

1 **RefSoil: A reference database of soil microbial genomes**

2 Running title: A reference database of soil microbial genomes

3

4 Jinlyung Choi¹, Fan Yang¹, Ramunas Stepanauskas², Erick Cardenas³, Aaron
5 Garoutte⁴, Ryan Williams¹, Jared Flater¹, James M Tiedje⁴, Kirsten S. Hofmockel^{5, 6},
6 Brian Gelder¹, Adina Howe^{1§}

7

8 ¹Agricultural and Biosystems Engineering, Iowa State University, Iowa, United
9 States

10 ²Bigelow Laboratory for Ocean Sciences, Maine, United States

11 Ramunas Stepanauskas

12 ³Department of Microbiology & Immunology, University of British Columbia,
13 Vancouver, Canada

14 ⁴Center for Microbial Ecology, Michigan State University, Michigan, United States

15 ⁵Environmental Molecular Sciences Laboratory, Pacific Northwest National
16 Laboratory, Washington, United States

17 ⁶Department of Ecology, Evolution and Organismal Biology, Iowa State University,
18 Iowa, United States

19

20 §Corresponding author

21

22 Corresponding author addresses:

23 Name: Adina Howe

24 Postal address: 1201 Sukup Hall Ames IA, 50011 United States

25 Telephone: +1-515-294-0176

26 E-mail: adina@iastate.edu

27

28

29 **Subject Categories**

30 Microbial population and community ecology

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46 **Abstract**

47

48 A database of curated genomes is needed to better assess soil microbial
49 communities and their processes associated with differing land management and
50 environmental impacts. Interpreting soil metagenomic datasets with existing
51 sequence databases is challenging because these datasets are biased towards
52 medical and biotechnology research and can result in misleading annotations. We
53 have curated a database of 922 genomes of soil-associated organisms (888 bacteria
54 and 34 archaea). Using this database, we evaluated phyla and functions that are
55 enriched in soils as well as those that may be underrepresented in RefSoil. Our
56 comparison of RefSoil to soil amplicon datasets allowed us to identify targets that if
57 cultured or sequenced would significantly increase the biodiversity represented
58 within RefSoil. To demonstrate the opportunities to access these underrepresented
59 targets, we employed single cell genomics in a pilot experiment to sequence 14
60 genomes. This effort demonstrates the value of RefSoil in the guidance of future
61 research efforts and the capability of single cell genomics as a practical means to fill
62 the existing genomic data gaps.

63

64 **Introduction**

65

66 Microbial populations in the soil impact the structure and fertility of our soils
67 by contributing to nutrient availability, soil stability, plant productivity, and
68 protection against disease and pathogens. Soil microbiology has a rich history of
69 isolating and characterizing representatives from the soil to assess their impacts on
70 soil health and stability. However, despite significant efforts to isolate microbes
71 from the soil, we have accessed only a small fraction of its biodiversity, with
72 estimates of less than 1 to 50 percent of species, even with novel isolation
73 techniques (Van H T Pham and Kim, 2012; Vartoukian *et al.*, 2010; Schoenborn *et al.*,
74 2004; Burmølle *et al.*, 2009; Janssen *et al.*, 2002; Ling *et al.*, 2015). With advances in
75 DNA sequencing technologies, we can now directly interrogate soil metagenomes,
76 allowing for the characterization of soil communities without the need to first
77 cultivate isolates. However, our ability to annotate and characterize the retrieved
78 genes is dependent on the availability of informative reference gene or genome
79 databases. The largest resource of full genome reference sequences for annotating
80 genes is the NCBI RefSeq database (Tatusova *et al.*, 2013), which contains a total of
81 57,993 genomes (Release 74). While broadly useful, this database is not
82 representative of the large majority of soil microbiomes, and the majority of genes
83 in previously published soil metagenomes (65-90%) cannot be annotated against
84 known genes (Delmont *et al.*, 2012; Fierer *et al.*, 2012). Further, contributions to the
85 NCBI databases have largely originated from human health and biotechnology
86 research efforts that can mislead annotations of genes originating from soil
87 microbiomes (e.g., annotations that are clearly not compatible with life in soil).

88 Soil microbiologists are not the first to face the problem of a limited
89 reference database. The NIH Human Microbiome Project (HMP) recognized the
90 critical need for a well-curated reference genome dataset and developed a reference

91 catalog of 3,000 genomes that were isolated and sequenced from human-associated
92 microbial populations(Huttenhower *et al.*, 2012). This publicly available reference
93 set of microbial isolates and their genomic sequences aids in the analysis of human
94 microbiome sequencing data(Segata *et al.*, 2012; Wu *et al.*, 2009) and also provides
95 strains for which isolates (both culture collections and nucleic acids) are available as
96 resources for experiments. Following this successful model of the HMP, we have
97 developed a curated database of reference genome sequences that originate from
98 soil. This database, called RefSoil, represents the current state of knowledge of
99 sequenced soil genomes. In addition to helping us to understand known soil
100 microbiology, RefSoil aids in the identification of knowledge gaps that need to be
101 filled to improve our understanding of soil biodiversity. We discuss the importance
102 the RefSoil database and demonstrate exciting opportunities to expand this
103 database and its applications in the future.

104

105 **Materials and Methods**

106

107 **Creation of RefSoil Database**

108 In order to create a soil-specific reference genome dataset, we targeted genome
109 sequences based on evidence that isolates had origins in soil systems. A total of
110 6,646 complete genomes were obtained from the Genomes OnLine Database (GOLD,
111 <https://gold.jgi.doe.gov>, October 9th, 2014); the GOLD database was chosen because
112 of the availability of metadata, particularly environmental origin, related to genome
113 sequences. These genomes were further selected based on soil-association with the
114 following criteria: (1) Within the GOLD database, organism information and
115 organism metadata (known habitats, ecosystem category, and ecosystem type) was
116 required to identify organism as originating in soil-associated categories. Organisms
117 from marine and deep sea environments were excluded, though these organisms
118 often were identified as soil-associated. Additionally, obligate host-associated
119 pathogens and extremophiles were excluded, as these organisms are unlikely to be
120 present in the absence of their host or in representative soil samples. We
121 considered these organisms to often be under very strong selective pressures that
122 can lead to reduced genomes or high rates of recombination that are difficult to
123 assess soil-specific trends. (2) For organisms that lacked appropriate metadata in
124 GOLD, a Google Custom Search was used to query all available websites for the
125 organism name and soil-related terms (rhizosphere, soil, sand, mud, or nodule).
126 Priorities were placed on the following websites:
127 <http://www.ncbi.nlm.nih.gov/pubmed/>, <http://www.straininfo.net/>,
128 <https://microbewiki.kenyon.edu/index.php/MicrobeWiki>,
129 <https://en.wikipedia.org/>, and <https://scholar.google.com>. Genomes with an
130 association with at least one webpage containing search query phrases were
131 included in RefSoil. This approach was tested with known soil-associated organisms,
132 verifying that it would reliably predict soil association for these organisms. Within
133 resulting genomes, duplicated chromosomes and strains were removed. If multiple
134 genome accession numbers were associated with a single strain, the most recent
135 genome sequence was chosen. NCBI Genbank annotation files associated with each

136 strain were obtained on 2/18/16. The genomes contained within RefSoil are
137 provided in Supplementary Table 1.

138

139 **Characterization of organisms in RefSoil**

140 All RefSoil genome sequences were associated with their NCBI RefSeq
141 accession ID and obtained from NCBI. Using NCBI Genbank CDS annotations, 16S
142 rRNA gene sequences were identified. If multiple 16S rRNA gene sequences were
143 present, the first sequence from the first chromosome (by NCBI index) was selected
144 as representative for the genome. Three genomes lacked annotations of a 16S rRNA
145 gene and a 16S rRNA gene HMM model was used and identified an additional 16S
146 rRNA sequence (Guo *et al.*, 2015). In total, 886 16S rRNA gene sequences were used
147 to build a phylogenetic tree for bacterial genomes within RefSoil (Fig. 1). All
148 bacterial 16S rRNA gene sequences were aligned using RDP's bacterial model
149 (Infernal 1.1.1 (Nawrocki and Eddy, 2013)), and a Maximum-likelihood
150 phylogenetic tree was constructed based on the Jukes-Cantor model by Fasttree
151 (Price *et al.*, 2010) and visualized with Graphlan (R 3.2.2, version 0.9.7) (Fig. 1). The
152 scripts for these approaches are publicly available on https://github.com/germs-lab/ref_soil.
153 RefSoil genomes were extracted from corresponding genomes
154 contained within NCBI RefSeq (release 74). Annotations of taxonomy for RefSoil and
155 RefSeq genes were obtained from NCBI (February 19, 2016) (Supplementary Fig. 1).
156 RefSoil genomes were submitted to RAST and genes were annotated using SEED
157 subsystem ontology (Supplementary Table 1). Among 1,811,233 unique genes,
158 approximately 39.1% of them were assigned to multiple SEED subsystem level 1
159 categories. For the purpose of this study, we included all annotations at subsystem
160 level 1 for each unique gene, which expanded the total gene counts to 2,619,643.
161 The percent abundance of genes from each phylum in RefSoil database was also
162 adjusted accordingly by including the abundance of genes that were assigned to
163 multiple subsystem level 1 categories (Fig. 2 and Supplement Table 2). The number
164 of phyla identified in each subsystem level 1 categories was summarized in
165 Supplemental Table 3.

166

167 **EMP databases used in this study**

168 A total of 15,481 amplicon datasets that form the EMP dataset were available
169 to compare environmental soil amplicons to RefSoil (total 5,594,412 OTUs,
170 clustered at 97%) (Rideout *et al.*, 2014). Soil samples were selected based on their
171 association with soil metadata resulting in a total of 2,476,795 OTUs from 3035
172 samples. A total of 2,158 unique taxonomy assignments were identified for EMP
173 OTUs using the RDP Classifier (Wang *et al.*, 2007) and used to construct a
174 phylogenetic tree with the Graphlan package (R 3.2.2, version 0.9.7) (Fig. 3).
175 Abundances for each taxonomic assignment were calculated as the sum of
176 abundance of OTUs associated with that taxonomy.

177 To evaluate the presence of RefSoil genomes in EMP, EMP and RefSoil 16S
178 rRNA gene amplicons were compared by alignment with BLAST, requiring an
179 alignment with greater than 97% similarity, a minimum alignment of 72 bp, and E-
180 value $\leq 1e-5$. If multiple hits were aligned, the hit with the lowest E-value was
181 selected as representative. Using these criteria, a total of 53,538 EMP OTUs were

182 associated RefSoil 16s rRNA genes (1.4% of total EMP OTUs; 10.2% of all EMP
183 amplicons).

184 Soil order for EMP samples located in the United States (1817 samples) was
185 obtained based on GPS location associated with sample metadata; locations were
186 only considered valid if the coordinates entered were actually located in the United
187 States (1627 samples). Valid GPS points were then located within a soil order by
188 querying the USDA NRCS Global Soil Regions map (Reich) (Accessed on January 26,
189 2016) for soil order plus a Rock/Sand/Ice category using ArcGIS 10.3.1. RefSoil
190 representatives for EMP OTUs were determined by similarity as described above
191 (Supplementary Table 4).

192

193 **Most-wanted soil OTUs**

194 The selection of the most wanted OTUs for RefSoil based on EMP-associated
195 amplicons was based on presence in EMP samples, using the criteria of EMP-RefSoil
196 similarity as described above, and abundance (number of amplicons associated with
197 the OTU in EMP samples). Candidate OTUs were ranked based on its observed
198 frequency in all EMP samples and abundance in EMP amplicons (Top 100 shown in
199 Supplementary Table 5 and 6). Taxonomy was assigned by the RDP classifier (Wang
200 *et al.*, 2007). The 21 OTUs present in both 'the most abundance' and 'the most
201 frequent' are listed in the Table 1.

202

203 **Single cell genomics**

204 For single cell genomics, a soil sample was collected from 0-10 cm depth in a
205 residential garden in Nobleboro, Maine (44° 5'48.10"N, 69°29'10.56"W) on May 5th,
206 2015. Approximately five grams of the sample were mixed with 30 mL sterile-
207 filtered phosphate-saline buffer (PBS), vortexed for 30 s at maximum speed, and
208 centrifuged for 30 s at 2,000 rpm. The obtained supernatant was diluted to below
209 10^5 cell x mL⁻¹ with PBS, pre-screened through a 40 µm mesh-size cell strainer (BD),
210 and incubated with SYTO-9 DNA stain (5 µM; Invitrogen) for 10-60 min. The
211 generation of single amplified genomes (SAGs) and their genomic sequencing were
212 performed by the Bigelow Laboratory Single Cell Genomics Center
213 (scgc.bigelow.org), as previously described (Stepanauskas *et al.*). SAGs representing
214 the "most-wanted list" were selected based on 16S rRNA gene BLAST alignments
215 with greater than 97% similarity over at least 72 bp.

216

217 **Code and sequencing data availability**

218 The analysis code used to generate the results is available from
219 https://github.com/germs-lab/ref_soil. All 14 single-cell sequencing data sets have
220 been deposited at the NCBI. Sequencing data with NCBI accession identifier are
221 listed in Table 2.

222

223

224

225

226

227

228 **Results**

229

230 *Phylogenetic and functional characterization of RefSoil*

231 RefSoil is a manually curated database containing a total 922 genomes of
232 soil-associated organisms, comprised of 888 bacteria and 34 archaea
233 (Supplementary Table 1). The genomes within RefSoil (soil-associated) and RefSeq
234 (all environments) reference databases were compared. Both genome databases
235 contained similar dominant phyla, including Proteobacteria, Firmicutes, and
236 Actinobacteria, comprising over 91% and 88% of genomes in RefSoil and RefSeq
237 bacteria, respectively. RefSoil contained higher proportions of Armatimonadetes,
238 Germmatimonadetes, Thermodesulfobacteria, Acidobacteria, Nitrospirae, and
239 Chloroflexi than RefSeq, suggesting that these phyla may be enriched in the soil or
240 underrepresented in the RefSeq database. A total of eleven RefSeq-associated
241 phylums were absent from RefSoil indicating that these phyla may be absent or
242 difficult to cultivate in soil environments (Supplementary Fig. 1). The 16S rRNA
243 genes for RefSoil bacterial genomes were obtained from NCBI Genbank annotations
244 and aligned to construct a phylogenetic tree representing RefSoil phylogenetic
245 diversity (Fig. 1).

246 RefSoil bacterial and archaeal genomes were further annotated with the
247 Rapid Annotation using Subsystem Technology (RAST, v 2.0(Aziz *et al.*, 2008))
248 (Supplementary Table 1), resulting in the annotation of a total of 1,811,233 RAST-
249 associated genes. RAST annotations, unlike GenBank annotations, include
250 classification into functional ontologies or subsystems, allowing for the comparison
251 of broad functions. Overall, 78% of the bacterial and archaeal genes could be
252 classified into functional subsystems (Supplementary Table 2). For genes associated
253 with each subsystem, we evaluated the phylogenetic origins of RefSoil genes and
254 compared the phyla distribution of annotated genes with those within the
255 cumulative RefSoil database (Fig. 2). If the proportion of genes associated with a
256 phylum within a functional subsystem was greater than its representation in
257 RefSoil, we considered the representation of the phyla *enriched* in this function (Fig.
258 2, Supplementary Table 3). For example, we observed that in most subsystems (15
259 out of 26 subsystems), Proteobacteria-associated genes were enriched relative to
260 their representation within RefSoil (56% of all RefSoil genes associated with
261 Proteobacteria). Additionally, Actinobacteria, Crenarchaeota, and Proteobacteria
262 genes were enriched in functions related to metabolism of aromatic compounds;
263 Firmicutes and Proteobacteria genes were enriched in functions related to iron
264 acquisition and metabolism; and function category dormancy and sporulation was
265 enriched with Firmicutes genes only. These enrichments indicate that organisms
266 from a small number of phyla may have advantages over other organisms in
267 environments where these functions are important.

268

269 *RefSoil compared to sequences originating from soil environments*

270 Genomes in RefSoil represent cultivated strains originating from soils that
271 have been isolated and often characterized in laboratory conditions. In order to
272 estimate their natural abundance in soils, we compared RefSoil to amplicon
273 sequencing datasets from global soil microbiomes in the Earth Microbiome Project

274 (EMP)(Rideout *et al.*, 2014; Gilbert *et al.*, 2014). Amplicons (16S rRNA genes) from
275 EMP datasets were clustered at 97% sequence similarity to determine
276 representative operational taxonomic units (OTUs)(Rideout *et al.*, 2014). We
277 selected only EMP samples that were associated with soil samples, resulting in a
278 total of 3,035 samples. Within these soil samples, we observed that the majority of
279 OTUs were rare (e.g., only observed in a few samples), with 76% of OTUs observed
280 in less than ten soil samples, and 1% of OTUs representing 81% of total abundance
281 in EMP.

282 Comparing RefSoil 16S rRNA genes to EMP amplicons, we observed that the
283 majority of RefSoil 16S rRNA sequences were highly similar to EMP amplicons (over
284 87% of RefSoil bacterial and archaeal 16S rRNA genes shared greater than 97%
285 similarity with a minimum 72 bp alignment length). In contrast, 99% (2,442,432 of
286 2,476,795) of EMP amplicons did not share high similarity (greater than 97%
287 similarity) to RefSoil genes, suggesting that EMP soil samples contain much higher
288 diversity than represented within RefSoil. While RefSoil genomes represent soil
289 microbes that have been well-studied in the laboratory, the EMP amplicons
290 represent microbes that have been observed in soils around the world but not
291 necessarily well-characterized (e.g., limited to the observation of the presence of
292 16S rRNA gene sequences).

293

294 *Soil microbiomes associated with different soil types*

295 In 1975, a soil taxonomy was developed by the United States Department of
296 Agriculture (USDA) and the National Resources Conservation Service, which
297 separates soils into twelve orders that are based on their physical, chemical, or
298 biological properties(Soil Survey Staff 1999, 1999). Despite the availability of this
299 classification, it is rarely incorporated into soil microbiome surveys. Using RefSoil
300 and estimated abundances from similar EMP OTUs, we evaluated microbial
301 distribution in various soil orders. We obtained GPS data from EMP soil samples
302 originating from the United States that allowed us to obtain the soil classification.
303 Within these EMP samples, the most represented soil orders included Mollisols
304 (58%, grassland fertile soils) and Alfisols (37%, fertile soils typically under forest
305 vegetation) (Supplementary Table 4). The abundance of OTUs similar to RefSoil
306 genes was used to estimate the membership of well-characterized isolates in these
307 soil samples (Supplementary Fig. 2). Mollisols, Alfisols and Vertisols (soils with high
308 clay content with pronounced changes in moisture) were associated with the most
309 RefSoil representatives, while Gelisols (cold climate soils), Ultisols (soils with low
310 cation exchange), and sand/rock/ice contained the least (Supplementary Table 4).

311

312 *Identification of the most beneficial future targets for RefSoil*

313 EMP OTUs that do not share high similarity with 16S rRNA genes from
314 RefSoil represent current knowledge gaps for which we lack cultivated isolates.
315 These gaps were visually identified by overlaying the presence of highly similar
316 relatives in RefSoil (similarity >97%) to branches in a phylogenetic tree of soil-
317 associated EMP amplicons (Fig. 3). We evaluated the presence of OTUs in EMP
318 libraries based on their presence in EMP samples (frequency) and their cumulative
319 observed abundance. We identified the most observed and abundant EMP OTUs

320 that did not share high similarity to Refsoil to generate a “RefSoil’s most wanted list”
321 of targets that, if isolated or sequenced, could provide information on prevalent but
322 uncharacterized lineages (Table 1, Supplementary Table 5-6). OTUs sharing
323 similarity to Verrucomicrobia (8 OTUs) and Acidobacteria (6 OTUs) were among the
324 most abundant and frequently observed EMP OTUs that are not currently
325 represented in RefSoil (Table 1). These targets, observed in nature but not in our
326 database, would be ideal to isolate and characterize to better understand the soil
327 biodiversity.

328

329 *Single cell genomics in soil*

330 Sequencing-based approaches provide an alternative to accessing the
331 genomes of soil organisms without cultivation. Previous efforts have used assembly
332 of genomes from metagenomes(Labonté *et al.*, 2015; Martijn *et al.*, 2015; Field *et al.*,
333 2014) and single cell genomics(Gawad *et al.*, 2016; Lasken, 2012; Stepanauskas,
334 2012) to obtain genomic blueprints of yet uncultured microbial groups. To evaluate
335 the effectiveness of single cell genomics on soil communities, we performed a pilot-
336 scale experiment on a residential garden soil in Maine. The 16S rRNA gene was
337 successfully recovered from 109 of the 317 single amplified genomes (SAGs). This
338 34% 16S rRNA gene recovery rate is comparable to single cell genomics studies in
339 marine, freshwater and other environments(Martinez-Garcia *et al.*, 2011; Rinke *et*
340 *al.*, 2013; Swan *et al.*, 2011). The 16S rRNA genes of 14 of these SAGs, belonging to
341 Proteobacteria, Actinobacteria, Nitrospirae, Verrucomicrobia, Planktomyces,
342 Acidobacteria and Chloroflexi, were selected based on their lack of representation
343 within RefSoil. Genomic sequencing of those SAGs resulted in a cumulative assembly
344 of 23 Mbp (Table 2, Supplementary Table 7). We estimated the abundance of OTUs
345 similar to single cell 16S rRNA genes to range from 5E-7 to 2E-2% of EMP OTU
346 abundances. These abundances are very low but comparable to the average and
347 median relative abundance of OTUs within EMP (4E-3% and 1E-6%, respectively).
348 If these draft genomes were added to RefSoil, these 14 SAGs would increase
349 RefSoil’s representation of EMP amplicons by 7% by abundance.

350

351 **Discussion**

352

353 Advances in sequencing techniques for utilizing culture-independent
354 approaches have created new opportunities for understanding soil microbiology
355 and its impact on soil health, stability, and management. We provide RefSoil as a
356 community genomic resource to provide high-quality curation of organisms that
357 originate in soil studies, allowing us to more accurately annotate soil sequencing
358 datasets. This database is an important initial effort to provide improved soil
359 references, and we acknowledge that it is far from a complete representation of
360 soil’s biodiversity. In fact, compared to EMP soils, RefSoil represents only 2% of
361 observed EMP OTUs and 10% by abundance of the observed soil OTUs, highlighting
362 the magnitude of soil biodiversity and the limitations of our current database based
363 on cultivated representatives. These results are consistent with a recent effort
364 demonstrating that we are very limited in our knowledge of not only soil but life’s
365 diversity and advocating that much can be learned by including uncultivated

366 organisms(Hug *et al.*, 2016). We show here how RefSoil can be used to evaluate
367 which organisms to target that would most effectively increase our knowledge of
368 soil biodiversity. For example, if genome references were available for the top most
369 wanted organisms identified in this effort (Table 1), we could expand RefSoil's
370 representation of EMP soils by 58% by abundance.

371 Notably, many of these targets have previously been observed to be
372 recalcitrant to cultivation with standard laboratory media, resulting in their absence
373 in current genome databases. Acidobacteria, for example, is known to be slow
374 growing and difficult to cultivate (Nunes da Rocha *et al.*, 2009) while also observed
375 to be highly abundant in soil (33% of EMP amplicons by abundance). Another
376 fastidious bacteria, Verrucomicrobia(Bergmann *et al.*, 2011), were also observed to
377 be highly abundant (12.5%) in EMP but not well represented in RefSoil (two of 888
378 bacterial genomes). Despite their absence from cultivated isolates, both
379 Acidobacteria and Verrucomicrobia have been observed to be critical for nutrient
380 cycling in soils (Nunes da Rocha *et al.*, 2009; Ward *et al.*, 2009; Fierer *et al.*, 2013;
381 Martinez-Garcia *et al.*, 2012).

382 Novel isolation and culturing techniques will help us to access these
383 previously difficult to grow bacteria and will also be complemented by emerging
384 sequencing technologies. In particular, single-cell genomics hold great promise to
385 provide genomic characterization of lineages that are difficult to culture (Gawad *et al.*,
386 2016; Lasken, 2012; Stepanauskas, 2012). In our pilot experiment, we
387 demonstrate, for the first time, that single-cell genomics is applicable on soil
388 samples and is well suited to recover the genomic information from abundant but
389 yet uncultured taxonomic groups. The 14 sequenced SAGs have significantly
390 increased the extent to which RefSoil represents the predominant soil lineages from
391 a single sample. Much larger single cell genomics projects are feasible and have been
392 employed in prior studies of other environments(Kashtan *et al.*, 2014; Rinke *et al.*,
393 2013; Swan *et al.*, 2013). The continued, rapid improvements in this technology are
394 likely to lead to further scalability, offering a practical means to fill the existing gaps
395 in the RefSoil database and biodiversity more broadly.

396 The RefSoil database is also a tool that allows us to characterize currently
397 known soil bacteria phylogeny and functions. This database spans 24 phyla of
398 bacteria and archaea. While genes related to microbial growth and reproduction
399 (e.g., DNA, RNA, and protein metabolism) originate from diverse phyla, key
400 functions related to metabolism of aromatic compounds, iron metabolism, and
401 dormancy and sporulation were observed to be enriched from only a few RefSoil
402 phyla, suggesting that these phyla may have specialized functions within soil
403 communities. Specifically, Proteobacteria-associated genes were enriched in
404 functions related to motility, chemotaxis, and membrane transportation (Figure 2),
405 suggesting that Proteobacteria are likely to be efficient at acquiring readily available
406 nutrients and elements for growth(Fierer *et al.*, 2007). Genes related to
407 Proteobacteria and Actinobacteria were also found dominant among RefSoil
408 genomes in subsystems related to the metabolism of aromatic compounds. This
409 observation is consistent with the association of Proteobacteria and humic
410 substances utilization and the contribution of Actinobacteria to plant material
411 degradation(Godden *et al.*, 1992; Fuchs *et al.*, 2011).

412 By comparing RefSoil to other databases, we identified biases for specific
413 phyla in RefSoil; in particular, Firmicutes are observed frequently in RefSoil but
414 were not observed to be highly abundant in soil environments (5.7% of all EMP
415 amplicons). Firmicutes have been well-studied as pathogens (Buffie and Pamer,
416 2013; Kamada *et al.*, 2013; Rupnik *et al.*, 2009), likely biasing their representation in
417 our databases and consequently their annotations in soil studies. A key advantage
418 to the development of the RefSoil database is the opportunity to evaluate these
419 biases and to identify targets for future curation that would create a more
420 representative database resource for the soil. To this end, we evaluated the
421 representation of RefSoil in various soils.

422 Consistent with previous observations that microbial community
423 composition is correlated with soil environments (Fierer *et al.*, 2012; Fierer and
424 Jackson, 2006), we observed specific RefSoil membership associated with soil
425 taxonomy. Not all soil taxonomic orders were equally represented in RefSoil, with
426 the most genomes associated with Mollisols, or grassland soils. This bias is likely
427 due to increased research in these soils due to their importance for agriculture and
428 productivity. In contrast, for Gellisol permafrost soils, which cover over 20% of
429 Earth's terrestrial surface (Koven *et al.*, 2011) and are a significant carbon source
430 from our biosphere to the atmosphere, we observe only a small fraction of known
431 soil microbes are present. Previous studies suggest that high levels of novelty can
432 be observed in permafrost microbiomes (Koven *et al.*, 2011; Mackelprang *et al.*,
433 2011), suggesting that these soils could significantly benefit from improved
434 references.

435 Another advantage to the RefSoil database is that it represents genomes for
436 which strains should currently be available and for which we have high-quality
437 genomes. As a consequence, genomes within RefSoil could help to inform and
438 design soil microbiology experiments. For example, it is known that nitrogen cycling
439 genes are abundant in agricultural soils (Xue *et al.*, 2013; Philippot *et al.*, 2007; Long
440 *et al.*, 2012). A mock community of isolates known for participating in nitrogen
441 cycling could be generated and proportionally designed to mimic soil conditions
442 using RefSoil: these strains might include microorganisms related to *Streptomyces*
443 *venezuelae* ATCC 10712 (assimilatory nitrate reductase), *Bacillus anthracis* (nitric
444 oxide reductase), *Pseudomonas brassicacearum* (nitrous oxide reductase),
445 *Halomonas elongata* DSM 2581 (ammonia monooxygenase), *Pseudomonas*
446 *fluorescens* SBW25 (ammonia monooxygenase), and *Pleurocapsa sp.* PCC 7327
447 (nitrogen fixation). Combining RefSoil and available amplicon datasets, one could
448 estimate the proportions of strains as observed in soil samples.

449 In conclusion, RefSoil is an important first step towards building a more
450 comprehensive, well-curated database. Though it is far from complete, we have
451 been able to use RefSoil as a tool to identify key phyla that represent gaps in known
452 soil diversity and underrepresented phyla. Going forward, this reference will be
453 expanded as isolation and cultivation-independent technologies continue to
454 improve.

455
456
457

458 **Acknowledgments**

459 We are grateful for the support of the NSF Terragenome International Soil
460 Metagenome Sequencing Consortium for providing a collaborative workshop with
461 the soil community for discussions to improve this project. This material is based
462 upon work supported by the U.S. Department of Energy, Office of Science, Office of
463 Biological and Environmental Research, under Award Number SC0010775.

464

465 **Contributions** E.C., A.G., A.H., and J.T. curated Refsoil; J.C. and A.H. collected the
466 genome information to build the database; J.C., J.F., A.H., R.W., and F.Y., characterized
467 and compared RefSoil to EMP datasets; B.G. provided soil orders for EMP samples;
468 R.S. performed single cell genomics; J.T., K.H. and A. H. supervised the work; J.C., F.Y.,
469 and A.H. wrote the manuscript with contributions from all other authors.

470

471 **Conflict of Interest**

472 The authors declare no conflict of interests.

473

474

475 **References**

- 476 Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, *et al.* (2008). The RAST
477 Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**: 75.
- 478 Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA, *et al.*
479 (2011). The under-recognized dominance of Verrucomicrobia in soil bacterial
480 communities. *Soil Biology and Biochemistry* **43**: 1450–1455.
- 481 Buffie CG, Pamer EG. (2013). Microbiota-mediated colonization resistance against
482 intestinal pathogens. *Nat Rev Immunol* **13**: 790–801.
- 483 Burmølle M, Johnsen K, Al-Soud WA, Hansen LH, Sørensen SJ. (2009). The presence
484 of embedded bacterial pure cultures in agar plates stimulate the culturability of soil
485 bacteria. *Journal of Microbiological Methods* **79**: 166–173.
- 486 Delmont TO, Prestat E, Keegan KP, Faubladièr M, Robe P, Clark IM, *et al.* (2012).
487 Structure, fluctuation and magnitude of a natural grassland soil metagenome. *ISME J*
488 **6**: 1677–1687.
- 489 Field EK, Sczyrba A, Lyman AE, Harris CC, Woyke T, Stepanauskas R, *et al.* (2014).
490 Genomic insights into the uncultivated marine Zetaproteobacteria at Loihi
491 Seamount. *ISME J* **9**: 857–870.
- 492 Fierer N, Bradford MA, Jackson RB. (2007). Toward an ecological classification of
493 soil bacteria. *Ecology* **88**: 1354–1364.
- 494 Fierer N, Jackson RB. (2006). The diversity and biogeography of soil bacterial
495 communities. *P Natl Acad Sci USA* **103**: 626–631.
- 496 Fierer N, Ladau J, Clemente JC, Leff JW, Owens SM, Pollard KS, *et al.* (2013).
497 Reconstructing the Microbial Diversity and Function of Pre-Agricultural Tallgrass
498 Prairie Soils in the United States. *Science* **342**: 621–624.
- 499 Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, *et al.* (2012). Cross-
500 biome metagenomic analyses of soil microbial communities and their functional
501 attributes. *P Natl Acad Sci USA* **109**: 21390–21395.
- 502 Fuchs G, Boll M, Heider J. (2011). Microbial degradation of aromatic compounds —
503 from one strategy to four. *Nat Rev Micro* **9**: 803–816.
- 504 Gawad C, Koh W, Quake SR. (2016). Single-cell genome sequencing: current state of
505 the science. *Nat Rev Genet* **17**: 175–188.
- 506 Gilbert JA, Jansson JK, Knight R. (2014). The Earth Microbiome project: successes
507 and aspirations. *BMC Biol* **12**: 69.
- 508 Godden B, Ball AS, Helvenstein P, Mccarthy AJ, Penninckx MJ. (1992). Towards

- 509 elucidation of the lignin degradation pathway in actinomycetes. *Journal of General*
510 *Microbiology* **138**: 2441–2448.
- 511 Guo J, Cole JR, Zhang Q, Brown CT, Tiedje JM. (2015). Microbial Community Analysis
512 with Ribosomal Gene Fragments from Shotgun Metagenomes Schloss PD (ed).
513 *Applied and Environmental Microbiology* **82**: 157–166.
- 514 Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, *et al.* (2016). A
515 new view of the tree of life. *Nat Microbiol* 16048.
- 516 Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, *et al.*
517 (2012). Structure, function and diversity of the healthy human microbiome. *Nature*
518 **486**: 207–214.
- 519 Janssen PH, Yates PS, Grinton BE, Taylor PM, Sait M. (2002). Improved Culturability
520 of Soil Bacteria and Isolation in Pure Culture of Novel Members of the Divisions
521 Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia. *Applied and*
522 *Environmental Microbiology* **68**: 2391–2396.
- 523 Kamada N, Chen GY, Inohara N, Núñez G. (2013). Control of pathogens and
524 pathobionts by the gut microbiota. *Nat Immunol* **14**: 685–690.
- 525 Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, *et al.* (2014).
526 Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild
527 *Prochlorococcus*. *Science* **344**: 416–420.
- 528 Koven CD, Ringeval B, Friedlingstein P, Ciais P, Cadule P, Khvorostyanov D, *et al.*
529 (2011). Permafrost carbon-climate feedbacks accelerate global warming. *P Natl*
530 *Acad Sci USA* **108**: 14769–14774.
- 531 Labonté JM, Swan BK, Poulos B, Luo H, Koren S, Hallam SJ, *et al.* (2015). Single-cell
532 genomics-based analysis of virus–host interactions in marine surface
533 bacterioplankton. *ISME J* **9**: 2386–2399.
- 534 Lasken RS. (2012). Genomic sequencing of uncultured microorganisms from single
535 cells. *Nat Rev Micro* **10**: 631–640.
- 536 Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, *et al.* (2015). A
537 new antibiotic kills pathogens without detectable resistance. *Nature* **517**: 455–459.
- 538 Long A, Heitman J, Tobias C, Philips R, Song B. (2012). Co-Occurring Anammox,
539 Denitrification, and Codenitrification in Agricultural Soils. *Applied and*
540 *Environmental Microbiology* **79**: 168–176.
- 541 Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ,
542 *et al.* (2011). Metagenomic analysis of a permafrost microbial community reveals a
543 rapid response to thaw. *Nature* **480**: 368–371.

- 544 Martijn J, Schulz F, Zaremba-Niedzwiedzka K, Viklund J, Stepanauskas R, Andersson
545 SGE, *et al.* (2015). Single-cell genomics of a rare environmental
546 alphaproteobacterium provides unique insights into Rickettsiaceae evolution. *ISME J*
547 **9**: 2373–2385.
- 548 Martinez-Garcia M, Brazel DM, Swan BK, Arnosti C, Chain PSG, Reitenga KG, *et al.*
549 (2012). Capturing Single Cell Genomes of Active Polysaccharide Degradors: An
550 Unexpected Contribution of Verrucomicrobia Ravel J (ed). *PLoS ONE* **7**: e35314.
- 551 Martinez-Garcia M, Swan BK, Poulton NJ, Gomez ML, Masland D, Sieracki ME, *et al.*
552 (2011). High-throughput single-cell sequencing identifies photoheterotrophs and
553 chemoautotrophs in freshwater bacterioplankton. *ISME J* **6**: 113–123.
- 554 Nawrocki EP, Eddy SR. (2013). Infernal 1.1: 100-fold faster RNA homology searches.
555 *Bioinformatics* **29**: 2933–2935.
- 556 Nunes da Rocha U, Van Overbeek L, Van Elsas JD. (2009). Exploration of hitherto-
557 uncultured bacteria from the rhizosphere. *FEMS Microbiology Ecology* **69**: 313–328.
- 558 Philippot L, Hallin S, Schloter M. (2007). Ecology of Denitrifying Prokaryotes in
559 Agricultural Soil. In: *Advances in Agronomy* Vol. 96. Elsevier, pp 249–305.
- 560 Price MN, Dehal PS, Arkin AP. (2010). FastTree 2 – Approximately Maximum-
561 Likelihood Trees for Large Alignments Poon AFY (ed). *PLoS ONE* **5**: e9490.
- 562 Reich P. Global soil suborders.
563 http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/use/?cid=nrcs142p2_054
564 013.
- 565 Rideout JR, He Y, Navas-Molina JA, Walters WA, Ursell LK, Gibbons SM, *et al.* (2014).
566 Subsampled open-reference clustering creates consistent, comprehensive OTU
567 definitions and scales to billions of sequences. *PeerJ* **2**: e545–25.
- 568 Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, *et al.* (2013).
569 Insights into the phylogeny and coding potential of microbial dark matter. *Nature*
570 **499**: 431–437.
- 571 Rupnik M, Wilcox MH, Gerding DN. (2009). Clostridium difficile infection: new
572 developments in epidemiology and pathogenesis. *Nat Rev Micro* **7**: 526–536.
- 573 Schoenborn L, Yates PS, Grinton BE, Hugenholtz P, Janssen PH. (2004). Liquid Serial
574 Dilution Is Inferior to Solid Media for Isolation of Cultures Representative of the
575 Phylum-Level Diversity of Soil Bacteria. *Applied and Environmental Microbiology* **70**:
576 4363–4366.
- 577 Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. (2012).
578 Metagenomic microbial community profiling using unique clade-specific marker

- 579 genes. *Nat Meth* **9**: 811–814.
- 580 Soil Survey Staff 1999. (1999). Soil Taxonomy: A basic system of soil classification
581 for making and interpreting soil surveys. 2nd ed. U.S. Department of Agriculture
582 Handbook 436: Natural Resources Conservation Service.
- 583 Stepanauskas R. (2012). Single cell genomics: an individual look at microbes.
584 *Current Opinion in Microbiology* **15**: 613–620.
- 585 Stepanauskas R, Fergusson E, Brown J, Poulton N, Tupper B, Labonté JM, *et al.*
586 Improved whole genome amplification and the integrated study of genomic and
587 optical properties of individual cells and viruses. *Manuscript in preparation*.
- 588 Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, *et al.* (2011).
589 Potential for Chemolithoautotrophy Among Ubiquitous Bacteria Lineages in the
590 Dark Ocean. *Science* **333**: 1296–1300.
- 591 Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, Gonzalez JM, *et al.*
592 (2013). Prevalent genome streamlining and latitudinal divergence of planktonic
593 bacteria in the surface ocean. *P Natl Acad Sci USA* **110**: 11463–11468.
- 594 Tatusova T, Ciufu S, Fedorov B, O'Neill K, Tolstoy I. (2013). RefSeq microbial
595 genomes database: new representation and annotation strategy. *Nucleic Acids Res*
596 **42**: D553–D559.
- 597 Van H T Pham, Kim J. (2012). Cultivation of unculturable soil bacteria. *Trends in*
598 *Biotechnology* **30**: 475–484.
- 599 Vartoukian SR, Palmer RM, Wade WG. (2010). Strategies for culture of ‘unculturable’
600 bacteria. *FEMS Microbiology Letters* **309**: 1–7.
- 601 Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian Classifier for Rapid
602 Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and*
603 *Environmental Microbiology* **73**: 5261–5267.
- 604 Ward NL, Challacombe JF, Janssen PH, Henrissat B, Coutinho PM, Wu M, *et al.*
605 (2009). Three Genomes from the Phylum Acidobacteria Provide Insight into the
606 Lifestyles of These Microorganisms in Soils. *Applied and Environmental Microbiology*
607 **75**: 2046–2056.
- 608 Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, *et al.* (2009). A
609 phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**:
610 1056–1060.
- 611 Xue K, Wu L, Deng Y, He Z, Van Nostrand J, Robertson PG, *et al.* (2013). Functional
612 Gene Differences in Soil Microbial Communities from Conventional, Low-Input, and
613 Organic Farmlands. *Applied and Environmental Microbiology* **79**: 1284–1292.

614 **Figure Legends**

615 **Figure 1. Phylogenetic tree of RefSoil**

616 Phylogenetic tree of aligned 16S rRNA genes originating from RefSoil bacterial
617 genomes. A: Acidobacteria, B: Actinobacteria, C: Aquificae, D: Armatimonadetes, E:
618 Bacteroidetes, F: Chlamydiae, G: Chlorobi, H: Chloroflexi, J: Cyanobacteria, K:
619 Deferribacteres, L: Deinococcus-Thermus, N: Firmicutes, O: Fusobacteria, P:
620 Gemmatimonadetes, Q: Nitrospirae, R: Planctomycetes, S: Proteobacteria, T:
621 Spirochaetes, U: Synergistetes, V: Tenericutes, W: Thermotogae, X: Verrucomicrobia.

622

623 **Figure 2. Functional analysis**

624 The distribution of phylogenetic origins associated with RefSoil functional
625 subsystems as annotated by RAST (panel A); the overall phylogenetic distribution of
626 all genes in RefSoil (panel B).

627

628 **Figure 3. Phylogenetic tree of EMP OTUs clustered by taxonomy.**

629 Ring I (green) represents the cumulative log-scaled abundance of OTUs in EMP soil
630 samples. Ring II (red) represents EMP OTUs that share greater than 97% gene
631 similarity (to RefSoil 16S rRNA genes; ring III (blue) indicates that these 16S rRNA
632 genes shared similarity to sorted cells that were selected for single cell genomics. A:
633 Acidobacteria, B: Actinobacteria, C: Aquificae, D: Armatimonadetes, E: Bacteroidetes,
634 F: Chlamydiae, G: Chlorobi, H: Chloroflexi, I: Crenarchaeota, K: Deferribacteres, L:
635 Deinococcus-Thermus, M: Euryarchaeota, N: Firmicutes, O: Fusobacteria, P:
636 Gemmatimonadetes, Q: Nitrospirae, R: Planctomycetes, S: Proteobacteria, T:
637 Spirochaetes, U: Synergistetes, V: Tenericutes, W: Thermotogae, X: Verrucomicrobia,
638 Y: Cyanobacteria/Chloroplast.

639

640 **Table Legends**

641 **Table 1. RefSoil's most wanted OTUs**

642 RefSoil's most wanted OTUs based on observed frequency and abundance in EMP
643 soil samples. Taxonomy for OTUs are assigned by RDP classifier(Wang *et al.*, 2007).
644 *:(Rideout *et al.*, 2014).

645

646 **Table 2. Single-cell amplified genomes**

647 Taxonomic classification of single-cell amplified genomes and the abundance of the
648 most similar EMP OTU. *: (Rideout *et al.*, 2014).

649

650 **Supplementary Description**

651 **Supplementary Figure 1. The abundance distribution of phyla**

652 The abundance distribution of phyla ($\log(\text{abundance}) + 1$) present in the RefSoil
653 database compared to NCBI's RefSeq (A) and phyla that are enriched in RefSoil
654 compared to RefSeq (B).

655

656 **Supplementary Figure 2. Average relative abundance for various soil orders**

657 Average relative abundance of EMP OTUs (sharing similarity with RefSoil genes) for
658 various soil orders as classified by NRCS Soil Taxonomy.

659

660

661 **Supplementary Table 1. The RefSoil Database**

662

663 **Supplementary Table 2. RefSoil genes in RAST subsystem functions**

664 Total number of RefSoil genes observed in RAST subsystem functions.

665

666 **Supplementary Table 3. Subsystem level functions in RefSoil**

667 The number of unique phyla and the number of enriched phyla associated with
668 encoded subsystem level functions in RefSoil. (Bold italicized: enriched with genes
669 associated with three or less phyla)

670

671 **Supplementary Table 4. Abundance of EMP OTUs in soil order**

672 The total abundance of EMP OTUs by soil order in soil samples originating in the
673 United States.

674

675 **Supplementary Table 5. 100 most abundant OTUs**

676 100 most abundant OTUs in EMP soil samples

677

678 **Supplementary Table 6. 100 most frequent OTUs**

679 100 most frequent OTUs in EMP soil samples

680

681 **Supplementary Table 7. Single cell genomic properties**

682

683 **Table 1. RefSoil's most wanted OTUs**
 684 RefSoil's most wanted OTUs based on observed frequency and abundance in EMP
 685 soil samples. Taxonomy for OTUs are assigned by RDP classifier(Wang *et al.*, 2007).
 686

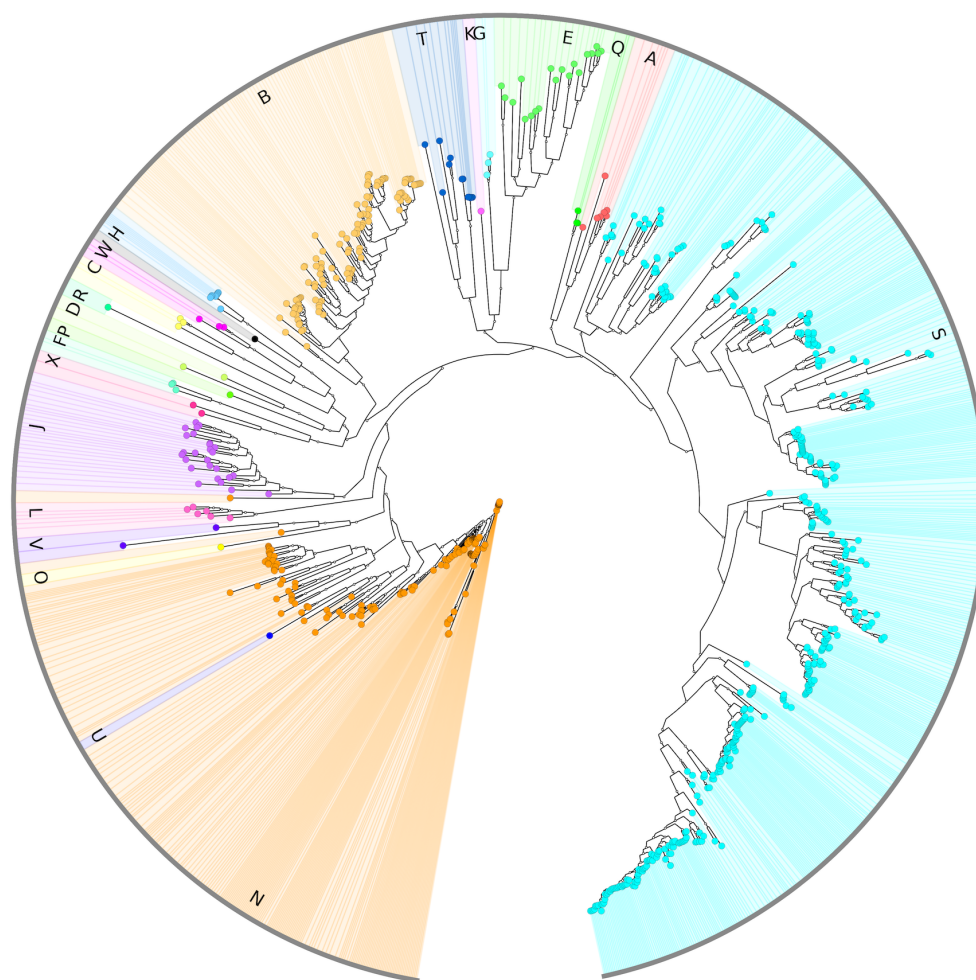
| OTU ID in (Rideout et al., 2014) | Closest Match in RDP Classifier | | RDP Classifier Similarity Score | Abundance (total amplicons) | Number of Samples |
|--|---------------------------------|---------------------|------------------------------------|-----------------------------------|-------------------------|
| | Phylum | Class | | | |
| 4457032 | Verrucomicrobia | Spartobacteria | 1 | 8,007,453 | 2,312 |
| 4471583 | Verrucomicrobia | Spartobacteria | 0.92 | 2,937,242 | 1,935 |
| 101868 | Verrucomicrobia | Spartobacteria | 1 | 1,828,706 | 2,151 |
| 559213 | Firmicutes | Bacilli | 1 | 1,606,757 | 2,410 |
| 1105039 | Verrucomicrobia | Spartobacteria | 1 | 1,295,847 | 2,102 |
| 807954 | Bacteroidetes | Sphingobacteriia | 0.98 | 875,988 | 2,546 |
| 4342972 | Verrucomicrobia | Spartobacteria | 0.97 | 750,557 | 2,386 |
| 4423681 | Gemmatimonadetes | Gemmatimonadetes | 0.25 | 689,209 | 1,954 |
| 1109646 | Verrucomicrobia | Spartobacteria | 1 | 554,748 | 2,012 |
| 610188 | Acidobacteria | Acidobacteria_Gp6 | 0.99 | 553,476 | 2,694 |
| 4373456 | Acidobacteria | Acidobacteria_Gp4 | 1 | 397,255 | 2,150 |
| 4341176 | Verrucomicrobia | Subdivision3 | 0.99 | 383,900 | 2,098 |
| 720217 | Proteobacteria | Deltaproteobacteria | 0.74 | 345,621 | 2,073 |
| 4314933 | Acidobacteria | Acidobacteria_Gp1 | 0.8 | 333,428 | 1,974 |
| 205391 | Acidobacteria | Acidobacteria_Gp3 | 0.87 | 327,162 | 2,190 |
| 4378940 | Acidobacteria | Acidobacteria_Gp6 | 1 | 310,421 | 2,499 |
| 946250 | Verrucomicrobia | Spartobacteria | 1 | 300,544 | 2,386 |
| 3122801 | Acidobacteria | Acidobacteria_Gp1 | 0.97 | 273,493 | 2,107 |
| 4450676 | Proteobacteria | Alphaproteobacteria | 1 | 255,424 | 1,963 |
| 206514 | Proteobacteria | Alphaproteobacteria | 0.92 | 207,460 | 2,075 |
| 4463040 | Firmicutes | Clostridia | 0.94 | 204,483 | 2,096 |

687
 688
 689

690 **Table 2. Single-cell amplified genomes**
691 Taxonomic classification of single-cell amplified genomes and the abundance of the
692 most similar EMP OTU.
693

| NCBI ID | Phylum | Class | EMP ID in (Rideout et al., 2014) | Abundance | Relative Abundance (%) |
|--------------|-----------------|--------------------------|------------------------------------|-----------|------------------------|
| LSTF00000000 | Proteobacteria | β -proteobacteria | New.52.CleanUp.ReferenceOTU25474 | 81,630 | 2.13E-02 |
| LSTI00000000 | Actinobacteria | Thermoleophila | 551344 | 35,500 | 9.27E-03 |
| LSTA00000000 | Verrucomicrobia | Spartobacteria | New.54.CleanUp.ReferenceOTU60738 | 5,076 | 1.33E-03 |
| LSTC00000000 | Nitrospirae | Nitrospira | New.22.CleanUp.ReferenceOTU3952 | 4,337 | 1.13E-03 |
| LSTH00000000 | Planctomycetes | Planctomycetia | 521158 | 4,161 | 1.09E-03 |
| LSSZ00000000 | Planctomycetes | Planctomycetia | New.5.CleanUp.ReferenceOTU188938 | 2,434 | 6.36E-04 |
| LSTE00000000 | Proteobacteria | γ -proteobacteria | NA | NA | NA |
| LSTB00000000 | Proteobacteria | α -proteobacteria | New.52.CleanUp.ReferenceOTU4588 | 309 | 8.07E-05 |
| LSSY00000000 | Actinobacteria | Thermoleophila | 904196 | 103 | 2.69E-05 |
| LSTG00000000 | Verrucomicrobia | Opiritae | New.5.CleanUp.ReferenceOTU291124 | 69 | 1.80E-05 |
| LSTD00000000 | Planctomycetes | Planctomycetia | New.59.CleanUp.ReferenceOTU1223798 | 44 | 1.15E-05 |
| LSTJ00000000 | Verrucomicrobia | Verrucomicrobiae | New.7.CleanUp.ReferenceOTU84651 | 11 | 2.87E-06 |
| LSTK00000000 | Acidobacteria | Acidobacteriia | New.59.CleanUp.ReferenceOTU999655 | 4 | 1.04E-06 |
| LSSX00000000 | Chloroflexi | TK17 | New.59.CleanUp.ReferenceOTU481621 | 2 | 5.22E-07 |

694
695
696



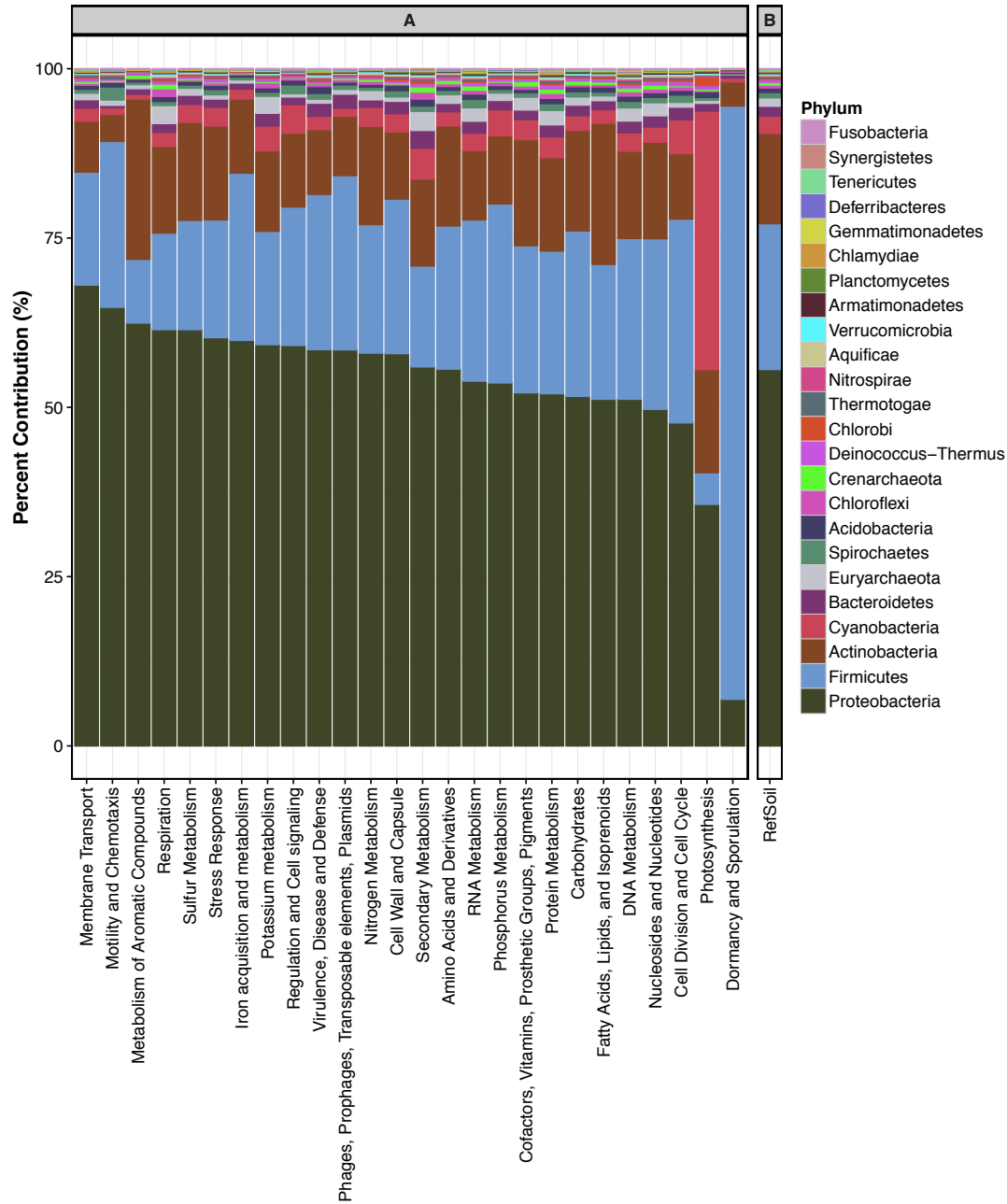
697

698

699 **Figure 1. Phylogenetic tree of RefSoil**

700 Phylogenetic tree of aligned 16S rRNA genes originating from RefSoil bacterial
701 genomes. A: Acidobacteria, B: Actinobacteria, C: Aquificae, D: Armatimonadetes, E:
702 Bacteroidetes, F: Chlamydiae, G: Chlorobi, H: Chloroflexi, J: Cyanobacteria, K:
703 Deferribacteres, L: Deinococcus-Thermus, N: Firmicutes, O: Fusobacteria, P:
704 Gemmatimonadetes, Q: Nitrospirae, R: Planctomycetes, S: Proteobacteria, T:
705 Spirochaetes, U: Synergistetes, V: Tenericutes, W: Thermotogae, X: Verrucomicrobia.

706



707

708 **Figure 2. Functional analysis**

709 The distribution of phylogenetic origins associated with RefSoil functional
 710 subsystems as annotated by RAST (panel A); the overall phylogenetic distribution of
 711 all genes in RefSoil (panel B).

712

713

714

715

716

717



718

719 **Figure 3. Phylogenetic tree of EMP OTUs clustered by taxonomy.**

720 Ring I (green) represents the cumulative log-scaled abundance of OTUs in EMP soil
721 samples. Ring II (red) represents EMP OTUs that share greater than 97% gene
722 similarity (to RefSoil 16S rRNA genes; ring III (blue) indicates that these 16S rRNA
723 genes shared similarity to sorted cells that were selected for single cell genomics. A:
724 Acidobacteria, B: Actinobacteria, C: Aquificae, D: Armatimonadetes, E: Bacteroidetes,
725 F: Chlamydiae, G: Chlorobi, H: Chloroflexi, I: Crenarchaeota, K: Deferribacteres, L:
726 Deinococcus-Thermus, M: Euryarchaeota, N: Firmicutes, O: Fusobacteria, P:
727 Gemmatimonadetes, Q: Nitrospirae, R: Planctomycetes, S: Proteobacteria, T:
728 Spirochaetes, U: Synergistetes, V: Tenericutes, W: Thermotogae, X: Verrucomicrobia,
729 Y: Cyanobacteria/Chloroplast.

730

731

732

733

734