NAR Genomics and Bioinformatics, 2024, **6**, 1–5
https://doi.org/10.1093/nargab/lqad110
Advance access publication date: 5 January 2024
Methods Article

Data-driven mathematical and visualization approaches for removing rare features for Compositional Data Analysis (CoDA)

Adrian Ortiz-Velez[1,2] **and Scott T. Kelley**[1,2,*,†]

[1]Biological and Medical Informatics Program, San Diego State University, San Diego, CA 92182, USA
[2]Department of Biology, San Diego State University, San Diego, CA 92182, USA

[*]To whom correspondence should be addressed. Tel: +1 619 206 8014; Fax: +1 619 594 5676; Email: skelley@sdsu.edu
[†]Membership list can be found in the Acknowledgments section.
Present address: Scott T. Kelley, Department of Biology, San Diego State University, San Diego, CA, USA.

Abstract

Sparse feature tables, in which many features are present in very few samples, are common in big biological data (e.g. metagenomics). Ignoring issues of zero-laden datasets can result in biased statistical estimates and decreased power in downstream analyses. Zeros are also a particular issue for compositional data analysis using log-ratios since the log of zero is undefined. Researchers typically deal with this issue by removing low frequency features, but the thresholds for removal differ markedly between studies with little or no justification. Here, we present CurvCut, an unsupervised data-driven approach with human confirmation for rare-feature removal. CurvCut implements two distinct approaches for determining natural breaks in the feature distributions: a method based on curvature analysis borrowed from thermodynamics and the Fisher-Jenks statistical method. Our results show that CurvCut rapidly identifies data-specific breaks in these distributions that can be used as cutoff points for low-frequency feature removal that maximizes feature retention. We show that CurvCut works across different biological data types and rapidly generates clear visual results that allow researchers to confirm and apply feature removal cutoffs to individual datasets.

Introduction

Advancement in next-generation sequencing (NGS) technology has made it possible to detect thousands of species, genes, transcripts, or polymorphisms in samples (1). This results in sparse feature tables dominated by zeros because many of the features are detected in only a few samples. These zero-laden datasets can cause overdispersion and decrease power in downstream analyses (2,3) or result in biased statistical estimates (4).

Zero-laden datasets can be problematic for many types of statistical data analysis (4) including compositional data analysis (CoDA) methods (5,6). CoDA methods, originally developed for the geological sciences, have been increasingly applied to multiomics analyses (e.g., marker-gene libraries, metagenomics, and metatranscriptomics) due to their unavoidable compositionality (6–8). CoDA approaches involve a log-ratio transformation (e.g., centered log-ratio, isometric log-ratio) and require replacement of all zeros in feature tables since the log of zero is undefined. Zero-replacement methods have been developed for this purpose (5,6), but these methods do not work well with highly sparse data. To overcome this limitation, researchers typically remove low-frequency features prior to zero-replacement. However, the process of feature removal is usually accomplished by setting an arbitrary threshold of percent presence among samples (i.e., only retain features present in at least 10% of samples), and there appears to be no rule or consistent approach across studies for identifying that threshold. Thresholds for feature removal have been set at 10% (9), ≥1% in at least one sample (10), 85% (11,12) or not reported or ignored (13–16). While setting a percentage threshold for feature retention seems appealing, such a process does not consider the sparsity of individual datasets.

Here, we present mathematical approaches for selecting dataset-specific feature removal cutoffs that maximize feature retention by detecting natural breaks in histogram distributions of dataset feature sparsity. The first method (CurvCut) uses curvature analysis from thermodynamics to detect discontinuities on the histogram with a sharp change in the characteristics of the distribution, while the second method uses the Fisher-Jenks statistical approach for detecting natural breaks in distributions. We tested our approach using four different NGS feature tables generated from small subunit ribosomal RNA (16S rRNA) amplicon datasets, a shotgun metagenomics dataset, and a single nucleotide polymorphism dataset. Our results show both the curvature approach and the Fisher-Jenks methods provide data-driven feature removal recommendations that consider the unique feature distribution of a given dataset.

Materials and methods

The first approach we describe in detail is a curvature analysis method for identifying regime change that is used in thermodynamics to identify the point where a fluid changes behavior. We are translating this approach to allow determination of

Received: March 18, 2023. **Revised:** November 25, 2023. **Editorial Decision:** December 11, 2023. **Accepted:** December 20, 2023

where the histogram curve changes behavior. In essence, we are identifying where the histogram becomes more uniform. The second method, known as Fisher-Jenks, is a statistical approach for identifying 'natural breaks' in distributions based on variance minimization.

Model derivation

To create a mathematical model based on curvature analysis to detect the regime change in the zero-count distribution, we visualized the problem as a ball rolling down a hill of the feature distribution. In our model, the radius of the ball decreases proportionally to the height of the features (as it rolls down the hill) until it reaches a minimum at the point of 'regime change' when it shifts between characteristics of distributions (Figure 1A). The radius of the ball is inversely proportional to the change of the tangents to the line from one point to the next. This means that when the line is smooth, the change in the tangents is small and the radius is large (big circle). When the line becomes less smooth, i.e. more curved, the change in the tangents is greater and the radius is small (small circle). The point when the ball is the smallest is when the line is the most curved, aka. the point of regime change.

The first step in our approach is to create an accumulated zero count from a feature histogram. This is done because accumulation is characteristically monotonically increasing, so it is easier to look for a maximal change in the curvature across the log-transformed zero count cumulative mass function. First, the zero-count cumulative mass is calculated as

$$\mathrm{M} = [H_1, \sum_i^2 H_i, \sum_i^3 H_i, ..., \sum_i^n H_i], \tag{1}$$

where H is the histogram array, and n is the number of features. Next, the data are log-transformed to maximize changes more than an order of magnitude and minimize changes less than an order of magnitude across the cumulative mass function. The log transform equation is

$$\mathrm{F} = log(M + k), \tag{2}$$

where F is the log-transformed array of M and k is a constant (default k: 100) used to minimize small changes across M. To perform the final curvature analysis, we used CubicSplines from SciPy (17) to create a continuous piecewise polynomial function from the discontinuous histogram array

$$\mathrm{F_{cs}} = CubicSplines(F), \tag{3}$$

where F[cs] is the continuous piecewise function. To find the curvature across the cubic spline, we implemented the curvature equation

$$\kappa = \frac{|F''_{cs}|}{(1 + F'^2_{cs})^{\frac{3}{2}}}, \tag{4}$$

where K is the curvature and F"cs is the second derivative, and F' is the first derivative of the cubic spline Fcs. This has many local maxima so we identify the last maxima where

$$\kappa' = 0, \tag{5}$$

and we plot for the histogram C with the diagnosed cut-off for the user discernment (4).

We also implemented the Fisher-Jenks statistical approach for identifying natural breaks in distributions (Figure 1B). Fisher-Jenks is an unordered grouping of the magnitude of each position of the histogram that operates effectively as a one-dimensional k-means method that uses an iterative approach to find the best grouping of numbers that minimizes in-group variance while maximizing between-group variance between a pre-selected number of groups (k). We set k equal to 2 to detect and separate potential abundant zero-laden features for removal. While the Fisher-Jenks algorithm has been around for many years, this is the first time it has been applied to zero-laden feature removal.

Implementation and datasets

We implemented our curvature analysis method, called CurvCut, as a command line Python program (https://github.com/aortizsax/curvcut) and tested it on five datasets: three 16S datasets, one metagenomic count table dataset, and one HIV site frequency spectrum (SFS) dataset. The program was implemented in Python 3.7 using Python packages Pandas 1.3.3 (18), Numpy 1.19.2 (19), Matplotlib 3.3.4 (20), SciPy 1.7.1 (17) and jenkspy 0.3.3 (21). Scripts and the test datasets analyzed in this paper can be found at github.com/aortizsax/curvcut. Two datasets, a 16S and metagenomic count tables, come from a periodontal study (22). The raw reads for the periodontal 16S rRNA sequences, and metagenomic OTUs classified by Kraken (23), were published previously (22,24,25). The second 16S dataset comes from an unpublished study. The third 16S SV dataset comes from a built environment (BE) study (26). The HIV histogram was created by making an SFS graph of a multiple sequence alignment of Pol genetic sequence data with 100 sequences and 4339 bases long collected from NCBI BLASTN (27–36). In this histogram, the heights of the bars represent the number of polymorphisms detected in one sequence (singletons), two sequences (doubletons), etc (polytons).

Results and discussion

Our results show that our data-driven modeling approaches identify points of regime change across various dataset types. CurvCut rapidly suggests a cutoff based on the distribution of features present rather than on an arbitrary cutoff that does not consider the characteristics of the data. Both the curvature analysis and the Fisher-Jenks method rapidly identify cutoff values, and for most datasets tested the results were very similar (Figure 2). After curvature analysis, the recommended cutoff removes those features that could contribute to overdispersion in downstream analyses. The cutoff value is data-driven in the sense that the cutoff is entirely dependent on the feature distribution. The recommended cutoff value differs by dataset, as expected by the clear differences in feature distributions between datasets (Figure 2). While most of the cutoffs suggested by our analysis were features that were in very few samples (5 or fewer), there was one dataset in which the curvature approach determined a very high feature cutoff recommendation (features in 46 samples or less; Figure 2B). However, for this same dataset, the Fisher-Jenks method recommended a cutoff of 3 or fewer samples, which seems more reasonable for this dataset. The importance of the cutoff being data-driven is very apparent when considering what would happen if the same cutoff was used for all the data sets in Figure 2. For example, a cutoff of features present in 3 or fewer samples would be appropriate for the allelic dataset (Figure 2 D) but would leave

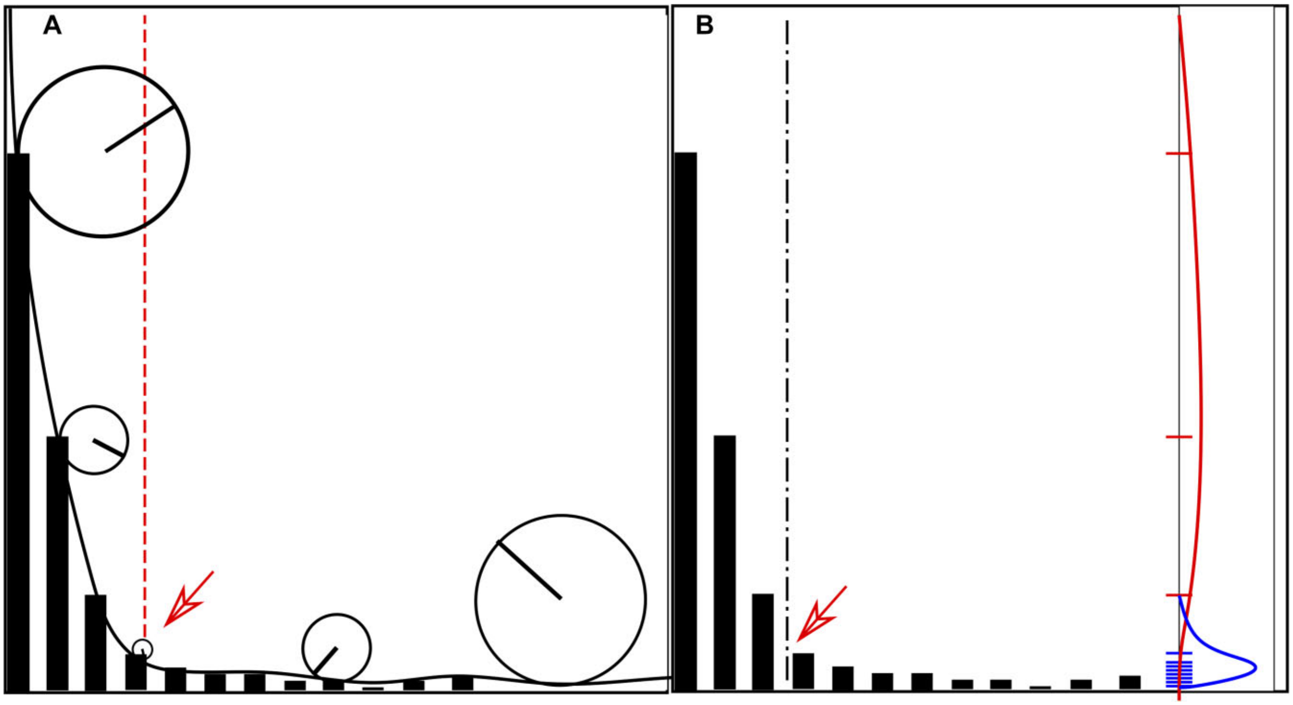

Figure 1. Artistic representation of the mathematical models. The histogram represents a hypothetical plot of features present per sample. The heights of the bars indicate the number of features (e.g., sequence variants, genes, single nucleotide polymorphisms) present in X samples. For example, the leftmost bar on the histogram represents features present in only one sample. The red arrows indicate the minimum feature cutoff recommendation (**A**) To detect the regime change, we visualized the problem as a ball rolling down the hill of features. The radius of the ball decreases proportionally to the change in the height of the features until it reaches a minimum at the point of regime change when the path of the ball reaches the maximum curvature (i.e. when the ball is the smallest). Then, after the regime change, the ball increases again proportional to the lack of change in the featur heights as they reach steady values. (**B**) We also implemented the Fisher-Jenks method, which uses an iterative *k*-means approach to find number groupings that maximize between group variance. The vertical red line indicates the feature trimming cutoff based on our curvature analysis, while the vertical black line indicates the feature trimming cutoff based on the Fisher-Jenks method. The vertical red and blue curves on the right indicate the feature groupings determined via Fisher-Jenks.

many zeros in the 16S and metagenomics datasets (Figure 2A and B).

A closer look at the trimmed sequence found many of them to be spurious. For example, the periodontal 16S dataset analyzed in Figure 2A trimmed two unidentified species of *Treponema*, most of which were missing in many or all the samples, while an identified species, *Treponema socranskii*, was identified in all samples. While most of the distributions were right-skewed, our approach also worked on a left-skewed metagenomics dataset (Supplementary Figure S1), which shows that the curvature method can detect these regime changes at either end of the distribution. A closer look at this metagenomic dataset found that the k-mer-based algorithm used to determine the number of species per sample, Kraken, identified certain species very readily in all the samples but also made many seemingly spurious identifications of closely related species or strains. For example, Kraken identified *Campylobacter gracilis* in every sample, with counts ranging between 900 and 330 000 (Avg. = 64 000). However, Kraken also identified 15 other species of *Campylobacter*, most of which were missing in many or all the samples. This helps explain the leftward skew of this distribution and suggests that a cutoff of features present in 45 or fewer samples would remove many spurious results. While most of the datasets were sequence count tables, our approach also worked on an SFS histogram from an HIV dataset (Figure 2D), which shows that our method can detect these regime changes in many types of datasets and is not exclusive to count tables. A closer look at the SFS dataset analyzed in Figure 2D found our analysis trimmed 661 and 266 of singletons and doubletons, removing possible noise and limiting overdispersion in downstream analyses.

Comparisons of the curvature method to the Fisher-Jenks approach found them to be very similar for most of these datasets. The one exception was the 16S dataset in Figure 2B, where the curvature analysis suggested a cutoff of 46, while Fisher-Jenks indicated a cutoff of three samples or fewer, which seems more in keeping with the other results. However, the curvature analysis appeared to find a clearer distribution break with the periodontal 16S dataset (Figure 2A). Having two methods allows users an additional unsupervised option for choosing cutoffs that remove zero-laden features while maximally retaining features.

In general, we suspect both methods work best with right- or left-skewed features distributions, i.e. in datasets in which many features are present in only a few samples, or the opposite where many features are in many samples. This is usually the case with metagenomics, metatranscriptomics, and allelic datasets. However, some datasets might have a bimodal distribution of features that does not conform to this clear regime pattern. We have also observed histograms that are both right- and left-skewed (data not shown), with many features present in very few samples and many others present in the majority or all of the samples. Thus, the graphical output produced

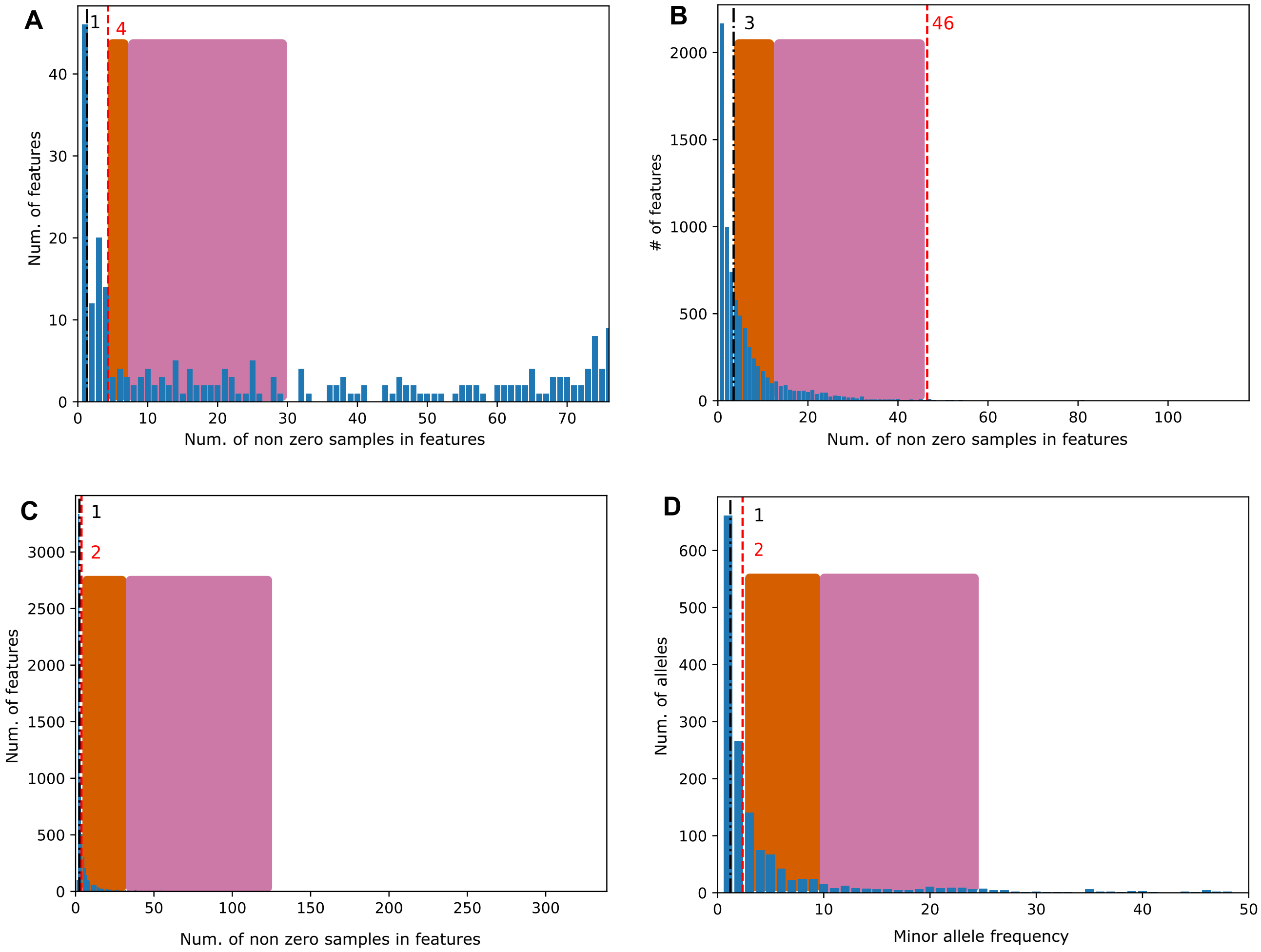

Figure 2. Curvature analysis of four datasets. The vertical red lines indicate the feature trimming cutoff based on our curvature analysis, and the vertical black lines indicate the feature trimming cutoff based on the Fisher-Jenk analysis. The pink background shows the range of current heuristic cutoffs 10-40%. The orange shows the minimum number of features that would be lost with current heuristics (10%). (**A**) 16S periodontal data contained 76 samples of 247 OTU features, (**B**) 16S unpublished data containing 118 samples of 5650 OTU features, (**C**) 16S built environment data containing 338 samples of 6467 SV features, and (**D**) HIV SFS data generated from an MSA of 100 Pol protein sequences, 4339 bases long.

by CurvCut also allows the user to visualize the cutoff value on the feature distribution to make an informed choice for an appropriate cutoff. Indeed, the graphical visualization of the feature histogram can point out methodological artifacts (e.g. the Kraken approach) that a blind reliance on a percentage cutoff would ignore. Our method can be easily integrated into common pipelines (e.g. QIIME2 (37) or mothur (38)) or run separately on datasets before further analysis. In addition to Fisher-Jenks, there are other more recently developed methods for one-dimensional clustering of data, such as Jiang's head/tail breaks method (39) and several algorithms implemented in the R package classInt (40) that could be implemented in a future version of CurvCut.

Data availability

The CurvCut Programming code, installation instructions, datasets used in this paper are available at https://doi.org/10.5281/zenodo.10366078.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We thank members of the Kelley Lab for testing CurvCut on numerous datasets, especially L Sisk-Hackworth. We also thank A. Sethuraman for preparing the SFS data, and V. Thackray for comments on the manuscript.

Conflict of interest statement

None declared.

References

1. Levy,S.E. and Myers,R.M. (2016) Advancements in next-feneration sequencing. *Annu. Rev. Genomics Hum. Genet.*, **17**, 95–115.
2. Van den Berge,K., Perraudeau,F., Soneson,C., Love,M.I., Risso,D., Vert,J.-P., Robinson,M.D., Dudoit,S. and Clement,L. (2018) Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.*, **19**, 24.
3. Calle,M.L. (2019) Statistical analysis of metagenomics data. *Genomics Inform*, **17**, e6.
4. Greenland,S., Mansournia,M.A. and Altman,D.G. (2016) Sparse data bias: a problem hiding in plain sight. *BMJ*, **352**, i1981.

5. Palarea-Albaladejo,J. and Martín-Fernández,J.A. (2015) zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometr. Intell. Lab. Syst.*, **143**, 85–96.
6. Erb,I., Gloor,G.B. and Quinn,T.P. (2020) Editorial: compositional data analysis and related methods applied to genomics—a first special issue from NAR Genomics and Bioinformatics. *NAR Genomics Bioinform.*, **2**, lqaa103.
7. Gloor,G.B., Macklaim,J.M., Pawlowsky-Glahn,V. and Egozcue,J.J. (2017) Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.*, **8**, 2224.
8. Sisk-Hackworth,L. and Kelley,S.T. (2020) An application of compositional data analysis to multiomic time-series data. *NAR Genomics Bioinform.*, **2**, lqaa079.
9. Sisk-Hackworth,L., Ortiz-Velez,A., Reed,M.B. and Kelley,S.T. (2021) Compositional data analysis of periodontal disease microbial communities. *Front. Microbiol.*, **12**, 617949.
10. Jervis-Bardy,J., Leong,L.E.X., Marri,S., Smith,R.J., Choo,J.M., Smith-Vaughan,H.C., Nosworthy,E., Morris,P.S., O'Leary,S., Rogers,G.B., *et al.* (2015) Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome*, **3**, 19.
11. Trego,A., Keating,C., Nzeteu,C., Graham,A., O'Flaherty,V. and Ijaz,U.Z. (2022) Beyond basic diversity estimates—analytical tools for mechanistic interpretations of amplicon sequencing data. *Microorganisms*, **10**, 1961.
12. Jalanka-Tuovinen,J., Salonen,A., Nikkilä,J., Immonen,O., Kekkonen,R., Lahti,L., Palva,A. and De Vos,W.M. (2011) Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms. *PLoS One*, **6**, e23035.
13. Liang,W., Yang,Y., Wang,H., Wang,H., Yu,X., Lu,Y., Shen,S. and Teng,L. (2019) Gut microbiota shifts in patients with gastric cancer in perioperative period. *Medicine (Baltimore)*, **98**, e16626.
14. Zverev,A.O., Kichko,A.A., Pinaev,A.G., Provorov,N.A. and Andronov,E.E. (2021) Diversity indices of plant communities and their rhizosphere microbiomes: an attempt to find the connection. *Microorganisms*, **9**, 2339.
15. Evdokimova,E.V., Gladkov,G.V., Kuzina,N.I., Ivanova,E.A., Kimeklis,A.K., Zverev,A.O., Kichko,A.A., Aksenova,T.S., Pinaev,A.G. and Andronov,E.E. (2020) The difference between cellulolytic 'culturomes' and microbiomes inhabiting two contrasting soil types. *PLoS One*, **15**, e0242060.
16. Brumfield,K.D., Raupp,M.J., Haji,D., Simon,C., Graf,J., Cooley,J.R., Janton,S.T., Meister,R.C., Huq,A., Colwell,R.R., *et al.* (2022) Gut microbiome insights from 16S rRNA analysis of 17-year periodical cicadas (Hemiptera: Magicicada spp. Broods II, VI, and X. *Sci. Rep.*, **12**, 16967.
17. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J., *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.
18. Reback,J., Mendel,B., McKinney,W., Van den Bossche,J., Augspurger,T., Cloud,P., Hawkins,S., Young,G., Roeschke,M., Sinhrks, *et al.* (2021) pandas-dev/pandas: Pandas 1.0.3. https://doi.org/10.5281/zenodo.3715232.
19. Harris,C.R., Millman,K.J., van der Walt,S.J., Gommers,R., Virtanen,P., Cournapeau,D., Wieser,E., Taylor,J., Berg,S., Smith,N.J., *et al.* (2020) Array programming with NumPy. *Nature*, **585**, 357–362.
20. Hunter,J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.
21. Viry,M.jenkspy: Compute Natural Breaks (Fisher-Jenks algorithm). *Python Package Index - PyPI*, https://pypi.org/project/jenkspy/.
22. Schwarzberg,K., Le,R., Bharti,B., Lindsay,S., Casaburi,G., Salvatore,F., Saber,M.H., Alonaizan,F., Slots,J., Gottlieb,R.A., *et al.* (2014) The personal human oral microbiome obscures the effects of treatment on periodontal disease. *PLoS One*, **9**, e86708.
23. Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.
24. Kumar,P.K.V., Gottlieb,R.A., Lindsay,S., Delange,N., Penn,T.E., Calac,D. and Kelley,S.T. (2018) Metagenomic analysis uncovers strong relationship between periodontal pathogens and vascular dysfunction in American Indian population. bioRxiv doi: https://doi.org/10.1101/250324, 20 January 2018, preprint: not peer reviewed.
25. Torres,P.J., Thompson,J., McLean,J.S., Kelley,S.T. and Edlund,A. (2019) Discovery of a novel periodontal disease-associated bacterium. *Microb Ecol*, **77**, 267–276.
26. Xu,Y., Tandon,R., Ancheta,C., Arroyo,P., Gilbert,J.A., Stephens,B. and Kelley,S.T. (2021) Quantitative profiling of built environment bacterial and fungal communities reveals dynamic material dependent growth patterns and microbial interactions. *Indoor Air*, **31**, 188–205.
27. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
28. Gish,W. and States,D.J. (1993) Identification of protein coding regions by database similarity search. *Nat. Genet.*, **3**, 266–272.
29. Madden,T.L., Tatusov,R.L. and Zhang,J. (1996) Applications of network BLAST server. *Methods Enzymol.*, **266**, 131–141.
30. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
31. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. J. Comput. Bio.*l*, **7**, 203–214.
32. Zhang,J. and Madden,T.L. (1997) PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res*, **7**, 649–656.
33. Morgulis,A., Coulouris,G., Raytselis,Y., Madden,T.L., Agarwala,R. and Schäffer,A.A. (2008) Database indexing for production MegaBLAST searches. *Bioinformatics*, **24**, 1757–1764.
34. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
35. Boratyn,G.M., Schäffer,A.A., Agarwala,R., Altschul,S.F., Lipman,D.J. and Madden,T.L. (2012) Domain enhanced lookup time accelerated BLAST. *Biol. Direct*, **7**, 12.
36. Boratyn,G.M., Thierry-Mieg,J., Thierry-Mieg,D., Busby,B. and Madden,T.L. (2019) Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinformatics*, **20**, 405.
37. Bolyen,E., Rideout,J.R., Dillon,M.R., Bokulich,N.A., Abnet,C.C., Al-Ghalith,G.A., Alexander,H., Alm,E.J., Arumugam,M., Asnicar,F., *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, **37**, 852–857.
38. Schloss,P.D., Westcott,S.L., Ryabin,T., Hall,J.R., Hartmann,M., Hollister,E.B., Lesniewski,R.A., Oakley,B.B., Parks,D.H., Robinson,C.J., *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.
39. Jiang,B. (2013) Head/tail Breaks: A New Classification Scheme for Data with a Heavy-tailed Distribution. *Prof. Geogr.*, **65**, 482–494.
40. Bivand,R., Denney,B., Dunlap,R., Hernangomez,D., Ono,H., Parry,J. and Stigler,M. (2023) classInt: Choose Univariate Class Intervals, https://r-spatial.github.io/classInt/.

Received: March 18, 2023. **Revised:** November 25, 2023. **Editorial Decision:** December 11, 2023. **Accepted:** December 20, 2023