FEMS Microbiology Ecology, 97, 2021, fiab068

doi: 10.1093/femsec/fiab068
Advance Access Publication Date: 7 May 2021
Research Article

RESEARCH ARTICLE

Functional trait relationships demonstrate life strategies in terrestrial prokaryotes

Damien R. Finn[1,2,3,*,†], Benoît Bergk-Pinto[2], Christina Hazard[2,‡], Graeme W. Nicol[2,§], Christoph C. Tebbe[3] and Timothy M. Vogel[2,¶]

[1]School of Agriculture and Food Sciences, University of Queensland, St Lucia, Brisbane 4072, Australia, [2]Environmental Microbial Genomics, Laboratoire Ampère, École Centrale de Lyon, Université de Lyon, Avenue Guy de Collongue 36 Écully 69134, France and [3]Thünen Institut für Biodiversität, Johann Heinrich von Thünen Institut, Bundesallee 65 Braunschweig 38116, Germany

[*]**Corresponding author:** Thünen Institut für Biodiversität, Johann Heinrich von Thünen Institut, Bundesallee 65 Braunschweig 38116, Germany. Tel: +49 531 596 2527; E-Mail: damien.finn@thuenen.de

One sentence summary: Functional trait based modelling identified specific traits to explain certain Families that fill terrestrial copiotroph and oligotroph niches, but that overall microbial life strategies are more complex than this framework.

Editor: Ian Anderson
[†]Damien R. Finn, http://orcid.org/0000-0002-0366-4422
[‡]Christina Hazard, http://orcid.org/0000-0002-0325-5856
[§]Graeme W. Nicol, http://orcid.org/0000-0002-3876-022X
[¶]Timothy M. Vogel, http://orcid.org/0000-0002-9542-3246

ABSTRACT

Functional, physiological traits are the underlying drivers of niche differentiation. A common framework related to niches occupied by terrestrial prokaryotes is based on copiotrophy or oligotrophy, where resource investment is primarily in either rapid growth or stress tolerance, respectively. A quantitative trait-based approach sought relationships between taxa, traits and niche in terrestrial prokaryotes. With 175 taxa from 11 Phyla and 35 Families ($n = 5$ per Family), traits were considered as discrete counts of shared genome-encoded proteins. Trait composition strongly supported non-random functional distributions as preferential clustering of related taxa via unweighted pair-group method with arithmetic mean. Trait similarity between taxa increased as taxonomic rank decreased. A suite of Random Forest models identified traits significantly enriched or depleted in taxonomic groups. These traits conveyed functions related to rapid growth, nutrient acquisition and stress tolerance consistent with their presence in copiotroph-oligotroph niches. Hierarchical clustering of traits identified a clade of competitive, copiotrophic Families resilient to oxidative stress versus glycosyltransferase-enriched oligotrophic Families resistant to antimicrobials and environmental stress. However, the formation of five clades suggested a more nuanced view to describe niche differentiation in terrestrial systems is necessary. We suggest considering traits involved in both resource investment and acquisition when predicting niche.

Keywords: theoretical ecology; niche differentiation; copiotroph-oligotroph; Random Forest modelling

INTRODUCTION

Niche differentiation, the process of physiologically distinct organisms adapting better to certain conditions, is a contributing factor to the high biodiversity inherent in microbial communities (Prosser 2012). Such differentiation is likely an inevitable consequence of the principles of competitive exclusion and natural selection working in tandem–no two

Received: 20 January 2021; **Accepted:** 5 May 2021

organisms can theoretically occupy the same niche, as the poorer competitor must either adapt to a unique niche or be driven to extinction in that system (Gause 1932; Hutchinson 1957; Leibold 1995). The physiological traits driving niche differentiation must have the capacity to convey an advantage to the organism's ability to survive and reproduce (i.e. fitness) and be inherited by successive generations (McGill *et al.* 2006). Importantly, this implies that microbial communities are not only diverse in terms of individual 16S rRNA gene sequences, commonly used to assess community diversity, but also diverse in regard to their physiological traits.

Explaining niche differentiation through the functional, physiological traits present in ecological community members has a long history in macroecology. For example, differences between beak size and shape in Galápagos finches was instrumental in Darwin's hypothesis that a common ancestor had differentiated into multiple, island-specific species. Within the past century, trait-based analyses have been particularly predominant in plant ecology, with seed germination in submerged soil, salt tolerance, carbon to nitrogen biomass stoichiometry, and leaf mass per unit area acting as examples of traits linked to niche differentiation (Gleason 1926; Grime 1979; Keddy 1992). In contrast, trait-based approaches to explain microbial ecology have only been performed in few instances, such as conceptualizing niches of methanotrophs based on abundance in high *versus* low methane environments or disturbed *versus* undisturbed soils (Ho *et al.* 2013), correlating increasing growth rate with increasing ribosomal gene and ribosome-associated gene copy number (Vieira-Silva and Rocha 2010), deterministic modelling of nitrification rate based on ammonia and oxygen uptake rate, temperature sensitivity and growth rate (Bouskill *et al.* 2012), defining distinct niches of 32 marine microorganisms based on clustering of genome-encoded functional proteins (Lauro *et al.* 2009), identifying habitat generalists and specialists based on taxon co-occurrence patterns (Barbéran *et al.* 2012) and recently comparisons of 23 'core' traits (*e.g.* motility, carbon metabolism, optimal pH for growth) across 15 000 diverse host-associated and environmental genomes (Madin *et al.* 2020).

A consistent trend noted in macroecology is that traits linked to how carbon and energy is processed and allocated to biomass can describe separate niches (Brown *et al.* 2004). The canonical example are *r* and *K* strategists, where carbon and energy are primarily invested in reproduction, or alternatively invested in tolerating biotic and/or abiotic stressors, respectively (Grime 1977). These dichotomous strategies have been observed in microbial ecology: copiotrophs are considered as microorganisms with relatively high growth rates that have relatively poor growth efficiency (as carbon incorporated to biomass per unit resource), relatively high cell maintenance energy costs, dependence on relatively high concentrations of organic carbon in their environment, demonstrate rapid population blooms upon the addition of organic matter and are not overly tolerant of abiotic stress (Semenov 1991; Koch 2001; Roller and Schmidt 2015; Ho, Paolo Di Lonardo and Bodelier 2017). Oligotrophs are considered as the inverse–low growth rate, high growth efficiency, low cell maintenance energy requirements, high substrate uptake affinity, slow growth yet at a consistent rate and are resilient to abiotic stress. Although the niche concept in macroecology has a formalized definition founded on where a taxon can maintain a stable population within multi-dimensional environmental space (Leibold 1995), in this study, niche is used simply to distinguish between prokaryotes being relatively more copiotrophic *versus* oligotrophic.

These distinct niches became associated with specific terrestrial taxa at high taxonomic rank based on recent molecular analyses. In complex microbial communities, the relative abundance of Gammaproteobacteria, Bacteroidetes and Actinobacteria were correlated with rapid growth in response to the addition of labile organic matter or nitrogen (copiotrophs) (Fierer, Bradford and Jackson 2007; Goldfarb *et al.* 2011; Fierer *et al.* 2012; Leff *et al.* 2015). Conversely, the Deltaproteobacteria, Acidobacteria, Verrucomicrobia and Planctomycetes were negatively correlated with the addition of organic matter or nitrogen (oligotrophs) (Fierer, Bradford and Jackson 2007; Fierer *et al.* 2012; Leff *et al.* 2015; Bastida *et al.* 2016). Conflicting reports exist of Beta- and Alphaproteobacteria, with some studies describing them as copiotrophic and others as oligotrophic highlighting that a consistent niche may not necessarily exist across species within a large taxonomic group (Ho, Paolo Di Lonardo and Bodelier 2017). A genomic basis for traits associated with soils dominated by putative copiotrophs and oligotrophs has been expertly reviewed elsewhere, and interested readers are referred to Trivedi, Anderson and Singh (2013) and references therein. Importantly, these observations suggest that specific traits that allow terrestrial prokaryotes to occupy these two niches should (generally) be associated with taxonomy. This is an example of ecological coherence at high taxonomic rank, whereby members within a taxon tend to have similar life strategies, niches and possess common traits compared to members of other taxa (Philippot *et al.* 2010). While ecological coherence of taxa has been considered previously, the shared, specific traits that drive niche differentiation in terrestrial prokaryotes remains an open question.

To identify the traits that differ between terrestrial prokaryote taxonomic groups, and whether these traits could describe the niches they occupy, a functional trait-based approach was adopted here. We posited that a trait must: (i) be associated with a physiological process that conveys a fitness advantage under certain environmental conditions; (ii) be measurable in well-defined units and (iii) vary more between taxonomic groups than within a taxonomic group (McGill *et al.* 2006; Kearney *et al.* 2010). Traits were measured as discrete counts of chromosome-encoded proteins shared between at least two of 175 terrestrial prokaryotes. Markov Chain clustering (MCL) was used to group proteins as traits based on amino acid sequence similarity (%) akin to a previous approach that confirmed close taxonomic relatives tend to share functional traits in 1374 genomes (Zhu *et al.* 2015). This was necessary to compare highly similar (but non-identical) proteins from separate genomes that carry out the same biological function. To better identify important, distinguishing traits of terrestrial prokaryotes, this study differed from Zhu *et al.* by: (i) comparing 175 publicly available terrestrial prokaryote genomes from 35 Families ($n = 5$ each), from 11 Phyla and two Kingdoms; (ii) selecting prokaryotes involved in terrestrial ecosystem processes of interest, including organic matter decomposition, nitrogen fixation, nitrification, denitrification, methane oxidation, plant-growth promotion, bioremediation of pollutants, pathogenesis and methanogenesis; (iii) selecting taxa isolated from a wide range of terrestrial environments, such as nutrient rich decomposing plant material and rhizosphere, submerged wetland and rice paddy soils, polluted soils and nutrient poor hot and cold arid environments and (iv) avoiding the inclusion of multiple subspecies and/or strains of a single species to prevent biases in analyses where highly over-represented species are compared with species that have fewer cultured representatives. The taxonomic system used here is from the NCBI, which is built upon a historical array of culture-dependent, physiological observations and genetic similarity to cultured isolates as average nucleotide identity, DNA-DNA hybridisation or 16S rRNA gene homology (ncbi.nlm.nih.gov/Taxonomy/Browser). The use of this classification system and comparison to others is discussed further below.

We hypothesised that: (i) traits are non-randomly distributed, with relatively closely related taxa demonstrating greater similarity than unrelated taxa (ecological coherence); (ii) traits that are differentially enriched between taxonomic groups would primarily be involved in metabolism, nutrient acquisition and/or tolerating environmental stress and (iii) copiotrophic and oligotrophic taxonomic groups would emerge based on collective trait enrichment.

METHODOLOGY

Collection of terrestrial prokaryote genomes

A collection of 175 sequenced and annotated genomes was collated (Table S1, Supporting Information). Listed are the genome ID, phylogenetic lineage, role in an ecosystem process if known, and isolation or genome sequencing reference. These genomes were sourced from the National Centre for Biotechnology Information (NCBI) and Joint Genome Institute (JGI) databases. Genomes were chosen based on several criteria: (i) five isolates per Family were chosen to have an equal minimum sample size per group, with this sample size being constrained by sequenced genomes of under-represented groups in public databases; (ii) only a single subspecies/strain per species was included to avoid bias due to over-representation of some species in public databases; (iii) an emphasis was placed to include isolates from diverse taxonomic lineages involved in terrestrial ecosystem processes of interest, such as ammonia oxidation and methanogenesis and (iv) there was an emphasis to include taxonomic groups frequently stated to be either copiotrophic (e.g. Actinobacteria, Gammaproteobacteria) or oligotrophic (e.g. Acidobacteria, Planctomycetes) based on observations from soil nutrient addition studies (Ho, Paolo Di Lonardo and Bodelier 2017). Taxonomic annotations for Phyla, Class, Order etc. were based on NCBI taxonomy as most genomes were sourced there. The authors recognise that taxonomy is constantly shifting, particularly so with the recent development of the Genome Taxonomy Database (GTDB) (Parks *et al.* 2018). Of note is that the vast majority of taxa here have the same taxonomy in NCBI as in GTDB, with the exceptions that GTDB considers the Sporomusaceae as split into three separate Families, the Leuconostocaceae to be Lactobacillaceae, the Promicromonosporaceae to be Cellulomonadaceae, and the Bradyrhizobiaceae and Methylococcaceae have been renamed as Xanthobacteraceae and Methylomonadaceae, respectively. Taxon selection was constrained by availability of genomes for under-represented groups, such as the Chloroflexi, Verrucomicrobia, Planctomycetes, Thaumarchaeota and Euryarchaeota. To meet the $n = 5$ requirement for balanced statistical analyses, it was necessary to consider these under-represented groups as 'Families'. Furthermore, due to the great diversity inherent within Proteobacterial Classes, Gamma-, Alpha-, Beta- and Deltaproteobacteria were considered as independent 'Phyla' for statistical analyses here. Indeed, GTDB now defines Deltaproteobacteria as its own Phylum, while Betaproteobacteria are considered as the Burkholderiales Order within the Gammaproteobacteria. The total of 175 genomes analysed here falls within the upper range of previous hypothesis-driven trait-based studies which varies from 11 isolates (Bouskill *et al.* 2012) to 214 genomes (Vieira-Silva and Rocha 2010).

Functional trait clustering

A step-by-step walkthrough of reproducible code to perform the following analyses on a subset of 12 genomes is available at: https://github.com/DamienFinn/Trait-based_analyses. First, a pairwise similarity comparison of all amino acid sequences (964 951 sequences) across the 175 genomes was performed with the all *versus* all basic local alignment tool function for proteins, BLAST-P (Altschul *et al.* 1990). Amino acid sequences were subsequently clustered as traits via MCL weighted by pairwise amino acid similarity (Enright, Van Dongen and Ouzounis 2002). Functional traits were grouped at a cluster value of 90.2, whereby $>$ 65 is considered 'fair' and confidence in accurately separating clusters cannot be higher than 100. The value of 90.2 is not chosen by the user but rather is a reflection of the quality of clustering in a given dataset. The MCL identified a total of 220664 traits shared between at least two genomes. A random subset of 1700 amino acid sequences were selected and the similarity of each sequence within its trait group (as determined by MCL) *versus* between other trait groups was visualised as a box and whisker plot (Fig. S1, Supporting Information) in R version 4.0.0 (R Core Team 2013). About 1700 sequences were chosen to maximise comparisons between trait groups under technical limitations, as increasing sequences led to exponential increases in trait combinations. A Student's T test was applied to determine whether sequences were more similar within their trait group relative to between trait groups in R. Finally, a matrix of genome *x* functional trait was generated in a two-step process by first associating genome IDs to the MCL output with a novel script 'MCLtoReshape2.py' (available at the above Github address) and secondly by casting the long data format to a wide data matrix with the 'reshape2' package in R (Wickham 2007). Box and whisker plots comparing counts of proteins per genome (input) and counts of functional traits shared by at least two genomes (output of computational workflow), for the 35 Families, is presented as Fig. 1.

UPGMA dendrogram of trait similarity between genomes

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

$$C = 1 - \left(\frac{\Sigma b_{ij}}{n^2}\right) \tag{1}$$

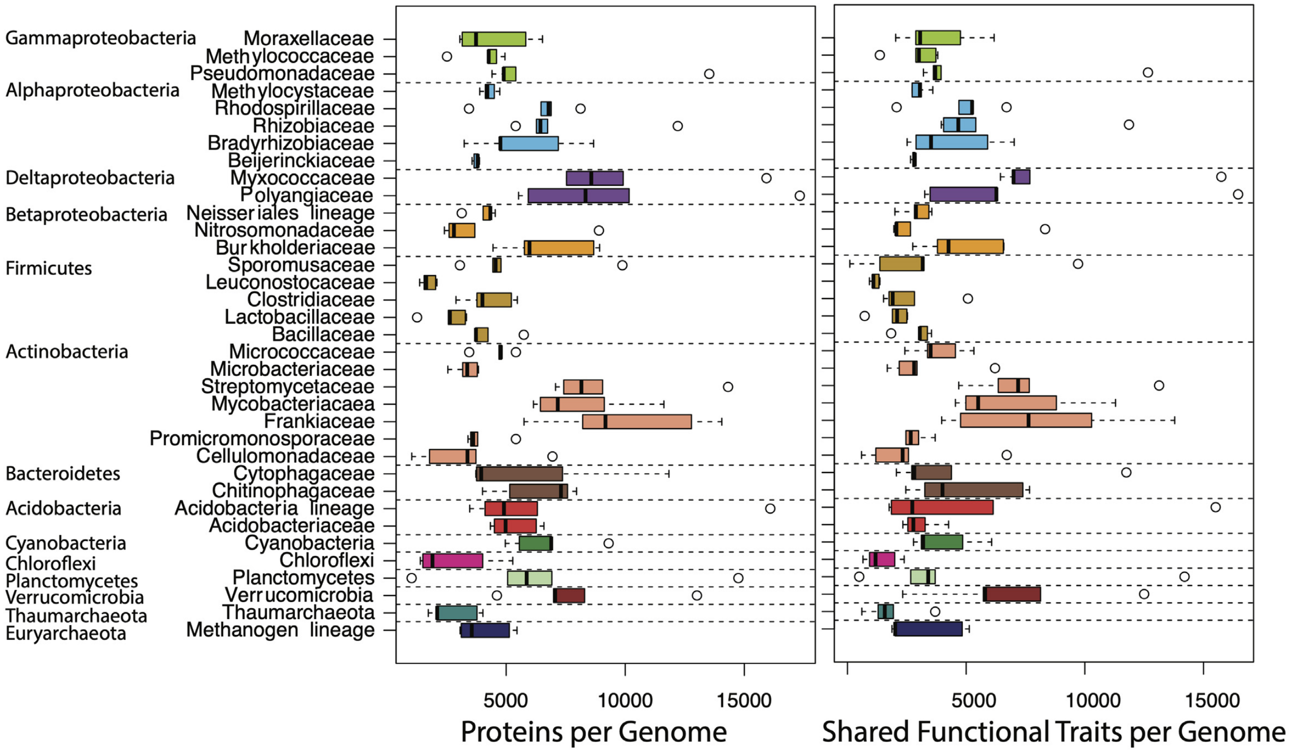

Figure 1. Box and whisker plots of the raw counts of proteins per genome beside the counts of shared functional traits per genome derived from MCL. Taxa $n = 5$ per Family.

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

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

Functional trait annotation

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

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 and 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: (i) Phylum, with Proteobacteria Classes separated due to their extensive diversity ($n = 14$); (ii) Family ($n = 35$); (iii) specifically for Families in the Proteobacteria ($n = 13$); (iv) Families in the Actinobacteria ($n = 7$); (v) Families in the Firmicutes ($n = 5$) and (iv) Families from 'Underrepresented' 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.

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 and Rocha 2010; Roller and Schmidt 2015; Pascual-Garcia and 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, while bacterial chemotaxis traits ranged from 0 to 15 copies) the trait copies were normalised for more appropriate comparisons. Normalised variance was calculated across the 35 Families for all traits with the 'decostand' function in the 'vegan' package (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).

RESULTS

Trait clustering and UPGMA

The 964951 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% to 100% for the 1st quartile, mean and 3rd quartile, respectively (Fig. S1, Supporting Information). Amino acid sequence similarity between traits ranged from 27.6%, 34% to 37.8% for the 1st quartile, mean and 3rd 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 220664 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 = 5533), found that there were fewer functional traits per genome (mean = 4240) (Fig. 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 (Fig. 1).

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

A simple index to measure trait similarity, as ecological coherence (C), within groups was devised (eq. 1). C increased as taxonomic rank decreased: Super Group < Phylum < Family (Fig. 2B). C was lowest for the larger, more diverse Proteobacteria and Terrabacteria (Super Group), and Firmicutes and Actinobacteria (Phylum). As the number of taxa being compared at the Super Group (*e.g.* Terrabacteria = 70 *versus* Acidobacteria = 10) and Phylum (*e.g.* Actinobacteria = 35 *versus* Thaumarchaeota = 5) were variable, the most meaningful comparisons between groups are at the Family level ($n = 5$ each). With the exception of the highly divergent 'Acidobacteria Lineage', all Families had a C greater than 0.3, with certain groups in the Alphaproteobacteria (Beijerinckiaceae), Firmicutes (Bacillaceae and Leuconostocaceae) and Actinobacteria (Promicromonosporaceae) being highly coherent ($C > 0.55$). Indeed, all individual Proteobacterial Families had $C > 0.4$, indicating that all five taxa within each of these Families had similar trait compositions. Despite not truly belonging to the same Family as per NCBI taxonomy, the $C > 0.33$ of the five Cyanobacteria, Thaumarchaeota and the Methanogen Lineage taxa was similar to other Families from the Bacteroidetes and Firmicutes. Thus, the UPGMA demonstrated that taxonomic relatives at the Phylum level tended to cluster with each other preferentially based on trait composition, and secondly that while similarity was broadly highest at low taxonomic rank, some Families were more coherent than others.

Phylogenetic distance (P) of each taxonomic group increased with decreasing taxonomic rank, and was highest in Proteobacteria, Actinobacteria and Firmicutes Families (Fig. S2B, Supporting Information). There was a strong positive linear relationship between P and C ($y = 0.86 \times - 0.25$, $R^2 = 0.39$, $P < 0.001$) supporting the result that taxonomic groups of closer related taxa tended to share more similar compositions of traits.

Random Forest trait identification

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

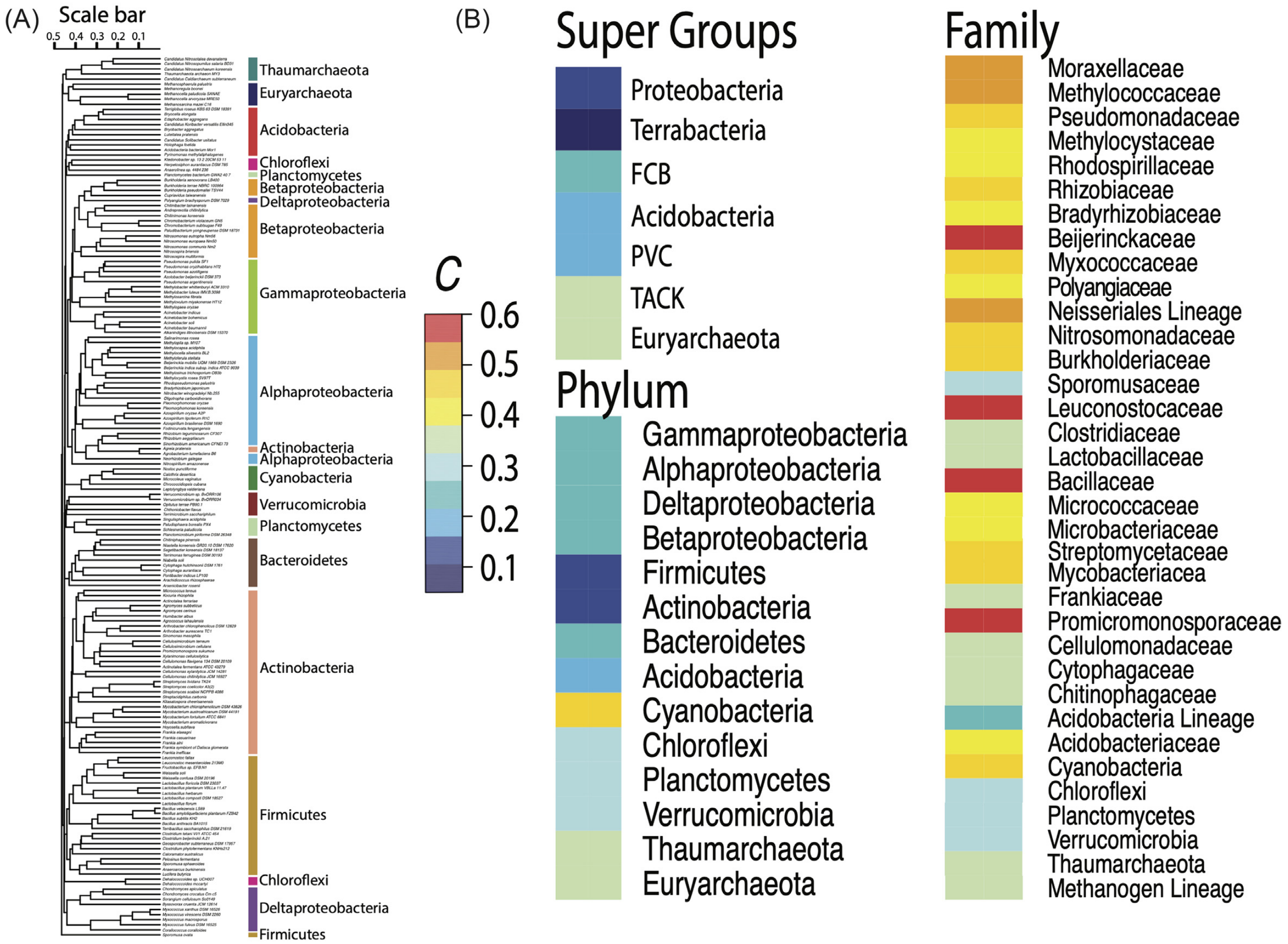

Figure 2. (A) Unweighted pair group method with arithmetic mean (UPGMA) dendrogram comparing similarity in composition of 220 664 traits across the 175 terrestrial microbial taxa. The Phylum of each taxon is highlighted. The scale bar units represent Bray–Curtis dissimilarity. **(B)** Comparisons of ecological coherence (C) between taxa belonging to the same Super Group, Phylum and Family. C, which varies between 0 and 1, was measured from branch lengths in a). Values of C approaching 1 indicate an ecologically coherent group with a similar composition of traits. Super groups were as follows: Proteobacteria, all Gamma-, Alpha-, Beta- and Deltaproteobacteria; Terrabacteria, all Actinobacteria, Firmicutes, Chloroflexi and Cyanobacteria; FCB were Bacteroidetes; Acidobacteria, all Acidobacteria; PVC, Verrucomicrobia and Planctomycetes; TACK were Thaumarchaeota; Euryarchaeota were Euryarchaeota.

Of the 260 BRITE 3 categories, the 60 most important traits in separating all Phyla and Families are ranked by importance measured as Mean Decrease in Accuracy (MDA) of the Random Forest models (Fig. 3). This is a measure of the average increase in classification error during permutation of trees ($n = 300$–400) when that particular trait is missing from the tree. For example, the accuracy of classifying Families was most improved by inclusion of the ABC transporters trait. Based on the identified traits, the Phylum model was capable of successfully classifying 81.14% of individual taxa. The Family model was capable of successfully classifying 71.43% of individual taxa. Confusion matrices for both models are presented as Tables S2–S4 (Supporting Information), and show that classification was particularly difficult for Chloroflexi and Planctomycetes (classification error >80%) in the Phylum model and for Cellulomonadaceae and the divergent Acidobacteria Lineage (classification error >80%) in the Family model. Random Forest models were robust against variation in P within Families, for example the nine families with all taxa perfectly classified ranged in P from the lowest (0.68) to highest (0.9).

The important traits in classifying the taxonomic groups were involved in: (i) metabolism and nutrient acquisition (oxidative phosphorylation, tricarboxylic acid (TCA) cycle, glyoxylate/decarboxylate, thermogenesis, propanoate, starch/sucrose, nitrogen, methane metabolism, synthesis of antioxidants such as glutathione, ATP-binding cassette (ABC) transporters, sugar uptake via phosphotransferase systems (PTS)); (ii) responding to environmental cues and stressors (protein kinases, two-component systems, transcription factors, proteasome, protein chaperones, RNA transport, chromosome repair via non-homologous DNA end joining); (iii) core cell physiology (flagella assembly, chemotaxis, sporulation, lipopolysaccharide (LPS), peptidoglycan, glycerolipid, sphingolipid and lipoarabinomannan (LAM) biosynthesis) and (iv) cell–cell interactions (beta-Lactam resistance, general secretion systems and Type IV secretion systems). Box and whisker plots of discrete counts of identified traits, and LSD results, are provided as Figs S4–S7 (Supporting Information). The Families significantly enriched and depleted in these traits are listed in Table 1. The Phyla significantly enriched and

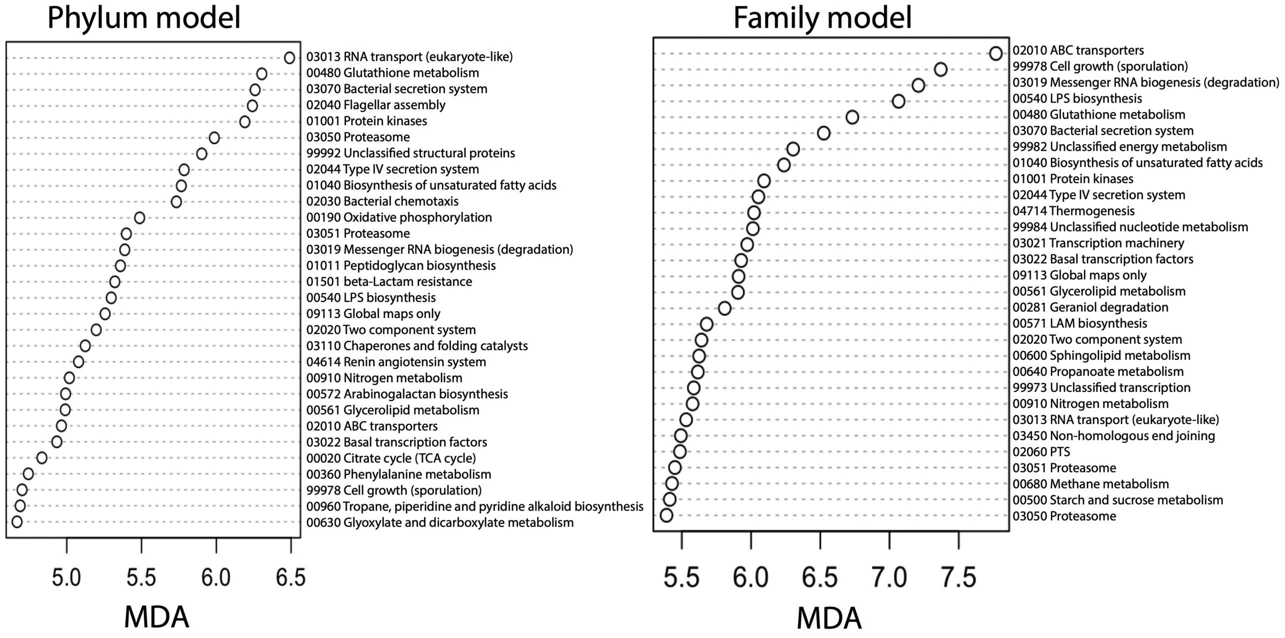

Figure 3. Random Forest identified traits that best explain separation of taxa groups at Phylum and Family taxonomic rank. Mean Decrease in Accuracy (MDA) is a measure of the average increase in classification error during permutation of trees if the trait is not included in the model. For example, the Family Random Forest was most accurate when ABC transporters were included in the model.

depleted in identified traits are listed in Table S9 (Supporting Information).

To better identify the more subtle differences between Families in the Proteobacteria, Actinobacteria, Firmicutes and the ‘Under-represented’ Phyla, individual Random Forest models were constructed for each of the four groups. The successful classification rates were 72.31%, 74.29%, 76% and 72%, respectively. Confusion matrices for each model are presented as Tables S5–S8 (Supporting Information). The models were unable to reliably classify Bradyrhizobiaceae and the divergent Acidobacteria Lineage (classification error >80%).

Figure S8 (Supporting Information) shows the most important traits in classifying the four groups. Unique traits not identified in the prior Phylum and Family models were: (i) for Proteobacteria, glycosyltransferases, butanoate metabolism, aminotransferases, ribosome biogenesis, mRNA biogenesis and degradation; (ii) for Actinobacteria, porphyrin and chlorophyll synthesis, pyruvate metabolism, aminotransferases, fatty acid and aliphatic hydrocarbon metabolism, polyketide and Type II polyketide biosynthesis, antimicrobial resistance genes; (iii) for Firmicutes, lysine, folate and varied amino acid synthesis, porphyrin and chlorophyll synthesis, DNA replication, bacterial toxins, penicillin and cephalosporin synthesis and (iv) for the ‘Under-represented’ taxa, glycosyltransferases, peptidases and inhibitors, photosynthesis and AMP-activated protein kinases. Box and whisker plots of discrete counts of identified traits, and LSD results, are provided as Figs S9–S12 (Supporting Information). Tables summarising Families enriched and depleted in these traits are included as Tables S10 and S11.

Hierarchical clustering of defining traits

Hierarchical clustering based on 60 traits, identified from Random Forest in this study and by previous copiotroph-oligotroph studies, indicated five general clades. The dendrogram on the y axis of Fig. 4 shows clustering of taxa as these five clades. The dendrogram on the *x* axis shows clustering of co-occurring traits. Clade I consisted of Proteobacteria, specifically the Pseudomonadaceae, Burkholderiaceae, Rhodospirillaceae, Bradyrhizobiaceae and Rhizobiaceae. These Families were uniquely enriched in flagellar assembly, chemotaxis, pyruvate metabolism, glutathione metabolism, ABC transporters, benzoate metabolism, transcription factors, glyoxylate and fatty acid metabolism. Clade II, also Proteobacteria, included Nitrosomonadaceae, Neisseriales Lineage, Methylocystaceae, Beijerinckiaceae, Methylococcaceae and Moraxellaceae. These Families clustered based on being enriched in Clade I traits, but to a lesser degree than the Pseudomonadaceae, Burkholderiaceae, Rhodospirillaceae, Bradyrhizobiaceae and Rhizobiaceae. Exceptions included the absence of benzoate metabolism and enrichment of methane metabolism in several Clade II Families.

Clade III, a diverse collection of Bacteroidetes (Chitinophagaceae, Cytophagaceae), Verrucomicrobia, Planctomycetes, divergent Acidobacteria Lineage and the Deltaproteobacteria (Polyangiaceae, Myxococcaceae), shared enrichment of sphingolipid metabolism, beta-Lactam resistance, penicillin and cephalosporin biosynthesis, LPS biosynthesis, glycosyltransferases and starch/sucrose metabolism. Many of these Families shared Clade I and II traits, including Type IV secretion system, oxidative phosphorylation, TCA cycle, PTS, nitrogen and glycerophospholipid metabolism. The absence of glutathione in non-Deltaproteobacterial Clade III Families was notable.

The three Actinobacteria Families in Clade IV, Mycobacteriaceae, Frankiaceae and Streptomycetaceae, were highly similar to each other because they were enriched in Type II polyketide biosynthesis. They also shared some Clade I traits (ABC transporters, transcription factors, pyruvate, benzoate and fatty acid metabolism) and Clade III traits (membrane trafficking, transcription machinery, polyketide biosynthesis and starch/sucrose metabolism). Similar to Clade III, these Actinobacteria were also depleted in glutathione traits.

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, Lactobacillacaceae
	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

Table 1. Continued

BRITE 1	BRITE 3	Significantly enriched	Significantly depleted
Environmental Information Processing	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, Streptomycetaceae, Mycobacteriaceae, Chloroflexi, Thaumarchaeota, Methanogen lineage
	02044 Type IV secretion system	Pseudomonadaceae, Myxococcaceae	Leuconostocaceae, all Bacteroidetes, Thaumarchaeota
Genetic Information Processing	03021 Transcription machinery	Myxococcaceae, Polyangiaceae, Streptomycetaceae, Mycobacteriaceae, Frankiaceae	Methylocystaceae, Rhizobiaceae, Bradyrhizobiaceae, Beijerinckiaceae, Leuconostocaceae, Lactobacillaceae, Thaumarchaeota, Methanogen lineage
	99973 Unclassified transcription factors	Pseudomonadaceae, Rhizobiaceae, Myxococcaceae, Streptomycetaceae	Chitinophagaceae, Cyanobacteria, Chloroflexi, Planctomycetaceae, 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, Streptomycetaceae, Verrucomicrobia	Moraxellaceae, Methylococcaceae, Neisseriales Lineage, Leuconostocaceae, Clostridaceae, Lactobacillaceae, Cyanobacteria
	03050 Proteasome	Micrococcaceae, Streptomycetaceae, Mycobacteriaceae, Frankiaceae, Promicromonosporaceae, Cellulomonadaceae, Thaumarchaeota, Methanogen Lineage	All other families
	03051 Prokaryote 20S Proteasome	Micrococcaceae, Streptomycetaceae, 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, Streptomycetaceae, Frankiaceae, all Bacteroidetes, all Acidobacteria, Chloroflexi, Planctomycetes, Thaumarchaeota, Methanogen Lineage

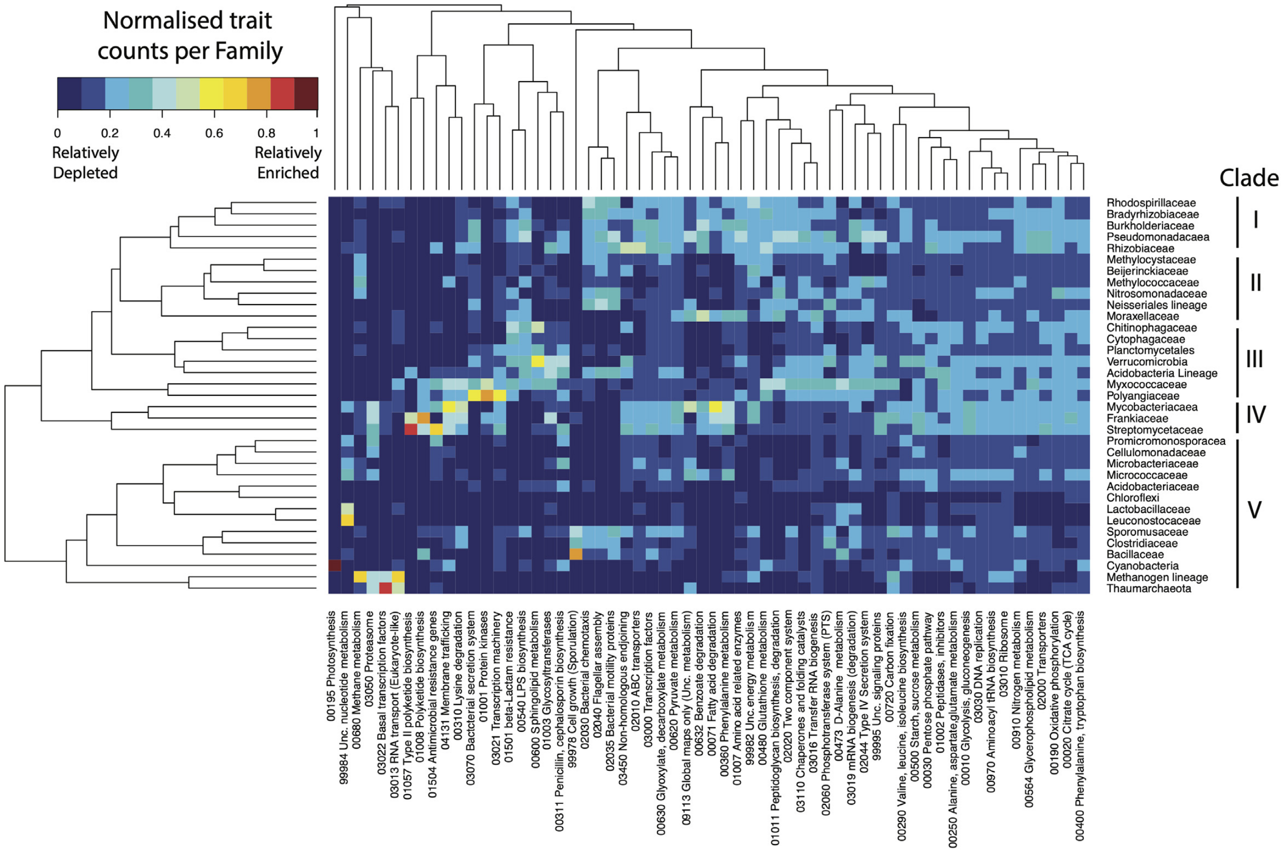

Figure 4. Hierarchical clustering of Families based on 60 traits selected from Random Forest and previously identified copiotroph-oligotroph traits. Families are clustered along the y axis, and specific co-occurring traits are clustered on the *x* axis. Five general Clades were identified based on Family clustering. Units are normalised variance in mean counts of traits per Family.

Finally, taxa within Clade V were similar to each other due to being depleted in traits shared among the other clades. Cyanobacteria were the only taxa that possessed photosynthesis traits. Lactic acid bacteria (Lactobacillaceae and Leuconostocaceae) were enriched in Unclassified nucleotide metabolism. The Archaea (Thaumarchaeota and Methanogen Lineage) shared eukaryote-like traits, proteasome, basal transcription factors and RNA transport. The Archaea were enriched in carbon fixation traits. Methanogens were also enriched in methane metabolism. Ammonia oxidising Thaumarchaeota were not enriched in nitrogen metabolism, however they were enriched in traits annotated by KEGG as Global maps only (unclassified metabolism). Further analysis found this to be the *fpr* gene, encoding a ferredoxin-flavodoxin NADP$^+$ reductase (K00528). Non-lactic acid bacteria of the Firmicutes (Bacillaceae, Sporomusaceae and Clostridiaceae) were enriched in sporulation and motility traits.

DISCUSSION

Non-random trait clustering demonstrates ecological coherence of taxa

A trait-based approach to investigate taxonomic relationships and potential biological function was carried out with a collection of 175 terrestrial prokaryotes. We hypothesised that traits would be non-randomly distributed amongst taxonomic groups, supported by previous observations that noted closely related taxa are isolated from similar habitats (Philippot *et al.* 2010). Similarity in the composition of 220664 traits, within 175 taxa, demonstrated strong agreement with established taxonomy at high (Phyla) and low (Family) rank (Fig. 2A). The exceptions to this ecological coherence at high rank were the division of the Chloroflexi and an individual from each of the Planctomycetes, Deltaproteobacteria, Actinobacteria and Firmicutes clustering with unrelated Phyla. These taxa were not mischaracterised, as based on phylogenetics of the full length 16S rRNA gene (Fig. 2A, Supporting Information). Trait similarity between related taxa, measured as C, tended to be highest at low rank (Fig. 2B). Uneven sample sizes between groups within Super Group and Phyla make comparisons at this level difficult–the inclusion of many diverse Firmicutes and Actinobacteria likely drove C to be lower here than in Thaumarchaeota and Euryarchaeota. However, equal comparisons at the Family level demonstrated interesting variability in coherence. All Proteobacterial Families had relatively high coherence (C = 0.4–0.6). The high C in Beijerinckiaceae is of particular interest as this group contained both specialist methanotroph (*Methylocapsa*, *Methylocella*), methylotroph (*Methyloferula*) and generalist heterotroph (*Beijerinckia* spp.) taxa. With such varied metabolic traits, one could reasonably expect C to be relatively low within this Family. The Beijerinckiaceae appear to have evolved from a common methylotrophic ancestor and still share traits for nitrogen fixation and tolerance for low pH soils (Tamas *et al.* 2014), and the high C measured here indicates that many additional shared traits remain. Both the relatively recent divergence of Families from a common ancestor and the higher number of shared traits are likely causes of the higher C observed at low taxonomic rank. The differing values of *C* for Methanogen and photosynthetic Cyanobacteria functional groups (0.33 and 0.44, respectively) is also worthy of note. Despite all five taxa in each group performing the same core role in a community, the individual isolates came from varying environments. The methanogens were isolated from a range of geographically separate wetlands, rice paddy soil and farm slurry and, while the Cyanobacteria were also isolated from geographically separate environments, they were all from sandy deserts or other nutrient poor, arid soils (Table S1, Supporting Information and references therein). Ultimately a taxon's trait composition will be affected by its functional role in a community, its evolutionary life-history (e.g. Beijerinckiaceae described above) and its local environment.

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

While the method of comparing taxa here differs from other studies, the results were not surprising as many 16S rRNA gene surveys of terrestrial systems consistently demonstrate ecological coherence at high rank. For example, independent studies of increasing agricultural intensity in soils show reductions in Actinobacteria abundance (Philippot *et al*. 2009; Jangid *et al*. 2011). Nitrogen addition to soils frequently enriches numerous taxa within the Actinobacteria and Proteobacterial Classes while negatively affecting taxa within the Verrucomicrobia and Planctomycetes (Wessen, Hallin and Philippot 2010; Fierer *et al*. 2012; Leff *et al*. 2015; Bastida *et al*. 2016). Arid, nutrient poor environments select for Actinobacteria-dominated communities (Cary *et al*. 2010; Crits-Christoph *et al*. 2013) and, in the absence of other primary producers, allow biological soil crust forming Cyanobacterial taxa to establish (Garcia-Pichel, Lopez-Cortes and Nübel 2001). Anoxic wetland and rice paddy environments support diverse communities of anaerobic Firmicutes, Chloroflexi and methanogenic Archaea (He *et al*. 2019; Finn *et al*. 2020b). These trends were noted prior to bioinformatic advances of metagenome assembled genomes (MAGs) that allow for the specific comparison of individual traits between uncultured environmental prokaryote genomes (Hug *et al*. 2013). 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 52515 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.

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

What can the Random Forests tell us?

The Random Forest works by identifying the traits that are most reliable in classifying individual Phyla and Families. It selects traits that tend to be: (i) of equal copy number per genome within a taxonomic group; and (ii) that differ markedly in copy number between taxonomic groups, since distinct separation of copies will maximise successful classification. A clear example of this is the consistent identification of eukaryote-like basal transcription factors, proteasome and RNA transport present in the Thaumarchaeota and Euryarchaeota, since they are absent from the majority of Bacterial Phyla. The binary nature of these traits (yes Archaea, no Bacteria) make them strong indicators. The presence of these and more eukaryote-like vesicular trafficking and actin traits have been noted in the Archaeal TACK super-phylum previously, and lend credence to the hypothesis that eukaryotes are descended from Archaea (Embley and Martin 2006; Spang *et al*. 2015). However, the Random Forest will not identify a trait unique to *Can*. Nitrosotalea devanterra that is absent from other Thaumarchaeota, as this single trait will not improve classification of the group as a whole. Consequently, the traits identified via Random Forest all tended to be core, fundamental traits shared by other members of a taxon's Phylum/Family.

Many of the best traits for distinguishing taxa have been historically used by microbiologists to do exactly that. These included fundamental cell physiology traits, such as oxidative phosphorylation, LPS biosynthesis, sporulation, flagellar assembly and chemotaxis. The Phylum model separated Betaproteobacteria, Actinobacteria and Bacteroidetes as taxa with the highest copies of oxidative phosphorylation traits. Firmicutes, Chloroflexi and Methanogens were identified as anaerobes depleted in oxidative phosphorylation, and all other taxa as sitting in between (Fig. S4 and Table S9, Supporting Information). Some Gram-negative Families were significantly enriched in LPS biosynthesis compared to others. These were the Pseudomonadaceae, Burkholderiaceae, Chitinophagaceae, divergent Acidobacteria and Verrucomicrobia (Table 1). This has been noted in soil communities previously. The extensive repertoire of LPS-associated genes in Chitinophagaceae, Acidobacteria and Verrucomicrobia likely play a critical role in enhancing soil aggregation (Cania *et al*. 2019) potentially through high LPS production and/or biofilm formation (discussed further below). In a demonstration of the robustness of the methods used here, the highly unusual Firmicute Sporomusaceae were shown to possess similar counts of LPS biosynthesis traits relative to most Gram-negative Families (Table S4, Supporting Information) in addition to sharing heat-resistant spore formation with Bacillaceae and Clostridiaceae (Table 1). The presence of both traits in a single Family have been remarked upon previously and used to conceptualise the evolution of Gram-negative *versus* Gram-positive lineages (Stackebrandt *et al*. 1985). The Sporomusaceae were also shown to have high Porphyrin and Chlorophyll metabolism traits in the Firmicutes model (Table S10, Supporting Information). The capacity to dechlorinate the soil pollutant perchloroethene to trichloroethylene via a porphyrin-based corrinoid is yet another interesting trait of this Family (Terzenbach and Blaut 1994).

Finally, flagella assembly and chemotaxis traits identified Alphaproteobacteria, Betaproteobacteria and Acidobacteria as Phyla that were particularly enriched with this mechanism of motility, while bacterial Actinobacteria, Bacteroidetes, Cyanobacteria, Chloroflexi and Verrucomicrobia were depleted (Table S9, Supporting Information). The Proteobacteria, Firmicutes and Under-represented models were better suited for identifying specific Families homogenously enriched or depleted in bacterial flagella and chemotaxis (Tables S10 and S11, Supporting Information). Enriched Families included the Rhodospirillaceae, Nitrosomonadaceae, Neisseriales lineage, divergent Acidobacteria lineage, Sporomusaceae, Bacillaceae, Clostridiaceae and Planctomycetes. Other forms of motility such as twitching and gliding have been noted in the Pseudomonadaceae, Myxococcaceae and Cyanobacteria (McBride 2001) but these traits were not identified by the Random Forest as being homogenously enriched in any Families. Furthermore, while Thaumarchaeota and Methanogens were both identified as being depleted in bacterial flagella assembly and chemotaxis traits (Tables S9 and S11, Supporting Information), Archaea possess a structurally distinct flagellum more similar to the Type IV bacterial pilus (Jarrell and Albers 2012). These taxa were not enriched with Type IV pilus, either, and it is possible that archaeal flagella may have failed proper characterisation by KEGG. Thus, while some Families were relatively enriched/depleted in bacterial flagella and chemotaxis traits, specific taxa depleted in these are not necessarily non-motile.

In summary, while the Random Forests may overlook certain traits in individual taxa, the models were highly robust in detecting conserved, shared traits within a Phylum/Family. Here the 'depth' of shared traits is limited by the number of taxa that could be considered as Phylum or Family. In future, if five (or more) taxa belonging to the same Genus or even Species could be compared, unique traits would be observed to explain how these subgroups have evolved from their respective Families to occupy distinct niches. Ideally the selection of individual taxa within groups for such future comparative analyses would also be standardised based on phylogenetic distance, either with *P* or a similar method, that would improve the robustness of trait-based comparisons at such a fine taxonomic level.

Plant-derived carbon metabolism and nutrient acquisition

Secondly, we hypothesised that the traits differentially enriched between taxonomic groups would largely reflect those associated with copiotrophs or oligotrophs, namely metabolism, nutrient acquisition and environmental stress response and tolerance. Of fundamental interest to soil microbiologists is the decomposition of plant biomass. This is the primary source of organic carbon to non-arid terrestrial systems (Kögel-Knabner 2002) and the transformation of plant material to substrates bioavailable for microorganisms is essential for community growth and activity. The traits involved in plant material catabolism belonged to the BRITE categories 'Starch and Sucrose Metabolism' (*e.g.* extracellular cellobiosidases, endoglucanases, glucosidases, trehalases, amylases) and 'Glycosyltransferases', all of which are carbohydrate activated enzymes (CAZy). The Families particularly enriched in these traits were the Polyangiaceae, Myxococcaceae, Rhizobiaceae, Streptomycetaceae, Mycobacteriaceae, Frankiaceae and Verrucomicrobia (Table 1; Table S11, Supporting Information). Genomic and culture-dependent analyses support *Sorangium cellulosum* (Polyangiaceae), *Streptomyces coelicolor* A3(2) (Streptomycetaceae) and *Chthoniobacter flavus* (Verrucomicrobia) as having particularly large genomes with extensive repertoires for cellulose, hemicellulose, pectin and lignin degradation (Bentley *et al.* 2002; Sangwan *et al.* 2004; Schneiker *et al.* 2007). Comparative genomics analyses have also identified Actinobacteria, Acidobacteria and Verrucomicrobia as being enriched in numerous enzymes for cellulose, hemicellulose and starch catabolism (Trivedi, Anderson and Singh 2013). *In situ* these Families likely play a critical role in making organic carbon bioavailable as di- and monosaccharides for the microbial community.

The Families equipped with many copies of high-affinity sugar uptake 'Phosphotransferase systems' (3–14 copies) did not necessarily correspond to those enriched with CAZy–only the Myxococcaceae and Verrucomicrobia were enriched in both. Pseudomonadaceae, Rhodospirillaceae, Neisseriales lineage and Clostridiaceae were only enriched in PTS. Despite being enriched in CAZy, the Frankiaceae were simultaneously depleted in PTS. The complex interplay between taxa capable of producing extracellular CAZy and competitors that rapidly scavenge available di- and monosaccharides has been well described by models that predict such competitive interactions exert important controls on the growth rate of the community as a whole (Freilich *et al.* 2011) and may even act to aid terrestrial carbon storage and limit carbon dioxide emissions from microbial respiration (Kaiser *et al.* 2015). Here, we identified the 'specialist' Families enriched in CAZy traits versus the 'opportunists' scavenging for sugars via PTS (Table 1; Table S11, Supporting Information).

ABC transporters facilitate the ATP-dependent uptake of soluble compounds across membranes or export waste metabolites, extracellular enzymes and toxins (Young and Holland 1999; Higgins 2001). This means of active transport allows microorganisms to acquire nutrients with high affinity at concentrations of 5–500 μg carbon L^{-1} *versus* the less efficient diffusion of nutrients across membranes, dependent on extracellular concentrations of 0.5–5 mg carbon L^{-1} (Kuznetsov, Dubinina and Lapteva 1979). In the spatially heterogenous soil environment where the concentration of bioavailable carbon substrate often limits growth (Blagodatsky and Richter 1998), possession of high affinity transporters likely provides a competitive advantage. The rhizosphere-associated Rhodospirillaceae, Rhizobiaceae and Burkholderiaceae tended to have the highest trait copies of ABC transporters (100–350 copies per genome, Fig. S4, Supporting Information). The diverse, non-rhizospheric Deltaproteobacteria, Actinobacteria, Firmicutes, Cyanobacteria and Verrucomicrobia all had greater than 50 copies per genome, highlighting the importance of these traits in soil. The particularly high gene copy number in rhizosphere-associated taxa from presumably nutrient-rich environments contrasts the assumption that ABC transporters are considered to play a greater role in nutrient-poor environments (Lauro *et al.* 2009). Comparative genomics analyses of soil bacteria have also found putatively copiotrophic Proteobacteria and Firmicutes to be particularly enriched in PTS and ABC transporters (Trivedi, Anderson and Singh 2013). In this study, the transporters enriched in rhizosphere-associated taxa were primarily aimed at scavenging maltose, phosphate, amino acids, oligopeptides and export of LPS, and these results suggest that these traits are not only for survival in nutrient-poor environments but also likely confer a competitive advantage in the rhizosphere. As prokaryotes compete simultaneously with other prokaryotes and plants for nitrogen and phosphorus in the rhizosphere, the high affinity acquisition of such nutrients is likely critical for growth.

Nitrogen and methane metabolism

The BRITE category 'Nitrogen metabolism' encompasses nitrogen fixation, denitrification, ammonia oxidation and synthesis of glutamate/glutamine which are critical amino acids for peptide synthesis. Since nitrogen limitation acts as an important control on soil microbial activity, these traits are also of interest to soil microbiologists. Three Proteobacterial Families, Rhodospirillaceae, Bradyrhizobiaceae and Burkholderiaceae, were particularly enriched in these traits. Genomic and culture-dependent analyses show these Families to be free-living or symbiotic diazotrophs in soil and freshwater environments (Madigan, Cox and Stegeman 1984; Itakura *et al.* 2009; de los Santos *et al.* 2018). Given their significantly greater copies of nitrogen-fixing genes, these Families may be a particularly important source of organic nitrogen for soil communities. Saprotrophic Mycobacteriaceae genomes, also identified as nitrogen cyclers, tend to have many copies of genes involved in ammonia uptake and glutamate synthesis (Amon, Titgemeyer and Burkovski 2010). This taxon may play an alternative role in converting mineral nitrogen to biomass where organic nitrogen as protein in excreted products or necromass can undergo proteolysis and uptake between other community members. The identification of Sporomusaceae as enriched in 'Nitrogen metabolism' traits is unusual as these obligate anaerobic fermenters cannot use nitrate as an electron acceptor (Möller *et al.* 1984). Nor were the Sporomusaceae enriched in ammonia uptake or glutamate synthesis genes (data not shown), and so it is uncertain what role this Family plays in nitrogen cycling. Thaumarchaeota and Nitrosomonadaceae, known ammonia oxidisers, were not enriched in 'Nitrogen metabolism' traits relative to other Families (Fig. S5, Supporting Information) despite Nitrosomonadaceae possessing multiple copies of the operon responsible for ammonia oxidation (Klotz and Norton 1998). Specific traits may be overlooked here if the BRITE category includes many diverse KOs (e.g. ammonia oxidation, nitrogen fixation, glutamate synthesis etc).

Another specialised metabolic pathway of interest involves 'Methane metabolism' that includes production and oxidation of a potent greenhouse gas. Unsurprisingly, the Methanogens and methanotrophic Methylococcaceae, Beijerinckiaceae, Methylocystaceae were all enriched in traits involved in methane metabolism. While methane oxidation can be present in some taxa from the Verrucomicrobia (Op den Camp *et al.* 2009) the above proteobacterial representatives act as the primary terrestrial methane sink (Dunfield 2007; Conrad 2009).

Sensing, responding and tolerating the environment

A particularly interesting divergence of traits were involved in how taxa detect and respond to environmental stimuli. Gram-negative Pseudomonadaceae, Rhodospirillaceae, Bradyrhizobiaceae and Myxococcaceae were enriched in two-component systems. These membrane-bound histidine kinases respond rapidly to extracellular stimuli (Galperin 2005) and these traits were primarily involved in nitrogen, potassium, initiating chemotaxis and C_4-dicarboxylate responses. Families enriched in transcription factors were the Myxococcaceae, Polyangiaceae, Streptomycetaceae, Mycobacteriaceae and Frankiaceae. These factors regulate transcription in response to intracellular cues and here these factors were primarily *rpoD* (housekeeping), *rpoH* (heat-shock/protein damage), *rpoE* (extra-cellular cytoplasmic stress) and *rpoS* (starvation) responses (Shimada, Tanaka and Ishihama 2017). The genomes of these taxa are also heavily enriched in regulatory genes for complex developmental stages, fruiting bodies and/or filamentous branching growth in soils (Bentley *et al.* 2002; Gao, Paramanathan and Gupta 2006; Schneiker *et al.* 2007). Thus, certain taxa may respond primarily to extracellular cues while others strictly monitor and respond to changes in cell homeostasis. This trend has been noted previously in 167 genomes across various Bacteria and Archaea, Proteobacteria had a higher ratio of sensors for external *versus* internal stimuli and were considered 'extroverts', while Cyanobacteria were considered strong 'introverts' focussed on responding to internal stimuli (Galperin 2005).

As mentioned above, Archaea exhibited unique traits in basal transcription and protein regulation via proteasome. These transcription factors were primarily involved in identifying DNA damage and excision repair: TFII-B, TFII-D, ERCC-2 and ERCC-3. DNA repair differs markedly between Bacteria and Archaea/eukaryotes. Specifically, Bacteria excise 12 nucleotides around a damaged site with a 3 polypeptide system whereas Archaea excise 24–32 nucleotides with a 13–16 polypeptide system (Sancar 1996). The use of ubiquitin-labelling and proteasome degradation of misfolded or 'old' proteins is arguably a more efficient system for recycling amino acids and regulating the 'lifespan' of a protein in eukaryotes, however, Bacteria are still fully capable of regulating protein misfolding or proteolysis with RpoH (and others) induced upon environmental stress (Goldberg 2003). From an ecological perspective, it is difficult to discern if these eukaryote-like traits confer any sort of competitive advantage to Archaea. They may simply be examples of convergent evolution for dealing with environmental stress.

Finally, most microbial cells likely exist within complex biofilms and/or assemblages adhered to surfaces with excreted exopolysaccharides, DNA and protein that serve to protect from adverse environmental factors (Flemming and Wingender 2010). Families with high copy numbers of exopolysaccharide biosynthesis and secretion systems may act as integral members of soil communities by predominantly contributing to biofilm/aggregate formation. The 'LPS biosynthesis' and 'Starch and sucrose metabolism' BRITE categories can synthesise N-acetyl glucosamine-based and cellulose-based exopolysaccharides, respectively. Taxa enriched in both these categories and secretion systems were the Polyangiaceae and Burkholderiaceae (Table 1), and in the refined 'Under-represented' model, Acidobacteria and Verrucomicrobia (Table S11, Supporting Information).

Direct cell–cell interactions

Type IV secretion systems were another important trait identified in the Random Forest models. These were enriched in Pseudomonadaceae and Myxococcaceae (Table 1) and Acidobacteria, Planctomycetes and Verrucomicrobia (Table S11, Supporting Information). These are highly specialised exporters that deliver DNA and/or toxins directly to other bacterial or plant cells, however, their role in ecology is poorly understood beyond root galls induced by *Agrobacterium tumefaciens* (Christie and Vogel 2000). These taxa should be explored for whether they utilise these traits for horizontal gene transfer or to inject toxins directly into other prokaryotes, and thus potentially provide a selective advantage for colonisation and competition.

Another archetypal trait for interactions between community members are production of antimicrobials and antimicrobial resistance genes. Penicillin and cephalosporin synthesis were enriched in the Sporomusaceae and Bacillaceae relative to other Firmicutes. Polyketide and Type II polyketide syntheses were important for separating Frankiaceae and Streptomycetaceae from other Actinobacteria (Table S10, Supporting Information). The Streptomycetaceae have a long history of use in biotechnology as prolific antimicrobial producers (Bentley *et al.* 2002). Bacillaceae (in particular *Bacillus subtilis* species) are also well known producers of a wide variety of antimicrobials (Caulier *et al.* 2019), but we noted that Sporomusaceae have an even greater number of these traits (Fig. S9, Supporting Information). To the authors' knowledge, antibiotic production in Sporomusaceae has not been investigated thoroughly and this may be a consequence of its obligate anaerobic nature and difficulties in culturing. In addition to prolific Type II polyketide producers, Streptomycetaceae were also enriched in antimicrobial resistance genes, while Bacteroidetes, Planctomycetes and Verrucomicrobia were specifically enriched in beta-Lactam resistance (Tables S9 and S10, Supporting Information).

Life strategies emerge from differentially enriched traits

We hypothesised that taxa would emerge as being inherently copiotrophic or oligotrophic based on trends in their enriched traits. Traits were chosen based on identification via Random Forest and identification as associated with copiotroph-oligotroph species or in mixed communities as described previously (Lauro *et al.* 2009; Vieira-Silva and Rocha 2010; Roller and Schmidt 2015; Pascual-Garcia and Bell 2020). Rhizosphere-associated Gamma-, Alpha- and Betaproteobacteria in Clade I fit the assumptions of a copiotrophic niche that invests in high metabolic rate–these taxa were uniquely enriched in competing for nutrient uptake via high-affinity ABC transporters, and energy generation from pyruvate, fatty acids, benzoate and glyoxylate carbon sources. Clade I was also enriched in glutathione metabolism, which acts as the major antioxidant for reducing intracellular free radicals produced during central carbon metabolism (Smirnova and Oktyabrsky 2005). Antioxidants have been hypothesised as an essential function for copiotrophs to survive their high metabolic rates (Koch 2001). All five Clade I Families were enriched in oxidative phosphorylation. The oxidative phosphorylation traits encompass a wide variety of electron transport chain proteins (oxidoreductases, dehydrogenases, cytochromes and ATPases) and are crucial for efficient energy production (Brochier-Armanet, Talla and Gribaldo 2009). All five Families were also enriched in nitrogen metabolism, which included both nitrogen fixation and glutamate (i.e. protein) synthesis. Nitrogen fixation is an energy intensive process requiring 20–30 ATP per reduced N_2 (Burris and Roberts 1993) and may be intrinsically linked to taxa with high oxidative phosphorylation. Finally, Clade I also shared motility and chemotaxis, which are also energy intensive traits. Clade II consisted of the remaining Gamma-, Alpha- and Betaproteobacteria, yet these were relatively less enriched in Clade I 'copiotroph' traits. These particular taxa may be responsible for the lack of a consistent copiotrophic response upon nutrient addition in Proteobacteria (Ho, Paolo Di Lonardo and Bodelier 2017).

Clade III was comprised of taxa generally considered as oligotrophs (Ho, Paolo Di Lonardo and Bodelier 2017) with the exception of Bacteroidetes (Fierer, Bradford and Jackson 2007). These taxa possessed high LPS and sphingolipid synthesis that can defend against desiccation and antimicrobials through biofilm and capsule/slime production (Flemming and Wingender 2010), beta-Lactam resistance, penicillin biosynthesis and several members had high pentose phosphate pathway for efficient carbon metabolism under starvation (Hodgson 2000). Clade III also possessed high CAZy traits, which Clade I largely lacked, and is consistent with observations of oligotrophs being primarily responsible for catabolising relatively recalcitrant plant material (Goldfarb *et al.* 2011). While Clade III were equally enriched in oxidative phosphorylation as Clade I, with the exception of the Deltaproteobacteria, these taxa were depleted in glutathione metabolism. The low copies per genome of this trait would explain why the abundance of oligotrophs drop rapidly in nutrient addition studies as they would be either out-competed by glutathione-rich taxa capable of exploiting plentiful nutrients or will lyse if their metabolic rate exceeds capacity to reduce free radicals (Koch 2001). Taken together, all of these traits indicate Clade III lead oligotrophic lifestyles whereby they are tolerant to adverse environmental conditions, can acquire carbon from recalcitrant plant material, and are incapable of rapid growth rates.

These results support previous observations that Rhodospirillaceae, Bradyrhizobiaceae, Burkholderiaceae, Pseudomonadaceae and Rhizobiaceae are copiotrophic, while Planctomycetes, Verrucomicrobia, Myxococcaceae, Polyangiaceae and Acidobacteria are oligotrophic (Ho, Paolo Di Lonardo and Bodelier 2017 and references therein). As has been proposed previously, the dominance of these groups in certain soils can provide inferences for ecosystem processes in that system, for example soils dominated by Verrucomicrobia, Planctomycetes and Acidobacteria will have greater capacity to degrade complex plant material while retaining most catabolised carbon in biomass (i.e. high growth or carbon use efficiency) or excreted byproducts that assist in soil aggregation (e.g. high LPS production) (Trivedi, Anderson and Singh 2013). Conversely, soils dominated by copiotrophic Proteobacteria Families will be systems primarily dependent on labile di- and monosaccharides that demonstrate low carbon use efficiency.

Streptomycetaceae, Mycobacteriaceae and Frankiaceae in Clade IV shared enrichment of several Clade I copiotroph traits. As mentioned above in Section 4.5, these Actinobacteria invest carbon and energy into complex filamentous growth and developmental cycles. They demonstrate classic copiotrophic responses to nutrient addition (Goldfarb *et al.* 2011; Leff *et al.* 2015) and their enriched ABC transporters, pyruvate, glyoxylate, benzoate and fatty acid metabolism all likely contribute to generating energy for complex lifecycles. Simultaneously, their enrichment of Clade III oligotroph traits for CAZy metabolism in addition to many traits for producing and resisting antimicrobials indicate a unique niche for these Actinobacteria that does not necessarily fall within the classical copiotroph-oligotroph framework.

Clade V differed markedly from all other clades and mostly consisted of 'specialist' metabolic functional groups involved in photosynthesis, ammonia oxidation, methanogenesis, lactic acid production and other fermentation. Similar to Clade III, these taxa would also be expected to have relatively low metabolic rates due to depletion of copiotroph traits associated with rapid metabolism and energy generation. Unlike Clade III, these taxa seemed to lack consistent mechanisms for stress tolerance. Thus, while certain taxa did invest in traits for rapid metabolic rate (Clades I and IV) and others primarily in stress tolerance (Clades III and IV), some taxa lacked these approaches altogether and pursued entirely distinct niches (Clade V). As an

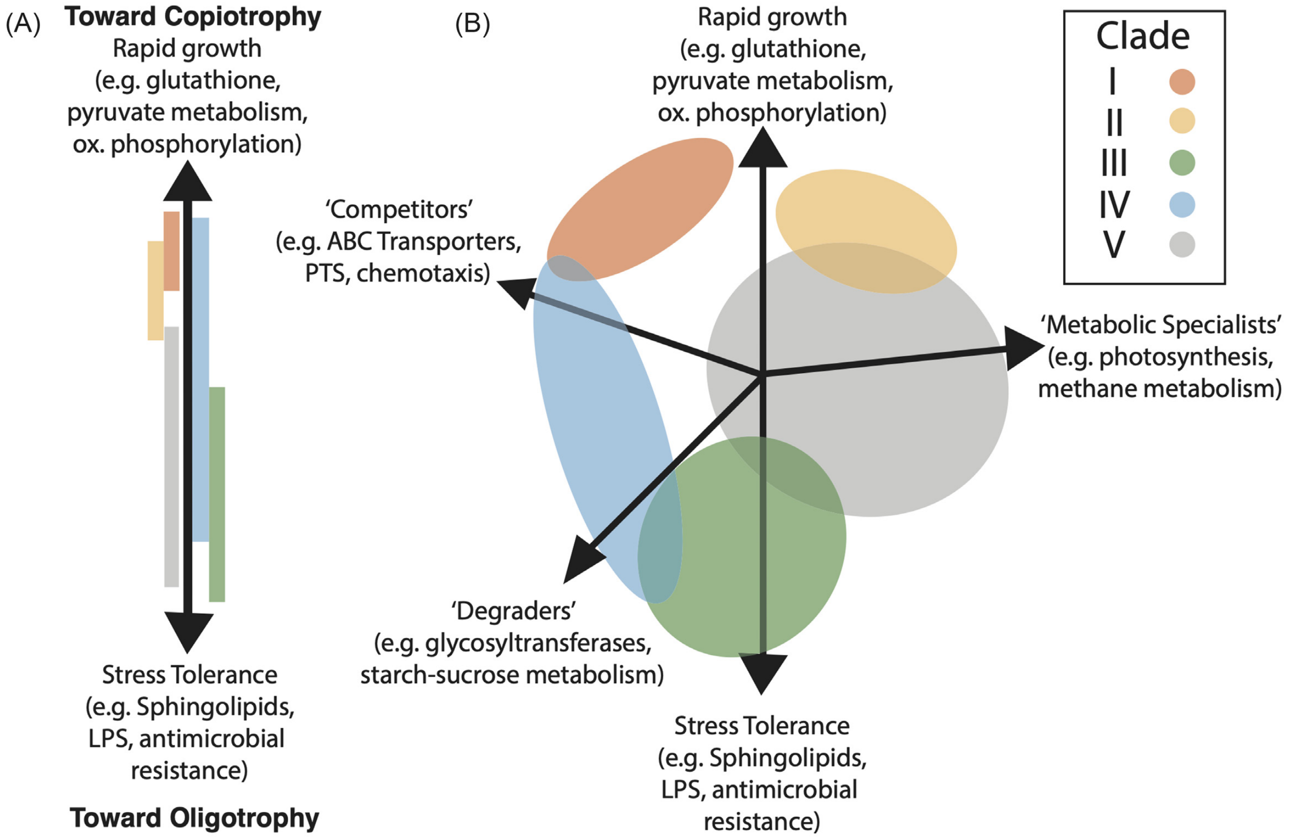

Figure 5. A conceptual diagram of overlapping life strategies between the five functional clades identified in this study. **(A)** The classical copiotroph-oligotroph dichotomy whereby the five clades are placed on a singular axis of 'resource investment', where growth strategies are either relatively more targeted toward rapid growth (copiotroph) or toward stress tolerance (oligotroph). BRITE categories are given as examples of functional traits contributing to either axis pole. The space occupied by clades along this axis is dependent on enrichment or depletion of these traits. All clades exhibit a great deal of overlap and certain clades, such as IV (including, for example, Actinobacteria), are hard to identify as either copiotroph or oligotroph. **(B)** A multidimensional concept where three axes for 'resource acquisition' are added, further separating taxa as either 'competitors', 'degraders' or 'metabolic specialists'. Again, BRITE categories are provided as examples of traits contributing to the additional axes and space occupied by clades is dependent on enrichment or depletion of these traits. These extra dimensions would suggest the niche space of clade IV exists between 'copiotrophic competitors' (I) and 'oligotrophic degraders' (III), potentially allowing IV to co-exist alongside both.

average of 47% of traits within each genome were Uncharacterised, Clade V is an over-simplification and that if novel, currently uncharacterised proteins and the traits they fulfil were incorporated into hierarchical clustering, this clade would separate more meaningfully.

If one were to consider taxa within the one-dimensional copiotroph-oligotroph spectrum, Clade I would represent one extreme, Clades III and V another, with Clades II and IV falling in between. Figure 5A is a conceptual diagram where these Clades have been placed on a singular axis of 'resource investment', with the niche space of Clades enriched in traits associated with rapid growth (e.g. glutathione) toward the 'copiotrophy' pole while Clades enriched in stress tolerance traits (e.g. LPS production) are placed toward the 'oligotrophy' pole. However, this approach overlooks the diverse functional potentials (and distinct niches) for carbon and energy metabolism associated with the various Clades. Furthermore, the large overlaps in niche space between Clades would suggest taxa from each group could not co-exist if 'resource investment' was the only important consideration (Gause 1932; Hutchinson 1957; Leibold 1995). A more meaningful perspective would be to consider the additional role of 'resource acquisition' that incorporates multiple axes for the life strategies identified via hierarchical clustering of traits. Figure 5B is a conceptual diagram of niche space where clades have been further separated along additional dimensions based on their enrichment of traits involved in competition, degradation or specialised metabolic pathways. The BRITE categories listed on the various axes are chosen to be useful markers in predicting the niche of a taxon. This expanded niche space would suggest taxa in Clade I are well equipped for nutrient acquisition (primarily, but not limited to, di- and monosaccharides), rapid growth and oxidative stress regulation as 'copiotrophic competitors'. Clade II, which may not be capable of competing directly with Clade I for carbon and energy, may thus occupy the niche space of specialist Proteobacterial methylotrophs, methane and ammonia oxidisers so as to be 'copiotrophic metabolic specialists'. The strategy of Clade III would be to decompose plant material via diverse CAZy and possess a variety of environmental stress tolerance traits as 'oligotrophic degraders'. Clade IV, which include for example Actinobacteria that share traits for competition, degradation and oligotrophy, could thus occupy niche space between Clades I and III. Finally, the strategy of Clade V would be to fill highly specialised, unrelated metabolic niches reliant on completely distinct carbon sources to other taxa. The large space conceptualised for Clade V in Fig. 5B, which does not imply either copiotrophic or oligotrophic resource investment, is an oversimplification and with improved understanding of traits in these taxa it may be possible to fracture and further separate them into more detailed groups. This would prove particularly beneficial for the poorly characterised Archaea.

Moving beyond a one-dimensional *r-K* spectrum to accommodate additional trait-driven life strategies has been proposed in plant ecology (Grime 1977). Specifically, Grime argued that plant taxa fall within a multi-dimensional space defined by extremes on three axes: 'competitors' that acquire nutrients, light, water etc. more effectively than neighbouring taxa in the same environment, 'stress tolerators' that are long-lived, slow growing taxa that resist desiccation, alkaline soils etc., and 'ruderals' that have very brief lifecycles between periods of disturbance and invest in environmentally hardy seeds. Despite these varied strategies for resource investment, plants are unified in that photosynthesis is their primary form of acquiring carbon and energy. The diversity of microbial strategies for acquiring carbon and energy enables them to explore a greater range of potential niche space, and in addition to growth traits that allow for a relatively more copiotrophic or oligotrophic investment of those resources, likely contributes to the high diversity of co-existing taxa observed in soil microbial communities.

However, to truly unravel differentiated niches and general microbial life strategies, two limitations must be overcome. First, 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. Second, robust trait-based analyses down to the finer scale of distinct genomes will likely be necessary to consider how individual taxonomic 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.

CONCLUSION

In a collection of 175 terrestrial prokaryotes that possess 220664 traits shared between at least two taxa, concepts in niche differentiation were explored. Non-random trait distributions were shown as preferential clustering of related taxa within most Phyla with a general trend of highest similarity at the level of Family. This strongly supported ecological coherence of shared traits within close relatives. Random Forest models successfully identified BRITE 3 categories that best explained differing traits between taxonomic groups. These traits were involved in a wide range of biological functions, including core physiological traits used historically to categorise taxa. Many traits were also involved in functions often associated with copiotrophs and oligotrophs, namely metabolism, nutrient acquisition and environmental stress tolerance. Hierarchical clustering of differential traits formed five distinct clusters, with Clade I representing the classical copiotrophic niche, Clades III and V as oligotrophic, and Clades II and IV in between. A more refined perspective would be to consider each Clade as its own life strategy in a niche space that considers both resource investment and acquisition simultaneously; for example, the strategy of Clade I is to invest in competition and rapid growth, while Clade V pursue highly distinct, specialised metabolic functions. The trait-based analyses here were effective in identifying general trends in potential function of terrestrial microbial taxa at the Phylum and Family level. Further investigation will be necessary to identify traits that give rise to niche differentiation at lower taxonomic ranks and, ultimately, the importance of this for ecosystem processes of interest.

SUPPLEMENTARY DATA

Supplementary data are available at FEMSEC online.

ACKNOWLEDGEMENTS

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

Conflicts of Interest. None declared.

REFERENCES

Altschul SF, Gish W, Miller W *et al.* Basic Local Alignment Search Tool. *J Mol Biol* 1990;**215**:403–10.

Amon J, Titgemeyer F, Burkovski A. Common patterns - unique features: nitrogen metabolism and regulation in Gram-positive bacteria. *FEMS Microbiol Rev* 2010;**34**:588–605.

Barbéran A, Bates ST, Casamayor EO *et al.* Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J* 2012;**6**:343–51.

Bastida F, Torres IF, Moreno JL *et al.* The active microbial diversity drives ecosystem multifunctionality and is physiologically related to carbon availability in Mediterranean semi-arid soils. *Mol Ecol* 2016;**25**:4660–73.

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

Blagodatsky SA, Richter O. Microbial growth in soil and nitrogen turnover: a theoretical model considering the activity state of microorganisms. *Soil Biol Biochem* 1998;**30**:1743–55.

Bouskill NJ, Tang J, Riley WJ *et al.* Trait-based representation of biological nitrification: model development, testing, and predicted community composition. *Front Microbiol* 2012;**3**:364.

Brochier-Armanet C, Talla E, Gribaldo S. The Multiple Evolutionary Histories of Dioxygen Reductases: implications for the Origin and Evolution of Aerobic Respiration. *Mol Biol Evol* 2009;**26**:285–97.

Brown JH, Gillooly JF, Allen AP *et al.* Toward a metabolic theory of ecology. *Ecology* 2004;**85**:1771–89.

Burris RH, Roberts GP. Biological nitrogen fixation. *Annu Rev Nutr* 1993;**13**:317–35.

Cania B, Vestergaard G, Krauss M *et al.* A long-term field experiment demonstrates the influence of tillage on the bacterial potential to produce soil structure-stabilizing agents such as exopolysaccharides and lipopolysaccharides. *Environmental Microbiome* 2019;**14**:1–14.

Cary SC, McDonald IR, Barrett JE *et al.* On the rocks: the microbiology of Antarctic Dry Valley soils. *Nat Rev Microbiol* 2010;**8**: 129–38.

Caulier S, Nannan C, Gillis A *et al.* Overview of the antimicrobial compounds produced by members of the Bacillus subtilis group. *Front Microbiol* 2019;**10**:302.

Christie PJ, Vogel JP. Bacterial type IV secretion: conjugation systems adapted to deliver effector molecules to host cells. *Trends Microbiol* 2000;**8**:354–60.

Conrad R. The global methane cycle: recent advances in understanding the microbial processes involved. *Environ Microbiol Rep* 2009;**1**:285–92.

Crits-Christoph A, Robinson CK, Barnum T *et al.* Colonization patterns of soil microbial communities in the Atacama Desert. *Microbiome* 2013;**1**:28.

de los Santos PE, Palmer M, Chávez-Ramirez B *et al.* Whole genome analyses suggests that Burkholderia sensu lato contains two additional novel genera (*Mycetohabitans* gen. nov., and *Trinickia* gen. nov.): implications for the evolution of diazotrophy and nodulation in the Burkholderiaceae. *Genes* 2018;**9**: DOI: 10.3390/genes9080389.

de Mendiburu F. *Agricolae: statistical procedures for agricultural research*, 2014. http://cran.r-project.org/web/packages/agricolae/index.html.

Dunfield PF. The Soil Methane Sink. *Greenhouse Gas Sinks* 2007;152–70.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**:1792–7.

Embley TM, Martin W. Eukaryotic evolution, changes and challenges. *Nature Reviews* 2006;**440**:623–30.

Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002;**30**:1575–84.

Fierer N, Bradford MA, Jackson RB. Toward an ecological classification of soil bacteria. *Ecology* 2007;**88**:1354–64.

Fierer N, Lauber CL, Ramirez KS *et al.* Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J* 2012;**6**: 1007–17.

Finn DR, Yu J, Ilhan ZE *et al.* MicroNiche: an R package for assessing microbial niche breadth and overlap from amplicon sequencing data. *FEMS Microbiol Ecol* 2020;**96**: fiaa131.

Finn DR, Ziv-el M, van Haren J *et al.* Methanogens and methanotrophs show nutrient-dependent community assemblage patterns across tropical peatlands of the Pastaza-Maranon Basin, Peruvian Amazonia. *Front Microbiol* 2020b;**11**:746.

Flemming HC, Wingender J. The biofilm matrix. *Nat Rev Microbiol* 2010;**8**:623–33.

Freilich S, Zarecki R, Eilam O *et al.* Competitive and cooperative metabolic interactions in bacterial communities. *Nat Commun* 2011;**2**.

Galperin MY. A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. *BMC Microbiol* 2005;**5**:35. DOI: 10.1186/1471-2180-1185-1135.

Gao B, Paramanathan R, Gupta RS. Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups. *Antonie Van Leeuwenhoek* 2006;**90**:69–91.

Garcia-Pichel F, Lopez-Cortes A, Nübel U. Phylogenetic and morphological diversity of Cyanobacteria in soil desert crusts from the Colorado Plateau. *Appl Environ Microbiol* 2001;**67**:1902–10.

Gause GF. Experimental studies on the struggle for existence I Mixed population of two species of yeast. *J Exp Biol* 1932;**9**:389–402.

Gleason HA. The individualistic concept of the plant association. *Bulletin of the Torrey Botanical Club* 1926;**53**:7–26.

Goldberg AL. Protein degradation and protection against misfolded or damaged proteins. *Nature* 2003;**426**:895–9.

Goldfarb KC, Karaoz U, Hanson CA *et al.* Differential growth responses of soil bacterial taxa to carbon substrates of varying chemical recalcitrance. *Front Microbiol* 2011;**2**.

Grime JP. Evidence for the existence of three primary strategies in plants and its relevance to ecological and evolutionary theory. *Am Nat* 1977;**111**:1169–94.

Grime JP. *Plant strategies and vegetation processes*. UK: Wiley, Chichester, 1979.

He M, Zhang J, Shen L *et al.* High-throughput sequencing analysis of microbial community diversity in response to indica and japonica bar-transgenic rice paddy soils. *PLoS One* 2019;**14**:e0222191.

Higgins CF. ABC transporters: physiology, structure and mechanism - an overview. *Res Microbiol* 2001;**152**:205–10.

Ho A, Kerckhof FM, Luke C *et al.* Conceptualizing functional traits and ecological characteristics of methane-oxidizing bacteria as life strategies. *Environ Microbiol Rep* 2013;**5**: 335–45.

Ho A, Paolo Di Lonardo D, Bodelier PL. Revisiting life strategy concepts in environmental microbial ecology. *FEMS Microbiol Ecol* 2017;**93**:1–14.

Hodgson DA. Primary metabolism and its control in streptomycetes: a most unusual group of bacteria. In: *Adv Microb Physiol*, **Vol 42**, Poole RK (ed.) 2000, 47–238.

Hug LA, Castelle CJ, Wrighton KC *et al.* Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* 2013;**1**:22.

Hutchinson GL. Concluding remarks. *Cold Spring Harb Symp Quant Biol* 1957;**22**:415–27.

Itakura M, Saeki K, Omori H *et al.* Genomic comparison of Bradyrhizobium japonicum strains with different symbiotic nitrogen-fixing capabilities and other Bradyrhizobiaceae members. *ISME J* 2009;**3**:326–39.

Jangid K, Williams MA, Franzluebbers A *et al.* Land-use history has a stronger impact on soil microbial community composition than aboveground vegetation and soil properties. *Soil Biol Biochem* 2011;**43**:2184–93.

Jarrell KF, Albers SV. The archaellum: an old motility structure with a new name. *Trends Microbiol* 2012;**20**:307–12.

Kaiser C, Franklin O, Richter A *et al.* Social dynamics within decomposer communities lead to nitrogen retention and organic matter build-up in soils. *Nat Commun* 2015;**6**.

Kanehisa M, Sato Y, Kawashima M *et al.* KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;**44**:D457–62.

Kearney M, Simpson SJ, Raubenheimer D *et al.* Modelling the ecological niche from functional traits. *Philos Trans Royal Soc B: Biolog Sci* 2010;**365**:3469–83.

Keddy PA. Assembly and response rules - two goals for predictive community ecology. *J Veg Sci* 1992;**3**:157–64.

Klotz MG, Norton JB. Multiple copies of ammonia monooxygenase (amo) operons have evolved under biased AT/GC mutational pressure in ammonia-oxidizing autotrophic bacteria. *FEMS Microbiol Lett* 1998;**168**:303–11.

Koch AL. Oligotrophs versus copiotrophs. *Bioessays* 2001;**23**: 657–61.

Kögel-Knabner I. The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter. *Soil Biol Biochem* 2002;**34**:139–62.

Kuznetsov SI, Dubinina GA, Lapteva NA. Biology of oligotrophic bacteria. *Annu Rev Microbiol* 1979;**33**:377–87.

Lauro FM, McDougald D, Thomas T *et al.* The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci* 2009;**106**:15527–33.

Leff JW, Jones SE, Prober SM *et al*. Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proc Natl Acad Sci* 2015;**112**:10967–72.

Leibold MA. The niche concept revisited: mechanistic models and community context. *Ecology* 1995;**76**:1371–82.

Liaw A, Weiner M. Classification and regression by randomForest. *R News* 2002;**2**:18–22.

Madigan M, Cox SS, Stegeman RA. Nitrogen fixation and nitrogenase activities in members of the Family Rhodospirillaceae. *J Bacteriol* 1984;**157**:73–78.

Madin JS, Nielsen DA, Brbic M *et al*. A synthesis of bacterial and archaeal phenotypic trait data. *Scientific Data* 2020;**7**:170.

McBride MJ. Bacterial gliding motility: mechanisms for cell movement over surfaces. *Annu Rev Microbiol* 2001;**55**:49–75.

McGill BJ, Enquist BJ, Weiher E *et al*. Rebuilding community ecology from functional traits. *Trends Ecol Evol* 2006;**21**:178–85.

Möller B, Oßmer R, Howard BH *et al*. Sporomusa, a new genus of Gram-negative anaerobic bacteria including Sporomusa sphaeroides spec. nov. and Sporomusa ovata spec. nov. *Arch Microbiol* 1984;**139**:388–96.

Nayfach S, Roux S, Seshadri R *et al*. A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2021;**39**:499–509.

Oksanen J, Guillaume Blanchet F, Kindt R *et al*. Vegan: community Ecology Package. R package version 2.0-10, 2013. http://CRAN.R-project.org/package=vegan.

Op den Camp HJM, Islam T, Stott MB *et al*. Environmental, genomic and taxonomic perspectives on methanotrophic Verrucomicrobia. *Environ Microbiol Rep* 2009;**1**:293–306.

Parks DH, Chuvochina M, Waite DW *et al*. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018;**36**:996–1004.

Pascual-Garcia A, Bell T. Community-level signatures of ecological succession in natural bacterial communities. *Nat Commun* 2020;**11**:2386.

Philippot L, Andersson SGE, Battin TJ *et al*. The ecological coherence of high bacterial taxonomic ranks. *Nat Rev Microbiol* 2010;**8**:523–9.

Philippot L, Bru D, Saby NPA *et al*. Spatial patterns of bacterial taxa in nature reflect ecological traits of deep branches of the 16S rRNA bacterial tree. *Environ Microbiol* 2009;**11**:3096–104.

Prosser JI. Ecosystem processes and interactions in a morass of diversity. *FEMS Microbiol Ecol* 2012;**81**:507–19.

R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for statistical computing, 2013.

Roller BRK, Schmidt TM. The physiology and ecological implications of efficient growth. *The ISME Journal* 2015;**9**:1481–7.

Sancar A. DNA excision repair. *Annu Rev Biochem* 1996;**65**:43–81.

Sangwan P, Chen XL, Hugenholtz P *et al*. Chthoniobacter flavus gen. nov., sp nov., the first pure-culture representative of subdivision two, Spartobacteria classis nov., of the phylum Verrucomicrobia. *Appl Environ Microbiol* 2004;**70**:5875–81.

Schliep KP, Potts AJ, Morrison DA *et al*. Intertwining phylogenetic trees and networks.. *Methods Ecol Evol* 2017;**8**:1212–20.

Schneiker S, Perlova O, Kaiser O *et al*. Complete genome of the myxobacterium Sorangium cellulosum. *Nat Biotechnol* 2007;**25**:1281–9.

Semenov AM. Physiological bases of oligotrophy of microorganisms and the concept of microbial community. *Microb Ecol* 1991;**22**:239–47.

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

Smirnova GV, Oktyabrsky ON. Glutathione in Bacteria. *Biochemistry* 2005;**70**:1199–211.

Spang A, Saw JH, Jorgensen SL *et al*. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 2015;**521**:173–9.

Stackebrandt E, Pohla H, Kroppenstedt RM *et al*. 16S rRNA analysis of Sporomusa, Selenomonas and Megasphaera: on the phylogenetic origin of Gram-positive Eubacteria. *Arch Microbiol* 1985;**143**:270–6.

Tamas I, Smirnova AV, He Z *et al*. The (d)evolution of methanotrophy in the Beijerinckiaceae - a comparative genomics analysis. *ISME J* 2014;**8**:369–82.

Terzenbach DP, Blaut M. Transformation of tetrachloroethylene to trichloroethylene by homoacetogenic bacteria. *FEMS Microbiol Lett* 1994;**123**:213–8.

Trivedi P, Anderson IC, Singh BK. Microbial modulators of soil carbon storage: integrating genomic and metabolic knowledge for global prediction. *Trends Microbiol* 2013;**21**:641–51.

van Rossum T, Ferretti P, Maistrenko OM *et al*. Diversity within species: interpreting strains in microbiomes. *Nat Rev Microbiol* 2020;**18**:491–506.

Vieira-Silva S, Rocha EPC. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLos Genet* 2010;**6**:e1000808.

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

Weins JJ. Testing phylogenetic methods with tree congruence: phylogenetic analysis of polymorphic morphological characters in Phrynosomatid lizards. *Syst Biol* 1998;**47**:427–44.

Wessen E, Hallin S, Philippot L. Differential responses of bacterial and archaeal groups at high taxonomical ranks to soil management. *Soil Biol Biochem* 2010;**42**:1759–65.

Wickham H. Reshaping data with the reshape package. *J Statis Soft* 2007;**21**:1–20.

Young J, Holland IB. ABC transporters: bacterial exporters-revisited five years on. *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1999;**1461**:177–200.

Zhu CS, Delmont TO, Vogel TM *et al*. Functional Basis of Microorganism Classification. *PLoS Comput Biol* e1004472, 2015;**11**.