCoA) (1) (see Fig. S1 in the supplemental material). Although alternatives such as conventional principal-component analysis (PCA), nonmetric multidimensional scaling (NMDS) (2), and t-distributed stochastic neighbor embedding (t-SNE) (3) are sometimes applied, PCoA in particular has been widely adopted by the microbiome community. Due to the high-dimensional and highly sparse nature of the data, which presents challenges on sequence count data (4, 5), one major benefit of PCoA over other methods on untransformed count data is that it

accommodates a generalized distance matrix (of beta diversities, for the microbiome). This allows use of distance metrics that are better suited for sparse data (e.g., Bray-Curtis [6], Jaccard [7], and UniFrac [8]).

Uniform Manifold Approximation and Projection (UMAP) (9) is a method that has gained traction in single-cell genomics analysis (10). Whereas PCoA performs an eigen-decomposition that focuses on linearly preserving the pairwise distances between the samples (global structure), UMAP uses a nonlinear graph construction and embedding method to optimize an objective that allows for a tradeoff between emphasizing local structures and preserving distances globally. This tradeoff is primarily controlled by the 'n_neighbors' and 'min_dist' parameters of UMAP. The 'n_neighbors' parameter controls the number of neighbors whose local topology is preserved, so global distances are preserved when it is high. The 'min_dist' parameter controls the minimum distance between samples in the embedding, which affects the spread of clusters. Low values of 'min_dist' allow UMAP to emphasize the similarity of dense clusters of samples, whereas larger values of 'min_dist' will focus on preserving distances more broadly.

Both UMAP and PCoA operate on a generalized distance (beta diversity) matrix, appropriate for microbiome data (Fig. S1). While the use of UMAP on microbiome data has been noted (11), the utility of UMAP on microbiome data remains underexplored. Using real data sets, we compared both visual qualities and quantitative measures of UMAP to those of PCoA on well-understood data sets. We additionally applied UMAP to data from the Human Microbiome Project (HMP) (12) to demonstrate its characteristics on a larger data set with more complex sources of variation.

Discrete clusters are one common pattern that microbial communities can exhibit (13). The "keyboard data" from reference 14 contain 16S samples (99 samples, 1,399 features, 5% dense) from the keyboards and fingers of 3 subjects. PCoA on the Aitchison distances on these samples can recover the cluster structure of the subjects in the data (Fig. 1a). We compared this to UMAP (n_neighbors = 15 and n_neighbors = 80, min_dist = 1) and found that UMAP can also recover the cluster structure of the subjects (Fig. 1b and c). We also saw that UMAP produced two-dimensional coordinates with improved separation within subjects by sample type. To quantitatively assess the dimensionality reduction, we performed a supervised classification with linear discriminant analysis (LDA) as well as an unsupervised evaluation of clustering using the silhouette measure on the low-dimensional representations. The LDA classification, which solely measures separability, demonstrated higher accuracy of sample type (stratified by subject) on UMAP with two components compared to PCoA with two or three components for all subjects (Table S1). Silhouette scores (15), which measure cluster separation and density, demonstrated that host separation is improved with UMAP with a low 'n_neighbors' value, but not for a higher 'n_neighbors' value, which is likely due to the reduced distance between clusters in the UMAP coordinates with higher 'n_neighbors'. The method with the highest within-host sample-type silhouette varied for each host. A simulated missing data analysis, where entries were randomly masked from samples, demonstrated that these results are sensitive to missing values (Fig. S3).

In dimensionality reduction, it is not only important for clusters to be separated; the positioning of clusters with respect to their similarity to other clusters, i.e., preserving global distances, is desirable. In the PCoA visualization (Fig. 1a), the samples of subjects M3 and M9 are similar to each other in the plot, and both are distant from M2. This corresponds with the expectation that M3 and M9 are more similar, because they shared an office. Additionally, this agrees with the original distances, where the mean Aitchison distance between M3 and M9 samples is 13.87 $\pm$ 0.11 (95% confidence interval [CI]), whereas the mean M2-M3 distance is 19.89 $\pm$ 0.11 (95% CI), and the mean M2-M9 distance is 18.94 $\pm$ 0.12 (95% CI). However, for UMAP with n_neighbors = 15 in Fig. 1b, the relative position of the clusters has changed (M9 is closer to M2 than it is to M3). Using the default 'spectral' initialization option, which is recommended for preserving global structure (16), we found that on only 34/50 initializations with different random seeds and n_neighbors = 15, UMAP produced clusters with the correct relative positioning.
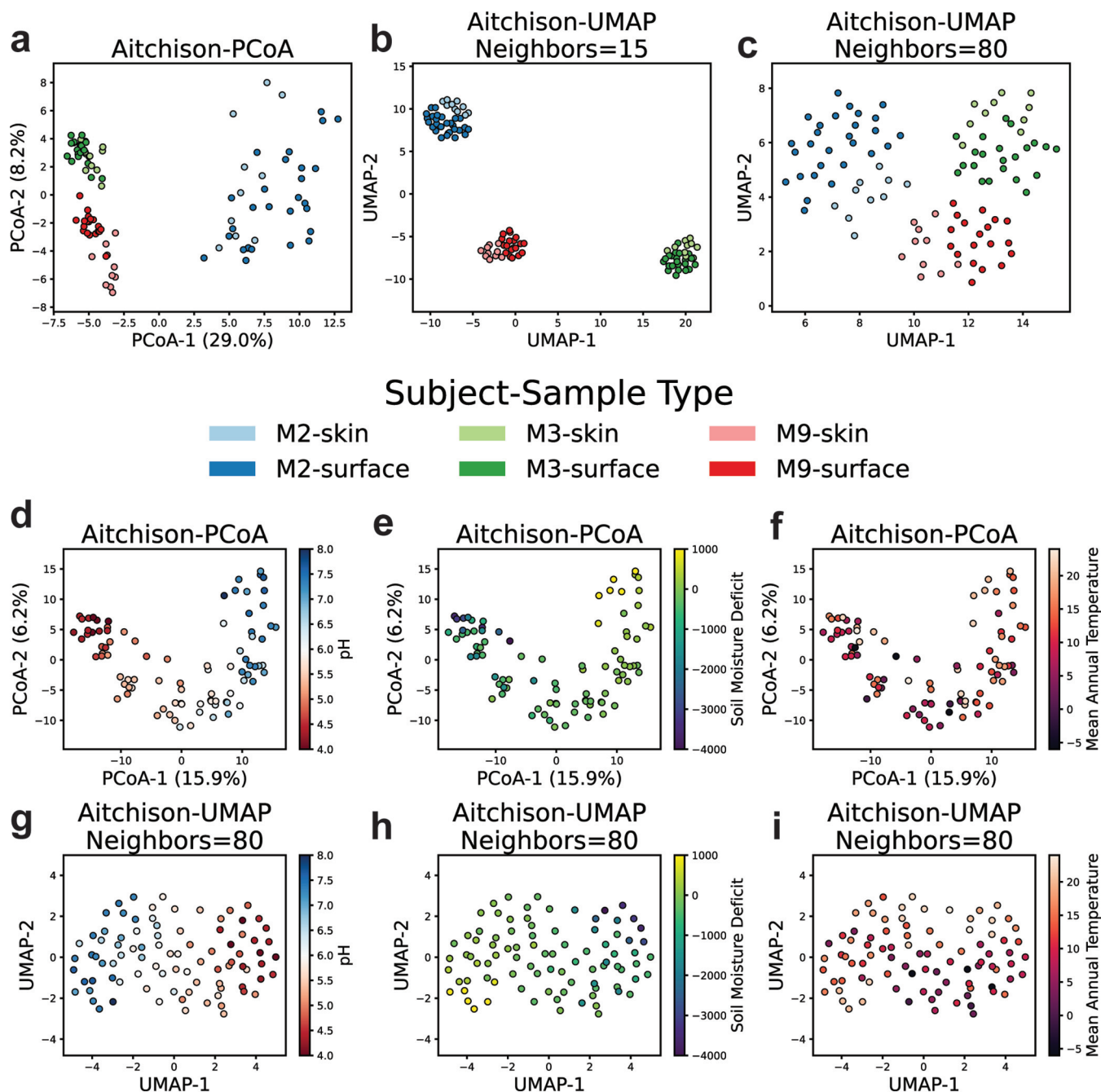
FIG 1 Comparison of PCoA and UMAP visualizations of cluster and gradient patterns on real data. The keyboard data set contains samples from three different subjects' keyboards (surface) and their hands (skin). (a) PCoA on Aitchison distances (pseudocount = 1) demonstrates a strong separation between M2 and the other subjects, as well as separation between subjects M3 and M9. (b) A UMAP (n_neighbors = 15, min_dist = 1) visualization demonstrates stronger clustering by subject, with a different relative positioning of the clusters by subject. The plot also emphasizes clustering by sample type. (c) UMAP with an increased n_neighbors parameter (n_neighbors = 80, min_dist = 1) reflects the same relative positioning of clusters as PCoA. It also demonstrates the improved localization by sample type within subjects. (d) On the "88 soils" data, PCoA on the Aitchison distances demonstrates a horseshoe pattern with pH distributed along the horseshoe. (e) Soil moisture deficit is also distributed along the horseshoe, and (f) there is not a strong association between mean annual temperature and position on the PCoA. (g) In the UMAP (n_neighbors = 80, min_dist = 1), followed by centering/rotation with PCA, using the same distances, pH appears correlated with the first coordinate, (h) soil moisture deficit appears correlated with a sloped line across the pH gradient, and (i) there is a correlation between mean annual temperature and the second coordinate.

However, when we increase the parameter to n_neighbors = 80, which represents a large majority of the samples, the visualization retains separation by subject (Fig. 1c), and 50/50 initializations produced clusters with the correct relative positioning.

Ecological gradients are another common pattern that microbial communities can exhibit (13). The "88 soils" data from reference 17 contain 16S samples (88 samples,

5,628 features, 4% dense) from 88 different soils with additional measurements of the soil. A Bio-Env test (18) reveals that the top three soil variates corresponding with the Aitchison distances are pH, moisture deficit, and mean annual temperature (Table S2). In the PCoA of the Aitchison distances, which displays a horseshoe artifact (19, 20), pH is distributed along the horseshoe (Fig. 1d). To quantitatively assess the visualization of gradients in the data, similarly to reference 13, we calculated the Spearman correlation of the components of the ordination with the ecological variable. We found that soil pH is strongly correlated with the first component (Spearman $r = 0.934$) (Table S3). Soil moisture deficit is also distributed along the horseshoe (Fig. 1e), with PCoA-1 (Spearman $r = 0.828$). There is a mild correlation between mean annual temperature and the second PCoA coordinate (Spearman $r = 0.313$), although a pattern is difficult to see visually due to the horseshoe artifact (Fig. 1f).

On the gradient problem, we fit UMAP with the parameters used with the keyboard data (min_dist = 1, n_neighbors = 80). Since the UMAP algorithm does not guarantee that the direction with the most variance in its output coordinates is axis aligned, we use PCA to identify the direction of maximum variance in the UMAP embedding and rotate the UMAP coordinates so that this direction is aligned with the x axis. The visualization shows reduced horseshoe-like warping, in contrast to the PCoA (Fig. 1g). Additionally, the pH gradient is highly correlated with the first principal component of the embedding (Spearman $r = -0.931$). Furthermore, the soil moisture deficit is displayed clearly across the diagonal of the embedding (Fig. 1h) and is correlated with both components of the axes (Table S3). Finally, the mean annual temperature has a much clearer association in two-dimensional UMAP coordinates compared to the first two components of PCoA, with a higher Spearman correlation with the second component ($r = 0.478$ for n_neighbors = 80, $r = -0.604$ for n_neighbors = 87). PCoA exhibits maximum Spearman correlation with mean annual temperature in its third component ($r = -0.567$). So, while a single axis of PCoA may be more correlated with the gradient, UMAP is able to display each of the gradients in fewer dimensions.

Next, we compared PCoA and UMAP on data from the HMP (8,280 samples, 13,318 features, 0.08% dense) (12). These samples are from various body sites and individuals, with a large portion of samples processed with primers for two different variable regions of 16S. As noted in reference 21, the PCoA on unweighted UniFrac distances shows that differences in primers are not visible in the first two coordinates (Fig. 2a). Localization by body sites, however, is more apparent (Fig. 2b). Clustering by primer is instead visible in the third component of the PCoA (Fig. S2a), where clustering by body site is also apparent (Fig. S2b). We also fit a two-dimensional UMAP (min_dist = 1, n_neighbors = 800) to the same data. UMAP is able to separate a majority of the samples by variable region (Fig. 2c) and produces more distinct clusters by body site.

To quantify the clustering in the HMP data, we trained a k-Nearest Neighbors (kNN) classifier on the respective variables with 10-fold cross validation and reported the mean accuracy on the test folds. We trained kNN models on the first one, two, and three components of the PCoA and fit UMAP embeddings for the respective number of dimensions. We found that kNN on a one-dimensional UMAP can outperform the sample site kNN for PCoA on up to 3 dimensions (Table S4). kNN trained on a two-dimensional UMAP was able to distinguish primers more accurately than kNN on the first two principal coordinates. This indicates that UMAP is capable of representing multiple sources of variability in microbiome data sets with thousands of samples more distinctly and in fewer dimensions than PCoA.

Finally, we explored a general-purpose recommendation for parameters. The parameters in this study were chosen to emphasize preserving the global structure of the data, by setting the 'min_dist' to its maximum of 1, increasing 'n_neighbors' from its default, and using default values for the rest of the parameters. In accordance with this goal, we set 'n_neighbors' to its maximum ($n - 1$ in general, 98 for soils, 87 for keyboard, and 8,279 for the HMP) and reran the previous analyses. With this parameter setting, the results remain largely unchanged (Table S4).
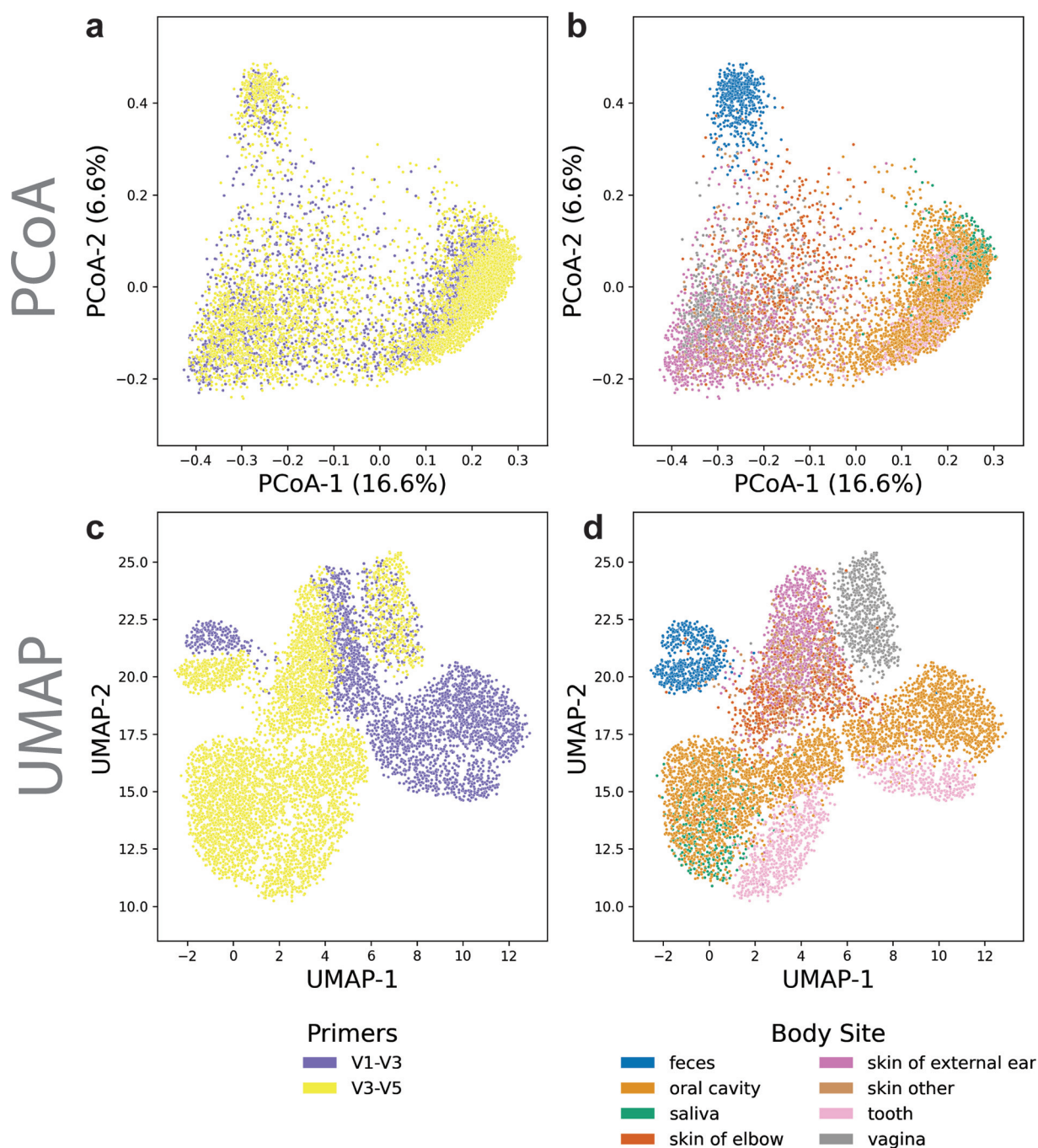
FIG 2 PCoA and UMAP comparison on 8,280 samples from the Human Microbiome Project (HMP). In the HMP data, when samples prepared with different primers are analyzed jointly, (a) there appears to be no separation between primers in the first two coordinates of PCoA and (b) mild separation by body site. In the same number of dimensions, UMAP is able to both (c) emphasize the differences between samples prepared with different variable regions and (d) improve clustering by body site. Both methods use the unweighted UniFrac distances on the HMP data rarefied to 1,000 sequences per sample.

Our benchmarks demonstrate the potential for improved performance and interpretability for both cluster and gradient microbiome data by using UMAP with its parameters set with the intent to preserve global geometry. Given that the two algorithms provide different guarantees with respect to the preservation of distances in embeddings, we conclude that UMAP should be routinely used for microbiome analyses as a complement to PCoA. In order to facilitate using UMAP, we have made it conveniently available via QIIME2 (22) and Qiita (23) plugins.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TEXT S1, DOCX file, 0.01 MB.
FIG S1, TIF file, 2.6 MB.
FIG S2, TIF file, 2.6 MB.
FIG S3, TIF file, 2.6 MB.
TABLE S1, XLSX file, 0.01 MB.
TABLE S2, XLSX file, 0.01 MB.
TABLE S3, XLSX file, 0.01 MB.
TABLE S4, XLSX file, 0.01 MB.

ACKNOWLEDGMENTS

REFERENCES

1. Kruskal JB, Wish M. 1978. Multidimensional scaling. SAGE Publications, Newbury Park, CA.
2. Kruskal JB. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1–27. https://doi.org/10.1007/BF02289565.
3. van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605.
4. Aitchison J. 1983. Principal component analysis of compositional data. Biometrika 70:57–65. https://doi.org/10.1093/biomet/70.1.57.
5. Martino C, Morton JT, Marotz CA, Thompson LR, Tripathi A, Knight R, Zengler K. 2019. A novel sparse compositional technique reveals microbial perturbations. mSystems 4:e00016-19. https://doi.org/10.1128/mSystems.00016-19.
6. Bray JR, Curtis JT. 1957. An ordination of the upland forest communities of Southern Wisconsin. Ecol Monogr 27:325–349. https://doi.org/10.2307/1942268.
7. Jaccard P. 1912. The distribution of the flora in the alpine zone. New Phytol 11:37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.
8. Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71:8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005.
9. McInnes L, Healy J, Melville J. 2018. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. arXiv https://arxiv.org/abs/1802.03426.
10. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zagar M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LB, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. 2020. Integrated analysis of multimodal single-cell data. bioRxiv. https://doi.org/10.1101/2020.10.12.335331.
11. Parker N. 2020. Visualizing high-dimensional microbiome data. Towards Data Science. https://towardsdatascience.com/visualizing-high-dimensional-microbiome-data-eacf02526c3a.
12. The Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214. https://doi.org/10.1038/nature11234.
13. Kuczynski J, Liu Z, Lozupone C, McDonald D, Fierer N, Knight R. 2010. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nat Methods 7:813–819. https://doi.org/10.1038/nmeth.1499.
14. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. 2010. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A 107:6477–6481. https://doi.org/10.1073/pnas.1000162107.
15. Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
16. Kobak D, Linderman GC. 2021. Initialization is critical for preserving global data structure in both t-SNE and UMAP. Nat Biotechnol 39:156–157. https://doi.org/10.1038/s41587-020-00809-z.
17. Lauber CL, Hamady M, Knight R, Fierer N. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol 75:5111–5120. https://doi.org/10.1128/AEM.00335-09.
18. Clarke KR, Ainsworth M. 1993. A method of linking multivariate community structure to environmental variables. Mar Ecol Prog Ser 92:205–219. https://doi.org/10.3354/meps092205.
19. Morton JT, Toran L, Edlund A, Metcalf JL, Lauber C, Knight R. 2017. Uncovering the horseshoe effect in microbial analyses. mSystems 2:e00166-16. https://doi.org/10.1128/mSystems.00166-16.
20. Diaconis P, Goel S, Holmes S. 2008. Horseshoes in multidimensional scaling and local kernel methods. Ann Appl Stat 2:777–807. https://doi.org/10.1214/08-AOAS165.
21. Debelius J, Song SJ, Vazquez-Baeza Y, Xu ZZ, Gonzalez A, Knight R. 2016. Tiny microbes, enormous impacts: what matters in gut microbiome studies? Genome Biol 17:217. https://doi.org/10.1186/s13059-016-1086-x.
22. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Titus BC, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857. https://doi.org/10.1038/s41587-019-0209-9.
23. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB, Sanders JG, Shorenstein J, Holste H, Petrus S, Robbins-Pianka A, Brislawn CJ, Wang M, Rideout JR, Bolyen E, Dillon M, Caporaso JG, Dorrestein PC, Knight R. 2018. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 15:796–798. https://doi.org/10.1038/s41592-018-0141-9.