



Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns

Amnon Amir,^a Daniel McDonald,^a Jose A. Navas-Molina,^{a,c} Evguenia Kopylova,^a James T. Morton,^a Zhenjiang Zech Xu,^a Eric P. Kightley,^b Luke R. Thompson,^a Embriette R. Hyde,^a Antonio Gonzalez,^a Rob Knight^{a,c,d}

Department of Pediatrics, University of California San Diego, La Jolla, California, USA^a; Department of Applied Mathematics, and Interdisciplinary Quantitative Biology Graduate Program, University of Colorado Boulder, Boulder, Colorado, USA^b; Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA^c; Center for Microbiome Innovation, University of California San Diego, San Diego, California, USA^d

ABSTRACT High-throughput sequencing of 16S ribosomal RNA gene amplicons has facilitated understanding of complex microbial communities, but the inherent noise in PCR and DNA sequencing limits differentiation of closely related bacteria. Although many scientific questions can be addressed with broad taxonomic profiles, clinical, food safety, and some ecological applications require higher specificity. Here we introduce a novel sub-operational-taxonomic-unit (sOTU) approach, Deblur, that uses error profiles to obtain putative error-free sequences from Illumina MiSeq and HiSeq sequencing platforms. Deblur substantially reduces computational demands relative to similar sOTU methods and does so with similar or better sensitivity and specificity. Using simulations, mock mixtures, and real data sets, we detected closely related bacterial sequences with single nucleotide differences while removing false positives and maintaining stability in detection, suggesting that Deblur is limited only by read length and diversity within the amplicon sequences. Because Deblur operates on a per-sample level, it scales to modern data sets and meta-analyses. To highlight Deblur's ability to integrate data sets, we include an interactive exploration of its application to multiple distinct sequencing rounds of the American Gut Project. Deblur is open source under the Berkeley Software Distribution (BSD) license, easily installable, and downloadable from <https://github.com/biocore/deblur>.

IMPORTANCE Deblur provides a rapid and sensitive means to assess ecological patterns driven by differentiation of closely related taxa. This algorithm provides a solution to the problem of identifying real ecological differences between taxa whose amplicons differ by a single base pair, is applicable in an automated fashion to large-scale sequencing data sets, and can integrate sequencing runs collected over time.

KEYWORDS DNA sequencing, microbiome

An important goal of microbiome research is identifying taxa present in a given sample. Next-generation sequencing of the 16S rRNA gene on Illumina instruments is commonly used for this task but suffers from an error rate of 0.1% per nucleotide (1). In a typical study spanning millions of sequences, many sequences contain at least one error, obscuring the underlying biology through inaccurate taxon identification and inflated diversity statistics. These errors seldom affect statistical tests for differences between two communities, but higher precision is becoming increasingly important as the field moves toward applications with clinical or regulatory significance.

Received 9 December 2016 Accepted 14 December 2016 Published 7 March 2017

Citation Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2:e00191-16. <https://doi.org/10.1128/mSystems.00191-16>.

Editor Jack A. Gilbert, Argonne National Laboratory

Copyright © 2017 Amir et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Rob Knight, robknight@ucsd.edu.

A.A. and D.M. contributed equally to this article.

 Deblur resolves single-nucleotide community sequencing patterns

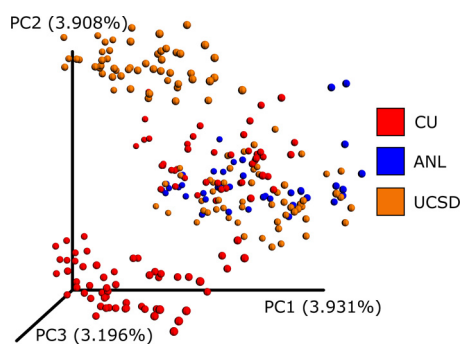


FIG 1 A principal-coordinate analysis plot of UniFrac distances from *de novo* OTUs as visualized by Emperor. A subset of American Gut Project samples spanning sequencing centers and rounds were selected. UCLUST (3) was run independently per round via QIIME. The resulting OTU tables were merged, normalizing sequencing identifiers (IDs) such that if the same sequence was observed in multiple rounds it would receive the same ID. Observations with fewer than 10 counts were dropped. The data were rarefied to 5,000 sequences per sample. The plot shown is based on unweighted UniFrac distances, and the samples are colored by the sequencing center. An interactive visualization can be viewed at https://nbviewer.jupyter.org/github/knightlab-analyses/deblur-manuscript/blob/master/embedded_figure_1.ipynb; the coloring used in the static image can be done by selecting “run_center” as the scatter field. CU, University of Colorado Boulder; ANL, Argonne National Laboratory; UCSD, University of California San Diego.

The classic approach to overcoming these errors is to cluster amplicon sequences into operational taxonomic units (OTUs) (2, 3) based on an arbitrary sequence identity threshold. This approach reduces problems caused by erroneous sequences but also reduces phylogenetic resolution because sequences below the identity threshold cannot be differentiated. Furthermore, OTUs picked within individual data sets cannot be merged when different centroid sequences are chosen in each data set, limiting the ability to combine OTU results. This problem of merging OTUs is exemplified by Fig. 1, for which *de novo* OTUs were assessed independently on separate rounds (i.e., “run_center”) of sequencing with a subset of American Gut Project data. Although approaches such as closed-reference and open-reference OTU picking (4) reduce this problem, integrating large data sets into a single OTU space remains a challenge. Here we describe Deblur, a novel sub-OTU (sOTU) method for fast and accurate identification of exact sequences in amplicon studies, and show how it can be used to integrate large data sets.

Similar in concept to AmpliconNoise (5), a denoising method for pyrosequencing, Deblur, like DADA2 (6) and UNOISE2 (7), attempts to obtain single-nucleotide resolution from Illumina data with statistical methods to infer the putative true sequences within a sample that give rise to the distribution of observed error-prone sequences. Unlike DADA2 and UNOISE2, Deblur operates on each sample independently. It compares sequence-to-sequence Hamming distances within a sample to an upper-bound error profile (see Table S1 in the supplemental material; mathematical derivation in Text S1 in the supplemental material) combined with a greedy algorithm to obtain single-nucleotide resolution. The Deblur algorithm is implemented as follows (see Fig. S1 in the supplemental material). First, sequences are sorted by abundance. Second, from the most to least abundant sequence, the number of predicted error-derived reads is subtracted from neighboring reads based on their Hamming distance, using an upper bound on the error probability. A parameterized maximal probability for indels (defaulting to 0.01) and a parameterized mean read error rate for normalization (defaulting to 0.5%) are included. Finally, any sequence whose abundance drops to 0 after a subtraction is removed from the list of valid sequences. Sequences not considered to be valid (i.e., noise) are removed. After applying Deblur, only reads likely to have been presented to the sequencer are retained. However, it is possible that the reads would still contain chimeras originating from PCR. Reads are filtered for *de novo* chimeras using UCHIME (8) as implemented by VSEARCH (9) using modified parameters (Text S1).

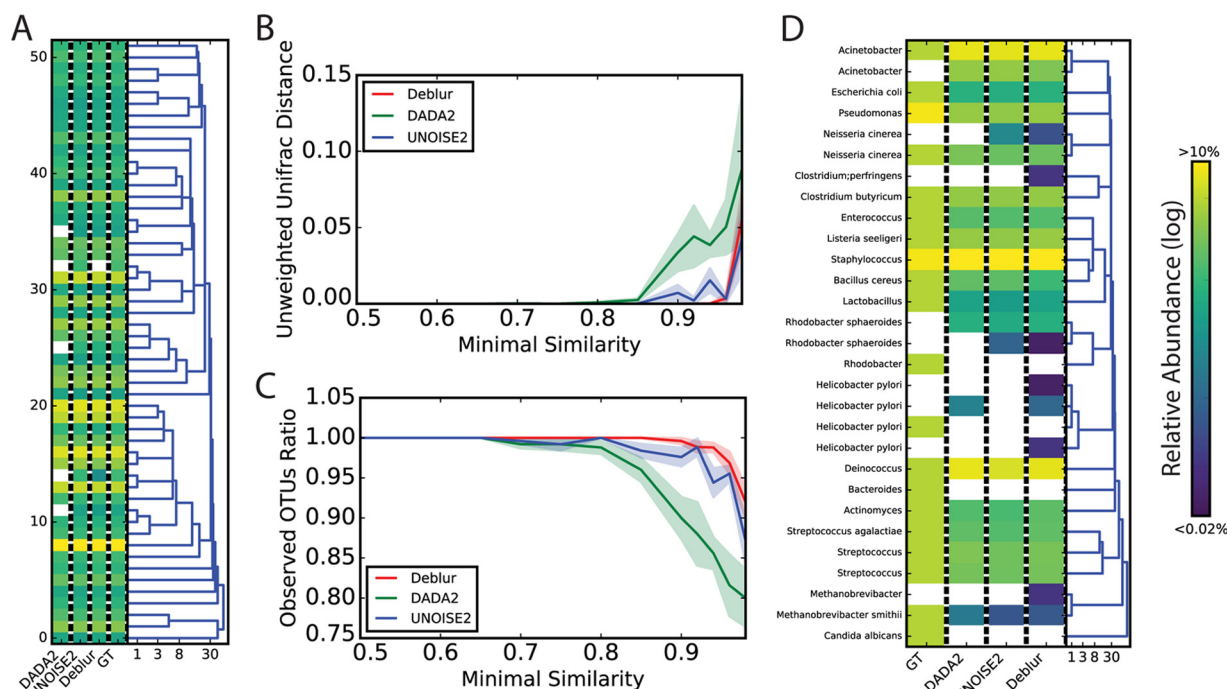


FIG 2 Benchmarks of OTU picking tools on artificial communities. (A) A simulation was performed on the basis of samples from a real fecal community (11) using the 52 most abundant bacterial species identified in this study. Reads were then simulated using an ART Illumina (12) read simulator. OTU picking was performed on these simulated reads using UNOISE2, DADA2, and Deblur. The relative abundances predicted by each of these tools and the ground truth (GT) are shown in the heat map. The dendrogram was built using hierarchical clustering based on the Hamming distance between the sequences, with numbers indicating sequence similarity (log scale). (B) Simulated communities with various levels of sequence-sequence similarity. Unweighted UniFrac distances of the predicted OTUs from UNOISE2, DADA2, and Deblur were compared to those of the original composition of the simulated communities. The x axis denotes the similarity radius for each community. The shaded area denotes the standard error of the mean distance estimation (based on 10 random repeats per community). (C) Similar to panel B but with the ratio of observed OTUs (predicted by UNOISE2, DADA2, and Deblur) to actual OTUs in each simulation indicated. (D) Performance of Deblur, UNOISE2, and DADA2 on the even1 community from mock-3 (14). GT data denote the expected ground truth relative frequency for each sOTU as informed by the design of the mock community. Dendrograms and colors are the same as described for panel A.

Using simulated, mock, and real data sets, we compared the performance of Deblur with the performance of DADA2 and UNOISE2. Most analyses using real data sets did not complete using the free version of UNOISE2. We omitted classic OTU methods and MED (10), given the benchmarks described in reference 6. Finally, we applied Deblur to multiple sequencing rounds of the American Gut Project samples to demonstrate integration across sequencing runs from multiple instruments.

We first compared methods using simulated communities based on bacterial taxa and frequencies obtained from Sanger sequencing of one stool sample (11). Reads were simulated from this real community using ART (12) to produce Illumina-like sequence data. All three methods identified sOTUs with single-nucleotide differences (Fig. 2A). We then simulated sequences over increasing levels of similarity between the real sequences, measuring unweighted UniFrac (13) distances to ground truth, and observed OTUs (Fig. 2B and C). Deblur, DADA2, and UNOISE2 were all close to the ground truth except at high similarity levels, but all three suffered from limitations in distinguishing the true reads. Using the MiSeq data generated from a 22-member community ("mock-3") (14) trimmed to the first 150 nucleotides (nt) due to poor 3' quality, we observed that all methods produced results that were close to the ground truth (Fig. 2D).

Stability (i.e., obtaining the same sOTU across different samples) is becoming critical as more study designs exploit existing samples from resources like the Earth Microbiome Project (15) or require integration of sequence data collected over time such as the American Gut Project (<http://americangut.org>). We compared the levels of stability of Deblur and DADA2 using technical replicates from a data set consisting of 40 individuals, each with one fecal sample sequenced twice on two separate MiSeq runs

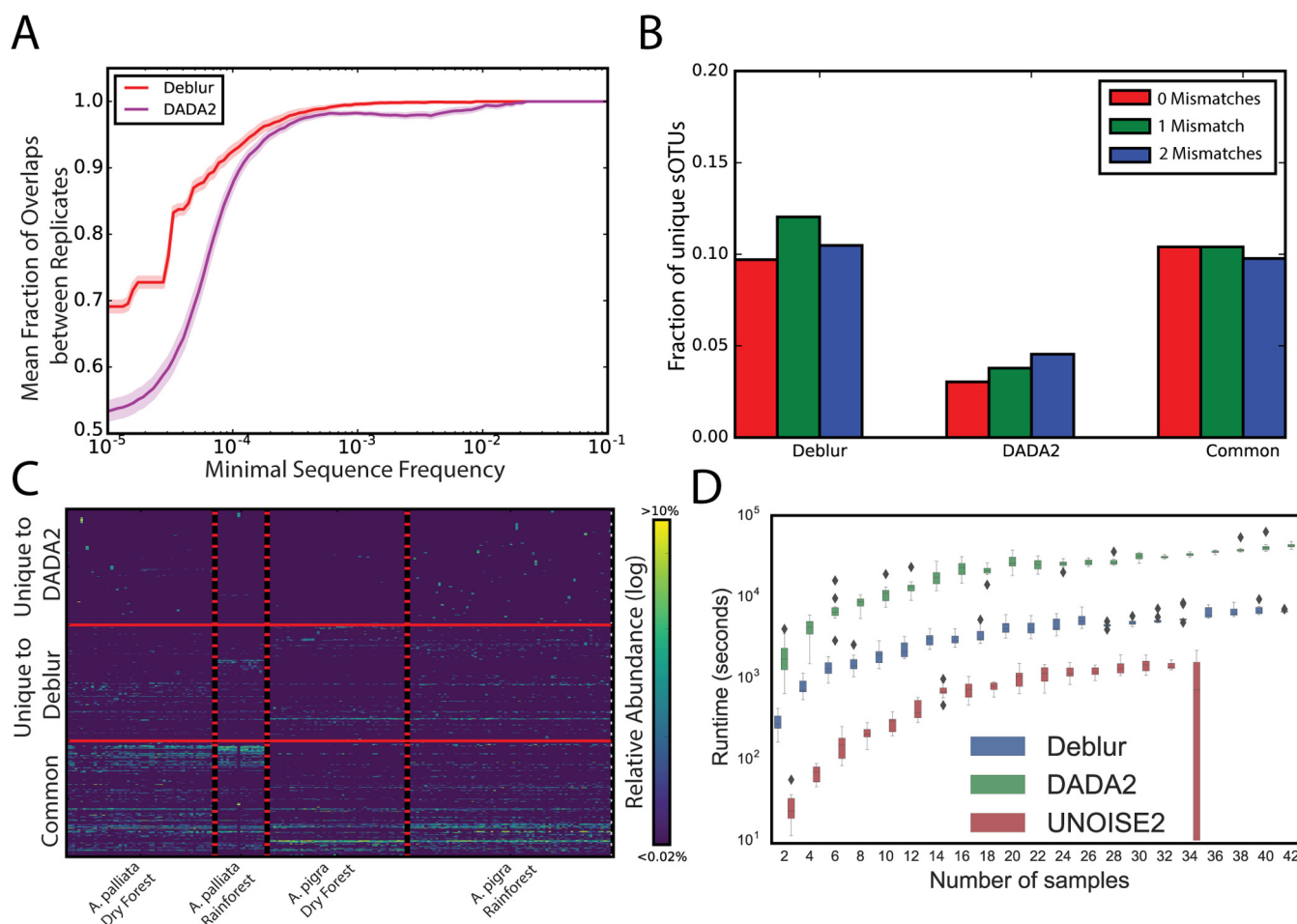


FIG 3 Benchmarks of OTU picking tools on natural communities. (A) Stability analysis on experimental technical repeats. Data indicate fractions of overlapping sOTUs from two technical replicates in all OTUs as a function of the minimal frequency threshold present in one of the repeats. (B and C) Application of Deblur in the howler monkey data set. (B) Fraction of sequences matching entries in the NCBI nr/nt database (as of 1 December 2016) with 0.1 or 2 mismatches (red, green, or blue, respectively) from sOTUs unique to Deblur or to DADA2 or present in both (left to right). (C) Heat maps showing sOTUs (rows) in common with Deblur and DADA2, as well as those unique to Deblur and DADA2 (bottom, middle, and top rows, respectively). Samples (columns) are sorted by species and habitat. A total of 200 sOTUs per group (i.e., common, unique to Deblur, or unique to DADA2) were randomly selected for visualization purposes. (D) Single-threaded runtime comparison of Deblur, DADA2, and UNOISE2 against one of the stability MiSeq runs at increasing numbers of samples.

(16). sOTUs for each run were assessed separately, and we compared the fractions of sOTUs from one run to those present in the second run, as a function of the minimal sOTU frequency. Deblur showed greater stability than DADA2 at a higher frequency cutoff (Fig. 3A), indicating that a larger fraction of sOTUs from the first run were also identified in the second run. To further test the stability of Deblur, we sourced previously sequenced fecal samples from the American Gut Project and selected fecal samples which spanned five distinct sequencing runs over multiple instruments from three geographically distinct locations (Table S2). As can be seen in the Emperor (17) plot of a principal-coordinate analysis of unweighted UniFrac distances, samples from different sequencing rounds (denoted by “center_project_name”) are integrated in the ordination, demonstrating that the sequencing rounds do not separate as is typical with OTU-based methods (Fig. 4). In contrast, performing the same experiment with UNOISE2 (running it per round and merging results with respect to commonly identified sequences) produced an observable effect mediated by the sequencing round and run center (Fig. 5), although the magnitude of the effect was notably less than that observed with *de novo* OTUs (Fig. 1). We note that this is in contradiction to the recommended way of running UNOISE2 (i.e., we did not run it on the full data set);

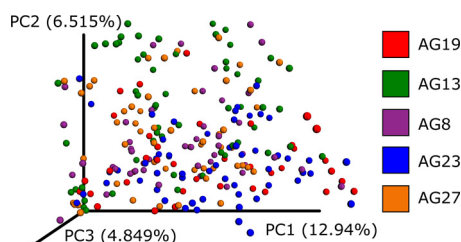


FIG 4 A principal-coordinate analysis plot of UniFrac distances from Deblur as visualized by Emperor. A subset of American Gut Project samples spanning sequencing centers and rounds were selected. Each sample was processed separately by Deblur. Observations with fewer than 10 counts were dropped. The data were rarefied to 5,000 sequences per sample. The plot shown is based on unweighted UniFrac distances and is colored according to the round of sequencing in the American Gut Project (AG). An interactive visualization can be viewed at https://nbviewer.jupyter.org/github/knightlab-analyses/deblur-manuscript/blob/master/embedded_figure_4.ipynb; the coloring used in the static image can be made by selecting the “center_project_name” as the scatter field.

however, we are unaware of a mode of operation (free or full version) capable of operating on modern large-scale data sets.

Next, we compared DADA2 and Deblur using a complex natural community and a previously published data set of fecal samples from two species of howler monkeys (18). Deblur and DADA2 detected 1,938 and 1,636 sOTUs, respectively, after removal of sOTUs with fewer than 10 total reads from each method (Fig. S2A). Following filtering, about 70% of the sOTUs were identical between the methods (Fig. S2B). As expected, both methods identified differential sOTUs (permutation-based rank mean test; 0.1 false-discovery rate–Benjamini-Hochberg method [FDR-BH] control value) with 61% of Deblur sOTUs differentiating between primate species (1,193/1,938), compared to 55% of DADA2 sOTUs (891/1,636). To assess whether the sOTUs unique to either method were from increased numbers of artifacts, we used BLAST (19) to compare each unique sequence against nt/nr and plotted the fraction of sOTUs with zero, one, or two mismatches. We observed that sOTUs unique to Deblur showed fewer mismatches than those unique to DADA2 (Fig. 3B). The distribution of sOTUs over the monkey samples suggests that the sOTUs unique to Deblur are more plausible because they show a pattern similar to those identified by both methods, whereas the sOTUs unique to

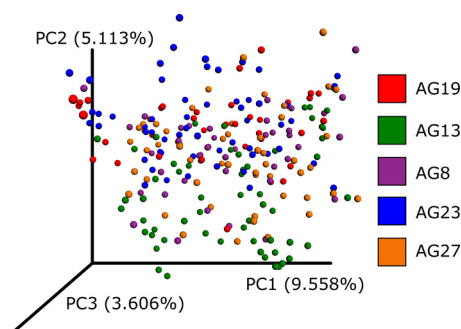


FIG 5 A principal-coordinate analysis plot of UniFrac distances from UNOISE2 as visualized by Emperor. A subset of American Gut Project samples spanning sequencing centers and rounds were selected. UNOISE2 was run independently per round. The resulting sOTU tables were merged, normalizing sequencing IDs such that if the same sequence were observed in multiple rounds it would receive the same ID. Observations with fewer than 10 counts were dropped. The data were rarefied to 5,000 sequences per sample. The plot shown is based on unweighted UniFrac distances and is colored according to the round of sequencing in the American Gut Project. An interactive visualization can be viewed at https://nbviewer.jupyter.org/github/knightlab-analyses/deblur-manuscript/blob/master/embedded_figure_5.ipynb; the coloring used in the static image can be made by selecting the “center_project_name” as the scatter field. The static shot is oriented to show PC1 versus PC2, and the separation is more pronounced if orienting the projection to look at PC2 versus PC3.

DADA2 have markedly different patterns of clusters of unique sOTUs within single samples (Fig. 3C).

Finally, to explore performance characteristics, we used a MiSeq run from the stability analysis in order to assess computational space and time demands of DADA2, Deblur, and UNOISE2 (where possible) over an increasing number of samples. UNOISE2 was an order of magnitude faster than Deblur, while Deblur was an order of magnitude faster than DADA2 (Fig. 3D). Deblur maintained a fairly flat memory profile (Fig. S3) in contrast to the growth observed with DADA2 and UNOISE2.

Like DADA2 and UNOISE2, which approach the same concept with different algorithms, Deblur produces stable sOTUs which can achieve single-nucleotide resolution and which can be used in place of OTUs. However, unlike DADA2 and UNOISE2, Deblur does not require operation on the full study and can thus be parallelized easily to very large projects. Finally, Deblur is released under the Berkeley Software Distribution (BSD) open source license, allowing easy commercial adoption and peer scrutiny. Consequently, because of its stability, ability to integrate, performance, and open-source license, Deblur is positioned to operate on present and future large-data sets as well as continued discovery through reuse of existing rich data sets.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/mSystems.00191-16>.

TEXT S1, DOCX file, 0.1 MB.

FIG S1, PDF file, 0.8 MB.

FIG S2, PDF file, 0.1 MB.

FIG S3, PDF file, 0.9 MB.

TABLE S1, XLSX file, 0.03 MB.

TABLE S2, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

We thank Justine Debelius for discussions on assessment of Deblur and Yoshiki Vázquez Baeza for help with embedding the Emperor plots.

This work was supported by the Sloan Foundation, the National Science Foundation (grants DBI-1565057 and DGE-1144086), the Department of Defense (grants W81XWH-15-1-0653 and N0001-14-15-1-2809), and the National Institutes of Health (grant P01DK078669).

REFERENCES

- Glenn TC. 2011. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* 11:759–769. <https://doi.org/10.1111/j.1755-0998.2011.03024.x>.
- Schloss PD, Handelsman J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* 71:1501–1506. <https://doi.org/10.1128/AEM.71.3.1501-1506.2005>.
- Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461. <https://doi.org/10.1093/bioinformatics/btq461>.
- Rideout JR, He Y, Navas-Molina JA, Walters WA, Ursell LK, Gibbons SM, Chase J, McDonald D, Gonzalez A, Robbins-Pianka A, Clemente JC, Gilbert JA, Huse SM, Zhou HW, Knight R, Caporaso JG. 2014. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* 2:e545. <https://doi.org/10.7717/peerj.545>.
- Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. 2011. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12:38. <https://doi.org/10.1186/1471-2105-12-38>.
- Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 13:581–583. <https://doi.org/10.1038/nmeth.3869>.
- Edgar RC. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* <https://doi.org/10.1101/003723>.
- Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27:2194–2200. <https://doi.org/10.1093/bioinformatics/btr381>.
- Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. <https://doi.org/10.7717/peerj.2584>.
- Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. 2015. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* 9:968–979. <https://doi.org/10.1038/ismej.2014.195>.
- Dethlefsen L, Huse S, Sogin ML, Relman DA. 2008. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* 6:e280. <https://doi.org/10.1371/journal.pbio.0060280>.
- Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. *Bioinformatics* 28:593–594. <https://doi.org/10.1093/bioinformatics/btr708>.
- 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>.
- Bokulich NA, Rideout JR, Mercurio WG, Shiffer A, Wolfe B, Maurice CF, Dutton RJ, Turnbaugh PJ, Knight R, Caporaso JG. 2016. mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems* 11:e00062-16. <https://doi.org/10.1128/mSystems.00062-16>.
- Gilbert JA, Jansson JK, Knight R. 2014. The Earth Microbiome project:

- successes and aspirations. *BMC Biol* 12:69. <https://doi.org/10.1186/s12915-014-0069-1>.
16. Hildebrand F, Tadeo R, Voigt AY, Bork P, Raes J. 2014. Lotus: an efficient and user-friendly OTU processing pipeline. *Microbiome* 2:30. <https://doi.org/10.1186/2049-2618-2-30>.
17. Vázquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. 2013. Emperor: a tool for visualizing high-throughput microbial community data. *Gigascience* 2:16. <https://doi.org/10.1186/2047-217X-2-16>.
18. Amato KR, Martinez-Mota R, Righini N, Raguette-Schofield M, Corcione FP, Marini E, Humphrey G, Gogul G, Gaffney J, Lovelace E, Williams L, Luong A, Dominguez-Bello MG, Stumpf RM, White B, Nelson KE, Knight R, Leigh SR. 2016. Phylogenetic and ecological factors impact the gut microbiota of two Neotropical primate species. *Oecologia* 180:717–733. <https://doi.org/10.1007/s00442-015-3507-z>.
19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
20. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, Caporaso JG. 2012. The biological observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience* 1:7. <https://doi.org/10.1186/2047-217X-1-7>.