



**HAL**  
open science

## Functional trait relationships demonstrate life strategies in terrestrial prokaryotes

Damien Finn, Benoît Bergk-Pinto, Christina Hazard, Graeme W. Nicol,  
Christoph Tebbe, Timothy Vogel

### ► To cite this version:

Damien Finn, Benoît Bergk-Pinto, Christina Hazard, Graeme W. Nicol, Christoph Tebbe, et al.. Functional trait relationships demonstrate life strategies in terrestrial prokaryotes. *FEMS Microbiology Ecology*, 2021, 97 (5), 10.1093/femsec/fiab068 . hal-03411983

**HAL Id: hal-03411983**

**<https://hal.science/hal-03411983v1>**

Submitted on 3 Nov 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1  
2  
3 1 Functional trait relationships demonstrate life strategies in terrestrial prokaryotes.  
4  
5 2

6 3 Damien R. Finn<sup>1,2,3</sup>, Benoît Bergk-Pinto<sup>2</sup>, Christina Hazard<sup>2</sup>, Graeme W. Nicol<sup>2</sup>,  
7 4 Christoph C. Tebbe<sup>3</sup>, Timothy M. Vogel<sup>2</sup>  
8  
9 5

10 6 <sup>1</sup>School of Agriculture and Food Sciences, University of Queensland, Brisbane,  
11 7 Australia 4072;

12 8 <sup>2</sup>Environmental Microbial Genomics, Laboratoire Ampère, École Centrale de Lyon,  
13 9 Université de Lyon, Écully, France 69134;

14 10 <sup>3</sup>Thünen Institut für Biodiversität, Johann Heinrich von Thünen Institut,  
15 11 Braunschweig, Germany 38116.  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

14 Author correspondence:

15 Damien Finn; current address: Thünen Institut für Biodiversität, Braunschweig,  
16 Germany 38116; email address: damien.finn@thuenen.de

18 Keywords:

19 Theoretical ecology, niche differentiation, copiotroph-oligotroph, Random Forest  
20 modelling  
21  
22

23 Abstract

24 Functional, physiological traits are the underlying drivers of niche differentiation. A  
25 common framework related to niches occupied by terrestrial prokaryotes is based on  
26 copiotrophy or oligotrophy, where resource investment is primarily in either rapid  
27 growth or stress tolerance, respectively. A quantitative trait-based approach sought  
28 relationships between taxa, traits and niche in terrestrial prokaryotes. With 175 taxa  
29 from 11 Phyla and 35 Families ( $n = 5$  per Family), traits were considered as discrete  
30 counts of shared genome-encoded proteins. Trait composition strongly supported non-  
31 random functional distributions as preferential clustering of related taxa via unweighted  
32 pair-group method with arithmetic mean. Trait similarity between taxa increased as  
33 taxonomic rank decreased. A suite of Random Forest models identified traits  
34 significantly enriched or depleted in taxonomic groups. These traits conveyed functions

1  
2  
3 35 related to rapid growth, nutrient acquisition and stress tolerance consistent with their  
4 36 presence in copiotroph-oligotroph niches. Hierarchical clustering of traits identified a  
5 37 clade of competitive, copiotrophic Families resilient to oxidative stress *versus*  
6 38 glycosyltransferase-enriched oligotrophic Families resistant to antimicrobials and  
7 39 environmental stress. However, the formation of five clades suggested a more  
8 40 nuanced view to describe niche differentiation in terrestrial systems is necessary. We  
9 41 suggest considering traits involved in both resource investment and acquisition when  
10 42 predicting niche.  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21

43  
44

## 45 1. Introduction

46 Niche differentiation, the process of physiologically distinct organisms adapting better  
47 48 to certain conditions, is a contributing factor to the high biodiversity inherent in microbial  
49 50 communities (Prosser, 2012). Such differentiation is likely an inevitable consequence  
51 52 of the principles of competitive exclusion and natural selection working in tandem – no  
53 54 two organisms can theoretically occupy the same niche, as the poorer competitor must  
55 56 either adapt to a unique niche or be driven to extinction in that system (Gause, 1932,  
57 58 Hutchinson, 1957, Leibold, 1995). The physiological traits driving niche differentiation  
59 60 must have the capacity to convey an advantage to the organism's ability to survive and  
61 62 reproduce (*i.e.* fitness) and be inherited by successive generations (McGill *et al.*, 2006).  
63 64 Importantly, this implies that microbial communities are not only diverse in terms of  
65 66 individual 16S rRNA gene sequences, commonly used to assess community diversity,  
67 68 but also diverse in regard to their physiological traits.

69 Explaining niche differentiation through the functional, physiological traits present in  
70 71 ecological community members has a long history in macroecology. For example,  
72 73 differences between beak size and shape in Galápagos finches was instrumental in  
74 75 Darwin's hypothesis that a common ancestor had differentiated into multiple, island-  
76 77 specific species. Within the past century, trait-based analyses have been particularly  
78 79 predominant in plant ecology, with seed germination in submerged soil, salt tolerance,  
80 81 carbon to nitrogen biomass stoichiometry, and leaf mass per unit area acting as  
82 83 examples of traits linked to niche differentiation (Gleason, 1926, Grime, 1979, Keddy,  
84 85 1992). In contrast, trait-based approaches to explain microbial ecology have only been  
86 87 performed in few instances, such as conceptualizing niches of methanotrophs based  
88 89 on abundance in high *versus* low methane environments or disturbed *versus*  
90 91

1  
2  
3 69 undisturbed soils (Ho *et al.*, 2013), correlating increasing growth rate with increasing  
4  
5 70 ribosomal gene and ribosome-associated gene copy number (Vieira-Silva & Rocha,  
6  
7 71 2010), deterministic modelling of nitrification rate based on ammonia and oxygen  
8  
9 72 uptake rate, temperature sensitivity and growth rate (Bouskill *et al.*, 2012), defining  
10  
11 73 distinct niches of 32 marine microorganisms based on clustering of genome-encoded  
12  
13 74 functional proteins (Lauro *et al.*, 2009), identifying habitat generalists and specialists  
14  
15 75 based on taxon co-occurrence patterns (Barbéran *et al.*, 2012) and recently  
16  
17 76 comparisons of 23 'core' traits (e.g. motility, carbon metabolism, optimal pH for growth)  
18  
19 77 across 15 000 diverse host-associated and environmental genomes (Madin *et al.*,  
20  
21 78 2020).

22 79 A consistent trend noted in macroecology is that traits linked to how carbon and  
23  
24 80 energy is processed and allocated to biomass can describe separate niches (Brown *et*  
25  
26 81 *al.*, 2004). The canonical example are *r* and *K* strategists, where carbon and energy  
27  
28 82 are primarily invested in reproduction, or alternatively invested in tolerating biotic  
29  
30 83 and/or abiotic stressors, respectively (Grime, 1977). These dichotomous strategies  
31  
32 84 have been observed in microbial ecology: copiotrophs are considered as  
33  
34 85 microorganisms with relatively high growth rates that have relatively poor growth  
35  
36 86 efficiency (as carbon incorporated to biomass per unit resource), relatively high cell  
37  
38 87 maintenance energy costs, dependence on relatively high concentrations of organic  
39  
40 88 carbon in their environment, demonstrate rapid population blooms upon the addition of  
41  
42 89 organic matter and are not overly tolerant of abiotic stress (Semenov, 1991, Koch,  
43  
44 90 2001, Roller & Schmidt, 2015, Ho *et al.*, 2017). Oligotrophs are considered as the  
45  
46 91 inverse – low growth rate, high growth efficiency, low cell maintenance energy  
47  
48 92 requirements, high substrate uptake affinity, slow growth yet at a consistent rate and  
49  
50 93 are resilient to abiotic stress. Although the niche concept in macroecology has a  
51  
52 94 formalized definition founded on where a taxon can maintain a stable population within  
53  
54 95 multi-dimensional environmental space (Leibold, 1995), in this study, niche is used  
55  
56 96 simply to distinguish between prokaryotes being relatively more copiotrophic *versus*  
57  
58 97 oligotrophic.

59 98 These distinct niches became associated with specific terrestrial taxa at high  
60  
61 99 taxonomic rank based on recent molecular analyses. In complex microbial  
62  
63 100 communities, the relative abundance of Gammaproteobacteria, Bacteroidetes and  
64  
65 101 Actinobacteria were correlated with rapid growth in response to the addition of labile  
66  
67 102 organic matter or nitrogen (copiotrophs) (Fierer *et al.*, 2007, Goldfarb *et al.*, 2011,



1  
2  
3 103 Fierer *et al.*, 2012, Leff *et al.*, 2015). Conversely, the Deltaproteobacteria,  
4 104 Acidobacteria, Verrucomicrobia and Planctomycetes were negatively correlated with  
5 105 the addition of organic matter or nitrogen (oligotrophs) (Fierer *et al.*, 2007, Fierer *et al.*,  
6 106 2012, Leff *et al.*, 2015, Bastida *et al.*, 2016). Conflicting reports exist of Beta- and  
7 107 Alphaproteobacteria, with some studies describing them as copiotrophic and others as  
8 108 oligotrophic highlighting that a consistent niche may not necessarily exist across  
9 109 species within a large taxonomic group (Ho *et al.*, 2017). A genomic basis for traits  
10 110 associated with soils dominated by putative copiotrophs and oligotrophs has been  
11 111 expertly reviewed elsewhere, and interested readers are referred to Trivedi *et al.*,  
12 112 (2013) and references therein. Importantly, these observations suggest that specific  
13 113 traits that allow terrestrial prokaryotes to occupy these two niches should (generally)  
14 114 be associated with taxonomy. This is an example of ecological coherence at high  
15 115 taxonomic rank, whereby members within a taxon tend to have similar life strategies,  
16 116 niches and possess common traits compared to members of other taxa (Philippot *et al.*,  
17 117 2010). While ecological coherence of taxa has been considered previously, the  
18 118 shared, specific traits that drive niche differentiation in terrestrial prokaryotes remains  
19 119 an open question.

20  
21  
22 120 To identify the traits that differ between terrestrial prokaryote taxonomic groups, and  
23 121 whether these traits could describe the niches they occupy, a functional trait-based  
24 122 approach was adopted here. We posited that a trait must: a) be associated with a  
25 123 physiological process that conveys a fitness advantage under certain environmental  
26 124 conditions; b) be measurable in well-defined units; and c) vary more between  
27 125 taxonomic groups than within a taxonomic group (McGill *et al.*, 2006, Kearney *et al.*,  
28 126 2010). Traits were measured as discrete counts of chromosome-encoded proteins  
29 127 shared between at least two of 175 terrestrial prokaryotes. Markov Chain clustering  
30 128 (MCL) was used to group proteins as traits based on amino acid sequence similarity  
31 129 (%) akin to a previous approach that confirmed close taxonomic relatives tend to share  
32 130 functional traits in 1 374 genomes (Zhu *et al.*, 2015). This was necessary to compare  
33 131 highly similar (but non-identical) proteins from separate genomes that carry out the  
34 132 same biological function. To better identify important, distinguishing traits of terrestrial  
35 133 prokaryotes, this study differed from Zhu *et al.*, by: a) comparing 175 publicly available  
36 134 terrestrial prokaryote genomes from 35 Families ( $n = 5$  each), from 11 Phyla and two  
37 135 Kingdoms; b) selecting prokaryotes involved in terrestrial ecosystem processes of  
38 136 interest, including organic matter decomposition, nitrogen fixation, nitrification,

1  
2  
3 137 denitrification, methane oxidation, plant-growth promotion, bioremediation of  
4  
5 138 pollutants, pathogenesis and methanogenesis; c) selecting taxa isolated from a wide  
6  
7 139 range of terrestrial environments, such as nutrient rich decomposing plant material and  
8  
9 140 rhizosphere, submerged wetland and rice paddy soils, polluted soils, and nutrient poor  
10  
11 141 hot and cold arid environments; and d) avoiding the inclusion of multiple subspecies  
12  
13 142 and/or strains of a single species to prevent biases in analyses where highly over-  
14  
15 143 represented species are compared with species that have fewer cultured  
16  
17 144 representatives. The taxonomic system used here is from the NCBI, which is built upon  
18  
19 145 a historical array of culture-dependent, physiological observations and genetic  
20  
21 146 similarity to cultured isolates as average nucleotide identity, DNA-DNA hybridisation or  
22  
23 147 16S rRNA gene homology ([ncbi.nlm.nih.gov/Taxonomy/Browser](http://ncbi.nlm.nih.gov/Taxonomy/Browser)). The use of this  
24  
25 148 classification system and comparison to others is discussed further below.

26  
27 149 We hypothesised that: 1) traits are non-randomly distributed, with relatively closely  
28  
29 150 related taxa demonstrating greater similarity than unrelated taxa (ecological  
30  
31 151 coherence); 2) traits that are differentially enriched between taxonomic groups would  
32  
33 152 primarily be involved in metabolism, nutrient acquisition and/or tolerating  
34  
35 153 environmental stress; and 3) copiotrophic and oligotrophic taxonomic groups would  
36  
37 154 emerge based on collective trait enrichment.

38 155

39 156

## 40 157 2. Methodology

### 41 158 2.1 Collection of terrestrial prokaryote genomes

42 159 A collection of 175 sequenced and annotated genomes was collated (Supplementary  
43  
44 160 Table 1). Listed are the genome ID, phylogenetic lineage, role in an ecosystem process  
45  
46 161 if known, and isolation or genome sequencing reference. These genomes were  
47  
48 162 sourced from the National Centre for Biotechnology Information (NCBI) and Joint  
49  
50 163 Genome Institute (JGI) databases. Genomes were chosen based on several criteria:  
51  
52 164 a) five isolates per Family were chosen to have an equal minimum sample size per  
53  
54 165 group, with this sample size being constrained by sequenced genomes of under-  
55  
56 166 represented groups in public databases; b) only a single subspecies/strain per species  
57  
58 167 was included to avoid bias due to over-representation of some species in public  
59  
60 168 databases; c) an emphasis was placed to include isolates from diverse taxonomic  
61  
62 169 lineages involved in terrestrial ecosystem processes of interest, such as ammonia  
63  
64 170 oxidation and methanogenesis; and d) there was an emphasis to include taxonomic

1  
2  
3 171 groups frequently stated to be either copiotrophic (e.g. Actinobacteria,  
4 172 Gammaproteobacteria) or oligotrophic (e.g. Acidobacteria, Planctomycetes) based on  
5 173 observations from soil nutrient addition studies (Ho *et al.*, 2017). Taxonomic  
6 174 annotations for Phyla, Class, Order etc. were based on NCBI taxonomy as most  
7 175 genomes were sourced there. The authors recognise that taxonomy is constantly  
8 176 shifting, particularly so with the recent development of the Genome Taxonomy  
9 177 Database (GTDB) (Parks *et al.*, 2018). Of note is that the vast majority of taxa here  
10 178 have the same taxonomy in NCBI as in GTDB, with the exceptions that GTDB  
11 179 considers the Sporomusaceae as split into three separate Families, the  
12 180 Leuconostocaceae to be Lactobacillaceae, the Promicromonosporaceae to be  
13 181 Cellulomonadaceae, and the Bradyrhizobiaceae and Methylococcaceae have been  
14 182 renamed as Xanthobacteraceae and Methylomonadaceae, respectively. Taxon  
15 183 selection was constrained by availability of genomes for under-represented groups,  
16 184 such as the Chloroflexi, Verrucomicrobia, Planctomycetes, Thaumarchaeota and  
17 185 Euryarchaeota. To meet the  $n = 5$  requirement for balanced statistical analyses, it was  
18 186 necessary to consider these under-represented groups as 'Families'. Furthermore, due  
19 187 to the great diversity inherent within Proteobacterial Classes, Gamma-, Alpha-, Beta-  
20 188 and Deltaproteobacteria were considered as independent 'Phyla' for statistical  
21 189 analyses here. Indeed, GTDB now defines Deltaproteobacteria as its own Phylum,  
22 190 while Betaproteobacteria are considered as the Burkholderiales Order within the  
23 191 Gammaproteobacteria. The total of 175 genomes analysed here falls within the upper  
24 192 range of previous hypothesis-driven trait-based studies which varies from 11 isolates  
25 193 (Bouskill *et al.*, 2012) to 214 genomes (Vieira-Silva & Rocha, 2010).

194

195

## 196 2.2 Functional trait clustering

197 A step-by-step walkthrough of reproducible code to perform the following analyses on  
198 a subset of 12 genomes is available at: [https://github.com/DamienFinn/Trait-](https://github.com/DamienFinn/Trait-based_analyses)  
199 [based\\_analyses](https://github.com/DamienFinn/Trait-based_analyses). Firstly, a pairwise similarity comparison of all amino acid sequences  
200 (964 951 sequences) across the 175 genomes was performed with the all *versus* all  
201 basic local alignment tool function for proteins, BLAST-P (Altschul *et al.*, 1990). Amino  
202 acid sequences were subsequently clustered as traits via MCL weighted by pairwise  
203 amino acid similarity (Enright *et al.*, 2002). Functional traits were grouped at a cluster  
204 value of 90.2, whereby  $> 65$  is considered 'fair' and confidence in accurately separating

1  
2  
3 205 clusters cannot be higher than 100. The value of 90.2 is not chosen by the user but  
4  
5 206 rather is a reflection of the quality of clustering in a given dataset. The MCL identified  
6  
7 207 a total of 220 664 traits shared between at least two genomes. A random subset of  
8  
9 208 1700 amino acid sequences were selected and the similarity of each sequence within  
10  
11 209 its trait group (as determined by MCL) *versus* between other trait groups was visualised  
12  
13 210 as a box and whisker plot (Supplementary Figure 1) in R version 4.0.0 (R Core Team,  
14  
15 211 2013). 1 700 sequences were chosen to maximise comparisons between trait groups  
16  
17 212 under technical limitations, as increasing sequences led to exponential increases in  
18  
19 213 trait combinations. A Student's T test was applied to determine whether sequences  
20  
21 214 were more similar within their trait group relative to between trait groups in R. Finally,  
22  
23 215 a matrix of genome x functional trait was generated in a two-step process by first  
24  
25 216 associating genome IDs to the MCL output with a novel script 'MCLtoReshape2.py'  
26  
27 217 (available at the above Github address) and secondly by casting the long data format  
28  
29 218 to a wide data matrix with the 'reshape2' package in R (Wickham, 2007). Box and  
30  
31 219 whisker plots comparing counts of proteins per genome (input) and counts of functional  
32  
33 220 traits shared by at least two genomes (output of computational workflow), for the 35  
34  
35 221 Families, is presented as Figure 1.

### 222 223 224 *2.3 UPGMA dendrogram of trait similarity between genomes*

225 The unweighted pair group method with arithmetic mean (UPGMA) was chosen to  
226  
227 compare distance-based similarity between taxa based on discrete counts of individual  
228  
229 traits per genome. This method is more robust for comparing similarity between sample  
230  
231 units (*i.e.* taxa) based on discrete counts of variables (*i.e.* individual traits per taxon)  
232  
233 rather than neighbour joining or maximum likelihood methods better suited for DNA or  
234  
235 amino acid sequence comparisons (Weins, 1998). The UPGMA was performed in R  
236  
237 with the 'phangorn' package as described (Schliep *et al.*, 2017) on a Bray-Curtis  
238  
239 transformed dissimilarity functional trait matrix, generated with the 'vegdist' function in  
240  
241 the 'vegan' package (Oksanen *et al.*, 2013). To measure ecological coherence (C) of  
242  
243 taxa within shared Super Groups, Phyla and Families, a similarity index was adapted  
244  
245 from Levins' Overlap (Finn *et al.*, 2020a), which measures pairwise similarity in  
246  
247 distributions of taxa, as the following:

$$C = 1 - \left( \frac{\sum b_{ij}}{n^2} \right) \quad \text{Equation 1.}$$

1  
2  
3 239  
4  
5 240 Where  $b_{ij}$  is the pairwise branch length between taxon  $i$  and  $j$  in the UPGMA tree,  
6 241 measured here as Bray-Curtis dissimilarity, which is summed for each taxon and its  
7  
8 242 relatives within a shared Super Group, Phylum or Family, and where  $n$  is the number  
9  
10 243 of taxa being compared within a shared Super Group, Phylum or Family.

11  
12 244 Furthermore, the full length 16S rRNA gene of each taxon was collated from NCBI.  
13 245 Genes were aligned with MUSCLE (Edgar, 2004) and a neighbour joining phylogenetic  
14 246 tree was constructed with the 'phangorn' package in R. Phylogenetic distance present  
15 247 in taxonomic groups ( $P$ ) was measured as per Equation 1., excepting that branch  
16  
17 248 length was in units of DNA sequence similarity as opposed to Bray-Curtis dissimilarity.  
18  
19 249 Finally, simple linear regression was used to test a relationship between  $P$  and  $C$ .

20  
21  
22 250

23  
24 251

#### 25 252 *2.4 Functional trait annotation*

26  
27 253 To inform the biological process a functional trait facilitated, traits were annotated with  
28  
29 254 the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. This was  
30  
31 255 performed in five steps: 1) a representative amino acid sequence from each trait was  
32  
33 256 extracted with the novel script 'IdentifyTraits.py'; 2) these sequences were annotated  
34  
35 257 with KEGG Orthology (KO) terms using the BlastKOALA database algorithm with a bit  
36  
37 258 score cut-off value of 75 (Kanehisa *et al.*, 2016); 3) BRITE functional hierarchies  
38  
39 259 associated with each KO term (e.g. KO1179 gene: endoglucanase, BRITE 1:  
40  
41 260 Metabolism, BRITE 2: Carbohydrate Metabolism, BRITE 3: Starch and Sucrose  
42  
43 261 Metabolism) were collected with the novel script 'GetBRITEinfo.py'; 4) Genome ID, trait  
44  
45 262 ID, KO term and BRITE metadata were all collated with the novel script 'MatchFCs.py';  
46  
47 263 and 5) the 'reshape2' package in R was used to create matrices of genome x BRITE  
48  
49 264 hierarchy. Where KEGG was unable to annotate a trait, it was considered as  
50  
51 265 'Uncharacterised'. As above, all novel scripts and a step-by-step walkthrough of  
52  
53 266 reproducible code is available at: [https://github.com/DamienFinn/Trait-](https://github.com/DamienFinn/Trait-based_analyses)  
54  
55 267 [based\\_analyses](https://github.com/DamienFinn/Trait-based_analyses).

56  
57 268

58  
59 269

#### 60 270 *2.5 Identifying traits differentially enriched in taxonomic groups*

61 271 Random Forest classification was chosen as a non-linear, multivariate cluster-based  
62  
63 272 method capable of identifying numerous predictor variables (*i.e.* traits) that define



1  
2  
3 273 different classes of a response variable (*i.e.* taxonomic group). This was performed  
4  
5 274 with the 'randomForest' package as described (Liaw & Weiner, 2002). Discrete counts  
6  
7 275 of traits at BRITE level 3 were used (*e.g.* Starch and Sucrose Metabolism) as this level  
8  
9 276 had the most accurate resolution of biological processes facilitated by traits. A total of  
10  
11 277 six Random Forest models were optimised to classify taxonomic groups at the level of:  
12  
13 278 a) Phylum, with Proteobacteria Classes separated due to their extensive diversity ( $n =$   
14  
15 279 14); b) Family ( $n = 35$ ); c) specifically for Families in the Proteobacteria ( $n = 13$ ); d)  
16  
17 280 Families in the Actinobacteria ( $n = 7$ ); e) Families in the Firmicutes ( $n = 5$ ); and f)  
18  
19 281 Families from 'Under-represented' groups, which were all other Families ( $n = 10$ ).  
20  
21 282 Optimal numbers of trees grown for each model were: 300, 400, 320, 300, 300 and  
22  
23 283 400, respectively. Six traits were randomly selected at each branch. As the Random  
24  
25 284 Forest only identifies traits that best explain separation of taxonomic groups, and does  
26  
27 285 not show whether traits have positive or negative associations with groups, box and  
28  
29 286 whisker plots and Fisher's Least Significant Difference (LSD) *post hoc* tests were  
30  
31 287 performed with the 'agricolae' package (de Mendiburu, 2014) to definitively state which  
32  
33 288 taxonomic groups were significantly enriched or depleted in traits identified via  
34  
35 289 Random Forest.

32  
33  
34  
35

## 292 *2.6 Hierarchical clustering of Families by defining traits*

37  
38 293 Finally, the relationship between Families based on similarity in counts of 60 traits was  
39  
40 294 assessed via hierarchical clustering. Traits were chosen based on being selected via  
41  
42 295 the above Random Forest models in this study, and from previous studies that  
43  
44 296 identified traits associated with copiotroph-oligotroph growth strategies in single  
45  
46 297 species or mixed communities (Lauro *et al.*, 2009, Vieira-Silva & Rocha, 2010, Roller  
47  
48 298 & Schmidt, 2015, Pascual-Garcia & Bell, 2020). The mean of trait discrete counts in  
49  
50 299 the five Family members was used as representative of each Family. Comparing trait  
51  
52 300 means between Families was considered acceptable as prior LSD *post hoc* tests had  
53  
54 301 demonstrated significant differences between Families. As traits had highly variable  
55  
56 302 copy numbers per Family (*e.g.* ABC transporter trait copies ranged from 10 – 350,  
57  
58 303 while bacterial chemotaxis traits ranged from 0 – 15 copies) the trait copies were  
59  
60 304 normalised for more appropriate comparisons. Normalised variance was calculated  
61  
62 305 across the 35 Families for all traits with the 'decostand' function in the 'vegan' package  
63  
64 306 (Oksanen *et al.*, 2013). Hierarchical clustering of Families based on normalised trait

counts was visualised with the 'heatmap.2' function in the 'gplots' package (Warnes *et al.*, 2019).

309

310

### 3. Results

#### 3.1 Trait clustering and UPGMA

The 964 951 amino acid sequences encoded by the 175 genomes were clustered as 220 664 traits by MCL. A random subset of 1 700 traits showed that amino acid sequence similarity within traits ranged from 62.4, 82.5 and 100% for the 1<sup>st</sup> quartile, mean and 3<sup>rd</sup> quartile, respectively (Supplementary Figure 1). Amino acid sequence similarity between traits ranged from 27.6, 34 and 37.8% for the 1<sup>st</sup> quartile, mean and 3<sup>rd</sup> quartile, respectively. A Student's T test found that sequences grouped together as a trait were significantly more similar to each other than to sequences grouped as different traits ( $t$  value = 61.3,  $p = 2 \times 10^{-16}$ ). Manual comparisons of amino acid sequences within several traits supported clustering of proteins with identical biological function based on KEGG annotation. Thus, the MCL was considered to perform well. However, the minimum amino acid sequence similarity within traits was 23.27% and maximum similarity between traits was 85.24%, indicating that across the 220 664 traits, a small proportion of dissimilar amino acid sequences were grouped as a trait incorrectly, while some amino acid sequences that were highly similar were considered different traits. This small number of incorrectly clustered sequences can be explained by the MCL clustering efficiency being 90.2, out of a possible 100.

A comparison of the number of proteins per genome at the Family level (mean = 5 533), found that there were fewer functional traits per genome (mean = 4 240) (Figure 1). These represent the input number of proteins per genome before trait clustering and the output number of traits after clustering, respectively. As only functional traits shared between at least two genomes were considered here, the loss of highly genome-specific traits that could not be compared between genomes was expected. Despite this drop in average traits per genome, this initial approach serves as a proof of concept to demonstrate that numbers of traits per genome at the Family level reflect trends in proteins per Family, and thus the MCL was not distorting trait clustering (Figure 1).

The UPGMA dendrogram comparing the 220 664 traits per genome showed that trait compositions were non-randomly distributed (Figure 2a). Specifically,



1  
2  
3 341 Thaumarchaeota, Euryarchaeota, Acidobacteria, Betaproteobacteria,  
4  
5 342 Gammaproteobacteria, Alphaproteobacteria, Cyanobacteria, Verrucomicrobia,  
6  
7 343 Planctomycetes, Bacteroidetes, Actinobacteria, Firmicutes and Deltaproteobacteria  
8  
9 344 clustered together preferentially. The Chloroflexi were split into two clusters: one  
10 345 *Dehalococcoides* and one *Ktedonobacter/Herpetosiphon/Anaerolinea* cluster. Several  
11 346 prokaryotes did not cluster with their high taxonomic rank, including a Planctomycetes  
12 347 bacterium, *Polyangium brachysporum* (Deltaproteobacteria), *Agreia pratensis*  
13 348 (Actinobacteria) and *Sporomusa ovata* (Firmicutes). Also of interest was that, in regard  
14 349 to distance between terminal nodes (as noted by the scale bar), the Betaproteobacteria  
15 350 and Gammaproteobacteria were more similar to each other than the  
16 351 Alphaproteobacteria, which formed its own large, diverse clade. A neighbour joining  
17 352 tree of full length 16S rRNA genes showed that all taxa clustered preferentially based  
18 353 on their taxonomic nomenclature at high taxonomic rank, including the Chloroflexi,  
19 354 indicating that the discrepancies in the UPGMA were not due to misclassification of the  
20 355 individual taxa (Supplementary Figure 2a).

21  
22  
23  
24  
25  
26  
27  
28  
29 356 A simple index to measure trait similarity, as ecological coherence ( $C$ ), within groups  
30 357 was devised (Equation 1).  $C$  increased as taxonomic rank decreased: Super Group <  
31 358 Phylum < Family (Figure 2b).  $C$  was lowest for the larger, more diverse Proteobacteria  
32 359 and Terrabacteria (Super Group), and Firmicutes and Actinobacteria (Phylum). As the  
33 360 number of taxa being compared at the Super Group (e.g. Terrabacteria = 70 *versus*  
34 361 Acidobacteria = 10) and Phylum (e.g. Actinobacteria = 35 *versus* Thaumarchaeota =  
35 362 5) were variable, the most meaningful comparisons between groups are at the Family  
36 363 level ( $n = 5$  each). With the exception of the highly divergent 'Acidobacteria Lineage',  
37 364 all Families had a  $C$  greater than 0.3, with certain groups in the Alphaproteobacteria  
38 365 (Beijerinckiaceae), Firmicutes (Bacillaceae and Leuconostocaceae) and  
39 366 Actinobacteria (Promicromonosporaceae) being highly coherent ( $C > 0.55$ ). Indeed, all  
40 367 individual Proteobacterial Families had  $C > 0.4$ , indicating that all five taxa within each  
41 368 of these Families had similar trait compositions. Despite not truly belonging to the same  
42 369 Family as per NCBI taxonomy, the  $C > 0.33$  of the five Cyanobacteria,  
43 370 Thaumarchaeota and the Methanogen Lineage taxa was similar to other Families from  
44 371 the Bacteroidetes and Firmicutes. Thus, the UPGMA demonstrated that taxonomic  
45 372 relatives at the Phylum level tended to cluster with each other preferentially based on  
46 373 trait composition, and secondly that while similarity was broadly highest at low  
47 374 taxonomic rank, some Families were more coherent than others.

1  
2  
3 375 Phylogenetic distance ( $P$ ) of each taxonomic group increased with decreasing  
4 376 taxonomic rank, and was highest in Proteobacteria, Actinobacteria and Firmicutes  
5 377 Families (Supplementary Figure 2b). There was a strong positive linear relationship  
6 378 between  $P$  and  $C$  ( $y = 0.86x - 0.25$ ,  $R^2 = 0.39$ ,  $p < 0.001$ ) supporting the result that  
7 379 taxonomic groups of closer related taxa tended to share more similar compositions of  
8 380 traits.

13 381

14 382

### 17 383 *3.2 Random Forest trait identification*

18 384 The KEGG annotated traits belonged to 260 different BRITE 3 categories. The  
19 385 percentage of traits that could not be annotated (and were termed 'Uncharacterised')  
20 386 ranged from 28 – 65% per genome, being particularly high in the Archaea. On average,  
21 387 47% of traits per genome were Uncharacterised with a standard deviation of 9.5%.

22 388 Of the 260 BRITE 3 categories, the 60 most important traits in separating all Phyla  
23 389 and Families are ranked by importance measured as Mean Decrease in Accuracy  
24 390 (MDA) of the Random Forest models (Figure 3). This is a measure of the average  
25 391 increase in classification error during permutation of trees ( $n = 300 - 400$ ) when that  
26 392 particular trait is missing from the tree. For example, the accuracy of classifying  
27 393 Families was most improved by inclusion of the ABC transporters trait. Based on the  
28 394 identified traits, the Phylum model was capable of successfully classifying 81.14% of  
29 395 individual taxa. The Family model was capable of successfully classifying 71.43% of  
30 396 individual taxa. Confusion matrices for both models are presented as Supplementary  
31 397 Tables 2 – 4, and show that classification was particularly difficult for Chloroflexi and  
32 398 Planctomycetes (classification error > 80%) in the Phylum model and for  
33 399 Cellulomonadaceae and the divergent Acidobacteria Lineage (classification error >  
34 400 80%) in the Family model. Random Forest models were robust against variation in  $P$   
35 401 within Families, for example the nine families with all taxa perfectly classified ranged  
36 402 in  $P$  from the lowest (0.68) to highest (0.9).

37 403 The important traits in classifying the taxonomic groups were involved in: a)  
38 404 metabolism and nutrient acquisition (oxidative phosphorylation, tricarboxylic acid  
39 405 (TCA) cycle, glyoxylate/decarboxylate, thermogenesis, propanoate, starch/sucrose,  
40 406 nitrogen, methane metabolism, synthesis of antioxidants such as glutathione, ATP-  
41 407 binding cassette (ABC) transporters, sugar uptake via phosphotransferase systems  
42 408 (PTS)); b) responding to environmental cues and stressors (protein kinases, two-

1  
2  
3 409 component systems, transcription factors, proteasome, protein chaperones, RNA  
4 410 transport, chromosome repair via non-homologous DNA end joining); c) core cell  
5 411 physiology (flagella assembly, chemotaxis, sporulation, lipopolysaccharide (LPS),  
6 412 peptidoglycan, glycerolipid, sphingolipid and lipoarabinomannan (LAM) biosynthesis);  
7 413 and d) cell-cell interactions (beta-Lactam resistance, general secretion systems and  
8 414 Type IV secretion systems). Box and whisker plots of discrete counts of identified traits,  
9 415 and LSD results, are provided as Supplementary Figures 4 – 7. The Families  
10 416 significantly enriched and depleted in these traits are listed in Table 1. The Phyla  
11 417 significantly enriched and depleted in identified traits are listed in Supplementary Table  
12 418 9.

13  
14  
15 419 To better identify the more subtle differences between Families in the  
16 420 Proteobacteria, Actinobacteria, Firmicutes and the 'Under-represented' Phyla,  
17 421 individual Random Forest models were constructed for each of the four groups. The  
18 422 successful classification rates were 72.31, 74.29, 76 and 72%, respectively. Confusion  
19 423 matrices for each model are presented as Supplementary Tables 5 – 8. The models  
20 424 were unable to reliably classify Bradyrhizobiaceae and the divergent Acidobacteria  
21 425 Lineage (classification error > 80%).

22 426 Supplementary Figure 8 shows the most important traits in classifying the four  
23 427 groups. Unique traits not identified in the prior Phylum and Family models were: a) for  
24 428 Proteobacteria, glycosyltransferases, butanoate metabolism, aminotransferases,  
25 429 ribosome biogenesis, mRNA biogenesis and degradation; b) for Actinobacteria,  
26 430 porphyrin and chlorophyll synthesis, pyruvate metabolism, aminotransferases, fatty  
27 431 acid and aliphatic hydrocarbon metabolism, polyketide and Type II polyketide  
28 432 biosynthesis, antimicrobial resistance genes; c) for Firmicutes, lysine, folate and varied  
29 433 amino acid synthesis, porphyrin and chlorophyll synthesis, DNA replication, bacterial  
30 434 toxins, penicillin and cephalosporin synthesis; and d) for the 'Under-represented' taxa,  
31 435 glycosyltransferases, peptidases and inhibitors, photosynthesis and AMP-activated  
32 436 protein kinases. Box and whisker plots of discrete counts of identified traits, and LSD  
33 437 results, are provided as Supplementary Figures 9 – 12. Tables summarising Families  
34 438 enriched and depleted in these traits are included as Supplementary Tables 10 and  
35 439 11.

36 440

37 441

38 442 **3.3 Hierarchical clustering of defining traits**

1  
2  
3 443 Hierarchical clustering based on 60 traits, identified from Random Forest in this study  
4  
5 444 and by previous copiotroph-oligotroph studies, indicated five general clades. The  
6  
7 445 dendrogram on the y axis of Figure 4 shows clustering of taxa as these five clades.  
8  
9 446 The dendrogram on the x axis shows clustering of co-occurring traits. Clade I consisted  
10  
11 447 of Proteobacteria, specifically the Pseudomonadaceae, Burkholderiaceae,  
12  
13 448 Rhodospirillaceae, Bradyrhizobiaceae and Rhizobiaceae. These Families were  
14  
15 449 uniquely enriched in flagellar assembly, chemotaxis, pyruvate metabolism, glutathione  
16  
17 450 metabolism, ABC transporters, benzoate metabolism, transcription factors, glyoxylate  
18  
19 451 and fatty acid metabolism. Clade II, also Proteobacteria, included Nitrosomonadaceae,  
20  
21 452 Neisseriales Lineage, Methylocystaceae, Beijerinckiaceae, Methylococcaceae and  
22  
23 453 Moraxellaceae. These Families clustered based on being enriched in Clade I traits, but  
24  
25 454 to a lesser degree than the Pseudomonadaceae, Burkholderiaceae,  
26  
27 455 Rhodospirillaceae, Bradyrhizobiaceae and Rhizobiaceae. Exceptions included the  
28  
29 456 absence of benzoate metabolism and enrichment of methane metabolism in several  
30  
31 457 Clade II Families.

32  
33 458 Clade III, a diverse collection of Bacteroidetes (Chitinophagaceae, Cytophagaceae),  
34  
35 459 Verrucomicrobia, Planctomycetes, divergent Acidobacteria Lineage and the  
36  
37 460 Deltaproteobacteria (Polyangiaceae, Myxococcaceae), shared enrichment of  
38  
39 461 sphingolipid metabolism, beta-Lactam resistance, penicillin and cephalosporin  
40  
41 462 biosynthesis, LPS biosynthesis, glycosyltransferases and starch/sucrose metabolism.  
42  
43 463 Many of these Families shared Clade I and II traits, including Type IV secretion system,  
44  
45 464 oxidative phosphorylation, TCA cycle, PTS, nitrogen and glycerophospholipid  
46  
47 465 metabolism. The absence of glutathione in non-Deltaproteobacterial Clade III Families  
48  
49 466 was notable.

50  
51 467 The three Actinobacteria Families in Clade IV, Mycobacteriaceae, Frankiaceae and  
52  
53 468 Streptomycetaceae, were highly similar to each other because they were enriched in  
54  
55 469 Type II polyketide biosynthesis. They also shared some Clade I traits (ABC  
56  
57 470 transporters, transcription factors, pyruvate, benzoate and fatty acid metabolism) and  
58  
59 471 Clade III traits (membrane trafficking, transcription machinery, polyketide biosynthesis  
60  
472 and starch/sucrose metabolism). Similar to Clade III, these Actinobacteria were also  
473 depleted in glutathione traits.

474 Finally, taxa within Clade V were similar to each other due to being depleted in traits  
475 shared among the other clades. Cyanobacteria were the only taxa that possessed  
476 photosynthesis traits. Lactic acid bacteria (Lactobacillaceae and Leuconostocaceae)

1  
2  
3 477 were enriched in Unclassified nucleotide metabolism. The Archaea (Thaumarchaeota  
4 478 and Methanogen Lineage) shared eukaryote-like traits, proteasome, basal  
5 479 transcription factors and RNA transport. The Archaea were enriched in carbon fixation  
6 480 traits. Methanogens were also enriched in methane metabolism. Ammonia oxidising  
7 481 Thaumarchaeota were not enriched in nitrogen metabolism, however they were  
8 482 enriched in traits annotated by KEGG as Global maps only (unclassified metabolism).  
9 483 Further analysis found this to be the *fpr* gene, encoding a ferredoxin-flavodoxin NADP<sup>+</sup>  
10 484 reductase (K00528). Non-lactic acid bacteria of the Firmicutes (Bacillaceae,  
11 485 Sporomusaceae and Clostridiaceae) were enriched in sporulation and motility traits.  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21

486  
487

## 488 4. Discussion

### 489 4.1 Non-random trait clustering demonstrates ecological coherence of taxa

490 A trait-based approach to investigate taxonomic relationships and potential biological  
491 function was carried out with a collection of 175 terrestrial prokaryotes. We  
492 hypothesised that traits would be non-randomly distributed amongst taxonomic groups,  
493 supported by previous observations that noted closely related taxa are isolated from  
494 similar habitats (Philippot *et al.*, 2010). Similarity in the composition of 220 664 traits,  
495 within 175 taxa, demonstrated strong agreement with established taxonomy at high  
496 (Phyla) and low (Family) rank (Figure 2a). The exceptions to this ecological coherence  
497 at high rank were the division of the Chloroflexi and an individual from each of the  
498 Planctomycetes, Deltaproteobacteria, Actinobacteria and Firmicutes clustering with  
499 unrelated Phyla. These taxa were not mischaracterised, as based on phylogenetics of  
500 the full length 16S rRNA gene (Supplementary Figure 2a). Trait similarity between  
501 related taxa, measured as *C*, tended to be highest at low rank (Figure 2b). Uneven  
502 sample sizes between groups within Super Group and Phyla make comparisons at this  
503 level difficult – the inclusion of many diverse Firmicutes and Actinobacteria likely drove  
504 *C* to be lower here than in Thaumarchaeota and Euryarchaeota. However, equal  
505 comparisons at the Family level demonstrated interesting variability in coherence. All  
506 Proteobacterial Families had relatively high coherence ( $C = 0.4 - 0.6$ ). The high *C* in  
507 Beijerinckiaceae is of particular interest as this group contained both specialist  
508 methanotroph (*Methylocapsa*, *Methylocella*), methylotroph (*Methyloferula*) and  
509 generalist heterotroph (*Beijerinckia* spp.) taxa. With such varied metabolic traits, one  
510 could reasonably expect *C* to be relatively low within this Family. The Beijerinckiaceae



1  
2  
3 511 appear to have evolved from a common methylo-trophic ancestor and still share traits  
4 512 for nitrogen fixation and tolerance for low pH soils (Tamas *et al.*, 2014), and the high  
5 513 C measured here indicates that many additional shared traits remain. Both the  
6 514 relatively recent divergence of Families from a common ancestor and the higher  
7 515 number of shared traits are likely causes of the higher C observed at low taxonomic  
8 516 rank. The differing values of C for Methanogen and photosynthetic Cyanobacteria  
9 517 functional groups (0.33 and 0.44, respectively) is also worthy of note. Despite all five  
10 518 taxa in each group performing the same core role in a community, the individual  
11 519 isolates came from varying environments. The methanogens were isolated from a  
12 520 range of geographically separate wetlands, rice paddy soil and farm slurry and, while  
13 521 the Cyanobacteria were also isolated from geographically separate environments, they  
14 522 were all from sandy deserts or other nutrient poor, arid soils (Supplementary Table 1  
15 523 and references therein). Ultimately a taxon's trait composition will be affected by its  
16 524 functional role in a community, its evolutionary life-history (e.g. Beijerinckiaceae  
17 525 described above) and its local environment.

18 526 However, these results are dependent on accurate taxonomic classification, and the  
19 527 C of Sporomusaceae, relatively low compared to other Families here, supports splitting  
20 528 this group into Sporomusaceae, Anaeromusaceae and Pelosinaceae by GTDB (Parks  
21 529 *et al.*, 2018). Finally worth noting, some groups at high rank were considered as  
22 530 'Families' here due to the number of available terrestrial genomes, e.g. Cyanobacteria  
23 531 and Chloroflexi. Even so, Cyanobacteria demonstrated a higher C than many  
24 532 taxonomically-defined Families, perhaps due to their common role as primary  
25 533 colonisers of nutrient-poor soils (Garcia-Pichel *et al.*, 2001). The number of Families  
26 534 are too numerous to discuss each at length here, but C was an effective means of  
27 535 measuring and comparing coherence between groups in the UPGMA tree.

28 536 While the method of comparing taxa here differs from other studies, the results were  
29 537 not surprising as many 16S rRNA gene surveys of terrestrial systems consistently  
30 538 demonstrate ecological coherence at high rank. For example, independent studies of  
31 539 increasing agricultural intensity in soils show reductions in Actinobacteria abundance  
32 540 (Philippot *et al.*, 2009, Jangid *et al.*, 2011). Nitrogen addition to soils frequently enriches  
33 541 numerous taxa within the Actinobacteria and Proteobacterial Classes while negatively  
34 542 affecting taxa within the Verrucomicrobia and Planctomycetes (Wessen *et al.*, 2010,  
35 543 Fierer *et al.*, 2012, Leff *et al.*, 2015, Bastida *et al.*, 2016). Arid, nutrient poor  
36 544 environments select for Actinobacteria-dominated communities (Cary *et al.*, 2010,

1  
2  
3 545 Crits-Christoph *et al.*, 2013) and, in the absence of other primary producers, allow  
4 546 biological soil crust forming Cyanobacterial taxa to establish (Garcia-Pichel *et al.*,  
5 547 2001). Anoxic wetland and rice paddy environments support diverse communities of  
6 548 anaerobic Firmicutes, Chloroflexi and methanogenic Archaea (He *et al.*, 2019, Finn *et*  
7 549 *al.*, 2020b). These trends were noted prior to bioinformatic advances of metagenome  
8 550 assembled genomes (MAGs) that allow for the specific comparison of individual traits  
9 551 between uncultured environmental prokaryote genomes (Hug *et al.*, 2013). The  
10 552 generation of MAGs has emerged as a useful tool for identifying traits necessary for  
11 553 life in such environments, and particularly for expanding knowledge of severely under-  
12 554 represented, difficult to culture taxonomic groups. For example, the recent  
13 555 reconstruction of 52 515 MAGs from a wide range of host-associated and  
14 556 environmental metagenomes was able to increase genomic information of  
15 557 Planctomycetes and Verrucomicrobia by 79% and 68%, respectively (Nayfach *et al.*,  
16 558 2021). Importantly, both 16S rRNA gene surveys and MAGs demonstrate that some  
17 559 functional traits that facilitate life under certain environmental conditions are  
18 560 intrinsically linked to taxonomy.

19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31 561 The ecological coherence observed in Figure 2 does not imply that closely related  
32 562 taxa (*e.g. Bacillus velezensis* LS69 and *Bacillus amyloliquefaciens plantarum* FZB42)  
33 563 have identical phenotypes. Close relatives possess a combination of core and  
34 564 accessory genes (traits) and the presence of even a single accessory gene is sufficient  
35 565 to dramatically alter a strain's phenotype (van Rossum *et al.*, 2020). Rather, our results  
36 566 (Figure 2) demonstrate that the composition of core and accessory traits in  
37 567 alphaproteobacterial Beijerinckiaceae are most similar to each other, relative to  
38 568 alphaproteobacterial Rhizobiaceae or to Actinobacteria, Firmicutes etc.

39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

#### 571 4.2 What can the Random Forests tell us?

572 The Random Forest works by identifying the traits that are most reliable in classifying  
573 individual Phyla and Families. It selects traits that tend to be: a) of equal copy number  
574 per genome within a taxonomic group; and b) that differ markedly in copy number  
575 between taxonomic groups, since distinct separation of copies will maximise  
576 successful classification. A clear example of this is the consistent identification of  
577 eukaryote-like basal transcription factors, proteasome and RNA transport present in  
578 the Thaumarchaeota and Euryarchaeota, since they are absent from the majority of



1  
2  
3 579 Bacterial Phyla. The binary nature of these traits (yes Archaea, no Bacteria) make  
4  
5 580 them strong indicators. The presence of these and more eukaryote-like vesicular  
6  
7 581 trafficking and actin traits have been noted in the Archaeal TACK super-phylum  
8  
9 582 previously, and lend credence to the hypothesis that eukaryotes are descended from  
10  
11 583 Archaea (Embley & Martin, 2006, Spang *et al.*, 2015). However, the Random Forest  
12  
13 584 will not identify a trait unique to *Can. Nitrosotalea devanterra* that is absent from other  
14  
15 585 Thaumarchaeota, as this single trait will not improve classification of the group as a  
16  
17 586 whole. Consequently, the traits identified via Random Forest all tended to be core,  
18  
19 587 fundamental traits shared by other members of a taxon's Phylum/Family.

20  
21 588 Many of the best traits for distinguishing taxa have been historically used by  
22  
23 589 microbiologists to do exactly that. These included fundamental cell physiology traits,  
24  
25 590 such as oxidative phosphorylation, LPS biosynthesis, sporulation, flagellar assembly  
26  
27 591 and chemotaxis. The Phylum model separated Betaproteobacteria, Actinobacteria and  
28  
29 592 Bacteroidetes as taxa with the highest copies of oxidative phosphorylation traits.  
30  
31 593 Firmicutes, Chloroflexi and Methanogens were identified as anaerobes depleted in  
32  
33 594 oxidative phosphorylation, and all other taxa as sitting in between (Supplementary  
34  
35 595 Figure 4 and Supplementary Table 9). Some Gram-negative Families were  
36  
37 596 significantly enriched in LPS biosynthesis compared to others. These were the  
38  
39 597 Pseudomonadaceae, Burkholderiaceae, Chitinophagaceae, divergent Acidobacteria  
40  
41 598 and Verrucomicrobia (Table 1). This has been noted in soil communities previously.  
42  
43 599 The extensive repertoire of LPS-associated genes in Chitinophagaceae, Acidobacteria  
44  
45 600 and Verrucomicrobia likely play a critical role in enhancing soil aggregation (Cania *et*  
46  
47 601 *al.*, 2019) potentially through high LPS production and/or biofilm formation (discussed  
48  
49 602 further below). In a demonstration of the robustness of the methods used here, the  
50  
51 603 highly unusual Firmicute Sporomusaceae were shown to possess similar counts of  
52  
53 604 LPS biosynthesis traits relative to most Gram-negative Families (Supplementary Table  
54  
55 605 4) in addition to sharing heat-resistant spore formation with Bacillaceae and  
56  
57 606 Clostridiaceae (Table 1). The presence of both traits in a single Family have been  
58  
59 607 remarked upon previously and used to conceptualise the evolution of Gram-negative  
60  
61 608 *versus* Gram-positive lineages (Stackebrandt *et al.*, 1985). The Sporomusaceae were  
62  
63 609 also shown to have high Porphyrin and Chlorophyll metabolism traits in the Firmicutes  
64  
65 610 model (Supplementary Table 10). The capacity to dechlorinate the soil pollutant  
66  
67 611 perchloroethene to trichloroethylene via a porphyrin-based corrinoid is yet another  
68  
69 612 interesting trait of this Family (Terzenbach & Blaut, 1994).

1  
2  
3 613 Finally, flagella assembly and chemotaxis traits identified Alphaproteobacteria,  
4 614 Betaproteobacteria and Acidobacteria as Phyla that were particularly enriched with this  
5 615 mechanism of motility, while bacterial Actinobacteria, Bacteroidetes, Cyanobacteria,  
6 616 Chloroflexi and Verrucomicrobia were depleted (Supplementary Table 9). The  
7 617 Proteobacteria, Firmicutes and Under-represented models were better suited for  
8 618 identifying specific Families homogenously enriched or depleted in bacterial flagella  
9 619 and chemotaxis (Supplementary Tables 10 and 11). Enriched Families included the  
10 620 Rhodospirillaceae, Nitrosomonadaceae, Neisseriales lineage, divergent Acidobacteria  
11 621 lineage, Sporomusaceae, Bacillaceae, Clostridiaceae and Planctomycetes. Other  
12 622 forms of motility such as twitching and gliding have been noted in the  
13 623 Pseudomonadaceae, Myxococcaceae and Cyanobacteria (McBride, 2001) but these  
14 624 traits were not identified by the Random Forest as being homogenously enriched in  
15 625 any Families. Furthermore, while Thaumarchaeota and Methanogens were both  
16 626 identified as being depleted in bacterial flagella assembly and chemotaxis traits  
17 627 (Supplementary Table 9 and 11), Archaea possess a structurally distinct flagellum  
18 628 more similar to the Type IV bacterial pilus (Jarrell and Albers, 2012). These taxa were  
19 629 not enriched with Type IV pilus, either, and it is possible that archaeal flagella may  
20 630 have failed proper characterisation by KEGG. Thus, while some Families were  
21 631 relatively enriched/depleted in bacterial flagella and chemotaxis traits, specific taxa  
22 632 depleted in these are not necessarily non-motile.

23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37 633 In summary, while the Random Forests may overlook certain traits in individual taxa,  
38 634 the models were highly robust in detecting conserved, shared traits within a  
39 635 Phylum/Family. Here the 'depth' of shared traits is limited by the number of taxa that  
40 636 could be considered as Phylum or Family. In future, if five (or more) taxa belonging to  
41 637 the same Genus or even Species could be compared, unique traits would be observed  
42 638 to explain how these subgroups have evolved from their respective Families to occupy  
43 639 distinct niches. Ideally the selection of individual taxa within groups for such future  
44 640 comparative analyses would also be standardised based on phylogenetic distance,  
45 641 either with *P* or a similar method, that would improve the robustness of trait-based  
46 642 comparisons at such a fine taxonomic level.

47 643

48 644

49 645 *4.3 Plant-derived carbon metabolism and nutrient acquisition*

50 646

51 647

52 648

53 649

54 650

1  
2  
3 646 Secondly, we hypothesised that the traits differentially enriched between taxonomic  
4  
5 647 groups would largely reflect those associated with copiotrophs or oligotrophs, namely  
6  
7 648 metabolism, nutrient acquisition and environmental stress response and tolerance. Of  
8  
9 649 fundamental interest to soil microbiologists is the decomposition of plant biomass. This  
10  
11 650 is the primary source of organic carbon to non-arid terrestrial systems (Kögel-Knabner,  
12  
13 651 2002) and the transformation of plant material to substrates bioavailable for  
14  
15 652 microorganisms is essential for community growth and activity. The traits involved in  
16  
17 653 plant material catabolism belonged to the BRITE categories 'Starch and Sucrose  
18  
19 654 Metabolism' (e.g. extracellular cellobiosidases, endoglucanases, glucosidases,  
20  
21 655 trehalases, amylases) and 'Glycosyltransferases', all of which are carbohydrate  
22  
23 656 activated enzymes (CAZy). The Families particularly enriched in these traits were the  
24  
25 657 Polyangiaceae, Myxococcaceae, Rhizobiaceae, Streptomycetaceae,  
26  
27 658 Mycobacteriaceae, Frankiaceae and Verrucomicrobia (Table 1, Supplementary Table  
28  
29 659 11). Genomic and culture-dependent analyses support *Sorangium cellulosum*  
30  
31 660 (Polyangiaceae), *Streptomyces coelicolor* A3(2) (Streptomycetaceae) and  
32  
33 661 *Chthoniobacter flavus* (Verrucomicrobia) as having particularly large genomes with  
34  
35 662 extensive repertoires for cellulose, hemicellulose, pectin and lignin degradation  
36  
37 663 (Bentley *et al.*, 2002, Sangwan *et al.*, 2004, Schneiker *et al.*, 2007). Comparative  
38  
39 664 genomics analyses have also identified Actinobacteria, Acidobacteria and  
40  
41 665 Verrucomicrobia as being enriched in numerous enzymes for cellulose, hemicellulose  
42  
43 666 and starch catabolism (Trivedi *et al.*, 2013). *In situ* these Families likely play a critical  
44  
45 667 role in making organic carbon bioavailable as di- and monosaccharides for the  
46  
47 668 microbial community.

48  
49 669 The Families equipped with many copies of high-affinity sugar uptake  
50  
51 670 'Phosphotransferase systems' (3 – 14 copies) did not necessarily correspond to those  
52  
53 671 enriched with CAZy – only the Myxococcaceae and Verrucomicrobia were enriched in  
54  
55 672 both. Pseudomonadaceae, Rhodospirillaceae, Neisseriales lineage and  
56  
57 673 Clostridiaceae were only enriched in PTS. Despite being enriched in CAZy, the  
58  
59 674 Frankiaceae were simultaneously depleted in PTS. The complex interplay between  
60  
61 675 taxa capable of producing extracellular CAZy and competitors that rapidly scavenge  
62  
63 676 available di- and monosaccharides has been well described by models that predict  
64  
65 677 such competitive interactions exert important controls on the growth rate of the  
66  
67 678 community as a whole (Freilich *et al.*, 2011) and may even act to aid terrestrial carbon  
68  
69 679 storage and limit carbon dioxide emissions from microbial respiration (Kaiser *et al.*,

1  
2  
3 680 2015). Here, we identified the ‘specialist’ Families enriched in CAZy traits *versus* the  
4  
5 681 ‘opportunists’ scavenging for sugars via PTS (Table 1, Supplementary Table 11).

6 682 ABC transporters facilitate the ATP-dependent uptake of soluble compounds across  
7  
8 683 membranes or export waste metabolites, extracellular enzymes and toxins (Young &  
9  
10 684 Holland, 1999, Higgins, 2001). This means of active transport allows microorganisms  
11  
12 685 to acquire nutrients with high affinity at concentrations of 5 – 500  $\mu\text{g carbon L}^{-1}$  *versus*  
13  
14 686 the less efficient diffusion of nutrients across membranes, dependent on extracellular  
15  
16 687 concentrations of 0.5 – 5 mg carbon  $\text{L}^{-1}$  (Kuznetsov *et al.*, 1979). In the spatially  
17  
18 688 heterogenous soil environment where the concentration of bioavailable carbon  
19  
20 689 substrate often limits growth (Blagodatsky & Richter, 1998), possession of high affinity  
21  
22 690 transporters likely provides a competitive advantage. The rhizosphere-associated  
23  
24 691 Rhodospirillaceae, Rhizobiaceae and Burkholderiaceae tended to have the highest  
25  
26 692 trait copies of ABC transporters (100 – 350 copies per genome, Supplementary Figure  
27  
28 693 4). The diverse, non-rhizospheric Deltaproteobacteria, Actinobacteria, Firmicutes,  
29  
30 694 Cyanobacteria and Verrucomicrobia all had greater than 50 copies per genome,  
31  
32 695 highlighting the importance of these traits in soil. The particularly high gene copy  
33  
34 696 number in rhizosphere-associated taxa from presumably nutrient-rich environments  
35  
36 697 contrasts the assumption that ABC transporters are considered to play a greater role  
37  
38 698 in nutrient-poor environments (Lauro *et al.*, 2009). Comparative genomics analyses of  
39  
40 699 soil bacteria have also found putatively copiotrophic Proteobacteria and Firmicutes to  
41  
42 700 be particularly enriched in PTS and ABC transporters (Trivedi *et al.*, 2013). In this  
43  
44 701 study, the transporters enriched in rhizosphere-associated taxa were primarily aimed  
45  
46 702 at scavenging maltose, phosphate, amino acids, oligopeptides and export of LPS, and  
47  
48 703 these results suggest that these traits are not only for survival in nutrient-poor  
49  
50 704 environments but also likely confer a competitive advantage in the rhizosphere. As  
51  
52 705 prokaryotes compete simultaneously with other prokaryotes and plants for nitrogen  
53  
54 706 and phosphorus in the rhizosphere, the high affinity acquisition of such nutrients is  
55  
56 707 likely critical for growth.

57 708

58 709

#### 59 710 *4.4 Nitrogen and methane metabolism*

60 711 The BRITE category ‘Nitrogen metabolism’ encompasses nitrogen fixation,  
61  
62 712 denitrification, ammonia oxidation and synthesis of glutamate/glutamine which are  
63  
64 713 critical amino acids for peptide synthesis. Since nitrogen limitation acts as an important

1  
2  
3 714 control on soil microbial activity, these traits are also of interest to soil microbiologists.  
4  
5 715 Three Proteobacterial Families, Rhodospirillaceae, Bradyrhizobiaceae and  
6  
7 716 Burkholderiaceae, were particularly enriched in these traits. Genomic and culture-  
8  
9 717 dependent analyses show these Families to be free-living or symbiotic diazotrophs in  
10  
11 718 soil and freshwater environments (Madigan *et al.*, 1984, Itakura *et al.*, 2009, de los  
12  
13 719 Santos *et al.*, 2018). Given their significantly greater copies of nitrogen-fixing genes,  
14  
15 720 these Families may be a particularly important source of organic nitrogen for soil  
16  
17 721 communities. Saprotrophic Mycobacteriaceae genomes, also identified as nitrogen  
18  
19 722 cyclers, tend to have many copies of genes involved in ammonia uptake and glutamate  
20  
21 723 synthesis (Amon *et al.*, 2010). This taxon may play an alternative role in converting  
22  
23 724 mineral nitrogen to biomass where organic nitrogen as protein in excreted products or  
24  
25 725 necromass can undergo proteolysis and uptake between other community members.  
26  
27 726 The identification of Sporomusaceae as enriched in 'Nitrogen metabolism' traits is  
28  
29 727 unusual as these obligate anaerobic fermenters cannot use nitrate as an electron  
30  
31 728 acceptor (Möller *et al.*, 1984). Nor were the Sporomusaceae enriched in ammonia  
32  
33 729 uptake or glutamate synthesis genes (data not shown), and so it is uncertain what role  
34  
35 730 this Family plays in nitrogen cycling. Thaumarchaeota and Nitrosomonadaceae, known  
36  
37 731 ammonia oxidisers, were not enriched in 'Nitrogen metabolism' traits relative to other  
38  
39 732 Families (Supplementary Figure 5) despite Nitrosomonadaceae possessing multiple  
40  
41 733 copies of the operon responsible for ammonia oxidation (Klotz & Norton, 1998).  
42  
43 734 Specific traits may be overlooked here if the BRITE category includes many diverse  
44  
45 735 KOs (e.g. ammonia oxidation, nitrogen fixation, glutamate synthesis etc).

46  
47 736 Another specialised metabolic pathway of interest involves 'Methane metabolism'  
48  
49 737 that includes production and oxidation of a potent greenhouse gas. Unsurprisingly, the  
50  
51 738 Methanogens and methanotrophic Methylococcaceae, Beijerinckiaceae,  
52  
53 739 Methylocystaceae were all enriched in traits involved in methane metabolism. While  
54  
55 740 methane oxidation can be present in some taxa from the Verrucomicrobia (Op den  
56  
57 741 Camp *et al.*, 2009) the above proteobacterial representatives act as the primary  
58  
59 742 terrestrial methane sink (Dunfield, 2007, Conrad, 2009).

60 743

744

#### 745 *4.5 Sensing, responding and tolerating the environment*

746 A particularly interesting divergence of traits were involved in how taxa detect and  
747 respond to environmental stimuli. Gram-negative Pseudomonadaceae,



1  
2  
3 748 Rhodospirillaceae, Bradyrhizobiaceae and Myxococcaceae were enriched in two-  
4  
5 749 component systems. These membrane-bound histidine kinases respond rapidly to  
6  
7 750 extracellular stimuli (Galperin, 2005) and these traits were primarily involved in  
8  
9 751 nitrogen, potassium, initiating chemotaxis and C<sub>4</sub>-dicarboxylate responses. Families  
10  
11 752 enriched in transcription factors were the Myxococcaceae, Polyangiaceae,  
12  
13 753 Streptomycetaceae, Mycobacteriaceae and Frankiaceae. These factors regulate  
14  
15 754 transcription in response to intracellular cues and here these factors were primarily  
16  
17 755 *rpoD* (housekeeping), *rpoH* (heat-shock/protein damage), *rpoE* (extra-cellular  
18  
19 756 cytoplasmic stress) and *rpoS* (starvation) responses (Shimada *et al.*, 2017). The  
20  
21 757 genomes of these taxa are also heavily enriched in regulatory genes for complex  
22  
23 758 developmental stages, fruiting bodies and/or filamentous branching growth in soils  
24  
25 759 (Bentley *et al.*, 2002, Gao *et al.*, 2006, Schneiker *et al.*, 2007). Thus, certain taxa may  
26  
27 760 respond primarily to extracellular cues while others strictly monitor and respond to  
28  
29 761 changes in cell homeostasis. This trend has been noted previously – in 167 genomes  
30  
31 762 across various Bacteria and Archaea, Proteobacteria had a higher ratio of sensors for  
32  
33 763 external *versus* internal stimuli and were considered ‘extroverts’, while Cyanobacteria  
34  
35 764 were considered strong ‘introverts’ focussed on responding to internal stimuli  
36  
37 765 (Galperin, 2005).

34 766 As mentioned above, Archaea exhibited unique traits in basal transcription and  
35  
36 767 protein regulation via proteasome. These transcription factors were primarily involved  
37  
38 768 in identifying DNA damage and excision repair: TFII-B, TFII-D, ERCC-2 and ERCC-3.  
39  
40 769 DNA repair differs markedly between Bacteria and Archaea/eukaryotes. Specifically,  
41  
42 770 Bacteria excise 12 nucleotides around a damaged site with a 3 polypeptide system  
43  
44 771 whereas Archaea excise 24 – 32 nucleotides with a 13 – 16 polypeptide system  
45  
46 772 (Sancar, 1996). The use of ubiquitin-labelling and proteasome degradation of  
47  
48 773 misfolded or ‘old’ proteins is arguably a more efficient system for recycling amino acids  
49  
50 774 and regulating the ‘lifespan’ of a protein in eukaryotes, however, Bacteria are still fully  
51  
52 775 capable of regulating protein misfolding or proteolysis with RpoH (and others) induced  
53  
54 776 upon environmental stress (Goldberg, 2003). From an ecological perspective, it is  
55  
56 777 difficult to discern if these eukaryote-like traits confer any sort of competitive advantage  
57  
58 778 to Archaea. They may simply be examples of convergent evolution for dealing with  
59  
60 779 environmental stress.

58 780 Finally, most microbial cells likely exist within complex biofilms and/or assemblages  
59  
60 781 adhered to surfaces with excreted exopolysaccharides, DNA and protein that serve to

1  
2  
3 782 protect from adverse environmental factors (Flemming & Wingender, 2010). Families  
4  
5 783 with high copy numbers of exopolysaccharide biosynthesis and secretion systems may  
6  
7 784 act as integral members of soil communities by predominantly contributing to  
8  
9 785 biofilm/aggregate formation. The 'LPS biosynthesis' and 'Starch and sucrose  
10  
11 786 metabolism' BRITE categories can synthesise N-acetyl glucosamine-based and  
12  
13 787 cellulose-based exopolysaccharides, respectively. Taxa enriched in both these  
14  
15 788 categories and secretion systems were the Polyangiaceae and Burkholderiaceae  
16  
17 789 (Table 1), and in the refined 'Under-represented' model, Acidobacteria and  
18  
19 790 Verrucomicrobia (Supplementary Table 11).

791

792

#### 793 4.6 Direct cell-cell interactions

794 Type IV secretion systems were another important trait identified in the Random Forest  
795 models. These were enriched in Pseudomonadaceae and Myxococcaceae (Table 1)  
796 and Acidobacteria, Planctomycetes and Verrucomicrobia (Supplementary Table 11).  
797 These are highly specialised exporters that deliver DNA and/or toxins directly to other  
798 bacterial or plant cells, however, their role in ecology is poorly understood beyond root  
799 galls induced by *Agrobacterium tumefaciens* (Christie & Vogel, 2000). These taxa  
800 should be explored for whether they utilise these traits for horizontal gene transfer or  
801 to inject toxins directly into other prokaryotes, and thus potentially provide a selective  
802 advantage for colonisation and competition.

803 Another archetypal trait for interactions between community members are  
804 production of antimicrobials and antimicrobial resistance genes. Penicillin and  
805 cephalosporin synthesis were enriched in the Sporomusaceae and Bacillaceae relative  
806 to other Firmicutes. Polyketide and Type II polyketide syntheses were important for  
807 separating Frankiaceae and Streptomycetaceae from other Actinobacteria  
808 (Supplementary Table 10). The Streptomycetaceae have a long history of use in  
809 biotechnology as prolific antimicrobial producers (Bentley *et al.*, 2002). Bacillaceae (in  
810 particular *Bacillus subtilis* species) are also well known producers of a wide variety of  
811 antimicrobials (Caulier *et al.*, 2019), but we noted that Sporomusaceae have an even  
812 greater number of these traits (Supplementary Figure 9). To the authors' knowledge,  
813 antibiotic production in Sporomusaceae has not been investigated thoroughly and this  
814 may be a consequence of its obligate anaerobic nature and difficulties in culturing. In  
815 addition to prolific Type II polyketide producers, Streptomycetaceae were also enriched



1  
2  
3 816 in antimicrobial resistance genes, while Bacteroidetes, Planctomycetes and  
4 817 Verrucomicrobia were specifically enriched in beta-Lactam resistance (Supplementary  
5 818 Tables 9 and 10).

8 819

10 820

#### 11 821 *4.7 Life strategies emerge from differentially enriched traits*

12 822 We hypothesised that taxa would emerge as being inherently copiotrophic or  
13 823 oligotrophic based on trends in their enriched traits. Traits were chosen based on  
14 824 identification via Random Forest and identification as associated with copiotroph-  
15 825 oligotroph species or in mixed communities as described previously (Lauro *et al.*, 2009,  
16 826 Vieira-Silva & Rocha, 2010, Roller & Schmidt, 2015, Pascual-Garcia & Bell, 2020).  
17 827 Rhizosphere-associated Gamma-, Alpha- and Betaproteobacteria in Clade I fit the  
18 828 assumptions of a copiotrophic niche that invests in high metabolic rate – these taxa  
19 829 were uniquely enriched in competing for nutrient uptake via high-affinity ABC  
20 830 transporters, and energy generation from pyruvate, fatty acids, benzoate and  
21 831 glyoxylate carbon sources. Clade I was also enriched in glutathione metabolism, which  
22 832 acts as the major antioxidant for reducing intracellular free radicals produced during  
23 833 central carbon metabolism (Smirnova & Oktyabrsky, 2005). Antioxidants have been  
24 834 hypothesised as an essential function for copiotrophs to survive their high metabolic  
25 835 rates (Koch, 2001). All five Clade I Families were enriched in oxidative phosphorylation.  
26 836 The oxidative phosphorylation traits encompass a wide variety of electron transport  
27 837 chain proteins (oxidoreductases, dehydrogenases, cytochromes and ATPases) and  
28 838 are crucial for efficient energy production (Brochier-Armanet *et al.*, 2009). All five  
29 839 Families were also enriched in nitrogen metabolism, which included both nitrogen  
30 840 fixation and glutamate (*i.e.* protein) synthesis. Nitrogen fixation is an energy intensive  
31 841 process requiring 20 – 30 ATP per reduced N<sub>2</sub> (Burris & Roberts, 1993) and may be  
32 842 intrinsically linked to taxa with high oxidative phosphorylation. Finally, Clade I also  
33 843 shared motility and chemotaxis, which are also energy intensive traits. Clade II  
34 844 consisted of the remaining Gamma-, Alpha- and Betaproteobacteria, yet these were  
35 845 relatively less enriched in Clade I 'copiotroph' traits. These particular taxa may be  
36 846 responsible for the lack of a consistent copiotrophic response upon nutrient addition in  
37 847 Proteobacteria (Ho *et al.*, 2017).

38 848 Clade III was comprised of taxa generally considered as oligotrophs (Ho *et al.*, 2017)  
39 849 with the exception of Bacteroidetes (Fierer *et al.*, 2007). These taxa possessed high

1  
2  
3 850 LPS and sphingolipid synthesis that can defend against desiccation and antimicrobials  
4  
5 851 through biofilm and capsule/slime production (Flemming & Wingender, 2010), beta-  
6  
7 852 Lactam resistance, penicillin biosynthesis and several members had high pentose  
8  
9 853 phosphate pathway for efficient carbon metabolism under starvation (Hodgson, 2000).  
10  
11 854 Clade III also possessed high CAZy traits, which Clade I largely lacked, and is  
12  
13 855 consistent with observations of oligotrophs being primarily responsible for catabolising  
14  
15 856 relatively recalcitrant plant material (Goldfarb *et al.*, 2011). While Clade III were equally  
16  
17 857 enriched in oxidative phosphorylation as Clade I, with the exception of the  
18  
19 858 Deltaproteobacteria, these taxa were depleted in glutathione metabolism. The low  
20  
21 859 copies per genome of this trait would explain why the abundance of oligotrophs drop  
22  
23 860 rapidly in nutrient addition studies as they would be either out-competed by  
24  
25 861 glutathione-rich taxa capable of exploiting plentiful nutrients or will lyse if their  
26  
27 862 metabolic rate exceeds capacity to reduce free radicals (Koch, 2001). Taken together,  
28  
29 863 all of these traits indicate Clade III lead oligotrophic lifestyles whereby they are tolerant  
30  
31 864 to adverse environmental conditions, can acquire carbon from recalcitrant plant  
32  
33 865 material, and are incapable of rapid growth rates.

34  
35 866 These results support previous observations that Rhodospirillaceae,  
36  
37 867 Bradyrhizobiaceae, Burkholderiaceae, Pseudomonadaceae and Rhizobiaceae are  
38  
39 868 copiotrophic, while Planctomycetes, Verrucomicrobia, Myxococcaceae,  
40  
41 869 Polyangiaceae and Acidobacteria are oligotrophic (Ho *et al.*, 2017 and references  
42  
43 870 therein). As has been proposed previously, the dominance of these groups in certain  
44  
45 871 soils can provide inferences for ecosystem processes in that system, for example soils  
46  
47 872 dominated by Verrucomicrobia, Planctomycetes and Acidobacteria will have greater  
48  
49 873 capacity to degrade complex plant material while retaining most catabolised carbon in  
50  
51 874 biomass (*i.e.* high growth or carbon use efficiency) or excreted byproducts that assist  
52  
53 875 in soil aggregation (*e.g.* high LPS production) (Trivedi *et al.*, 2013). Conversely, soils  
54  
55 876 dominated by copiotrophic Proteobacteria Families will be systems primarily  
56  
57 877 dependent on labile di- and monosaccharides that demonstrate low carbon use  
58  
59 878 efficiency.

60  
61 879 Streptomycetaceae, Mycobacteriaceae and Frankiaceae in Clade IV shared  
62  
63 880 enrichment of several Clade I copiotroph traits. As mentioned above in Section 4.5,  
64  
65 881 these Actinobacteria invest carbon and energy into complex filamentous growth and  
66  
67 882 developmental cycles. They demonstrate classic copiotrophic responses to nutrient  
68  
69 883 addition (Goldfarb *et al.*, 2011, Leff *et al.*, 2015) and their enriched ABC transporters,

1  
2  
3 884 pyruvate, glyoxylate, benzoate and fatty acid metabolism all likely contribute to  
4  
5 885 generating energy for complex lifecycles. Simultaneously, their enrichment of Clade III  
6  
7 886 oligotroph traits for CAZy metabolism in addition to many traits for producing and  
8  
9 887 resisting antimicrobials indicate a unique niche for these Actinobacteria that does not  
10  
11 888 necessarily fall within the classical copiotroph-oligotroph framework.

12 889 Clade V differed markedly from all other clades and mostly consisted of 'specialist'  
13  
14 890 metabolic functional groups involved in photosynthesis, ammonia oxidation,  
15  
16 891 methanogenesis, lactic acid production and other fermentation. Similar to Clade III,  
17  
18 892 these taxa would also be expected to have relatively low metabolic rates due to  
19  
20 893 depletion of copiotroph traits associated with rapid metabolism and energy generation.  
21  
22 894 Unlike Clade III, these taxa seemed to lack consistent mechanisms for stress  
23  
24 895 tolerance. Thus, while certain taxa did invest in traits for rapid metabolic rate (Clades I  
25  
26 896 and IV) and others primarily in stress tolerance (Clades III and IV), some taxa lacked  
27  
28 897 these approaches altogether and pursued entirely distinct niches (Clade V). As an  
29  
30 898 average of 47% of traits within each genome were Uncharacterised, Clade V is an  
31  
32 899 over-simplification and that if novel, currently uncharacterised proteins and the traits  
33  
34 900 they fulfil were incorporated into hierarchical clustering, this clade would separate more  
35  
36 901 meaningfully.

37 902 If one were to consider taxa within the one-dimensional copiotroph-oligotroph  
38  
39 903 spectrum, Clade I would represent one extreme, Clades III and V another, with Clades  
40  
41 904 II and IV falling in between. Figure 5a is a conceptual diagram where these Clades  
42  
43 905 have been placed on a singular axis of 'resource investment', with the niche space of  
44  
45 906 Clades enriched in traits associated with rapid growth (e.g. glutathione) toward the  
46  
47 907 'copiotrophy' pole while Clades enriched in stress tolerance traits (e.g. LPS production)  
48  
49 908 are placed toward the 'oligotrophy' pole. However, this approach overlooks the diverse  
50  
51 909 functional potentials (and distinct niches) for carbon and energy metabolism  
52  
53 910 associated with the various Clades. Furthermore, the large overlaps in niche space  
54  
55 911 between Clades would suggest taxa from each group could not co-exist if 'resource  
56  
57 912 investment' was the only important consideration (Gause, 1932, Hutchinson, 1957,  
58  
59 913 Leibold, 1995). A more meaningful perspective would be to consider the additional role  
60  
61 914 of 'resource acquisition' that incorporates multiple axes for the life strategies identified  
62  
63 915 via hierarchical clustering of traits. Figure 5b is a conceptual diagram of niche space  
64  
65 916 where clades have been further separated along additional dimensions based on their  
66  
67 917 enrichment of traits involved in competition, degradation or specialised metabolic

1  
2  
3 918 pathways. The BRITE categories listed on the various axes are chosen to be useful  
4  
5 919 markers in predicting the niche of a taxon. This expanded niche space would suggest  
6  
7 920 taxa in Clade I are well equipped for nutrient acquisition (primarily, but not limited to,  
8  
9 921 di- and monosaccharides), rapid growth and oxidative stress regulation as  
10  
11 922 'copiotrophic competitors'. Clade II, which may not be capable of competing directly  
12  
13 923 with Clade I for carbon and energy, may thus occupy the niche space of specialist  
14  
15 924 Proteobacterial methylotrophs, methane and ammonia oxidisers so as to be  
16  
17 925 'copiotrophic metabolic specialists'. The strategy of Clade III would be to decompose  
18  
19 926 plant material via diverse CAZy and possess a variety of environmental stress  
20  
21 927 tolerance traits as 'oligotrophic degraders'. Clade IV, which include for example  
22  
23 928 Actinobacteria that share traits for competition, degradation and oligotrophy, could thus  
24  
25 929 occupy niche space between Clades I and III. Finally, the strategy of Clade V would  
26  
27 930 be to fill highly specialised, unrelated metabolic niches reliant on completely distinct  
28  
29 931 carbon sources to other taxa. The large space conceptualised for Clade V in Figure  
30  
31 932 5b, which does not imply either copiotrophic or oligotrophic resource investment, is an  
32  
33 933 oversimplification and with improved understanding of traits in these taxa it may be  
34  
35 934 possible to fracture and further separate them into more detailed groups. This would  
36  
37 935 prove particularly beneficial for the poorly characterised Archaea.

34 936 Moving beyond a one-dimensional *r-K* spectrum to accommodate additional trait-  
35  
36 937 driven life strategies has been proposed in plant ecology (Grime, 1977). Specifically,  
37  
38 938 Grime argued that plant taxa fall within a multi-dimensional space defined by extremes  
39  
40 939 on three axes: 'competitors' that acquire nutrients, light, water etc. more effectively  
41  
42 940 than neighbouring taxa in the same environment, 'stress tolerators' that are long-lived,  
43  
44 941 slow growing taxa that resist desiccation, alkaline soils etc., and 'ruderals' that have  
45  
46 942 very brief lifecycles between periods of disturbance and invest in environmentally  
47  
48 943 hardy seeds. Despite these varied strategies for resource investment, plants are  
49  
50 944 unified in that photosynthesis is their primary form of acquiring carbon and energy. The  
51  
52 945 diversity of microbial strategies for acquiring carbon and energy enables them to  
53  
54 946 explore a greater range of potential niche space, and in addition to growth traits that  
55  
56 947 allow for a relatively more copiotrophic or oligotrophic investment of those resources,  
57  
58 948 likely contributes to the high diversity of co-existing taxa observed in soil microbial  
59  
60 949 communities.

58 950 However, to truly unravel differentiated niches and general microbial life strategies,  
59  
60 951 two limitations must be overcome. Firstly, a better understanding of the many

1  
2  
3 952 'Uncharacterised' traits in environmental isolates is required. For example, the recent  
4  
5 953 large-scale MAG study by Nayfach *et al.*, (2021) identified 5.8 million protein clusters  
6  
7 954 (traits), of which over 75% could not be annotated meaningfully by current protein  
8  
9 955 databases. Secondly, robust trait-based analyses down to the finer scale of distinct  
10  
11 956 genomes will likely be necessary to consider how individual taxonomic members of a  
12  
13 957 community have either differentiated in order to co-exist or are in the throes of  
14  
15 958 competition that will ultimately exclude one of the competitors.

15 959

16 960

## 18 961 5. Conclusion

20 962 In a collection of 175 terrestrial prokaryotes that possess 220 664 traits shared  
21  
22 963 between at least two taxa, concepts in niche differentiation were explored. Non-random  
23  
24 964 trait distributions were shown as preferential clustering of related taxa within most  
25  
26 965 Phyla with a general trend of highest similarity at the level of Family. This strongly  
27  
28 966 supported ecological coherence of shared traits within close relatives. Random Forest  
29  
30 967 models successfully identified BRITE 3 categories that best explained differing traits  
31  
32 968 between taxonomic groups. These traits were involved in a wide range of biological  
33  
34 969 functions, including core physiological traits used historically to categorise taxa. Many  
35  
36 970 traits were also involved in functions often associated with copiotrophs and oligotrophs,  
37  
38 971 namely metabolism, nutrient acquisition and environmental stress tolerance.  
39  
40 972 Hierarchical clustering of differential traits formed five distinct clusters, with Clade I  
41  
42 973 representing the classical copiotrophic niche, Clades III and V as oligotrophic, and  
43  
44 974 Clades II and IV in between. A more refined perspective would be to consider each  
45  
46 975 Clade as its own life strategy in a niche space that considers both resource investment  
47  
48 976 and acquisition simultaneously; for example, the strategy of Clade I is to invest in  
49  
50 977 competition and rapid growth, while Clade V pursue highly distinct, specialised  
51  
52 978 metabolic functions. The trait-based analyses here were effective in identifying general  
53  
54 979 trends in potential function of terrestrial microbial taxa at the Phylum and Family level.  
55  
56 980 Further investigation will be necessary to identify traits that give rise to niche  
57  
58 981 differentiation at lower taxonomic ranks and, ultimately, the importance of this for  
59  
60 982 ecosystem processes of interest.

56 983

58 984

## 60 985 Acknowledgements



1  
2  
3 986 The authors wish to thank the Australian Government Endeavour research program  
4  
5 987 ID: 6123\_2017 for funding D.R. Finn. Funding support for B. Bergk-Pinto came from  
6  
7 988 the European Union's Horizon 2020 research and innovation program under the Marie  
8  
9 989 Sklodowska-Curie ITN 675546-MicroArctic. The authors also wish to thank the  
10  
11 990 microbiologists involved in the isolation and genome sequencing of the taxa studied  
12  
13 991 here; it is hoped the majority of those involved are cited within the Supplementary  
14  
15 992 Material. Without this diligent culture-dependent work by the scientific community, the  
16  
17 993 theoretical ecology performed here would not be possible.

18 994

## 19 995 Conflict of Interest

20 996 The authors declare no conflict of interest regarding this study or its outcomes.

21 997

## 22 998 References

23  
24 999 Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic Local Alignment Search  
25  
26 1000 Tool. *Journal of Molecular Biology* **215**: 403-410.27 1001 Amon J, Titgemeyer F & Burkovski A (2010) Common patterns - unique features: nitrogen  
28 1002 metabolism and regulation in Gram-positive bacteria. *FEMS Microbiology Reviews* **34**: 588-  
29 1003 605.30 1004 Barbéran A, Bates ST, Casamayor EO & Fierer N (2012) Using network analysis to explore co-  
31 1005 occurrence patterns in soil microbial communities. *The ISME Journal* **6**: 343-351.32 1006 Bastida F, Torres IF, Moreno JL, *et al.* (2016) The active microbial diversity drives ecosystem  
33 1007 multifunctionality and is physiologically related to carbon availability in Mediterranean semi-  
34 1008 arid soils. *Molecular Ecology* **25**: 4660-4673.35 1009 Bentley SD, Chater KF, Cerdeno-Tarraga AM, *et al.* (2002) Complete genome sequence of the  
36 1010 model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**: 141-147.37 1011 Blagodatsky SA & Richter O (1998) Microbial growth in soil and nitrogen turnover: A  
38 1012 theoretical model considering the activity state of microorganisms. *Soil Biology & Biochemistry*  
39 1013 **30**: 1743-1755.40 1014 Bouskill NJ, Tang J, Riley WJ & Brodie EL (2012) Trait-based representation of biological  
41 1015 nitrification: model development, testing, and predicted community composition. *Frontiers in*  
42 1016 *Microbiology* **3**: 364.43 1017 Brochier-Armanet C, Talla E & Gribaldo S (2009) The Multiple Evolutionary Histories of  
44 1018 Dioxygen Reductases: Implications for the Origin and Evolution of Aerobic Respiration.  
45 1019 *Molecular Biology and Evolution* **26**: 285-297.46 1020 Brown JH, Gillooly JF, Allen AP, Savage VM & West GB (2004) Toward a metabolic theory of  
47 1021 ecology. *Ecology* **85**: 1771-1789.48 1022 Burris RH & Roberts GP (1993) Biological nitrogen fixation. *Annual Reviews Nutrition* **13**: 317-  
49 1023 335.50 1024 Cania B, Vestergaard G, Krauss M, Fliessbach A, Schloter M & Schulz S (2019) A long-term field  
51 1025 experiment demonstrates the influence of tillage on the bacterial potential to produce soil  
52 1026 structure-stabilizing agents such as exopolysaccharides and lipopolysaccharides.  
53 1027 *Environmental Microbiome* **14**: 1-14.

- 1  
2  
3 1028 Cary SC, McDonald IR, Barrett JE & Cowan DA (2010) On the rocks: the microbiology of  
4 1029 Antarctic Dry Valley soils. *Nature Reviews Microbiology* **8**: 129-138.
- 5 1030 Caulier S, Nannan C, Gillis A, Licciardi F, Bragard C & Mahillon J (2019) Overview of the  
6 1031 antimicrobial compounds produced by members of the *Bacillus subtilis* group. *Frontiers in*  
7 1032 *Microbiology* **10**: 302.
- 8 1033 Christie PJ & Vogel JP (2000) Bacterial type IV secretion: conjugation systems adapted to  
9 1034 deliver effector molecules to host cells. *Trends in Microbiology* **8**: 354-360.
- 10 1035 Conrad R (2009) The global methane cycle: recent advances in understanding the microbial  
11 1036 processes involved. *Environmental Microbiology Reports* **1**: 285-292.
- 12 1037 Crits-Christoph A, Robinson CK, Barnum T, Fricke WF, Davila AF, Jedynek B, McKay CP &  
13 1038 DiRuggiero J (2013) Colonization patterns of soil microbial communities in the Atacama  
14 1039 Desert. *Microbiome* **1**: 28.
- 15 1040 de los Santos PE, Palmer M, Chávez-Ramirez B, Beukes C, Steenkamp ET, Briscoe L, Khan N,  
16 1041 Maluk M *et al.*, (2018) Whole genome analyses suggests that *Burkholderia sensu lato* contains  
17 1042 two additional novel genera (*Mycetohabitans* gen. nov., and *Trinickia* gen. nov.): implications  
18 1043 for the evolution of diazotrophy and nodulation in the Burkholderiaceae. *Genes* **9**:  
19 1044 doi:10.3390/genes9080389.
- 20 1045 de Mendiburu F (2014) *Agricolae*: statistical procedures for agricultural research.  
21 1046 <http://cran.r-project.org/web/packages/agricolae/index.html>.
- 22 1047 Dunfield PF (2007) The Soil Methane Sink. *Greenhouse Gas Sinks* 152-170.
- 23 1048 Embley TM & Martin W (2006) Eukaryotic evolution, changes and challenges. *Nature Reviews*  
24 1049 **440**: 623-630.
- 25 1050 Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high  
26 1051 throughput. *Nucleic Acid Research* **32**: 1792-1797.
- 27 1052 Enright AJ, Van Dongen S & Ouzounis CA (2002) An efficient algorithm for large-scale detection  
28 1053 of protein families. *Nucleic Acids Research* **30**: 1575-1584.
- 29 1054 Fierer N, Bradford MA & Jackson RB (2007) Toward an ecological classification of soil bacteria.  
30 1055 *Ecology* **88**: 1354-1364.
- 31 1056 Fierer N, Lauber CL, Ramirez KS, Zaneveld J, Bradford MA & Knight R (2012) Comparative  
32 1057 metagenomic, phylogenetic and physiological analyses of soil microbial communities across  
33 1058 nitrogen gradients. *The ISME Journal* **6**: 1007-1017.
- 34 1059 Finn DR, Yu J, Ilhan ZE, Fernandes VMC, Penton CR, Krajmalnik-Brown R, Garcia-Pichel F &  
35 1060 Vogel TM (2020a) MicroNiche: an R package for assessing microbial niche breadth and overlap  
36 1061 from amplicon sequencing data. *Fems Microbiology Ecology* **96**: fiae131.
- 37 1062 Finn DR, Ziv-el M, van Haren J, Park JG, del Aguila-Pasquel J, Urquiza-Munoz JD & Cadillo-  
38 1063 Quiroz H (2020b) Methanogens and methanotrophs show nutrient-dependent community  
39 1064 assemblage patterns across tropical peatlands of the Pastaza-Maranon Basin, Peruvian  
40 1065 Amazonia. *Frontiers in Microbiology* **11**: 746.
- 41 1066 Flemming HC & Wingender J (2010) The biofilm matrix. *Nature Reviews Microbiology* **8**: 623-  
42 1067 633.
- 43 1068 Freilich S, Zarecki R, Eilam O, Segal ES, Henry CS, Kupiec M, Gophna U, Sharan R & Ruppin E  
44 1069 (2011) Competitive and cooperative metabolic interactions in bacterial communities. *Nature*  
45 1070 *Communications* **2**.
- 46 1071 Galperin MY (2005) A census of membrane-bound and intracellular signal transduction  
47 1072 proteins in bacteria: bacterial IQ, extroverts and introverts. *BMC Microbiology* **5**:  
48 1073 doi:10.1186/1471-2180-1185-1135.
- 49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1074 Gao B, Paramanathan R & Gupta RS (2006) Signature proteins that are distinctive  
1075 characteristics of Actinobacteria and their subgroups. *Antonie Van Leeuwenhoek International*  
1076 *Journal of General and Molecular Microbiology* **90**: 69-91.
- 1077 Garcia-Pichel F, Lopez-Cortes A & Nübel U (2001) Phylogenetic and morphological diversity of  
1078 Cyanobacteria in soil desert crusts from the Colorado Plateau. *Applied and Environmental*  
1079 *Microbiology* **67**: 1902-1910.
- 1080 Gause GF (1932) Experimental studies on the struggle for existence I Mixed population of two  
1081 species of yeast. *Journal of Experimental Biology* **9**: 389-402.
- 1082 Gleason HA (1926) The individualistic concept of the plant association. *Bull Torrey Botany Club*  
1083 **53**: 7-26.
- 1084 Goldberg AL (2003) Protein degradation and protection against misfolded or damaged  
1085 proteins. *Nature* **426**: 895-890.
- 1086 Goldfarb KC, Karaoz U, Hanson CA, Santee CA, Bradford MA, Treseder KK, Wallenstein MD &  
1087 Brodie EL (2011) Differential growth responses of soil bacterial taxa to carbon substrates of  
1088 varying chemical recalcitrance. *Frontiers in Microbiology* **2**.
- 1089 Grime JP (1977) Evidence for the existence of three primary strategies in plants and its  
1090 relevance to ecological and evolutionary theory. *The American Naturalist* **111**: 1169-1194.
- 1091 Grime JP (1979) Plant strategies and vegetation processes. *Wiley, Chichester, UK*.
- 1092 He M, Zhang J, Shen L, *et al.* (2019) High-throughput sequencing analysis of microbial  
1093 community diversity in response to indica and japonica bar-transgenic rice paddy soils. *PLOS*  
1094 *One* **14**: e0222191.
- 1095 Higgins CF (2001) ABC transporters: physiology, structure and mechanism - an overview.  
1096 *Research in Microbiology* **152**: 205-210.
- 1097 Ho A, Paolo Di Lonardo D & Bodelier PL (2017) Revisiting life strategy concepts in  
1098 environmental microbial ecology. *FEMS Microbiology Ecology* **93**: 1-14.
- 1099 Ho A, Kerckhof FM, Luke C, Reim A, Krause S, Boon N & Bodelier PL (2013) Conceptualizing  
1100 functional traits and ecological characteristics of methane-oxidizing bacteria as life strategies.  
1101 *Environmental Microbiology Reports*  
1102 **5**: 335-345.
- 1103 Hodgson DA (2000) Primary metabolism and its control in streptomycetes: A most unusual  
1104 group of bacteria. *Advances in Microbial Physiology, Vol 42, Vol. 42* (Poole RK, ed.) p. 47-  
1105 238.
- 1106 Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR, Williams KH, Tringe SG  
1107 & Banfield JF (2013) Community genomic analyses constrain the distribution of metabolic  
1108 traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome*  
1109 **1**.
- 1110 Hutchinson GL (1957) Concluding remarks *Cold Spring Harbor Symposia on Quantitative*  
1111 *Biology* **22**: 415-427.
- 1112 Itakura M, Saeki K, Omori H, Yokoyama T, Kaneko T, Tabata S, Ohwada T, Tajima S & . ea (2009)  
1113 Genomic comparison of Bradyrhizobium japonicum strains with different symbiotic nitrogen-  
1114 fixing capabilities and other Bradyrhizobiaceae members. *The ISME journal* **3**: 326-339.
- 1115 Jangid K, Williams MA, Franzluebbers A, Schmidt TM, Coleman DC & Whitman WB (2011)  
1116 Land-use history has a stronger impact on soil microbial community composition than  
1117 aboveground vegetation and soil properties. *Soil Biol Biochem* **43**: 2184-2193.
- 1118 Jarrell KF, Albers SV (2012) The archaellum: an old motility structure with a new name. *Trends*  
1119 *Microbiol* **20**: 307-312.

- 1  
2  
3 1120 Kaiser C, Franklin O, Richter A & Dieckmann U (2015) Social dynamics within decomposer  
4 1121 communities lead to nitrogen retention and organic matter build-up in soils. *Nature*  
5 1122 *Communications* **6**.
- 7 1123 Kanehisa M, Sato Y, Kawashima M, Furumichi M & Tanabe M (2016) KEGG as a reference  
8 1124 resource for gene and protein annotation. *Nucleic Acids Research* **44**: D457-D462.
- 9 1125 Kearney M, Simpson SJ, Raubenheimer D & Helmuth B (2010) Modelling the ecological niche  
10 1126 from functional traits. *Philosophical Transactions of the Royal Society B-Biological Sciences*  
11 1127 **365**: 3469-3483.
- 13 1128 Keddy PA (1992) Assembly and response rules - two goals for predictive community ecology.  
14 1129 *Journal of Vegetation Science* **3**: 157-164.
- 15 1130 Klotz MG & Norton JB (1998) Multiple copies of ammonia monooxygenase (amo) operons  
16 1131 have evolved under biased AT/GC mutational pressure in ammonia-oxidizing autotrophic  
17 1132 bacteria. *Fems Microbiology Letters* **168**: 303-311.
- 19 1133 Koch AL (2001) Oligotrophs versus copiotrophs. *BioEssays* **23**: 657-661.
- 20 1134 Kögel-Knabner I (2002) The macromolecular organic composition of plant and microbial  
21 1135 residues as inputs to soil organic matter. *Soil Biol Biochem* **34**: 139-162.
- 23 1136 Kuznetsov SI, Dubinina GA & Lapteva NA (1979) Biology of oligotrophic bacteria. *Annual*  
24 1137 *Review of Microbiology* **33**: 377-387.
- 25 1138 Lauro FM, McDougald D, Thomas T, *et al.* (2009) The genomic basis of trophic strategy in  
26 1139 marine bacteria. *Proceedings of the National Academy of Sciences of the United States of*  
27 1140 *America* **106**: 15527-15533.
- 29 1141 Leff JW, Jones SE, Prober SM, *et al.* (2015) Consistent responses of soil microbial communities  
30 1142 to elevated nutrient inputs in grasslands across the globe. *Proceedings of the National*  
31 1143 *Academy of Sciences of the United States of America* **112**: 10967-10972.
- 32 1144 Leibold MA (1995) The niche concept revisited: mechanistic models and community context.  
33 1145 *Ecology* **76**: 1371-1382.
- 35 1146 Liaw A & Weiner M (2002) Classification and regression by randomForest. *R News* **2**: 18-22.
- 36 1147 Madigan M, Cox SS & Stegeman RA (1984) Nitrogen fixation and nitrogenase activities in  
37 1148 members of the Family Rhodospirillaceae. *Journal of Bacteriology* **157**: 73-78.
- 38 1149 Madin JS, Nielsen DA, Brbic M, Corkrey R, Danko D, Edwards K *et al.*, (2020) A synthesis of  
39 1150 bacterial and archaeal phenotypic trait data. *Nature Scientific Data* **7**: 170.
- 41 1151 McBride MJ (2001) Bacterial gliding motility: mechanisms for cell movement over surfaces.  
42 1152 *Annual Review of Microbiology* **55**: 49-75.
- 43 1153 McGill BJ, Enquist BJ, Weiher E & Westoby M (2006) Rebuilding community ecology from  
44 1154 functional traits. *Trends in Ecology & Evolution* **21**: 178-185.
- 46 1155 Möller B, Oßmer R, Howard BH, Gottschalk G & Hippe H (1984) Sporomusa, a new genus of  
47 1156 Gram-negative anaerobic bacteria including *Sporomusa sphaeroides* spec. nov. and  
48 1157 *Sporomusa ovata* spec. nov. *Archives of Microbiology* **139**: 388-396.
- 49 1158 Nayfach S, Roux S, Seshadri R, Udwaray D, Varghese N, Schulz F *et al.*, (2021) A genomic catalog  
50 1159 of Earth's microbiomes. *Nature Biotechnology* **39**: 499-509.
- 52 1160 Oksanen J, Guillaume Blanchet F, Kindt R, *et al.* (2013) Vegan: Community Ecology Package. R  
53 1161 package version 2.0-10. <http://CRAN.R-project.org/package=vegan>.
- 54 1162 Op den Camp HJM, Islam T, Stott MB, Harhangi HR, Hynes A, Schouten S, Jetten MSM,  
55 1163 Birkeland N-K, Pol A & Dunfield PF (2009) Environmental, genomic and taxonomic perspectives  
56 1164 on methanotrophic Verrucomicrobia. *Environmental Microbiology Reports* **1**: 293-306.
- 58 1165 Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA & Hugenholtz P  
59 1166 (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises  
60 1167 the tree of life. *Nature Biotechnology* **36**: 996-1004.

1

2

3 1168 Pascual-Garcia A & Bell T (2020) Community-level signatures of ecological succession in  
4 1169 natural bacterial communities. *Nature Communications* **11**: 2386.

5 1170 Philippot L, Bru D, Saby NPA, Cuhel J, Arrouays D, Simek M & Hallin S (2009) Spatial patterns  
6 1171 of bacterial taxa in nature reflect ecological traits of deep branches of the 16S rRNA bacterial  
7 1172 tree. *Environmental Microbiology* **11**: 3096-3104.

8 1173 Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, Whitman WB & Hallin S (2010)  
9 1174 The ecological coherence of high bacterial taxonomic ranks. *Nature Reviews Microbiology* **8**:  
10 1175 523-529.

11 1176 Prosser JI (2012) Ecosystem processes and interactions in a morass of diversity. *Fems*  
12 1177 *Microbiology Ecology* **81**: 507-519.

13 1178 R Core Team (2013) R: A language and environment for statistical computing. *R Foundation*  
14 1179 *for statistical computing, Vienna, Austria.*

15 1180 Roller BRK & Schmidt TM (2015) The physiology and ecological implications of efficient  
16 1181 growth. *The ISME journal* **9**: 1481-1487.

17 1182 Sancar A (1996) DNA excision repair. *Annual Review of Biochemistry* **65**: 43-81.

18 1183 Sangwan P, Chen XL, Hugenholtz P & Janssen PH (2004) *Chthoniobacter flavus* gen. nov., sp  
19 1184 nov., the first pure-culture representative of subdivision two, Spartobacteria classis nov., of  
20 1185 the phylum Verrucomicrobia. *Applied and Environmental Microbiology* **70**: 5875-5881.

21 1186 Schliep KP, Potts AJ, Morrison DA & Grimm WA (2017) Intertwining phylogenetic trees and  
22 1187 networks. *Methods in Ecology and Evolution* **8**: 1212-1220.

23 1188 Schneiker S, Perlova O, Kaiser O, Gerth K, Alici A, Altmeyer MO, Bartels D, Bekel T *et al.*, (2007)  
24 1189 Complete genome of the myxobacterium *Sorangium cellulosum*. *Nature Biotechnology* **25**:  
25 1190 1281-1290.

26 1191 Semenov AM (1991) Physiological bases of oligotrophy of microorganisms and the concept of  
27 1192 microbial community. *Microbial Ecology* **22**: 239-247.

28 1193 Shimada T, Tanaka K & Ishihama A (2017) The whole set of the constitutive promoters  
29 1194 recognized by four minor sigma subunits of *Escherichia coli* RNA polymerase. *Plos One* **12**:  
30 1195 e0179181.

31 1196 Smirnova GV & Oktyabrsky ON (2005) Glutathione in Bacteria. *Biochemistry* **70**: 1199-1211.

32 1197 Spang A, Poehlein A, Offre P, Zumbrägel S, Haider S, Rychlik N, Nowka B *et al.*, (2012) The  
33 1198 genome of the ammonia-oxidizing Candidatus *Nitrososphaera gargensis*: insights into  
34 1199 metabolic versatility and environmental adaptations. *Environmental Microbiology* **14**(12):  
35 1200 3122-3145.

36 1201 Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R,  
37 1202 Schleper C, Guy L & Ettema TJG (2015) Complex archaea that bridge the gap between  
38 1203 prokaryotes and eukaryotes. *Nature* **521**: 173-185.

39 1204 Stackebrandt E, Pohla H, Kroppenstedt RM, Hippe H & Woese CR (1985) 16S rRNA analysis of  
40 1205 *Sporomusa*, *Selenomonas* and *Megasphaera*: on the phylogenetic origin of Gram-positive  
41 1206 Eubacteria. *Archives of Microbiology* **143**: 270-276.

42 1207 Tamas I, Smirnova AV, He Z & Dunfield PF (2014) The (d)evolution of methanotrophy in the  
43 1208 Beijerinckiaceae - a comparative genomics analysis. *The ISME journal* **8**: 369-382.

44 1209 Terzenbach DP & Blaut M (1994) Transformation of tetrachloroethylene to trichloroethylene  
45 1210 by homoacetogenic bacteria. *Fems Microbiology Letters* **123**: 213-218.

46 1211 Trivedi P, Anderson IC & Singh BK (2013) Microbial modulators of soil carbon storage:  
47 1212 integrating genomic and metabolic knowledge for global prediction. *Trends in Microbiology*  
48 1213 **21**: 641-651.

49 1214 van Rossum T, Ferretti P, Maistrenko OM & Bork P (2020) Diversity within species: interpreting  
50 1215 strains in microbiomes. *Nature Reviews Microbiology* **18**: 491-506.

- 1  
2  
3 1216 Vieira-Silva S & Rocha EPC (2010) The systemic imprint of growth and its uses in ecological  
4 1217 (meta)genomics. *Plos Genetics* **6**: e1000808.  
5 1218 Warnes GR, Bolker B, Bonebakker L, *et al.* (2019) gplots: various R programming tools for  
6 1219 plotting data. <https://cran.r-project.org/web/packages/gplots/index.html>.  
7 1220 Weins JJ (1998) Testing phylogenetic methods with tree congruence: phylogenetic analysis of  
8 1221 polymorphic morphological characters in Phrynosomatid lizards. *Systematic Biology* **47**: 427-  
9 1222 444.  
10 1223 Wessen E, Hallin S & Philippot L (2010) Differential responses of bacterial and archaeal groups  
11 1224 at high taxonomical ranks to soil management. *Soil Biol Biochem* **42**: 1759-1765.  
12 1225 Wickham H (2007) Reshaping data with the reshape package. *Journal of Statistical Software*  
13 1226 **21**: 1-20.  
14 1227 Young J & Holland IB (1999) ABC transporters: bacterial exporters-revisited five years on.  
15 1228 *Biochimica Et Biophysica Acta-Biomembranes* **1461**: 177-200.  
16 1229 Zhu CS, Delmont TO, Vogel TM & Bromberg Y (2015) Functional Basis of Microorganism  
17 1230 Classification. *Plos Computational Biology* **11**.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46

For Peer Review



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Peer Review

1  
2  
3 1 Figure Legends:  
4

5 2 Figure 1. Box and whisker plots of the raw counts of proteins per genome beside the  
6 3 counts of shared functional traits per genome derived from MCL. Taxa  $n = 5$  per  
7 4 Family.  
8  
9

10 5  
11 6 Figure 2. a) Unweighted pair group method with arithmetic mean (UPGMA)  
12 7 dendrogram comparing similarity in composition of 220 664 traits across the 175  
13 8 terrestrial microbial taxa. The Phylum of each taxon is highlighted. The scale bar units  
14 9 represent Bray-Curtis dissimilarity. b) Comparisons of ecological coherence ( $C$ )  
15 10 between taxa belonging to the same Super Group, Phylum and Family.  $C$ , which varies  
16 11 between 0 and 1, was measured from branch lengths in a). Values of  $C$  approaching  
17 12 1 indicate an ecologically coherent group with a similar composition of traits. Super  
18 13 groups were as follows: Proteobacteria, all Gamma-, Alpha-, Beta- and  
19 14 Deltaproteobacteria; Terrabacteria, all Actinobacteria, Firmicutes, Chloroflexi and  
20 15 Cyanobacteria; FCB were Bacteroidetes; Acidobacteria, all Acidobacteria; PVC,  
21 16 Verrucomicrobia and Planctomycetes; TACK were Thaumarchaeota; Euryarchaeota  
22 17 were Euryarchaeota.  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35

36 20 Figure 3. Random Forest identified traits that best explain separation of taxa groups  
37 21 at Phylum and Family taxonomic rank. Mean Decrease in Accuracy (MDA) is a  
38 22 measure of the average increase in classification error during permutation of trees if  
39 23 the trait is not included in the model. For example, the Family Random Forest was  
40 24 most accurate when ABC transporters were included in the model.  
41  
42  
43  
44  
45  
46  
47

48 27 Figure 4. Hierarchical clustering of Families based on 60 traits selected from Random  
49 28 Forest and previously identified copiotroph-oligotroph traits. Families are clustered  
50 29 along the  $y$  axis, and specific co-occurring traits are clustered on the  $x$  axis. Five  
51 30 general Clades were identified based on Family clustering. Units are normalised  
52 31 variance in mean counts of traits per Family.  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 34 Figure 5: A conceptual diagram of overlapping life strategies between the five  
4 35 functional clades identified in this study. a) the classical copiotroph-oligotroph  
5 36 dichotomy whereby the five clades are placed on a singular axis of 'resource  
6 37 investment', where growth strategies are either relatively more targeted toward rapid  
7 38 growth (copiotroph) or toward stress tolerance (oligotroph). BRITE categories are  
8 39 given as examples of functional traits contributing to either axis pole. The space  
9 40 occupied by clades along this axis is dependent on enrichment or depletion of these  
10 41 traits. All clades exhibit a great deal of overlap and certain clades, such as IV  
11 42 (including, for example, Actinobacteria), are hard to identify as either copiotroph or  
12 43 oligotroph. b) a multidimensional concept where three axes for 'resource acquisition'  
13 44 are added, further separating taxa as either 'competitors', 'degraders' or 'metabolic  
14 45 specialists'. Again, BRITE categories are provided as examples of traits contributing  
15 46 to the additional axes and space occupied by clades is dependent on enrichment or  
16 47 depletion of these traits. These extra dimensions would suggest the niche space of  
17 48 clade IV exists between 'copiotrophic competitors' (I) and 'oligotrophic degraders' (III),  
18 49 potentially allowing IV to co-exist alongside both.  
19 50

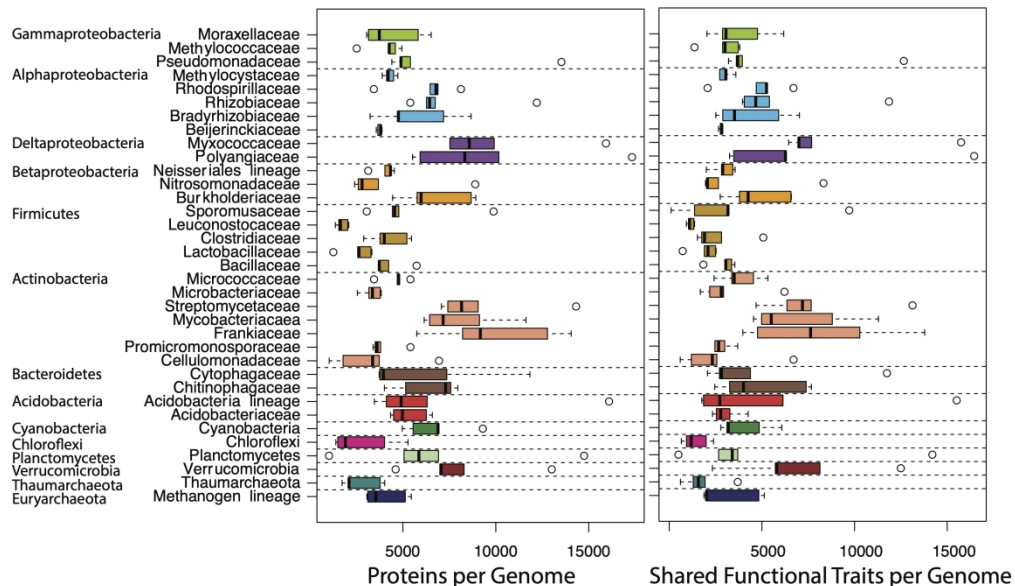


Figure 1

268x157mm (300 x 300 DPI)

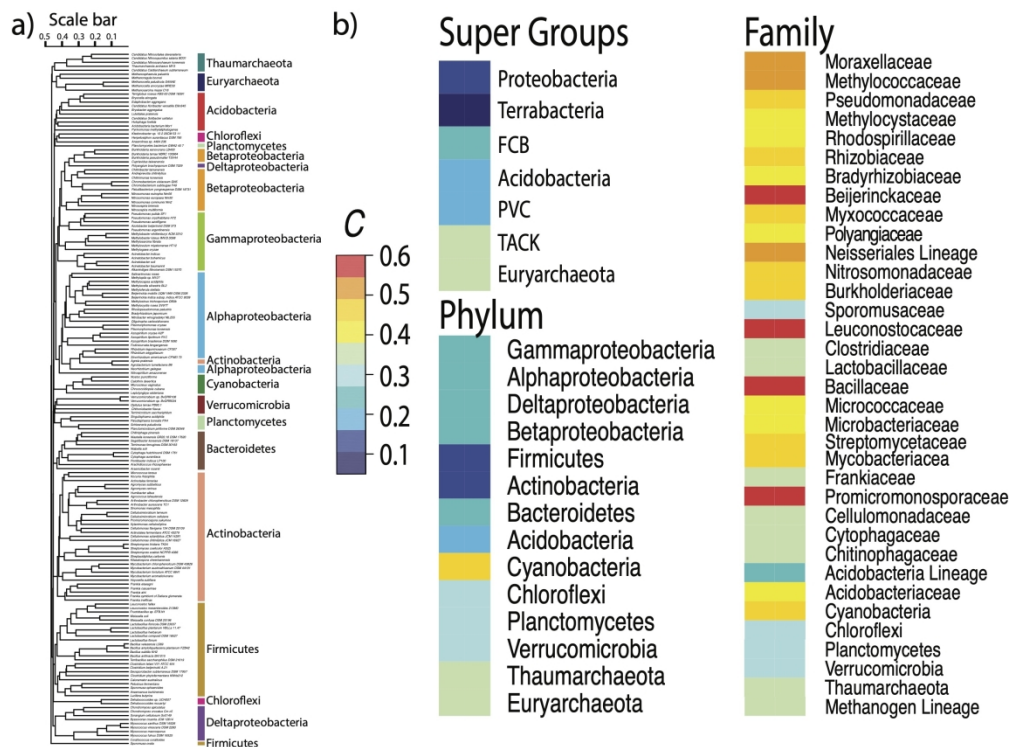


Figure 2

232x172mm (300 x 300 DPI)



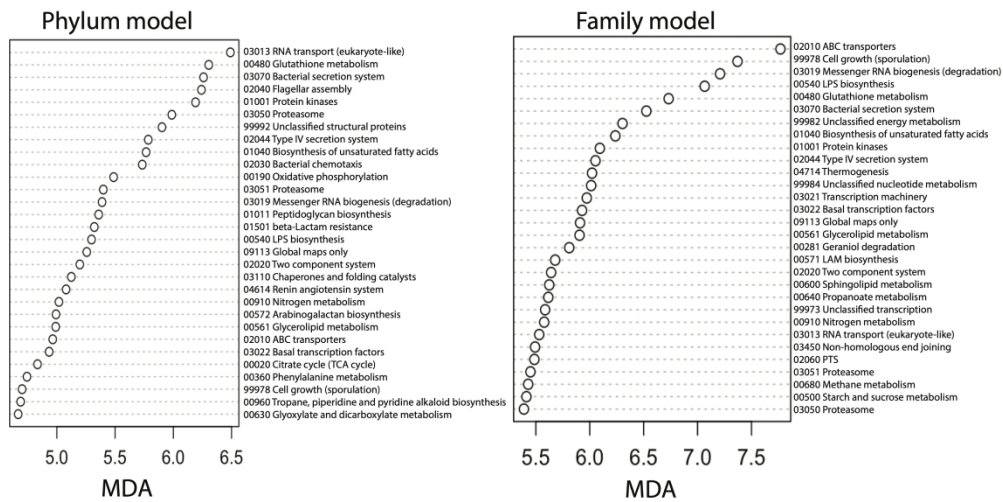


Figure 3

338x170mm (300 x 300 DPI)



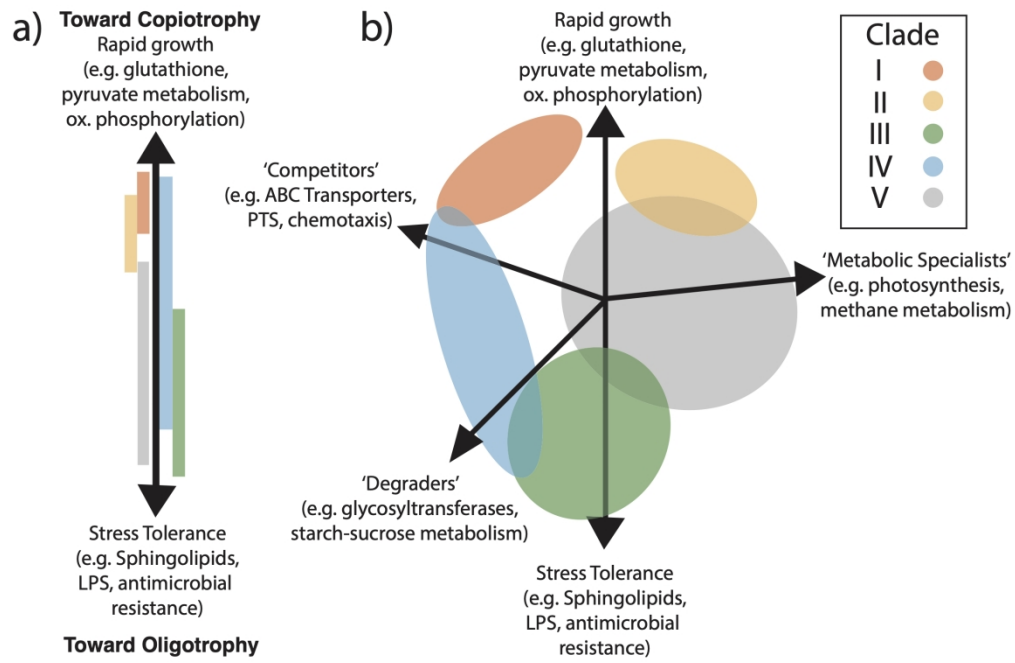


Figure 5

203x134mm (300 x 300 DPI)

Table 1: Summary of traits significantly enriched and depleted in the Family Random Forest model.

BRITE 1	BRITE 3	Significantly enriched	Significantly depleted
Metabolism	00540 LPS biosynthesis	Pseudomonadaceae, Burkholderiaceae, Chitinophagaceae, Acidobacteria Lineage, Verrucomicrobia	Leuconostocaceae, Clostridiaceae, Lactobacillaceae, Bacillaceae, all Actinobacteria, Chloroflexi, Thaumarchaeota, Methanogen Lineage
	01040 Biosynthesis of unsaturated fatty acids	Mycobacteriaceae, Frankiaceae, Pseudomonadaceae	Leuconostocaceae, Clostridiaceae, Lactobacillaceae, Bacillaceae, Cyanobacteria, Chloroflexi, Thaumarchaeota, Methanogen lineage
	00561 Glycerolipid metabolism	Myxococcaceae, Mycobacteriaceae, Frankiaceae	Thaumarchaeota, Methanogen Lineage
	00571 LAM biosynthesis	Streptomycetaceae, Mycobacteriaceae, Frankiaceae	All other Families
	00600 Sphingolipid metabolism	Myxococcaceae, Chitinophagaceae, Planctomycetaceae, Verrucomicrobia	All Gammaproteobacteria, Rhodospirillaceae, Bradyrhizobiaceae, Neisseriales Lineage, Nitrosomonadaceae, Sporomusaceae, Leuconostocaceae, Lactobacillaceae, Cellulomonadaceae, Chloroflexi, Thaumarchaeota, Methanogen Lineage
	00500 Starch and sucrose metabolism	Polyangiaceae, Streptomycetaceae, Mycobacteriaceae, Frankiaceae, Verrucomicrobia	Moraxcellaceae, Thaumarchaeota, Methanogen Lineage
	04714 Thermogenesis	Methylocystaceae, Rhodospirillaceae, Bradyrhizobiaceae, Beijerinckaceae, Mycobacteriaceae	All other Families
	00640 Propanoate metabolism	Burkholderiaceae, Mycobacteriaceae, Frankiaceae	Methylococcaceae, Leuconostocaceae, Lactobacillaceae, Thaumarchaeota
	00910 Nitrogen metabolism	Rhodospirillaceae, Bradyrhizobiaceae, Burkholderiaceae, Sporomusaceae, Mycobacteriaceae	Leuconostocaceae, Lactobacillaceae, Chloroflexi
	00680 Methane metabolism	Methylococcaceae, Methylocystaceae, Methanogen Lineage	Leuconostocaceae, Lactobacillaceae
Environmental Information Processing	99982 Unclassified energy metabolism	Rhodospirillaceae, Beijerinckiaceae, Mycobacteriaceae, Frankiaceae	Leuconostocaceae, Lactobacillaceae, Microbacteriaceae, Promicromonosporaceae, Cellulomonadaceae, Chloroflexi, Planctomycetales, Thaumarchaeota
	09113 Global maps only (unclassified metabolism)	Pseudomonadaceae, Mycobacteriaceae	Methylococcaceae, Beijerinckiaceae, Nitrosomonadaceae, Leuconostocaceae, Bacillaceae, Cytophagaceae, Acidobacteriaceae, Cyanobacteria, Verrucomicrobia
	00480 Glutathione metabolism	Rhizobiaceae, Bradyrhizobiaceae, Myxococcaceae, Polyangiaceae, Burkholderiaceae	All Firmicutes, Acidobacteriaceae, Chloroflexi, Planctomycetales, Verrucomicrobia, Thaumarchaeota, Methanogen Lineage
	00281 Geraniol degradation	Moraxellaceae, Myxococcaceae, Mycobacteriaceae	Methylococcaceae, Pseudomonadaceae, Rhizobiaceae, Beijerinckiaceae, Neisseriales Lineage, Nitrosomonadaceae, all Firmicutes, Streptomycetaceae, Promicromonosporaceae, Cellulomonadaceae, Acidobacteriaceae, Cyanobacteria, Chloroflexi, Planctomycetaceae, Verrucomicrobia, Thaumarchaeota, Methanogen Lineage
	01001 Protein kinases	Myxococcaceae, Polyangiaceae, Frankiaceae	Lactobacillaceae
	02010 ABC transporters	Rhodospirillaceae, Rhizobiaceae, Burkholderiaceae	Acidobacteriaceae, Thaumarchaeota, Methanogen Lineage
	02020 Two component system	Pseudomonadaceae, Rhodospirillaceae, Bradyrhizobiaceae, Myxococcaceae	Leuconostocaceae, Lactobacillaceae, Bacillaceae, Promicromonosporaceae, Cellulomonadaceae, Chloroflexi, Thaumarchaeota
	02060 Phosphotransferase system (PTS)	Pseudomonadaceae, Rhodospirillaceae, Myxococcaceae, Neisseriales lineage, Clostridiaceae, Verrucomicrobia	Methylococcaceae, Frankiaceae, Acidobacteria Lineage, Thaumarchaeota, Methanogen Lineage

	03070 Bacterial secretion system	Polyangiaceae, Myxococcaceae, Burkholderiaceae	Leuconostocaceae, Lactobacillaceae, Bacillaceae, Streptomyetaceae, Mycobacteriaceae, Chloroflexi, Thaumarchaeota, Methanogen lineage
	02044 Type IV secretion system	Pseudomonadaceae, Myxococcaceae	Leuconostocaceae, all Bacteroidetes, Thaumarchaeota
Genetic Information Processing	03021 Transcription machinery	Myxococcaceae, Polyangiaceae, Streptomyetaceae, Mycobacteriaceae, Frankiaceae	Methylocystaceae, Rhizobiaceae, Bradyrhizobiaceae, Beijerinckiaceae, Leuconostocaceae, Lactobacillaceae, Thaumarchaeota, Methanogen lineage
	99973 Unclassified transcription factors	Pseudomonadaceae, Rhizobiaceae, Myxococcaceae, Streptomyetaceae	Chitinophagaceae, Cyanobacteria, Chloroflexi, Planctomyetaceae, Verrucomicrobia
	03022 Basal transcription factors	Thaumarchaeota, Methanogen Lineage	All other Families
	03013 RNA transport (eukaryote-like)	Myxococcaceae, Thaumarchaeota, Methanogen Lineage	All other Families
	03019 Messenger RNA biogenesis (degradation)	Myxococcaceae, Burkholderiaceae	Methylocystaceae, Rhizobiaceae, Bradyrhizobiaceae, Beijerinckiaceae, Cyanobacteria, Chloroflexi, Thaumarchaeota
	03450 Non-homologous end joining	Rhizobiaceae, Streptomyetaceae, Verrucomicrobia	Moraxellaceae, Methylococcaceae, Neisseriales Lineage, Leuconostocaceae, Clostridaceae, Lactobacillaceae, Cyanobacteria
	03050 Proteasome	Micrococcaceae, Streptomyetaceae, Mycobacteriaceae, Frankiaceae, Promicromonosporaceae, Cellulomonadaceae, Thaumarchaeota, Methanogen Lineage	All other families
	03051 Prokaryote 20S Proteasome	Micrococcaceae, Streptomyetaceae, Mycobacteriaceae, Frankiaceae	All other Families
Signalling and Cellular Processes	99978 Cell growth (sporulation)	Sporomusaceae, Clostridiaceae, Bacillaceae	All other families
Not included in pathway or BRITE	99984 Unclassified nucleotide metabolism	Leuconostocaceae, Lactobacillaceae, Micrococcaceae	Moraxellaceae, Methylococcaceae, Methylocystaceae, Rhodospirillaceae, Bradyrhizobiaceae, Beijerinckiaceae, all Deltaproteobacteria, all Betaproteobacteria, Sporomusaceae, Clostridiaceae, Streptomyetaceae, Frankiaceae, all Bacteroidetes, all Acidobacteria, Chloroflexi, Planctomyetes, Thaumarchaeota, Methanogen Lineage



## Supplementary Material

Supplementary Table 1: List of microorganisms utilised in this study.

ID	Phylum	Class	Order	Family	Genus	Species	Strain	Ecosystem Process <sup>a</sup>	Reference
Ga0074846 _101	Firmicutes	Bacilli	Bacillales	Bacillaceae	<i>Terribacillus</i>	<i>saccharophilus</i>	DSM21619	ND	(An et al 2007b)
Ga0067102 _11	Firmicutes	Bacilli	Bacillales	Bacillaceae	<i>Bacillus</i>	<i>amyloliquefaciens</i>	Plantarum FZB42	Rhizosphere	(Chen et al 2007)
Ga0077930 _11	Firmicutes	Bacilli	Bacillales	Bacillaceae	<i>Bacillus</i>	<i>anthracis</i>	BA1015	Pathogen	(Helgason et al 2000)
Ga0175433 _11	Firmicutes	Bacilli	Bacillales	Bacillaceae	<i>Bacillus</i>	<i>subtilis</i>	KH2	ND	(Graham and Istock 1978)
CP015911 LAC_LBA_D RAFT__cont	Firmicutes	Bacilli	Bacillales	Bacillaceae	<i>Bacillus</i>	<i>velezensis</i>	LS69	Rhizosphere	(Liu et al 2017)
ig00001 Ga0106160 _101	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	<i>Lactobacillus</i>	<i>plantarum</i>		Rhizosphere	(Chen et al 2005)
Ga0106062 _101	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	<i>Lactobacillus</i>	<i>composti</i>	DSM18527	Rhizosphere	(Endo and Okada 2007)
Ga0106062 _101	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	<i>Lactobacillus</i>	<i>floricola</i>	DSM23037	Rhizosphere	(Kawasaki et al 2011)
1423745	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	<i>Lactobacillus</i>	<i>florum</i>	F91	Rhizosphere	(Endo et al 2010)
1670446 Ga0126468 _101	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	<i>Lactobacillus</i>	<i>herbarum</i>		ND	(Mao et al 2015)
Ga0106620 _101	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	<i>Leuconostoc</i>	<i>mesenteroides</i>	213MO	Rhizosphere	(Chen et al 2005)
Ga0114202 _101	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	<i>Weissella</i>	<i>confusa</i>	DSM20196	Rhizosphere	(Chen et al 2005)
155866	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	<i>Fructobacillus</i>	<i>sp</i>	EFBN1	Decomposition	(Endo et al 2009)
907931 BO10DRAF T_scf71800	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	<i>Leuconostoc</i>	<i>fallax</i>	KCTC3537	Rhizosphere	(Magnusson et al 2002)
00000004_q uiver.1	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	<i>Clostridium</i>	<i>phytofermentans</i>	KNHs212	Decomposition	(Warnick et al 2002)
CLOBL_conti g000001	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	<i>Clostridium</i>	<i>beijerinckii</i>	A21	Rhizosphere	(Johnson et al 1997)

1										
2										
3										
4										
5										
6	ct454_4544									
7	54contig000									
8	01	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	<i>Clostridium</i>	<i>tetani</i>	ATCC454	Pathogen	(Cohen et al 2017)
9	1121919	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	<i>Geosporobacter</i>	<i>subterraneus</i>	DSM17957	ND	(Klouche et al 2007)
10	857293	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	<i>Caloramator</i>	<i>australicus</i>	RC3	ND	(Ogg and Patel 2009)
11	GCA_90047									
12	6375.1	Firmicutes	Negativicutes	Selenomonadales	Sporomusaceae	<i>Lucifera</i>	<i>butyrica</i>		ND	(Sanchez-Andrea et al 2018)
13	NZ_CP0109									
14	78.1	Firmicutes	Negativicutes	Selenomonadales	Sporomusaceae	<i>Pelosinus</i>	<i>fermentans</i>	JBW45	Decomposition	(de Leon et al 2012)
15	16187	Firmicutes	Negativicutes	Selenomonadales	Sporomusaceae	<i>Sporomusa</i>	<i>ovata</i>	DSM2662	ND	(Möller et al 1984)
16	51809	Firmicutes	Negativicutes	Selenomonadales	Sporomusaceae	<i>Sporomusa</i>	<i>sphaeroides</i>	DSM2875	ND	(Möller et al 1984)
17	GCA_00043									
18	0605.1	Firmicutes	Negativicutes	Selenomonadales	Sporomusaceae	<i>Anaeroarcus</i>	<i>burkinensis</i>	DSM6283	Rhizosphere	(Strömpl et al 1999)
19	Ga0074713									
20	_11	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	<i>Arthrobacter</i>	<i>chlorophenolicus</i>	DSM12829	Bioremediation	(Westerberg et al 2000)
21	Ga0067230									
22	_11	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	<i>Arthrobacter</i>	<i>aurescens</i>	TC1	Bioremediation	(Mongodin et al 2006)
23	1531955	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	<i>Sinomonas</i>	<i>mesophila</i>	MPKL26	ND	(Prabhu et al 2015)
24	378753	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	<i>Kocuria</i>	<i>rhizophila</i>	DC2201	Rhizosphere	(Takarada et al 2008)
25	574650	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	<i>Micrococcus</i>	<i>terreus</i>		ND	(Zhang et al 2010)
26	Ga0080924									
27	_101	Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	<i>Mycobacterium</i>	<i>chlorophenolicum</i>	DSM43826	Bioremediation	(Das et al 2015)
28	Ga0124153									
29	_1001	Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	<i>Mycobacterium</i>	<i>fortuitum</i>	ATCC6841	Bioremediation	(Nohynek et al 1993)
30	Ga0056865									
31	_101	Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	<i>Mycobacterium</i>	<i>austroafricanum</i>	DSM44191	Bioremediation	(Leys et al 2005)
32	443218	Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	<i>Hoyosella</i>	<i>subflava</i>	DQS39A1	Bioremediation	(Hamada et al 2016)
33	1440774	Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	<i>Mycobacterium</i>	<i>aromaticivorans</i>	JCM16368	Bioremediation	(Hennessee et al 2009)
34	NC_014151									
35	Ga0128466	Actinobacteria	Actinobacteria	Actinomycetales	Cellulomonadaceae	<i>Cellulomonas</i>	<i>flavigena</i>	DSM20109	Decomposition	(Abt et al 2010)
36	_10001	Actinobacteria	Actinobacteria	Actinomycetales	Cellulomonadaceae	<i>Cellulomonas</i>	<i>chitinilytica</i>	JCM16927	Decomposition	(Yoon et al 2008)
37	Ga0128457									
38	_10001	Actinobacteria	Actinobacteria	Actinomycetales	Cellulomonadaceae	<i>Cellulomonas</i>	<i>xylanilytica</i>	JCM14281	Decomposition	(Rivas et al 2004 <sup>a</sup> )
39	862422	Actinobacteria	Actinobacteria	Actinomycetales	Cellulomonadaceae	<i>Actinotalea</i>	<i>fermentans</i>	DSM3133	Decomposition	(Bagnara et al 1985)
40										
41										
42										
43										
44										
45										
46										
47										

1										
2										
3										
4										
5										
6										
7	948458	Actinobacteria	Actinobacteria	Actinomycetales	Cellulomonadaceae	<i>Actinotalea</i>	<i>ferrariae</i>	CF54	ND	(Li et al 2013)
8	Ga0057460	Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	<i>Streptomyces</i>	<i>scabiei</i>	NCPB4086	Pathogen	(Dees et al 2013)
9	_1001	Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	<i>Streptomyces</i>	<i>lividans</i>	TK24	Decomposition	(Ruckert et al 2015)
10	Ga0057466	Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	<i>Streptomyces</i>	<i>coelicolor</i>	A3(2)	Decomposition	(Bentley et al 2002)
11	_gi6723671	Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	<i>Streptomyces</i>	<i>coelicolor</i>	A3(2)	Decomposition	(Bentley et al 2002)
12	50.1	Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	<i>Streptomyces</i>	<i>coelicolor</i>	A3(2)	Decomposition	(Bentley et al 2002)
13	Ga0076575	Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	<i>Streptomyces</i>	<i>coelicolor</i>	A3(2)	Decomposition	(Bentley et al 2002)
14	_11	Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	<i>Streptomyces</i>	<i>coelicolor</i>	A3(2)	Decomposition	(Bentley et al 2002)
15	105422	Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	<i>Streptacidiphilus</i>	<i>carbonis</i>	DSM41754	Decomposition	(Kim et al 2003)
16	1348663	Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	<i>Kitasatospora</i>	<i>cheerisanensis</i>	KCTC2395	ND	(Hwang et al 2015)
17	326424	Actinobacteria	Actinobacteria	Frankiales	Frankiaceae	<i>Frankia</i>	<i>alni</i>	ACN14a	Nitrogen fixation	(Normand and Lalonde 1982)
18	106370	Actinobacteria	Actinobacteria	Frankiales	Frankiaceae	<i>Frankia</i>	<i>casuarinae</i>	Ccl3	Nitrogen fixation	(Normand et al 2007)
19	298654	Actinobacteria	Actinobacteria	Frankiales	Frankiaceae	<i>Frankia</i>	<i>inefficax</i>	Eul1c	ND	(Nouioui et al 2017)
20	656024	Actinobacteria	Actinobacteria	Frankiales	Frankiaceae	<i>Frankia</i>	<i>sp</i>	Datisca symbiont	Nitrogen fixation	(Persson et al 2015)
21	222534	Actinobacteria	Actinobacteria	Frankiales	Frankiaceae	<i>Frankia</i>	<i>elaeagni</i>	BMG512	Nitrogen fixation	(Nouioui et al 2016)
22	15565	Actinobacteria	Actinobacteria	Micrococcales	Promicromonosporaceae	<i>Cellulosimicrobium</i>	<i>cellulans</i>	J36	Decomposition	(Schumann et al 2001)
23	75317	Actinobacteria	Actinobacteria	Micrococcales	Promicromonosporaceae	<i>Cellulosimicrobium</i>	<i>terreum</i>		Decomposition	(Yoon et al 2007)
24	NC_013530.	Actinobacteria	Actinobacteria	Micrococcales	Promicromonosporaceae	<i>Xylanimonas</i>	<i>cellulosilytica</i>	DSM15894	Decomposition	(Rivas et al 2003)
25	_1	Actinobacteria	Actinobacteria	Micrococcales	Promicromonosporaceae	<i>Xylanimonas</i>	<i>cellulosilytica</i>	DSM15894	Decomposition	(Takahashi et al 1987)
26	16475	Actinobacteria	Actinobacteria	Micrococcales	Promicromonosporaceae	<i>Promicromonospora</i>	<i>sukumoe</i>	327MFSHa3.1	Decomposition	(Rivas et al 2004 <sup>b</sup> )
27	76147	Actinobacteria	Actinobacteria	Micrococcales	Promicromonosporaceae	<i>Xylanibacterium</i>	<i>ulmi</i>		Decomposition	(Zgurskaya et al 1992)
28	50964	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	<i>Agromyces</i>	<i>cerinus</i>		ND	(Jurado et al 2005)
29	16552	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	<i>Agromyces</i>	<i>subbeticus</i>	DSM16689	ND	(Schumann et al 2003)
30	54323	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	<i>Agreia</i>	<i>pratensis</i>		Rhizosphere	(Schumann et al 2003)
31	16541	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	<i>Agrococcus</i>	<i>lahaulensis</i>	DSM17612	ND	(Mayilraj et al 2006)
32	16623	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	<i>Humibacter</i>	<i>albus</i>	DSM18994	Decomposition	(Vaz-Moreira et al 2008)
33	Ga0065637	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	<i>Humibacter</i>	<i>albus</i>	DSM18994	Decomposition	(Vaz-Moreira et al 2008)
34	_11	Acidobacteria	Acidobacteria	Acidobacteriales	Acidobacteriaceae	<i>Koribacter</i>	<i>versatilis</i>	Ellin345	Decomposition	(Ward et al 2009)
35										
36										
37										
38										
39										
40										
41										
42										
43										
44										
45										
46										
47										

1										
2										
3										
4										
5										
6										
7	Terro_Conti g51.1	Acidobacteria	Acidobacteria	Acidobacteriales	Acidobacteriaceae	<i>Terriglobus</i>	<i>roseus</i>	DSM18391	Decomposition	(Eichorst et al 2007)
8	Ga0133452 _11	Acidobacteria	Acidobacteria	Acidobacteriales	Acidobacteriaceae	<i>Luteitalea</i>	<i>pratensis</i>	DSM100886	Decomposition	(Vieira et al 2017)
9	863522	Acidobacteria	Acidobacteria	Acidobacteriales	Acidobacteriaceae	<i>Bryocella</i>	<i>elongata</i>	DSM22489	Decomposition	(Dedysh et al 2012)
10	1121860	Acidobacteria	Acidobacteria	Acidobacteriales	Acidobacteriaceae	<i>Edaphobacter</i>	<i>aggregans</i>	DSM19364	Decomposition	(Koch et al 2008)
11					Divergent					
12					Acidobacteria	<i>Candidatus</i>				
13	234267	Acidobacteria	Acidobacteria	Solibacteriales	lineage	<i>Solibacter</i>	<i>usitatus</i>	Ellin6076	Decomposition	(Ward et al 2009)
14					Divergent					
15	1340493	Acidobacteria	Acidobacteria	Solibacteriales	Acidobacteria	<i>Bryobacter</i>	<i>aggregatus</i>	MPL3	Decomposition	(Kulichevskaya et al 2010)
16					lineage					
17	454194	Acidobacteria	Acidobacteria	Blastocatellales	Acidobacteria	<i>Pyrinomonas</i>	<i>methyhaliphatogenes</i>	K22	Decomposition	(Crowe et al 2014)
18					Divergent					
19					Acidobacteria					
20	903818	Acidobacteria	Acidobacteria	Holophagales	lineage	<i>Holophaga</i>	<i>foetida</i>	DSM6591	ND	(Liesack et al 1994)
21					Divergent					
22	1660251	Acidobacteria	Acidobacteria	Unclassified	Acidobacteria	<i>Acidobacteria</i>	<i>bacterium</i>	Mor1	Rhizosphere	(Tanaka et al 2017)
23	Ga0077356 _1001	Proteobacteria	Beta	Burkholderiales	Burkholderiaceae	<i>Burkholderia</i>	<i>pseudomallei</i>	TSV44	Pathogen	(Brook et al 1997)
24	Ga0065534 _11	Proteobacteria	Beta	Burkholderiales	Burkholderiaceae	<i>Burkholderia</i>	<i>xenovorans</i>	LB400	Rhizosphere	(Chain et al 2006)
25	Ga0056711 _1001	Proteobacteria	Beta	Burkholderiales	Burkholderiaceae	<i>Burkholderia</i>	<i>terrae</i>	NBRC100964	Rhizosphere	(Yang et al 2006)
26	1121278	Proteobacteria	Beta	Burkholderiales	Burkholderiaceae	<i>Chitinimonas</i>	<i>koreensis</i>	DSM17726	ND	(Kim et al 2006)
27	1040978	Proteobacteria	Beta	Burkholderiales	Burkholderiaceae	<i>Cupriavidus</i>	<i>taiwanensis</i>	STM6018	Nitrogen fixation	(Chen et al 2001)
28	Ga0105865 _101	Proteobacteria	Beta	Nitrosomonadales	Nitrosomonadaceae	<i>Nitrosomonas</i>	<i>europaea</i>	Nm50	Nitrification	(Koops et al 1991)
29	Ga0105862 _1001	Proteobacteria	Beta	Nitrosomonadales	Nitrosomonadaceae	<i>Nitrosomonas</i>	<i>communis</i>	Nm2	Nitrification	(Koops et al 1991)
30	Ga0105869 _101	Proteobacteria	Beta	Nitrosomonadales	Nitrosomonadaceae	<i>Nitrosomonas</i>	<i>eutropha</i>	Nm56	Nitrification	(Koops et al 1991)
31	1266925	Proteobacteria	Beta	Nitrosomonadales	Nitrosomonadaceae	<i>Nitrospira</i>	<i>briensis</i>	C128	Nitrification	(Rice et al 2016)
32	323848	Proteobacteria	Beta	Nitrosomonadales	Nitrosomonadaceae	<i>Nitrospira</i>	<i>multiformis</i>	ATCC25196	Nitrification	(Norton et al 2008)
33										
34										
35										
36										
37										
38										
39										
40										
41										
42										
43										
44										
45										
46										
47										

1										
2										
3										
4										
5										
6	G520DRAF									
7	T_scaffold00									
8	001.1	Proteobacteria	Beta	Neisseriales	Neisseriales_lineage	<i>Paludibacterium</i>	<i>yongneupense</i>	DSM19731	ND	(Kwon et al 2008)
9	Ga0077161									
10	_101	Proteobacteria	Beta	Neisseriales	Neisseriales_lineage	<i>Chromobacterium</i>	<i>subtsugae</i>	F49	ND	(Soby et al 2013)
11	Ga0080895									(Koburger and May 1982)
12	_101	Proteobacteria	Beta	Neisseriales	Neisseriales_lineage	<i>Chromobacterium</i>	<i>violaceum</i>	GN5	Pathogen	
13	1121274	Proteobacteria	Beta	Neisseriales	Neisseriales_lineage	<i>Chitinibacter</i>	<i>tainanensis</i>	DSM15459	Decomposition	(Chern et al 2004)
14	1120999	Proteobacteria	Beta	Neisseriales	Neisseriales_lineage	<i>Andreprevotia</i>	<i>chitinilytica</i>	DSM18519	Decomposition	(Weon et al 2007)
15	Ga0077767									
16	_1001	Proteobacteria	Gamma	Pseudomonadales	Pseudomonadaceae	<i>Pseudomonas</i>	<i>putida</i>	SF1	Rhizosphere	(Tikariha et al 2016)
17	Ga0061119									
18	_scaffold000									
19	01.1	Proteobacteria	Gamma	Pseudomonadales	Pseudomonadaceae	<i>Azotobacter</i>	<i>beijerinckii</i>	DSM373	Nitrogen fixation	(Lipman 1904)
20	Ga0129289									(Lymeropoulou et al 2017)
21	_101	Proteobacteria	Gamma	Pseudomonadales	Pseudomonadaceae	<i>Pseudomonas</i>	<i>oryzihabitans</i>	H72	Rhizosphere	
22	289370	Proteobacteria	Gamma	Pseudomonadales	Pseudomonadaceae	<i>Pseudomonas</i>	<i>argentinensis</i>	CECT7010	Rhizosphere	(Peix et al 2005)
23	1292027	Proteobacteria	Gamma	Pseudomonadales	Pseudomonadaceae	<i>Pseudomonas</i>	<i>azotifigens</i>	DSM17556	Nitrogen fixation	(Hatayama et al 2005)
24	Ga0044516									
25	_101	Proteobacteria	Gamma	Pseudomonadales	Moraxellaceae	<i>Acinetobacter</i>	<i>solii</i>		ND	(Kim et al 2008)
26	Ga0080829									(Davenport et al 2014)
27	_1001	Proteobacteria	Gamma	Pseudomonadales	Moraxellaceae	<i>Acinetobacter</i>	<i>baumannii</i>		Pathogen	
28	Q337DRAF									
29	T_scaffold00									
30	010.10	Proteobacteria	Gamma	Pseudomonadales	Moraxellaceae	<i>Alkanindiges</i>	<i>illinoisensis</i>	DSM15370	Bioremediation	(Bogan et al 2003)
31	1217715	Proteobacteria	Gamma	Pseudomonadales	Moraxellaceae	<i>Acinetobacter</i>	<i>bohemicus</i>	ANC3994	ND	(Krizova et al 2014)
32	1217660	Proteobacteria	Gamma	Pseudomonadales	Moraxellaceae	<i>Acinetobacter</i>	<i>indicus</i>	ANC4215	Bioremediation	(Malhotra et al 2012)
33	GY38DRAF									
34	T_scf71800									
35	00000021_q									
36	uiver.1	Proteobacteria	Gamma	Methylococcales	Methylococcaceae	<i>Methylobacter</i>	<i>whittenburyi</i>	ACM3310	Methane oxidation	(Whittenbury et al 1970)
37	MetmiDRAF									
38	T_scaffold1.									
39	1	Proteobacteria	Gamma	Methylococcales	Methylococcaceae	<i>Methylovulum</i>	<i>miyakonense</i>	HT12	Methane oxidation	(Iguchi et al 2011)
40	MetluDRAF									
41	T_scaffold1.									
42	1	Proteobacteria	Gamma	Methylococcales	Methylococcaceae	<i>Methylobacter</i>	<i>luteus</i>	IMVB3098	Methane oxidation	(Whittenbury et al 1970)
43	1132443	Proteobacteria	Gamma	Methylococcales	Methylococcaceae	<i>Methylosarcina</i>	<i>fibrata</i>	AMLC10	Methane oxidation	(Wise et al 2001)



1										
2										
3										
4										
5										
6										
7	493385	Proteobacteria	Gamma	Methylococcales	Methylococcaceae	<i>Methylogaea</i>	<i>oryzae</i>	JCM16910	Methane oxidation	(Geymonat et al 2011)
8	Ga0061138									
9	_scaffold000									
10	01.1	Proteobacteria	Alpha	Rhodospirillales	Rhodospirillaceae	<i>Azospirillum</i>	<i>brasiliense</i>	DSM1690	Nitrogen fixation	(Kuklinsky-Sobral et al 2004)
11	Ga0048939	Proteobacteria	Alpha	Rhodospirillales	Rhodospirillaceae	<i>Azospirillum</i>	<i>oryzae</i>	A2P	Nitrogen fixation	(Xie and Yokota 2005a)
12	_110									
13	Ga0048941	Proteobacteria	Alpha	Rhodospirillales	Rhodospirillaceae	<i>Azospirillum</i>	<i>lipoferum</i>	R1C	Nitrogen fixation	(Van et al 1997)
14	_11									
15	570967	Proteobacteria	Alpha	Rhodospirillales	Rhodospirillaceae	<i>Fodinicurvata</i>	<i>fenggangensis</i>	DSM21160	ND	(Wang et al 2009)
16										(Magalhaes et al 1983)
17	1003237	Proteobacteria	Alpha	Rhodospirillales	Rhodospirillaceae	<i>Nitrospirillum</i>	<i>amazonense</i>	Y2	Nitrogen fixation	
18	DL88DRAFT									
19	_scaffold000									
20	01.1	Proteobacteria	Alpha	Rhizobiales	Beijerinckiaceae	<i>Beijerinckia</i>	<i>mobilis</i>	DSM2326	Nitrogen fixation	(Derx 1950)
21	182677003	Proteobacteria	Alpha	Rhizobiales	Beijerinckiaceae	<i>Beijerinckia</i>	<i>indica</i>	ATCC9039	Nitrogen fixation	(Derx 1950)
22	217976201	Proteobacteria	Alpha	Rhizobiales	Beijerinckiaceae	<i>Methylocella</i>	<i>silvestris</i>	BL2	Methane oxidation	(Dunfield et al 2003)
23	395964	Proteobacteria	Alpha	Rhizobiales	Beijerinckiaceae	<i>Methylocapsa</i>	<i>acidiphila</i>	B2	Methane oxidation	(Dedysh et al 2002)
24	876269	Proteobacteria	Alpha	Rhizobiales	Beijerinckiaceae	<i>Methyloferula</i>	<i>stellata</i>	AR4	Methane oxidation	(Vorobev et al 2011)
25	A3OODRAF									
26	T_scaffold1.									
27	1	Proteobacteria	Alpha	Rhizobiales	Methylocystaceae	<i>Methylocystis</i>	<i>rosea</i>	SV97T	Methane oxidation	(Wartiainen et al 2006)
28	MettrDRAFT									
29	_Contig106	Proteobacteria	Alpha	Rhizobiales	Methylocystaceae	<i>Methylosinus</i>	<i>trichosporium</i>	OB3b	Methane oxidation	(Whittenbury et al 1970)
30	A3OUDRAF									
31	T_chromosome1.	Proteobacteria	Alpha	Rhizobiales	Methylocystaceae	<i>Methylopila</i>	<i>sp</i>	M107	Methane oxidation	(Chistoserdova 2011)
32										(Xie and Yokota 2005b)
33	1122963	Proteobacteria	Alpha	Rhizobiales	Methylocystaceae	<i>Pleomorphomonas</i>	<i>oryzae</i>	DSM16300	Nitrogen fixation	
34	1122962	Proteobacteria	Alpha	Rhizobiales	Methylocystaceae	<i>Pleomorphomonas</i>	<i>korensis</i>	DSM23070	Nitrogen fixation	(Im et al 2006)
35	Ga0065594	Proteobacteria	Alpha	Rhizobiales	Bradyrhizobiaceae	<i>Nitrobacter</i>	<i>winogradskyi</i>	Nb255	Nitrification	(Starkenbourg et al 2006)
36	_11									
37	Ga0098245	Proteobacteria	Alpha	Rhizobiales	Bradyrhizobiaceae	<i>Bradyrhizobium</i>	<i>japonicum</i>	E109	Nitrogen fixation	(Brunel et al 1988)
38	_11									
39	Ga0077765	Proteobacteria	Alpha	Rhizobiales	Bradyrhizobiaceae	<i>Rhodopseudomonas</i>	<i>palustris</i>	42OL	Bioremediation	(Larimer et al 2004)
40	_1001									
41	1031710	Proteobacteria	Alpha	Rhizobiales	Bradyrhizobiaceae	<i>Oligotropha</i>	<i>carboxidovorans</i>	OM4	Decomposition	(Volland et al 2011)
42										
43										
44										
45										
46										
47										

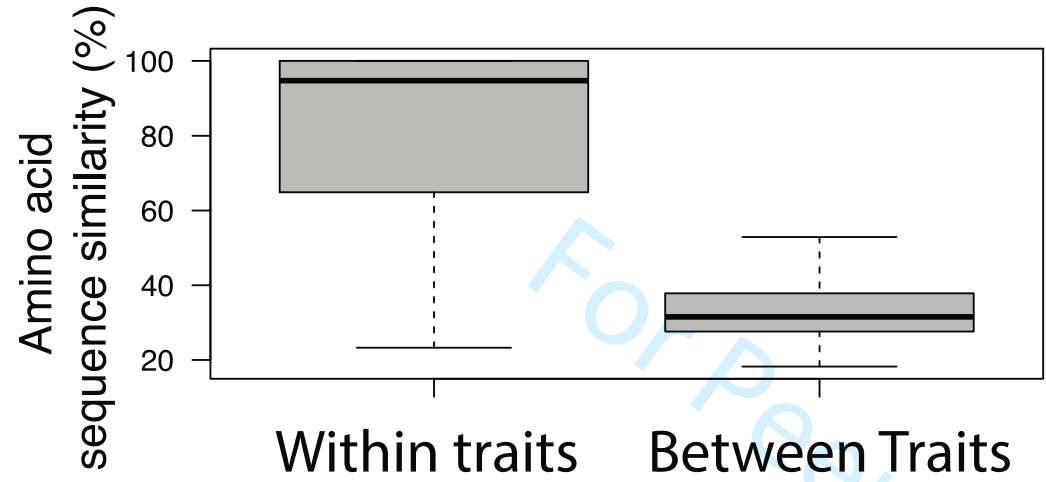
1										
2										
3										
4										
5										
6										
7	1123229	Proteobacteria	Alpha	Rhizobiales	Bradyrhizobiaceae	<i>Salinarimonas</i>	<i>rosea</i>	DSM21201	ND	(Liu et al 2010)
8	Ga0127029	Proteobacteria	Alpha	Rhizobiales	Rhizobiaceae	<i>Agrobacterium</i>	<i>tumefaciens</i>	B6	Rhizosphere	(Levanon 1978)
9	EW91DRAFT									
10	T_scaffold00									
11	_001.1	Proteobacteria	Alpha	Rhizobiales	Rhizobiaceae	<i>Rhizobium</i>	<i>leguminosarum</i>	CF307	Nitrogen fixation	(Naeem et al 2004)
12	Ga0174993	Proteobacteria	Alpha	Rhizobiales	Rhizobiaceae	<i>Sinorhizobium</i>	<i>americanum</i>	CFNEI73	Rhizosphere	(Toledo et al 2003)
13	_11									(Osterman et al 2014)
14	1028800	Proteobacteria	Alpha	Rhizobiales	Rhizobiaceae	<i>Neorhizobium</i>	<i>galegae</i>	HAMBI540	Nitrogen fixation	(Shamseldin et al 2016)
15	1764550	Proteobacteria	Alpha	Rhizobiales	Rhizobiaceae	<i>Rhizobium</i>	<i>aegyptiacum</i>	1010	Nitrogen fixation	
16	Ga0128794	Proteobacteria	Delta	Myxococcales	Polyangiaceae	<i>Sorangium</i>	<i>cellulosum</i>	So0149	Decomposition	(Li et al 2014)
17	_10470									(Zaburannyi et al 2016)
18	Ga0098230	Proteobacteria	Delta	Myxococcales	Polyangiaceae	<i>Chondromyces</i>	<i>crocatus</i>	Cmc5	ND	
19	_11									
20	Ga0081738	Proteobacteria	Delta	Myxococcales	Polyangiaceae	<i>Polyangium</i>	<i>brachysporum</i>	DSM7029	ND	(Tang et al 2015)
21	_11									(Kleinig and Reichenbach 1973)
22	1192034	Proteobacteria	Delta	Myxococcales	Polyangiaceae	<i>Chondromyces</i>	<i>apiculatus</i>	DSM436	ND	
23	Ga0128396	Proteobacteria	Delta	Myxococcales	Polyangiaceae	<i>Byssovorax</i>	<i>cruenta</i>		ND	(Reichenbach et al 2006)
24	_10001									(Voelz and Dworkin 1962)
25	Ga0131205	Proteobacteria	Delta	Myxococcales	Myxococcaceae	<i>Myxococcus</i>	<i>xanthus</i>	DSM16526	ND	
26	_101									
27	Ga0070493	Proteobacteria	Delta	Myxococcales	Myxococcaceae	<i>Myxococcus</i>	<i>virescens</i>	DSM2260	ND	(Lang et al 2008)
28	_101									
29	Ga0131203	Proteobacteria	Delta	Myxococcales	Myxococcaceae	<i>Myxococcus</i>	<i>fulvus</i>	DSM16525	ND	(Jahn 1911)
30	_101									
31	1144275	Proteobacteria	Delta	Myxococcales	Myxococcaceae	<i>Corallococcus</i>	<i>coralloides</i>	DSM2259	ND	(Huntley et al 2012)
32	_101									
33	1189310	Proteobacteria	Delta	Myxococcales	Myxococcaceae	<i>Myxococcus</i>	<i>macrosporus</i>	DSM14697	ND	(Lang and Stackebrandt 2009)
34	_101									(Sangkhobol and Skerman 1981)
35	NC_013132	Bacteroidetes	Bacteroidetes	Chitinophagales	Chitinophagaceae	<i>Chitinophaga</i>	<i>pinensis</i>	DSM2588	Decomposition	
36	B154DRAFT									
37	_scaffold_0.									
38	1	Bacteroidetes	Bacteroidetes	Chitinophagales	Chitinophagaceae	<i>Segetibacter</i>	<i>koreensis</i>	DSM18137	ND	(An et al 2007a)
39	H608DRAFT									
40	_scaffold000									
41	01.1	Bacteroidetes	Bacteroidetes	Chitinophagales	Chitinophagaceae	<i>Terrimonas</i>	<i>ferruginea</i>	DSM30193	ND	(Xie and Yokota 2006)
42	551991	Bacteroidetes	Bacteroidetes	Chitinophagales	Chitinophagaceae	<i>Arachidicoccus</i>	<i>rhizosphaerae</i>	Vu144	Rhizosphere	(Madhaiyan et al 2015)
43	929713	Bacteroidetes	Bacteroidetes	Chitinophagales	Chitinophagaceae	<i>Niabella</i>	<i>solii</i>	DSM19437	ND	(Weon et al 2008)
44										
45										
46										
47										

1										
2										
3										
4										
5										
6										
7	2506615776 Ga0066359	Bacteroidetes	Sphingobacteria	Sphingobacteriales	Cytophagaceae	<i>Niastella</i>	<i>koreensis</i>	GR2010	ND	(Weon et al 2006)
8	_101	Bacteroidetes	Sphingobacteria	Sphingobacteriales	Cytophagaceae	<i>Cytophaga</i>	<i>hutchinsonii</i>	DSM1761	Decomposition	(Xie et al 2007)
9	Ga0116983	Bacteroidetes	Sphingobacteria	Sphingobacteriales	Cytophagaceae	<i>Pontibacter</i>	<i>indicus</i>	LP100	Bioremediation	(Singh et al 2014)
10	_11	Bacteroidetes	Sphingobacteria	Sphingobacteriales	Cytophagaceae	<i>Arsenicibacter</i>	<i>rosenii</i>	SM1	Bioremediation	(Huang et al 2017)
11	1750698	Bacteroidetes	Sphingobacteria	Sphingobacteriales	Cytophagaceae	<i>Arsenicibacter</i>	<i>rosenii</i>	SM1	Bioremediation	(Huang et al 2017)
12	1121373	Bacteroidetes	Sphingobacteria	Sphingobacteriales	Cytophagaceae	<i>Cytophaga</i>	<i>aurantiaca</i>	DSM3654	Decomposition	(Winogradsky 1929)
13	Ga0056857	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobia lineage	<i>Verrucomicrobium</i>	<i>sp</i>	BvORR106	Rhizosphere	Unpublished
14	_1001	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobia lineage	<i>Verrucomicrobium</i>	<i>sp</i>	BvORR034	Rhizosphere	Unpublished
15	Ga0056855	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobia lineage	<i>Verrucomicrobium</i>	<i>sp</i>	BvORR034	Rhizosphere	Unpublished
16	_1001	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobia lineage	<i>Verrucomicrobium</i>	<i>sp</i>	BvORR034	Rhizosphere	Unpublished
17	182411827	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobia lineage	<i>Opitutus</i>	<i>terrae</i>	PB90	ND	(Chin et al 2001)
18	690879	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobia lineage	<i>Terrimicrobium</i>	<i>sacchariphilum</i>	NM5	ND	(Qiu et al 2014)
19	497964	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobia lineage	<i>Chthoniobacter</i>	<i>flavus</i>	Ellin428	ND	(Sangwan et al 2004)
20	Ga0154327	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetales lineage	<i>Planctomyces</i>	<i>sp</i>	GWA2407	ND	Unpublished
21	_1001	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetales lineage	<i>Planctomyces</i>	<i>sp</i>	GWA2407	ND	Unpublished
22	Ga0111584	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetales lineage	<i>Planctomicrobium</i>	<i>piriforme</i>	DSM26348	ND	(Kulichevskaya et al 2015)
23	_101	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetales lineage	<i>Planctomicrobium</i>	<i>piriforme</i>	DSM26348	ND	(Kulichevskaya et al 2015)
24	Ga0198709	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetales lineage	<i>Paludisphaera</i>	<i>borealis</i>	PX4	ND	(Kulichevskaya et al 2016)
25	_11	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetales lineage	<i>Paludisphaera</i>	<i>borealis</i>	PX4	ND	(Kulichevskaya et al 2016)
26	1123242	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetales lineage	<i>Schlesneria</i>	<i>paludicola</i>	DSM18645	ND	(Kulichevskaya et al 2007)
27	886293	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetales lineage	<i>Singulisphaera</i>	<i>acidiphila</i>	DSM18658	ND	(Kulichevskaya et al 2008)
28	GCA_00021	Cyanobacteria	Cyanobacteria	Oscillatoriales	Microcoleaceae	<i>Microcoleus</i>	<i>vaginatus</i>	FGP-2	Phototroph	(Starkenburg et al 2011)
29	4075.2	Cyanobacteria	Cyanobacteria	Oscillatoriales	Microcoleaceae	<i>Microcoleus</i>	<i>vaginatus</i>	FGP-2	Phototroph	(Starkenburg et al 2011)
30		Cyanobacteria	Cyanobacteria	Nostocales	Nostocaceae	<i>Nostoc</i>	<i>punctiforme</i>	PCC73102	Phototroph	(Rippka and Herdman, 1992)
31	67533	Cyanobacteria	Cyanobacteria	Chroococcidiopsidales	Chroococcidiopsidaceae	<i>Chroococcidiopsis</i>	<i>cubana</i>		Phototroph	(Komarek and Hindak, 1975)
32		Cyanobacteria	Cyanobacteria	Nostocales	Calotrichaceae	<i>Calothrix</i>	<i>desertica</i>		Phototroph	(Schwabe, 1960)
33	74975	Cyanobacteria	Cyanobacteria	Nostocales	Calotrichaceae	<i>Calothrix</i>	<i>desertica</i>		Phototroph	(Schwabe, 1960)
34	GCA_00163	Cyanobacteria	Cyanobacteria	Synechococcales	Leptolyngbyaceae	<i>Leptolyngbya</i>	<i>valderiana</i>	BDU20041	Phototroph	(Anagnostidis and Komarek, 1988)
35	7395.1	Cyanobacteria	Cyanobacteria	Synechococcales	Leptolyngbyaceae	<i>Leptolyngbya</i>	<i>valderiana</i>	BDU20041	Phototroph	(Anagnostidis and Komarek, 1988)
36	NC-002936.3	Chloroflexi	Dehalococcoidia	Dehalococcoidales	Dehalococcoidaceae	<i>Dehalococcoides</i>	<i>mccartyi</i>	CG5	Bioremediation	(He et al 2003)
37	50999	Chloroflexi	Ktedonobacteria	Ktedonobacteriales	Ktedonobacteraceae	<i>Ktedonobacter</i>	<i>Sp.</i>		ND	Unpublished
38										
39										
40										
41										
42										
43										
44										
45										
46										
47										

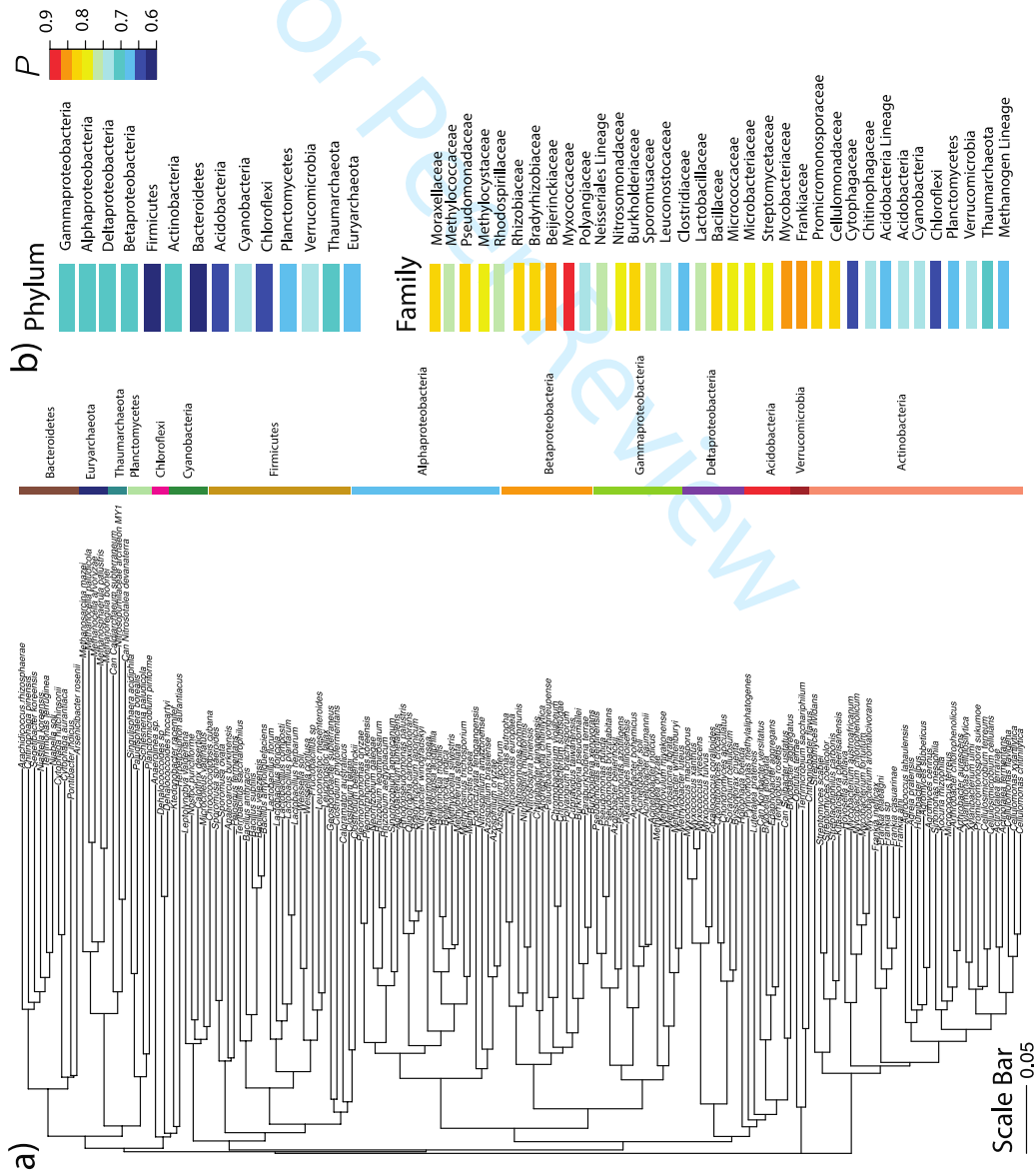
1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

46725	Chloroflexi	Anaerolineae	Anaerolineales	Anaerolineaceae	<i>Anaerolinea</i>	<i>Sp.</i>		Decomposition	Unpublished
13657	Chloroflexi	Dehalococcoidia	Dehalococcoidales	Dehalococcoidaceae	<i>Dehalococcoides</i>	<i>Sp.</i>		Bioremediation	Unpublished
GCA_000018565.1	Chloroflexi	Chloroflexia	Herpetosiphonales	Herpetosiphonaceae	<i>Herpetosiphon</i>	<i>aurantiacus</i>		ND	(Kiss et al 2011)
LRC50LV_640427159s	Euryarchaeota	Methanomicrobia	Methanocellales	Methanogen lineage	<i>Methanocella</i>	<i>arvoryzae</i>	MRE50	Methanogenesis	(Sakai et al 2010)
MCPLV_646311984standard	Euryarchaeota	Methanomicrobia	Methanocellales	Methanogen lineage	<i>Methanocella</i>	<i>paludicola</i>	SANAE	Methanogenesis	(Sakai et al 2008)
Ga0072472_11	Euryarchaeota	Methanomicrobia	Methanosarcinales	Methanogen lineage	<i>Methanosarcina</i>	<i>mazei</i>	C16	Methanogenesis	(Maestrojuán et al 1992)
456442	Euryarchaeota	Methanomicrobia	Methanosarcinales	Methanogen lineage	<i>Methanoregula</i>	<i>boonei</i>	6A8	Methanogenesis	(Brauer et al 2011)
521011	Euryarchaeota	Methanomicrobia	Methanosarcinales	Methanogen lineage	<i>Methanosphaerula</i>	<i>palustris</i>	E19c	Methanogenesis	(Cadillo-Quiroz et al 2009)
Ga0073019_101	Thaumarchaeota	Nitrososphaeria	Nitrososphaerales	Thaumarchaeota lineage	<i>Nitrocosmicus</i>	<i>oleophilus</i>	MY3	Nitrification	(Jung et al 2016)
CCsub_BA00048_1	Thaumarchaeota	Nitrososphaeria	Nitrososphaerales	Thaumarchaeota lineage	<i>Caldiarchoaeum</i>	<i>subterraneum</i>		Nitrification	(Takami et al 2015)
bd31_gi386807657.1	Thaumarchaeota	Nitrososphaeria	Nitrososphaerales	Thaumarchaeota lineage	<i>Nitrosopumilus</i>	<i>salaria</i>	BD31	Nitrification	(Mosier et al 2012)
1078905	Thaumarchaeota	Nitrososphaeria	Nitrososphaerales	Thaumarchaeota lineage	<i>Nitrosotalea</i>	<i>devanterra</i>		Nitrification	(Lehtovirta-Morley et al 2014)
1001994	Thaumarchaeota	Nitrososphaeria	Nitrososphaerales	Thaumarchaeota lineage	<i>Nitrosoarchaeum</i>	<i>koreensis</i>	MY1	Nitrification	(Kim et al 2011)

<sup>a</sup>ND: Not defined.

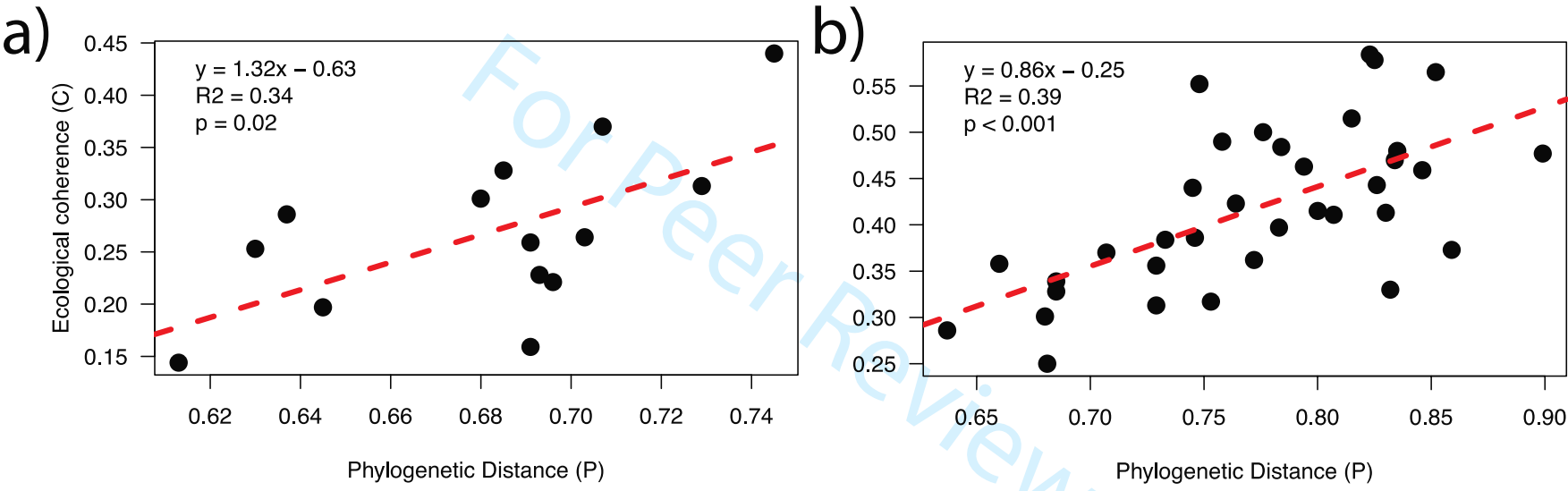


Supplementary Figure 1: Box and whisker plot comparing amino acid sequence similarity (%) within traits versus between traits, as determined by MCL. Amino acid sequences grouped together as traits were significantly more similar than those in different traits (Student's T test,  $p = 2 \times 10^{-16}$ ).





Supplementary Figure 2: a) Neighbour joining tree of full length 16S rRNA genes of taxa analysed in this study. b) Phylogenetic distance (*P*) of each taxonomic group, ranging between 0.6 (dissimilar) and 0.9 (similar) determined from branch lengths of 16S rRNA gene analysis in a). *P* was calculated as ecological coherence (Eq 1.) excepting that branch lengths are in units of gene sequence similarity and not Bray-Curtis distance (*C*).



Supplementary Figure 3: a) Linear correlation between Phylogenetic Distance (*P*) derived from 16S rRNA genes and ecological coherence (*C*) of functional traits present in Phyla analysed in this study. b) Linear correlation at the level of Family.

Supplementary Table 2: Confusion matrix of Phylum level predictions based on Random Forest selection of important traits.

## Successful classifications

	Acid.	Actino.	Alpha.	Bacter.	Beta.	Chloro.	Cyano.	Delta.	Eury.	Firm.	Gamma.	Planc.	Thaum.	Verruc.	Success rate (%)
Acidobacteria	6			1				1		2					60
Actinobacteria		33	1							1					95
Alphaproteobacteria			25												100
Bacteroidetes				9						1					90
Betaproteobacteria			2		11						2				74
Chloroflexi		1				0				3			1		0
Cyanobacteria		1					3			1					60
Deltaproteobacteria	1				1			8							80
Euryarchaeota									5						100
Firmicutes						2				23					92
Gammaproteobacteria			1		1					1	12				80
Planctomycetes	2							1		1		1			20
Thaumarchaeota										1			4		80
Verrucomicrobia	1			2										2	40

Supplementary Table 3: Confusion matrix of Family level predictions.

Phylum	Family	Mor.	Methyloc.	Pseu.	Methylocy.	Rhod.	Rhiz.	Brad.	Beij.	Myxo.	Poly.	Neis.	Nitr.	Burk.	Spor.	Leuc.	Clos.	Lact.	Bac.	Micrococ.	Microbac.	Success Rate (%)
Gammaproteobacteria	Moraxellaceae	5																				100
	Methylococcaceae		4																			80
Alphaproteobacteria	Pseudomonadaceae			4									1									80
	Methylocystaceae				3				2													60
	Rhodospirillaceae				1	4																80
	Rhizobiaceae						4	1														80
	Bradyrhizobiaceae				2		1	2														40
	Beijerinckiaceae				1					4												
Deltaproteobacteria	Myxococcaceae									5												100
	Polyangiaceae									1	3			1								60
Betaproteobacteria	Neisseriales Lineage											5										100
	Nitrosomonadaceae			1									4									80
Firmicutes	Burkholderiaceae			1																		60
	Sporomusaceae													3								40
	Leuconostocaceae														2							100
	Clostridiaceae																4		1			80
	Lactobacillaceae															2		3				60
	Bacillaceae																1		4			80
Actinobacteria	Micrococcaceae																			3		60
	Microbacteriaceae						1														4	80
	Streptomycetaceae																			1		60
	Mycobacteriaceae																					100
	Frankiaceae																					60
Promicromonosporaceae																					100	

For Peer Review

1																			
2																			
3																			
4																			
5																			
6		Cellulomonada																	
7		ceae														1			20
8	Bacteroidetes	Cytophagaceae																	40
9		Chitinophagaceae																	40
10		Divergent																	
11	Acidobacteria	Acidobacteria						1			1					1			0
12		Acidobacteriaceae																	100
13	Cyanobacteria	Cyanobacteria																	100
14	Chloroflexi	Chloroflexi										1							80
15		Planctomycetales Lineage										1							40
16	Planctomycetes	Verrucomicrobia Lineage																	80
17	Verrucomicrobia	Thaumarchaeota Lineage																	80
18	Thaumarchaeota	Methanogen Lineage																	80
19	Euryarchaeota																		100

Supplementary Table 4: Confusion matrix of Family level predictions.

Phylum	Family	Strep.	Myco.	Fran.	Prom.	Cell.	Cyto.	Chitin.	Acido.Lin.	Acido.	Cyano.	Chloro.	Planc.	Verru.	Thaum.	Methan.	Success Rate (%)	
Gammaproteobacteria	Moraxellaceae																	100
	Methylococcaceae											1						80
	Pseudomonadaceae																	80
Alphaproteobacteria	Methylocystaceae																	60
	Rhodospirillaceae																	80
	Rhizobiaceae																	80
	Bradyrhizobiaceae																	40
	Beijerinckiaceae																	80
Deltaproteobacteria	Myxococcaceae																	100
	Polyangiaceae																	60
Betaproteobacteria	Neisseriales Lineage																	100



Supplementary Table 5: Confusion matrix of Proteobacterial Family predictions.

	Successful classifications													Success rate (%)
	Beij.	Brady.	Burkh.	Methylococ.	Methylocys.	Morax.	Myxo.	Neiss.	Nitroso.	Polyan.	Pseudo.	Rhizo.	Rhodo.	
Beijerinckiaceae	5													100
Bradyrhizobiaceae		1			2						1	1		20
Burkholderiaceae			4					1						80
Methylococcaceae				4					1					80
Methylocystaceae	3				2									40
Moraxellaceae						4		1						80
Myxococcaceae							5							100
Neisseriales lineage								4	1					80
Nitrosomonadaceae									4		1			80
Polyangiaceae			1				1			3				60
Pseudomonadaceae						1					3	1		60
Rhizobiaceae											1	4		80
Rhodospirillaceae					1								4	80



Supplementary Table 6: Confusion matrix of Actinobacterial Family predictions.

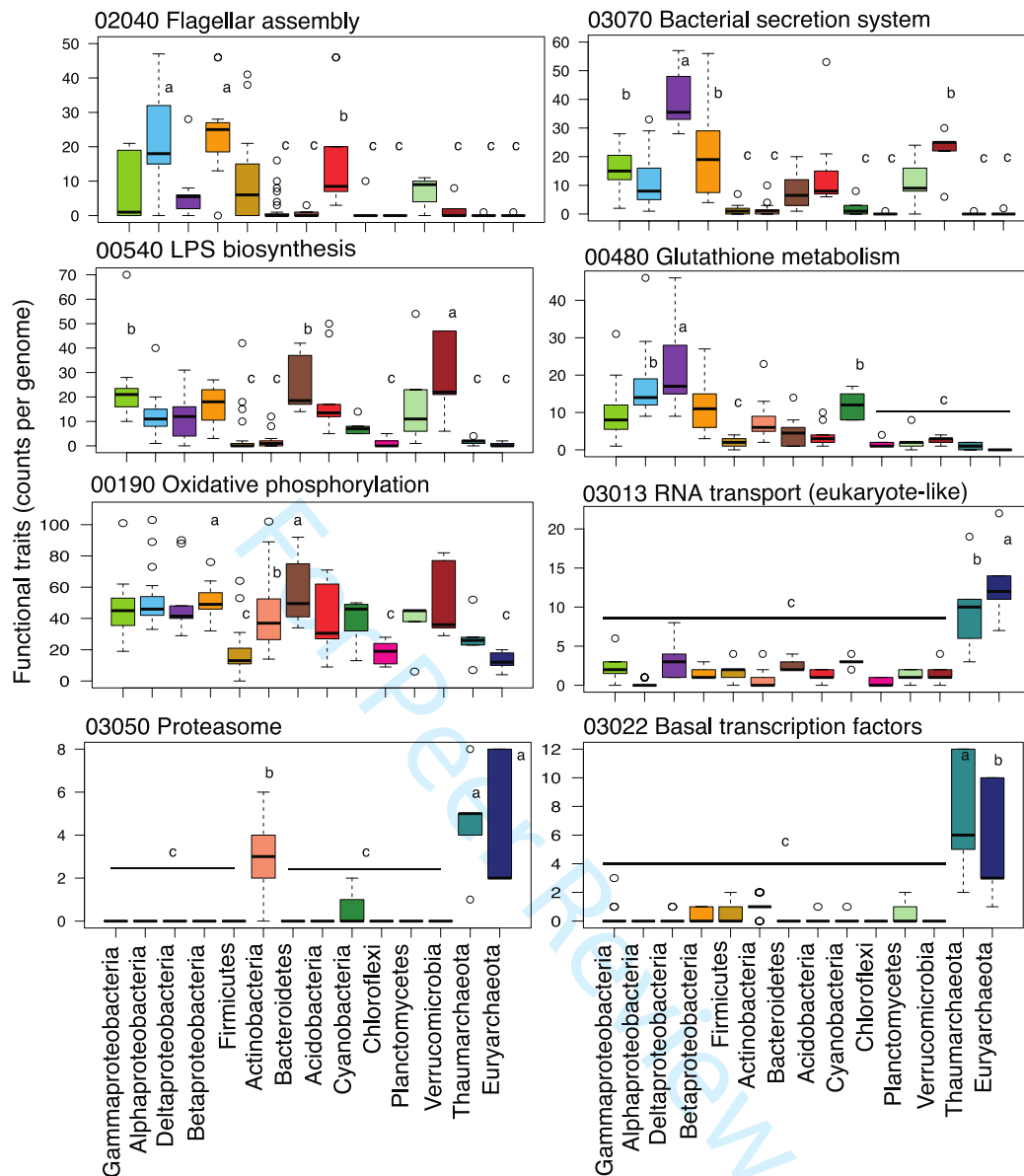
	Successful classifications							
	Cellu.	Frank.	Microbac.	Micrococ.	Mycobac.	Promicro.	Strepto.	Success rate (%)
Cellulomonadaceae	2		1	1		1		40
Frankiaceae		3		1	1			60
Microbacteriaceae	1		3				1	60
Micrococcaceae				4			1	80
Mycobacteriaceae					5			100
Promicromonosporaceae						5		100
Streptomycetaceae				1			4	80

Supplementary Table 7: Confusion matrix of Firmicutes Family predictions.

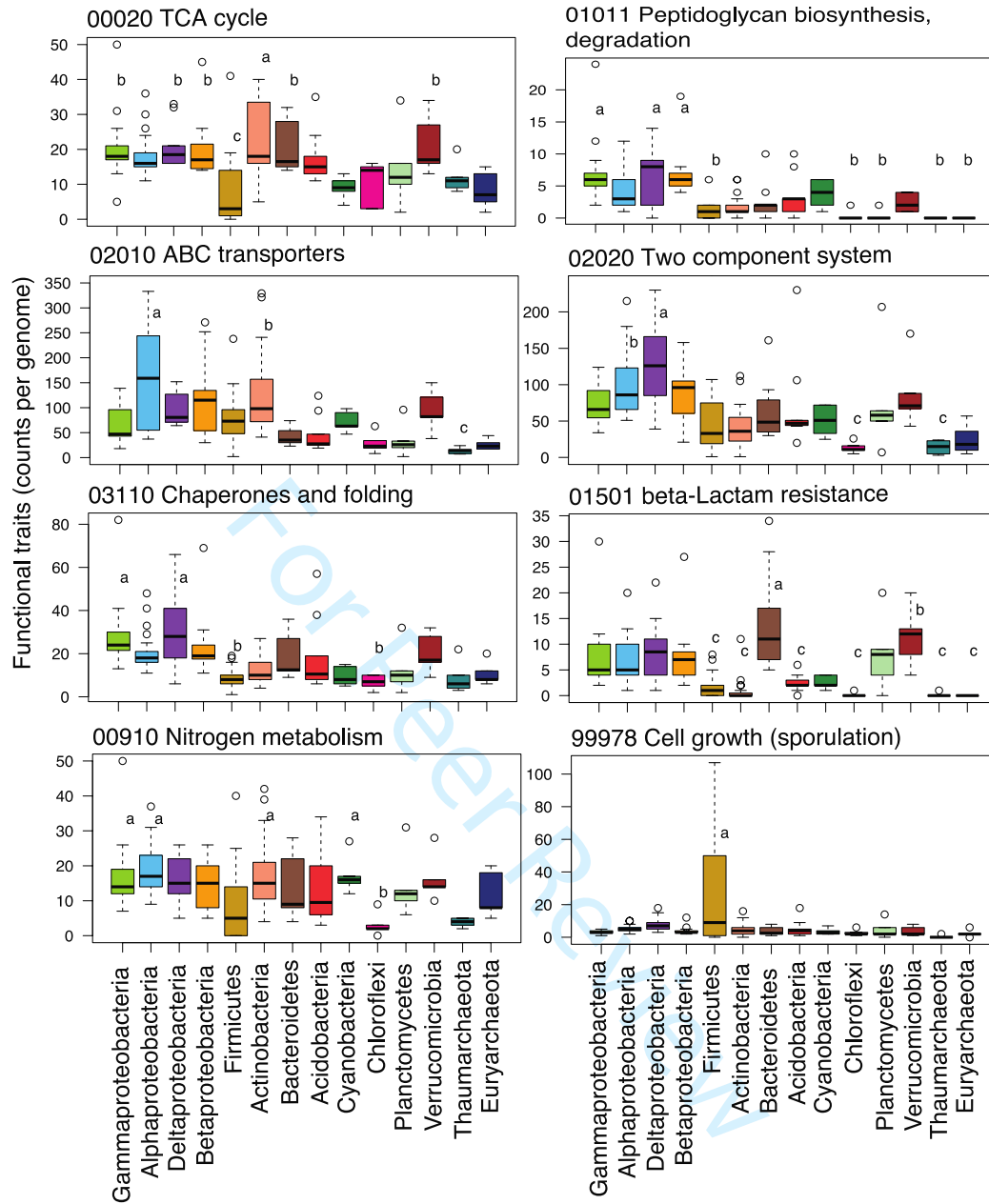
	Successful classifications					
	Bacill.	Clostrid.	Lacto.	Leuco.	Sporo.	Success rate (%)
Bacillaceae	4	1				80
Clostridiaceae	1	3	1			60
Lactobacillaceae			4	1		80
Leuconostocaceae				5		100
Sporomusaceae		1		1	3	60

Supplementary Table 8: Confusion matrix of 'Under-represented' Family predictions.  
Successful classifications

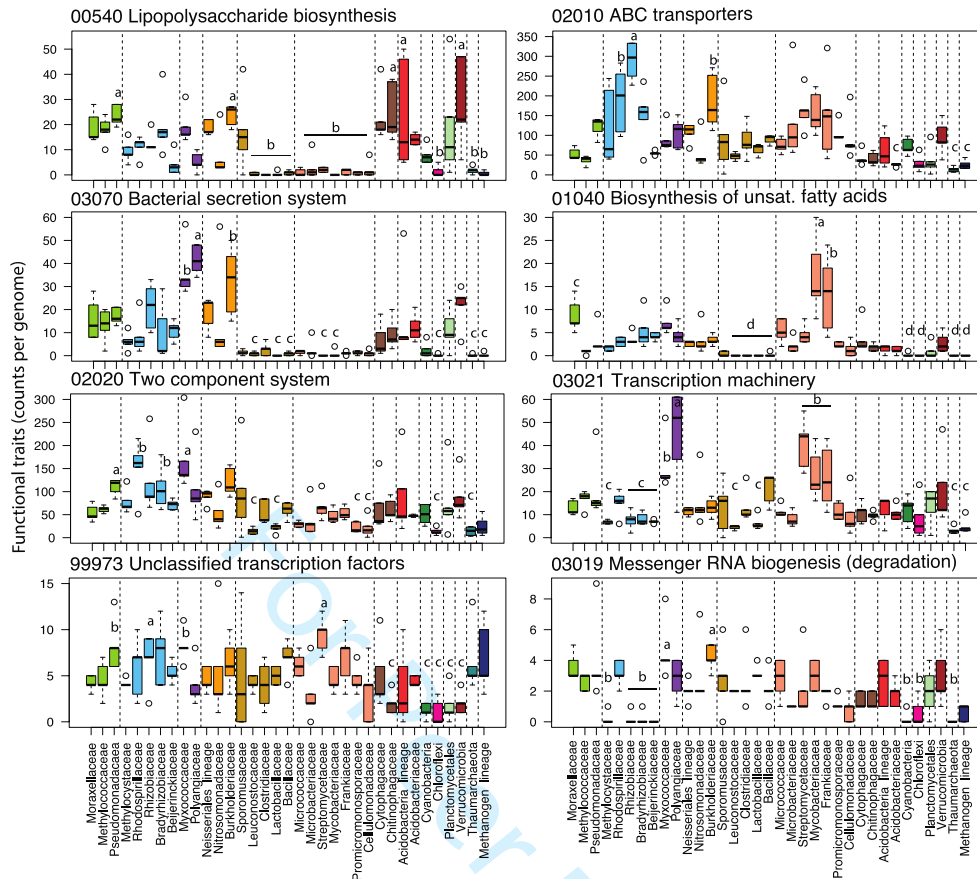
	Acid.	Chitin.	Chloro.	Cyano.	Cytoph.	Acido. Lin.	Methan.	Plancto.	Thaum.	Verruc.	Success rate (%)
Acidobacteriaceae	5										100
Chitinophagaceae		3			2						60
Chloroflexi			4					1			80
Cyanobacteria				5							100
Cytophagaceae		2			3						60
Acidobacteria Lineage	1		2							2	0
Methanogen Lineage							5				100
Planctomycetales			1			1		3			60
Thaumarchaeota			1						4		80
Verrucomicrobia	1									4	80



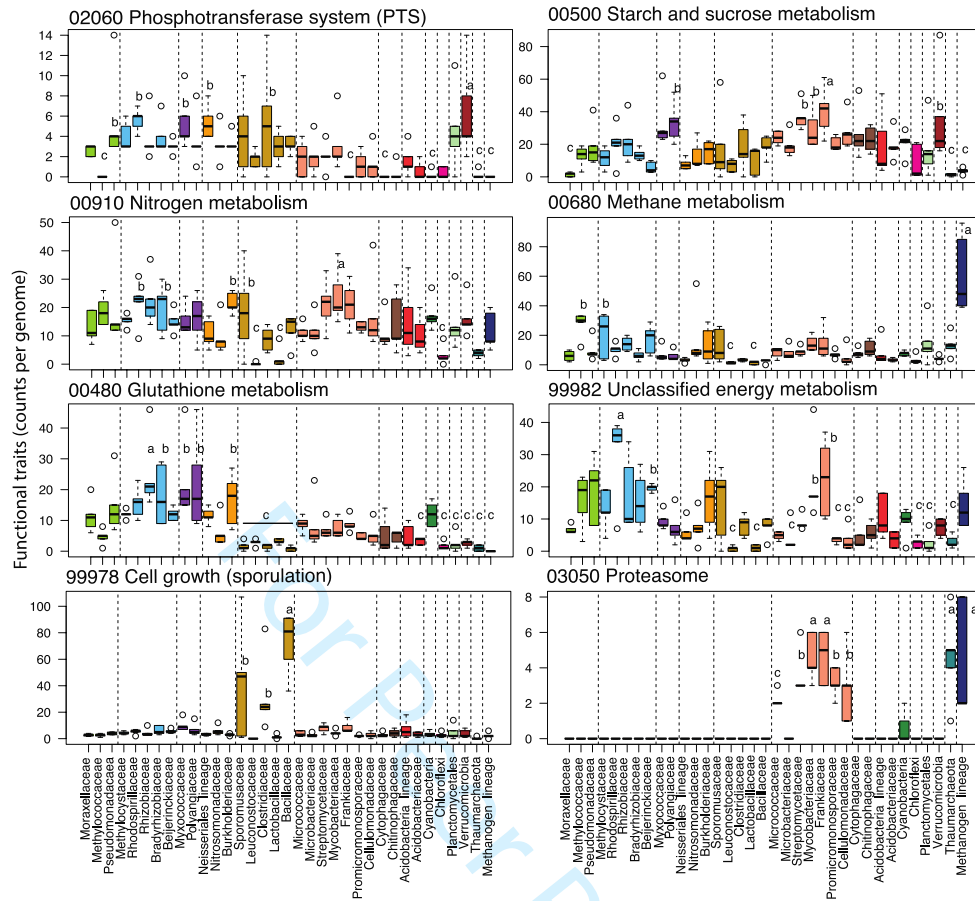
Supplementary Figure 4: Box and whisker plots of functional traits present in Phyla. Traits were selected based on being selected by the Phylum Random Forest model. Letters indicate significantly high or low ( $p < 0.05$ ) counts based on LSD *post hoc* tests.



Supplementary Figure 5: Box and whisker plots of functional traits present in Phyla. Traits were selected based on being selected by the Phylum Random Forest model. Letters indicate significantly high or low ( $p < 0.05$ ) counts based on LSD *post hoc* tests.

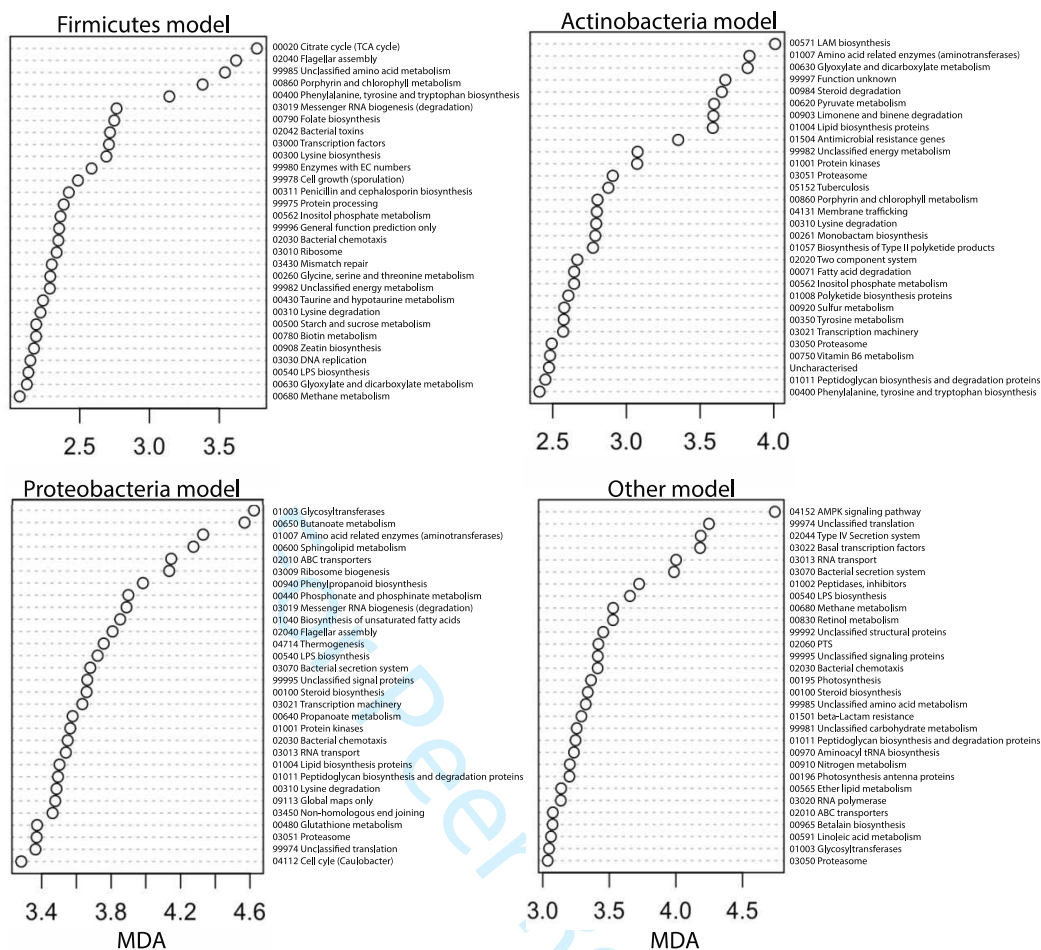


Supplementary Figure 6: Box and whisker plots of functional traits present in Families. Traits were selected based on being selected by the Family Random Forest model. Letters indicate significantly high or low ( $p < 0.05$ ) counts based on LSD *post hoc* tests.

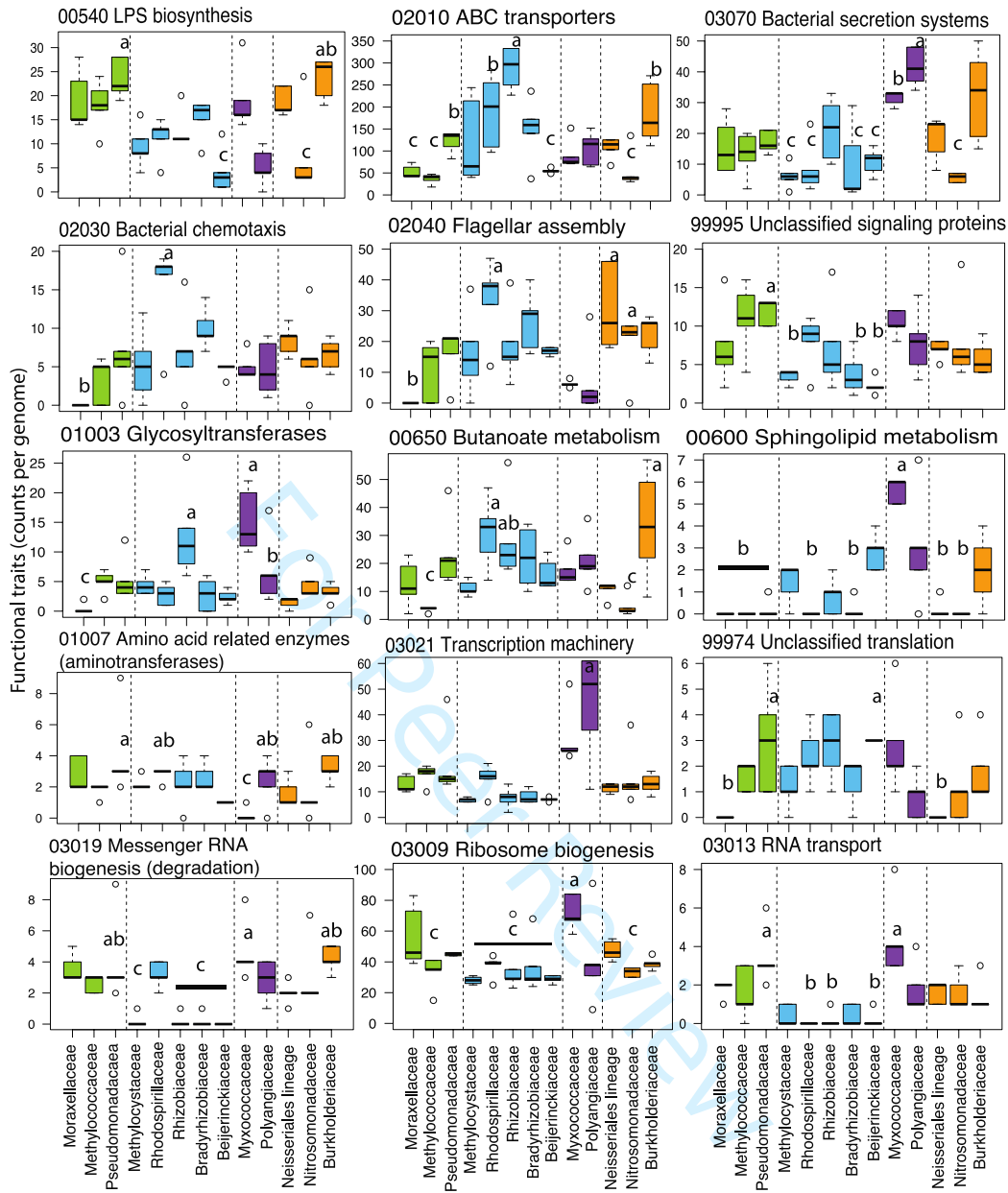


Supplementary Figure 7: Box and whisker plots of functional traits present in Families. Traits were selected based on being selected by the Family Random Forest model. Letters indicate significantly high or low ( $p < 0.05$ ) counts based on LSD *post hoc* tests.

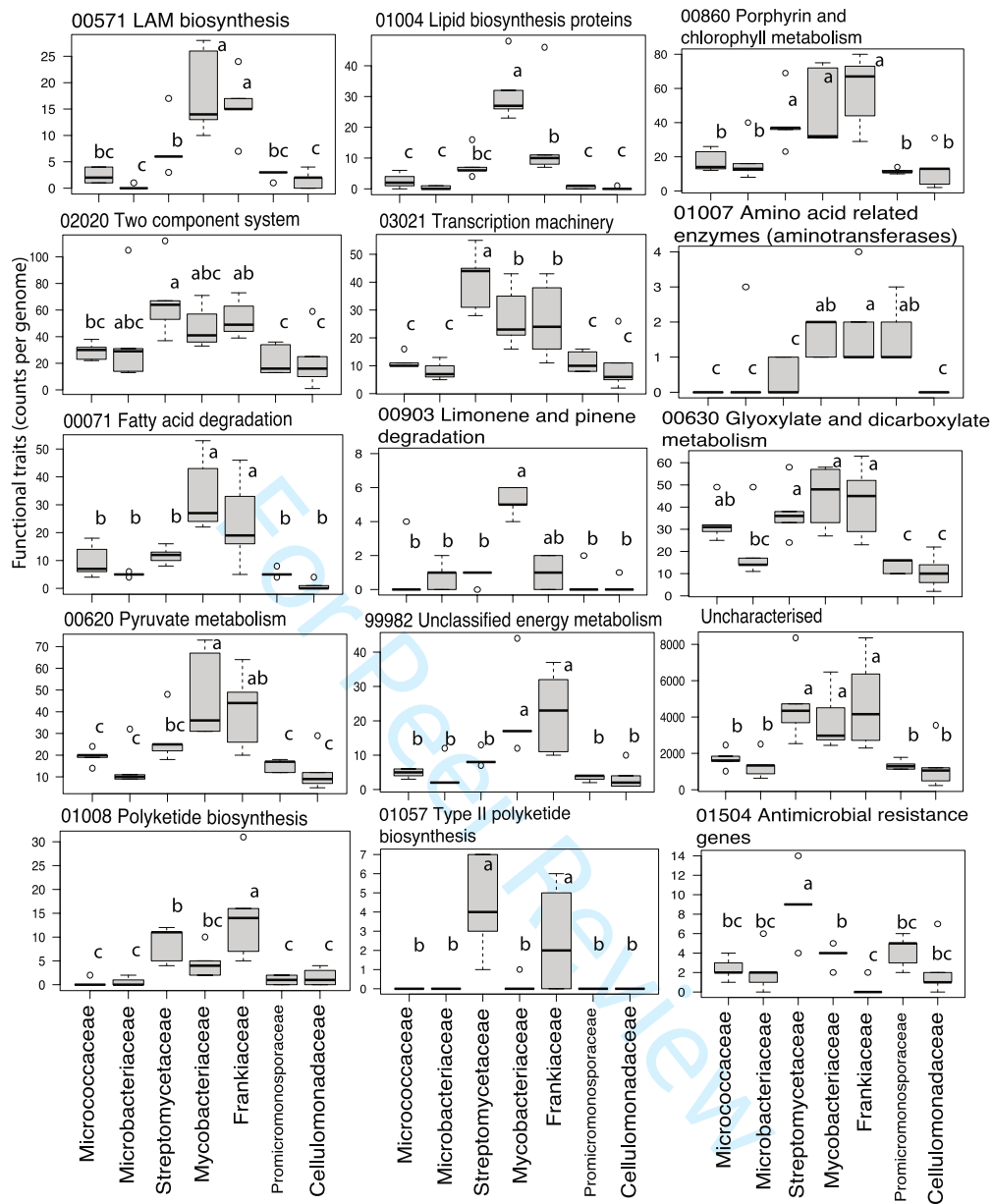




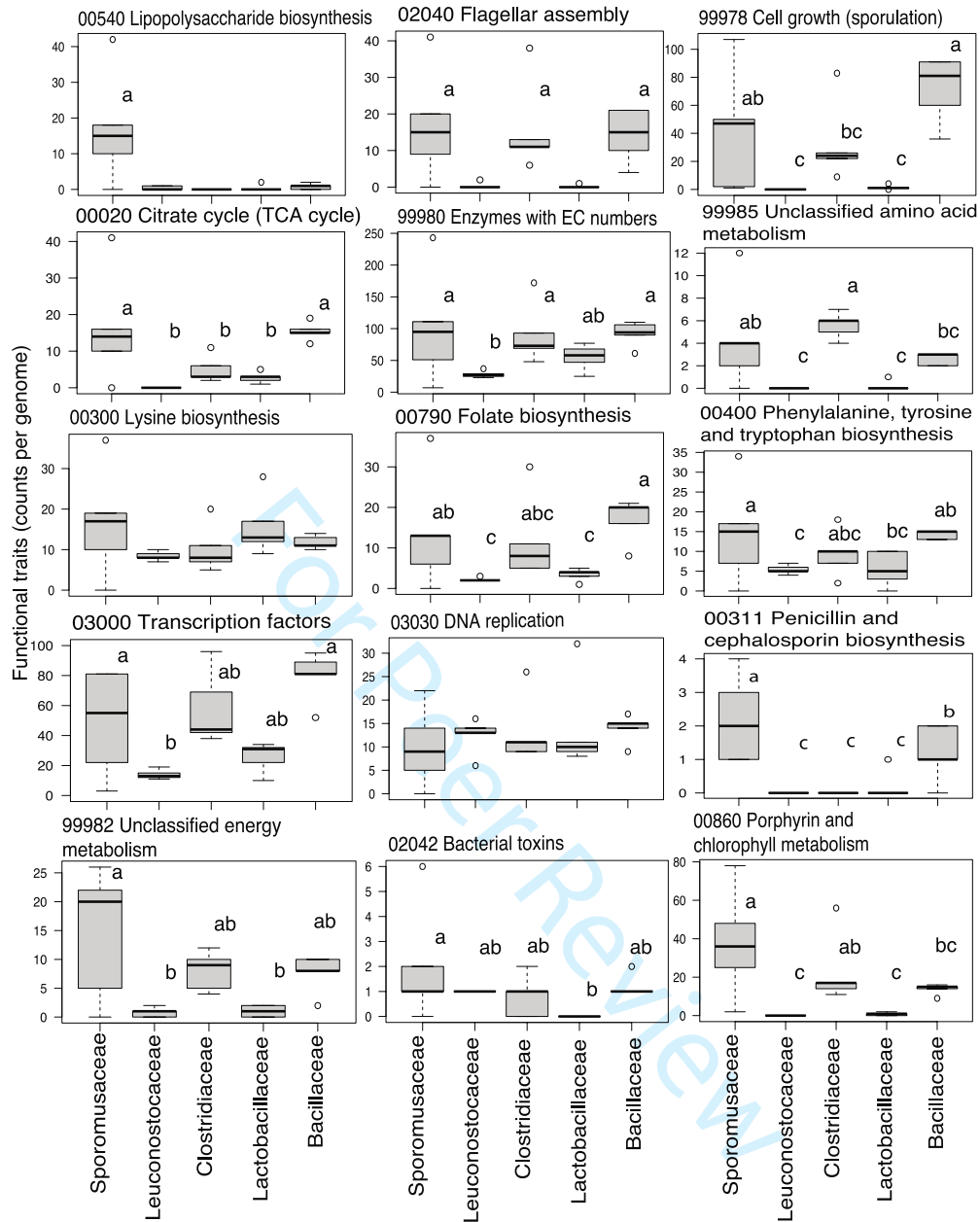
Supplementary Figure 8: Important predictor traits in classifying subgroups of Families as selected by the Proteobacteria, Actinobacteria, Firmicutes and Others model. Traits are ranked in order of decreasing Mean Decrease in Accuracy (MDA) of the four models.



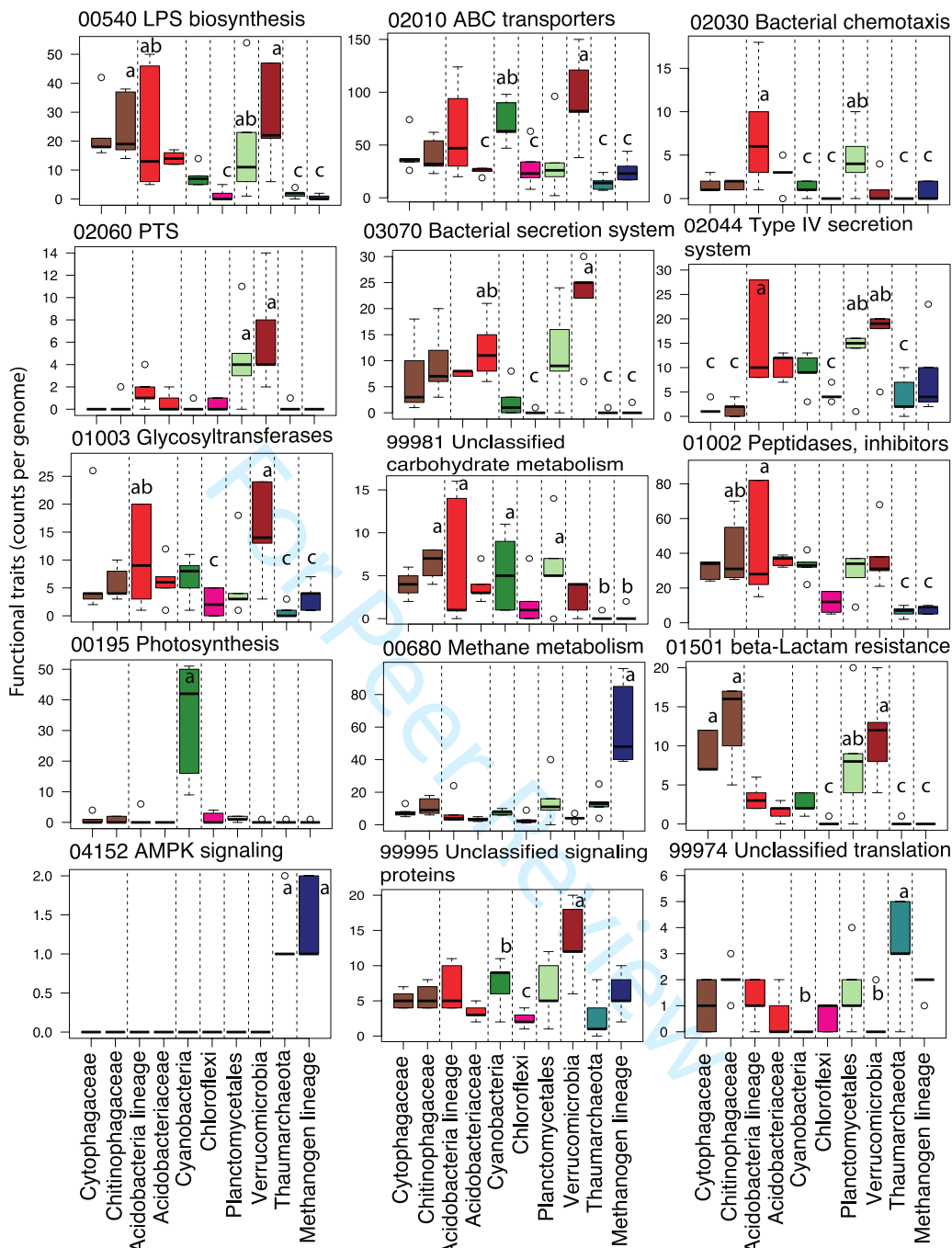
Supplementary Figure 9: Box and whisker plots of functional traits present in Proteobacterial Families. Traits were selected based on being selected by the Proteobacteria Family Random Forest model. Letters indicate significantly high or low ( $p < 0.05$ ) counts based on LSD *post hoc* tests.



Supplementary Figure 10: Box and whisker plots of functional traits present in Actinobacterial Families. Traits were selected based on being selected by the Actinobacterial Family Random Forest model. Letters indicate significantly high or low ( $p < 0.05$ ) counts based on LSD *post hoc* tests.



Supplementary Figure 11: Box and whisker plots of functional traits present in Firmicutes Families. Traits were selected based on being selected by the Firmicutes Family Random Forest model. Letters indicate significantly high or low ( $p < 0.05$ ) counts based on LSD *post hoc* tests.



Supplementary Figure 12: Box and whisker plots of functional traits present in Under-represented Families. Traits were selected based on being selected by the Under-represented Family Random Forest model. Letters indicate significantly high or low ( $p < 0.05$ ) counts based on LSD *post hoc* tests.

Supplementary Table 9: Summary of traits significantly enriched and depleted in the Phylum Random Forest model.

Model	BRITE 1	BRITE 3	Significantly enriched	Significantly depleted
Phylum	Metabolism	00020 TCA cycle	Gammaproteobacteria, Deltaproteobacteria, Betaproteobacteria, Actinobacteria, Bacteroidetes, Verrucomicrobia	Firmicutes, Euryarchaeota
		00190 Oxidative phosphorylation	Betaproteobacteria, Actinobacteria, Bacteroidetes	Firmicutes, Chloroflexi, Euryarchaeota
		00480 Glutathione metabolism	Alphaproteobacteria, Deltaproteobacteria, Cyanobacteria	Firmicutes, Chloroflexi, Planctomycetales, Verrucomicrobia, Thaumarchaeota, Euryarchaeota
		00910 Nitrogen metabolism	Gammaproteobacteria, Alphaproteobacteria, Actinobacteria, Cyanobacteria	Chloroflexi
		00540 LPS biosynthesis	Gammaproteobacteria, Bacteroidetes, Verrucomicrobia	Firmicutes, Actinobacteria, Chloroflexi, Thaumarchaeota, Euryarchaeota
		01011 Peptidoglycan biosynthesis, degradation	Gammaproteobacteria, Deltaproteobacteria, Betaproteobacteria	Firmicutes, Chloroflexi, Planctomycetales, Thaumarchaeota, Euryarchaeota
		01001 Protein kinases	Deltaproteobacteria, Planctomycetales	Alphaproteobacteria, Firmicutes, Thaumarchaeota, Euryarchaeota
		01040 Biosynthesis of unsaturated fatty acids	Deltaproteobacteria, Actinobacteria	Firmicutes, Chloroflexi, Thaumarchaeota, Euryarchaeota
		09113 Global maps only (unclassified metabolism)	Gammaproteobacteria, Actinobacteria, Thaumarchaeota	Firmicutes, Bacteroidetes, Acidobacteria, Cyanobacteria, Chloroflexi, Planctomycetales, Verrucomicrobia
		04614 Renin-angiotensin system	Deltaproteobacteria, Bacteroidetes	Gammaproteobacteria, Firmicutes, Thaumarchaeota
		00572 Arabinogalactan biosynthesis	Actinobacteria, Planctomycetales	Gammaproteobacteria, Alphaproteobacteria, Chloroflexi, Euryarchaeota
		00561 Glycerolipid metabolism	Deltaproteobacteria, Actinobacteria, Verrucomicrobia	Thaumarchaeota, Euryarchaeota
		00360 Phenylalanine metabolism	Gammaproteobacteria, Actinobacteria	Firmicutes, Acidobacteria, Cyanobacteria, Chloroflexi
		00960 Tropane, piperidine and pyridine alkaloid biosynthesis	Alphaproteobacteria, Betaproteobacteria, Euryarchaeota	Firmicutes, Planctomycetales
		00630 Glyoxylate and dicarboxylate metabolism	Alphaproteobacteria, Acidobacteria	Firmicutes, Chloroflexi, Thaumarchaeota, Euryarchaeota
	Environmental Information Processing	02010 ABC transporters	Alphaproteobacteria, Actinobacteria	Thaumarchaeota
		02020 Two component system	Alphaproteobacteria, Deltaproteobacteria	Chloroflexi, Thaumarchaeota
		03070 Bacterial secretion system	Gammaproteobacteria, Deltaproteobacteria, Betaproteobacteria, Verrucomicrobia	Firmicutes, Actinobacteria, Cyanobacteria, Chloroflexi, Thaumarchaeota, Euryarchaeota
		02044 Type IV secretion system	Gammaproteobacteria, Deltaproteobacteria, Betaproteobacteria, Verrucomicrobia	Firmicutes, Bacteroidetes, Chloroflexi, Thaumarchaeota
	Genetic Information Processing	03013 RNA transport (eukaryote-like)	Thaumarchaeota, Euryarchaeota	All other phyla



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

	03022 Basal transcription factors	Thaumarchaeota, Euryarchaeota	All other phyla
	03050 Proteasome	Actinobacteria, Thaumarchaeota, Euryarchaeota	All other phyla
	03110 Chaperones and folding	Gammaproteobacteria, Deltaproteobacteria	Firmicutes, Chloroflexi
	03051 proteasome 20S subunit	Deltaproteobacteria, Actinobacteria	All other phyla
	03019 mRNA biogenesis (degradation)	Gammaproteobacteria, Deltaproteobacteria	Alphaproteobacteria, Cyanobacteria, Chloroflexi, Thaumarchaeota, Euryarchaeota
Signaling and Cellular Processes	99978 Cell growth (sporulation)	Firmicutes	All other phyla
	01501 beta-Lactam resistance	Bacteroidetes, Verrucomicrobia	Firmicutes, Actinobacteria, Acidobacteria, Chloroflexi, Thaumarchaeota, Euryarchaeota
	02040 Flagellar assembly	Alphaproteobacteria, Betaproteobacteria, Acidobacteria	Actinobacteria, Bacteroidetes, Cyanobacteria, Chloroflexi, Verrucomicrobia, Thaumarchaeota, Euryarchaeota
	02030 bacterial chemotaxis	Alphaproteobacteria, Betaproteobacteria	Actinobacteria, Chloroflexi, Thaumarchaeota
Not included in pathway or BRITE	99992 Unclassified structural proteins	Gammaproteobacteria, Betaproteobacteria	Firmicutes, Actinobacteria, Chloroflexi, Thaumarchaeota, Euryarchaeota

Supplementary Table 10: Summary of traits significantly enriched and depleted in Firmicutes and Actinobacteria models.

Model	BRITE 1	BRITE 3	Significantly high	Significantly low	
Firmicutes	Metabolism	00540 LPS biosynthesis	Sporomusaceae	All other families	
		00020 Citrate cycle (TCA cycle)	Sporomusaceae, Bacillaceae	All other families	
		99985 Unclassified amino acid metabolism	Clostridiaceae	Leuconostocaceae, Lactobacillaceae	
		00300 Lysine biosynthesis	All equal	All equal	
		00790 Folate biosynthesis	Bacillaceae	Leuconostocaceae, Lactobacillaceae	
		00400 Phenylalanine, tyrosine and tryptophan biosynthesis	Sporomusaceae	Leuconostocaceae	
		99982 Unclassified energy metabolism	Sporomusaceae	Leuconostocaceae, Lactobacillaceae	
		00860 Porphyrin and chlorophyll metabolism	Sporomusaceae	Leuconostocaceae, Lactobacillaceae	
		Genetic Information Processing	03000 Transcription factors	Sporomusaceae, Bacillaceae	Leuconostocaceae
			03030 DNA replication	All equal	All equal
		Signaling and Cellular Processes	02040 Flagellar assembly	Sporomusaceae, Clostridiaceae, Bacillaceae	Leuconostocaceae, Lactobacillaceae
			99978 Cell growth (sporulation)	Bacillaceae	Leuconostocaceae, Lactobacillaceae
			02042 Bacterial toxins	Sporomusaceae	Lactobacillaceae
			00311 Penicillin and cephalosporin biosynthesis	Sporomusaceae, Bacillaceae	Leuconostocaceae, Lactobacillaceae, Clostridiaceae
Enzymes	99980 Enzymes with EC numbers	Sporomusaceae, Clostridiaceae, Bacillaceae	Leuconostocaceae		
Actinobacteria	Metabolism	00571 LAM biosynthesis	Mycobacteriaceae, Frankiaceae	Microbacteriaceae, Cellulomonadaceae	
		01004 Lipid biosynthesis proteins	Mycobacteriaceae	Micrococcaceae, Microbacteriaceae, Promicromonosporaceae, Cellulomonadaceae	
		00860 Porphyrin and chlorophyll metabolism	Streptomycetaceae, Mycobacteriaceae, Frankiaceae	All other families	
		01007 Amino acid related enzymes (aminotransferases)	Frankiaceae	Micrococcaceae, Microbacteriaceae, Streptomycetaceae, Cellulomonadaceae	
		00071 Fatty acid degradation	Mycobacteriaceae, Frankiaceae	All other families	
		00903 Limonene and pinene degradation	Mycobacteriaceae	Micrococcaceae, Microbacteriaceae, Streptomycetaceae, Promicromonosporaceae, Cellulomonadaceae	
		00630 Glyoxylate and dicarboxylate metabolism	Streptomycetaceae, Mycobacteriaceae, Frankiaceae	Promicromonosporaceae, Cellulomonadaceae	

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

		00620 Pyruvate metabolism	Mycobacteriaceae	Micrococcaceae, Microbacteriaceae, Promicromonosporaceae, Cellulomonadaceae
		99982 Unclassified energy metabolism	Mycobacteriaceae, Frankiaceae	All other families
		01008 Polyketide biosynthesis	Frankiaceae	Micrococcaceae, Microbacteriaceae, Promicromonosporaceae, Cellulomonadaceae
		01057 Type II polyketide biosynthesis	Streptomycetaceae, Frankiaceae	All other families
Environmental Information Processing		02020 Two component system	Streptomycetaceae	Promicromonosporaceae, Cellulomonadaceae
Genetic Information Processing		03021 Transcription machinery	Streptomycetaceae	Micrococcaceae, Microbacteriaceae, Promicromonosporaceae, Cellulomonadaceae
Signaling and Cellular Processes		01504 Antimicrobial resistance genes	Streptomycetaceae	Frankiaceae
	Uncharacterised	Uncharacterised	Streptomycetaceae, Mycobacteriaceae, Frankiaceae	All other families

Supplementary Table 11: Summary of traits significantly enriched and depleted in Proteobacteria and Others models.

Model	BRITE 1	BRITE 3	Significantly high	Significantly low	
Proteobacteria	Metabolism	00540 LPS biosynthesis	Pseudomonadaceae, Burkholderiaceae	Beijerinckiaceae, Nitrosomonadaceae	
		01003 Glycosyltransferases	Rhizobiaceae, Myxococcaceae, Polyangiaceae	Moraxellaceae	
		00650 Butanoate metabolism	Rhodospirillaceae, Rhizobiaceae, Burkholderiaceae	Methylococcaceae, Nitrosomonadaceae	
		00600 Sphingolipid metabolism	Myxococcaceae	All Gammaproteobacteria, Rhodospirillaceae, Bradyrhizobiaceae, Neisseriales Lineage, Nitrosomonadaceae	
		01007 Amino acid related enzymes (aminotransferases)	Pseudomonadaceae, Rhodospirillaceae, Polyangiaceae, Burkholderiaceae	Myxococcaceae	
		Environmental Information Processing	02010 ABC transporters	Pseudomonadaceae, Rhizobiaceae, Burkholderiaceae	Moraxellaceae, Methylococcaceae, Beijerinckiaceae, Nitrosomonadaceae
			03070 Bacterial secretion systems	Myxococcaceae, Polyangiaceae	Methylocystaceae, Rhodospirillaceae, Bradyrhizobiaceae, Beijerinckiaceae, Nitrosomonadaceae
		Genetic Information Processing	03021 Transcription machinery	Polyangiaceae	All other families
			99974 Unclassified translation	Pseudomonadaceae, Beijerinckiaceae	Moraxellaceae, Neisseriales Lineage
			03019 Messenger RNA biogenesis (degradation)	Pseudomonadaceae, Myxococcaceae, Burkholderiaceae	Methylocystaceae, Rhizobiaceae, Bradyrhizobiaceae, Beijerinckiaceae
			03009 Ribosome biogenesis	Myxococcaceae	Methylococcaceae, all Alphaproteobacteria, Nitrosomonadaceae
		Signaling and Cellular Processes	03013 RNA transport	Pseudomonadaceae, Myxococcaceae	Rhodospirillaceae, Rhizobiaceae, Beijerinckiaceae
			02030 Bacterial chemotaxis	Rhodospirillaceae	Moraxellaceae
			02040 Flagellar assembly	Rhodospirillaceae, Neisseriales Lineage, Nitrosomonadaceae	
99995 Unclassified signaling proteins	Pseudomonadaceae		Methylocystaceae, Bradyrhizobiaceae, Beijerinckiaceae		
Other	Metabolism	00540 LPS biosynthesis	Chitinophagaceae, Acidobacteria Lineage, Planctomycetales, Verrucomicrobia	Chloroflexi, Thaumarchaeota, Methanogen Lineage	
		01003 Glycosyltransferases	Acidobacteria Lineage, Verrucomicrobia	Chloroflexi, Thaumarchaeota, Methanogen Lineage	
		99981 Unclassified carbohydrate metabolism	Chitinophagaceae, Acidobacteria Lineage, Planctomycetales	Thaumarchaeota, Methanogen Lineage	
		01002 Peptidases, inhibitors	Chitinophagaceae, Acidobacteria Lineage	Thaumarchaeota, Methanogen Lineage	
		00195 Photosynthesis	Cyanobacteria	All other families	
		00680 Methane metabolism	Methanogen Lineage	All other families	
		Environmental Information Processing	02010 ABC transporters	Cyanobacteria, Verrucomicrobia	Acidobacteriaceae, Chloroflexi, Thaumarchaeota, Methanogen Lineage

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

	02060 PTS	Planctomycetales, Verrucomicrobia	All other families
	04152 AMPK signaling	Thaumarchaeota, Methanogen Lineage	All other families
	03070 Bacterial secretion system	Acidobacteriaceae, Verrucomicrobia	Chloroflexi, Thaumarchaeota, Methanogen Lineage
	02044 Type IV Secretion system	Acidobacteria Lineage, Planctomycetales, Verrucomicrobia	Cytophagaceae, Chitinophagaceae, Chloroflexi, Thaumarchaeota
Genetic Information Processing	99974 Unclassified translation	Thaumarchaeota	Cyanobacteria, Verrucomicrobia
Signaling and Cellular Processes	02030 Bacterial chemotaxis	Acidobacteria Lineage, Planctomycetales	Cyanobacteria, Chloroflexi, Thaumarchaeota, Methanogen Lineage
	01501 beta-Lactam resistance	Cytophagaceae, Chitinophagaceae, Planctomycetales, Verrucomicrobia	Chloroflexi, Thaumarchaeota, Methanogen Lineage
	99995 Unclassified signaling proteins	Cyanobacteria	Chloroflexi

## References

Abt B, Foster B, Lapidus A, Clum A, Sun H, Pukall R *et al* (2010). Complete genome sequence of *Cellulomonas flavigena* type strain (134(T)). *Standards in Genomic Sciences* **3**: 15-25.

An DS, Lee HG, Im WT, Liu OM, Lee ST (2007a). *Segetibacter koreensis* gen. nov., sp nov., a novel member of the phylum Bacteroidetes, isolated from the soil of a ginseng field in South Korea. *International Journal of Systematic and Evolutionary Microbiology* **57**: 1828-1833.

An SY, Asahara M, Goto K, Kasai H, Yokota A (2007b). *Terribacillus saccharophilus* gen. nov., sp nov and *Terribacillus halophilus* sp nov., spore-forming bacteria isolated from field soil in Japan. *International Journal of Systematic and Evolutionary Microbiology* **57**: 51-55.

Anagnostidis K, Komarek J (1988). Modern approach to the classification system of cyanophytes. 3. Oscillatoriales. *Archiv für Hydrobiologie, Supplement* **80**: 327-472.

Bagnara C, Toci R, Gaudin C, Belaich JP (1985). Isolation and characterization of a cellulolytic microorganism, *Cellulomonas-fermentans* sp-nov. *International Journal of Systematic Bacteriology* **35**: 502-507.

1  
2  
3  
4  
5  
6  
7 Barrangou R, Yoon SS, Breidt F, Fleming HP, Klaenhammer TR (2002). Identification and characterization of *Leuconostoc fallax* strains  
8 isolated from an industrial sauerkraut fermentation. *Applied and Environmental Microbiology* **68**: 2877-2884.

9  
10 Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD *et al* (2002). Complete genome sequence of the model  
11 actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**: 141-147.

12  
13 Bogan BW, Sullivan WR, Kayser KJ, Derr KD, Aldrich HC, Peterek JR (2003). *Alkanindiges illinoisensis* gen. nov., sp nov., an obligately  
14 hydrocarbonoclastic, aerobic squalane-degrading bacterium isolated from oilfield soils. *International Journal of Systematic and*  
15 *Evolutionary Microbiology* **53**: 1389-1395.

16  
17 Brauer SL, Cadillo-Quiroz H, Ward RJ, Yavitt JB, Zinder SH (2011). *Methanoregula boonei* gen. nov., sp. nov., an acidiphilic methanogen  
18 isolated from an acidic peat bog. *International Journal of Systematic and Evolutionary Microbiology* **61**: 45-52.

19  
20 Brook MD, Currie B, Desmarchelier PM (1997). Isolation and identification of *Burkholderia pseudomallei* from soil using selective  
21 culture techniques and the polymerase chain reaction. *Journal of Applied Microbiology* **82**: 589-596.

22  
23 Brunel B, Cleyetmarel JC, Normand P, Bardin R (1988). Stability of *Bradyrhizobium-japonicum* inoculants after introduction into soil.  
24 *Applied and Environmental Microbiology* **54**: 2636-2642.

25  
26 Cadillo-Quiroz H, Yavitt JB, Zinder SH (2009). *Methanosphaerula palustris* gen. nov., sp nov., a hydrogenotrophic methanogen isolated  
27 from a minerotrophic fen peatland. *International Journal of Systematic and Evolutionary Microbiology* **59**: 928-935.

28  
29 Chain PSG, Denev VJ, Konstantinidis KT, Vergez LM, Agullo L, Reyes VL *et al* (2006). *Burkholderia xenovorans* LB400 harbors a multi-  
30 replicon, 9.73-Mbp genome shaped for versatility. *Proceedings of the National Academy of Sciences of the United States of America* **103**:  
31 15280-15287.

32  
33 Chen WM, Laevens S, Lee TM, Coenye T, De Vos P, Mergeay M *et al* (2001). *Ralstonia taiwanensis* sp nov., isolated from root nodules of  
34 *Mimosa* species and sputum of a cystic fibrosis patient. *International Journal of Systematic and Evolutionary Microbiology* **51**: 1729-1735.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

Chen XH, Koumoutsi A, Scholz R, Eisenreich A, Schneider K, Heinemeyer I *et al* (2007). Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42. *Nature Biotechnology* **25**: 1007-1014.

Chen YS, Yanagida F, Shinohara T (2005). Isolation and identification of lactic acid bacteria from soil using an enrichment procedure. *Letters in Applied Microbiology* **40**: 195-200.

Chern LL, Stackebrandt E, Lee SF, Lee FL, Chen JK, Fu HM (2004). *Chitinibacter tainanensis* gen. nov., sp. nov., a chitin-degrading aerobe from soil in Taiwan. *International Journal of Systematic and Evolutionary Microbiology* **54**: 1387-1391.

Chin KJ, Liesack W, Janssen PH (2001). *Opiritatus terrae* gen nov, sp nov, to accommodate novel strains of the division 'Verrucomicrobia' isolated from rice paddy soil. *International Journal of Systematic and Evolutionary Microbiology* **51**: 1965-1968.

Chistoserdova L (2011). Methylotrophy in a Lake: from Metagenomics to Single-Organism Physiology. *Applied and Environmental Microbiology* **77**: 4705-4711.

Cohen JE, Wang R, Shen RF, Wu WW, Keller JE (2017). Comparative pathogenomics of *Clostridium tetani*. *Plos One* **12**.

Crowe MA, Power JF, Morgan XC, Dunfield PF, Lagutin K, Rijpstra IC *et al* (2014). *Pyrinomonas methylaliphatogenes* gen. nov., sp. nov., a novel group 4 thermophilic member of the phylum Acidobacteria from geothermal soils (vol 64, pg 220, 2014). *International Journal of Systematic and Evolutionary Microbiology* **64**: 1074-1074.

Das S, Pettersson BMF, Behra PRK, Ramesh M, Dasgupta S, Bhattacharya A *et al* (2015). Characterization of Three *Mycobacterium* spp. with Potential Use in Bioremediation by Genome Sequencing and Comparative Genomics. *Genome Biology and Evolution* **7**: 1871-1886.

Davenport KW, Daligault HE, Minogue TD, Bruce DC, Chain PSG, Coyne SR *et al* (2014). Draft genome assembly of *Acinetobacter baumannii* ATCC 19606. *Genome Announcements* **2**: e00832-00814.



1  
2  
3  
4  
5  
6 Dedysh SN, Khmelenina VN, Suzina NE, Trotsenko YA, Semrau JD, Liesack W *et al* (2002). *Methylocapsa acidiphila* gen. nov., sp nov., a  
7 novel methane-oxidizing and dinitrogen-fixing acidophilic bacterium from Sphagnum bog. *International Journal of Systematic and*  
8 *Evolutionary Microbiology* **52**: 251-261.  
9

10 Dedysh SN, Kulichevskaya IS, Serkebaeva YM, Mityaeva MA, Sorokin VV, Suzina NE *et al* (2012). *Bryocella elongata* gen. nov., sp nov., a  
11 member of subdivision 1 of the Acidobacteria isolated from a methanotrophic enrichment culture, and emended description of  
12 *Edaphobacter aggregans* Koch *et al.* 2008. *International Journal of Systematic and Evolutionary Microbiology* **62**: 654-664.  
13  
14

15 Dees MW, Sletten A, Hermansen A (2013). Isolation and characterization of *Streptomyces* species from potato common scab lesions in  
16 Norway. *Plant Pathology* **62**: 217-225.  
17

18 De Leon KB, Young ML, Camilleri LB, Brown SD, Skerker JM, Deutschbauer AM *et al* (2012). Draft genome sequence of *Pelosinus*  
19 *fermentans* JBW45, isolated during in situ stimulation for Cr(VI) reduction. *Journal of Bacteriology* **194**: 5456-5457.  
20

21 Derx HG (1950). *Beijerinckia*, a new genus of nitrogen-fixing bacteria in tropical soils. *Repr Meded Konink Ned Akad Wetens* **53**: 140-147.  
22  
23

24 Dunfield PF, Khmelenina VN, Suzina NE, Trotsenko YA, Dedysh SN (2003). *Methylocella silvestris* sp nov., a novel methanotroph isolated  
25 from an acidic forest cambisol. *International Journal of Systematic and Evolutionary Microbiology* **53**: 1231-1239.  
26

27 Eichorst SA, Breznak JA, Schmidt TM (2007). Isolation and characterization of soil bacteria that define *Teniglobus* gen. nov., in the  
28 phylum Acidobacteria. *Applied and Environmental Microbiology* **73**: 2708-2717.  
29

30 Endo A, Okada S (2007). *Lactobacillus composti* sp nov, a lactic acid bacterium isolated from a compost of distilled shochu residue.  
31 *International Journal of Systematic and Evolutionary Microbiology* **57**: 870-872.  
32  
33

34 Endo A, Futagawa-Endo Y, Dicks LMT (2009). Isolation and characterization of fructophilic lactic acid bacteria from fructose-rich niches.  
35 *Systematic and Applied Microbiology* **32**: 593-600.  
36  
37  
38  
39  
40  
41  
42  
43  
44

1  
2  
3  
4  
5  
6 Endo A, Futagawa-Endo Y, Sakamoto M, Kitahara M, Dicks LMT (2010). *Lactobacillus florum* sp. nov., a fructophilic species isolated from  
7 flowers. *International Journal of Systematic and Evolutionary Microbiology* **60**: 2478-2482.

8  
9  
10 Geymonat E, Ferrando L, Tarlera SE (2011). *Methylogaea oryzae* gen. nov., sp nov., a mesophilic methanotroph isolated from a rice  
11 paddy field. *International Journal of Systematic and Evolutionary Microbiology* **61**: 2568-2572.

12  
13 Graham JB, Istock CA (1978). Genetic exchange in *Bacillus-subtilis* in soil. *Molecular & General Genetics* **166**: 287-290.

14  
15 Hamada M, Shibata C, Sakurai K, Hosoyama A, Oji S, Teramoto K *et al* (2016). Reclassification of *Amycolicococcus subflavus* as *Hoyosella*  
16 *subflava* comb. nov and emended descriptions of the genus *Hoyosella* and *Hoyosella altamirensis*. *International Journal of Systematic and*  
17 *Evolutionary Microbiology* **66**: 4711-4715.

18  
19  
20 Hatayama K, Kawai S, Shoun H, Ueda Y, Nakamura A (2005). *Pseudomonas azotifigens* sp nov., a novel nitrogen-fixing bacterium isolated  
21 from a compost pile. *International Journal of Systematic and Evolutionary Microbiology* **55**: 1539-1544.

22  
23 He J, Ritalahti KM, Yang KL, Koenigsberg SS, Loffler FE (2003). Detoxification of vinyl chloride to ethene coupled to growth of an  
24 anaerobic bacterium. *Nature* **424**: 62-65.

25  
26 Helgason E, Okstad OA, Caugant DA, Johansen HA, Fouet A, Mock M *et al* (2000). *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus*  
27 *thuringiensis* - One species on the basis of genetic evidence. *Applied and Environmental Microbiology* **66**: 2627-2630.

28  
29  
30 Hennessee CT, Seo JS, Alvarez AM, Li QX (2009). Polycyclic aromatic hydrocarbon-degrading species isolated from Hawaiian soils:  
31 *Mycobacterium crocinum* sp nov., *Mycobacterium pallens* sp nov., *Mycobacterium rutilum* sp nov., *Mycobacterium rufum* sp nov and  
32 *Mycobacterium aromaticivorans* sp nov. *International Journal of Systematic and Evolutionary Microbiology* **59**: 378-387.

33  
34 Huang K, Xu Y, Zhang J, Chen C, Gao F, Zhao FJ (2017). *Arsenicibacter rosenii* gen. nov., sp. nov., an efficient arsenic methylating and  
35 volatilizing bacterium isolated from an arsenic-contaminated paddy soil. *International Journal of Systematic and Evolutionary*  
36 *Microbiology* **67**: 3186-3191.

1  
2  
3  
4  
5  
6  
7 Huntley S, Zhang Y, Treuner-Lange A, Kneip S, Sensen CW, Sogaard-Andersen L (2012). Complete Genome Sequence of the Fruiting  
8 Myxobacterium *Coralloccoccus coralloides* DSM 2259. *Journal of Bacteriology* **194**: 3012-3013.

9  
10 Hwang JY, Kim SH, Oh HR, Kwon E, Nam DH (2015). Analysis of a draft genome sequence of *Kitasatospora cheerisanensis* KCTC 2395  
11 producing bafilomycin antibiotics. *Journal of Microbiology* **53**: 84-89.

12  
13 Iguchi H, Yurimoto H, Sakai Y (2011). *Methylovulum miyakonense* gen. nov., sp. nov., a type I methanotroph isolated from forest soil.  
14 *International Journal of Systematic and Evolutionary Microbiology* **61**: 810-815.

15  
16 Im WT, Kim SH, Kim MK, Ten LN, Lee ST (2006). *Pleomorphomonas koreensis* sp nov., a nitrogen-fixing species in the order Rhizobiales.  
17 *International Journal of Systematic and Evolutionary Microbiology* **56**: 1663-1666.

18  
19 Jahn E (1911). Myxobacteriales. In: *Kryptogamenflora der Mark Brandenburg und angrenzender Gebiete, Vol V, pp 187-206 Leipzig:*  
20 *Borntraeger.*

21  
22  
23 Johnson JL, Toth J, Santiwatanakul S, Chen JS (1997). Cultures of "*Clostridium acetobutylicum*" from various collections comprise  
24 *Clostridium acetobutylicum*, *Clostridium beijerinckii*, and two other distinct types based on DNA-DNA reassociation. *International*  
25 *Journal of Systematic Bacteriology* **47**: 420-424.

26  
27 Jung MY, Kim JG, Damste JSS, Rijpstra WIC, Madsen EL, Kim SJ *et al* (2016). A hydrophobic ammonia-oxidizing archaeon of the  
28 *Nitrosocosmicus* clade isolated from coal tar-contaminated sediment. *Environmental Microbiology Reports* **8**: 983-992.

29  
30 Jurado V, Groth I, Gonzalez JM, Laiz L, Saiz-Jimenez C (2005). *Agromyces subbeticus* sp. nov., isolated from a cave in southern Spain.  
31 *International Journal of Systematic and Evolutionary Microbiology* **55**: 1897-1901.

32  
33  
34 Kawasaki S, Kurosawa K, Miyazaki M, Yagi C, Kitajima Y, Tanaka S *et al* (2011). *Lactobacillus floricola* sp nov., lactic acid bacteria isolated  
35 from mountain flowers. *International Journal of Systematic and Evolutionary Microbiology* **61**: 1356-1359.

1  
2  
3  
4  
5  
6  
7 Kim BK, Jung MY, Yu DS, Park SJ, Oh TK, Rhee SK *et al* (2011). Genome Sequence of an Ammonia-Oxidizing Soil Archaeon, "Candidatus Nitrosoarchaeum koreensis" MY1. *Journal of Bacteriology* **193**: 5539-5540.

8  
9  
10 Kim BY, Weon HY, Yoo SH, Chen WM, Kwon SW, Go SJ *et al* (2006). Chitinimonas koreensis sp nov., isolated from greenhouse soil in Korea. *International Journal of Systematic and Evolutionary Microbiology* **56**: 1761-1764.

11  
12  
13 Kim D, Baik KS, Kim MS, Park SC, Kim SS, Rhee MS *et al* (2008). Acinetobacter soli sp nov., isolated from forest soil. *Journal of Microbiology* **46**: 396-401.

14  
15  
16 Kim SB, Lonsdale J, Seong CN, Goodfellow M (2003). Streptacidiphilus gen. nov., acidophilic actinomycetes with wall chemotype I and emendation of the family Streptomycetaceae (Waksman and Henrici (1943)(AL)) emend. Rainey *et al*. 1997. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology* **83**: 107-116.

17  
18  
19  
20  
21 Kiss H, Nett M, Domin N, Martin K, Maresca JA, Copeland A *et al* (2011). Complete genome sequence of the filamentous gliding predatory bacterium Herpetosiphon aurantiacus type strain (114-95<sup>T</sup>). *Standards in Genomic Sciences* **5**: 356-370.

22  
23  
24 Kleinig H, Reichenbach H (1973). New carotenoid glucoside ester from Chondromyces-apiculatus. *Phytochemistry* **12**: 2483-2485.

25  
26 Klouche N, Fardeau ML, Lascourreges JF, Cayol JL, Hacene H, Thomas P *et al* (2007). Geosporobacter subterraneus gen. nov., sp nov., a spore-forming bacterium isolated from a deep subsurface aquifer. *International Journal of Systematic and Evolutionary Microbiology* **57**: 1757-1761.

27  
28  
29  
30  
31 Koburger JA, May SO (1982). Isolation of Chromobacterium spp from foods, soil, and water. *Applied and Environmental Microbiology* **44**: 1463-1465.

32  
33  
34 Koch IH, Gich F, Dunfield PF, Overmann J (2008). Edaphobacter modestus gen. nov., sp nov., and Edaphobacter aggregans sp nov., acidobacteria isolated from alpine and forest soils. *International Journal of Systematic and Evolutionary Microbiology* **58**: 1114-1122.

1  
2  
3  
4  
5  
6 Komarek J, Hindak F (1975). Taxonomy of the new isolated strains of Chroococciopsis (Cyanophyceae). *Archiv für Hydrobiologie* **13**:  
7 311-329.  
8

9  
10 Koops HP, Bottcher B, Moller UC, Pommereningroser A, Stehr G (1991). Classification of 8 new species of Ammonia-Oxidizing bacteria -  
11 Nitrosomonas-communis sp-nov, Nitrosomonas-ureae sp-nov, Nitrosomonas-aestuarii sp-nov, Nitrosomonas-marina sp-nov,  
12 Nitrosomonas-nitrosa sp-nov, Nitrosomonas-eutropha sp-nov, Nitrosomonas-oligotropha sp-nov and Nitrosomonas-halophila sp-nov.  
13 *Journal of General Microbiology* **137**: 1689-1699.  
14

15 Krizova L, Maixnerova M, Sedo O, Nemeč A (2014). Acinetobacter bohemicus sp nov widespread in natural soil and water ecosystems in  
16 the Czech Republic. *Systematic and Applied Microbiology* **37**: 467-473.  
17

18 Kuklinsky-Sobral J, Araujo WL, Mendes R, Geraldi IO, Pizzirani-Kleiner AA, Azevedo JL (2004). Isolation and characterization of soybean-  
19 associated bacteria and their potential for plant growth promotion. *Environmental Microbiology* **6**: 1244-1251.  
20

21 Kulichevskaya IS, Ivanova AO, Belova SE, Baulina OI, Bodelier PLE, Rijkstra WIC *et al* (2007). Schlesneria paludicola gen. nov., sp nov.,  
22 the first acidophilic member of the order Planctomycetales, from Sphagnum-dominated boreal wetlands. *International Journal of*  
23 *Systematic and Evolutionary Microbiology* **57**: 2680-2687.  
24  
25

26 Kulichevskaya IS, Ivanova AO, Baulina OI, Bodelier PLE, Damste JSS, Dedysh SN (2008). Singulisphaera acidiphila gen. nov., sp nov., a  
27 non-filamentous, Isosphaera-like planctomycete from acidic northern wetlands. *International Journal of Systematic and Evolutionary*  
28 *Microbiology* **58**: 1186-1193.  
29

30 Kulichevskaya IS, Suzina NE, Liesack W, Dedysh SN (2010). Bryobacter aggregatus gen. nov., sp nov., a peat-inhabiting, aerobic chemo-  
31 organotroph from subdivision 3 of the Acidobacteria. *International Journal of Systematic and Evolutionary Microbiology* **60**: 301-306.  
32  
33

34 Kulichevskaya IS, Ivanova AA, Detkova EN, Rijkstra WIC, Damste JSS, Dedysh SN (2015). Planctomicrobium piriforme gen. nov., sp nov., a  
35 stalked planctomycete from a littoral wetland of a boreal lake. *International Journal of Systematic and Evolutionary Microbiology* **65**:  
36 1659-1665.  
37  
38  
39  
40  
41  
42  
43  
44

1  
2  
3  
4  
5  
6 Kulichevskaya IS, Ivanova AA, Suzina NE, Rijpstra WIC, Damste JSS, Dedysh SN (2016). Paludisphaera borealis gen. nov., sp nov., a  
7 hydrolytic planctomycete from northern wetlands, and proposal of Isosphaeraceae fam. nov. *International Journal of Systematic and*  
8 *Evolutionary Microbiology* **66**: 837-844.  
9

10 Kwon SW, Kim YY, Kim WG, Yoo KH, Yoo SH, Son JA *et al* (2008). Paludibacterium yongneupense gen. nov., sp nov., isolated from a  
11 wetland, Yongneup, in Korea. *International Journal of Systematic and Evolutionary Microbiology* **58**: 190-194.  
12  
13

14 Lang E, Kroppenstedt RM, Straubler B, Stackebrandt E (2008). Reclassification of Myxococcus flavescens Yamanaka et al. 1990(vp) as a  
15 later synonym of Myxococcus virescens Thaxter 1892(AL). *International Journal of Systematic and Evolutionary Microbiology* **58**: 2607-  
16 2609.  
17

18 Lang E, Stackebrandt E (2009). Emended descriptions of the genera Myxococcus and Coralloccoccus, typification of the species  
19 Myxococcus stipitatus and Myxococcus macrosporus and a proposal that they be represented by neotype strains. Request for an  
20 Opinion. *International Journal of Systematic and Evolutionary Microbiology* **59**: 2122-2128.  
21  
22

23 Larimer FW, Chain P, Hauser L, Lamerdin J, Malfatti S, Do L *et al* (2004). Complete genome sequence of the metabolically versatile  
24 photosynthetic bacterium Rhodospseudomonas palustris. *Nature Biotechnology* **22**: 55-61.  
25

26 Lehtovirta-Morley LE, Ge CR, Ross J, Yao HY, Nicol GW, Prosser JI (2014). Characterisation of terrestrial acidophilic archaeal ammonia  
27 oxidisers and their inhibition and stimulation by organic compounds. *Fems Microbiology Ecology* **89**: 542-552.  
28

29 Levanon D (1978). Isolation and properties of cell-walls from Agrobacterium-tumefaciens-B6. *Experientia* **34**: 354-356.  
30

31 Leys NM, Ryngaert A, Bastiaens L, Wattiau P, Top EM, Verstraete W *et al* (2005). Occurrence and community composition of fast-  
32 growing Mycobacterium in soils contaminated with polycyclic aromatic hydrocarbons. *Fems Microbiology Ecology* **51**: 375-388.  
33  
34

35 Li SG, Zhao L, Han K, Li PF, Li ZF, Hu W *et al* (2014). Diversity of epothilone producers among Sorangium strains in producer-positive  
36 soil habitats. *Microbial Biotechnology* **7**: 130-141.  
37  
38  
39  
40  
41  
42  
43  
44

1  
2  
3  
4  
5  
6  
7 Li YZ, Chen F, Dong K, Wang GJ (2013). *Actinotalea ferrariae* sp nov., isolated from an iron mine, and emended description of the genus  
8 *Actinotalea*. *International Journal of Systematic and Evolutionary Microbiology* **63**: 3398-3403.

9  
10 Liesack W, Bak F, Kreft JU, Stackebrandt E (1994). *Holophaga foetida* gen-nov, sp-nov, a new, homoacetogenic bacterium degrading  
11 methoxylated aromatic-compounds. *Archives of Microbiology* **162**: 85-90.

12  
13 Lipman JG (1904). Further contributions to the physiology and morphology of members of the *Azotobacter* group. *New Jersey*  
14 *Agricultural Experimental Station 25th Annual Report*, pp 237-289.

15  
16 Liu GQ, Kong YY, Fan YJ, Geng C, Peng DH, Sun M (2017). Whole-genome sequencing of *Bacillus velezensis* LS69, a strain with a broad  
17 inhibitory spectrum against pathogenic bacteria. *Journal of Biotechnology* **249**: 20-24.

18  
19 Liu JH, Wang YX, Zhang XX, Wang ZG, Chen YG, Wen ML *et al* (2010). *Salinarimonas rosea* gen. nov., sp nov., a new member of the alpha-2  
20 subgroup of the Proteobacteria. *International Journal of Systematic and Evolutionary Microbiology* **60**: 55-60.

21  
22 Lymperopoulou DS, Coil DA, Schichnes D, Lindow SE, Jospin G, Eisen JA *et al* (2017). Draft genome sequences of eight bacteria isolated  
23 from the indoor environment: *Staphylococcus capitis* strain H36, *S. capitis* strain H65, *S. cohnii* strain H62, *S. hominis* strain H69,  
24 *Microbacterium* sp strain H83, *Mycobacterium iranicum* strain H39, *Plantibacter* sp strain H53, and *Pseudomonas oryzihabitans* strain  
25 H72. *Standards in Genomic Sciences* **12**.

26  
27 Madhaiyan M, Poonguzhali S, Senthilkumar M, Pragatheswari D, Lee JS, Lee KC (2015). *Arachidicoccus rhizosphaerae* gen. nov., sp nov., a  
28 plant-growth-promoting bacterium in the family Chitinophagaceae isolated from rhizosphere soil. *International Journal of Systematic*  
29 *and Evolutionary Microbiology* **65**: 578-586.

30  
31 Maestrojuán GM, Boone JE, Mah RA, Mienaia JAGF, Sachs MS, Boone DR (1992). Taxonomy and halotolerance of mesophilic  
32 *Methanosarcina* strains, assignment of strains to species, and synonymy of *Methanosarcina mazei* and *Methanosarcina frisia*.  
33 *International Journal of Systematic Bacteriology* **42**: 561-567.



- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47
- Magalhaes FM, Baldani JI, Souto SM, Kuykendall JR, Dobereiner J (1983). A new acid-tolerant *Azospirillum* species. *Anais Da Academia Brasileira De Ciencias* **55**: 417-430.
- Magnusson J, Jonsson H, Schnurer J, Roos S (2002). *Weisselia soli* sp nov., a lactic acid bacterium isolated from soil. *International Journal of Systematic and Evolutionary Microbiology* **52**: 831-834.
- Malhotra J, Anand S, Jindal S, Rajagopal R, Lal R (2012). *Acinetobacter indicus* sp nov., isolated from a hexachlorocyclohexane dump site. *International Journal of Systematic and Evolutionary Microbiology* **62**: 2883-2890.
- Mao YJ, Chen M, Horvath P (2015). *Lactobacillus herbarum* sp nov., a species related to *Lactobacillus plantarum*. *International Journal of Systematic and Evolutionary Microbiology* **65**: 4682-4688.
- Mayilraj S, Suresh K, Schumann P, Kroppenstedt RM, Saini HS (2006). *Agrococcus lahaulensis* sp. nov., isolated from a cold desert of the Indian Himalayas. *International Journal of Systematic and Evolutionary Microbiology* **56**: 1807-1810.
- Möller B, Ossmer R, Howard BH, Gottschalk G, Hippe H (1984). *Sporomusa*, a new genus of Gram-negative anaerobic bacteria including *Sporomuusa sphaeroides* spec. nov. and *Sporomusa ovata* spec. nov. *Archives of Microbiology* **139**: 388-396.
- Mongodin EF, Shapir N, Daugherty SC, Deboy RT, Emerson JB, Shvartzbeyn A *et al* (2006). Secrets of soil survival revealed by the genome sequence of *Arthrobacter aurescens* TC1. *Plos Genetics* **2**: 2094-2106.
- Mosier AC, Allen EE, Kim M, Ferriera S, Francis CA (2012). Genome Sequence of "Candidatus *Nitrosopumilus salaria*" BD31, an Ammonia-Oxidizing Archaeon from the San Francisco Bay Estuary. *Journal of Bacteriology* **194**: 2121-2122.
- Naeem FI, Ashraf MM, Malik KA, Hafeez FY (2004). Competitiveness of introduced *Rhizobium* strains for nodulation in fodder legumes. *Pakistan Journal of Botany* **36**: 159-166.

- 1  
2  
3  
4  
5  
6 Nohynek LJ, Haggblom MM, Palleroni NJ, Kronqvist K, Nurmiolahassila EL, Salkinojasalonen M (1993). Characterization of a  
7 Mycobacterium-fortuitum strain capable of degrading polychlorinated phenolic-compounds. *Systematic and Applied Microbiology* **16**:  
8 126-134.  
9
- 10 Normand P, Lalonde M (1982). Evaluation of Frankia strains isolated from provenances of 2 Alnus species. *Canadian Journal of*  
11 *Microbiology* **28**: 1133-1142.  
12
- 13  
14 Normand P, Lapierre P, Tisa LS, Gogarten JP, Alloisio N, Bagnarol E *et al* (2007). Genome characteristics of facultatively symbiotic  
15 Frankia sp strains reflect host range and host plant biogeography. *Genome Research* **17**: 7-15.  
16
- 17 Norton JM, Klotz MG, Stein LY, Arp DJ, Bottomley PJ, Chain PSG *et al* (2008). Complete genome sequence of Nitrosospira multiformis, an  
18 ammonia-oxidizing bacterium from the soil environment. *Applied and Environmental Microbiology* **74**: 3559-3572.  
19
- 20  
21 Nouioui I, Ghodhbane-Gtari F, Montero-Calasanz MD, Goker M, Meier-Kolthoff JP, Schumann P *et al* (2016). Proposal of a type strain for  
22 Frankia alni (Woronin 1866) Von Tubeuf 1895, emended description of Frankia alni, and recognition of Frankia casuarinae sp nov and  
23 Frankia elaeagni sp nov. *International Journal of Systematic and Evolutionary Microbiology* **66**: 5201-5210.  
24
- 25 Nouioui I, Ghodhbane-Gtari F, Montero-Calasanz MD, Rohde M, Tisa LS, Gtari M *et al* (2017). Frankia inefficax sp nov., an actinobacterial  
26 endophyte inducing ineffective, non nitrogen-fixing, root nodules on its actinorhizal host plants. *Antonie Van Leeuwenhoek International*  
27 *Journal of General and Molecular Microbiology* **110**: 313-320.  
28
- 29  
30 Ogg CD, Patel BKC (2009). Caloramator australicus sp nov., a thermophilic, anaerobic bacterium from the Great Artesian Basin of  
31 Australia. *International Journal of Systematic and Evolutionary Microbiology* **59**: 95-101.  
32
- 33 Osterman J, Marsh J, Laine PK, Zeng Z, Alatalo E, Sullivan JT *et al* (2014). Genome sequencing of two Neorhizobium galegae strains  
34 reveals a noeT gene responsible for the unusual acetylation of the nodulation factors. *Bmc Genomics* **15**.  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

1  
2  
3  
4  
5  
6 Peix A, Berge O, Rivas R, Abril A, Velazquez E (2005). *Pseudomonas argentinensis* sp nov., a novel yellow pigment-producing bacterial  
7 species, isolated from rhizospheric soil in Cordoba, Argentina. *International Journal of Systematic and Evolutionary Microbiology* **55**:  
8 1107-1112.  
9

10 Persson T, Battenberg K, Demina IV, Vigil-Stenman T, Heuvel BV, Pujic P *et al* (2015). Candidatus *Frankia Datiscae* Dg1, the  
11 Actinobacterial Microsymbiont of *Datisca glomerata*, Expresses the Canonical nod Genes nodABC in Symbiosis with Its Host Plant. *Plos*  
12 *One* **10**.  
13

14  
15 Prabhu DM, Quadri SR, Cheng J, Liu L, Chen W, Yang Y *et al* (2015). *Sinomonas mesophila* sp nov., isolated from ancient fort soil. *Journal*  
16 *of Antibiotics* **68**: 318-321.  
17

18 Qiu YL, Kuang XZ, Shi XS, Yuan XZ, Guo RB (2014). *Terrimicrobium sacchariphilum* gen. nov., sp nov., an anaerobic bacterium of the class  
19 'Spartobacteria' in the phylum Verrucomicrobia, isolated from a rice paddy field. *International Journal of Systematic and Evolutionary*  
20 *Microbiology* **64**: 1718-1723.  
21

22  
23 Reichenbach H, Lang E, Schumann P, Sproer C (2006). *Byssovorax cruenta* gen. nov., sp nov., nom. rev., a cellulose-degrading  
24 myxobacterium: rediscovery of 'Myxococcus cruentus' Thaxter 1897. *International Journal of Systematic and Evolutionary Microbiology*  
25 **56**: 2357-2363.  
26

27 Rice MC, Norton JM, Valois F, Bollmann A, Bottomley PJ, Klotz MG *et al* (2016). Complete genome of *Nitrosospira briensis* C-128, an  
28 ammonia-oxidizing bacterium from agricultural soil. *Standards in Genomic Sciences* **11**.  
29

30 Rippka R, Herdman M (1992). *Pasteur Culture Collection of Cyanobacterial Strains in Axenic Culture, Catalogue and Taxonomic Handbook*.  
31 Institut Pasteur, Paris. 103 pp.  
32

33  
34 Rivas R, Sanchez M, Trujillo ME, Zurdo-Pineiro JL, Mateos PF, Martinez-Molina E, Velazquez E (2003). *Xylanimonas cellulositytica* gen.  
35 nov., sp. nov., a xylanolytic bacterium isolated from a decayed tree (*Ulmus nigra*). *International Journal of Systematic and Evolutionary*  
36 *Microbiology* **53**: 99-103.  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

<sup>a</sup>Rivas R, Trujillo ME, Mateos PF, Martinez-Molina E, Velazquez E (2004). Cellulomonas xylanilytica sp nov., a cellulolytic and xylanolytic bacterium isolated from a decayed elm tree. *International Journal of Systematic and Evolutionary Microbiology* **54**: 533-536.

<sup>b</sup>Rivas R, Trujillo ME, Schumann P, Kroppenstedt RM, Sanchez M, Mateos PF et al (2004). Xylanibacterium ulmi gen. nov., sp. nov., a novel xylanolytic member of the family Promicromonosporaceae. *International Journal of Systematic and Evolutionary Microbiology* **54**: 557-561.

Ruckert C, Albersmeier A, Busche T, Jaenicke S, Winkler A, Friojonsson OH *et al* (2015). Complete genome sequence of Streptomyces lividans TK24. *Journal of Biotechnology* **199**: 21-22.

Sanchez-Andrea I, Florentino AP, Semerel J, Strepis N, Sousa DZ, Stams AJM (2018). Co-culture of a novel fermentative bacterium, Lucifera butyrica gen. nov. sp. nov., with the sulfur reducer Desulfurella amilsii for enhanced sulfidogenesis. *Frontiers in Microbiology* **9**: 3108.

Sakai S, Imachi H, Hanada S, Ohashi A, Harada H, Kamagata Y (2008). Methanocella paludicola gen. nov., sp nov., a methane-producing archaeon, the first isolate of the lineage 'Rice Cluster I', and proposal of the new archaeal order Methanocellales ord. nov. *International Journal of Systematic and Evolutionary Microbiology* **58**: 929-936.

Sakai S, Conrad R, Liesack W, Imachi H (2010). Methanocella arvoryzae sp nov., a hydrogenotrophic methanogen isolated from rice field soil. *International Journal of Systematic and Evolutionary Microbiology* **60**: 2918-2923.

Sangkhobol V, Skerman VBD (1981). Chitinophaga, a new genus of chitinolytic Myxobacteria. *International Journal of Systematic Bacteriology* **31**: 285-293.

Sangwan P, Chen XL, Hugenholtz P, Janssen PH (2004). Chthoniobacter flavus gen. nov., sp nov., the first pure-culture representative of subdivision two, Spartobacteria classis nov., of the phylum Verrucomicrobia. *Applied and Environmental Microbiology* **70**: 5875-5881.

Schumann P, Weiss N, Stackebrandt E (2001). Reclassification of Cellulomonas cellulans (Stackebrandt and Keddie 1986) as Cellulosimicrobium cellulans gen. nov., comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **51**: 1007-1010.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

Schumann P, Behrendt U, Ulrich A, Suzuki K (2003). Reclassification of *Subtercola pratensis* Behrendt et al. 2002 as *Agreia pratensis* comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **53**: 2041-2044.

Schwabe GH (1960). Zur autotrophen vegetation in ariden Böden. Blaualgen und Lebensraum IV. *Österreichische Botanische Zeitschrift* **107**: 281-309.

Shamseldin A, Carro L, Peix A, Velazquez E, Moawad H, Sadowsky MJ (2016). The symbiovar trifolii of *Rhizobium bangladeshense* and *Rhizobium aegyptiacum* sp nov nodulate *Trifolium alexandrinum* in Egypt. *Systematic and Applied Microbiology* **39**: 275-279.

Singh AK, Garg N, Lata P, Kumar R, Negi V, Vikram S *et al* (2014). *Pontibacter indicus* sp nov., isolated from hexachlorocyclohexane-contaminated soil. *International Journal of Systematic and Evolutionary Microbiology* **64**: 254-259.

Soby SD, Gadagkar SR, Contreras C, Caruso FL (2013). *Chromobacterium vaccinii* sp nov., isolated from native and cultivated cranberry (*Vaccinium macrocarpon* Ait.) bogs and irrigation ponds. *International Journal of Systematic and Evolutionary Microbiology* **63**: 1840-1846.

Starkenburger SR, Chain PSG, Sayavedra-Soto LA, Hauser L, Land ML, Larimer FW *et al* (2006). Genome sequence of the chemolithoautotrophic nitrite-oxidizing bacterium *Nitrobacter winogradskyi* Nb-255. *Applied and Environmental Microbiology* **72**: 2050-2063.

Starkenburger SR, Reitenga KG, Freitas T, Johnson S, Chain PSG, Garcia-Pichel F, Kusek CR (2011). Genome of the Cyanobacterium *Microcoleus vaginatus* FGP-2, a photosynthetic ecosystem engineer of arid land soil biocrusts worldwide. *Journal of Bacteriology* doi.org/10.1128/JB.05138-11.

Strömpl C, Tindall BJ, Jarvis GN, Lünsdorf H, Moore ERB, Hippe H (1999). A re-evaluation of the taxonomy of the genus *Anaerovibrio*, with the reclassification of *Anaerovibrio glycerini* as *Anaerosinus glycerini* gen. nov., comb. Nov., and *Anaerovibrio burkinabensis* as *Anaeroarcus burkinensis* [corrig.] gen. nov., comb. Nov. *International Journal of Systematic Bacteriology* **49**: 1861-1872.

- 1  
2  
3  
4  
5  
6 Takahashi Y, Tanaka Y, Iwai Y, Omura S (1987). Promicromonospora sukumoe sp. nov., a new species of the Actinomycetales. *Journal of*  
7 *General and Applied Microbiology* **33**: 507-519.
- 8  
9 Takami H, Arai W, Takemoto K, Uchiyama I, Taniguchi T (2015). Functional Classification of Uncultured "Candidatus Caldiarchaeum  
10 subterraneum" Using the Maple System. *Plos One* **10**.
- 11  
12 Takarada H, Sekine M, Kosugi H, Matsuo Y, Fujisawa T, Omata S *et al* (2008). Complete genome sequence of the soil actinomycete  
13 Kocuria rhizophila. *Journal of Bacteriology* **190**: 4139-4146.
- 14  
15 Tanaka Y, Matsuzawa H, Tamaki H, Tagawa M, Toyama T, Kamagata Y *et al* (2017). Isolation of Novel Bacteria Including Rarely  
16 Cultivated Phyla, Acidobacteria and Verrucomicrobia, from the Roots of Emergent Plants by Simple Culturing Method. *Microbes and*  
17 *Environments* **32**: 288-292.
- 18  
19 Tang B, Yu YC, Zhang YM, Zhao GP, Ding XM (2015). Complete genome sequence of the glidobactin producing strain Polyangium  
20 brachysporum DSM 7029. *Journal of Biotechnology* **210**: 83-84.
- 21  
22 Tikariha H, Pal RR, Qureshi A, Kapley A, Purohit HJ (2016). In silico analysis for prediction of degradative capacity of Pseudomonas  
23 putida SF1. *Gene* **591**: 382-392.
- 24  
25 Toledo I, Lloret L, Martinez-Romero E (2003). Sinorhizobium americanus sp nov., a new Sinorhizobium species nodulating native Acacia  
26 spp. in Mexico. *Systematic and Applied Microbiology* **26**: 54-64.
- 27  
28 Van VT, Ngoke S, Berge O, Faure D, Bally R, Hebbar P *et al* (1997). Isolation of Azospirillum lipoferum from the rhizosphere of rice by a  
29 new, simple method. *Canadian Journal of Microbiology* **43**: 486-490.
- 30  
31 Vaz-Moreira I, Nobre MF, Ferreira ACS, Schumann P, Nunes OC, Manaia CM (2008). Humibacter albus gen. nov., sp. nov., isolated from  
32 sewage sludge compost. *International Journal of Systematic and Evolutionary Microbiology* **58**: 1014-1018.
- 33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

1  
2  
3  
4  
5  
6  
7 Vieira S, Luckner M, Wanner G, Overmann J (2017). *Luteitalea pratensis* gen. nov., sp nov a new member of subdivision 6 Acidobacteria  
8 isolated from temperate grassland soil. *International Journal of Systematic and Evolutionary Microbiology* **67**: 1408-1414.

9  
10 Voelz H, Dworkin M (1962). Fine structure of *Myxococcus xanthus* during morphogenesis. *Journal of Bacteriology* **84**: 943-&.

11  
12 Volland S, Rachinger M, Strittmatter A, Daniel R, Gottschalk G, Meyer O (2011). Complete Genome Sequences of the  
13 Chemolithoautotrophic Oligotropha carboxidovorans Strains OM4 and OM5. *Journal of Bacteriology* **193**: 5043-5043.

14  
15 Vorobev AV, Baani M, Doronina NV, Brady AL, Liesack W, Dunfield PF *et al* (2011). *Methyloferula stellata* gen. nov., sp nov., an  
16 acidophilic, obligately methanotrophic bacterium that possesses only a soluble methane monooxygenase. *International Journal of*  
17 *Systematic and Evolutionary Microbiology* **61**: 2456-2463.

18  
19 Wang YX, Liu JH, Zhang XX, Chen YG, Wang ZG, Chen Y *et al* (2009). *Fodinicurvata sediminis* gen. nov., sp nov and *Fodinicurvata*  
20 *fenggangensis* sp nov., poly-beta-hydroxybutyrate-producing bacteria in the family Rhodospirillaceae. *International Journal of*  
21 *Systematic and Evolutionary Microbiology* **59**: 2575-2581.

22  
23 Ward NL, Challacombe JF, Janssen PH, Henrissat B, Coutinho PM, Wu M *et al* (2009). Three Genomes from the Phylum Acidobacteria  
24 Provide Insight into the Lifestyles of These Microorganisms in Soils. *Applied and Environmental Microbiology* **75**: 2046-2056.

25  
26 Warnick TA, Methe BA, Leschine SB (2002). *Clostridium phytofermentans* sp nov., a cellulolytic mesophile from forest soil. *International*  
27 *Journal of Systematic and Evolutionary Microbiology* **52**: 1155-1160.

28  
29 Wartiainen I, Hestnes AG, McDonald IR, Svenning MM (2006). *Methylocystis rosea* sp nov., a novel methanotrophic bacterium from  
30 Arctic wetland soil, Svalbard, Norway (78 degrees N). *International Journal of Systematic and Evolutionary Microbiology* **56**: 541-547.

31  
32 Weon HY, Kim BY, Yoo SH, Lee SY, Kwon SW, Go SJ *et al* (2006). *Niastella koreensis* gen. nov., sp nov and *Niastella yeongjuensis* sp nov.,  
33 novel members of the phylum Bacteroidetes, isolated from soil cultivated with Korean ginseng. *International Journal of Systematic and*  
34 *Evolutionary Microbiology* **56**: 1777-1782.



- 1  
2  
3  
4  
5  
6 Weon HY, Kim BY, Yoo SH, Joa JH, Kwon SW, Kim WG (2007). *Andreprevotia chitinilytica* gen. nov., sp nov., isolated from forest soil from  
7 Halla Mountain, Jeju Island, Korea. *International Journal of Systematic and Evolutionary Microbiology* **57**: 1572-1575.  
8
- 9 Weon HY, Kim BY, Joa JH, Kwon SW, Kim WG, Koo BS (2008). *Niabella soli* sp nov., isolated from soil from Jeju Island, Korea.  
10 *International Journal of Systematic and Evolutionary Microbiology* **58**: 467-469.  
11
- 12  
13 Westerberg K, Elvang AM, Stackebrandt E, Jansson JK (2000). *Arthrobacter chlorophenolicus* sp nov., a new species capable of degrading  
14 high concentrations of 4-chlorophenol. *International Journal of Systematic and Evolutionary Microbiology* **50**: 2083-2092.  
15
- 16 Whittenbury R, Phillips KC, Wilkinso J (1970). Enrichment, isolation and some properties of Methane-utilizing bacteria. *Journal of*  
17 *General Microbiology* **61**: 205-&.  
18
- 19 Winogradsky S (1929). Études sur la microbiologie du sol - sur la dégradation de la cellulose dans le sol. *Annales de l'Institut Pasteur* **43**:  
20 549-633.  
21
- 22  
23 Wise MG, McArthur JV, Shimkets LJ (2001). *Methylosarcina fibrata* gen. nov., sp nov and *Methylosarcina quisquiliarum* sp nov., novel  
24 type I methanotrophs. *International Journal of Systematic and Evolutionary Microbiology* **51**: 611-621.  
25
- 26 Xie CH, Yokota A (2005a). *Azospirillum oryzae* sp nov., a nitrogen-fixing bacterium isolated from the roots of the rice plant *Oryza sativa*.  
27 *International Journal of Systematic and Evolutionary Microbiology* **55**: 1435-1438.  
28
- 29 Xie CH, Yokota A (2005b). *Pleomorphomonas oryzae* gen. nov., sp nov., a nitrogen-fixing bacterium isolated from paddy soil of *Oryza*  
30 *sativa*. *International Journal of Systematic and Evolutionary Microbiology* **55**: 1233-1237.  
31
- 32  
33 Xie CH, Yokota A (2006). Reclassification of *Flavobacterium ferrugineum* as *Terrimonas ferruginea* gen. nov., comb. nov., and description  
34 of *Terrimonas lutea* sp nov., isolated from soil. *International Journal of Systematic and Evolutionary Microbiology* **56**: 1117-1121.  
35
- 36 Xie G, Bruce DC, Challacombe JF, Chertkov O, Detter JC, Gilna P *et al* (2007). Genome sequence of the cellulolytic gliding bacterium  
37 *Cytophaga hutchinsonii*. *Applied and Environmental Microbiology* **73**: 3536-3546.  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

1  
2  
3  
4  
5  
6  
7  
8 Yang HC, Im WT, Kim KK, An DS, Lee ST (2006). Burkholderia terrae sp nov, isolated from a forest soil. *International Journal of*  
9 *Systematic and Evolutionary Microbiology* **56**: 453-457.

10  
11 Yoon JH, Kang SJ, Schumann P, Oh TK (2007). Cellulosimicrobium terreum sp. nov., isolated from soil. *International Journal of*  
12 *Systematic and Evolutionary Microbiology* **57**: 2493-2497.

13  
14 Yoon MH, Ten LN, Im WT, Lee ST (2008). Cellulomonas chitinilytica sp nov., a chitinolytic bacterium isolated from cattle-farm compost.  
15 *International Journal of Systematic and Evolutionary Microbiology* **58**: 1878-1884.

16  
17 Zaburannyi N, Bunk B, Maier J, Overmann J, Muller R (2016). Genome Analysis of the Fruiting Body-Forming Myxobacterium  
18 Chondromyces crocatus Reveals High Potential for Natural Product Biosynthesis. *Applied and Environmental Microbiology* **82**: 1945-  
19 1957.

20  
21  
22 Zhang JY, Liu XY, Liu SJ (2010). Agrococcus terreus sp nov and Micrococcus terreus sp nov., isolated from forest soil. *International*  
23 *Journal of Systematic and Evolutionary Microbiology* **60**: 1897-1903.

24  
25 Zgurskaya HI, Evtushenko LI, Akimov VN, Voyevoda HV, Dobrovolskaya TG, Lysak LV *et al* (1992). Emended description of the  
26 genus Agromyces and description of Agromyces cerinus subsp. cerinus sp. nov., subsp. nov., Agrocymces cerinus subsp. nitratius  
27 sp. nov., subsp. nov., Agromyces fucosus subsp. fucosus sp. nov., subsp. nov., and Agromyces fucosus subsp. nov., and  
28 Agromyces fucosus subsp. hippuratus sp. nov., subsp. nov. *International Journal of Systematic Bacteriology* **42**: 635-641.  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

1  
2  
3 Manuscript id: FEMSEC-21-01-0026

4 Title: Functional trait relationships demonstrate generalisable life strategies in  
5 terrestrial prokaryotes.

6 Authors: Finn, Damien; Bergk-Pinto, Benoit; Hazard, Christina; Nicol, Graeme;  
7 Tebbe, Christoph; Vogel, Tim  
8  
9

10  
11 Reviewers' Comments to Authors:  
12  
13

14 **Reviewer: 1**

15 Comments to the Author

16 Review of the paper: Functional trait relationships demonstrate generalisable life  
17 strategies in terrestrial prokaryotes.

18 This paper aims at identifying copiotrophic vs oligotrophic strategies among 170  
19 species of bacteria and archaea by comparing their similarity in their traits (based on  
20 clustering of functionally similar amino acid sequences). Instead of finding two clear  
21 cluster reflecting copiotrophic-oligotrophic framework, they identify 5 clusters. Only  
22 one of these clusters can be fit as copiotrophic. All other ones, reflect, as the author  
23 suggest, unique life-history strategies on their own.

24 I really like the paper and I think it is worth publication for two reasons: 1) It  
25 challenges simplistic views that entire kingdoms (or other higher taxa) of  
26 microorganisms can be placed along a unique axis of energy use. 2) the method they  
27 use will surely be used by other researchers. I also send my kudos to the authors to  
28 develop a step-by-step repo explaining their method. Doing so, not only ensures  
29 reproducibility but also encourage other scientist to test similar questions with their  
30 approach.

31 I like it as it is. However, I want to acknowledge that I cannot judge on the  
32 information of the natural history and ecology of many of the taxa included.  
33  
34  
35

36  
37  
38  
39  
40 Comment:

41 My only major suggestion is to make more explicit the way they test hypothesis 3:  
42 "copiotrophic and oligotrophic groups would emerge based on trait enrichment".  
43 (L149.150)  
44

45 From what I understood in the method section (283-287) that hypothesis was tested  
46 by subsetting 60 traits that emerge from their random forest clustering that are  
47 "associated with copiotroph-oligotroph growth strategies". This association is based  
48 "previous studies". Since this is a major hypothesis in the paper, I suggest to  
49 summarize the traits as a table (at least in the supplementary) where it is indicated  
50 which ones are expected to belong in copiotroph-oligotroph axis. Even better, it would  
51 be great to have them grouped as a figure. A way to simplify this large number of traits  
52 in a figure would be to group them according the Brite categories. Ultimately, I am  
53 talking about some conceptual figure that shows the idealized clustering of copiotroph-  
54 oligotroph based exactly on the 60 traits you use.  
55  
56  
57  
58  
59

60 Response:

1  
2  
3 The authors would like to thank the reviewer for their incredibly generous comments  
4 and the suggestion to create a conceptual figure summarising the Brite category  
5 clustering per Family. Indeed, by trying to consider the best way to communicate our  
6 results it has helped us to further synthesise them. Figure 5 is now a conceptual  
7 diagram that attempts to show how the five identified clades overlap based on their  
8 trait composition. This is either as a) a one-dimensional axis of 'resource investment',  
9 analogous to the classical dichotomy of copiotroph-oligotroph, and b) a multi-  
10 dimensional space where clades are further separated based on traits for both  
11 'resource investment' and different means of 'resource acquisition'. By considering  
12 both investment and acquisition together, we hope to show that the clades show less  
13 overlap in space, which would theoretically allow them to occupy separate niches and  
14 co-exist. We do not include all 60 Brite categories in this figure, but rather have chosen  
15 some select traits (14 in total) that may be useful in demonstrating our point e.g. if a  
16 taxon possesses many glycosyltransferases in addition to spore formation or  
17 antimicrobial resistance genes it could be considered an 'oligotrophic degrader'. The  
18 14 traits were partially chosen based on a comment from Reviewer 2, which suggested  
19 to name major functional markers that could be used to predict a taxon's niche. Section  
20 4.7 has been updated to incorporate and explain Figure 5 in the context of our results.

21 We are acutely aware that this conceptual figure may be met with criticism from the  
22 scientific community, and we are more than happy to modify, adjust or otherwise take  
23 on board any further constructive comments from the reviewer.  
24  
25  
26

27  
28  
29  
30  
31  
32  
33  
34 Minor comment

35 L40-42. I think you can be bolder. I think your study shows that the copio-oligotroph  
36 view is too simplistic and that more life-history strategies emerge once you compare  
37 species based on their traits and not pre-conceived ideas of how microorganisms  
38 should behave. I think that stating that the study "warrants a more nuanced view of  
39 life-histories" is too soft and vague.  
40  
41  
42

43 Response:

44 The concluding sentences of the abstract have been amended to emphasise that the  
45 copiotroph-oligotroph framework is insufficient to describe prokaryote niche, and that  
46 we suggest considering traits involved in growth and resource acquisition in tandem  
47 when predicting niche:  
48  
49

50 "However, the formation of five clades suggested a more nuanced view to describe  
51 niche differentiation in terrestrial systems is necessary. We suggest considering traits  
52 involved both in resource investment and acquisition when predicting niche."  
53  
54  
55  
56

57 **Reviewer: 2**

58 Comments to the Author

59 I reviewed the manuscript "Functional trait relationships demonstrate generalisable  
60

1  
2  
3 life strategies in terrestrial prokaryotes” by Finn et al. The objective of the manuscript  
4 is to identify the traits within soil bacterial and archaeal species that can distinguish  
5 their lifestyle and niche occupancy. Genome sequence of 175 taxa from 11 Phyla  
6 and 35 Families were used for analysis. The findings strongly suggested a non-  
7 random distribution of traits across phylogenetic lineages. Random forest analysis  
8 further identified traits that were directly related to niche colonization, substrate use  
9 efficiency and metabolism within phylogenetic lineages.

10  
11 I enjoyed reading the manuscript. The manuscript fits very well within the scope of  
12 journal. Proper statistical tests have been performed to reach the conclusions.

13  
14 Manuscript is very well written and presented. I have few suggestions that will further  
15 strengthen the manuscript.  
16  
17  
18

19  
20 Comment:

21  
22 1. There has been some work that has explored the genomic basis of tropic lifestyle  
23 in soil bacteria. This work should be cited and discussed appropriately.  
24 e.g. Trivedi et al. 2013. Trends in Microbiology.  
25

26  
27 Response:

28  
29 The authors would like to thank the reviewer for their generous comments. In regard  
30 to this specific comment, references and discussions of Trivedi et al., 2013 have been  
31 incorporated several times throughout the manuscript, for example:

32  
33 Ln 121: “A genomic basis for traits associated with soils dominated by putative  
34 copiotrophs and oligotrophs has been expertly reviewed elsewhere, and interested  
35 readers are referred to Trivedi *et al.*, 2013 and references therein.”  
36

37  
38 Ln 639: “Comparative genomics analyses have also identified Actinobacteria,  
39 Acidobacteria and Verrucomicrobia as being enriched in numerous enzymes for  
40 cellulose, hemicellulose and starch catabolism (Trivedi *et al.*, 2013).”  
41

42  
43 And Ln 674: “Comparative genomics analyses of soil bacteria have also found  
44 putatively copiotrophic Proteobacteria and Firmicutes to be particularly enriched in  
45 PTS and ABC transporters (Trivedi *et al.*, 2013).”  
46

47  
48 Furthermore, the study by Barbéran et al., 2012 to define ‘generalist’ and ‘specialist’  
49 soil taxa based on co-occurrence has been incorporated in the Introduction:  
50

51  
52 Ln 86: “... identifying habitat generalists and specialists based on taxon co-occurrence  
53 patterns (Barbéran *et al.*, 2012) ...”  
54  
55

56  
57 Comment:

58  
59 2. Only five strains from family level classification were selected. Authors have  
60 mentioned multiple times that some families will have a significant phylogenetic

1  
2  
3 diversity within strains, while for others, it will be restricted. How do these differences  
4 impact overall results? What was the criteria of selecting only 5 and not more?  
5  
6

7 Response:

8 A minimum of five genomes per taxonomic group were selected to accommodate  
9 Thaumarchaeota, Verrucomicrobia Chloroflexi and some others, which had the least  
10 number of publicly available sequenced genomes of soil taxa. We specifically wanted  
11 to avoid generating a dataset which favoured frequently isolated and sequenced soil  
12 organisms (e.g. Bacilli, Rhizobia) and thus five genomes per group were used to  
13 balance statistical analyses. This is mentioned at line 177 in the Methods section,  
14 under Collection of terrestrial prokaryote genomes.  
15  
16  
17  
18

19 Questions 2 and 3 from the reviewer have prompted additional analyses to quantify  
20 the phylogenetic (dis)similarity of taxonomic groups based on full length 16S rRNA  
21 genes, and to compare this to trait similarity ( $C$ ). The 16S genes were aligned with  
22 MUSCLE, simple Neighbour Joining trees were constructed in R with ape/phangorn  
23 packages, and phylogenetic distance ( $P$ ) calculated as per Equation 1, excepting that  
24  $P$  is in units of gene sequence similarity as opposed to Bray-Curtis distance for trait  
25 similarity ( $C$ ). Important considerations between  $P$  and  $C$  are as follows:  
26  
27  
28

29  
30 1) Taxa were originally chosen in the hope of maximising  $P$  (and presumably  $C$ ) per  
31 Family by choosing different Genera where possible. This was not always possible,  
32 and ultimately choices were constrained by what genomes were publicly available.  
33 However, looking at values of  $P$  per Phylum and Family (Supplementary Figure 2),  
34 there does not appear to be a consistent relationship between Families where different  
35 Genera were included and  $P$ . For example, Lactobacillaceae consisted of 5  
36 *Lactobacillus* spp. ( $P$  of 0.772) whereas the Beijerinckiaceae consisted of 4 different  
37 genera that appear to be actually more similar to each other in regard to the 16S rRNA  
38 gene ( $P$  of 0.852). In retrospect this lack of consistent variation in phylogenetic  
39 distance between taxa based on pre-existing nomenclature should not be a surprise  
40 at it is the reason for the efforts of the researchers involved with the Genome  
41 Taxonomy Database who seek to standardise the current state of microbial taxonomy.  
42  
43  
44  
45  
46

47 2) There is a very nice linear relationship between  $P$  and  $C$  when considering taxa at  
48 both the Phyla and Family levels (Supplementary Figure 3). As implied by the  
49 reviewer, this will mean any Families that consist of more closely related taxa (*i.e.* high  
50  $P$ ) will be more likely to be similar in regard to their traits.  
51  
52

53 3) The importance of this underlying relationship between the variation in phylogenetic  
54 distance between taxa in a group and the results of the Random Forest models is,  
55 however, uncertain. Supplementary Table 3 shows the Confusion Matrix for the Family  
56 level Random Forest. Some Families with the poorest predictive capacity (< 50% taxa  
57 correctly predicted) had quite high  $P$  (> 0.8), e.g. Bradyrhizobiaceae,  
58 Cellulomonadaceae. Some Families with low  $P$  (< 0.7) had poor predictive capacity,  
59  
60



1  
2  
3 e.g. Divergent Acidobacteria. Conversely, Families that had the highest predictive  
4 capacity (100%) ranged in  $P$  from 0.68 – 0.9. It is the opinion of the authors that the  
5 presence of specific, unique and consistent traits in a taxonomic group is far more  
6 important for the RF results than the underlying relationship between phylogenetic  
7 distance and trait composition. For example, the five Methanogen taxa had very low  
8 phylogenetic distance (0.68) yet the RF was perfectly accurate in classifying them,  
9 probably due to their methanogen metabolism and Eukaryote-like genetic traits.  
10  
11  
12

13  
14 Nevertheless, we feel the reviewer raises a very important point when considering  
15 taxon selection for these types of comparative genomic analyses, and in future studies  
16 we shall place emphasis on standardising phylogenetic distance within groups. The  
17 following has been added to the manuscript to highlight this:  
18  
19

20 In the Methods section 2.3 discussing UPGMA of trait similarity analysis:

21 Ln 291: “Furthermore, the full length 16S rRNA gene of each taxon was collated from  
22 NCBI. Genes were aligned with MUSCLE (Edgar, 2004) and a Neighbour-Joining  
23 phylogenetic tree was constructed with the ‘phangorn’ package in R. Phylogenetic  
24 distance present in taxonomic groups ( $P$ ) was measured as per Equation 1., excepting  
25 that branch length was in units of DNA sequence similarity as opposed to Bray-Curtis  
26 dissimilarity. Finally, simple linear regression was used to test a relationship between  
27  $P$  and  $C$ .”  
28  
29  
30  
31

32 In the Results section discussing the UPGMA trait tree and RF models:

33 Ln 364: “A Neighbour-Joining tree of full length 16S rRNA genes showed that all taxa  
34 clustered preferentially based on their taxonomic nomenclature at high taxonomic  
35 rank, including the Chloroflexi, indicating that the discrepancies in the UPGMA were  
36 not due to misclassification of the individual taxa (Supplementary Figure 2a).”  
37  
38  
39

40 Ln 388: “Phylogenetic distance ( $P$ ) of each taxonomic group increased with  
41 decreasing taxonomic rank, and was highest in Proteobacteria, Actinobacteria and  
42 Firmicutes Families (Supplementary Figure 2b). There was a strong positive linear  
43 relationship between  $P$  and  $C$  ( $y = 0.86x - 0.25$ ,  $R^2 = 0.39$ ,  $p < 0.001$ ) demonstrating  
44 that taxonomic groups of closer related taxa tended to share more similar  
45 compositions of traits.”  
46  
47  
48

49 Ln 412: “Random Forest models were robust against variation in  $P$  within Families, for  
50 example the nine families with 0% classification error ranged in  $P$  from the lowest  
51 (0.68) to highest (0.9).”  
52  
53  
54

55 And finally in the Discussion section in regard to the robustness and interpretation of  
56 the Random Forest approach:  
57  
58

59 Ln 643: “Ideally the selection of individual taxa within groups for such future  
60 comparative analyses would also be standardised based on phylogenetic distance,  
61



1  
2  
3 either with  $P$  or a similar method, which would improve the robustness of trait-based  
4 comparisons at such a fine taxonomic level.”  
5  
6

7 Comment:

8 **3. What were the criteria of selection of 5 species within a family e.g., phylogenetic**  
9 **distance?**  
10  
11

12 Response:

13 As mentioned above, the selection criteria of the five taxa per Family was based  
14 primarily on public availability of genomes. However, the authors acknowledge and  
15 fully agree with the reviewers point that phylogenetic distance should be considered  
16 as a means to standardise selection of taxa. While we are confident that our methods  
17 are robust to the underlying relationship between  $P$  and  $C$  in the coarse Phyla and  
18 Family taxonomic groups analysed here, we will incorporate a method for  
19 standardising taxon selection in future studies that consider a finer detail of niche  
20 differentiation between closely related Species and Strains.  
21  
22  
23  
24  
25  
26

27 Comment:

28 **4. Lines 165-166: I like the idea of choosing strains that perform particular**  
29 **functions. However, these functions are performed by specific bacterial lineages. My**  
30 **question is whether these biased selections for traits inflate the closeness within a**  
31 **group. Or in other words, if the other members within the families of the selected**  
32 **functional groups were selected, do the results would have been the same?**  
33  
34  
35

36 Response:

37 This is an interesting question, and fortunately our results have several functional  
38 groups that can be compared to address it. For example, to our knowledge, all  
39 characterised members of Nitrosomonadaceae are chemolithoautotrophic nitrifiers  
40 while the Cyanobacteria perform photosynthesis. Thus, the relatively high value of  $C$   
41 measured here for these groups that share common functions is not surprising.  
42 However, trait variation certainly exists within other functional groups, for example the  
43 five Methanogens (isolated from diverse anaerobic habitats including rice rhizosphere,  
44 tropical and boreal wetlands) had an intermediate value of  $C$ . Perhaps the best  
45 example is that the inclusion of functionally distinct methanotrophs, methylotrophs and  
46 heterotrophs within the Beijerinckiaceae did not have a negative effect on its  $C$ , which  
47 yielded one of the highest values (discussed at Ln 524). Ultimately trait composition  
48 will be influenced by a taxon's functional group (e.g. ammonia oxidiser, heterotroph,  
49 etc), its evolutionary life-history (as a very general example Bacteria, Archaea) and  
50 the environment from where the taxon was isolated.  
51  
52  
53  
54  
55  
56

57 We have included a statement this effect in the Discussion, which is an interesting and  
58 important concept to consider:  
59  
60

1  
2  
3 Ln 569: “The differing values of C for Methanogen and photosynthetic Cyanobacteria  
4 functional groups (0.33 and 0.44, respectively) is also worthy of note. Despite all five  
5 taxa in each group performing the same core role in a community, the individual  
6 isolates came from varying environments. The methanogens were isolated from a  
7 range of geographically separate wetlands, rice paddy soil and farm slurry and, while  
8 the Cyanobacteria were also isolated from geographically separate environments,  
9 they were all from sandy deserts or other nutrient poor, arid soils (Supplementary  
10 Table 1 and references therein). Ultimately a taxon’s trait composition will be affected  
11 by its functional role in a community, its evolutionary life-history (e.g. Beijerinckia  
12 described above) and its local environment.”  
13  
14  
15  
16  
17  
18

19 Comment:

20 5. Like Lauro et al. 2009, is there a way to do a self-organizing map or a constrained  
21 analysis showing the separation of different phylogenetic groups according to the  
22 traits?  
23  
24  
25

26 Response:

27 The constrained analyses of Lauro *et al.*, 2009 revolved around the *a priori*  
28 classification of one genome as ‘copiotrophic’ and one genome as ‘oligotrophic’.  
29 Unknown genomes were then placed somewhere between these two based on  
30 similarity of shared traits. In this study we specifically sought to avoid any  
31 preconceived *a priori* assumptions of ‘taxon A is a copiotroph’, and instead opted for  
32 unconstrained, exploratory cluster-based statistical techniques (e.g. dendrograms,  
33 Random Forest) that would let the data speak for itself. We hoped that the UPGMA  
34 dendrogram (Figure 2a) would be the simplest and most flexible way to demonstrate  
35 relationships between different phylogenetic groups and their trait composition, while  
36 the clustering of taxonomic groups based on the counts of specific copiotroph-  
37 oligotroph traits would validate whether certain taxa could be considered copiotrophic  
38 (Figure 4).  
39  
40  
41  
42  
43  
44

45 Comment:

46 6. Based on the results, can you substantiate the phylogenetic grouping of  
47 copiotroph/oligotroph groups? I think yes, there is much information, but I would like  
48 to have a strong statement suggesting that based on the results, the inclusion of X, Y,  
49 and Z phylum as copiotroph was validated. Then can you resolve the debate about  
50 the inclusion of groups such as Alpha and betaproteobacteria as copiotroph/oligotroph  
51 (Lines 104-105). See 5 as the analysis suggested there can help you to address this  
52 question. This information will help the researchers that want to link microbial  
53 community composition to soil processes, particularly soil C cycling.  
54  
55  
56  
57  
58

59 Response:  
60

1  
2  
3 A definitive statement outlining which groups were confirmed as copiotrophs and  
4 oligotrophs has been included in the Discussion. This also explicitly links to how these  
5 results can give insight into soil processes such as C cycling:  
6  
7

8 Ln 872: "These results support previous observations that Rhodospirillaceae,  
9 Bradyrhizobiaceae, Burkholderiaceae, Pseudomonadaceae and Rhizobiaceae are  
10 copiotrophic, while Planctomycetes, Verrucomicrobia, Myxococcaceae,  
11 Polyangiaceae and Acidobacteria are oligotrophic (Ho *et al.*, 2017 and references  
12 therein). As has been proposed previously, the dominance of these groups in certain  
13 soils can provide inferences for ecosystem processes in that system, for example soils  
14 dominated by Verrucomicrobia, Planctomycetes and Acidobacteria will have greater  
15 capacity to degrade complex plant material while retaining most catabolised carbon in  
16 biomass (*i.e.* high growth or carbon use efficiency) or excreted byproducts that assist  
17 in soil aggregation (*e.g.* high LPS production) (Trivedi *et al.*, 2013). Conversely, soils  
18 dominated by copiotrophic Proteobacteria Families will be systems primarily  
19 dependent on labile di- and monosaccharides that demonstrate low carbon use  
20 efficiency."  
21  
22  
23  
24  
25

26  
27 Resolving the outstanding question of taxonomic groups that do not fit into the  
28 classical copiotroph-oligotroph framework is addressed further below as a conceptual  
29 figure (Comment 11).  
30  
31  
32

33 Comment:

34  
35 7. One important trait that distinguish copiotrophs/oligotrophs is the carbon use  
36 efficiency (CUE). There have been some efforts to incorporate microbial CUE to  
37 parameterize models for soil carbon. While CUE's topic is touched briefly in  
38 Introduction, it has not been explicitly discussed in later sections.  
39  
40

41 Response:

42 Carbon use efficiency has now been mentioned in regard to the specific  
43 copiotrophic/oligotrophic taxonomic groups that will have a high CUE (please see the  
44 above response to Comment 6).  
45  
46  
47  
48

49 Comment:

50  
51 8. Authors talk about MAGs in the discussion section. It will be worth discussing  
52 how MAGs will help to elaborate similar studies that is reported. Where we are and  
53 what needs to be done further?  
54  
55

56 Response:

57 The authors agree that this is worth expanding upon. The following two statements  
58 have been added to the Discussion to better incorporate the value of MAGs and  
59 current gaps for trait-based analyses:  
60

Ln 560: “The generation of MAGs has emerged as a useful tool for identifying traits necessary for life in such environments, and particularly for expanding knowledge of severely under-represented, difficult to culture taxonomic groups. For example, the recent reconstruction of 52, 515 MAGs from a wide range of host-associated and environmental metagenomes was able to increase genomic information of Planctomycetes and Verrucomicrobia by 79% and 68%, respectively (Nayfach et al., 2021). Importantly, both 16S rRNA gene surveys and MAGs demonstrate that some functional traits that facilitate life under certain environmental conditions are intrinsically linked to taxonomy.”

Ln 970: “However, to truly unravel differentiated niches and general microbial life strategies, two limitations must be overcome. Firstly, a better understanding of the many ‘Uncharacterised’ traits in environmental isolates is required. For example, the recent large-scale MAG study by Nayfach *et al.*, (2021) identified 5.8 million protein clusters (traits), of which over 75% could not be annotated meaningfully by current protein databases. Secondly, robust trait-based analyses down to the finer scale of distinct genomes will likely be necessary to consider how individual members of a community have either differentiated in order to co-exist or are in the throes of competition that will ultimately exclude one of the competitors.”

Comment:

9. Based on the results, where will you fit archaea (copiotrophs or oligotrophs)?

Response:

The authors are hesitant to place Archaea as either copiotrophs or oligotrophs as we consider the one-dimensional framework to be insufficient for describing niche. Please see further below for a detailed response (Comment 11).

Comment:

10. How this type of data will be used to distill the complexity within the soil microbiome functions? Can we use the functional signatures to determine the metabolic potential of microbial communities from metagenome data? How can this information help us to incorporate microbial activities into process models? I suggest touching on these in conclusions.

Response:

Again, this will be addressed further in Comment 11.

Comment:

11. One of the hallmarks of impactful papers is to move beyond information to

1  
2  
3 knowledge. Can you think of making a conceptual figure showing different genes and  
4 how they are responsible for different lifestyles or niche colonization? This is just a  
5 thought as I understand how difficult it can be.  
6  
7

8 Response:

9  
10 The authors particularly appreciated this comment, and in particular the expression  
11 'moving beyond information to knowledge'! Indeed, our attempts to consider a  
12 conceptual figure that conveys our results in a simple manner has been extremely  
13 helpful in further synthesising them.  
14  
15

16 Figure 5 is now a conceptual diagram that attempts to show how the five identified  
17 clades overlap based on their trait composition. This is either as a) a one-dimensional  
18 axis of 'resource investment', analogous to the classical dichotomy of copiotroph-  
19 oligotroph, and b) a multi-dimensional space where clades are further separated  
20 based on traits for both 'resource investment' and different means of 'resource  
21 acquisition'. By considering both investment and acquisition together, we hope to  
22 show that the clades show less overlap in space, which would theoretically allow them  
23 to occupy separate niches and co-exist. The final section of the Discussion, from Ln  
24 916, has been expanded to discuss this figure and its significance for interpreting  
25 microbial niche.  
26  
27  
28  
29  
30

31 In regard to Comment 10, which asks how our results may provide signatures for  
32 process-based models or linking to community function, we have chosen the BRITE  
33 categories in Figure 5 to act as signatures for defining, for example, Rhizobiaceae as  
34 'copiotrophic competitors'. This is explicitly stated:  
35  
36  
37

38 Ln 929: "Figure 5b is a conceptual diagram of niche space where clades have been  
39 further separated along additional axes based on their enrichment of traits involved in  
40 competition, degradation or specialised metabolic pathways. The BRITE categories  
41 listed on the various axes are chosen to be useful markers in predicting the niche of a  
42 taxon. This expanded niche space would suggest taxa in Clade I are well equipped for  
43 nutrient acquisition (primarily, but not limited to, di- and monosaccharides), rapid  
44 growth and oxidative stress regulation as 'copiotrophic competitors'. ..."  
45  
46  
47

48 And finally, specifically in regard to whether Archaea should be copiotrophs or  
49 oligotrophs:  
50  
51

52 Ln 956: "The large space conceptualised for Clade V in Figure 5b, which does not  
53 imply either copiotrophic or oligotrophic resource investment, is an oversimplification  
54 and with improved understanding of traits in these taxa it may be possible to fracture  
55 and further separate them into more detailed groups. This would prove particularly  
56 beneficial for the poorly characterised Archaea."  
57  
58  
59  
60

1  
2  
3 Comment:

4 12. I appreciate the authors of acknowledging the tedious work of other researchers  
5 on culturing. This is very important.  
6  
7

8 Response:

9 Of course we are happy to do so, as it is essential work.  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Peer Review

1  
2  
3 1 Functional trait relationships demonstrate ~~generalisable~~-life strategies in terrestrial  
4  
5 2 prokaryotes.  
6  
7 3

8 4 Damien R. Finn<sup>1,2,3</sup>, Benoît Bergk-Pinto<sup>2</sup>, Christina Hazard<sup>2</sup>, Graeme W. Nicol<sup>2</sup>,  
9 5 Christoph C. Tebbe<sup>3</sup>, Timothy M. Vogel<sup>2</sup>  
10  
11 6

12  
13 7 <sup>1</sup>School of Agriculture and Food Sciences, University of Queensland, Brisbane,  
14 8 Australia 4072;

15 9 <sup>2</sup>Environmental Microbial Genomics, Laboratoire Ampère, École Centrale de Lyon,  
16 10 Université de Lyon, Écully, France 69134;

17 11 <sup>3</sup>Thünen Institut für Biodiversität, Johann Heinrich von Thünen Institut,  
18 12 Braunschweig, Germany 38116.  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28

29 15 Author correspondence:

30 16 Damien Finn; current address: Thünen Institut für Biodiversität, Braunschweig,  
31 17 Germany 38116; email address: damien.finn@thuenen.de  
32  
33  
34  
35

36 19 Keywords:

37 20 Theoretical ecology, niche differentiation, copiotroph-oligotroph, Random Forest  
38 21 modelling  
39  
40  
41  
42  
43

44 24 Abstract

45 25 Functional, physiological traits are the underlying drivers of niche differentiation. A  
46 26 common framework related to niches occupied by terrestrial prokaryotes is based on  
47 27 copiotrophy or oligotrophy, where ~~genomic-resource~~ investment is primarily in either  
48 28 ~~metabolism-rapid growth~~ or stress tolerance, respectively. A quantitative trait-based  
49 29 approach ~~was employed to demonstrate quantitativesought~~ relationships between  
50 30 taxa, traits and niche in terrestrial prokaryotes. With 175 taxa from 11 Phyla and 35  
51 31 Families ( $n = 5$  per Family), traits were considered as discrete counts of shared  
52 32 genome-encoded proteins. ~~The composition of 220 664 traits~~ Trait composition strongly  
53 33 supported non-random ~~trait-functional~~ distributions as preferential clustering of related  
54 34 taxa via unweighted pair-group method with arithmetic mean. Trait similarity between  
55  
56  
57  
58  
59  
60



1  
2  
3 35 taxa increased as taxonomic rank decreased. A suite of Random Forest models  
4 identified traits significantly enriched or depleted in taxonomic groups. Many  
5 traits These traits conveyed functions related to metabolism rapid growth, nutrient  
6 acquisition and stress tolerance consistent with their presence in copiotroph-oligotroph  
7 niches. Finally, h Hierarchical clustering of these traits identified a clade of competitive,  
8 copiotrophic Families resilient to oxidative stress *versus* glycosyltransferase-enriched  
9 oligotrophic Families resistant to antimicrobials and environmental stress. However,  
10 the formation of five clades, including an Actinobacteria and 'specialised metabolic  
11 function' clade, warrants suggested a more nuanced view of 'life strategies' to describe  
12 niche differentiation in terrestrial systems is necessary. We suggest considering traits  
13 involved in both resource investment and acquisition when predicting niche.  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26

## 27 48 1. Introduction

28 49 Niche differentiation, the process of physiologically distinct organisms adapting better  
29 50 to certain conditions, is a contributing factor to the high biodiversity inherent in microbial  
30 51 communities (Prosser, 2012). Such differentiation is likely an inevitable consequence  
31 52 of the principles of competitive exclusion and natural selection working in tandem – no  
32 53 two organisms can theoretically occupy the same niche, as the poorer competitor must  
33 54 either adapt to a unique niche or be driven to extinction in that system (Gause, 1932,  
34 55 Hutchinson, 1957, Leibold, 1995). The physiological traits driving niche differentiation  
35 56 must have the capacity to convey an advantage to the organism's ability to survive and  
36 57 reproduce (*i.e.* fitness) and be inherited by successive generations (McGill *et al.*, 2006).  
37 58 Importantly, this implies that microbial communities are not only diverse in terms of  
38 59 individual 16S rRNA gene sequences, commonly used to assess community diversity,  
39 60 but also diverse in regard to their physiological traits.

40 61 Explaining niche differentiation through the functional, physiological traits present in  
41 62 ecological community members has a long history in macroecology. For example,  
42 63 differences between beak size and shape differed between in Galápagos finches was  
43 64 instrumental in Darwin's hypothesis that a common ancestor had differentiated into  
44 65 multiple, island-specific species. Within the past century, trait-based analyses have  
45 66 been particularly predominant in plant ecology, with seed germination in submerged  
46 67 soil, salt tolerance, carbon to nitrogen biomass stoichiometry, and leaf mass per unit  
47 68 area acting as examples of traits linked to niche differentiation (Gleason, 1926, Grime,

1  
2  
3 69 1979, Keddy, 1992). In contrast, trait-based approaches to explain microbial ecology  
4  
5 70 have only been performed in few instances, such as conceptualizing niches of  
6  
7 71 methanotrophs based on abundance in high *versus* low methane environments or  
8  
9 72 disturbed *versus* undisturbed soils (Ho *et al.*, 2013), correlating increasing growth rate  
10  
11 73 with increasing ribosomal gene and ribosome-associated gene copy number (Vieira-  
12  
13 74 Silva & Rocha, 2010), deterministic modelling of nitrification rate based on ammonia  
14  
15 75 and oxygen uptake rate, temperature sensitivity and growth rate (Bouskill *et al.*, 2012),  
16  
17 76 defining distinct niches of 32 marine microorganisms based on clustering of genome-  
18  
19 77 encoded functional proteins (Lauro *et al.*, 2009), [identifying habitat generalists and](#)  
20  
21 78 [specialists based on taxon co-occurrence patterns \(Barb eran \*et al.\*, 2012\)](#) and recently  
22  
23 79 comparisons of 23 'core' traits (e.g. motility, carbon metabolism, optimal pH for growth)  
24  
25 80 across 15 000 diverse host-associated and environmental genomes (Madin *et al.*,  
26  
27 81 2020).

28  
29 82 A consistent trend noted in macroecology is that traits linked to how carbon and  
30  
31 83 energy is processed and allocated to biomass can describe separate niches (Brown *et*  
32  
33 84 *al.*, 2004). The canonical example are *r* and *K* strategists, where carbon and energy  
34  
35 85 are primarily invested in reproduction, or alternatively invested in tolerating biotic  
36  
37 86 and/or abiotic stressors, respectively (Grime, 1977). These dichotomous strategies  
38  
39 87 have been observed in microbial ecology: copiotrophs are considered as  
40  
41 88 microorganisms with relatively high growth rates that have relatively poor growth  
42  
43 89 efficiency (as carbon incorporated to biomass per unit resource), relatively high cell  
44  
45 90 maintenance energy costs, dependence on relatively high concentrations of organic  
46  
47 91 carbon in their environment, demonstrate rapid population blooms upon the addition of  
48  
49 92 organic matter and are not overly tolerant of abiotic stress (Semenov, 1991, Koch,  
50  
51 93 2001, Roller & Schmidt, 2015, Ho *et al.*, 2017). Oligotrophs are considered as the  
52  
53 94 inverse – low growth rate, high growth efficiency, low cell maintenance energy  
54  
55 95 requirements, high substrate uptake affinity, [slow](#) growth ~~at a slow~~ yet [at a](#) consistent  
56  
57 96 rate and are resilient to abiotic stress. Although the niche concept in macroecology has  
58  
59 97 a formalized definition founded on where a taxon can maintain a stable population  
60  
61 98 within multi-dimensional environmental space (Leibold, 1995), in this study, niche is  
62  
63 99 used simply to distinguish between prokaryotes being relatively more copiotrophic  
64  
65 100 *versus* oligotrophic.

66  
67 101 These distinct niches became associated with specific terrestrial taxa at high  
68  
69 102 taxonomic rank based on recent molecular analyses. In complex microbial

1  
2  
3 103 communities, the relative abundance of Gammaproteobacteria, Bacteroidetes and  
4 104 Actinobacteria were correlated with rapid growth in response to the addition of labile  
5 105 organic matter or nitrogen (copiotrophs) (Fierer *et al.*, 2007, Goldfarb *et al.*, 2011,  
6 106 Fierer *et al.*, 2012, Leff *et al.*, 2015). Conversely, the Deltaproteobacteria,  
7 107 Acidobacteria, Verrucomicrobia and Planctomycetes were negatively correlated with  
8 108 the addition of organic matter or nitrogen (oligotrophs) (Fierer *et al.*, 2007, Fierer *et al.*,  
9 109 2012, Leff *et al.*, 2015, Bastida *et al.*, 2016). Conflicting reports exist of Beta- and  
10 110 Alphaproteobacteria, with some studies describing them as copiotrophic and others as  
11 111 oligotrophic highlighting that a consistent niche may not necessarily exist across  
12 112 species within a large taxonomic group (Ho *et al.*, 2017). A genomic basis for traits  
13 113 associated with soils dominated by putative copiotrophs and oligotrophs has been  
14 114 expertly reviewed elsewhere, and interested readers are referred to Trivedi *et al.*,  
15 115 (2013) and references therein. Importantly, these observations suggest that specific  
16 116 traits that allow terrestrial prokaryotes to occupy these two niches should (generally)  
17 117 be associated with taxonomy. This is an example of ecological coherence at high  
18 118 taxonomic rank, whereby members within a taxon tend to have similar life strategies,  
19 119 niches and possess common traits compared to members of other taxa (Philippot *et*  
20 120 *al.*, 2010). While ecological coherence of taxa has been considered previously, the  
21 121 shared, specific traits that drive niche differentiation in terrestrial prokaryotes remains  
22 122 an open question.

23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38 123 To identify the traits that differ between terrestrial prokaryote taxonomic groups, and  
39 124 whether these traits could describe the niches they occupy, a functional trait-based  
40 125 approach was adopted here. We posited that a trait must: a) be associated with a  
41 126 physiological process that conveys a fitness advantage under certain environmental  
42 127 conditions; b) be measurable in well-defined units; and c) vary more between  
43 128 taxonomic groups than within a taxonomic group (McGill *et al.*, 2006, Kearney *et al.*,  
44 129 2010). Traits were measured as discrete counts of chromosome-encoded proteins  
45 130 shared between at least two of 175 terrestrial prokaryotes. Markov Chain clustering  
46 131 (MCL) was used to group proteins as traits based on amino acid sequence similarity  
47 132 (%) akin to a previous approach that confirmed close taxonomic relatives tend to share  
48 133 functional traits in 1,374 genomes (Zhu *et al.*, 2015). This was necessary to compare  
49 134 highly similar (but non-identical) proteins from separate genomes that carry out the  
50 135 same biological function. To better identify important, distinguishing traits of terrestrial  
51 136 prokaryotes, this study differed from Zhu *et al.*, by: a) comparing 175 publicly available

1  
2  
3 137 terrestrial prokaryote genomes from 35 Families ( $n = 5$  each), from 11 Phyla and two  
4  
5 138 Kingdoms; b) ~~the selecting~~ prokaryotes ~~selected are all~~ involved in terrestrial  
6  
7 139 ecosystem processes of interest, including organic matter decomposition, nitrogen  
8  
9 140 fixation, nitrification, denitrification, methane oxidation, plant-growth promotion,  
10  
11 141 bioremediation of pollutants, pathogenesis and methanogenesis; c) ~~selected selecting~~  
12  
13 142 taxa ~~were~~ isolated from a wide range of terrestrial environments, such as nutrient rich  
14  
15 143 decomposing plant material and rhizosphere, submerged wetland and rice paddy soils,  
16  
17 144 polluted soils, and nutrient poor hot and cold arid environments; and d) ~~avoiding~~ the  
18  
19 145 inclusion of multiple subspecies and/or strains of a single species ~~was specifically~~  
20  
21 146 ~~avoided~~ to prevent biases in analyses where highly over-represented species are  
22  
23 147 compared with species that have fewer cultured representatives. The taxonomic  
24  
25 148 system used here is from the NCBI, which is built upon a historical array of culture-  
26  
27 149 dependent, physiological observations and genetic similarity to cultured isolates as  
28  
29 150 average nucleotide identity, DNA-DNA hybridisation or 16S rRNA gene homology  
30  
31 151 (ncbi.nlm.nih.gov/Taxonomy/Browser). The use of this classification system and  
32  
33 152 comparison to others is discussed further below.

34  
35 153 We hypothesised that: 1) traits are non-randomly distributed, with relatively closely  
36  
37 154 related taxa demonstrating greater similarity than unrelated taxa (ecological  
38  
39 155 coherence); 2) traits that are differentially enriched between taxonomic groups would  
40  
41 156 primarily be involved in metabolism, nutrient acquisition and/or tolerating  
42  
43 157 environmental stress; and 3) copiotrophic and oligotrophic taxonomic groups would  
44  
45 158 emerge based on collective trait enrichment.

46  
47 159

48  
49 160

## 161 2. Methodology

### 162 2.1 Collection of terrestrial prokaryote genomes

163 A collection of 175 sequenced and annotated genomes was collated (Supplementary  
164 Table 1). Listed are the genome ID, phylogenetic lineage, role in an ecosystem process  
165 if known, and isolation or genome sequencing reference. These genomes were  
166 sourced from the National Centre for Biotechnology Information (NCBI) and Joint  
167 Genome Institute (JGI) databases. Genomes were chosen based on several criteria:  
168 a) five isolates per Family were chosen to have an equal minimum sample size per  
169 group, with this sample size being constrained by sequenced genomes of under-  
170 represented groups in public databases; b) only a single subspecies/strain per species

1  
2  
3 171 was included to avoid bias due to over-representation of some species in public  
4  
5 172 databases; c) an emphasis was placed to include isolates from diverse taxonomic  
6  
7 173 lineages involved in terrestrial ecosystem processes of interest, such as ammonia  
8  
9 174 oxidation and methanogenesis; and d) there was an emphasis to include taxonomic  
10  
11 175 groups frequently stated to be either copiotrophic (e.g. Actinobacteria,  
12  
13 176 Gammaproteobacteria) or oligotrophic (e.g. Acidobacteria, Planctomycetes) based on  
14  
15 177 observations from soil nutrient addition studies (Ho *et al.*, 2017). Taxonomic  
16  
17 178 annotations for Phyla, Class, Order etc. were based on NCBI taxonomy as most  
18  
19 179 genomes were sourced there. The authors recognise that taxonomy is constantly  
20  
21 180 shifting, particularly so with the recent development of the Genome Taxonomy  
22  
23 181 Database (GTDB) (Parks *et al.*, 2018). Of note is that the vast majority of taxa here  
24  
25 182 have the same taxonomy in NCBI as in GTDB, with the exceptions that GTDB  
26  
27 183 considers the Sporomusaceae as split into three separate Families, the  
28  
29 184 Leuconostocaceae to be Lactobacillaceae, the Promicromonosporaceae to be  
30  
31 185 Cellulomonadaceae, and the Bradyrhizobiaceae and Methylococcaceae have been  
32  
33 186 renamed as Xanthobacteraceae and Methylomonadaceae, respectively. Taxon  
34  
35 187 selection was constrained by availability of genomes for under-represented groups,  
36  
37 188 such as the Chloroflexi, Verrucomicrobia, Planctomycetes, Thaumarchaeota and  
38  
39 189 Euryarchaeota. To meet the  $n = 5$  requirement for balanced statistical analyses, it was  
40  
41 190 necessary to consider these under-represented groups as 'Families'. Furthermore, due  
42  
43 191 to the great diversity inherent within Proteobacterial Classes, Gamma-, Alpha-, Beta-  
44  
45 192 and Deltaproteobacteria were considered as independent 'Phyla' for statistical  
46  
47 193 analyses here. Indeed, GTDB now defines Deltaproteobacteria as its own Phylum,  
48  
49 194 while Betaproteobacteria are considered as the Burkholderiales Order within the  
50  
51 195 Gammaproteobacteria. The total of 175 genomes analysed here falls within the upper  
52  
53 196 range of previous hypothesis-driven trait-based studies which varies from 11 isolates  
54  
55 197 (Bouskill *et al.*, 2012) to 214 genomes (Vieira-Silva & Rocha, 2010).

198

199

## 200 *2.2 Functional trait clustering*

201 A step-by-step walkthrough of reproducible code to perform the following analyses on  
202 a subset of 12 genomes is available at: [https://github.com/DamienFinn/Trait-](https://github.com/DamienFinn/Trait-based_analyses)  
203 [based\\_analyses](https://github.com/DamienFinn/Trait-based_analyses). Firstly, a pairwise similarity comparison of all amino acid sequences  
204 (964 951 sequences) across the 175 genomes was performed with the all *versus* all



1  
2  
3 205 basic local alignment tool function for proteins, BLAST-P (Altschul *et al.*, 1990). Amino  
4  
5 206 acid sequences were subsequently clustered as traits via MCL weighted by pairwise  
6  
7 207 amino acid similarity (Enright *et al.*, 2002). Functional traits were grouped at a cluster  
8  
9 208 value of 90.2, whereby > 65 is considered 'fair' and confidence in accurately separating  
10  
11 209 clusters cannot be higher than 100. The value of 90.2 is not chosen by the user but  
12  
13 210 rather is a reflection of the quality of clustering in a given dataset. The MCL identified  
14  
15 211 a total of 220 664 traits shared between at least two genomes. A random subset of  
16  
17 212 1700 amino acid sequences were selected and the similarity of each sequence within  
18  
19 213 its trait group (as determined by MCL) *versus* between other trait groups was visualised  
20  
21 214 as a box and whisker plot (Supplementary Figure 1) in R version 4.0.0 (R Core Team,  
22  
23 215 2013). 1\_700 sequences were chosen to maximise comparisons between trait groups  
24  
25 216 under technical limitations, as increasing sequences led to exponential increases in  
26  
27 217 trait combinations. A Student's T test was applied to determine whether sequences  
28  
29 218 were more similar within their trait group relative to between trait groups in R. Finally,  
30  
31 219 a matrix of genome x functional trait was generated in a two-step process by first  
32  
33 220 associating genome IDs to the MCL output with a novel script 'MCLtoReshape2.py'  
34  
35 221 (available at the above Github address) and secondly by casting the long data format  
36  
37 222 to a wide data matrix with the 'reshape2' package in R (Wickham, 2007). Box and  
38  
39 223 whisker plots comparing counts of proteins per genome (input) and counts of functional  
40  
41 224 traits shared by at least two genomes (output of computational workflow), for the 35  
42  
43 225 Families, is presented as Figure 1.

44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

### 228 *2.3 UPGMA dendrogram of trait similarity between genomes*

229 The unweighted pair group method with arithmetic mean (UPGMA) was chosen to  
230 compare distance-based similarity between taxa based on discrete counts of individual  
231 traits per genome. This method is more robust for comparing similarity between sample  
232 units (*i.e.* taxa) based on discrete counts of variables (*i.e.* individual traits per taxon)  
233 rather than neighbour joining or maximum likelihood methods better suited for DNA or  
234 amino acid sequence comparisons (Weins, 1998). The UPGMA was performed in R  
235 with the 'phangorn' package as described (Schliep *et al.*, 2017) on a Bray-Curtis  
236 transformed dissimilarity functional trait matrix, generated with the 'vegdist' function in  
237 the 'vegan' package (Oksanen *et al.*, 2013). To measure ecological coherence (C) of  
238 taxa within shared Super Groups, Phyla and Families, a similarity index was adapted

239 from Levins' Overlap (Finn *et al.*, 2020a), which measures pairwise similarity in  
240 distributions of taxa, as the following:

$$241$$
$$242 \quad C = 1 - \left( \frac{\sum b_{ij}}{n^2} \right) \quad \text{Equation 1.}$$
$$243$$

244 Where  $b_{ij}$  is the pairwise branch length between taxon  $i$  and  $j$  in the UPGMA tree,  
245 measured here as Bray-Curtis dissimilarity, which is summed for each taxon and its  
246 relatives within a shared Super Group, Phylum or Family, and where  $n$  is the number  
247 of taxa being compared within a shared Super Group, Phylum or Family.

248 Furthermore, the full length 16S rRNA gene of each taxon was collated from NCBI.  
249 Genes were aligned with MUSCLE (Edgar, 2004) and a neighbour joining phylogenetic  
250 tree was constructed with the 'phangorn' package in R. Phylogenetic distance present  
251 in taxonomic groups ( $P$ ) was measured as per Equation 1., excepting that branch  
252 length was in units of DNA sequence similarity as opposed to Bray-Curtis dissimilarity.  
253 Finally, simple linear regression was used to test a relationship between  $P$  and  $C$ .

#### 256 2.4 Functional trait annotation

257 To inform the biological process a functional trait facilitated, traits were annotated with  
258 the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. This was  
259 performed in five steps: 1) a representative amino acid sequence from each trait was  
260 extracted with the novel script 'IdentifyTraits.py'; 2) these sequences were annotated  
261 with KEGG Orthology (KO) terms using the BlastKOALA database algorithm with a bit  
262 score cut-off value of 75 (Kanehisa *et al.*, 2016); 3) BRITE functional hierarchies  
263 associated with each KO term (e.g. KO1179 gene: endoglucanase, BRITE 1:  
264 Metabolism, BRITE 2: Carbohydrate Metabolism, BRITE 3: Starch and Sucrose  
265 Metabolism) were collected with the novel script 'GetBRITEinfo.py'; 4) Genome ID, trait  
266 ID, KO term and BRITE metadata were all collated with the novel script 'MatchFCs.py';  
267 and 5) the 'reshape2' package in R was used to create matrices of genome x BRITE  
268 hierarchy. Where KEGG was unable to annotate a trait, it was considered as  
269 'Uncharacterised'. As above, all novel scripts and a step-by-step walkthrough of  
270 reproducible code is available at: [https://github.com/DamienFinn/Trait-](https://github.com/DamienFinn/Trait-based_analyses)  
271 [based\\_analyses](https://github.com/DamienFinn/Trait-based_analyses).



273

## 2.5 Identifying traits differentially enriched in taxonomic groups

Random Forest classification was chosen as a non-linear, multivariate cluster-based method capable of identifying numerous predictor variables (*i.e.* traits) that define different classes of a response variable (*i.e.* taxonomic group). This was performed with the 'randomForest' package as described (Liaw & Weiner, 2002). Discrete counts of traits at BRITE level 3 were used (*e.g.* Starch and Sucrose Metabolism) as this level had the most accurate resolution of biological processes facilitated by traits. A total of six Random Forest models were optimised to classify taxonomic groups at the level of: a) Phylum, with Proteobacteria Classes separated due to their extensive diversity ( $n = 14$ ); b) Family ( $n = 35$ ); c) specifically for Families in the Proteobacteria ( $n = 13$ ); d) Families in the Actinobacteria ( $n = 7$ ); e) Families in the Firmicutes ( $n = 5$ ); and f) Families from 'Under-represented' groups, which were all other Families ( $n = 10$ ). Optimal numbers of trees grown for each model were: 300, 400, 320, 300, 300 and 400, respectively. Six traits were randomly selected at each branch. As the Random Forest only identifies traits that best explain separation of taxonomic groups, and does not show whether traits have positive or negative associations with groups, box and whisker plots and Fisher's Least Significant Difference (LSD) *post hoc* tests were performed with the 'agricolae' package (de Mendiburu, 2014) to definitively state which taxonomic groups were significantly enriched or depleted in traits identified via Random Forest.

294

295

## 2.6 Hierarchical clustering of Families by defining traits

Finally, the relationship between Families based on similarity in counts of 60 traits was assessed via hierarchical clustering. Traits were chosen based on being selected via the above Random Forest models in this study, and from previous studies that identified traits associated with copiotroph-oligotroph growth strategies in single species or mixed communities (Lauro *et al.*, 2009, Vieira-Silva & Rocha, 2010, Roller & Schmidt, 2015, Pascual-Garcia & Bell, 2020). The mean of trait discrete counts in the five Family members was used as representative of each Family. Comparing trait means between Families was considered acceptable as prior LSD *post hoc* tests had demonstrated significant differences between Families. As traits had highly variable copy numbers per Family (*e.g.* ABC transporter trait copies ranged from 10 – 350,

1  
2  
3 307 while bacterial chemotaxis traits ranged from 0 – 15 copies) the trait copies were  
4  
5 308 normalised for more appropriate comparisons. Normalised variance was calculated  
6  
7 309 across the 35 Families for all traits with the 'decostand' function in the 'vegan' package  
8  
9 310 (Oksanen *et al.*, 2013). Hierarchical clustering of Families based on normalised trait  
10  
11 311 counts was visualised with the 'heatmap.2' function in the 'gplots' package (Warnes *et*  
12 312 *al.*, 2019).

13 313

14 314

15 315 

### 3. Results

16 316 

#### 3.1 Trait clustering and UPGMA

17 317 The 964 951 amino acid sequences encoded by the 175 genomes were clustered as  
18  
19 318 220 664 traits by MCL. A random subset of 1\_700 traits showed that amino acid  
20  
21  
22 319 sequence similarity within traits ranged from 62.4, 82.5 and 100% for the 1<sup>st</sup> quartile,  
23  
24 320 mean and 3<sup>rd</sup> quartile, respectively (Supplementary Figure 1). Amino acid sequence  
25  
26 321 similarity between traits ranged from 27.6, 34 and 37.8% for the 1<sup>st</sup> quartile, mean and  
27  
28 322 3<sup>rd</sup> quartile, respectively. A Student's T test found that sequences grouped together as  
29  
30 323 a trait were significantly more similar to each other than to sequences grouped as  
31  
32 324 different traits ( $t$  value = 61.3,  $p = 2 \times 10^{-16}$ ). Manual comparisons of amino acid  
33  
34 325 sequences within several traits supported clustering of proteins with identical biological  
35  
36 326 function based on KEGG annotation. Thus, the MCL was considered to perform well.  
37  
38 327 However, the minimum amino acid sequence similarity within traits was 23.27% and  
39  
40 328 maximum similarity between traits was 85.24%, indicating that across the 220 664  
41  
42 329 traits, a small proportion of dissimilar amino acid sequences were grouped as a trait  
43  
44 330 incorrectly, while some amino acid sequences that were highly similar were considered  
45  
46 331 different traits. This small number of incorrectly clustered sequences can be explained  
47  
48 332 by the MCL clustering efficiency being 90.2, out of a possible 100.

49 333 A comparison of the number of proteins per genome, per Family at the Family level  
50 334 (mean = 5\_533), found that there were fewer functional traits per genome (mean = 4  
51 335 240) (Figure 1). These represent the input number of proteins per genome before trait  
52 336 clustering and the output number of traits after clustering, respectively. As only  
53 337 functional traits shared between at least two genomes were considered here, the loss  
54 338 of highly genome-specific traits that could not be compared between genomes was  
55 339 expected. Despite this drop in average traits per genome, this initial approach serves  
56  
57 340 as a proof of concept to demonstrate that numbers of traits per genome at the Family

341 level reflect trends in proteins per Family, and thus the MCL was not distorting trait  
342 clustering (Figure 1).

343 The UPGMA dendrogram comparing the 220 664 traits per genome showed that  
344 trait compositions were non-randomly distributed (Figure 2a). Specifically,  
345 Thaumarchaeota, Euryarchaeota, Acidobacteria, Betaproteobacteria,  
346 Gammaproteobacteria, Alphaproteobacteria, Cyanobacteria, Verrucomicrobia,  
347 Planctomycetes, Bacteroidetes, Actinobacteria, Firmicutes and Deltaproteobacteria  
348 clustered together preferentially. The Chloroflexi were split into two clusters: one  
349 *Dehalococcoides* and one *Ktedonobacter/Herpetosiphon/Anaerolinea* cluster. Several  
350 prokaryotes did not cluster with their high taxonomic rank, including a Planctomycetes  
351 bacterium, *Polyangium brachysporum* (Deltaproteobacteria), *Agreia pratensis*  
352 (Actinobacteria) and *Sporomusa ovata* (Firmicutes). Also of interest was that, in regard  
353 to distance between terminal nodes (as noted by the scale bar), the Betaproteobacteria  
354 and Gammaproteobacteria were more similar to each other than the  
355 Alphaproteobacteria, which formed its own large, diverse clade. A neighbour joining  
356 tree of full length 16S rRNA genes showed that all taxa clustered preferentially based  
357 on their taxonomic nomenclature at high taxonomic rank, including the Chloroflexi,  
358 indicating that the discrepancies in the UPGMA were not due to misclassification of the  
359 individual taxa (Supplementary Figure 2a).

360 A simple index to measure trait similarity, as ecological coherence ( $C$ ), within groups  
361 was devised (Equation 1).  $C$  increased as taxonomic rank decreased: Super Group <  
362 Phylum < Family (Figure 2b).  $C$  was lowest for the larger, more diverse Proteobacteria  
363 and Terrabacteria (Super Group), and Firmicutes and Actinobacteria (Phylum). As the  
364 number of taxa being compared at the Super Group (e.g. Terrabacteria = 70 *versus*  
365 Acidobacteria = 10) and Phylum (e.g. Actinobacteria = 35 *versus* Thaumarchaeota =  
366 5) were variable, the most meaningful comparisons between groups are at the Family  
367 level ( $n = 5$  each). With the exception of the highly divergent 'Acidobacteria Lineage',  
368 all Families had a  $C$  greater than 0.3, with certain groups in the Alphaproteobacteria  
369 (Beijerinckiaceae), Firmicutes (Bacillaceae and Leuconostocaceae) and  
370 Actinobacteria (Promicromonosporaceae) being highly coherent ( $C > 0.55$ ). Indeed, all  
371 individual Proteobacterial Families had  $C > 0.4$ , indicating that all five taxa within each  
372 of these Families had similar trait compositions. Despite not truly belonging to the same  
373 Family as per NCBI taxonomy, the  $C > 0.335$  of the five Cyanobacteria,  
374 Thaumarchaeota and the Methanogen Lineage taxa was similar to other Families from

1  
2  
3 375 the Bacteroidetes and Firmicutes. Thus, the UPGMA demonstrated that taxonomic  
4 376 relatives at the Phylum level tended to cluster with each other preferentially based on  
5 377 trait composition, and secondly that while similarity was broadly highest at low  
6 378 taxonomic rank, some Families were more coherent than others.

7  
8  
9  
10 379 Phylogenetic distance ( $P$ ) of each taxonomic group increased with decreasing  
11 380 taxonomic rank, and was highest in Proteobacteria, Actinobacteria and Firmicutes  
12 381 Families (Supplementary Figure 2b). There was a strong positive linear relationship  
13 382 between  $P$  and  $C$  ( $y = 0.86x - 0.25$ ,  $R^2 = 0.39$ ,  $p < 0.001$ ) supporting the result that  
14 383 taxonomic groups of closer related taxa tended to share more similar compositions of  
15 384 traits.

16 385

17 386

### 18 387 *3.2 Random Forest trait identification*

19 388 The KEGG annotated traits belonged to 260 different BRITE 3 categories. The  
20 389 percentage of traits that could not be annotated (and were termed 'Uncharacterised')  
21 390 ranged from 28 – 65% per genome, being particularly high in the Archaea. On average,  
22 391 47% of traits per genome were Uncharacterised with a standard deviation of 9.5%.

23 392 Of the 260 BRITE 3 categories, the 60 most important traits in separating all Phyla  
24 393 and Families are ranked by importance measured as Mean Decrease in Accuracy  
25 394 (MDA) of the Random Forest models (Figure 3). This is a measure of the average  
26 395 increase in classification error during permutation of trees ( $n = 300 - 400$ ) when that  
27 396 particular trait is missing from the tree. For example, the accuracy of classifying  
28 397 Families was most improved by inclusion of the ABC transporters trait. Based on the  
29 398 identified traits, the Phylum model was capable of successfully classifying 81.14% of  
30 399 individual taxa. The Family model was capable of successfully classifying 71.43% of  
31 400 individual taxa. Confusion matrices for both models are presented as Supplementary  
32 401 Tables 2 – 4, and show that classification was particularly difficult for Chloroflexi and  
33 402 Planctomycetes (classification error > 80%) in the Phylum model and for  
34 403 Cellulomonadaceae and the divergent Acidobacteria Lineage (classification error >  
35 404 80%) in the Family model. Random Forest models were robust against variation in  $P$   
36 405 within Families, for example the nine families with all taxa perfectly classified ranged  
37 406 in  $P$  from the lowest (0.68) to highest (0.9).

38 407 The important traits in classifying the taxonomic groups were involved in: a)  
39 408 metabolism and nutrient acquisition (oxidative phosphorylation, tricarboxylic acid

1  
2  
3 409 (TCA) cycle, glyoxylate/decarboxylate, thermogenesis, propanoate, starch/sucrose,  
4  
5 410 nitrogen, methane metabolism, synthesis of antioxidants such as glutathione, ATP-  
6  
7 411 binding cassette (ABC) transporters, sugar uptake via phosphotransferase systems  
8  
9 412 (PTS)); b) responding to environmental cues and stressors (protein kinases, two-  
10  
11 413 component systems, transcription factors, proteasome, protein chaperones, RNA  
12  
13 414 transport, chromosome repair via non-homologous DNA end joining); c) core cell  
14  
15 415 physiology (flagella assembly, chemotaxis, sporulation, lipopolysaccharide (LPS),  
16  
17 416 peptidoglycan, glycerolipid, sphingolipid and lipoarabinomannan (LAM) biosynthesis);  
18  
19 417 and d) cell-cell interactions (beta-Lactam resistance, general secretion systems and  
20  
21 418 Type IV secretion systems). Box and whisker plots of discrete counts of identified traits,  
22  
23 419 and LSD results, are provided as Supplementary Figures [42](#) – [75](#). The Families  
24  
25 420 significantly enriched and depleted in these traits are listed in Table 1. The Phyla  
26  
27 421 significantly enriched and depleted in identified traits are listed in Supplementary Table  
28  
29 422 9.

30  
31 423 To better identify the more subtle differences between Families in the  
32  
33 424 Proteobacteria, Actinobacteria, Firmicutes and the 'Under-represented' Phyla,  
34  
35 425 individual Random Forest models were constructed for each of the four groups. The  
36  
37 426 successful classification rates were 72.31, 74.29, 76 and 72%, respectively. Confusion  
38  
39 427 matrices for each model are presented as Supplementary Tables 5 – 8. The models  
40  
41 428 were unable to reliably classify Bradyrhizobiaceae and the divergent Acidobacteria  
42  
43 429 Lineage (classification error > 80%).

44  
45 430 Supplementary Figure [86](#) shows the most important traits in classifying the four  
46  
47 431 groups. Unique traits not identified in the prior Phylum and Family models were: a) for  
48  
49 432 Proteobacteria, glycosyltransferases, butanoate metabolism, aminotransferases,  
50  
51 433 ribosome biogenesis, mRNA biogenesis and degradation; b) for Actinobacteria,  
52  
53 434 porphyrin and chlorophyll synthesis, pyruvate metabolism, aminotransferases, fatty  
54  
55 435 acid and aliphatic hydrocarbon metabolism, polyketide and Type II polyketide  
56  
57 436 biosynthesis, antimicrobial resistance genes; c) for Firmicutes, lysine, folate and varied  
58  
59 437 amino acid synthesis, porphyrin and chlorophyll synthesis, DNA replication, bacterial  
60  
438 toxins, penicillin and cephalosporin synthesis; and d) for the 'Under-represented' taxa,  
439  
440 glycosyltransferases, peptidases and inhibitors, photosynthesis and AMP-activated  
441  
442 protein kinases. Box and whisker plots of discrete counts of identified traits, and LSD  
443  
444 results, are provided as Supplementary Figures [97](#) – [120](#). Tables summarising



1  
2  
3 442 Families enriched and depleted in these traits are included as Supplementary Tables  
4 443 10 and 11.

5 444

6 445

### 7 446 *3.3 Hierarchical clustering of defining traits*

8 447 Hierarchical clustering based on 60 traits, identified from Random Forest in this study  
9 448 and by previous copiotroph-oligotroph studies, indicated five general clades. The  
10 449 dendrogram on the y axis of Figure 4 shows clustering of taxa as these five clades.  
11 450 The dendrogram on the x axis shows clustering of co-occurring traits. Clade I consisted  
12 451 of Proteobacteria, specifically the Pseudomonadaceae, Burkholderiaceae,  
13 452 Rhodospirillaceae, Bradyrhizobiaceae and Rhizobiaceae. These Families were  
14 453 uniquely enriched in flagellar assembly, chemotaxis, pyruvate metabolism, glutathione  
15 454 metabolism, ABC transporters, benzoate metabolism, transcription factors, glyoxylate  
16 455 and fatty acid metabolism. Clade II, also Proteobacteria, included Nitrosomonadaceae,  
17 456 Neisseriales Lineage, Methylocystaceae, Beijerinckiaceae, Methylococcaceae and  
18 457 Moraxellaceae. These Families clustered based on being enriched in Clade I traits, but  
19 458 to a lesser degree than the Pseudomonadaceae, Burkholderiaceae,  
20 459 Rhodospirillaceae, Bradyrhizobiaceae and Rhizobiaceae. Exceptions included the  
21 460 absence of benzoate metabolism and enrichment of methane metabolism in several  
22 461 Clade II Families.

23 462 Clade III, a diverse collection of Bacteroidetes (Chitinophagaceae, Cytophagaceae),  
24 463 Verrucomicrobia, Planctomycetes, divergent Acidobacteria Lineage and the  
25 464 Deltaproteobacteria (Polyangiaceae, Myxococcaceae), shared enrichment of  
26 465 sphingolipid metabolism, beta-Lactam resistance, penicillin and cephalosporin  
27 466 biosynthesis, LPS biosynthesis, glycosyltransferases and starch/sucrose metabolism.  
28 467 Many of these Families shared Clade I and II traits, including Type IV secretion system,  
29 468 oxidative phosphorylation, TCA cycle, PTS, nitrogen and glycerophospholipid  
30 469 metabolism. The absence of glutathione in non-Deltaproteobacterial Clade III Families  
31 470 was notable.

32 471 The three Actinobacteria Families in Clade IV, Mycobacteriaceae, Frankiaceae and  
33 472 Streptomycetaceae, were highly similar to each other because they were enriched in  
34 473 Type II polyketide biosynthesis. They also shared some Clade I traits (ABC  
35 474 transporters, transcription factors, pyruvate, benzoate and fatty acid metabolism) and  
36 475 Clade III traits (membrane trafficking, transcription machinery, polyketide biosynthesis

1  
2  
3 476 and starch/sucrose metabolism). Similar to Clade III, these Actinobacteria were also  
4  
5 477 depleted in glutathione traits.

6 478 Finally, taxa within Clade V were similar to each other due to being depleted in traits  
7  
8 479 shared among the other clades. Cyanobacteria were the only taxa that possessed  
9  
10 480 photosynthesis traits. Lactic acid bacteria (Lactobacillaceae and Leuconostocaceae)  
11  
12 481 were enriched in Unclassified nucleotide metabolism. The Archaea (Thaumarchaeota  
13  
14 482 and Methanogen Lineage) shared eukaryote-like traits, proteasome, basal  
15  
16 483 transcription factors and RNA transport. The Archaea were enriched in carbon fixation  
17  
18 484 traits. Methanogens were also enriched in methane metabolism. Ammonia oxidising  
19  
20 485 Thaumarchaeota were not enriched in nitrogen metabolism, however they were  
21  
22 486 enriched in traits annotated by KEGG as Global maps only (unclassified metabolism).  
23  
24 487 Further analysis found this to be the *fpr* gene, encoding a ferredoxin-flavodoxin NADP<sup>+</sup>  
25  
26 488 reductase (K00528). Non-lactic acid bacteria of the Firmicutes (Bacillaceae,  
27  
28 489 Sporomusaceae and Clostridiaceae) were enriched in sporulation and motility traits.

29

30

## 31 492 4. Discussion

### 32 493 4.1 Non-random trait clustering demonstrates ecological coherence of taxa

34 494 A trait-based approach to investigate taxonomic relationships and potential biological  
35  
36 495 function was carried out with a collection of 175 terrestrial prokaryotes. We  
37  
38 496 hypothesised that traits would be non-randomly distributed amongst taxonomic groups,  
39  
40 497 ~~supportive of~~ supported by previous observations that noted closely related taxa are  
41  
42 498 isolated from similar habitats (Philippot *et al.*, 2010). Similarity in the composition of  
43  
44 499 220 664 traits, within 175 taxa, demonstrated strong agreement with established  
45  
46 500 taxonomy at high (Phyla) and low (Family) rank (Figure 2a). The exceptions to this  
47  
48 501 ecological coherence at high rank were the division of the Chloroflexi and an individual  
49  
50 502 from each of the Planctomycetes, Deltaproteobacteria, Actinobacteria and Firmicutes  
51  
52 503 clustering with unrelated Phyla. ~~It is possible that these four individual taxa have~~  
53  
54 504 ~~been~~ These taxa were not mischaracterised, as based on phylogenetics of the full  
55  
56 505 length 16S rRNA gene (Supplementary Figure 2a). Trait similarity between related  
57  
58 506 taxa, measured as C, tended to be highest at low rank (Figure 2b). Uneven sample  
59  
60 507 sizes between groups within Super Group and Phyla make comparisons at this level  
508  
509 508 difficult – the inclusion of many diverse Firmicutes and Actinobacteria likely drove C to  
509  
509 509 be lower here than in Thaumarchaeota and Euryarchaeota. However, equal



1  
2  
3 510 comparisons at the Family level demonstrated interesting variability in coherence. All  
4  
5 511 Proteobacterial Families had relatively high coherence ( $C = 0.4 - 0.6$ ). The high  $C$  in  
6  
7 512 Beijerinckiaceae is of particular interest as this group contained both specialist  
8  
9 513 methanotroph (*Methylocapsa*, *Methylocella*), methylotroph (*Methyloferula*) and  
10  
11 514 generalist heterotroph (*Beijerinckia* spp.) taxa. With such varied metabolic traits, one  
12  
13 515 could reasonably expect  $C$  to be relatively low within this Family. The Beijerinckiaceae  
14  
15 516 appear to have evolved from a common methylotrophic ancestor and still share traits  
16  
17 517 for nitrogen fixation and tolerance for low pH soils (Tamas *et al.*, 2014), and the high  
18  
19 518  $C$  measured here indicates that many additional shared traits remain. Both the  
20  
21 519 relatively recent divergence of Families from a common ancestor and the higher  
22  
23 520 number of shared traits are likely causes of the higher  $C$  observed at low taxonomic  
24  
25 521 rank. The differing values of  $C$  for Methanogen and photosynthetic Cyanobacteria  
26  
27 522 functional groups (0.33 and 0.44, respectively) is also worthy of note. Despite all five  
28  
29 523 taxa in each group performing the same core role in a community, the individual  
30  
31 524 isolates came from varying environments. The methanogens were isolated from a  
32  
33 525 range of geographically separate wetlands, rice paddy soil and farm slurry and, while  
34  
35 526 the Cyanobacteria were also isolated from geographically separate environments, they  
36  
37 527 were all from sandy deserts or other nutrient poor, arid soils (Supplementary Table 1  
38  
39 528 and references therein). Ultimately a taxon's trait composition will be affected by its  
40  
41 529 functional role in a community, its evolutionary life-history (e.g. Beijerinckiaceae  
42  
43 530 described above) and its local environment.

44  
45 531 \_\_ However, these results are dependent on accurate taxonomic classification, and the  
46  
47 532  $C$  of Sporomusaceae, relatively low compared to other Families here, supports splitting  
48  
49 533 this group into Sporomusaceae, Anaeromusaceae and Pelosinaceae by GTDB (Parks  
50  
51 534 *et al.*, 2018). Finally worth noting, some groups at high rank were considered as  
52  
53 535 'Families' here due to the number of available terrestrial genomes, e.g. Cyanobacteria  
54  
55 536 and Chloroflexi. Even so, Cyanobacteria demonstrated a higher  $C$  than many  
56  
57 537 taxonomically-defined Families, perhaps due to their common role as primary  
58  
59 538 colonisers of nutrient-poor soils (Garcia-Pichel *et al.*, 2001). The number of Families  
60  
539 are too numerous to discuss each at length here, but  $C$  was an effective means of  
540  
541 measuring and comparing coherence between groups in the UPGMA tree.

542 While the method of comparing taxa here differs from other studies, the results were  
543 not surprising as many 16S rRNA gene surveys of terrestrial systems consistently  
demonstrate ecological coherence at high rank. For example, independent studies of

1  
2  
3 544 increasing agricultural intensity in soils show reductions in Actinobacteria abundance  
4 545 (Philippot *et al.*, 2009, Jangid *et al.*, 2011). Nitrogen addition to soils frequently enriches  
5 546 numerous taxa within the Actinobacteria and Proteobacterial Classes while negatively  
6 547 affecting taxa within the Verrucomicrobia and Planctomycetes (Wessen *et al.*, 2010,  
7 548 Fierer *et al.*, 2012, Leff *et al.*, 2015, Bastida *et al.*, 2016). Arid, nutrient poor  
8 549 environments select for Actinobacteria-dominated communities (Cary *et al.*, 2010,  
9 550 Crits-Christoph *et al.*, 2013) and, in the absence of other primary producers, allow  
10 551 biological soil crust forming Cyanobacterial taxa to establish (Garcia-Pichel *et al.*,  
11 552 2001). Anoxic wetland and rice paddy environments support diverse communities of  
12 553 anaerobic Firmicutes, Chloroflexi and methanogenic Archaea (He *et al.*, 2019, Finn *et*  
13 554 *al.*, 2020**b**). These trends were noted prior to bioinformatic advances of metagenome  
14 555 assembled genomes (MAGs) that allow for the specific comparison of individual traits  
15 556 between uncultured environmental prokaryote genomes (Hug *et al.*, 2013). The  
16 557 generation of MAGs has emerged as a useful tool for identifying traits necessary for  
17 558 life in such environments, and particularly for expanding knowledge of severely under-  
18 559 represented, difficult to culture taxonomic groups. For example, the recent  
19 560 reconstruction of 52 515 MAGs from a wide range of host-associated and  
20 561 environmental metagenomes was able to increase genomic information of  
21 562 Planctomycetes and Verrucomicrobia by 79% and 68%, respectively (Nayfach *et al.*,  
22 563 2021). ~~I-but~~ importantly, both 16S rRNA gene surveys and MAGs ~~the earlier studies~~  
23 564 demonstrated that some functional traits that facilitate life under certain environmental  
24 565 conditions are intrinsically linked to taxonomy.

25 566 The ecological coherence observed in Figure 2 does not imply that closely related  
26 567 taxa (e.g. *Bacillus velezensis* LS69 and *Bacillus amyloliquefaciens plantarum* FZB42)  
27 568 have identical phenotypes. Close relatives possess a combination of core and  
28 569 accessory genes (traits) and the presence of even a single accessory gene is sufficient  
29 570 to dramatically alter a strain's phenotype (van Rossum *et al.*, 2020). Rather, our results  
30 571 (Figure 2) demonstrate that the composition of core and accessory traits in  
31 572 alphaproteobacterial Beijerinckiaceae are most similar to each other, relative to  
32 573 alphaproteobacterial Rhizobiaceae or to Actinobacteria, Firmicutes etc.

33 574

34 575

35 576 *4.2 What can the Random Forests tell us?*

1  
2  
3 577 The Random Forest works by identifying the traits that are most reliable in classifying  
4 578 individual Phyla and Families. It selects traits that tend to be: a) of equal copy number  
5 579 per genome within a taxonomic group; and b) that differ markedly in copy number  
6 580 between taxonomic groups, since distinct separation of copies will maximise  
7 581 successful classification. A clear example of this is the consistent identification of  
8 582 eukaryote-like basal transcription factors, proteasome and RNA transport present in  
9 583 the Thaumarchaeota and Euryarchaeota, since they are absent from the majority of  
10 584 Bacterial Phyla. The binary nature of these traits (yes Archaea, no Bacteria) make  
11 585 them strong indicators. The presence of these and more eukaryote-like vesicular  
12 586 trafficking and actin traits have been noted in the Archaeal TACK super-phylum  
13 587 previously, and lend credence to the hypothesis that eukaryotes are descended from  
14 588 Archaea (Embley & Martin, 2006, Spang *et al.*, 2015). However, ~~it must be noted that~~  
15 589 the Random Forest will not identify a trait unique to *Can. Nitrosotalea devanterra*, ~~and~~  
16 590 ~~that is~~ absent from other Thaumarchaeota, as this single trait will not improve  
17 591 classification of the group as a whole. Consequently, the traits identified via Random  
18 592 Forest all tended to be core, fundamental traits shared by other members of a taxon's  
19 593 Phylum/Family.

20  
21  
22 594 Many of the best traits for distinguishing taxa have been historically used by  
23 595 microbiologists to do exactly that. These included fundamental cell physiology traits,  
24 596 such as oxidative phosphorylation, LPS biosynthesis, sporulation, flagellar assembly  
25 597 and chemotaxis. The Phylum model separated Betaproteobacteria, Actinobacteria and  
26 598 Bacteroidetes as taxa with the highest copies of oxidative phosphorylation traits.  
27 599 Firmicutes, Chloroflexi and Methanogens were identified as anaerobes depleted in  
28 600 oxidative phosphorylation, and all other taxa as sitting in between (Supplementary  
29 601 Figure 42 and Supplementary Table 9). Some Gram-negative Families were  
30 602 significantly enriched in LPS biosynthesis compared to others. These were the  
31 603 Pseudomonadaceae, Burkholderiaceae, Chitinophagaceae, divergent Acidobacteria  
32 604 and Verrucomicrobia (Table 1). This has been noted in soil communities previously.  
33 605 The extensive repertoire of LPS-associated genes in Chitinophagaceae, Acidobacteria  
34 606 and Verrucomicrobia likely play a critical role in enhancing soil aggregation (Cania *et*  
35 607 *al.*, 2019) potentially through high LPS production and/or biofilm formation (discussed  
36 608 further below). In a demonstration of the robustness of the methods used here, the  
37 609 highly unusual Firmicute Sporomusaceae were shown to possess similar counts of  
38 610 LPS biosynthesis traits relative to most Gram-negative Families (Supplementary Table

1  
2  
3 611 4) in addition to sharing heat-resistant spore formation with Bacillaceae and  
4 612 Clostridiaceae (Table 1). The presence of both traits in a single Family have been  
5 613 remarked upon previously and used to conceptualise the evolution of Gram-negative  
6 614 *versus* Gram-positive lineages (Stackebrandt *et al.*, 1985). The Sporomusaceae were  
7 615 also shown to have high Porphyrin and Chlorophyll metabolism traits in the Firmicutes  
8 616 model (Supplementary Table 10). The capacity to dechlorinate the soil pollutant  
9 617 perchloroethene to trichloroethylene via a porphyrin-based corrinoid is yet another  
10 618 interesting trait of this Family (Terzenbach & Blaut, 1994).

11 619 Finally, flagella assembly and chemotaxis traits identified Alphaproteobacteria,  
12 620 Betaproteobacteria and Acidobacteria as Phyla that were particularly enriched with this  
13 621 mechanism of motility, while bacterial Actinobacteria, Bacteroidetes, Cyanobacteria,  
14 622 Chloroflexi and Verrucomicrobia were depleted (Supplementary Table 9). The  
15 623 Proteobacteria, Firmicutes and Under-represented models were better suited for  
16 624 identifying specific Families homogenously enriched or depleted in bacterial flagella  
17 625 and chemotaxis (Supplementary Tables 10 and 11). Enriched Families included the  
18 626 Rhodospirillaceae, Nitrosomonadaceae, Neisseriales lineage, divergent Acidobacteria  
19 627 lineage, Sporomusaceae, Bacillaceae, Clostridiaceae and Planctomycetes. Other  
20 628 forms of motility such as twitching and gliding have been noted in the  
21 629 Pseudomonadaceae, Myxococcaceae and Cyanobacteria (McBride, 2001) but these  
22 630 traits were not identified by the Random Forest as being homogenously enriched in  
23 631 any Families. Furthermore, while Thaumarchaeota and Methanogens were both  
24 632 identified as being depleted in bacterial flagella assembly and chemotaxis traits  
25 633 (Supplementary Table 9 and 11), Archaea possess a structurally distinct flagellum  
26 634 more similar to the Type IV bacterial pilus (Jarrell and Albers, 2012). These taxa were  
27 635 not enriched with Type IV pilus, either, and it is possible that archaeal flagella may  
28 636 have failed proper characterisation by KEGG. Thus, while some Families were  
29 637 relatively enriched/depleted in bacterial flagella and chemotaxis traits, specific taxa  
30 638 depleted in these are not necessarily non-motile.

31 639 In summary, while the Random Forests may overlook certain traits in individual taxa,  
32 640 the models were highly robust in detecting conserved, shared traits within a  
33 641 Phylum/Family. Here the 'depth' of shared traits is limited by the number of taxa that  
34 642 could be considered as Phylum or Family. In future, if five (or more) taxa belonging to  
35 643 the same Genus or even Species could be compared, unique traits would be observed  
36 644 to explain how these subgroups have evolved from their respective Families to occupy

1  
2  
3 645 distinct niches. Ideally the selection of individual taxa within groups for such future  
4 comparative analyses would also be standardised based on phylogenetic distance,  
5 646 either with *P* or a similar method, that would improve the robustness of trait-based  
6 comparisons at such a fine taxonomic level.  
7  
8  
9 648

10 649

11 650

### 13 651 4.3 Plant-derived carbon metabolism and nutrient acquisition

15 652 Secondly, we hypothesised that the traits differentially enriched between taxonomic  
16 653 groups would largely reflect those associated with copiotrophs or oligotrophs, namely  
17 654 metabolism, nutrient acquisition and environmental stress response and tolerance. Of  
18 655 fundamental interest to soil microbiologists is the decomposition of plant biomass. This  
19 656 is the primary source of organic carbon to non-arid terrestrial systems (Kögel-Knabner,  
20 657 2002) and the transformation of plant material to substrates bioavailable for  
21 658 microorganisms is essential for community growth and activity. The traits involved in  
22 659 plant material catabolism belonged to the BRITE categories 'Starch and Sucrose  
23 660 Metabolism' (e.g. extracellular cellobiosidases, endoglucanases, glucosidases,  
24 661 trehalases, amylases) and 'Glycosyltransferases', all of which are carbohydrate  
25 662 activated enzymes (CAZy). The Families particularly enriched in these traits were the  
26 663 Polyangiaceae, Myxococcaceae, Rhizobiaceae, Streptomycetaceae,  
27 664 Mycobacteriaceae, Frankiaceae and Verrucomicrobia (Table 1, Supplementary Table  
28 665 11). Genomic and culture-dependent analyses support *Sorangium cellulosum*  
29 666 (Polyangiaceae), *Streptomyces coelicolor* A3(2) (Streptomycetaceae) and  
30 667 *Chthoniobacter flavus* (Verrucomicrobia) as having particularly large genomes with  
31 668 extensive repertoires for cellulose, hemicellulose, pectin and lignin degradation  
32 669 (Bentley *et al.*, 2002, Sangwan *et al.*, 2004, Schneiker *et al.*, 2007). Comparative  
33 670 genomics analyses have also identified Actinobacteria, Acidobacteria and  
34 671 Verrucomicrobia as being enriched in numerous enzymes for cellulose, hemicellulose  
35 672 and starch catabolism (Trivedi *et al.*, 2013). *In situ* these Families likely play a critical  
36 673 role in making organic carbon bioavailable as di- and monosaccharides for the  
37 674 microbial community.

38 675 The Families equipped with many copies of high-affinity sugar uptake  
39 676 'Phosphotransferase systems' (3 – 14 copies) did not necessarily correspond to those  
40 677 enriched with CAZy – only the Myxococcaceae and Verrucomicrobia were enriched in  
41 678 both. Pseudomonadaceae, Rhodospirillaceae, Neisseriales lineage and



1  
2  
3 679 Clostridiaceae were only enriched in PTS. Despite being enriched in CAZy, the  
4  
5 680 Frankiaceae were simultaneously depleted in PTS. The complex interplay between  
6  
7 681 taxa capable of producing extracellular CAZy and competitors that rapidly scavenge  
8  
9 682 available di- and monosaccharides has been well described by models that predict  
10  
11 683 such competitive interactions exert important controls on the growth rate of the  
12  
13 684 community as a whole (Freilich *et al.*, 2011) and may even act to aid terrestrial carbon  
14  
15 685 storage and limit carbon dioxide emissions from microbial respiration (Kaiser *et al.*,  
16  
17 686 2015). Here, we identified the ‘specialist’ Families enriched in CAZy traits *versus* the  
18  
19 687 ‘opportunists’ scavenging for sugars via PTS (Table 1, Supplementary Table 11).

20  
21 688 ABC transporters facilitate the ATP-dependent uptake of soluble compounds across  
22  
23 689 membranes or export waste metabolites, extracellular enzymes and toxins (Young &  
24  
25 690 Holland, 1999, Higgins, 2001). This means of active transport allows microorganisms  
26  
27 691 to acquire nutrients with high affinity at concentrations of 5 – 500  $\mu\text{g carbon L}^{-1}$  *versus*  
28  
29 692 the less efficient diffusion of nutrients across membranes, dependent on extracellular  
30  
31 693 concentrations of 0.5 – 5 mg carbon  $\text{L}^{-1}$  (Kuznetsov *et al.*, 1979). In the spatially  
32  
33 694 heterogenous soil environment where the concentration of bioavailable carbon  
34  
35 695 substrate often limits growth (Blagodatsky & Richter, 1998), possession of high affinity  
36  
37 696 transporters likely provides a competitive advantage. The rhizosphere-associated  
38  
39 697 Rhodospirillaceae, Rhizobiaceae and Burkholderiaceae tended to have the highest  
40  
41 698 trait copies of ABC transporters (100 – 350 copies per genome, Supplementary Figure  
42  
43 699 4). The diverse, non-rhizospheric Deltaproteobacteria, Actinobacteria, Firmicutes,  
44  
45 700 Cyanobacteria and Verrucomicrobia all had greater than 50 copies per genome,  
46  
47 701 highlighting the importance of these traits in soil. The particularly high gene copy  
48  
49 702 number in rhizosphere-associated taxa from presumably nutrient-rich environments  
50  
51 703 contrasts the assumption that ABC transporters are considered to play a greater role  
52  
53 704 in nutrient-poor environments (Lauro *et al.*, 2009). Comparative genomics analyses of  
54  
55 705 soil bacteria have also found putatively copiotrophic Proteobacteria and Firmicutes to  
56  
57 706 be particularly enriched in PTS and ABC transporters (Trivedi *et al.*, 2013). In this  
58  
59 707 study, tThe transporters enriched in ~~these~~ rhizosphere-associated taxa were primarily  
60  
708 aimed at scavenging maltose, phosphate, amino acids, oligopeptides and export of  
709 LPS, and these results suggest that these traits are not only for survival in nutrient-  
710 poor environments but also likely confer a competitive advantage in the rhizosphere.  
711 As prokaryotes compete simultaneously with other prokaryotes and plants for nitrogen

1  
2  
3 712 and phosphorus in the rhizosphere, the high affinity acquisition of such nutrients is  
4  
5 713 likely critical for growth.

6 714

7  
8 715

9  
10 716 *4.4 Nitrogen and methane metabolism*

11 717 The BRITE category 'Nitrogen metabolism' encompasses nitrogen fixation,  
12  
13 718 denitrification, ammonia oxidation and synthesis of glutamate/glutamine which are,  
14  
15 719 critical amino acids for peptide synthesis. Since nitrogen limitation acts as an important  
16  
17 720 control on soil microbial activity, these traits are also of interest to soil microbiologists.  
18  
19 721 Three Proteobacterial Families, Rhodospirillaceae, Bradyrhizobiaceae and  
20  
21 722 Burkholderiaceae, were particularly enriched in these traits. Genomic and culture-  
22  
23 723 dependent analyses show these Families to be free-living or symbiotic diazotrophs in  
24  
25 724 soil and freshwater environments (Madigan *et al.*, 1984, Itakura *et al.*, 2009, de los  
26  
27 725 Santos *et al.*, 2018). Given their significantly greater copies of nitrogen-fixing genes,  
28  
29 726 these Families may be a particularly important source of organic nitrogen for soil  
30  
31 727 communities. Saprotrophic Mycobacteriaceae genomes, also identified as nitrogen  
32  
33 728 cyclers, tend to have many copies of genes involved in ammonia uptake and glutamate  
34  
35 729 synthesis (Amon *et al.*, 2010). This taxon may play an alternative role in converting  
36  
37 730 mineral nitrogen to biomass where organic nitrogen as protein in excreted products or  
38  
39 731 necromass can undergo proteolysis and uptake between other community members.  
40  
41 732 The identification of Sporomusaceae as enriched in 'Nitrogen metabolism' traits is  
42  
43 733 unusual as these obligate anaerobic fermenters cannot use nitrate as an electron  
44  
45 734 acceptor (Möller *et al.*, 1984). Nor were the Sporomusaceae enriched in ammonia  
46  
47 735 uptake or glutamate synthesis genes (data not shown), and so it is uncertain what role  
48  
49 736 this Family plays in nitrogen cycling. Thaumarchaeota and Nitrosomonadaceae, known  
50  
51 737 ammonia oxidisers, were not enriched in 'Nitrogen metabolism' traits relative to other  
52  
53 738 Families (Supplementary Figure 5) despite Nitrosomonadaceae possessing multiple  
54  
55 739 copies of the operon responsible for ammonia oxidation (Klotz & Norton, 1998).  
56  
57 740 Specific traits may be overlooked here if the BRITE category includes many diverse  
58  
59 741 KOs (e.g. ammonia oxidation, nitrogen fixation, glutamate synthesis etc).

60 742 Another specialised metabolic pathway of interest involves 'Methane metabolism'  
61  
62 743 that includes production and oxidation of a potent greenhouse gas. Unsurprisingly, the  
63  
64 744 Methanogens and methanotrophic Methylococcaceae, Beijerinckiaceae,  
65  
66 745 Methylocystaceae were all enriched in traits involved in methane metabolism. While



1  
2  
3 746 methane oxidation can be present in some taxa from the Verrucomicrobia (Op den  
4 747 Camp *et al.*, 2009) the above proteobacterial representatives act as the primary  
5 748 terrestrial methane sink (Dunfield, 2007, Conrad, 2009).  
6  
7  
8  
9

749

750

#### 751 *4.5 Sensing, responding and tolerating the environment*

752 A particularly interesting divergence of traits were involved in how taxa detect and  
753 respond to environmental stimuli. Gram-negative Pseudomonadaceae,  
754 Rhodospirillaceae, Bradyrhizobiaceae and Myxococcaceae were enriched in two-  
755 component systems. These membrane-bound histidine kinases respond rapidly to  
756 extracellular stimuli (Galperin, 2005) and these traits were primarily involved in  
757 nitrogen, potassium, initiating chemotaxis and C<sub>4</sub>-dicarboxylate responses. Families  
758 enriched in transcription factors were the Myxococcaceae, Polyangiaceae,  
759 Streptomycetaceae, Mycobacteriaceae and Frankiaceae. These factors regulate  
760 transcription in response to intracellular cues and here these factors were primarily  
761 *rpoD* (housekeeping), *rpoH* (heat-shock/protein damage), *rpoE* (extra-cellular  
762 cytoplasmic stress) and *rpoS* (starvation) responses (Shimada *et al.*, 2017). The  
763 genomes of these taxa are also heavily enriched in regulatory genes for complex  
764 developmental stages, fruiting bodies and/or filamentous branching growth in soils  
765 (Bentley *et al.*, 2002, Gao *et al.*, 2006, Schneiker *et al.*, 2007). Thus, certain taxa may  
766 respond primarily to extracellular cues while others strictly monitor and respond to  
767 changes in cell homeostasis. This trend has been noted previously – in 167 genomes  
768 across various Bacteria and Archaea, Proteobacteria had a higher ratio of sensors for  
769 external *versus* internal stimuli and were considered ‘extroverts’, while Cyanobacteria  
770 were considered strong ‘introverts’ focussed on responding to internal stimuli  
771 (Galperin, 2005).

772 As mentioned above, Archaea exhibited unique traits in basal transcription and  
773 protein regulation via proteasome. These transcription factors were primarily involved  
774 in identifying DNA damage and excision repair: TFII-B, TFII-D, ERCC-2 and ERCC-3.  
775 DNA repair differs markedly between Bacteria and Archaea/eukaryotes. Specifically,  
776 Bacteria excise 12 nucleotides around a damaged site with a 3 polypeptide system  
777 whereas Archaea excise 24 – 32 nucleotides with a 13 – 16 polypeptide system  
778 (Sancar, 1996). The use of ubiquitin-labelling and proteasome degradation of  
779 misfolded or ‘old’ proteins is arguably a more efficient system for recycling amino acids

1  
2  
3 780 and regulating the 'lifespan' of a protein in eukaryotes, however, Bacteria are still fully  
4  
5 781 capable of regulating protein misfolding or proteolysis with RpoH (and others) induced  
6  
7 782 upon environmental stress (Goldberg, 2003). From an ecological perspective, it is  
8  
9 783 difficult to discern if these eukaryote-like traits confer any sort of competitive advantage  
10  
11 784 to Archaea. They may simply be examples of convergent evolution for dealing with  
12  
13 785 environmental stress.

14 786 Finally, most microbial cells likely exist within complex biofilms and/or assemblages  
15  
16 787 adhered to surfaces with excreted exopolysaccharides, DNA and protein that serve to  
17  
18 788 protect from adverse environmental factors (Flemming & Wingender, 2010). Families  
19  
20 789 with high copy numbers of exopolysaccharide biosynthesis and secretion systems may  
21  
22 790 act as integral members of soil communities by predominantly contributing to  
23  
24 791 biofilm/aggregate formation. The 'LPS biosynthesis' and 'Starch and sucrose  
25  
26 792 metabolism' BRITE categories can synthesise N-acetyl glucosamine-based and  
27  
28 793 cellulose-based exopolysaccharides, respectively. Taxa enriched in both these  
29  
30 794 categories and secretion systems were the Polyangiaceae and Burkholderiaceae  
31  
32 795 (Table 1), and in the refined 'Under-represented' model, Acidobacteria and  
33  
34 796 Verrucomicrobia (Supplementary Table 11).

35 797

36 798

#### 37 799 *4.6 Direct cell-cell interactions*

38 800 Type IV secretion systems were another important trait identified in the Random Forest  
39  
40 801 models. These were enriched in Pseudomonadaceae and Myxococcaceae (Table 1)  
41  
42 802 and Acidobacteria, Planctomycetes and Verrucomicrobia (Supplementary Table 11).  
43  
44 803 These are highly specialised exporters that deliver DNA and/or toxins directly to other  
45  
46 804 bacterial or plant cells, however, their role in ecology is poorly understood beyond root  
47  
48 805 galls induced by *Agrobacterium tumefaciens* (Christie & Vogel, 2000). These taxa  
49  
50 806 should be explored for whether they utilise these traits for horizontal gene transfer or  
51  
52 807 to inject toxins directly into other prokaryotes, and thus potentially provide a selective  
53  
54 808 advantage for colonisation and competition.

55 809 Another archetypal trait for interactions between community members are  
56  
57 810 production of antimicrobials and antimicrobial resistance genes. Penicillin and  
58  
59 811 cephalosporin synthesis were enriched in the Sporomusaceae and Bacillaceae relative  
60  
812 to other Firmicutes. Polyketide and Type II polyketide syntheses were important for  
813 separating Frankiaceae and Streptomycetaceae from other Actinobacteria

1  
2  
3 814 (Supplementary Table 10). The Streptomycetaceae have a long history of use in  
4  
5 815 biotechnology as prolific antimicrobial producers (Bentley *et al.*, 2002). Bacillaceae (in  
6  
7 816 particular *Bacillus subtilis* species) are also well known producers of a wide variety of  
8  
9 817 antimicrobials (Caulier *et al.*, 2019), but we noted that Sporomusaceae have an even  
10  
11 818 greater number of these traits (Supplementary Figure 9). To the authors' knowledge,  
12  
13 819 antibiotic production in Sporomusaceae has not been investigated thoroughly and this  
14  
15 820 may be a consequence of its obligate anaerobic nature and difficulties in culturing. In  
16  
17 821 addition to prolific Type II polyketide producers, Streptomycetaceae were also enriched  
18  
19 822 in antimicrobial resistance genes, while Bacteroidetes, Planctomycetes and  
20  
21 823 Verrucomicrobia were specifically enriched in beta-Lactam resistance (Supplementary  
22  
23 824 Tables 9 and 10).

825

826

#### 827 4.7 Generalisable life strategies emerge from differentially enriched traits

828 We hypothesised that taxa would emerge as being inherently copiotrophic or  
829 oligotrophic based on trends in their enriched traits. Traits were chosen based on  
830 identification via Random Forest and identification as associated with copiotroph-  
831 oligotroph species or in mixed communities as described previously (Lauro *et al.*, 2009,  
832 Vieira-Silva & Rocha, 2010, Roller & Schmidt, 2015, Pascual-Garcia & Bell, 2020).  
833 Rhizosphere-associated Gamma-, Alpha- and Betaproteobacteria in Clade I fit the  
834 assumptions of a copiotrophic niche that invests in high metabolic rate – these taxa  
835 were uniquely enriched in competing for nutrient uptake via high-affinity ABC  
836 transporters, and energy generation from pyruvate, fatty acids, benzoate and  
837 glyoxylate carbon sources. Clade I was also enriched in glutathione metabolism, which  
838 acts as the major antioxidant for reducing intracellular free radicals produced during  
839 central carbon metabolism (Smirnova & Oktyabrsky, 2005). Antioxidants have been  
840 hypothesised as an essential function for copiotrophs to survive their high metabolic  
841 rates (Koch, 2001). All five Clade I Families were enriched in oxidative phosphorylation.  
842 The oxidative phosphorylation traits encompass a wide variety of electron transport  
843 chain proteins (oxidoreductases, dehydrogenases, cytochromes and ATPases) and  
844 are crucial for efficient energy production (Brochier-Armanet *et al.*, 2009). All five  
845 Families were also enriched in nitrogen metabolism, which included both nitrogen  
846 fixation and glutamate (*i.e.* protein) synthesis. Nitrogen fixation is an energy intensive  
847 process requiring 20 – 30 ATP per reduced N<sub>2</sub> (Burriss & Roberts, 1993) and may be

1  
2  
3 848 intrinsically linked to taxa with high oxidative phosphorylation. Finally, Clade I also  
4  
5 849 shared motility and chemotaxis, which are also energy intensive traits. Clade II  
6  
7 850 consisted of the remaining Gamma-, Alpha- and Betaproteobacteria, yet these were  
8  
9 851 relatively less enriched in Clade I 'copiotroph' traits. These particular taxa may be  
10  
11 852 responsible for the lack of a consistent copiotrophic response upon nutrient addition in  
12  
13 853 Proteobacteria (Ho *et al.*, 2017).

14 854 Clade III was comprised of taxa generally considered as oligotrophs (Ho *et al.*, 2017)  
15  
16 855 with the exception of Bacteroidetes (Fierer *et al.*, 2007). These taxa possessed high  
17  
18 856 LPS and sphingolipid synthesis that can defend against desiccation and antimicrobials  
19  
20 857 through biofilm and capsule/slime production (Flemming & Wingender, 2010), beta-  
21  
22 858 Lactam resistance, penicillin biosynthesis and several members had high pentose  
23  
24 859 phosphate pathway for efficient carbon metabolism under starvation (Hodgson, 2000).  
25  
26 860 Clade III also possessed high CAZy traits, which Clade I largely lacked, and is  
27  
28 861 consistent with observations of oligotrophs being primarily responsible for catabolising  
29  
30 862 relatively recalcitrant plant material (Goldfarb *et al.*, 2011). While Clade III were equally  
31  
32 863 enriched in oxidative phosphorylation as Clade I, with the exception of the  
33  
34 864 Deltaproteobacteria, these taxa were depleted in glutathione metabolism. The low  
35  
36 865 copies per genome of this trait would explain why the abundance of oligotrophs drop  
37  
38 866 rapidly in nutrient addition studies as they would be either out-competed by  
39  
40 867 glutathione-rich taxa capable of exploiting plentiful nutrients or will lyse if their  
41  
42 868 metabolic rate exceeds capacity to reduce free radicals (Koch, 2001). Taken together,  
43  
44 869 all of these traits indicate Clade III lead oligotrophic lifestyles whereby they are tolerant  
45  
46 870 to adverse environmental conditions, can acquire carbon from recalcitrant plant  
47  
48 871 material, and are incapable of rapid growth rates.

49  
50 872 These results support previous observations that Rhodospirillaceae,  
51  
52 873 Bradyrhizobiaceae, Burkholderiaceae, Pseudomonadaceae and Rhizobiaceae are  
53  
54 874 copiotrophic, while Planctomycetes, Verrucomicrobia, Myxococcaceae,  
55  
56 875 Polyangiaceae and Acidobacteria are oligotrophic (Ho *et al.*, 2017 and references  
57  
58 876 therein). As has been proposed previously, the dominance of these groups in certain  
59  
60 877 soils can provide inferences for ecosystem processes in that system, for example soils  
878 dominated by Verrucomicrobia, Planctomycetes and Acidobacteria will have greater  
879 capacity to degrade complex plant material while retaining most catabolised carbon in  
880 biomass (*i.e.* high growth or carbon use efficiency) or excreted byproducts that assist  
881 in soil aggregation (*e.g.* high LPS production) (Trivedi *et al.*, 2013). Conversely, soils

1  
2  
3 882 dominated by copiotrophic Proteobacteria Families will be systems primarily  
4 dependent on labile di- and monosaccharides that demonstrate low carbon use  
5 883 efficiency.  
6  
7 884

8 885 Streptomycetaceae, Mycobacteriaceae and Frankiaceae in Clade IV shared  
9 enrichment of several Clade I copiotroph traits. As mentioned above in Section 4.5,  
10 886 these Actinobacteria invest carbon and energy into complex filamentous growth and  
11 887 these Actinobacteria invest carbon and energy into complex filamentous growth and  
12 888 developmental cycles. They demonstrate classic copiotrophic responses to nutrient  
13 889 addition (Goldfarb *et al.*, 2011, Leff *et al.*, 2015) and their enriched ABC transporters,  
14 890 pyruvate, glyoxylate, benzoate and fatty acid metabolism all likely contribute to  
15 891 generating energy for complex lifecycles. Simultaneously, their enrichment of Clade III  
16 892 oligotroph traits for CAZy metabolism in addition to many traits for producing and  
17 893 resisting antimicrobials indicate a unique niche for these Actinobacteria that does not  
18 894 necessarily fall within the classical copiotroph-oligotroph framework.

19 895 Clade V differed markedly from all other clades and mostly consisted of 'specialist'  
20 896 metabolic functional groups involved in photosynthesis, ammonia oxidation,  
21 897 methanogenesis, lactic acid production and other fermentation. Similar to Clade III,  
22 898 these taxa would also be expected to have relatively low metabolic rates due to  
23 899 depletion of copiotroph traits associated with rapid metabolism and energy generation.  
24 900 Unlike Clade III, these taxa seemed to lack consistent mechanisms for stress  
25 901 tolerance. Thus, while certain taxa did invest in traits for rapid metabolic rate (Clades I  
26 902 and IV) and others primarily in stress tolerance (Clades III and IV), some taxa lacked  
27 903 these approaches altogether and pursued entirely distinct niches (Clade V). As an  
28 904 average of 47% of traits within each genome were Uncharacterised, Clade V is an  
29 905 over-simplification and that if novel, currently uncharacterised proteins and the traits  
30 906 they fulfil were incorporated into hierarchical clustering, this clade would separate more  
31 907 meaningfully.

32 908 If one were to consider taxa within the one-dimensional copiotroph-oligotroph  
33 909 spectrum, Clade I would represent one extreme, Clades III and V another, with Clades  
34 910 II and IV falling in between. Figure 5a is a conceptual diagram where these Clades  
35 911 have been placed on a singular axis of 'resource investment', with the niche space of  
36 912 Clades enriched in traits associated with rapid growth (e.g. glutathione) toward the  
37 913 'copiotrophy' pole while Clades enriched in stress tolerance traits (e.g. LPS production)  
38 914 are placed toward the 'oligotrophy' pole. However, this approach overlooks the diverse  
39 915 functional potentials (and distinct niches) for carbon and energy metabolism in Clade



1  
2  
3 916 ~~V~~associated with the various Clades. Furthermore, the large overlaps in niche space  
4  
5 917 between Clades would suggest taxa from each group could not co-exist if 'resource  
6  
7 918 investment' was the only important consideration (Gause, 1932, Hutchinson, 1957,  
8  
9 919 Leibold, 1995). A more meaningful perspective would be to consider the additional role  
10 920 of 'resource acquisition' that incorporates multiple axes for the life strategies identified  
11 921 via hierarchical clustering of traits. Figure 5b is a conceptual diagram of niche space  
12 922 where clades have been further separated along additional dimensions based on their  
13 923 enrichment of traits involved in competition, degradation or specialised metabolic  
14 924 pathways. The BRITE categories listed on the various axes are chosen to be useful  
15 925 markers in predicting the niche of a taxon. -This expanded niche space would suggest  
16 926 hierarchical clustering in Figure 4 with each clade as its own 'life strategy'. The strategy  
17 927 of taxa in Clade I follow that of classical 'copiotrophs' and are well equipped for nutrient  
18 928 acquisition (primarily, but not limited to, di- and monosaccharides), rapid growth and  
19 929 oxidative stress regulation as 'copiotrophic competitors'. Clade II, which may not be  
20 930 capable of competing directly with Clade I for carbon and energy, ~~and perhaps the~~  
21 931 ~~evolution of~~may thus occupy the niche space of specialist Proteobacterial  
22 932 methylotrophs, methane and ammonia oxidisers ~~may be a responseso as~~ to ~~avoid~~  
23 933 ~~competition over saccharides~~be 'copiotrophic metabolic specialists'. The strategy of  
24 934 Clade III ~~is~~would be to ~~specialise in decomposing~~decompose plant material via diverse  
25 935 CAZy and possess a variety of environmental stress tolerance traits as 'oligotrophic  
26 936 degraders'. Clade IV, which include for example ~~are unique~~ Actinobacteria that share  
27 937 traits for competition, degradation and oligotrophy, could thus occupy niche space  
28 938 between Clades I and III ~~traits, yet likely fill a distinct niche due to complex~~  
29 939 ~~developmental cycles and filamentous growth~~. Finally, the strategy of Clade V ~~is~~would  
30 940 be to fill highly specialised, unrelated metabolic niches reliant on completely distinct  
31 941 carbon sources to other taxa. The large space conceptualised for Clade V in Figure  
32 942 5b, which does not imply either copiotrophic or oligotrophic resource investment, is an  
33 943 oversimplification and with improved understanding of traits in these taxa it may be  
34 944 possible to fracture and further separate them into more detailed groups. This would  
35 945 prove particularly beneficial for the poorly characterised Archaea. -and are unified by  
36 946 an absence of classical copiotroph and oligotroph traits.

37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57 947 \_\_ Moving beyond a one-dimensional *r-K* spectrum to accommodate additional trait-  
58 948 driven life strategies has been proposed in plant ecology (Grime, 1977). Specifically,  
59 949 Grime argued that plant taxa fall within a multi-dimensional space defined by extremes

1  
2  
3 950 on three axes: 'competitors' that acquire nutrients, light, water etc. more effectively  
4 951 than neighbouring taxa in the same environment, 'stress tolerators' that are long-lived,  
5 952 slow growing taxa that resist desiccation, alkaline soils etc., and 'ruderals' that have  
6 953 very brief lifecycles between periods of disturbance and invest in environmentally  
7 954 hardy seeds. Despite these varied strategies for resource investment, plants are  
8 955 unified in that photosynthesis is their primary form of acquiring carbon and energy. The  
9 956 diversity of microbial strategies for acquiring carbon and energy enables them to  
10 957 explore a greater range of potential niche space, and in addition to growth traits that  
11 958 allow for a relatively more copiotrophic or oligotrophic investment of those resources,  
12 959 likely contributes to the high diversity of co-existing taxa observed in soil microbial  
13 960 communities. ~~complexity of co-occurring traits across microbial Families indicate more~~  
14 961 ~~generalisable life strategies than the copiotroph-oligotroph spectrum are needed.~~  
15 962 However, to truly unravel differentiated niches— and general microbial life  
16 963 strategies within these generalisable life strategies, two limitations must be overcome.  
17 964 Firstly, will require a better understanding of the many 'Uncharacterised' traits in  
18 965 environmental isolates, ~~in addition to robust trait-based analyses at the Species or~~  
19 966 ~~even Strain taxonomic levels~~ is required. For example, the recent large-scale MAG study  
20 967 by Nayfach *et al.*, (2021) identified 5.8 million protein clusters (traits), of which over  
21 968 75% could not be annotated meaningfully by current protein databases. Secondly,  
22 969 robust trait-based analyses down to the finer scale of distinct genomes will likely be  
23 970 necessary to consider how individual taxonomic members of a community have either  
24 971 differentiated in order to co-exist or are in the throes of competition that will ultimately  
25 972 exclude one of the competitors.

973

974

## 975 5. Conclusion

976 In a collection of 175 terrestrial prokaryotes that possess 220 664 traits shared  
977 between at least two taxa, concepts in niche differentiation were explored. Non-random  
978 trait distributions were shown as preferential clustering of related taxa within most  
979 Phyla with a general trend of highest similarity at the level of Family. This strongly  
980 supported ecological coherence of shared traits within close relatives. Random Forest  
981 models successfully identified BRIT 3 categories that best explained differing traits  
982 between taxonomic groups. These traits were involved in a wide range of biological  
983 functions, including core physiological traits used historically to categorise taxa. Many



1  
2  
3 984 traits were also involved in functions often associated with copiotrophs and oligotrophs,  
4  
5 985 namely metabolism, nutrient acquisition and environmental stress tolerance.  
6  
7 986 Hierarchical clustering of differential traits formed five distinct clusters, with Clade I  
8  
9 987 representing the classical copiotrophic niche, Clades III and V as oligotrophic, and  
10  
11 988 Clades II and IV in between. A more refined perspective would be to consider each  
12  
13 989 Clade as its own ~~generalisable~~ life strategy in a niche space that considers both  
14  
15 990 resource investment and acquisition simultaneously; for example, the strategy of Clade  
16  
17 991 I is to invest in competition and rapid growth, while Clade V pursue highly distinct,  
18  
19 992 specialised metabolic functions. The trait-based analyses here were effective in  
20  
21 993 identifying general trends in potential function of terrestrial microbial taxa at the Phylum  
22  
23 994 and Family level. Further investigation will be necessary to identify traits that give rise  
24  
25 995 to niche differentiation at lower taxonomic ranks and, ultimately, the importance of this  
26  
27 996 for ecosystem processes of interest.

997

998

### 999 Acknowledgements

1000 The authors wish to thank the Australian Government Endeavour research program  
1001 ID: 6123\_2017 for funding D.R. Finn. Funding support for B. Bergk-Pinto came from  
1002 the European Union's Horizon 2020 research and innovation program under the Marie  
1003 Sklodowska-Curie ITN 675546-MicroArctic. The authors also wish to thank the  
1004 microbiologists involved in the isolation and genome sequencing of the taxa studied  
1005 here; it is hoped the majority of those involved are cited within the Supplementary  
1006 Material. Without this diligent culture-dependent work by the scientific community, the  
1007 theoretical ecology performed here would not be possible.

1008

### 1009 Conflict of Interest

1010 The authors declare no conflict of interest regarding this study or its outcomes.

1011

### 1012 References

- 1013 Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic Local Alignment Search  
1014 Tool. *Journal of Molecular Biology* **215**: 403-410.  
1015 Amon J, Titgemeyer F & Burkovski A (2010) Common patterns - unique features: nitrogen  
1016 metabolism and regulation in Gram-positive bacteria. *FEMS Microbiology Reviews* **34**: 588-  
1017 605.  
1018 Barbérán A, Bates ST, Casamayor EO & Fierer N (2012) Using network analysis to explore co-  
1019 occurrence patterns in soil microbial communities. *The ISME Journal* **6**: 343-351.

1

2

3 1020 Bastida F, Torres IF, Moreno JL, *et al.* (2016) The active microbial diversity drives ecosystem  
4 1021 multifunctionality and is physiologically related to carbon availability in Mediterranean semi-  
5 1022 arid soils. *Molecular Ecology* **25**: 4660-4673.

7 1023 Bentley SD, Chater KF, Cerdeno-Tarraga AM, *et al.* (2002) Complete genome sequence of the  
8 1024 model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**: 141-147.

9 1025 Blagodatsky SA & Richter O (1998) Microbial growth in soil and nitrogen turnover: A  
10 1026 theoretical model considering the activity state of microorganisms. *Soil Biology & Biochemistry*  
11 1027 **30**: 1743-1755.

13 1028 Bouskill NJ, Tang J, Riley WJ & Brodie EL (2012) Trait-based representation of biological  
14 1029 nitrification: model development, testing, and predicted community composition. *Frontiers in*  
15 1030 *Microbiology* **3**: 364.

17 1031 Brochier-Armanet C, Talla E & Gribaldo S (2009) The Multiple Evolutionary Histories of  
18 1032 Dioxygen Reductases: Implications for the Origin and Evolution of Aerobic Respiration.  
19 1033 *Molecular Biology and Evolution* **26**: 285-297.

20 1034 Brown JH, Gillooly JF, Allen AP, Savage VM & West GB (2004) Toward a metabolic theory of  
21 1035 ecology. *Ecology* **85**: 1771-1789.

23 1036 Burris RH & Roberts GP (1993) Biological nitrogen fixation. *Annual Reviews Nutrition* **13**: 317-  
24 1037 335.

25 1038 Cania B, Vestergaard G, Krauss M, Fliessbach A, Schloter M & Schulz S (2019) A long-term field  
26 1039 experiment demonstrates the influence of tillage on the bacterial potential to produce soil  
27 1040 structure-stabilizing agents such as exopolysaccharides and lipopolysaccharides.  
28 1041 *Environmental Microbiome* **14**: 1-14.

30 1042 Cary SC, McDonald IR, Barrett JE & Cowan DA (2010) On the rocks: the microbiology of  
31 1043 Antarctic Dry Valley soils. *Nature Reviews Microbiology* **8**: 129-138.

32 1044 Caulier S, Nannan C, Gillis A, Licciardi F, Bragard C & Mahillon J (2019) Overview of the  
33 1045 antimicrobial compounds produced by members of the *Bacillus subtilis* group. *Frontiers in*  
34 1046 *Microbiology* **10**: 302.

36 1047 Christie PJ & Vogel JP (2000) Bacterial type IV secretion: conjugation systems adapted to  
37 1048 deliver effector molecules to host cells. *Trends in Microbiology* **8**: 354-360.

38 1049 Conrad R (2009) The global methane cycle: recent advances in understanding the microbial  
39 1050 processes involved. *Environmental Microbiology Reports* **1**: 285-292.

41 1051 Crits-Christoph A, Robinson CK, Barnum T, Fricke WF, Davila AF, Jedynak B, McKay CP &  
42 1052 DiRuggiero J (2013) Colonization patterns of soil microbial communities in the Atacama  
43 1053 Desert. *Microbiome* **1**: 28.

45 1054 de los Santos PE, Palmer M, Chávez-Ramirez B, Beukes C, Steenkamp ET, Briscoe L, Khan N,  
46 1055 Maluk M *et al.*, (2018) Whole genome analyses suggests that *Burkholderia sensu lato* contains  
47 1056 two additional novel genera (*Mycetohabitans* gen. nov., and *Trinickia* gen. nov.): implications  
48 1057 for the evolution of diazotrophy and nodulation in the Burkholderiaceae. *Genes* **9**:  
49 1058 doi:10.3390/genes9080389.

51 1059 de Mendiburu F (2014) *Agricolae*: statistical procedures for agricultural research.  
52 1060 <http://cran.r-project.org/web/packages/agricolae/index.html>.

53 1061 Dunfield PF (2007) The Soil Methane Sink. *Greenhouse Gas Sinks* 152-170.

54 1062 Embley TM & Martin W (2006) Eukaryotic evolution, changes and challenges. *Nature Reviews*  
55 1063 **440**: 623-630.

57 1064 Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high  
58 1065 throughput. *Nucleic Acid Research* **32**: 1792-1797.

59 1066 Enright AJ, Van Dongen S & Ouzounis CA (2002) An efficient algorithm for large-scale detection  
60 1067 of protein families. *Nucleic Acids Research* **30**: 1575-1584.

- 1  
2  
3 1068 Fierer N, Bradford MA & Jackson RB (2007) Toward an ecological classification of soil bacteria.  
4 1069 *Ecology* **88**: 1354-1364.
- 5 1070 Fierer N, Lauber CL, Ramirez KS, Zaneveld J, Bradford MA & Knight R (2012) Comparative  
6 1071 metagenomic, phylogenetic and physiological analyses of soil microbial communities across  
7 1072 nitrogen gradients. *The ISME Journal* **6**: 1007-1017.
- 8 1073 ~~Finn DR, Yu J, Ilhan ZE, Fernandes VMC, Penton CR, Krajmalnik-Brown R, Garcia-Pichel F &~~  
9 1074 ~~Vogel TM (2020a) MicroNiche: an R package for assessing microbial niche breadth and overlap~~  
10 1075 ~~from amplicon sequencing data. *Fems Microbiology Ecology* **96**: fiae131.~~  
11 1076 ~~Flemming HC & Wingender J (2010) The biofilm matrix. *Nature Reviews Microbiology* **8**: 623-~~  
12 1077 ~~633.~~
- 13 1078
- 14 1079 Finn DR, Ziv-el M, van Haren J, Park JG, del Aguila-Pasquel J, Urquiza-Munoz JD & Cadillo-  
15 1080 Quiroz H (2020b) Methanogens and methanotrophs show nutrient-dependent community  
16 1081 assemblage patterns across tropical peatlands of the Pastaza-Maranon Basin, Peruvian  
17 1082 Amazonia. *Frontiers in Microbiology* **11**: 746.
- 18 1083 ~~Flemming HC & Wingender J (2010) The biofilm matrix. *Nature Reviews Microbiology* **8**: 623-~~  
19 1084 ~~633.~~
- 20 1085 ~~Finn DR, Yu J, Ilhan ZE, Fernandes VMC, Penton CR, Krajmalnik-Brown R, Garcia-Pichel F &~~  
21 1086 ~~Vogel TM (2020) MicroNiche: an R package for assessing microbial niche breadth and overlap~~  
22 1087 ~~from amplicon sequencing data. *Fems Microbiology Ecology* **96**: fiae131.~~  
23 1088 ~~Flemming HC & Wingender J (2010) The biofilm matrix. *Nature Reviews Microbiology* **8**: 623-~~  
24 1089 ~~633.~~
- 25 1090 Freilich S, Zarecki R, Eilam O, Segal ES, Henry CS, Kupiec M, Gophna U, Sharan R & Ruppin E  
26 1091 (2011) Competitive and cooperative metabolic interactions in bacterial communities. *Nature*  
27 1092 *Communications* **2**.
- 28 1093 Galperin MY (2005) A census of membrane-bound and intracellular signal transduction  
29 1094 proteins in bacteria: bacterial IQ, extroverts and introverts. *BMC Microbiology* **5**:  
30 1095 doi:10.1186/1471-2180-1185-1135.
- 31 1096 Gao B, Paramanathan R & Gupta RS (2006) Signature proteins that are distinctive  
32 1097 characteristics of Actinobacteria and their subgroups. *Antonie Van Leeuwenhoek International*  
33 1098 *Journal of General and Molecular Microbiology* **90**: 69-91.
- 34 1099 Garcia-Pichel F, Lopez-Cortes A & Nübel U (2001) Phylogenetic and morphological diversity of  
35 1100 Cyanobacteria in soil desert crusts from the Colorado Plateau. *Applied and Environmental*  
36 1101 *Microbiology* **67**: 1902-1910.
- 37 1102 Gause GF (1932) Experimental studies on the struggle for existence I Mixed population of two  
38 1103 species of yeast. *Journal of Experimental Biology* **9**: 389-402.
- 39 1104 Gleason HA (1926) The individualistic concept of the plant association. *Bull Torrey Botany Club*  
40 1105 **53**: 7-26.
- 41 1106 Goldberg AL (2003) Protein degradation and protection against misfolded or damaged  
42 1107 proteins. *Nature* **426**: 895-890.
- 43 1108 Goldfarb KC, Karaoz U, Hanson CA, Santee CA, Bradford MA, Treseder KK, Wallenstein MD &  
44 1109 Brodie EL (2011) Differential growth responses of soil bacterial taxa to carbon substrates of  
45 1110 varying chemical recalcitrance. *Frontiers in Microbiology* **2**.
- 46 1111 Grime JP (1977) Evidence for the existence of three primary strategies in plants and its  
47 1112 relevance to ecological and evolutionary theory. *The American Naturalist* **111**: 1169-1194.
- 48 1113 Grime JP (1979) Plant strategies and vegetation processes. *Wiley, Chichester, UK*.
- 49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1114 He M, Zhang J, Shen L, *et al.* (2019) High-throughput sequencing analysis of microbial  
1115 community diversity in response to indica and japonica bar-transgenic rice paddy soils. *PLOS*  
1116 *One* **14**: e0222191.
- 1117 Higgins CF (2001) ABC transporters: physiology, structure and mechanism - an overview.  
1118 *Research in Microbiology* **152**: 205-210.
- 1119 Ho A, Paolo Di Lonardo D & Bodelier PL (2017) Revisiting life strategy concepts in  
1120 environmental microbial ecology. *FEMS Microbiology Ecology* **93**: 1-14.
- 1121 Ho A, Kerckhof FM, Luke C, Reim A, Krause S, Boon N & Bodelier PL (2013) Conceptualizing  
1122 functional traits and ecological characteristics of methane-oxidizing bacteria as life strategies.  
1123 *Environmental Microbiology Reports*  
1124 **5**: 335-345.
- 1125 Hodgson DA (2000) Primary metabolism and its control in streptomycetes: A most unusual  
1126 group of bacteria. *Advances in Microbial Physiology, Vol 42*, Vol. 42 (Poole RK, ed.) p. 47-  
1127 238.
- 1128 Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR, Williams KH, Tringe SG  
1129 & Banfield JF (2013) Community genomic analyses constrain the distribution of metabolic  
1130 traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome*  
1131 **1**.
- 1132 Hutchinson GL (1957) Concluding remarks *Cold Spring Harbor Symposia on Quantitative*  
1133 *Biology* **22**: 415-427.
- 1134 Itakura M, Saeki K, Omori H, Yokoyama T, Kaneko T, Tabata S, Ohwada T, Tajima S & . ea (2009)  
1135 Genomic comparison of Bradyrhizobium japonicum strains with different symbiotic nitrogen-  
1136 fixing capabilities and other Bradyrhizobiaceae members. *The ISME journal* **3**: 326-339.
- 1137 Jangid K, Williams MA, Franzluebbers A, Schmidt TM, Coleman DC & Whitman WB (2011)  
1138 Land-use history has a stronger impact on soil microbial community composition than  
1139 aboveground vegetation and soil properties. *Soil Biol Biochem* **43**: 2184-2193.
- 1140 Jarrell KF, Albers SV (2012) The archaeellum: an old motility structure with a new name. *Trends*  
1141 *Microbiol* **20**: 307-312.
- 1142 Kaiser C, Franklin O, Richter A & Dieckmann U (2015) Social dynamics within decomposer  
1143 communities lead to nitrogen retention and organic matter build-up in soils. *Nature*  
1144 *Communications* **6**.
- 1145 Kanehisa M, Sato Y, Kawashima M, Furumichi M & Tanabe M (2016) KEGG as a reference  
1146 resource for gene and protein annotation. *Nucleic Acids Research* **44**: D457-D462.
- 1147 Kearney M, Simpson SJ, Raubenheimer D & Helmuth B (2010) Modelling the ecological niche  
1148 from functional traits. *Philosophical Transactions of the Royal Society B-Biological Sciences*  
1149 **365**: 3469-3483.
- 1150 Keddy PA (1992) Assembly and response rules - two goals for predictive community ecology.  
1151 *Journal of Vegetation Science* **3**: 157-164.
- 1152 Klotz MG & Norton JB (1998) Multiple copies of ammonia monooxygenase (amo) operons  
1153 have evolved under biased AT/GC mutational pressure in ammonia-oxidizing autotrophic  
1154 bacteria. *Fems Microbiology Letters* **168**: 303-311.
- 1155 Koch AL (2001) Oligotrophs versus copiotrophs. *BioEssays* **23**: 657-661.
- 1156 Kögel-Knabner I (2002) The macromolecular organic composition of plant and microbial  
1157 residues as inputs to soil organic matter. *Soil Biol Biochem* **34**: 139-162.
- 1158 Kuznetsov SI, Dubinina GA & Lapteva NA (1979) Biology of oligotrophic bacteria. *Annual*  
1159 *Review of Microbiology* **33**: 377-387.



- 1  
2  
3 1160 Lauro FM, McDougald D, Thomas T, *et al.* (2009) The genomic basis of trophic strategy in  
4 1161 marine bacteria. *Proceedings of the National Academy of Sciences of the United States of*  
5 1162 *America* **106**: 15527-15533.
- 7 1163 Leff JW, Jones SE, Prober SM, *et al.* (2015) Consistent responses of soil microbial communities  
8 1164 to elevated nutrient inputs in grasslands across the globe. *Proceedings of the National*  
9 1165 *Academy of Sciences of the United States of America* **112**: 10967-10972.
- 11 1166 Leibold MA (1995) The niche concept revisited: mechanistic models and community context.  
12 1167 *Ecology* **76**: 1371-1382.
- 13 1168 Liaw A & Weiner M (2002) Classification and regression by randomForest. *R News* **2**: 18-22.
- 14 1169 Madigan M, Cox SS & Stegeman RA (1984) Nitrogen fixation and nitrogenase activities in  
15 1170 members of the Family Rhodospirillaceae. *Journal of Bacteriology* **157**: 73-78.
- 17 1171 Madin JS, Nielsen DA, Brbic M, Corkrey R, Danko D, Edwards K & *et al.*, (2020) A synthesis of  
18 1172 bacterial and archaeal phenotypic trait data. *Nature Scientific Data* **7**: 170.
- 19 1173 McBride MJ (2001) Bacterial gliding motility: mechanisms for cell movement over surfaces.  
20 1174 *Annual Review of Microbiology* **55**: 49-75.
- 21 1175 McGill BJ, Enquist BJ, Weiher E & Westoby M (2006) Rebuilding community ecology from  
22 1176 functional traits. *Trends in Ecology & Evolution* **21**: 178-185.
- 24 1177 Möller B, Oßmer R, Howard BH, Gottschalk G & Hippe H (1984) Sporomusa, a new genus of  
25 1178 Gram-negative anaerobic bacteria including *Sporomusa sphaeroides* spec. nov. and  
26 1179 *Sporomusa ovata* spec. nov. *Archives of Microbiology* **139**: 388-396.
- 28 1180 [Nayfach S, Roux S, Seshadri R, Udworthy D, Varghese N, Schulz F \*et al.\*, \(2021\) A genomic catalog](#)  
29 1181 [of Earth's microbiomes. \*Nature Biotechnology\* \*\*39\*\*: 499-509.](#)
- 30 1182 Oksanen J, Guillaume Blanchet F, Kindt R, *et al.* (2013) Vegan: Community Ecology Package. R  
31 1183 package version 2.0-10. <http://CRAN.R-project.org/package=vegan>.
- 32 1184 Op den Camp HJM, Islam T, Stott MB, Harhangi HR, Hynes A, Schouten S, Jetten MSM,  
33 1185 Birkeland N-K, Pol A & Dunfield PF (2009) Environmental, genomic and taxonomic perspectives  
34 1186 on methanotrophic Verrucomicrobia. *Environmental Microbiology Reports* **1**: 293-306.
- 36 1187 Parks DH, Chuvochina M, Waite DW, Rinke C, Skarszewski A, Chaumeil PA & Hugenholtz P  
37 1188 (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises  
38 1189 the tree of life. *Nature Biotechnology* **36**: 996-1004.
- 40 1190 Pascual-Garcia A & Bell T (2020) Community-level signatures of ecological succession in  
41 1191 natural bacterial communities. *Nature Communications* **11**: 2386.
- 42 1192 Philippot L, Bru D, Saby NPA, Cuhel J, Arrouays D, Simek M & Hallin S (2009) Spatial patterns  
43 1193 of bacterial taxa in nature reflect ecological traits of deep branches of the 16S rRNA bacterial  
44 1194 tree. *Environmental Microbiology* **11**: 3096-3104.
- 46 1195 Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, Whitman WB & Hallin S (2010)  
47 1196 The ecological coherence of high bacterial taxonomic ranks. *Nature Reviews Microbiology* **8**:  
48 1197 523-529.
- 49 1198 Prosser JI (2012) Ecosystem processes and interactions in a morass of diversity. *Fems*  
50 1199 *Microbiology Ecology* **81**: 507-519.
- 52 1200 R Core Team (2013) R: A language and environment for statistical computing. *R Foundation*  
53 1201 *for statistical computing, Vienna, Austria*.
- 54 1202 Roller BRK & Schmidt TM (2015) The physiology and ecological implications of efficient  
55 1203 growth. *The ISME journal* **9**: 1481-1487.
- 57 1204 Sancar A (1996) DNA excision repair. *Annual Review of Biochemistry* **65**: 43-81.
- 58 1205 Sangwan P, Chen XL, Hugenholtz P & Janssen PH (2004) *Chthoniobacter flavus* gen. nov., sp  
59 1206 nov., the first pure-culture representative of subdivision two, Spartobacteria classis nov., of  
60 1207 the phylum Verrucomicrobia. *Applied and Environmental Microbiology* **70**: 5875-5881.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Schliep KP, Potts AJ, Morrison DA & Grimm WA (2017) Intertwining phylogenetic trees and networks. *Methods in Ecology and Evolution* **8**: 1212-1220.

Schneiker S, Perlova O, Kaiser O, Gerth K, Alici A, Altmeyer MO, Bartels D, Bekel T ~~et al.~~, (2007) Complete genome of the myxobacterium *Sorangium cellulosum*. *Nature Biotechnology* **25**: 1281-1290.

Semenov AM (1991) Physiological bases of oligotrophy of microorganisms and the concept of microbial community. *Microbial Ecology* **22**: 239-247.

Shimada T, Tanaka K & Ishihama A (2017) The whole set of the constitutive promoters recognized by four minor sigma subunits of *Escherichia coli* RNA polymerase. *Plos One* **12**: e0179181.

Smirnova GV & Oktyabrsky ON (2005) Glutathione in Bacteria. *Biochemistry* **70**: 1199-1211.

Spang A, Poehlein A, Offre P, Zumbärgel S, Haider S, Rychlik N, Nowka B *et al.*, (2012) The genome of the ammonia-oxidizing Candidatus *Nitrososphaera gargensis*: insights into metabolic versatility and environmental adaptations. *Environmental Microbiology* **14**(12): 3122-3145.

Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L & Ettema TJG (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**: 173-185.

Stackebrandt E, Pohla H, Kroppenstedt RM, Hippe H & Woese CR (1985) 16S rRNA analysis of *Sporomusa*, *Selenomonas* and *Megasphaera*: on the phylogenetic origin of Gram-positive Eubacteria. *Archives of Microbiology* **143**: 270-276.

Tamas I, Smirnova AV, He Z & Dunfield PF (2014) The (d)evolution of methanotrophy in the Beijerinckiaceae - a comparative genomics analysis. *The ISME journal* **8**: 369-382.

Terzenbach DP & Blaut M (1994) Transformation of tetrachloroethylene to trichloroethylene by homoacetogenic bacteria. *Fems Microbiology Letters* **123**: 213-218.

Trivedi P, Anderson IC & Singh BK (2013) Microbial modulators of soil carbon storage: integrating genomic and metabolic knowledge for global prediction. *Trends in Microbiology* **21**: 641-651.

van Rossum T, Ferretti P, Maistrenko OM & Bork P (2020) Diversity within species: interpreting strains in microbiomes. *Nature Reviews Microbiology* **18**: 491-506.

Vieira-Silva S & Rocha EPC (2010) The systemic imprint of growth and its uses in ecological (meta)genomics. *Plos Genetics* **6**: e1000808.

Warnes GR, Bolker B, Bonebakker L, *et al.* (2019) gplots: various R programming tools for plotting data. <https://cran.r-project.org/web/packages/gplots/index.html>.

Weins JJ (1998) Testing phylogenetic methods with tree congruence: phylogenetic analysis of polymorphic morphological characters in Phrynosomatid lizards. *Systematic Biology* **47**: 427-444.

Wessen E, Hallin S & Philippot L (2010) Differential responses of bacterial and archaeal groups at high taxonomical ranks to soil management. *Soil Biol Biochem* **42**: 1759-1765.

Wickham H (2007) Reshaping data with the reshape package. *Journal of Statistical Software* **21**: 1-20.

Young J & Holland IB (1999) ABC transporters: bacterial exporters-revisited five years on. *Biochimica Et Biophysica Acta-Biomembranes* **1461**: 177-200.

Zhu CS, Delmont TO, Vogel TM & Bromberg Y (2015) Functional Basis of Microorganism Classification. *Plos Computational Biology* **11**.

For Peer Review

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Peer Review