Title: Greengenes2 enables a shared data universe for microbiome studies

Authors: Daniel McDonald[1], Yueyu Jiang[2], Metin Balaban[3], Kalen Cantrell[4], Qiyun Zhu[5,6], Antonio Gonzalez[1], James T. Morton[7], Giorgia Nicolaou[8], Donovan H. Parks[9], Søren Karst[10], Mads Albertsen[11], Philip Hugenholtz[9], Todd DeSantis[12], Se Jin Song[13], Andrew Bartko[13], Aki S. Havulinna[14,15], Pekka Jousilahti[14], Susan Cheng[16,17], Mike Inouye[18,19], Teemu Niiranen[14,20], Mohit Jain[21], Veikko Salomaa[14], Leo Lahti[22], Siavash Mirarab[2], Rob Knight[1,4,13,23,*]

Affiliations:
[1] Department of Pediatrics, UC San Diego School of Medicine, La Jolla, CA, USA
[2] Department of Electrical and Computer Engineering, UC San Diego, La Jolla, CA, USA
[3] Bioinformatics and System Biology Program, UC San Diego, La Jolla, CA, USA
[4] Department of Computer Science and Engineering, UC San Diego, La Jolla, CA, USA
[5] School of Life Sciences, Arizona State University, Tempe, AZ, USA
[6] Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University, Tempe, AZ, USA
[7] Biostatistics & Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, USA
[8] Halicioglu Data Science Institute, UC San Diego, La Jolla, CA, USA
[9] Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia
[10] Department of Obstetrics and Gynecology, Columbia University, New York, NY, USA
[11] Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark
[12] Department of Informatics, Second Genome, Brisbane, CA, USA
[13] Center for Microbiome Innovation, Jacobs School of Engineering, UC San Diego, La Jolla, CA, USA
[14] Finnish Institute for Health and Welfare, Helsinki, Finland
[15] Institute for Molecular Medicine Finland, FIMM-HiLIFE, Helsinki, Finland
[16] Division of Cardiology, Brigham and Women's Hospital, Boston, MA, USA
[17] Cedars-Sinai Medical Center, Los Angeles, CA, USA
[18] Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, VIC, Australia
[19] Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[20] Division of Medicine, Turku University Hospital and University of Turku, Turku, Finland
[21] Sapient Bioanalytics LLC, San Diego, CA, USA
[22] Department of Computing, University of Turku, Turku, Finland
[23] Department of Bioengineering, UC San Diego, La Jolla, CA, USA
[*] Corresponding author

Address correspondence to Rob Knight (robknight@eng.ucsd.edu)

Abstract:

16S rRNA and shotgun metagenomics studies typically yield different results, usually attributed to biases in PCR amplification of 16S rRNA genes. Here, we introduce Greengenes2 and show that differences in reference phylogeny are more important. By inserting sequences into a

47 whole-genome phylogeny, we show that 16S rRNA and shotgun metagenomic data generated
48 from the same samples agree in principal coordinates space, taxonomy, and in phenotype effect
49 size when analyzed with the same tree.
50
51 **Body:**
52 Shotgun metagenomics and 16S rRNA gene amplicon (16S) studies are widely used in
53 microbiome research, but investigators using different methods typically find their results hard to
54 reconcile. This lack of standardization across methods limits the utility of the microbiome for
55 reproducible biomarker discovery.
56
57 A key problem is that whole-genome resources and rRNA resources depend on different
58 taxonomies and phylogenies. For example, Web of Life (WoL) [1] and the Genome Taxonomy
59 Database (GTDB) [2] provide whole-genome trees that cover only a small fraction of known
60 bacteria and archaea, while SILVA [3] and Greengenes [4] are more comprehensive but most often
61 not linked to genome records.
62
63 We reasoned that an iterative approach could yield a single massive reference tree that unifies
64 these different data layers (e.g., genome and 16S rRNA records). We began with a whole-
65 genome catalog of 15,953 bacterial and archeal genomes evenly sampled from NCBI, and
66 reconstructed an accurate phylogenomic tree by summarizing evolutionary trajectories of 380
67 global marker genes using the new workflow uDance. This work, namely Web of Life version 2
68 (WoL2), represents a significant upgrade from the previously released WoL1 (10,575 genomes)
69 [15]. Then, we added 18,356 full-length 16S rRNA sequences from the Living Tree Project
70 January 2022 release [6] ,1,725,274 near-complete 16S rRNA genes from Karst et al. [7] and the
71 EMP500 [8], and all full-length 16S rRNA sequences from GTDB r207 to the genome-based
72 backbone with uDance v1.1.0, producing a genome supported phylogeny with 16S rRNA

73 explicitly represented. Finally, we inserted 23,113,447 short V4 16S rRNA Deblur v1.1.0 [9]
74 amplicon sequence variants (ASVs) from Qiita (retrieved Dec. 14, 2021) [10] as well as
75 mitochondria and chloroplast 16S from SILVA v138 using DEPP v0.3[11]. This final step
76 represents ASVs from over 300,000 public and private samples in Qiita including the entirety of
77 the Earth Microbiome Project [12] and American Gut Project/Microsetta [13] (Fig. 1A). Our use of
78 uDance ensured the genome-based relationships are kept fixed and relationships between full-
79 length 16S rRNA sequences are inferred. For short fragments, we kept genome and full-length
80 relationships fixed and inserted fragments independently from each other. Following
81 deduplication and quality control on fragment placement, this yielded a tree covering 21,074,442
82 sequences from 31 different Earth Microbiome Project Ontology (EMPO) EMPO_3
83 environments, of which 46.5% of species-level leaves were covered by a complete genome.
84 Taxonomic labels were decorated onto the phylogeny using tax2tree v1.1 [4]. The input taxonomy
85 for decoration used GTDB r207, combined with the Living Tree Project January 2022 release.
86 Taxonomy was harmonized prioritizing GTDB including preserving the polyphyletic labelings of
87 GTDB (see also Online Methods). Taxonomy will be updated every six months using the latest
88 versions of GTDB and LTP.
89
90 Greengenes2 is much larger than past resources in its phylogenetic coverage, as compared to
91 SILVA (Fig. 1B), Greengenes (Fig. S1A) or GTDB (Fig. S1B). Moreover, because our amplicon
92 library is linked to environments labeled with Earth Microbiome Project Ontology (EMPO)
93 categories, we can easily identify the environments that contain samples that can fill out the
94 tree. Because MAG assembly efforts can only cover abundant taxa, we plotted for each EMPO
95 category the amount of new branch length added to the tree by taxa whose minimum
96 abundance is 1% in each sample (Fig. 1C). The results show which environment types on
97 average will best yield new metagenome assembled genomes (MAGs), and also show which
98 environments harbor individual samples that will have a large impact when sequenced.

Past efforts to reconcile 16S and shotgun datasets have led to non-overlapping distributions and only techniques such as Procrustes analysis can even show relationships between the results [14]. On two large human stool cohorts [13,15] where both 16S and shotgun data were generated on the same samples, we find that Bray-Curtis [16] (non-phylogenetic) ordination fails to reconcile at the feature level (Fig. 1D) and is poor at the genus level (Fig. 1E, S1C). However, UniFrac [17], a phylogenetic method, used with our Greengenes2 tree provides better concordance (Fig. 1F, S1D). To examine applicability of Greengenes2 to non-human environments, we next computed both Bray-Curtis and weighted UniFrac at the feature level on the 16S and shotgun data from the Earth Microbiome Project [8]. As with the human data, we observe better concordance with the use of the Greengenes2 phylogeny (Fig. S2), despite limited representation of whole genomes from non-human sources as these environments are not as well characterized in general.

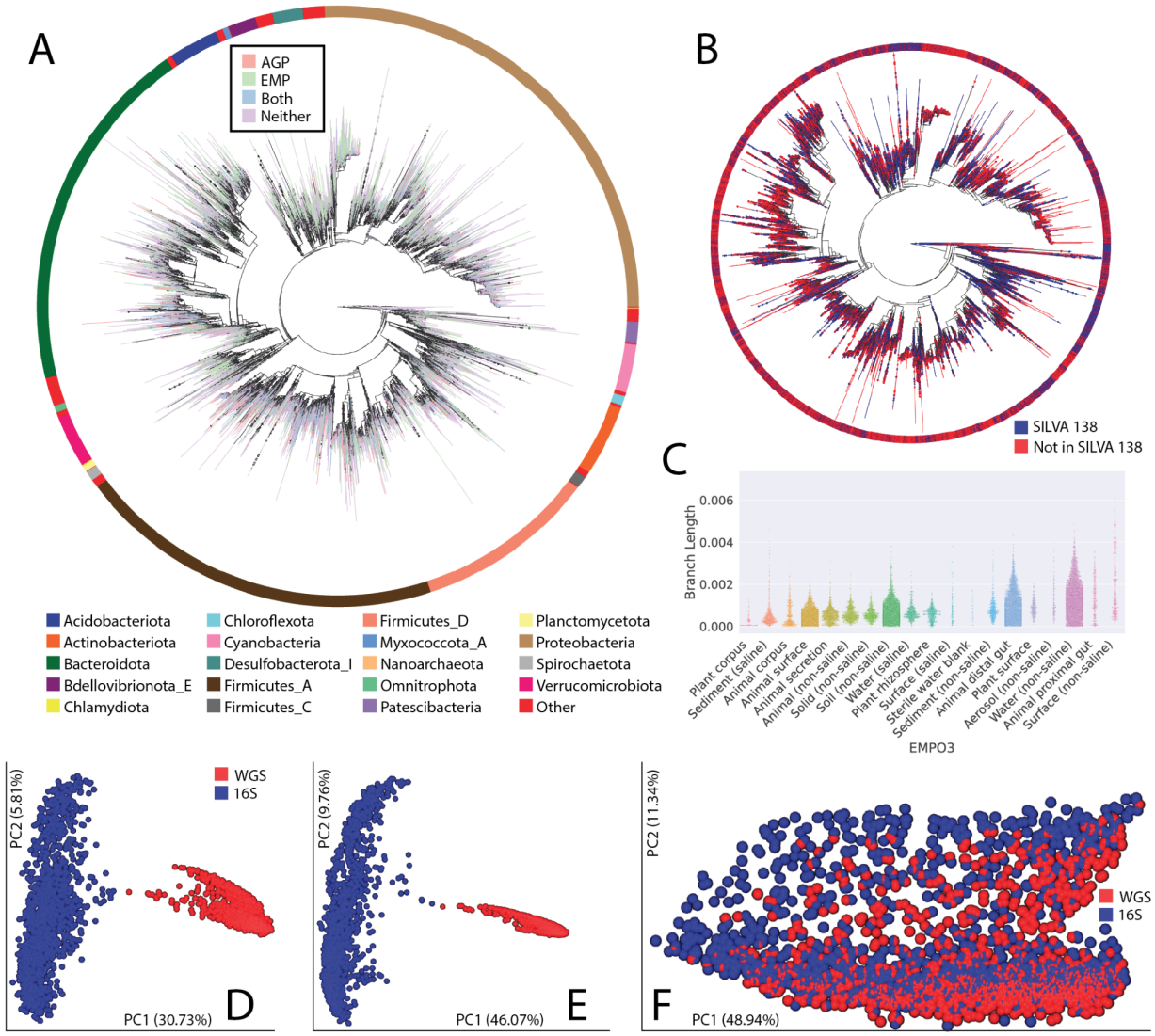


Figure 1. (A) The Greengenes2 phylogeny rendered using Empress [18] with amplicon sequence variant multifurcations collapsed, tip color indicating representation in the American Gut Project (AGP), the Earth Microbiome Project (EMP), both or neither, and with the top 20 represented phyla depicted in the outer bar. (B) The same collapsed phylogeny, colored by the presence or absence of a best BLAST [19] hit from SILVA 138.The bar depicts the same coloring as the tips. (C) Earth Microbiome Project samples and the amount of novel branch length, normalized by the total backbone branch length, added to the tree through amplicon sequence variant fragment placement. Note that sample counts are not even across EMPO3 categories. (D) Bray-

Curtis applied to paired 16S V4 rRNA amplicon sequence variants and whole genome shotgun samples from The Human Diet Microbiome Initiative subset of The Microsetta Initiative. (E) Same data as (D) but computing Bray-Curtis on genus collapsed data. (F) Same data as (D-E) but using weighted UniFrac at the ASV and genome identifier level.

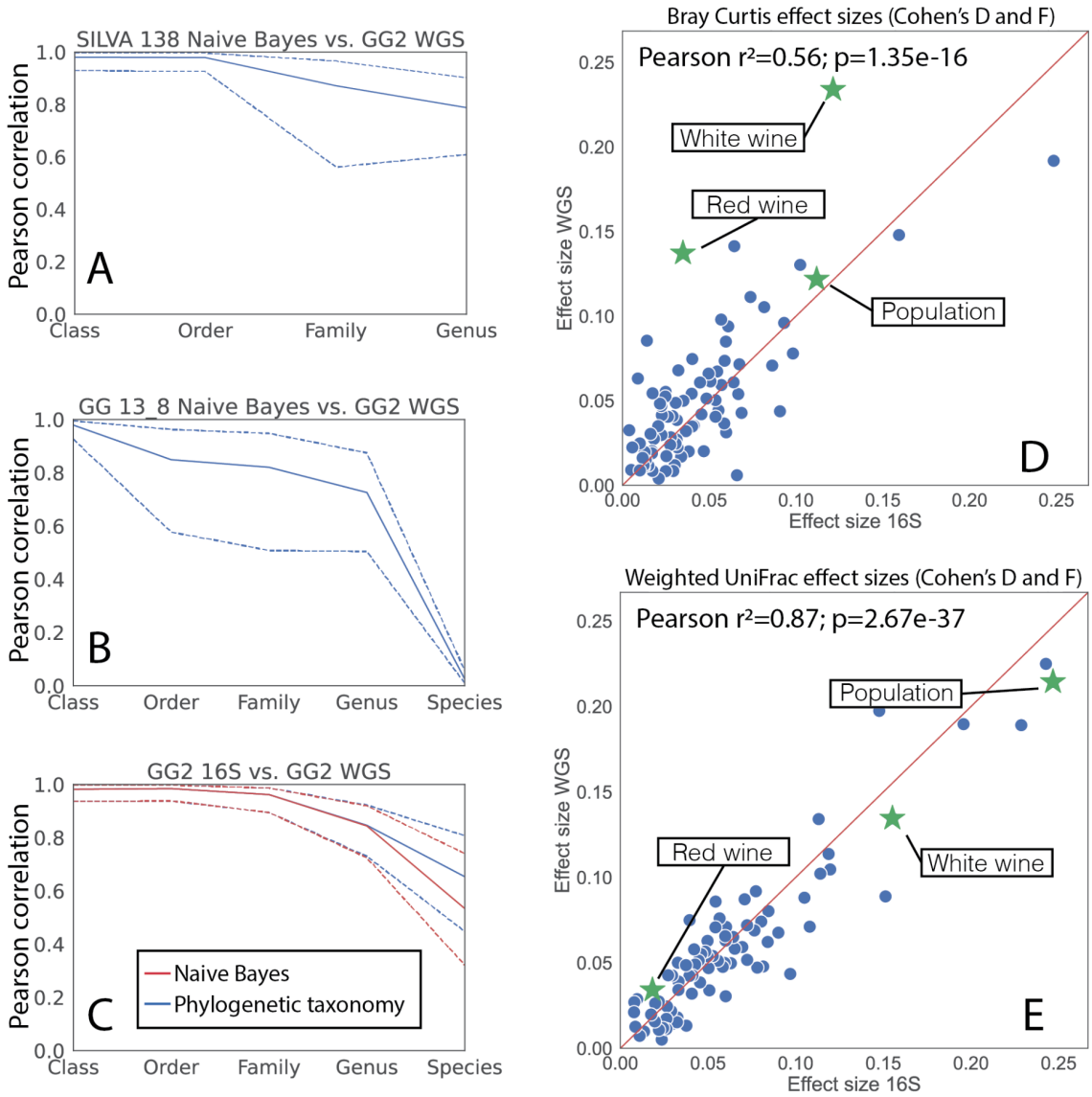
We also find that the per-sample shotgun and 16S taxonomy relative abundance profiles are concordant even to the species level. We first computed taxonomy profiles for shotgun data using the Woltka pipeline [20]. Using a Naive Bayes classifier from q2-feature-classifier v2022.2 [21] to compare GTDB r207 taxonomy results at each level down to genus against SILVA v138 (Fig. 2A) or down to species against Greengenes v13_8 (Fig. 2B), no species-level reconciliation was possible. In contrast, Greengenes2 provided excellent concordance at the genus level (Pearson r=0.85) and good concordance at the species level (Pearson r=0.65) (Fig. 2C). Interestingly, the tree is now sufficiently complete that exact matching of 16S ASVs followed by reading the taxonomy off the tree performs even better than the Naive Bayes Classifier (Naive Bayes; Pearson r=0.54 at species, r=0.84 at genus).

Finally, a critical reason to assign taxonomy is downstream use of biomarkers and indicator taxa. Microbiome science has been described as having a reproducibility crisis [22], but much of this problem stems from incompatible methods [23]. We initially used the The Human Diet Microbiome Initiative (THDMI) dataset, which is a multipopulation expansion of The Microsetta Initiative [13] that contains samples with paired 16S and shotgun preparations, to test whether a harmonized resource would provide concordant rankings for the variables that affect the human microbiome similarly. Using Greengenes2, the concordance was good with Bray-Curtis (Fig. 2D; Pearson r^2=0.56), better using UniFrac with different phylogenies (SILVA 138 and Greengenes2; Fig S1E; Pearson r^2=0.77), and excellent with UniFrac on the same phylogeny

146 (Fig. 2E; Pearson r^2=0.87). We confirmed these results with an additional cohort [15] (Fig. S1FG).

147 Intriguingly, the ranked effect sizes across different cohorts were concordant.

148



149

150 *Figure 2. (A-C) Per sample taxonomy comparisons between 16S and whole genome shotgun*

151 *profiles from THDMI. The solid bar depicts the 50th percentile, the dashed lines are 25th and*

152 *75th percentiles. (A) 16S taxonomy assessed with SILVA 138 using the default q2-feature-*

153 *classifier Naive Bayes model (note: SILVA does not annotate at the species level). (B) 16S*

154 *taxonomy assessment with Greengenes 13_8 using the default q2-feature-classifier Naive*

155 *Bayes model. (C) 16S taxonomy assessment performed by reading the lineages directly from*
156 *the phylogeny or through Naive Bayes trained on the V4 regions of the Greengenes2 backbone.*
157 *(D-E) Effect size calculations performed with Evident on paired 16S and whole genome shotgun*
158 *samples from THDMI. Calculations performed at maximal resolution, using ASVs for 16S and*
159 *genome identifiers for shotgun. The data represented here are human gut microbiome samples.*
160 *The stars denote variables which are drawn out specifically in the plot (e.g., Population), and*
161 *were arbitrarily selected as comparison points to help highlight differences between panels (D)*
162 *and (E). (D) Bray-Curtis distances. (E) Weighted normalized UniFrac.*
163
164 Taken together, these results show that use of a consistent, integrated taxonomic resource
165 dramatically improves the reproducibility of microbiome studies using different data types, and
166 allows variables of large versus small effect to be reliably recovered in different populations.
167
168 ONLINE METHODS
169
170 *Phylogeny construction*
171 Web of Life version 2 [1] (a tree inferred using genome-wide data) was used as the starting
172 backbone. Full length 16S sequences from the Living Tree Project [6], full length mitochondria
173 and chloroplast from SILVA 138 [3], full length 16S from GTDB r207 [2], full length 16S from Karst
174 et al [7], and full length 16S from the EMP 500 [8] (samples selected and sequenced specifically for
175 Greengenes2) were collected and deduplicated. Sequences were then aligned using UPP [24]
176 and gappy sequences with less than 1000bp were removed. The resulting set of 321,210
177 unique sequences were used with uDance v1.1.0 to update the Web of Life 2 (WoL2) backbone.
178 Briefly, uDance updates an existing tree with new sequences and (unlike placement methods)
179 also infers the relationship of existing sequences. uDance has two modes: one that allows
180 updates to the backbone and one that keeps the backbone fixed, where the former mode is

intended for use with whole genomes. In our analyses, we kept the backbone tree (inferred using genomic data) fixed. To extend the genomic tree with 16S data, we identified 13,249 (out of 15,953 total) genomes in the WoL2 backbone tree with at least one 16S copy and used them to train a Deep-learning Enabled Phylogenetic Placement (DEPP) model with the weighted average method detailed below to handle multiple copies. We then used DEPP to insert all 16S copies of all genomes into the backbone and measured the distance between the genome position and the 16S position. We removed copies that were placed far further than others, as identified using a 2-means approach with centroids equals to at least 13 branches. We repeated this process in a second round. Then, for every remaining genome, we selected as its representative the copy with the minimum placement error and computing the consensus with ties. At the end, we were left with 12,344 unique 16S sequences across all the WoL2 genomes. For tree inference, uDance used IQ-TREE2 [25] in fast tree search with model GTR+Γ after removing duplicate sequences.

Next, we collected 16S V4 ASVs from Qiita [10] using redbiom [26] (query performed December 14, 2021) from contexts “Deblur_2021.09-Illumina-16S-V4-90nt-dd6875”, “Deblur_2021.09-Illumina-16S-V4-100nt-50b3a2”, “Deblur_2021.09-Illumina-16S-V4-125nt-92f954”, “Deblur_2021.09-Illumina-16S-V4-150nt-ac8c0b”, “Deblur_2021.09-Illumina-16S-V4-200nt-0b8b48”, “Deblur_2021.09-Illumina-16S-V4-250nt-8b2bff” and aligned them to the existing 16S alignment of sequences in WoL2 using UPP, setting the maximum alignment subset size to 200 (to help with scalability). The collected 16S V4 ASVs are aligned to the V4 region of the existing "backbone" alignments. A DEPP model was then trained on the full length 16S sequences from the backbone. DEPP constructs a Neural network model that embeds sequences in high dimensional spaces such that embedded points resemble the phylogeny in their distances. Such a model then allows insertion of new sequences into a tree using distance-based phylogenetic insertion method APPLES-2 [27]. The ASVs from redbiom were then inserted into

207 the backbone using the trained DEPP model. To enable analyses of large datasets, we used a
208 clustering approach with DEPP: we trained an ensemble of DEPP models corresponding to
209 different parts of the tree and used a classifier to detect the correct subtree. During training, for
210 species with multiple 16S, all the copies are mapped to the same leaf in the backbone tree. To
211 train the DEPP models with multiple sequences mapped to a leaf, each site in the sequences is
212 encoded as a probability vector of four nucleotides across all the copies.
213
214 *Integrating the GTDB and Living Tree Project taxonomies*
215 GTDB and Living Tree Project are not directly compatible due to differences in their curation. As
216 a result, it is not always possible to map a species from one resource to the other, either
217 because parts of a species lineage are not present, are described using different names, or
218 have an ambiguous association due to polyphyletic taxa in GTDB (e.g., Firmicutes_A,
219 Firmicutes_B, etc, see https://gtdb.ecogenomic.org/faq#why-do-some-family-and-higher-rank-
220 names-end-with-an-alphabetic-suffix for more detail). We integrated taxonomic data from LTP
221 into GTDB as LTP includes species that are not yet represented in GTDB. Additionally, GTDB is
222 actively curated, while LTP generally uses the NCBI taxonomy. To account for these
223 differences, we first mapped any species that had a perfect species name association and
224 revised its ancestral lineage to match GTDB. Next, we generated lineage rewrite rules using the
225 GTDB record metadata. Specifically, we limited the metadata to records that are GTDB
226 representatives and NCBI type material, and then defined a lineage renaming from the recorded
227 NCBI taxonomy to the GTDB taxonomy. These rewrite rules were applied from most to least
228 specific taxa, and through this mechanism we could revise much of the higher ranks of LTP. We
229 then identified *incertae sedis* records in LTP that we could not map, removed their lineage
230 strings and did not attempt to provide taxonomy for them, instead opting to rely on downstream
231 taxonomy decoration to resolve their lineages. Next, any record that was ambiguous to map
232 was split into a secondary taxonomy for use in backfilling in the downstream taxonomy

233 decoration. Finally, we instrumented numerous consistency checks in the taxonomy through the
234 process to capture inconsistent parents in the taxonomic hierarchy, consistent numbers of ranks
235 in a lineage and ensuring the resulting taxonomy was a strict hierarchy.
236
237 *Taxonomy decoration*
238 The original tax2tree algorithm was not well suited for a large volume of species level records in
239 the backbone, as the algorithm requires an internal node to place a name. If two species are
240 siblings, the tree would lack a node to contain the species label for both taxa. To account for
241 this, we updated the algorithm to insert “placeholder” nodes with zero branch length as the
242 parents of backbone records, which could accept these species labels. We further updated
243 tax2tree to operate directly on .jplace data [28], preserving edge numbering of the original edges
244 prior to adding “placeholder” nodes. To support LTP records which could not be integrated into
245 GTDB, we instrumented a secondary taxonomy mode for tax2tree. Specifically, following the
246 standard decoration, backfilling and name promotion procedures, we determine on a per record
247 basis for the secondary taxonomy what portion of the lineage is missing, and place the missing
248 labels on the placeholder node. We then issue a second round of name promotion using the
249 existing tax2tree methods.
250
251 The actual taxonomy decoration occurs on the backbone tree, which contains only full length
252 16S records, and does not contain the amplicon sequence variants (ASV). This is done as ASV
253 placements are independent, do not modify the backbone, and would substantially increase the
254 computational resources required. After the backbone is decorated, fragment placements from
255 DEPP are resolved using a multifurcation strategy using the balanced-parentheses library
256 (https://github.com/biocore/improved-octo-waddle/).
257
258 *Phylogenetic collapse for visualization*

259 We are unaware of phylogenetic visualization software that can display a tree with over
260 20,000,000 tips. To produce the visualizations in figure 1, we reduced the dimension of the tree
261 by collapsing fragment multifurcations to single nodes, dropping the tree to 522,849 tips.
262
263 *MAG target environments*
264 A feature table for the 27,015 16S rRNA V4 90nt Earth Microbiome Project samples was
265 obtained from redbiom. The amplicon sequence variants (ASV) were filtered to the overlap of
266 ASVs present in Greengenes2. Any feature with < 1% relative abundance within a sample was
267 removed. The feature table was then rarefied to 1,000 sequences per sample. The amount of
268 novel branch length was then computed, per sample, by summing the branch length of each
269 ASV's placement edge. The per sample branch length was then normalized by the total tree
270 branch length (excluding length contributed by ASVs).
271
272 *Per sample taxonomy correlations*
273 All comparisons used the THDMI[13] 16S and Woltka processed shotgun data. These data were
274 accessed from Qiita study 10317, and filtered the set of features that overlap with Greengenes2
275 using the QIIME 2[29] q2-greengenes2 plugin. 16S taxonomy was assessed using either a
276 traditional Naive Bayes classifier with q2-feature-classifier and default references from QIIME 2
277 2022.2, or by reading the lineage directly from the phylogeny. To help improve correlation
278 between SILVA and Greengenes2, and Greengenes and Greengenes2, we stripped
279 polyphyletic labelings from those data; we did not strip polyphyletic labels from the phylogenetic
280 taxonomy comparison or the Greengenes2 16S vs. Greengenes2 WGS Naive Bayes
281 comparison. Shotgun taxonomy was determined by the specific observed genome records.
282 Once the 16S taxonomy was assigned, those tables as well as the WGS Woltka WoL version 2
283 table were collapsed at the species, genus, family, order, and class levels. We then computed a
284 minimum relative abundance per sample in the THDMI dataset. In each sample, we removed

285 any feature, either 16S or WGS, below the per sample minimum (i.e., max(min(16S),
286 min(WGS))), forming a common minimal basis for taxonomy comparison. Following filtering,
287 Pearson correlation was computed per sample using SciPy [30]. These correlations were
288 aggregated per 16S taxonomy assignment method, and by each taxonomic rank. The 25th, 50th
289 and 75th percentiles were then plotted with Matplotlib [31].
290
291 *Principal coordinates*
292 THDMI Deblur 16S and Woltka processed shotgun sequencing data, against WoL version 2,
293 were obtained from Qiita study 10317. Both feature tables were filtered against Greengenes2
294 2022.10, removing any feature not present in the tree. For the genus collapsed plot (figure 1e),
295 both the 16S and WGS data features were collapsed using the same taxonomy. For all three
296 figures, the 16S data were subsampled, with replacement, to 10,000 sequences per sample.
297 The WGS data were subsampled, with replacement, to 1,000,000 sequences per sample. Bray-
298 Curtis and Weighted UniFrac, and PCoA were computed using q2-diversity 2022.2. The
299 resulting coordinates were visualized with q2-emperor [32].
300
301 The Earth Microbiome Project "EMP500" 16S and Woltka processed shotgun sequencing data,
302 against WoL version 2, were obtained from Qiita study 13114. Both feature tables were filtered
303 against Greengenes2 2022.10. The 16S data were subsampled, with replacement, to 1,000
304 sequences per sample. The WGS data were subsampled, with replacement, to 50,000
305 sequences per sample. The sequencing depth for WGS data was selected based on Figure S6
306 of Shaffer et al.[8] which noted low levels of read recruitment to publicly available whole genomes.
307 Bray-Curtis and Weighted UniFrac, and PCoA were computed using q2-diversity 2022.2. The
308 resulting coordinates were visualized with q2-emperor.
309
310 *Effect size calculations*

311 Similar to principal coordinates, the THDMI data were rarefied to 9,000 and 2,000,000
312 sequences per sample for 16S and WGS respectively. Bray-Curtis and weighted normalized
313 UniFrac were computed on both sets of data. The variables for THDMI were subset to those
314 with at least two category values having more than 50 samples. For UniFrac with SILVA, figure
315 S1E, we performed fragment insertion using q2-fragment-insertion [33] into the standard QIIME 2
316 SILVA reference, followed by rarefaction to 9,000 sequences per sample, and then computed
317 weighted normalized UniFrac.
318
319 For FINRISK, the data were rarefied to 1,000 and 500,000 sequences per sample for 16S and
320 WGS, respectively. A different depth was used to account for the overall lower amount of
321 sequencing data for FINRISK. As with THDMI, the variables selected were reduced to those
322 with at least two category values having more than 50 samples.
323
324 Support for computing paired effect sizes is part of the QIIME2 Greengenes2 plugin, q2-
325 greengenes2, which performs effect size calculations using Evident
326 (https://github.com/biocore/evident/).
327
328 *Data access*
329 The official location of the Greengenes2 releases is http://ftp.microbio.me/greengenes_release/.
330 The data are released under a BSD-3 clause license. A QIIME 2 plugin is available to facilitate
331 use with the resource that can be obtained from https://github.com/biocore/q2-greengenes2/
332 (version 2023.3; DOI: 10.5281/zenodo.7758134). Taxonomy construction, decoration, and
333 release processing is part of https://github.com/biocore/greengenes2 (version 2023.3; DOI:
334 10.5281/zenodo.7758138). uDance is available at GitHub:
335 https://github.com/balabanmetin/uDance (version v1.1.0; DOI: 10.5281/zenodo.7758289).
336 Phylogeny insertion using DEPP is available at https://github.com/yueyujiang/DEPP (version

0.3; DOI: 10.5281/zenodo.7768798); the trained model accessioned with Zenodo at 10.5281/zenodo.7416684. The THDMI data are part of Qiita study 10317, and EBI accession PRJEB11419. The FINRISK data are available under EGAD00001007035. Finally, an interactive website to explore the Greengenes2 data is available at https://greengenes2.ucsd.edu.

Acknowledgements

This work was supported in part by NSF XSEDE BIO210103, NSF RAPID 20385.09, NIH 1R35GM14272, NIH U19AG063744, NIH U24DK131617, NIH DP1-AT010885 and Emerald Foundation 3022. JTM was funded by the intramural research program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). The Human Diets and Microbiome Initiative dataset was generated through support from Danone Nutricia Research and the Center for Microbiome Innovation.

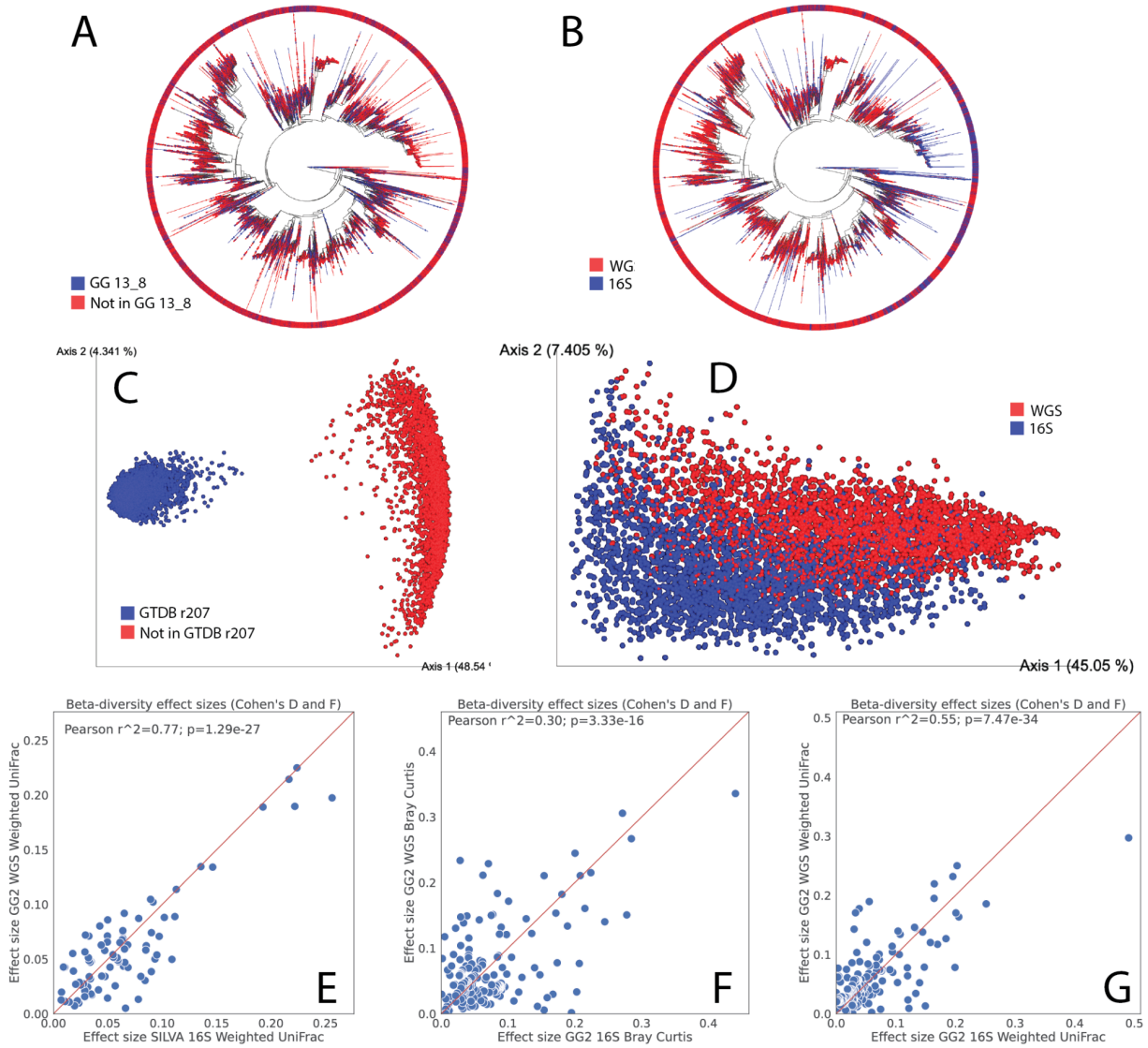


Figure S1. (A) Best BLAST hit for Greengenes 13_8 99% OTUs against Greengenes2. (B) Best BLAST hit for GTDB r207 SSU sequences against Greengenes2. (C) The FINRISK 16S and WGS data combined, collapsed to genus relative to Greengenes2, with Bray-Curtis computed followed by Principal Coordinates Analysis, colored by technical preparation. (D) The same data as (C) but using weighted UniFrac. (E) Effect sizes of the THDMI data using the SILVA 138 phylogeny for 16S data, and the Greengenes2 phylogeny for WGS data. (F) Effect sizes of the FINRISK data using Bray-Curtis. (G) The same data as (E) but using Weighted UniFrac.

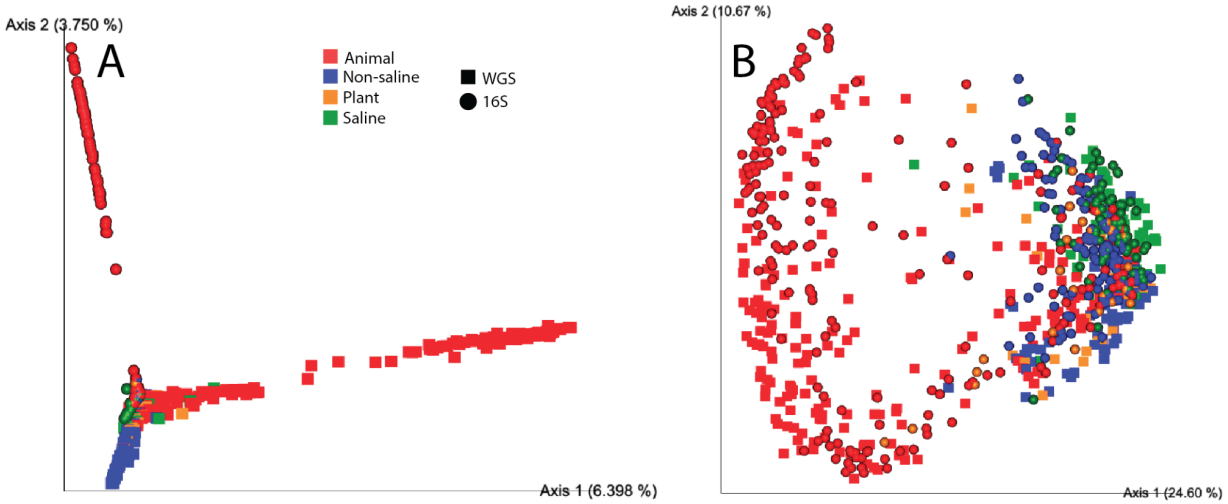


Figure S2. (A) The Earth Microbiome Project 16S and WGS data combined, and with Bray-Curtis computed at the feature level followed by Principal Coordinates Analysis, colored by Earth Microbiome Project Ontology level 2. Squares represent WGS data, and spheres represent 16S data. (B) The same data and coloring as in (A) but using weighted UniFrac.

REFERENCES

1. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).
2. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
3. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).
4. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).
5. Hugenholtz, P., Chuvochina, M., Oren, A., Parks, D. H. & Soo, R. M. Prokaryotic taxonomy and nomenclature in the age of big sequence data. *ISME J.* **15**, 1879–1892 (2021).

6. Ludwig, W. *et al.* Release LTP_12_2020, featuring a new ARB alignment and improved 16S rRNA tree for prokaryotic type strains. *Syst. Appl. Microbiol.* **44**, 126218 (2021).
7. Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat. Methods* **18**, 165–169 (2021).
8. Shaffer, J. P. *et al.* Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. *Nat Microbiol* **7**, 2128–2150 (2022).
9. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2**, (2017).
10. Gonzalez, A. *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**, 796–798 (2018).
11. Jiang, Y., McDonald, D., Knight, R. & Mirarab, S. Scaling deep phylogenetic embedding to ultra-large reference trees: a tree-aware ensemble approach. *bioRxiv* 2023.03.27.534201 (2023) doi:10.1101/2023.03.27.534201.
12. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
13. McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3**, (2018).
14. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
15. Salosensaari, A. *et al.* Taxonomic signatures of cause-specific mortality risk in human gut microbiome. *Nat. Commun.* **12**, 2671 (2021).
16. Bray, J. R. & Curtis, J. T. An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* **27**, 325–349 (1957).
17. Sfiligoi, I., Armstrong, G., Gonzalez, A., McDonald, D. & Knight, R. Optimizing UniFrac with OpenACC Yields Greater Than One Thousand Times Speed Increase. *mSystems* **7**, e0002822 (2022).

403 18. Cantrell, K. *et al.* EMPress Enables Tree-Guided, Interactive, and Exploratory Analyses of
404 Multi-omic Data Sets. *mSystems* **6**, (2021).
405 19. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
406 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
407 20. Zhu, Q. *et al.* Phylogeny-Aware Analysis of Metagenome Community Ecology Based on
408 Matched Reference Genomes while Bypassing Taxonomy. *mSystems* **7**, e0016722 (2022).
409 21. Bokulich, N. A. *et al.* Optimizing taxonomic classification of marker-gene amplicon
410 sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90 (2018).
411 22. Schloss, P. D. Identifying and Overcoming Threats to Reproducibility, Replicability,
412 Robustness, and Generalizability in Microbiome Research. *MBio* **9**, (2018).
413 23. Sinha, R. *et al.* Assessment of variation in microbial community amplicon sequencing by
414 the Microbiome Quality Control (MBQC) project consortium. *Nat. Biotechnol.* **35**, 1077–
415 1086 (2017).
416 24. Nguyen, N.-P. D., Mirarab, S., Kumar, K. & Warnow, T. Ultra-large alignments using
417 phylogeny-aware profiles. *Genome Biol.* **16**, 124 (2015).
418 25. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference
419 in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
420 26. McDonald, D. *et al.* redbiom: a Rapid Sample Discovery and Feature Characterization
421 System. *mSystems* **4**, (2019).
422 27. Balaban, M., Jiang, Y., Roush, D., Zhu, Q. & Mirarab, S. Fast and accurate distance-based
423 phylogenetic placement using divide and conquer. *Mol. Ecol. Resour.* **22**, 1213–1227
424 (2022).
425 28. Matsen, F. A., Hoffman, N. G., Gallagher, A. & Stamatakis, A. A format for phylogenetic
426 placements. *PLoS One* **7**, e31009 (2012).
427 29. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data
428 science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).

30. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
31. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
32. Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience* **2**, 16 (2013).
33. Janssen, S. *et al.* Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. *mSystems* **3**, (2018).