1   **Title:** Greengenes2 enables a shared data universe for microbiome studies

2

3   **Authors:** Daniel McDonald[1], Yueyu Jiang[2], Metin Balaban[3], Kalen Cantrell[4], Qiyun Zhu[5,6],

4   Antonio Gonzalez[1], James T. Morton[7], Giorgia Nicolaou[8], Donovan H. Parks[9], Søren Karst[10],

5   Mads Albertsen[11], Phil Hugenholtz[9], Todd DeSantis[12], Siavash Mirarab[2], Rob Knight[1,4,13,14,*]

6

7   **Affiliations:**

8   [1] Department of Pediatrics, UC San Diego School of Medicine, La Jolla, CA, USA

9   [2] Department of Electrical and Computer Engineering, UC San Diego, La Jolla, CA, USA

10   [3] Bioinformatics and System Biology Program, UC San Diego, La Jolla, CA, USA

11   [4] Department of Computer Science and Engineering, UC San Diego, La Jolla, CA, USA

12   [5] School of Life Sciences, Arizona State University, Tempe, AZ, USA

13   [6] Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University,

14   Tempe, AZ, USA

15   [7] Biostatistics & Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child

16   Health and Human Development, National Institutes of Health, Bethesda, MD, USA

17   [8] Halicioglu Data Science Institute, UC San Diego, La Jolla, CA, USA

18   [9] Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The

19   University of Queensland, St Lucia, QLD 4072, Australia

20   [10] Center for Microbial Communities, Aalborg University, Aalborg, Denmark

21   [11] Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark

22   [12] Department of Informatics, Second Genome, Brisbane, CA, USA

23   [13] Center for Microbiome Innovation, Jacobs School of Engineering, UC San Diego, La Jolla,

24   CA, USA

25   [14] Department of Bioengineering, UC San Diego, La Jolla, CA, USA

26   [*] Corresponding author

27

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

29

30   **Abstract:**

31   16S rRNA and shotgun metagenomics studies typically yield different results, traditionally

32   thought to be due to biases in amplification. We show that differences in reference phylogeny

33   are more important. By inserting sequences into a whole-genome phylogeny, we show that 16S

34   rRNA and shotgun metagenomic data generated from the same samples agree in principal

35   coordinates space, taxonomy, and in phenotype effect size when analyzed with the same tree.

36

37   **Body:**

38   Shotgun metagenomics and 16S rRNA gene amplicon (16S) studies are widely used in

39   microbiome research, but investigators using different methods typically find their results hard to

40   reconcile. This lack of standardization across methods limits the utility of the microbiome for

41   reproducible biomarker discovery.

42

43   A key problem is that whole-genome resources and rRNA resources depend on different

44   taxonomies and phylogenies. For example, Web of Life (WoL) [1] and the Genome Taxonomy

45   Database (GTDB) [2] provide whole-genome trees that cover only a small fraction of known

46   bacteria and archaea, while SILVA [3] and Greengenes [4] are more comprehensive but not fully

47   linked to genome records.

48

49   We reasoned that an iterative approach could yield a massive reference tree that unifies these

50   different data layers. We began with a whole-genome catalog of 15,953 bacterial and archeal

51   genomes evenly sampled from NCBI, and reconstructed an accurate phylogenomic tree by

52   summarizing evolutionary trajectories of 380 global marker genes using the new workflow

53   uDance. This work, namely Web of Life version 2 (WoL2), represents a significant upgrade from

54   the previously released WoL1 (10,575 genomes) [1]. Then, we added 18,356 full-length rRNA

55   amplicons from the Living Tree Project January 2022 release [5] and 1,725,274 near-complete

56   16S rRNA genes from Karst et al. [6] and the EMP500 [7] with uDance v1.1.0, then added all full-

57   length 16S sequences from GTDB r207, and finally inserted 23,113,447 short V4 16S rRNA

58   Deblur v1.1.0 [8] amplicon sequence variants from Qiita (retrieved Dec. 14, 2021) [9] as well as

59   mitochondria and chloroplast 16S from SILVA v138 using DEPP v0.3 [10], including everything

60   from the Earth Microbiome Project [11] and American Gut Project/Microsetta [12] (Fig. 1A). Our use

61   of uDance ensured the genome-based relationships are kept fixed and relationships between

62   full-length 16S sequences are inferred. For short fragments, we kept genome and full length

63   relationships fixed and inserted fragments independently from each other. Following

64   deduplication and quality control on fragment placement, this yielded a tree covering 21,074,442

65   sequences from 31 different Earth Microbiome Project Ontology (EMPO) EMPO_3

66   environments, of which 46.5% of species-level leaves were covered by a complete genome.

67   Taxonomic labels were decorated onto the phylogeny using tax2tree v1.1 [4]. The input taxonomy

68   for decoration used GTDB r207, combined with the Living Tree Project January 2022 release.

69   Taxonomy was harmonized prioritizing GTDB including preserving the polyphyletic labelings of

70   GTDB (see also Online Methods). Taxonomy will be updated every six months using the latest

71   versions of GTDB and LTP.

72

73   Our resource is much larger than past resources in its phylogenetic coverage, as compared to

74   the last release of Greengenes (Fig. 1B), SILVA (Fig. S1A) or GTDB (Fig. S1B). However,

75   because our amplicon library is linked to environments labeled with Earth Microbiome Project

76   Ontology (EMPO) categories, we can easily identify the environments that contain samples that

77   can fill out the tree. Because MAG assembly efforts can only cover abundant taxa, we plotted

78   for each EMPO category the amount of new branch length added to the tree by taxa whose

79   minimum abundance is 1% in each sample (Fig. 1C). The results show which environment

80   types on average will best yield new metagenome assembled genomes (MAGs), and also show

81   which environments harbor individual samples that will have a large impact when sequenced.

82

83   Past efforts to reconcile 16S and shotgun datasets have led to non-overlapping distributions and

84   only techniques such as Procrustes analysis can even show relationships between the results

85   [13]. On two large human stool cohorts [12,14] where both 16S and shotgun data were generated on

86   the same samples, we find that Bray-Curtis [15] (non-phylogenetic) ordination fails to reconcile at

87   the feature level (Fig. 1D) and is poor at the genus level (Fig. 1E, S1C). However, UniFrac [16], a

88   phylogenetic method, used with our Greengenes2 tree provides far better concordance (Fig. 1F,
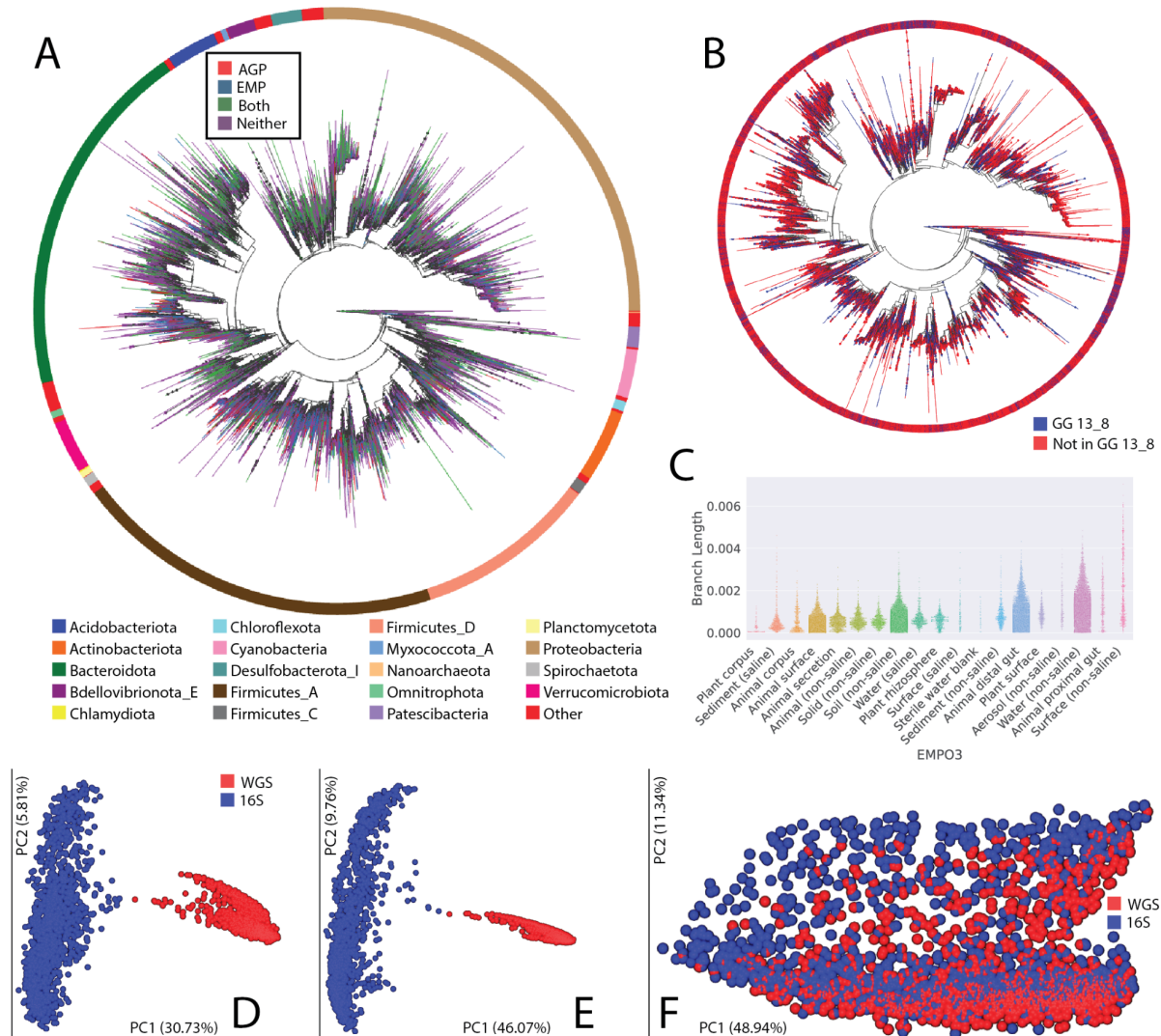
89   S1D).

Figure 1. (A) The Greengenes2 phylogeny rendered using Empress [17] 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 [18] hit from Greengenes 13_8 99% OTUs. 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. (D) Bray Curtis applied to paired 16S V4 rRNA amplicon sequence variants and whole genome shotgun samples from The Healthy Microbiome Diet

100   *Initiative subset of The Microsetta Initiative. (E) Same data as (D) but computing Bray Curtis on*

101   *genus collapsed data. (F) Same data as (D-E) but using weighted UniFrac.*

102

103   We also find that the per-sample shotgun and 16S taxonomy concordances are excellent even

104   to the species level. We first computed taxonomy profiles for shotgun data using the Woltka

105   pipeline [19]. Using a Naive Bayes classifier from q2-feature-classifier v2022.2 [20] to compare

106   GTDB r207 taxonomy results at each level against SILVA v138 (Fig. 2A) or Greengenes v13_8

107   (Fig. 2B), no species-level reconciliation was possible. In contrast, Greengenes2 provided

108   excellent concordance at the genus level (Pearson r=0.85) and good concordance at the

109   species level (Pearson r=0.65) (Fig. 2C). Interestingly, the tree is now sufficiently complete that

110   exact matching of 16S ASVs followed by reading the taxonomy off the tree performs even better

111   than the Naive Bayes Classifier (Naive Bayes; Pearson r=0.54 at species, r=0.84 at genus).

112

113   Finally, a critical reason to assign taxonomy is downstream use of biomarkers and indicator

114   taxa. Microbiome science has been described as having a reproducibility crisis [21], but much of

115   this problem stems from incompatible methods [22]. We initially used the The Human Diet

116   Microbiome Initiative (THDMI) dataset, which is a multipopulation expansion of The Microsetta

117   Initiative [12] that contains samples with paired 16S and shotgun preparations, to test whether a

118   harmonized resource would provide concordant rankings for the variables that affect the human

119   microbiome similarly. Using Greengenes2, the concordance was good with Bray-Curtis (Fig. 2D;

120   Pearson r^2=0.56), better using UniFrac with different phylogenies (SILVA 138 and

121   Greengenes2; Fig S1E; Pearson r^2=0.77), and excellent with UniFrac on the same phylogeny

122   (Fig. 2E; Pearson r^2=0.87). We confirmed these results with an additional cohort [14] (Fig. S1FG).

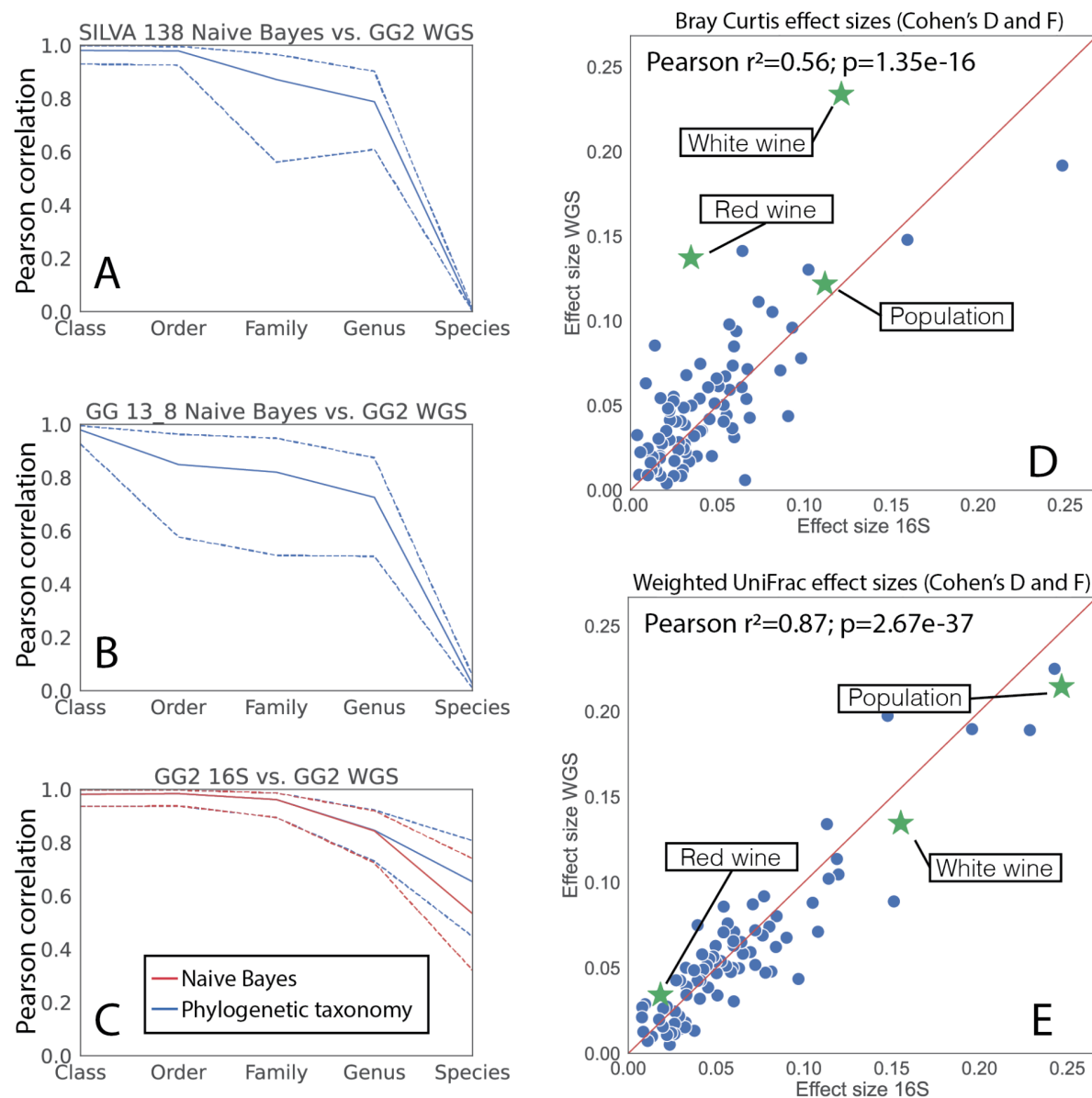123   Intriguingly, the ranked effect sizes across different cohorts were concordant.

124

125

Figure 2. (A-C) Per sample taxonomy comparisons between 16S and whole genome shotgun profiles from THDMI. The solid bar depicts the 50th percentile, the dashed lines are 25th and 75th percentiles. (A) 16S taxonomy assessed with SILVA 138 using the default q2-feature-classifier Naive Bayes model. (B) 16S taxonomy assessment with Greengenes 13_8 using the default q2-feature-classifier Naive Bayes model. (C) 16S taxonomy assessment performed by reading the lineages directly from the phylogeny or through Naive Bayes trained on the V4

132   *regions of the Greengenes2 backbone. (D-E) Effect size calculations performed with Evident on*

133   *paired 16S and whole genome shotgun samples from THDMI. Calculations performed at*

134   *maximal resolution, using ASVs for 16S and genome identifiers for shotgun. (D) Bray Curtis*

135   *distances. (E) Weighted normalized UniFrac.*

136

137   Taken together, these results show that use of a consistent, integrated taxonomic resource

138   dramatically improves the reproducibility of microbiome studies using different data types, and

139   allows variables of large versus small effect to be reliably recovered in different populations.

140

141   ONLINE METHODS

142

143   *Phylogeny construction*

144   Web of Life version 2 [1] (a tree inferred using genome-wide data) was used as the starting

145   backbone. Full length 16S sequences from the Living Tree Project [5], full length mitochondria

146   and chloroplast from SILVA 138 [3], full length 16S from GTDB r207 [2], full length 16S from Karst

147   et al [6], and full length 16S from the EMP 500 [7] (samples selected and sequenced specifically for

148   Greengenes2) were collected and deduplicated. Sequences were then aligned using UPP [23]

149   and gappy sequences with less than 1000bp were removed. The resulting set of 321,210

150   unique sequences were used with uDance v1.1.0 to update the Web of Life 2 (WoL2) backbone.

151   Briefly, uDance updates an existing tree with new sequences and (unlike placement methods)

152   also infers the relationship of existing sequences. uDance has two modes: one that allows

153   updates to the backbone and one that keeps the backbone fixed. In our analyses, we kept the

154   backbone tree (inferred using genomic data) fixed. To extend the genomic tree with 16S data,

155   we identified 13,249 genomes in the WoL2 backbone tree with at least one 16S copy and used

156   them to train a DEPP model with the weighted average method detailed below to handle

157   multiple copies. We then used DEPP to insert all 16S copies of all genomes into the backbone

158 and measured the distance between the genome position and the 16S position. We removed

159 copies that were placed far further than others, as identified using a 2-means approach with

160 centroids equals to at least 13 branches. We repeated this process a second round. Then, for

161 every remaining genome, we selected as its representative the copy with the minimum

162 placement error and computing the consensus when there were ties. At the end, we are left with

163 12,344 unique 16S sequences across all the WoL2 genomes. For tree inference, uDance used

164 IQ-TREE2 [24] in fast tree search with model GTR+Γ after removing duplicate sequences.

165

166 Next, we collected 16S V4 ASVs from Qiita [9] using redbiom [25] (query performed December 14,

167 2021) from contexts "Deblur_2021.09-Illumina-16S-V4-90nt-dd6875", "Deblur_2021.09-Illumina-

168 16S-V4-100nt-50b3a2", "Deblur_2021.09-Illumina-16S-V4-125nt-92f954", "Deblur_2021.09-

169 Illumina-16S-V4-150nt-ac8c0b", "Deblur_2021.09-Illumina-16S-V4-200nt-0b8b48",

170 "Deblur_2021.09-Illumina-16S-V4-250nt-8b2bff" and aligned them to the existing 16S alignment

171 of sequences in WoL2 using UPP, setting the maximum alignment subset size to 200 (to help

172 with scalability). The collected 16S V4 ASVs are aligned to the V4 region of the existing

173 "backbone" alignments. A DEPP model was then trained on the full length 16S sequences from

174 the backbone. DEPP constructs a Neural network model that embeds sequences in high

175 dimensional spaces such that embedded points resemble the phylogeny in their distances.

176 Such a model then allows insertion of new sequences into a tree using distance-based

177 phylogenetic insertion method APPLES-2 [26]. The ASVs from redbiom were then inserted into

178 the backbone using the trained DEPP model. To enable analyses of large datasets, we used a

179 clustering approach with DEPP: we trained an ensemble of DEPP models corresponding to

180 different parts of the tree and used a classifier to detect the correct subtree. During training, for

181 species with multiple 16S, all the copies are mapped to the same leaf in the backbone tree. To

182 train the DEPP models with multiple sequences mapped to a leaf, each site in the sequences is

183 encoded as a probability vector of four nucleotides across all the copies.

184

185   *Integrating the GTDB and Living Tree Project taxonomies*

186   GTDB and Living Tree Project are not directly compatible due to differences in their curation. As

187   a result, it is not always possible to map a species from one resource to the other, either

188   because parts of a species lineage are not present, are described using different names, or

189   have an ambiguous association due to polyphyletic taxa in GTDB. GTDB is actively curated,

190   while LTP generally uses the NCBI taxonomy. To account for these differences, we first mapped

191   any species that had a perfect species name association and revised its ancestral lineage to

192   match GTDB. Next, we generated lineage rewrite rules using the GTDB record metadata.

193   Specifically, we limited the metadata to records which are GTDB representatives and NCBI type

194   material, and then defined a lineage renaming from the recorded NCBI taxonomy to the GTDB

195   taxonomy. These rewrite rules were applied from most to least specific taxa, and through this

196   mechanism we could revise much of the higher ranks of LTP. We then identified incertae sedis

197   records in LTP which we could not map, removed their lineage strings and did not attempt to

198   provide taxonomy for them, instead opting to rely on downstream taxonomy decoration to

199   resolve their lineages. Next, any record which was ambiguous to map was split into a secondary

200   taxonomy for use in backfilling in the downstream taxonomy decoration. Finally, we

201   instrumented numerous consistency checks in the taxonomy through the process to capture

202   inconsistent parents in the taxonomy hierarchy, consistent numbers of ranks in a lineage and

203   ensuring the resulting taxonomy was a strict hierarchy.

204

205   *Taxonomy decoration*

206   The original tax2tree algorithm was not well suited for a large volume of species level records in

207   the backbone, as the algorithm requires an internal node to place a name. If two species are

208   siblings, the tree would lack a node to contain the species label for both taxa. To account for

209   this, we updated the algorithm to insert "placeholder" nodes with zero branch length as the

210   parents of backbone records, which could accept these species labels. We further updated

211   tax2tree to operate directly on .jplace data [27], preserving edge numbering of the original edges

212   prior to adding "placeholder" nodes. To support LTP records which could not be integrated into

213   GTDB, we instrumented a secondary taxonomy mode for tax2tree. Specifically, following the

214   standard decoration, backfilling and name promotion procedures, we determine on a per record

215   basis for the secondary taxonomy what portion of the lineage is missing, and place the missing

216   labels on the placeholder node. We then issue a second round of name promotion using the

217   existing tax2tree methods.

218

219   The actual taxonomy decoration occurs on the backbone tree, which contains only full length

220   16S records, and does not contain the amplicon sequence variants (ASV). This is done as ASV

221   placements are independent, do not modify the backbone, and would substantially increase the

222   computational resources required. After the backbone is decorated, fragment placements from

223   DEPP are resolved using a multifurcation strategy using the balanced-parentheses library

224   (https://github.com/biocore/improved-octo-waddle/).

225

226   *Phylogenetic collapse for visualization*

227   We are unaware of phylogenetic visualization software that can display a tree with over

228   20,000,000 tips. To produce the visualizations in figure 1, we reduced the dimension of the tree

229   by collapsing fragment multifurcations to single nodes, dropping the tree to 522,849 tips.

230

231   *MAG target environments*

232   A feature table for the 27,015 16S rRNA V4 90nt Earth Microbiome Project samples was

233   obtained from redbiom. The amplicon sequence variants (ASV) were filtered to the overlap of

234   ASVs present in Greengenes2. Any feature with < 1% relative abundance within a sample was

235   removed. The feature table was then rarefied to 1,000 sequences per sample. The amount of

236   novel branch length was then computed, per sample, by summing the branch length of each

237   ASV's placement edge. The per sample branch length was then normalized by the total tree

238   branch length (excluding length contributed by ASVs).

239

240   *Per sample taxonomy correlations*

241   All comparisons used the THDMI 16S and Woltka processed shotgun data. These data were

242   accessed from Qiita study 10317, and filtered the set of features which overlap with

243   Greengenes2 using the QIIME 2[28] q2-greengenes2 plugin. 16S taxonomy was assessed using

244   either a traditional Naive Bayes classifier with q2-feature-classifier and default references from

245   QIIME 2 2022.2, or by reading the lineage directly from the phylogeny. To help improve

246   correlation between SILVA and Greengenes2, and Greengenes and Greengenes2, we stripped

247   polyphyletic labelings from those data; we did not strip polyphyletic labels from the phylogenetic

248   taxonomy comparison or the Greengenes2 16S vs. Greengenes2 WGS Naive Bayes

249   comparison. Shotgun taxonomy was determined by the specific observed genome records.

250   Once the 16S taxonomy was assigned, those tables as well as the WGS Woltka WoL version 2

251   table were collapsed at the species, genus, family, order, and class levels. We then computed a

252   minimum relative abundance per sample in the THDMI dataset. In each sample, we removed

253   any feature, either 16S or WGS, below the per sample minimum (i.e., max(min(16S),

254   min(WGS))), forming a common minimal basis for taxonomy comparison. Following filtering,

255   Pearson correlation was computed per sample using SciPy [29]. These correlations were

256   aggregated per 16S taxonomy assignment method, and by each taxonomic rank. The 25th, 50th

257   and 75th percentiles were then plotted with Matplotlib [30].

258

259   *Principal coordinates*

260   THDMI Deblur 16S and Woltka processed shotgun sequencing data, against WoL version 2,

261   were obtained from Qiita study 10317. Both feature tables were filtered against Greengenes2

262   2022.10, removing any feature not present in the tree. For the genus collapsed plot (figure 1e),

263   both the 16S and WGS data features were collapsed using the same taxonomy. For all three

264   figures, the 16S data were subsampled, with replacement, to 10,000 sequences per sample.

265   The WGS data were subsampled, with replacement, to 1,000,000 sequences per sample. Bray

266   Curtis and Weighted UniFrac, and PCoA were computed using q2-diversity 2022.2. The

267   resulting coordinates were visualized with q2-emperor [31].

268

269   *Effect size calculations*

270   Similar to principal coordinates, the THDMI data were rarefied to 9,000 and 2,000,000

271   sequences per sample for 16S and WGS respectively. Bray Curtis and weighted normalized

272   UniFrac were computed on both sets of data. The variables for THDMI were subset to those

273   with at least two category values having more than 50 samples. For UniFrac with SILVA, figure

274   S1E, we performed fragment insertion using q2-fragment-insertion [32] into the standard QIIME 2

275   SILVA reference, followed by rarefaction to 9,000 sequences per sample, and then computed

276   weighted normalized UniFrac.

277

278   For FinRISK, the data were rarefied to 1,000 and 500,000 sequences per sample for 16S and

279   WGS. A different depth was used to account for the overall lower amount of sequencing data for

280   FinRISK. As with THDMI, the variables selected were reduced to those with at least two

281   category values having more than 50 samples.

282

283   Support for computing paired effect sizes is part of the QIIME2 Greengenes2 plugin, q2-

284   greengenes2, which performs effect size calculations using Evident

285   (https://github.com/biocore/evident/).

286

287   *Data access*

288   The official location of the Greengenes2 releases is http://ftp.microbio.me/greengenes_release/.

289   The data are released under a BSD-3 clause license. A QIIME 2 plugin is available to facilitate

290   use with the resource, which can be obtained from https://github.com/biocore/q2-greengenes2/.

291   Taxonomy construction, decoration, and release processing is part of

292   https://github.com/biocore/greengenes2. uDance release v1.1.0 is available at GitHub:

293   https://github.com/balabanmetin/uDance. Phylogeny insertion using DEPP is available at

294   https://github.com/yueyujiang/DEPP; the trained model accessioned with Zenodo at

295   10.5281/zenodo.7416684. The THDMI data are part of Qiita study 10317, and EBI accession

296   PRJEB11419. The FinRISK data are available under EGAD00001007035. Finally, an interactive

297   website to explore the Greengenes2 data is available at https://greengenes2.ucsd.edu.
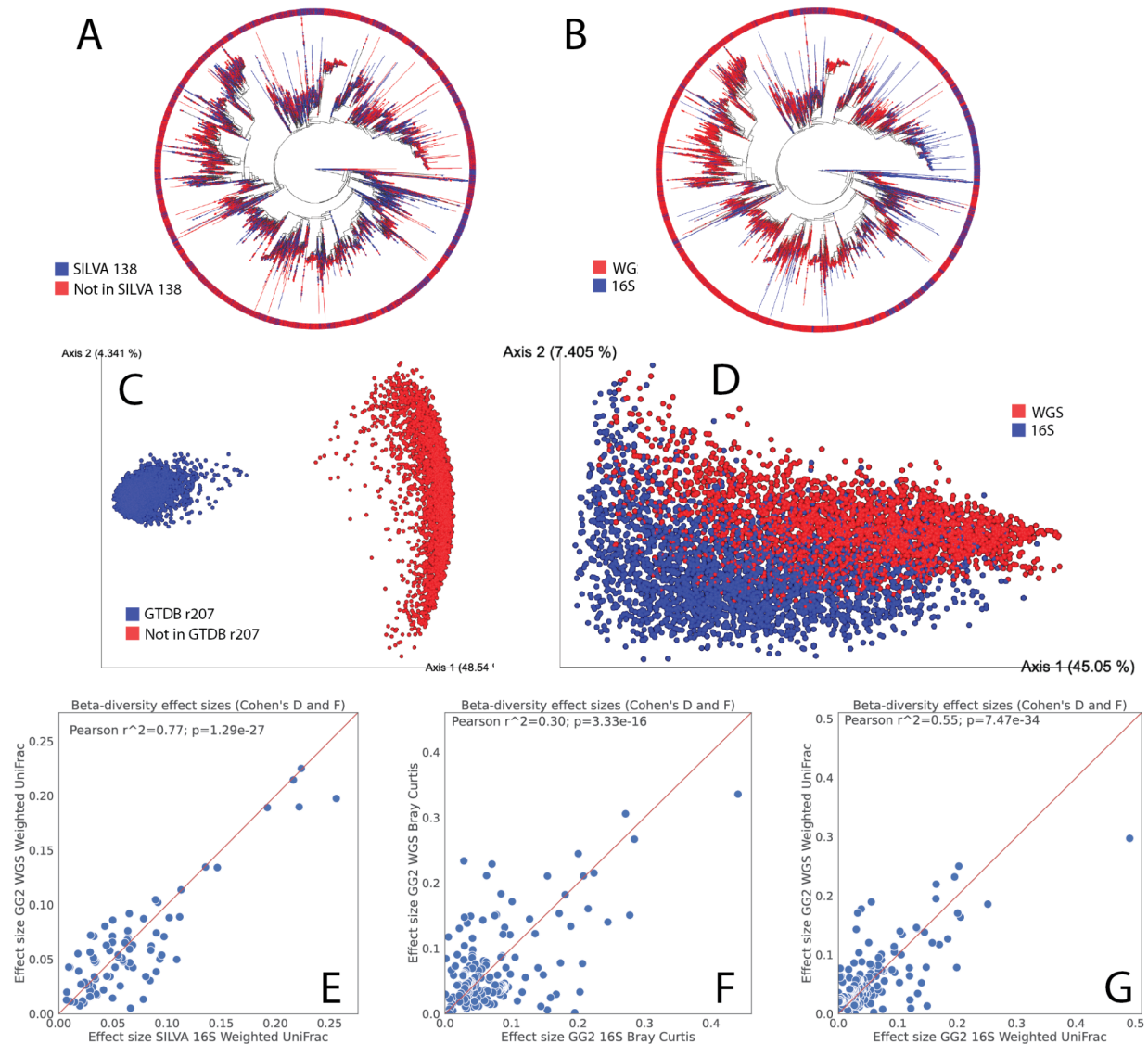
298

299   *Acknowledgements*

Figure S1. (A) Best BLAST hit for SILVA 138 against Greengenes2. (B) Best BLAST hit for GTDB r207 SSU sequences against Greengenes2. (C) The FinRISK 16S and WGS data combined, collapsed to genus, 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.

313

314   REFERENCES

315   1.   Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between

316   domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).

317   2.   Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a

318   phylogenetically consistent, rank normalized and complete genome-based taxonomy.

319   *Nucleic Acids Res.* **50**, D785–D794 (2022).

320   3.   Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data

321   processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).

322   4.   McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological

323   and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).

324   5.   Ludwig, W. *et al.* Release LTP_12_2020, featuring a new ARB alignment and improved

325   16S rRNA tree for prokaryotic type strains. *Syst. Appl. Microbiol.* **44**, 126218 (2021).

326   6.   Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique molecular

327   identifiers with Nanopore or PacBio sequencing. *Nat. Methods* **18**, 165–169 (2021).

328   7.   Shaffer, J. P. *et al.* Standardized multi-omics of Earth's microbiomes reveals microbial and

329   metabolite diversity. *Nat Microbiol* **7**, 2128–2150 (2022).

330   8.   Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns.

331   *mSystems* **2**, (2017).

332   9.   Gonzalez, A. *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**,

333   796–798 (2018).

334   10.   Jiang, Y., Balaban, M., Zhu, Q. & Mirarab, S. DEPP: Deep learning enables extending

335   species trees using single genes. *bioRxiv* (2021) doi:10.1101/2021.01.22.427808.

336   11.   Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity.

337   *Nature* **551**, 457–463 (2017).

338   12. McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome

339   Research. *mSystems* **3**, (2018).

340   13. Human Microbiome Project Consortium. Structure, function and diversity of the healthy

341   human microbiome. *Nature* **486**, 207–214 (2012).

342   14. Salosensaari, A. *et al.* Taxonomic signatures of cause-specific mortality risk in human gut

343   microbiome. *Nat. Commun.* **12**, 2671 (2021).

344   15. Bray, J. R. & Curtis, J. T. An ordination of the upland forest communities of southern

345   Wisconsin. *Ecol. Monogr.* **27**, 325–349 (1957).

346   16. Sfiligoi, I., Armstrong, G., Gonzalez, A., McDonald, D. & Knight, R. Optimizing UniFrac with

347   OpenACC Yields Greater Than One Thousand Times Speed Increase. *mSystems* **7**,

348   e0002822 (2022).

349   17. Cantrell, K. *et al.* EMPress Enables Tree-Guided, Interactive, and Exploratory Analyses of

350   Multi-omic Data Sets. *mSystems* **6**, (2021).

351   18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment

352   search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

353   19. Zhu, Q. *et al.* Phylogeny-Aware Analysis of Metagenome Community Ecology Based on

354   Matched Reference Genomes while Bypassing Taxonomy. *mSystems* **7**, e0016722 (2022).

355   20. Bokulich, N. A. *et al.* Optimizing taxonomic classification of marker-gene amplicon

356   sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90 (2018).

357   21. Schloss, P. D. Identifying and Overcoming Threats to Reproducibility, Replicability,

358   Robustness, and Generalizability in Microbiome Research. *MBio* **9**, (2018).

359   22. Sinha, R. *et al.* Assessment of variation in microbial community amplicon sequencing by

360   the Microbiome Quality Control (MBQC) project consortium. *Nat. Biotechnol.* **35**, 1077–

361   1086 (2017).

362   23. Nguyen, N.-P. D., Mirarab, S., Kumar, K. & Warnow, T. Ultra-large alignments using

363   phylogeny-aware profiles. *Genome Biol.* **16**, 124 (2015).

364   24. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference

365   in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

366   25. McDonald, D. *et al.* redbiom: a Rapid Sample Discovery and Feature Characterization

367   System. *mSystems* **4**, (2019).

368   26. Balaban, M., Jiang, Y., Roush, D., Zhu, Q. & Mirarab, S. Fast and accurate distance-based

369   phylogenetic placement using divide and conquer. *Mol. Ecol. Resour.* **22**, 1213–1227

370   (2022).

371   27. Matsen, F. A., Hoffman, N. G., Gallagher, A. & Stamatakis, A. A format for phylogenetic

372   placements. *PLoS One* **7**, e31009 (2012).

373   28. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data

374   science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).

375   29. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat.*

376   *Methods* **17**, 261–272 (2020).

377   30. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

378   31. Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. EMPeror: a tool for visualizing

379   high-throughput microbial community data. *Gigascience* **2**, 16 (2013).

380   32. Janssen, S. *et al.* Phylogenetic Placement of Exact Amplicon Sequences Improves

381   Associations with Clinical Information. *mSystems* **3**, (2018).

382