

1 **Supporting Information**

2 **Interpreting UniFrac with Absolute Abundance: A Conceptual and Practical Guide**

3 Augustus Pendleton^{1*} & Marian L. Schmidt^{1*}

4 ¹Department of Microbiology, Cornell University, 123 Wing Dr, Ithaca, NY 14850, USA

5 Corresponding Authors: Augustus Pendleton: arp277@cornell.edu; Marian L. Schmidt:
6 marschmi@cornell.edu

7

8 **Supporting Methods**

9 *ASV Generation and Phylogenetic Tree Construction*

10 Sequencing data and identifying metadata were downloaded from the Sequence Read
11 Archive (SRA), from BioProject IDs PRJNA815056, PRJNA575097, PRJNA1212049, and
12 PRJNA302180 [1–4]. Full details and code of the Pendleton et al. 2025 data analysis are
13 included within that paper and associated Github repository and not shown here. Each dataset
14 varied substantially in terms of which 16S region it targeted, sequencing strategy, and read
15 quality, so ASV generation varied between them in terms of primer removal, filtering, and
16 trimming (see code for full description of these steps). Post trimming, all ASVs were generated
17 using the same methods within the standard DADA2 workflow [5]. Chimaeras were removed,
18 and ASVs were size selected (252/253 bp for V4 datasets, >400bp for V3-V4 datasets).
19 Taxonomy was assigned via the Silva v138.2 database, and used to remove mitochondrial and
20 chloroplast sequences [6]. When sequencing positives or negatives were present, they were
21 removed.

22 Phylogenetic trees were built using alignment via MAFFT followed by FastTree under a
23 generalized time-reversible model [7, 8]. Trees were visualized via ggtree in R, and anomalously
24 long branches were removed using ape [9]. Trees, metadata, taxonomy, and ASV abundances
25 (OTU tables) were organized and analyzed using phyloseq [10].

26 *Rarefaction and β -diversity*

27 To generate rarefied ASV tables of equal sequencing depth, ASV abundance matrices
28 were subsampled using a multivariate hypergeometric distribution via the rmvhyper function in
29 the extraDistr package (see generate_rarefied_abs_tables.R) [11]. Each ASV was then converted
30 to relative abundances, and then to absolute abundances by multiplying the relative abundance
31 by each samples cell count or 16S copy number. Bray-Curtis dissimilarities were calculated via
32 the vegdist function in vegan [12]. Unless otherwise noted, all Unifrac distances were calculated
33 via the GUnifrac package [13]. Final distance matrices were the average of all rarefied distance
34 matrices. All samples within each dataset were used for contour plots in Figure 2.

35 For the analysis of rarefaction itself (Fig. S4), datasets were rarefied to multiple
36 sequencing depths (250, 500, 1000, 5000, 10,000 reads per sample), as well as the minimum
37 sequencing depth observed in each dataset (ranging from 1,086 reads per sample in the cooling
38 reactor dataset to 48,601 reads per sample in the soil dataset). In addition, a non-rarefied control
39 from each dataset was generated as a control. Absolute abundance normalization and GU^A
40 calculation were then performed as described above.

41 *PERMANOVAs, Ordinations, and Mantel Tests*

42 PERMANOVAs were conducted via the adonis2 function in vegan (Fig. 3 and Fig. S2).
43 To limit confounding variables, not all samples were used in these analyses. From the cooling
44 water dataset, just samples from Reactor cycle 1 were used. For the mouse gut, just stool samples
45 were used. For the soil dataset, just mature samples were used. All PERMANOVAs were run
46 with 1,000 iterations. These same, simplified datasets were used for Principal Coordinates
47 Analysis in Fig S3. When estimating the correlation between distance matrices, we used the
48 mantel function from the vegan package with 999 permutations.

49

50 *Timing Analysis*

51 To estimate computational time, we subsampled the soil dataset to a set number of ASVs,
52 samples, and α numbers. When testing ASV number, we used 10 samples and one α value, when
53 testing sample or α values, ASVs were held constant at 2,000. Each case was replicated 20
54 times, and computation time was calculated via the microbenchmark function from the
55 microbenchmark package, with two replications each time [14].

56 *Error Analysis*

57 To estimate the impact of random error on quantification methods, we used the mouse
58 gut dataset, focusing only on the stool samples. These samples ranged in 16S copy number from
59 10^{11} - 10^{12} copies/gram. We tested a range of potential error, from 1% up to 50%. For each error
60 percentage, the amount of error was selected from a normal distribution with a mean of that error
61 percent and a standard deviation $1/10^{\text{th}}$ of that error percentage. This error was then randomly
62 assigned a direction (by multiplying by a binomial distribution of -1 and 1), and multiplied by
63 the copy number to create a deviation from the true value, which was added to the original value.
64 For example, in the case of 50% error, we first drew a random selection of error values from a
65 distribution with mean 0.5 and standard deviation of 0.05. These errors were then randomly
66 assigned to be negative or positive, and multiplied by the original cell counts, plus the cell count
67 itself. We repeated this fifty times. Across these fifty iterations, we first rarefied the ASV
68 abundances to relative abundance and then normalized to absolute abundance using these error-
69 added values. We then compared the absolute difference in GU^A or BC^A from these error-added
70 datasets compared to the original data to produce Fig. 4C-D.

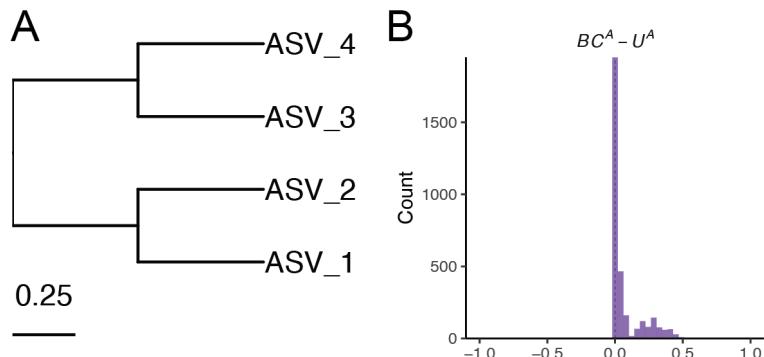
71 *Copy Number Normalization*

72 We used PICRUSt2 (v2.6, with the updated PICRUSt2-SC database) to normalize ASV
73 abundances for each dataset, using the default settings ($\text{NSTI} = 2$) in the picrust2_pipeline.py
74 function [15]. These normalized ASV tables (in /EC_metagenome_out/seqtan_norm.tsv.gz) were
75 then used to calculate GU^A using our standard pipeline, including rarefaction to the minimum
76 sequencing depth (across 10 iterations), normalization by absolute abundance, and averaging of
77 GU^A across iterations. The predicted 16S rRNA gene copy number of each ASV (for Fig. S6)
78 was accessed from combined_marker_predicted_and_nsti.tsv.gz.

79 *Other Coding Packages*

80 Other packages used for general coding and visualization include tidyverse, purr,
81 patchwork, NatParksPalette, broom, corrr, ggpibr, biomformat, and renv. All packages and
82 version numbers are listed in Table S1.

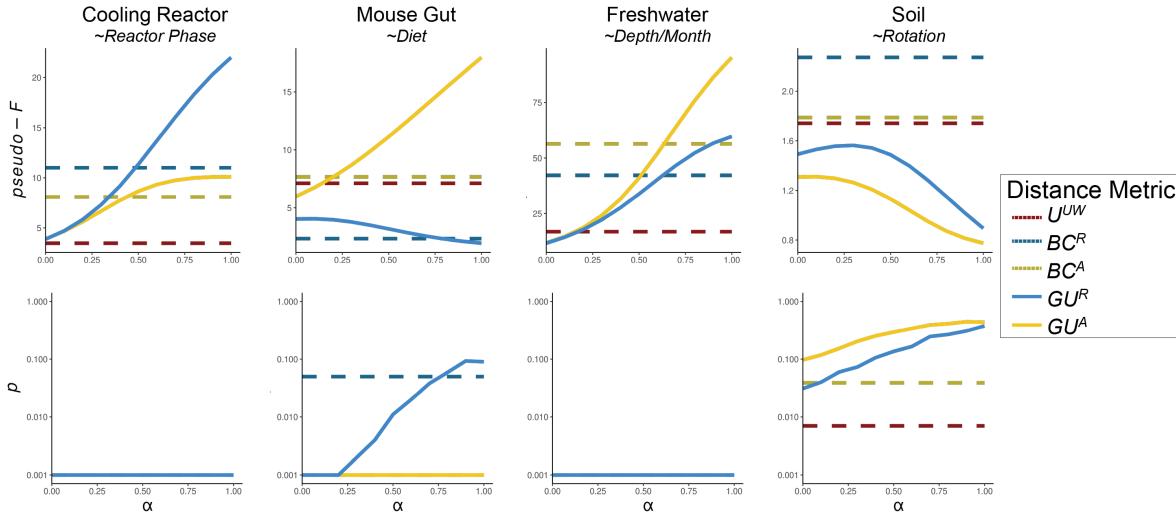
83 **Supplemental Figures**



84

85 *Figure S1.* U^A is always less than BC^A when branch lengths are fully symmetrical. (A)
86 Symmetrical tree used for simulations as opposed to non-symmetrical tree in Fig. 1A. (B)
87 Distribution of differences between BC^A and U^A . As the differences are never negative, U^A is
88 always less than or equal to BC^A .

89

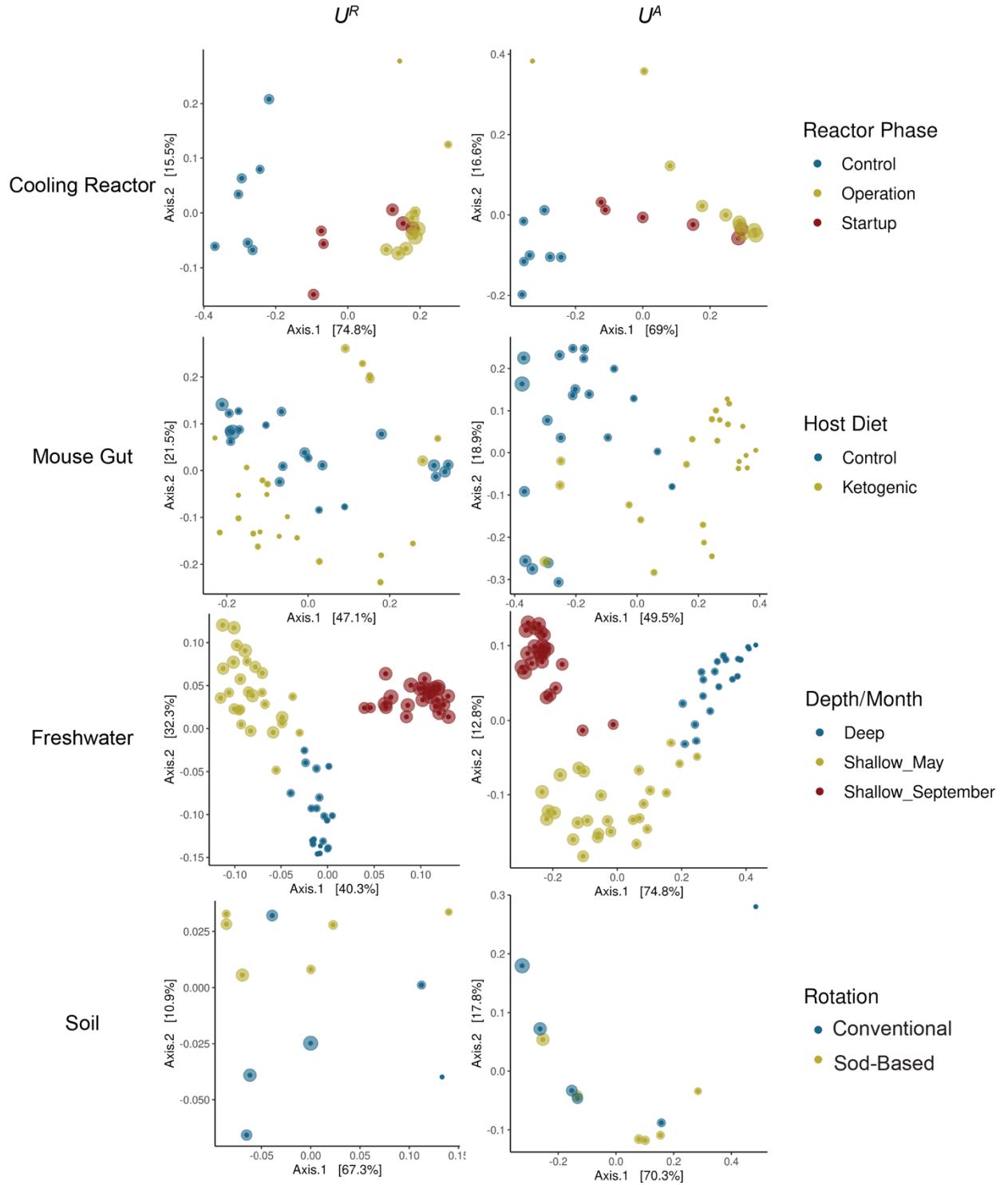


90

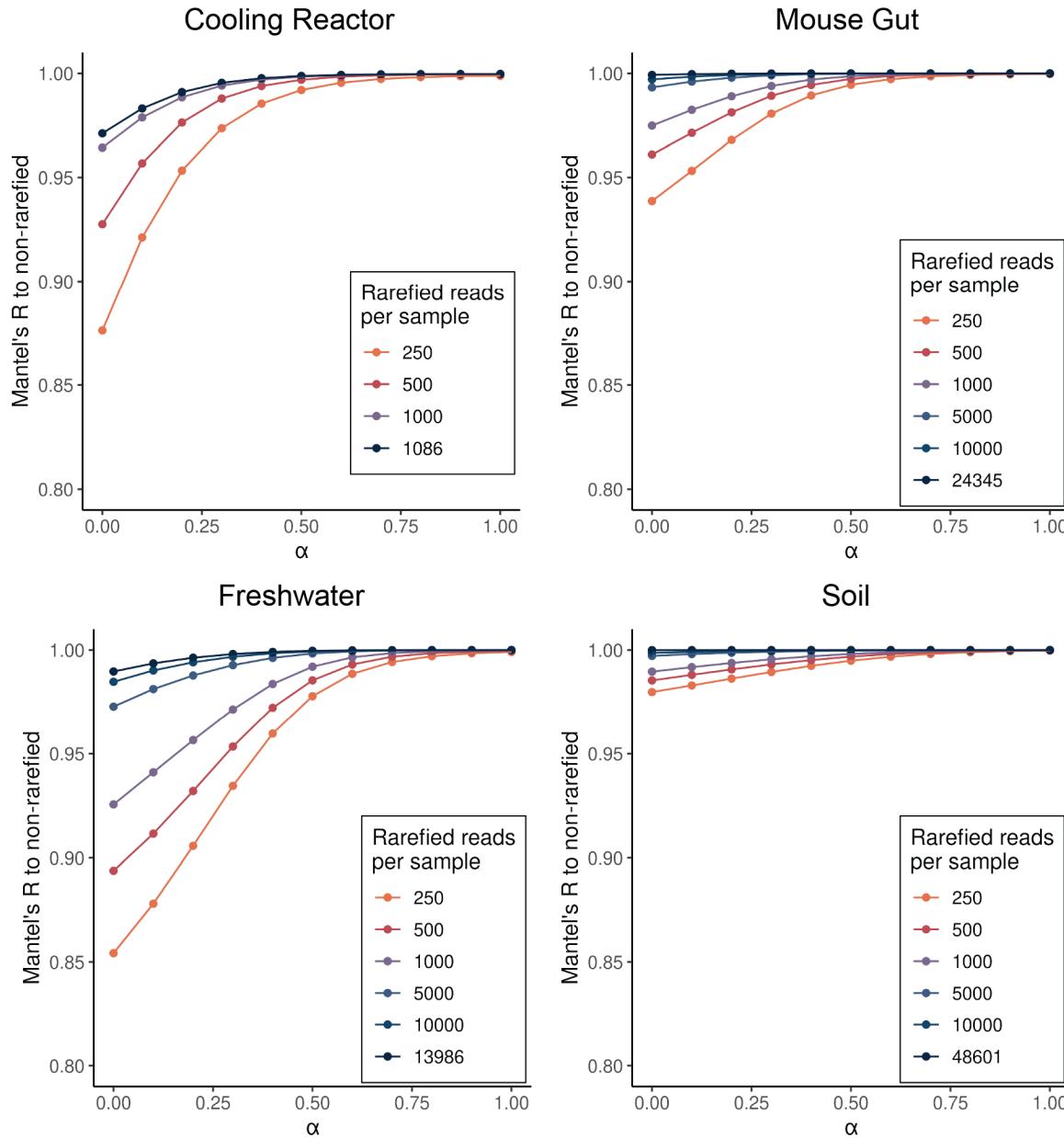
91 *Figure S2. Additional PERMANOVA results when using GU^A across a range of α values.*
 92 PERMANOVAs were run testing the significance of two-three category groups from each
 93 dataset (provided in italics beneath data names). Results indicate *pseudo-F* statistics and *p*-values
 94 after 1,000 iterations. In the cooling reactor, only samples from Reactor cycle 1 were used; in the
 95 mouse gut, only stool samples were used, and in the soil, only mature samples were used. Note
 96 the y-axes for *pseudo-F* plots are variable between datasets, and y-axes for the *p*-value plots are
 97 log-scaled.

98

99



100 *Figure S3. Principal Coordinate Analysis ordinations of each dataset using U^R and U^A .* Points
101 are colored using the same categorical variable tested in the PERMANOVAs of Fig. 2 and Fig.
102 S2 (for additional details on experimental design, see [1, 3, 4, 16]). Both U^R and U^A were
103 calculated at an $\alpha = 1$.

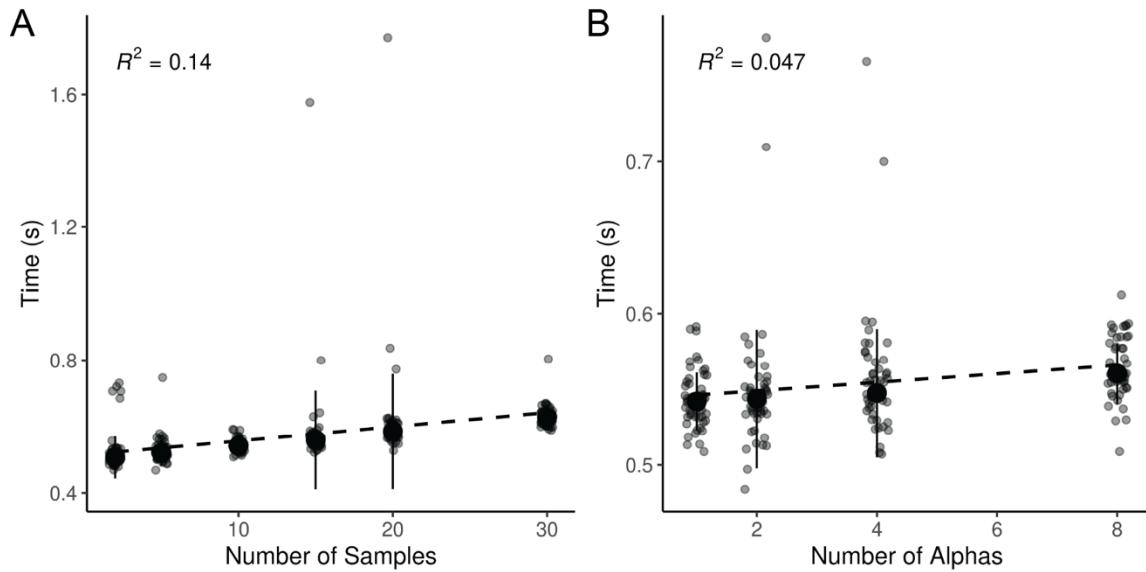


105

106 *Figure S4. Effect of rarefaction on GU^A .* GU^A was calculated across α values from 0 up to 1 at
 107 multiple rarefaction depths. Depths ranged from 250 reads/sample, up to minimum sequencing
 108 depth for each dataset (1,086 reads/sample in the cooling reactor, up to 48,601 reads/sample in
 109 the soil dataset). Mantel tests were then used to calculate the correlation between the rarefied
 110 GU^A distance matrices compared to GU^A calculated on the non-rarefied control. Each rarefaction
 111 depth was calculated across 10 iterations.

112

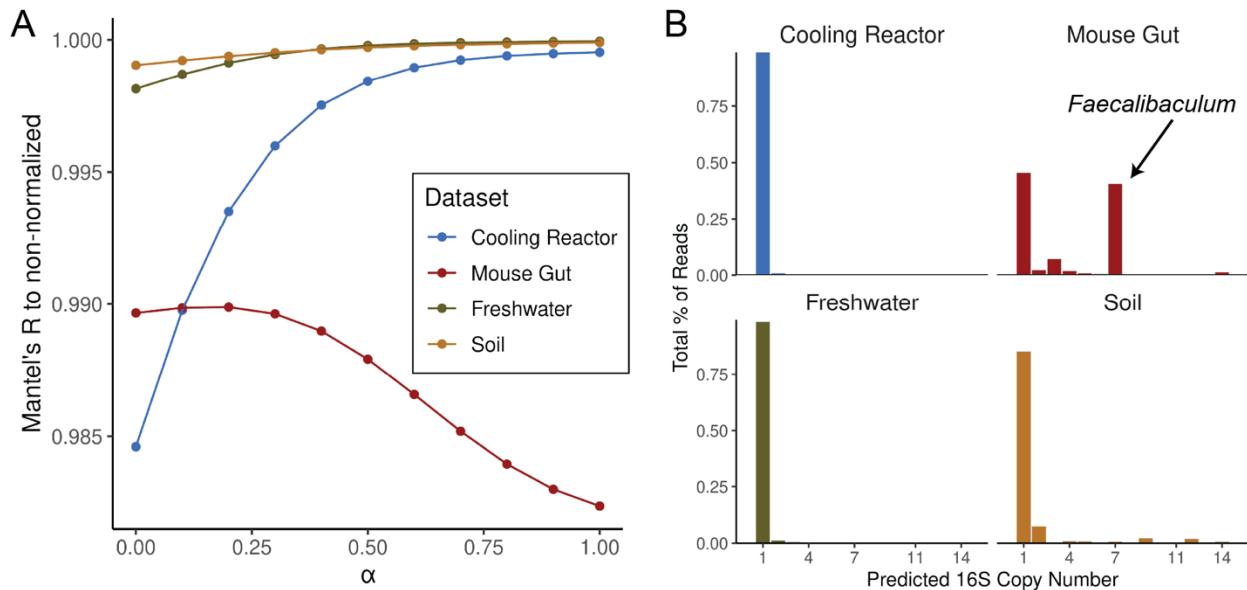
113



114

115 *Figure S5. Additional parameters which weakly influence computation time for GU^A.* A) GU^A
 116 was calculated 50 times across six sample sizes (2, 5, 15, 20, 30) with a constant of 2,000 ASVs
 117 and one calculated α (though unweighted Unifrac is also calculated by default). B) GU^A was
 118 calculated 50 times across four alpha parameter sizes (1, 2, 4, 8; note unweighted Unifrac is also
 119 calculated by default) with a constant of 2,000 ASVs and 10 samples. In both panels, R^2 is
 120 derived from a linear model between the x and y axes.

121



122

123 *Figure S6. 16S rRNA gene copy number-normalization had a negligible effect on GU⁴.* A)
124 PICRUSt2 (v2.6) was used to normalize sequencing reads within each sample by the predicted
125 16S rRNA gene copy number [15]. GU⁴ was then calculated across α values from 0 to 1. Mantel
126 tests were used to assess correlations between the GU⁴ distance matrices calculated from copy
127 number-normalized dataset and those calculated from non-normalized controls. For both
128 analyses, rarefaction was carried out to the minimum sequencing depth for each dataset across 10
129 iterations. The y-axis is truncated to highlight the differences among correlations, all of which
130 exceed a Mantel's R of 0.98. B) Percentage of total reads assigned to ASVs at each predicted
131 16S copy number. Predicted copy numbers extended up to 27 copies per genome, however,
132 categories with negligible representation are not shown. In the mouse gut dataset, ASVs (and
133 reads) with seven predicted 16S copies per genome primarily belonged to the genus
134 *Faecalibaculum*, with one ASV assigned to *Escherichia-Shigella* and one to the class Clostridia.

135

136

Package/Software	Version	Citation
R	4.3.3	[17]
RStudio	2024.12.1+563	[18]
tidyverse	2.0.0	[19]
phyloseq	1.52.0	[10]
vegan	2.7-1	[12]
GUniFrac*	1.8.1	[13]
ggtree	3.16.0	[20]
patchwork	1.3.1	[21]
NatParksPalettes	0.2.0	[22]
ape	5.8-1	[9]
broom	1.0.8	[23]
corrr	0.4.4	[24]
renv	1.0.5	[25]
microbenchmark	1.5.0	[14]
ggpubr	0.6.1	[26]
dada2	1.36.0	[5]
MAFFT	7.520	[7]
FastTree	2.1.11	[8]
cutadapt	5.1	[27]
extraDistr	1.10.0	[11]
PICRUSt2	2.6.0	[15, 28]
biomformat	1.30.0	[29]

137 *Table S1.* Software and packages used in analysis. Note that GUniFrac was modified slightly to
 138 make incorporating absolute abundances more apparent; this version can be installed via Github
 139 at <https://github.com/MarschmiLab/GUniFrac>.

140

141 **Supporting References**

- 142 1. Zhang K et al. Absolute microbiome profiling highlights the links among microbial
143 stability, soil health, and crop productivity under long-term sod-based rotation. *Biol Fertil
144 Soils* 2022;58:883–901. <https://doi.org/10.1007/s00374-022-01675-4>
- 145 2. Props R et al. Measuring the biodiversity of microbial communities by flow cytometry.
146 *Methods in Ecology and Evolution* 2016;7:1376–1385. [https://doi.org/10.1111/2041-210X.12607](https://doi.org/10.1111/2041-
147 210X.12607)
- 148 3. Pendleton A, Wells M, Schmidt ML. Upwelling periodically disturbs the ecological
149 assembly of microbial communities in the Laurentian Great Lakes. 2025. bioRxiv, 2025. ,
150 2025.01.17.633667
- 151 4. Barlow JT, Bogatyrev SR, Ismagilov RF. A quantitative sequencing framework for absolute
152 abundance measurements of mucosal and luminal microbial communities. *Nat Commun*
153 2020;11:2590. <https://doi.org/10.1038/s41467-020-16224-6>
- 154 5. Callahan BJ et al. DADA2: High-resolution sample inference from Illumina amplicon data.
155 *Nat Methods* 2016;13:581–583. <https://doi.org/10.1038/nmeth.3869>
- 156 6. Quast C et al. The SILVA ribosomal RNA gene database project: improved data processing
157 and web-based tools. *Nucleic Acids Research* 2013;41:D590–D596.
158 <https://doi.org/10.1093/nar/gks1219>
- 159 7. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7:
160 Improvements in Performance and Usability. *Molecular Biology and Evolution*
161 2013;30:772–780. <https://doi.org/10.1093/molbev/mst010>
- 162 8. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees
163 for Large Alignments. *PLOS ONE* 2010;5:e9490.
164 <https://doi.org/10.1371/journal.pone.0009490>

- 165 9. Paradis E et al. ape: Analyses of phylogenetics and evolution. 2023.
- 166 10. McMurdie PJ, Holmes S. phyloseq: An R package for reproducible interactive analysis and
167 graphics of microbiome census data. *PLoS ONE* 2013;8:e61217.
- 168 11. Wolodzko T. extraDistr: Additional univariate and multivariate distributions. 2023.
- 169 12. Oksanen J et al. vegan: Community ecology package. 2022.
- 170 13. Chen J et al. Associating microbiome composition with environmental covariates using
171 generalized UniFrac distances. *Bioinformatics* 2012;28:2106–2113.
<https://doi.org/10.1093/bioinformatics/bts342>
- 172 14. Mersmann O. microbenchmark: Accurate timing functions. 2024.
- 173 15. Wright RJ, Langille MGI. PICRUSt2-SC: an update to the reference database used for
174 functional prediction within PICRUSt2. *Bioinformatics* 2025;41:btaf269.
<https://doi.org/10.1093/bioinformatics/btaf269>
- 175 16. Props R et al. Absolute quantification of microbial taxon abundances. *ISME J* 2017;11:584–
176 587. <https://doi.org/10.1038/ismej.2016.117>
- 177 17. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R
178 Foundation for Statistical Computing, 2022.
- 179 18. RStudio Team. RStudio: Integrated Development Environment for R. Boston, MA:
180 RStudio, PBC., 2020.
- 181 19. Wickham H. tidyverse: Easily install and load the tidyverse. 2023.
- 182 20. Xu S et al. Ggtree: A serialized data object for visualization of a phylogenetic tree and
183 annotation data. *iMeta* 2022;1:e56. <https://doi.org/10.1002/imt2.56>
- 184 21. Pedersen TL. patchwork: The composer of plots. 2024.
- 185 22. Blake K. NatParksPalettes: Color palettes inspired by national parks. 2022.

- 188 23. Robinson D, Hayes A, Couch S. broom: Convert statistical objects into tidy tibbles. 2023.
- 189 24. Kuhn M, Jackson S, Cimentada J. corrr: Correlations in R. 2022.
- 190 25. Ushey K, Wickham H. renv: Project environments. 2024.
- 191 26. Kassambara A. ggpubr: ggplot2 based publication ready plots. 2023.
- 192 27. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
- 193 *EMBnet.journal* 2011;17:10–12. <https://doi.org/10.14806/ej.17.1.200>
- 194 28. Douglas GM et al. PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol*
- 195 2020;38:685–688. <https://doi.org/10.1038/s41587-020-0548-6>
- 196 29. McMurdie PJ, Paulson JN. biomformat: An interface package for the BIOM file format.
- 197 2023.
- 198