technologies to generate hypotheses or plan new experiments. If investigators choose to download data and/or code, they can access and analyze disparate data with the same functionality and syntax, which allows for faster comparisons and scientific discoveries. Because all of the code is open source, anybody can download, set up, and modify this ecosystem.

Data availability

All code is available from https://neurodata.io/tools/ under an Apache 2.0 license unless otherwise specified. All publicly available data are accessible at https://neurodata.io/data/ under an ODC-By v1.0 license, unless otherwise specified. ❐

Joshua T. Vogelstein[1]*, Eric Perlman[1], Benjamin Falk[1], Alex Baden[1], William Gray Roncal[2], Vikram Chandrashekhar[1], Forrest Collman[3], Sharmishtaa Seshamani[3], Jesse L. Patsolic[1], Kunal Lillaney[1], Michael Kazhdan[1], Robert Hider Jr.[2], Derek Pryor[2], Jordan Matelsky[2], Timothy Gion[2], Priya Manavalan[2], Brock Wester[2], Mark Chevillet[4], Eric T. Trautman[5], Khaled Khairy[5], Eric Bridgeford[1], Dean M. Kleissas[6], Daniel J. Tward[1], Ailey K. Crow[7], Brian Hsueh[7], Matthew A. Wright[7], Michael I. Miller[1], Stephen J. Smith[3], R. Jacob Vogelstein[6], Karl Deisseroth[7] and Randal Burns[1]

[1]*Johns Hopkins University, Baltimore, MD, USA.* [2]*Applied Physics Laboratory, Johns Hopkins University, Laurel, MD, USA.* [3]*Allen Institute for Brain Sciences, Seattle, WA, USA.* [4]*Facebook, Menlo Park, CA, USA.* [5]*Janelia Research Campus, Ashburn, VA, USA.* [6]*Gigantum, Washington, DC, USA.* [7]*Stanford University, Stanford, CA, USA.* **e-mail: jovo@jhu.edu*

Published online: 30 October 2018
https://doi.org/10.1038/s41592-018-0181-1

References

1. Burns, R. et al. The open connectome project data cluster: scalable analysis and vision for high-throughput neuroscience. In *Proc. 25th International Conference on Scientific and Statistical Database Management* (eds Szalay, A. et al.) Article 27 (ACM, New York, 2013).
2. Collman, F. et al. *J. Neurosci.* **35**, 5792–5807 (2015).
3. Simhal, A. K. et al. *PLoS Comput. Biol.* **13**, e1005493 (2017).
4. Shen, C. et al. *arXiv* Preprint at https://arxiv.org/abs/1609.05148 (2016).
5. Chung, K. et al. *Nature* **497**, 332–337 (2013).
6. Kutten, K. S. et al. A large deformation diffeomorphic approach to registration of CLARITY images via mutual information. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2017* Vol. 10433 (eds Descoteaux, M. et al.) 275–282 (Springer, Cham, 2017).

Acknowledgements

R.B., E.P., B.F., A.B., V.C., K.L., M.K., E.B., J.L.P., D.J.T., M.I.M., and J.T.V. were supported by the Defense Advanced Research Projects Agency (DARPA) SIMPLEX program, SPAWAR contract N66001-15-C-4041, NIH-NINDS TRA 1R01NS092474, "Synaptomes of Mouse and Man," and NSF 16-569 NeuroNex contract 1707298. W.G.R., B.W., R.H., D.P., T.G., P.M., and J.M. were supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA contract no. 2017-17032700004-005 under the MICrONS program and APL Internal Research and Development Funds. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the US Government. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation.

Competing interests

D.M.K. and R.J.V. are employed by Gigantum; this company provided support in the form of salaries for these authors, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Striped UniFrac: enabling microbiome analysis at unprecedented scale

To the Editor — The UniFrac metric is used frequently in microbiome research, but it does not scale to today's large datasets. We propose a new algorithm, Striped UniFrac, which produces results identical to those of previous algorithms but requires dramatically less memory and computing power. A BSD-licensed implementation is available that produces a C shared library linkable by any programming language (Supplementary Software and https://github.com/biocore/unifrac).

UniFrac[1] is a phylogenetic distance metric used to compare pairs of microbiome profiles. Microbiome studies now encompass tens of thousands of samples, such as the 27,751-sample Earth Microbiome Project (EMP)[2] and the 15,096-sample American Gut Project[3]. Existing algorithms for UniFrac computation cannot scale in time or space to these study designs. For example, Fast UniFrac with the EMP was projected to take months. Striped UniFrac produces results identical to those of other existing algorithms, shows >30-fold improvement in single-threaded performance and near-linear parallel scaling (Supplementary Fig. 1a,b), and can process the EMP dataset on a laptop in less than 24 hours. It can enable scientists to derive new biological insights, as shown by a meta-analysis[3] of the American Gut Project and EMP. To demonstrate the utility of the algorithm, we computed UniFrac on 113,721 public samples in Qiita[4] in less than 48 hours using 256 CPUs (an interactive plot is available at https://bit.ly/2LHMDFC).

The key advances with Striped UniFrac are improved space complexity, obtained through aggregation of metric constituents in post-order traversal, and rotation of proportion vectors for pairwise comparisons (methods, pseudocode, and complexity analysis are provided in the Supplementary Note). The post-order aggregation removes the dominant scaling factor for space complexity in Fast UniFrac (Supplementary Fig. 1c,d). Vector rotation (expressed by embedding; Fig. 1a) allows compilers to use single instruction multiple data (SIMD) operations. As a consequence of rotation, pairwise distances are computed along diagonals of the distance matrix (Fig. 1b), which results in more cache utilization, task-level parallelism, and hardware-level prefetch. We also introduce an optional heuristic that reduces compute by 50% by ignoring tips of the phylogeny (correlations with exact calculations are shown in Supplementary Fig. 1e). Empirical scaling results show that Striped UniFrac outperforms current popular implementations of UniFrac (time in Fig. 1c; space in Fig. 1d; benchmark and algorithm in the Supplementary Note).

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

An optimized implementation of Striped UniFrac is available on GitHub (https://github.com/biocore/unifrac) under a BSD license (implementation details are provided in the Supplementary Note),

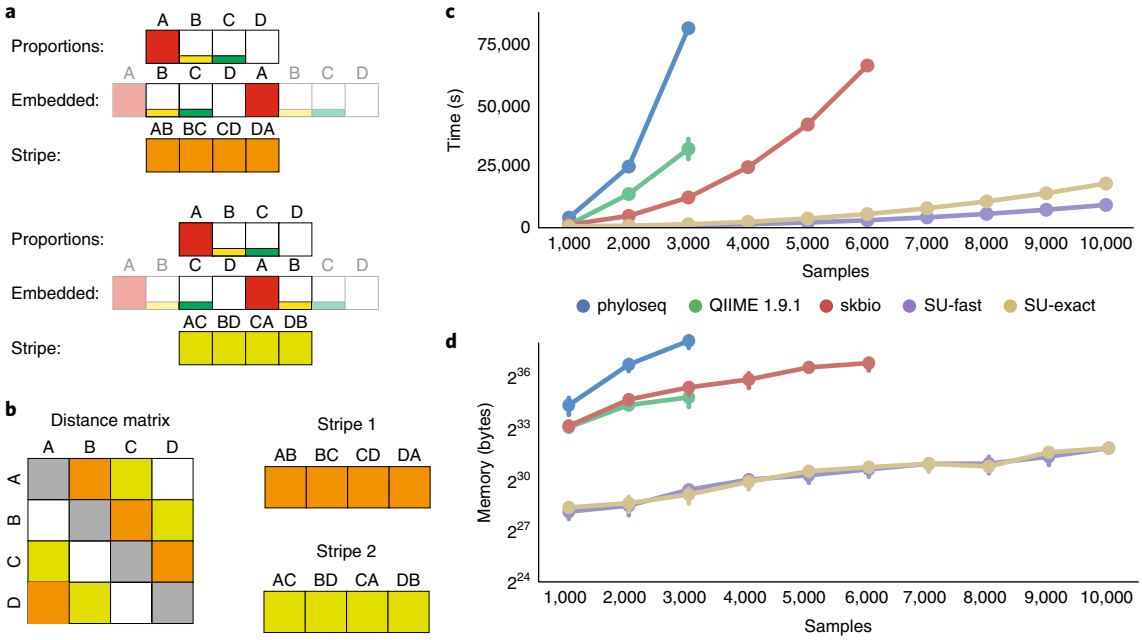

Fig. 1 | Algorithm description and empirical performance results. a, For a given node in a phylogeny, sample proportions are embedded to duplicate the proportion information. This duplication mimics rotation of the sample proportions to compute many pairwise distances at once in contiguous blocks of memory. **b**, A schematic of the two stripes in a four-sample logical distance matrix; the labels above the stripes indicate the pairwise comparison represented (for example, "AB" indicates the distance between samples A and B). **c**,**d**, Empirical time (**c**) and space (**d**) comparisons of weighted un-normalized UniFrac[1] with phyloseq[8], QIIME v1.9.1[5], scikit-bio (skbio), Striped UniFrac in fast mode (SU-fast), and Striped UniFrac in exact mode (SU-exact). Each data point represents the mean of $n = 10$ independent experiments (random subsets) using the EMP 90-nt (317,314 unique Deblur[9] sub-OTUs (operational taxonomic units)) dataset with increasing numbers of samples. All methods were run single-threaded on nonshared compute nodes that were not running other compute tasks. A job was killed if it exceeded 24 hours of wall time or 256 GB of memory (system max). Error bars indicate the 95% confidence interval around the mean. Source data for **c** and **d** are provided in Supplementary Data 2.

is part of QIIME 2[5], and is the UniFrac implementation used for meta-analyses in Qiita[4]. The source code builds a command-line version, a C shared library, and a Python application program interface. All known variants of UniFrac are implemented[1,6,7].

Data availability

The datasets analyzed during the current study are available in the Qiita repository with the specific study accessions in Supplementary Data 1, and were extracted with Qiita's redbiom interface. ❐

Daniel McDonald[1], Yoshiki Vázquez-Baeza[1], David Koslicki[2], Jason McClelland[2], Nicolai Reeve[1,6], Zhenjiang Xu[1], Antonio Gonzalez[1] and Rob Knight[1,3,4,5*]

*[1]Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA. [2]Mathematics Department, Oregon State University, Corvallis, OR, USA. [3]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA. [4]Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA, USA. [5]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA. [6]Present address: Biota Technology Inc., La Jolla, CA, USA. *e-mail: robknight@ucsd.edu*

Published online: 30 October 2018
https://doi.org/10.1038/s41592-018-0187-8

References

1. Lozupone, C. & Knight, R. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
2. Thompson, L. R. et al. *Nature* **551**, 457–463 (2017).
3. McDonald, D. et al. *mSystems* **3**, e00031-18 (2018).
4. Gonzalez, A. et al. *Nat. Methods* **15**, 796–798 (2018).
5. Caporaso, J. G. et al. *Nat. Methods* **7**, 335–336 (2010).
6. Chang, Q., Luan, Y. & Sun, F. *BMC Bioinformatics* **12**, 118 (2011).
7. Chen, J. et al. *Bioinformatics* **28**, 2106–2113 (2012).
8. McMurdie, P. J. & Holmes, S. *PLoS One* **8**, e61217 (2013).
9. Amir, A. et al. *mSystems* **2**, e00191-16 (2017).

Acknowledgements

This work was supported by the NSF (grant DBI-1565100 to D.M., Y.V.-B., Z.X., A.G., and R.K.; award 1664803 to D.K and J.M.), the Alfred P. Sloan Foundation (G-2017-9838 to D.M., Y.V.-B., A.G., and R.K.; G-2015-13933 to A.G. and R.K.), ONR (grant N00014-15-1-2809 to D.M., A.G., and R.K.), and NIH–NIDDK (grant P01DK078669 to A.G. and R.K.). This work was partially supported by XSEDE resource grant BIO150043. Additional support was provided by CRISP, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

Author contributions

D.M. designed Striped UniFrac, planned the study, analyzed data, and wrote the manuscript. Y.V.-B. integrated Striped UniFrac with QIIME 2 and contributed to the manuscript. D.K. and J.M. contributed to the proof. N.R. contributed language interface code. Z.X. contributed to the manuscript. A.G integrated Striped UniFrac with Qiita. R.K. planned the study and wrote the manuscript.

Competing interests

R.K. is a founder and CSO of Biota Technology Inc. D.M. is a consultant with Biota Technology Inc.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0187-8.

nature research

Corresponding author(s): *Rob Knight*

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	redbiom 0.2.0 was used to obtain data from Qiita , and is open source under the BSD license (https://github.com/biocore/redbiom/).
Data analysis	Striped UniFrac 0 9 3 (BSD) scikit bio 0 5 1 (BSD) pandas 0 20 2 (BSD) seaborn 0 8 1 (BSD) Python 6 2 1 (BSD) EMPeror 1 0beta16dev (BSD) Q ME 1 9 1 (GPL) phyloseq 1 16 2 (AGPL 3) Cython 0 26 (Apache 2 0) R 3 3 0 (GPL) Python 3 5 (PSFL)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets analysed during the current study are available publicly in the Qiita repository. For the analysis using the public data in Qiita, we provide a supplementary table with the specific Qiita study IDs used.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The Earth Microbiome Project and Qiita represent the largest microbiome sample sizes we are aware of.
Data exclusions	For empirical space and time measurements processing jobs were killed if they exceeded 24 hours walltime or 256GB RAM
Replication	For empirical space and time analysis each method was run on 10 subsets of data at a given level (e g 1000 samples) Variation in the results is represented in the figures with 95% confidence intervals All attempts at replication were successful
Randomization	For scalability testing, tips or samples were randomly selected using NumPy's random module.
Blinding	Blinding was obtained by randomization

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging