A global perspective on microbial diversity in the terrestrial deep subsurface

Authors:

Soares, A.[1,2,3], Edwards, A.[2,3*], An, D.[4], Bagnoud, A.[5], Bomberg, M.[6], Budwill, K.[7], Caffrey, S. M.[8], Fields, M.[9], Gralnick., J.[10], Kadnikov, V.[11,12], Momper, L.[13], Osburn, M.[14], Moreau, J.W.[15], Moser, D.[19], Mu, A.[15,16,17,18], Purkamo, L.[6,20,21], Rassner, S. M.[1,3], Sheik, C. S.[22], Sherwood Lollar, B. [23], Toner, B. M.[24], Voordouw, G.[4], Wouters, K.[26], Mitchell, A. C.[1,3*]

Affiliations:

1. Department of Geography and Earth Sciences (DGES), Aberystwyth University (AU), Wales, UK
2. Institute of Biology, Environmental and Rural Sciences (IBERS), AU
3. Interdisciplinary Centre for Environmental Microbiology (iCEM), AU
4. Department of Biological Sciences, University of Calgary, Canada
5. Institut de Génie Thermique (IGT), Haute École d'Ingénierie et de Gestion du Canton de Vaud (HEIG-VD), Yverdon-les-Bains, Switzerland
6. VTT Technical Research Centre of Finland, Finland
7. Alberta Innovates, Canada
8. University of Toronto, Canada (UT)
9. Center for Biofilm Engineering (CBE), Montana State University (MSU), USA
10. Department of Plant and Microbial Biology, UM, USA
11. Faculty of Biology, Moscow State University (MoSU), Russia
12. Institute of Bioengineering, Research Center of Biotechnology, Russian Academy of Sciences, Russia
13. Department of Earth, Atmospheric and Planetary Sciences (DEAPS), The Massachusetts Institute of Technology (MIT), United States of America (USA)
14. Department of Earth and Planetary Sciences (DEPS), Northwestern University (NWU), USA
15. School of Earth Sciences, The University of Melbourne (UM), Parkville, Australia
16. Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, UM
17. Doherty Applied Microbial Genomics, Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, UM

18. Microbiological Diagnostic Unit Public Health Laboratory, Department of Microbiology and Immunology, UM
19. Division of Hydrologic Sciences, Desert Research Institute (DRI), Las Vegas, NV, USA
20. School of Earth and Environmental Sciences (SEES), University of St. Andrews (USA), Scotland, UK
21. Geological Survey of Finland (GTK), Finland
22. Large Lakes Observatory, University of Minnesota – Duluth (UMD)
23. Department of Earth Sciences, UT, Canada
24. Department of Soil, Water & Climate, University of Minnesota
25. Institute for Environment, Health and Safety (EHS), Belgian Nuclear Research Centre SCK•CEN, Mol, Belgium

*__Corresponding Authors__: AC Mitchell nem@aber.ac.uk; A Edwards aye@aber.ac.uk

Summary

While recent efforts to catalogue Earth's microbial diversity have focused upon surface and marine habitats, 12% to 20% of Earth's bacterial and archaeal biomass is suggested to inhabit the terrestrial deep subsurface, compared to ~1.8% in the deep subseafloor [1–3]. Metagenomic studies of the terrestrial deep subsurface have yielded a trove of divergent and functionally important microbiomes from a range of localities [4–6]. However, a wider perspective of microbial diversity and its relationship to environmental conditions within the terrestrial deep subsurface is still required. Here, we show the diversity of bacterial communities in deep subsurface groundwater is controlled by aquifer lithology globally, by using 16S rRNA gene datasets collected across five countries on two continents and from fifteen rock types over the past decade. Furthermore, our meta-analysis reveals that terrestrial deep subsurface microbiota are dominated by Betaproteobacteria, Gammaproteobacteria and Firmicutes, likely as a function of the diverse metabolic strategies of these taxa. Despite this similarity, evidence was found not only for aquifer-specific microbial communities, but also for a common small consortium of prevalent Betaproteobacteria and Gammaproteobacterial OTUs across the localities. This finding implies a core terrestrial deep subsurface community, irrespective of aquifer lithology, that may play an important role in colonising and sustaining microbial habitats in the deep terrestrial subsurface. An *in-silico* contamination-aware approach to analysing this dataset underscores the importance of downstream methods for assuring that robust conclusions can be reached from deep subsurface-derived sequencing data. Understanding the global panorama of microbial diversity and ecological dynamics in the deep terrestrial subsurface provides a first step towards understanding the role of microbes in global subsurface element and nutrient cycling.

Main text

Understanding the distribution of microbial diversity is pivotal for advancing our knowledge of deep subsurface global biogeochemical cycles [7,8]. Subsurface biomass is suggested to have exceeded that of the Earth's surface by an order of magnitude (~45% of Earth's total biomass) before land plants evolved, at ca. 0.5 billion years ago [9]. Integrative modelling of cell count and quantitative PCR (qPCR) data and geophysical factors indicated in late 2018 that the bacterial and archaeal biomass found in the global deep subsurface may range from 23 to 31 petagrams of carbon (PgC). These values halved previous efforts from earlier that year[10] but maintained the notion that the terrestrial deep subsurface holds ca. 5-fold more bacterial and archaeal biomass than the deep marine

subsurface. Further, it is expected that 20-80% of the possible 2-6 x 10^{29} prokaryotic cells present in the terrestrial subterranean biome exist as biofilms and play crucial roles in global biogeochemical cycles [10,11].

Cataloguing microbial diversity and functionality in the terrestrial deep subsurface has mostly been achieved by means of marker gene and metagenome sequencing in coals, sandstones, carbonates, and clays, as well as deep igneous and metamorphic rocks [4–6,12–20] . Only recently has the first comprehensive database of 16S rRNA gene-based studies targeting terrestrial subsurface environments been compiled [10]. This work focused on updating estimates for bacterial and archaeal biomass, and cell numbers across the terrestrial deep subsurface, but also linked the identified bacterial and archaeal phylum-level compositions to host-rock type, and to 16S rRNA gene region primer targets [10]. While highlighting Firmicutes and Proteobacterial dominance in the bacterial component of terrestrial deep subsurface, no further taxonomic insights were gained. However, genus-level identification is critical for understanding community composition, inferred metabolism and hence microbial contributions of distinct community members to biogeochemical cycling in the deep subsurface [18,21–23]. Indeed, such genus-specific traits have been demonstrated as critical for understanding crucial biological functions in other microbiomes [24], and genus-specific functions of relevance for deep subsurface biogeochemistry are clear [25,26].

So far, the potential biogeochemical impacts of microbial activity in the deep subsurface have been inferred through shotgun metagenomics, as well as from incubation experiments of primary geological samples amended with molecules or minerals of interest [13,19,20,27–30]. Recent studies of deep terrestrial subsurface microbial communities further suggest that these are metabolically active, generally associated with novel uncultured phyla, and potentially directly involved in carbon and sulphur cycling [31–36]. Concomitant advancements in subsurface drilling, molecular methods and computational techniques have aided the exploration of the subsurface biosphere, but serious challenges remain mostly related to deciphering sample contamination by drilling methods and sample transportation to laboratories for processing [37,38]. The logistical challenges inherent to accessing and recovering *in situ* samples from hundreds to thousands of metres below surface complicate our view of terrestrial subsurface microbial ecology [39].

In this study, we capitalize on the increased availability of 16S rRNA gene amplicon data from multiple studies of the terrestrial deep subsurface conducted over the last decade. We apply bespoke bioinformatic scripts to generate insights into the microbial community structure and controls upon bacterial microbiomes of the terrestrial deep subsurface across a large distribution of habitat types on

multiple continents. The deep biosphere is as-yet undefined as a biome - elevated temperature, anoxic conditions, low levels of organic carbon, and measures of isolation from the surface photosphere are some of the criteria used albeit without a consensus. For this work a more general approach has been taken to define the terrestrial deep subsurface as the zone at least 100 m from the surface [40,41].

Meta-analysis of the terrestrial deep subsurface microbiome

Here, we were able to compare datasets encompassing different 16S rRNA gene hyper-variable regions, and derived from different DNA extraction methodologies, facilitated by closed-reference Operational Taxonomic Unit (OTU)-picking of each study individually using the same 16S rRNA gene reference database. This procedure begins to address technically confounding variables by limiting taxonomy assignments to only the archaeal and bacterial diversity listed in the chosen database and precludes the discovery of novel taxa.

The finalized meta-analysis dataset comprised of 16S rRNA data from seventeen aquifers in either sedimentary- or crystalline-host rocks, from depths spanning 94 m to 2300 m below land surface (mbls), targeting mostly groundwater across 5 countries and two continents (**Supplementary Table 3**). Nine DNA extraction techniques were used in these studies, ranging from standard and modified kit protocols (e.g. MOBIO® PowerSoil, see **Table 1**) to phenol-chloroform and CTAB/NaCl based methods [42–47]. Finally, 6 different primer pair amplified regions of the 16S rRNA gene, in 454 pyrosequencing and Illumina sequencing, were used to generate the datasets.

Table 1. Metadata table for the studies utilized in this meta-analysis (*cf.* **Supplementary Table 3** and **Supplementary Figure 3** for more details). NA is used as an acronym for "not available". The dataset unavailable through SRA is available through http://hmp.ucalgary.ca/HMP/.

SRA Accession	Originator	DOI	Year of Sampling	Final no. of samples	Location	Depth gradient (m)	Host-rock	Final no. of sequences	Final no. of OTUs
PRJNA262938	Magdalena Osburn, Lily Momper	10.3389/fmicb.2014.00610	2013/2014	6	South Dakota, USA	243.84-1478.28	Sulphide-rich schists	170364	1367
PRJNA268940	Duane Moser	NA	2007-2011	8	California, USA; Ontario, Canada	94-2383	Dolomite, Tuff, Rhyolite/tuff-breccia	68071	741
PRJNA248749	Jeffrey Gralnick	NA	2011	6	Minnesota, USA	730	Hematite	69757	511
PRJNA251746	Rick Colwell	NA	2009	6	Washington, USA	393-1135	Basalt	30121	613
PRJNA375701	Elliot Barnhart	10.1016/j.coal.2016.05.001	2013	5	Montana, USA	109-114.7	Sub-bituminous coal, Shale, Sandstone, Siltstone	35926	154
NA	Karen Budwill	10.1128/AEM.01737-15	2009	38	Alberta, Canada	140-1064.4	Sub-bituminous coal, Volatile bituminous coal	100618	5110
PRJEB1468	Katinka Wouters	10.1111/1574-6941.12171		6	Mol, Belgium	217-232	Kaolinite/illite and smectite	47123	497
PRJEB10822	Lotta Purkamo	10.5194/bg-13-3091-2016	2009-2011	7	Outokumpu, Finland	180-2300	Mica-schist, Biotite-gneiss, Chlorite-sericite schist	12290	177

Initial processing of 187 retrieved samples revealed 24,632,035 chimera-checked sequences [17,28,43,44,48–50]. SILVA 123-aided closed-reference OTU-picking yielded 6,975 OTUs associated to 598,341 sequences following exclusion of singleton OTUs and samples containing 2 or less OTUs. The final dataset following stricter contamination-aware filtering (*cf.* **Methodology**) was comprised of 70 samples and 2,207 OTUs (513,929 sequences, 2.54% of the initial sequences), where Archaeal reads comprised 1.5% of the total number of reads.

Trends in taxonomic diversity

Among a total of 45 detected bacterial phyla, Proteobacteria were seen to dominate most community profiles in this dataset (**Figure 1**). The most abundant proteobacterial classes (Alpha-, Betaproteobacteria, Delta-, Gammaproteobacteria) represented 57.2% of the total number of reads, with 13.4% of these assigned to class Clostridia (Firmicutes). A general prevalence of Betaproteobacteria and Gammaproteobacteria in the deep biosphere may be explained by the diverse metabolic capabilities of taxa within these clades. Families Gallionellaceae, Pseudomonadaceae, Rhodocyclaceae and Hydrogeniphillaceae within Betaproteobacteria and Gammaproteobacteria are suggested to play crucial roles in deep subsurface iron, nitrogen, sulphur and carbon cycling across the world [43,51,52]. The relative abundance of order Burkholderiales (Betaproteobacteria) in surficial soils has previously been correlated (R^2=0.92, ANOVA p-value <0.005) with mineral dissolution rates, while genus *Pseudomonas* (Gammaproteobacteria) is widely known to playing a key role in hydrocarbon-degradation, denitrification and coal solubilisation in different locations[53–55].

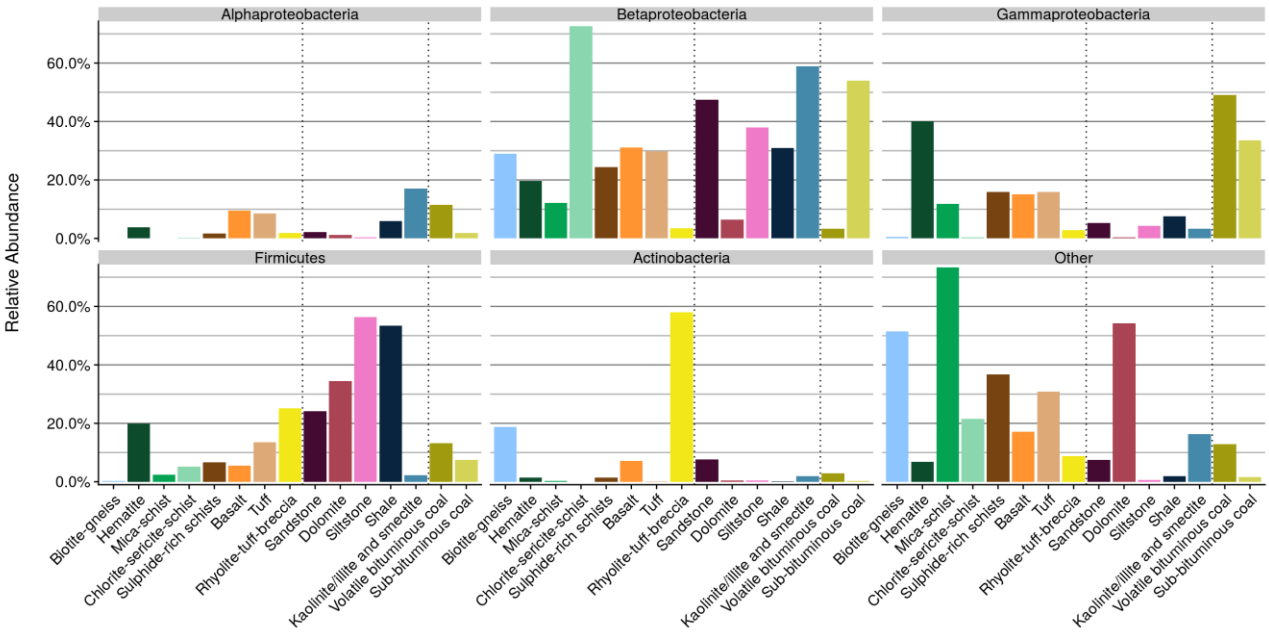


Figure 1. Mean relative abundances (%, y-axis) of the most abundant taxonomic groups across the dataset across all analysed aquifer lithologies (x-axis). Vertical dashed lines divide crystalline and

sedimentary rocks. Coals ranks are also separated due to their higher sample contribution to the dataset.

Mean grouped proportion values indicated that Betaproteobacteria were the most abundant proteobacterial class in most host rocks, representing 26.1% of all reads in the dataset. While Betaproteobacteria accounted for 53.96% of the community profile for sub-bituminous coals, Gammaproteobacteria dominated volatile bituminous coals (49.1% of the profile, **Figure 1**). The dominance of Betaproteobacteria and Gammaproteobacteria in coals builds on culture-based evidence of widespread degradation of coal-associated complex organic compounds by these classes [56–59].

Firmicutes were represented in large part by class Clostridia and mostly associated with sedimentary aquifers (i.e. sandstone, dolomite, siltstone, shale – **Figure 1**). This class includes ubiquitous anaerobic hydrogen-driven sulphate reducers also known to sporulate and metabolize a wide range of organic carbon compounds that have been found to dominate extremely deep subsurface ecosystems beneath South Africa and likely globally given the pervasive high levels of H_2 in similar geologic settings [13,60–62]. Clostridia from metagenomes have been detected from the terrestrial deep subsurface and inferred to have the physiological capabilities needed to thrive in these environments [20]. Adaptation to extreme environments in Clostridia is posed to be driven by varied metabolic potential, sporulation ability, and capacity for CO_2- or sulphur-based autotrophic H_2-dependent growth [22,63].

Lithological controls on community structure

Microbial community structure and composition in soils depend on fine-tuned geochemical, physical and hydrogeological conditions that influence microbial presence and metabolism [64]. This relationship also appears to be reflected in the global subsurface, where host-rock lithology is evident as a primary control on community structure (**Figures 2** and **3**). Indeed, most host-rocks (10 out of 15 in this dataset) have, on average, more unique OTUs than they share with other host-rocks (**Figure 2**). Particularly, in sulphide-rich schists, 73% of the OTUs are, on average, unique to the host-rock. The role of host-rock lithology is further evidenced (**Figure 3**) as some of the host-rocks clustered at a 95% confidence interval, suggesting closely related microbial communities within similar lithologies, despite other environmental factors such as depth or location. Further, 50.6% of Jensen-Shannon distances ordinated (**Figure 3**) were significantly explained by aquifer lithology (ADONIS/PERMANOVA, F-statistic=4.65, p-value <0.001, adjusted Bonferroni correction p-value <0.001); thus showing that rock type was the primary variable defining microbial community

structure. Other environmental features such as absolute sample depth and medium-scale location (i.e. state, region of the sampling site) explained only 3.08% and 2.78% of the significant metadata-driven variance in microbial community structure, respectively (ADONIS/PERMANOVA, F-statistic=3.95, 3.57 p-value <0.001, adjusted Bonferroni correction p-value <0.001). This suggests that depth-related changes in temperature and pressure are not significant controls on community structure. The relationship of community structure to hydrogeochemical parameters remains an area for future investigation – since hydrogeology and fluid geochemistry are also strongly controlled by lithology via water-rock reactions.

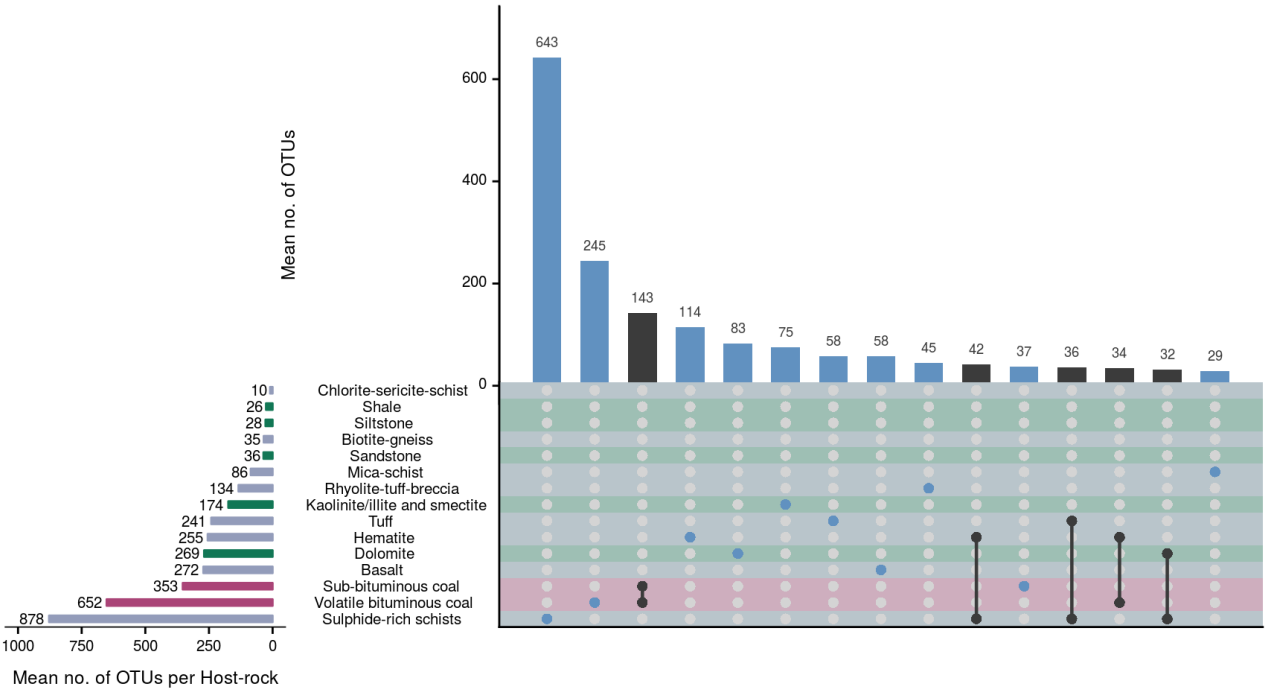


Figure 2. *UpsetR* plot of mean numbers of OTU interactions among rock types. Only interactions involving 25 or more OTUs on average are shown. Coloured matrix rows correspond to host-rocks, and are coloured according to rock type: blue for crystalline, green for sedimentary rocks and pink for coals, which were highlighted due to their higher sample contribution to the dataset. Columns depict OTU interactions: blue dots mark independent (mean number of non-shared OTUs) interactions and black dots connected by black lines mark shared OTUs between two or more host-rocks. Shared interactions are composed of only the host-rocks marked by dots. Vertical bars on top of the coloured matrix correspond to mean OTU numbers present in the described interactions and are coloured black or blue if depicting shared or non-shared interactions, respectively. Horizontal bars by the left of the coloured matrix depict mean total numbers of OTUs per host-rock.

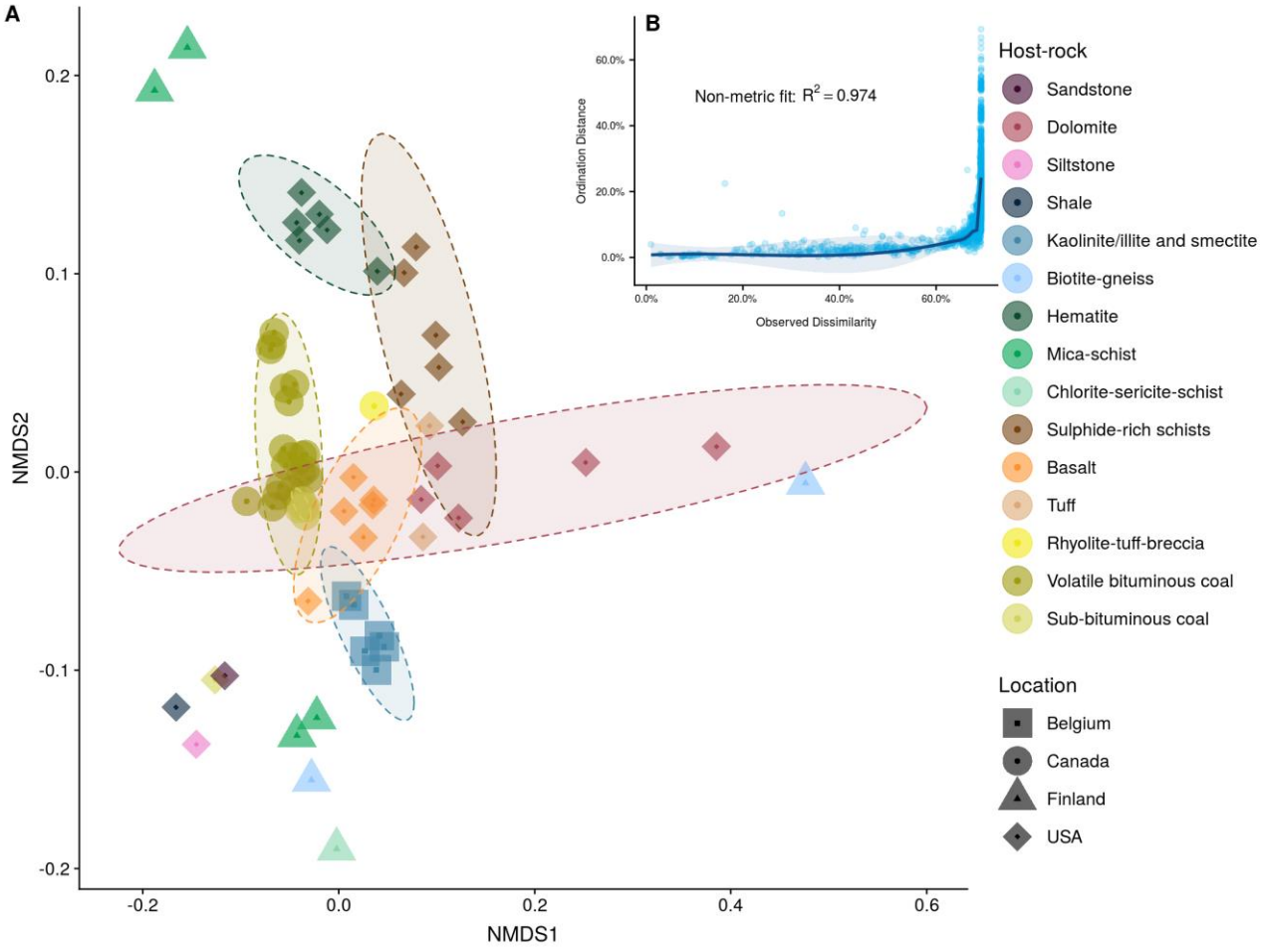


Figure 3. Non-metric Multidimensional Scaling (nMDS) of Jensen-Shannon distances between samples (A). Shapes correspond to different locations, whereas colours depict host-rocks targeted in this study. B depicts a Shepard's stress plot of observed (original) dissimilarities and ordination distances. An R^2 measure of stress is further shown for the non-metric fit of the variables. Confidence interval ellipses were plotted at a level of 95% according to host-rock.

Metadata variables that were unavailable for all samples in the dataset were excluded from the statistical analyses, thus further insights into the significance of other environmental variables was not possible. Nevertheless, this is the first large-scale evidence that deep subsurface microbial community taxonomy appears host-rock-specific. Given the importance of chemolithotrophic metabolisms in dark, subsurface environments, the unique chemical and mineralogical compositions within different aquifer lithologies impart strong controls over microbiomes associated with mineral surfaces and porewaters [43,62,65–67]. Indeed, direct utilisation of mineral surfaces or dissolved species from minerals for respiration and/or metabolism has been shown to be critical in localised subsurface environments [30,35,68,69]. Due to low numbers of samples for some host-rock lithologies in this dataset (e.g. one sample each for siltstone, sandstone, shale and chlorite-sericite-schist), it is not possible to ascertain that microbiome specificity is generalisable to all deep subsurface aquifer types on Earth

(see **Figure 3**). Nevertheless, this study provides the first large-scale evidence that, at a global scale, lithology surpasses depth in shaping deep subterranean microbial communities.

A core terrestrial deep subsurface microbial community?

Analysis of prevalence across the dataset revealed that seven OTUs, all affiliated to genus *Pseudomonas*, were present in more than 25 and up to 41 samples (see **Supplementary Figure 1**, **Supplementary Table 2**). Network analysis (**Table 2**) highlighted a *Pseudomonas* OTU highly connected to other OTUs in the dataset. Further, BLAST[70] results indicated that recovered sequences for OTUs affiliated to this genus were generally associated to marine and terrestrial soil and sediments (*cf* **Supplementary Table 4, Supplementary Figure 4**). Four OTUs affiliated to Burkholderiales (Betaproteobacteria), the second most prevalent order in the dataset, were also found to be connected to up to 34 other OTUs. Genus *Thauera* (Betaproteobacteria, Rhodocyclales), represented by a single OTU, was the second most central to the dataset. Finally, network and prevalence analysis highlighted the putative importance classes Betaproteobacteria and Gammaproteobacteria may have in the deep subsurface, since taxa affiliated to these were highly connected across the dataset (**Table 2**). These observations suggest genus-level taxonomy may be relevant to shaping subterranean microbial communities irrespective of host-rock lithology.

Table 2. Top 10 most central OTUs in the Jaccard distances network (as defined by eigenvector centrality scores, or the scored value of the centrality of each connected neighbour of an OTU) and correspondent closeness centrality (scores of shortest paths to and from an OTU to all the remainder in a network) and degree (number of directly connected edges, or OTUs) values.

OTU ID	OTU Classification	Centrality	Closeness	Degree
EF554871.1.1486	Proteobacteria; **Gammaproteobacteria**; Pseudomonadales; Pseudomonadaceae; Pseudomonas	1.0000000	2.13e-05	38
HH792638.1.1492	Proteobacteria; **Betaproteobacteria**; Rhodocyclales; Rhodocyclaceae; Thauera	0.9753542	2.13e-05	36
HQ681977.1.1496	Proteobacteria; **Betaproteobacteria**; Burkholderiales; Comamonadaceae; Diaphorobacter	0.9445053	2.13e-05	34
KF465077.1.1336	Proteobacteria; **Betaproteobacteria**; Burkholderiales; Comamonadaceae; Acidovorax	0.8887751	2.13e-05	30
JQ072853.1.1348	Proteobacteria; **Betaproteobacteria**; Rhodocyclales; Rhodocyclaceae; Thauera	0.8808435	2.13e-05	30
KM200734.1.1449	Proteobacteria; **Alphaproteobacteria**; Rhizobiales; Rhizobiaceae; Rhizobium	0.8716886	2.13e-05	31
KC758926.1.1392	Proteobacteria; **Betaproteobacteria**; Burkholderiales; Comamonadaceae; Acidovorax	0.8662805	2.13e-05	29
FJ032194.1.1456	Proteobacteria; **Betaproteobacteria**; Burkholderiales; Comamonadaceae; Rhodoferax	0.8662805	2.13e-05	29
EU771645.1.1366	**Firmicutes**; Bacilli; Bacillales; Planococcaceae; Planomicrobium	0.8476970	2.13e-05	30
JN245782.1.1433	Proteobacteria; **Alphaproteobacteria**.; Rhodobacterales; Rhodobacteraceae; Defluviimonas	0.8356655	2.13e-05	29

The metabolic plasticity of Pseudomonadales and Burkholderiales orders has been demonstrated [71–74] and may be a catalyst for their apparent centrality across the terrestrial deep subsurface microbiomes analysed in this study. These bacterial orders may represent important keystone taxa in microbial consortia responsible for providing key substrates to other colonizers in deep subsurface environments [75,76]. In particular, given the number of highly central *Pseudomonas*-affiliated OTUs and the prevalence of this genus in the dataset, we suggest that this genus may be key in establishing conditions for microbial colonization in many terrestrial subsurface environments. Genus *Pseudomonas* and possibly several members of Burkholderiales may therefore comprise an important component of the global core terrestrial deep subsurface microbial community.

Challenges from contamination

16S rRNA gene PCR-based approaches for characterizing microbial diversity in low biomass environments benefit from the sensitivity afforded by PCR, at the cost of vulnerability to contamination [77]. Here, we used the prominence of sequences associated with phototrophic taxa as an indicator of either ingress of surface microbiota or contamination during sample processing. The discovery of potentially photosynthetic taxa in the initial dataset, namely 46 OTUs classified as Chloroplast (Cyanobacteria) was read as a sign that further bioinformatics-driven precautions should be taken, despite recent evidence of some cyanobacterial presence in some locations within the deep subsurface [78,79]. Specifically, the presence of other phototrophic members of phyla Chloroflexi and Chlorobi as well as classes Rhodospirillales (Alphaproteobacteria) and Chromatiales (Gammaproteobacteria) informed the decision to filter the dataset to hold only OTUs represented by more than 500 sequences and present in at least 10 samples. Recent recommendations for quality control of 16S rRNA gene datasets also support filtering-based approaches when applied to low biomass subsurface environments [38]. This constraint reduced the dataset to the 70 samples and 2,207 OTUs (513,929 sequences) used for the meta-analysis (**Table 1)**, and also reduced the number of prospective contaminants by half, although only ~26% of the reads associated to Chloroplast-like sequences were removed (17 OTUs, 1958 reads).

Collecting contamination-free samples from the deep subsurface is difficult but important for cataloguing the authentic microbial diversity of the terrestrial subsurface. This study follows recent recommendations for downstream processing of contaminant-prone samples originated in the deep subsurface (Census of Deep Life project - http://codl.coas.oregonstate.edu/), where physical, chemical and biological, but also *in-silico* bioinformatics strategies to prevent erroneous conclusions have been highlighted [38,80,81]. This study also follows frequency-based OTU filtration techniques

similar to those recommended in Sheik *et al.* (2018) [38] and designed to remove possible contaminants introduced during sampling or during the various steps related to sample processing. The pre-emptive quality control steps hereby undertaken support a non-contaminant origin for taxa analysed in this dataset following careful in-field and laboratorial contamination-aware procedures carried out in each study. As such, the predominance of typically contaminant taxa affiliated *e.g.* to genus *Pseudomonas* was accepted as a true trend in the microbial ecology of the terrestrial deep subsurface.

No evidence was found for DNA extraction and PCR procedures significantly affecting microbial community structure in this meta-analysis (6.01% of the microbial community structure cumulatively [ADONIS/PERMANOVA, F-statistic=3.85, 3.23, p-value <0.01, adjusted Bonferroni correction p-value <0.001] vs. 50.6% from host rock lithology). In spite of this, a general convergence in DNA extraction methods would help further reduce methodology based variation and to standardize downstream analysis of deep subsurface microbial datasets [82], despite the practical challenges of each host-rock matrix and local geochemical conditions.

In the near future, the advent of recently developed techniques for primer bias-free long read 16S rRNA and 16S rRNA-ITS gene amplicon long-read-based sequencing may initiate a convergence of molecular methods from which the deep subsurface microbiology community would benefit greatly [83,84]. The future of large-scale, collaborative deep subsurface microbial diversity studies should encompass not only an effort towards standardization of several molecular biology techniques but also the long-term archival of samples [85]. This will permit re-analyses using updated or unified methods after collection, where methodological variations would be controlled, and robust conclusions would more easily be achieved.

Conclusions

A global scale meta-analysis addressing the available 16S rRNA gene-based studies of the deep terrestrial subsurface revealed the dominance of Betaproteobacteria, Gammaproteobacteria and Firmicutes across this biome. Further, aquifer lithology was identified as the main driver of deep subterranean microbial communities. Depth and location were not significant controls of microbial community structure at this scale. Finally, evidence for a core terrestrial deep subsurface microbiome population was recognised through the prevalence and centrality of genus *Pseudomonas* (Gammaproteobacteria) and several other genera affiliated to class Betaproteobacteria. The adaptable metabolic capabilities associated to the above-mentioned taxa may be critical for colonizing the deep subsurface and sustaining communities. The terrestrial deep subsurface is a hard-to-reach complex

ecosystem crucial to global biogeochemical cycles. This study attempts to consolidate a global-scale understanding of taxonomical trends underpinning terrestrial deep subsurface microbial ecology and geomicrobiology.

Methodology

Data acquisition

The Sequence Read Archive database of the National Center for Biotechnology Information (SRA-NCBI) was queried for 16S rRNA-based deep subsurface datasets (excluding marine and ice samples, as well as any human-impacted samples); available studies, were downloaded using the SRA Run Selector. Studies were selected considering the metadata and information on sequencing platform used – i.e., only samples derived from 454 pyrosequencing and Illumina sequencing were considered. Analysis of related literature resulted in the detection of other deposited studies previous search efforts in NCBI-SRA failed to detect. Further private contacts allowed access to unpublished data included in this study. The final list of NCBI accession numbers**,** totalling 222 samples, was downloaded using *fastq-dump* from the SRA toolkit (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/).

As seen in **Table 1**, required metadata included host-rock lithology, general and specific geographical locations, depth of sampling, DNA extraction method, sequenced 16S rRNA gene region and sequencing method. Any samples for which the above-mentioned metadata could not be found were discarded and not considered for downstream analyses.

Pre-processing of 16S rRNA gene datasets

A customised pipeline was created in *bash* language making use of *python* scripts developed for QIIME v1.9.1 [86], to facilitate bioinformatic analyses in this study (see https://github.com/GeoMicroSoares/mads_scripts for scripts). Briefly, demultiplexed FASTQ files were processed to create an OTU table. Quality control steps involved trimming, quality-filtering and chimera checking by means of USEARCH 6.1 [87]. Sequence data that passed quality control were then subjected to closed-reference (CR) OTU-picking on a per-study basis using UCLUST [87] and reverse strand matching against the SILVA v123 taxonomic references (https://www.arb-silva.de/documentation/release-123/). Closed-reference OTU picking excludes OTUs whose taxonomy has not been found in the 16S rRNA gene database used. Although this limits the recovery of prokaryotic diversity to the recorded in the database, cross-study comparisons of microbial communities generated by different 16S rRNA gene primers are made possible. This conservative

approach classified OTUs in each study individually to the common 16S rRNA gene reference database the merge of all classification outputs. A single BIOM (Biological Observation Matrix) file was generated using QIIME's *merge_otu_tables.py* script. The BIOM file was then filtered to exclude samples represented by less than 2 OTUs using *filter_samples_from_otu_table.py*, as well as OTUs represented by one sequence (singleton OTUs) by using *filter_otus_from_otu_table.py*. In an attempt to reduce the impacts of potential contaminant OTUs from the dataset, the post-singleton filtered dataset was further filtered to include only OTUs represented by at least 500 sequences and present in at least 10 samples overall using *filter_otus_from_otu_table.py*.

Data analysis

All downstream analyses were conducted using the *phyloseq* (https://github.com/joey711/phyloseq) package within R, which allowed for simple handling of metadata and taxonomy and abundance data [88–90]. Merged and filtered BIOM files were imported into R using internal *phyloseq* functions, which allowed further filtering, transformation and plotting of the dataset (see https://github.com/GeoMicroSoares/mads_scripts for scripts).

Briefly, following a general assessment of the number of reads across samples and OTUs, *tax_glom* (*phyloseq*) allowed the agglomeration of the OTU table at phylum-level. For the metadata category-directed analyses, function *merge_samples* (*phyloseq*) created averaged OTU tables, which permitted testing of hypotheses for whether geology or depth had significant impacts on microbial community structure and composition. Computation of a Jensen-Shannon divergence PCoA (Principal Coordinate Analysis) was achieved with ordinate (*phyloseq*) which makes use of metaMDS (*vegan*) [91,92]. All figures were plotted making use of the *ggplot2* R package (https://github.com/tidyverse/ggplot2), except for the UpsetR plot in **Figure 2**, which was plotted with package *UpsetR* (https://github.com/hms-dbmi/UpSetR).

References

1. McMahon, S. & Parnell, J. Weighing the deep continental biosphere. *FEMS Microbiol. Ecol.* **87,** 113–120 (2014).
2. Bar-On, Y. M., Phillips, R., Milo, R. & Falkowski, P. G. The biomass distribution on Earth. *PNAS* (2018). doi:10.1073/pnas.1711842115
3. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551,** 457–463 (2017).
4. Long, P. E., Williams, K. H., Hubbard, S. S. & Ban, J. F. Microbial Metagenomics Reveals Climate-Relevant Subsurface Biogeochemical Processes. *Trends Microbiol.* **24,** 1–11 (2016).
5. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7,** 1–11 (2016).
6. Hug, L. A. *et al.* Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environ. Microbiol.* **18,** 159–173 (2016).
7. Jelen, B. I., Giovannelli, D. & Falkowski, P. G. The Role of Microbial Electron Transfer in the Coevolution of the Biosphere and Geosphere. 45–62 (2016). doi:10.1146/annurev-micro-102215-095521
8. Falkowski, P. G., Fenchel, T. & Delong, E. F. The Microbial Engines That Drive Earth ' s Biogeochemical Cycles. *Science (80-.).* **320,** 1034–1039 (2008).
9. McMahon, S. & Parnell, J. The Deep History of Earth's Biomass. *J. Geol. Soc. London.* jgs2018-061 (2018). doi:10.1144/jgs2018-061
10. Magnabosco, C. *et al.* The biomass and biodiversity of the continental subsurface. *Nat. Geosci.* 1 (2018). doi:10.1038/s41561-018-0221-6
11. Flemming, H.-C. & Wuertz, S. Bacteria and archaea on Earth and their abundance in biofilms. *Nat. Rev. Microbiol.* 1 (2019). doi:10.1038/s41579-019-0158-9
12. Gihring, T. M. *et al.* The distribution of microbial taxa in the subsurface water of the kalahari shield, south africa. *Geomicrobiol. J.* **23,** 415–430 (2006).
13. Moser, D. P. *et al.* Desulfotomaculum spp. and Methanobacterium spp. Dominate 4-5 km Deep Fault. *Appl. Environ. Microbiol.* **71,** 8773–8783 (2005).
14. Miettinen, H. *et al.* Microbiome composition and geochemical characteristics of deep subsurface high-pressure environment, Pyhäsalmi mine Finland. *Front. Microbiol.* **6,** 1–16 (2015).
15. Bagnoud, A. *et al.* Reconstructing a hydrogen-driven microbial metabolic network in Opalinus Clay rock. *Nat. Commun.* **7,** 12770 (2016).
16. Dong, Y. *et al.* Halomonas sulfidaeris-dominated microbial community inhabits a 1.8 km-

417 deep subsurface Cambrian Sandstone reservoir. *Environ. Microbiol.* **16,** 1695–708 (2014).
418 17. Purkamo, L., Bomberg, M., Kietäväinen, R., Salavirta, H. & Nyyssönen, M. Microbial co-
419 occurrence patterns in deep Precambrian bedrock fracture fluids. *Biogeosciences* 3091–3108
420 (2016). doi:10.5194/bg-13-3091-2016
421 18. Purkamo, L. *et al.* Diversity and functionality of archaeal, bacterial and fungal communities
422 in deep Archaean bedrock groundwater. *FEMS Microbiol. Ecol.* (2018).
423 doi:10.1093/femsec/fiy116
424 19. Baker, B. J. *et al.* Related assemblages of sulphate-reducing bacteria associated with
425 ultradeep gold mines of South Africa and deep basalt aquifers of Washington State. *Environ.*
426 *Microbiol.* **5,** 267–277 (2003).
427 20. Chivian, D. *et al.* Environmental Genomics Reveals a Single-Species Ecosystem Deep
428 Within Earth. *Science (80-.).* **322,** 275–278 (2008).
429 21. Ueno, A., Shimizu, S., Tamamura, S., Okuyama, H. & Naganuma, T. Anaerobic
430 decomposition of humic substances by Clostridium from the deep subsurface. *Nat. Publ. Gr.*
431 1–9 (2016). doi:10.1038/srep18990
432 22. Sousa, D. Z. *et al.* The deep-subsurface sulfate reducer Desulfotomaculum kuznetsovii
433 employs two methanol-degrading pathways. *Nat. Commun.* **9,** 239 (2018).
434 23. Brazelton, W. J., Morrill, P. L., Szponar, N. & Schrenk, M. O. Bacterial communities
435 associated with subsurface geochemical processes in continental serpentinite springs. *Appl.*
436 *Environ. Microbiol.* **79,** 3906–3916 (2013).
437 24. Vieira-Silva, S. *et al.* Species–function relationships shape ecological properties of the
438 human gut microbiome. *Nat. Microbiol.* **1,** 16088 (2016).
439 25. Nuppunen-Puputti, M. *et al.* Rare Biosphere Archaea Assimilate Acetate in Precambrian
440 Terrestrial Subsurface at 2.2 km Depth. *Geosciences* **8,** 418 (2018).
441 26. Hubalek, V. *et al.* Connectivity to the surface determines diversity patterns in subsurface
442 aquifers of the Fennoscandian shield. *ISME J.* **10,** 2447–2458 (2016).
443 27. Emerson, J. B., Thomas, B. C., Alvarez, W. & Banfield, J. F. Metagenomic analysis of a high
444 carbon dioxide subsurface microbial community populated by chemolithoautotrophs and
445 bacteria and archaea from candidate phyla. *Environ. Microbiol.* n/a-n/a (2015).
446 doi:10.1111/1462-2920.12817
447 28. Wouters, K., Moors, H., Boven, P. & Leys, N. Evidence and characteristics of a diverse and
448 metabolically active microbial community in deep subsurface clay borehole water. *FEMS*
449 *Microbiol. Ecol.* **86,** 458–473 (2013).
450 29. Ünal, B. *et al.* Trace elements affect methanogenic activity and diversity in enrichments from
451 subsurface coal bed produced water. *Front. Microbiol.* **3,** 1–14 (2012).

30. Hernsdorf, A. W. *et al.* Potential for microbial H 2 and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments. *ISME J.* **11,** 1915–1929 (2017).
31. Lopez-Fernandez, M. *et al.* Investigation of viable taxa in the deep terrestrial biosphere suggests high rates of nutrient recycling. *FEMS Microbiol. Ecol.* **94,** (2018).
32. Lloyd, K. G., Ladau, J., Steen, A. D., Yin, J. & Crosby, L. Phylogenetically novel uncultured microbial cells dominate Earth microbiomes. *bioRxiv* 303602 (2018). doi:10.1101/303602
33. Lollar, B. S. *et al.* Unravelling abiogenic and biogenic sources of methane in the Earth's deep subsurface. *Chem. Geol.* **226,** 328–339 (2006).
34. Lin, L. *et al.* Long-Term Sustainability of a High-Energy, Low-Diversity Crustal Biome. *Science (80-.).* **314,** (2006).
35. Li, L. *et al.* Sulfur mass-independent fractionation in subsurface fracture waters indicates a long-standing sulfur cycle in Precambrian rocks. *Nat. Commun.* **7,** 13252 (2016).
36. Lollar, G. S., Warr, O., Telling, J., Osburn, M. R. & Sherwood Lollar, B. "Follow the Water": Hydrogeochemical Constraints on Microbial Investigations 2.4 km below surface at the Kidd Creek Deep Fluid and Deep Life Observatory. *Geomicrobiol. J.* **In review,** (2019).
37. Griffin, W. T., Phelps, T. J., Colwell, F. S. & Fredrickson, J. K. in *Microbiology of the Terrestrial Deep Subsurface* 23–44 (CRC Press, 2018). doi:10.1201/9781351074568-3
38. Sheik, C. S. *et al.* Identification and removal of contaminant sequences from ribosomal gene databases: Lessons from the Census of Deep Life. *Front. Microbiol.* **9,** 840 (2018).
39. Wilkins, M. J. *et al.* Trends and future challenges in sampling the deep terrestrial biosphere. *Front. Microbiol.* **5,** 1–8 (2014).
40. Brockman, F. J., Murray, C. J. & Murray, C. J. in *The Microbiology of the Terrestrial Deep Subsurface* 75–102 (CRC Press, 2018). doi:10.1201/9781351074568-6
41. Russell, B. F., Phelps, T. J., Griffin, W. T. & Sargent, K. A. Procedures for Sampling Deep Subsurface Microbial Communities in Unconsolidated Sediments. *Groundw. Monit. Remediat.* **12,** 96–104 (1992).
42. Andreou, L. V. in *Methods in Enzymology* **529,** 143–151 (John Wiley & Sons, Inc., 2013).
43. Osburn, M. R., LaRowe, D. E., Momper, L. M. & Amend, J. P. Chemolithotrophy in the continental deep subsurface: Sanford Underground Research Facility (SURF), USA. *Front. Microbiol.* **5,** 1–14 (2014).
44. Barnhart, E. P. *et al.* Hydrogeochemistry and coal-associated bacterial populations from a methanogenic coal bed. *Int. J. Coal Geol.* **162,** 14–26 (2016).
45. Foght, J. *et al.* Culturable Bacteria in Subglacial Sediments and Ice from Two Southern Hemisphere Glaciers. *Microb. Ecol.* **47,** 329–40 (2004).

46. Tillett, D. & Neilan, B. A. Xanthogenate nucleic acid isolation from cultured and environmental cyanobacteria. *J. Phycol.* **36,** 251–258 (2000).
47. Leuko, S. *et al.* Lysis efficiency of standard DNA extraction methods for Halococcus spp. in an organic rich environment. *Extremophiles* **12,** 301–308 (2008).
48. Frank, Y. A. *et al.* Stable and variable parts of microbial community in Siberian deep subsurface thermal aquifer system revealed in a long-term monitoring study. *Front. Microbiol.* **7,** 1–15 (2016).
49. Lawson, C. E. *et al.* Patterns of endemism and habitat selection in coalbed microbial communities. *Appl. Environ. Microbiol.* AEM.01737-15 (2015). doi:10.1128/AEM.01737-15
50. Bomberg, M., Lamminmäki, T. & Itävaara, M. Microbial communities and their predicted metabolic characteristics in deep fracture groundwaters of the crystalline bedrock at Olkiluoto, Finland. *Biogeosciences* **13,** 6031–6047 (2016).
51. Rajala, P. & Bomberg, M. Reactivation of Deep Subsurface Microbial Community in Response to Methane or Methanol Amendment. *Front. Microbiol.* **08,** 431 (2017).
52. Blanco, Y. *et al.* Deciphering the Prokaryotic Community and Metabolisms in South African Deep-Mine Biofilms through Antibody Microarrays and Graph Theory. *PLoS One* **9,** e114180 (2014).
53. Singh, P. K., Singh, A. L., Kumar, A. & Singh, M. P. Control of different pyrite forms on desulfurization of coal with bacteria. *Fuel* **106,** 876–879 (2013).
54. Machnikowska, H., Pawelec, K. & Podgórska, A. Microbial degradation of low rank coals. *Fuel Process. Technol.* **77–78,** 17–23 (2002).
55. Lepleux, C., Turpault, M. P., Oger, P., Frey-Klett, P. & Uroz, S. Correlation of the abundance of betaproteobacteria on mineral surfaces with mineral weathering in forest soils. *Appl. Environ. Microbiol.* **78,** 7114–9 (2012).
56. Gründger, F. *et al.* Microbial methane formation in deep aquifers of a coal-bearing sedimentary basin, Germany. *Front. Microbiol.* **6,** (2015).
57. Yagi, J. M., Sims, D., Brettin, T., Bruce, D. & Madsen, E. L. The genome of Polaromonas naphthalenivorans strain CJ2, isolated from coal tar-contaminated sediment, reveals physiological and metabolic versatility and evolution through extensive horizontal gene transfer. *Environ. Microbiol.* **11,** 2253–2270 (2009).
58. Kostka, J. E. *et al.* Hydrocarbon-degrading bacteria and the bacterial community response in gulf of Mexico beach sands impacted by the deepwater horizon oil spill. *Appl. Environ. Microbiol.* **77,** 7962–74 (2011).
59. Posman, K. M., DeRito, C. M. & Madsen, E. L. Benzene Degradation by a Variovorax Species within a Coal Tar-Contaminated Groundwater Microbial Community. *Appl. Environ.*

Microbiol. **83,** (2017).

60. Dürre, P. in *Encyclopedia of Life Sciences* 1–11 (John Wiley & Sons, Ltd, 2015). doi:10.1002/9780470015902.a0020370.pub2

61. Jungbluth, S. P., Glavina del Rio, T., Tringe, S. G., Stepanauskas, R. & Rappé, M. S. Genomic comparisons of a bacterial lineage that inhabits both marine and terrestrial deep subsurface systems. *PeerJ* **5,** e3134 (2017).

62. Lollar, B. S., Onstott, T. C., Lacrampe-Couloume, G. & Ballentine, C. J. The contribution of the Precambrian continental lithosphere to global H2 production. *Nature* **516,** 379–382 (2014).

63. Aüllo, T., Ranchou-Peyruse, A., Ollivier, B. & Magot, M. Desulfotomaculum spp. and related gram-positive sulfate-reducing bacteria in deep subsurface environments. *Front. Microbiol.* **4,** 1–12 (2013).

64. Stegen, J. C. *et al.* Coupling among Microbial Communities , Biogeochemistry , and Mineralogy across Biogeochemical Facies. *Nat. Publ. Gr.* 1–14 (2016). doi:10.1038/srep30553

65. Schrenk, M. O., Brazelton, W. J. & Lang, S. Q. Serpentinization, Carbon, and Deep Life. *Rev. Mineral. Geochemistry* **75,** 575–606 (2013).

66. Fredrickson, J. K. & Balkwill, D. L. Geomicrobial Processes and Biodiversity in the Deep Terrestrial Subsurface. *Geomicrobiol. J.* **23,** 345–356 (2006).

67. Parnell, J. & McMahon, S. Physical and chemical controls on habitats for life in the deep subsurface beneath continents and ice. *Philos. Trans. R. Soc. A* **374,** 20140293 (2016).

68. Momper, L., Jungbluth, S. P., Lee, M. D. & Amend, J. P. Energy and carbon metabolisms in a deep terrestrial subsurface fluid microbial community. *ISME J.* **11,** 2319–2333 (2017).

69. Wu, X. *et al.* Microbial metagenomes from three aquifers in the Fennoscandian shield terrestrial deep biosphere reveal metabolic partitioning among populations. *ISME J.* **10,** 1192–1203 (2016).

70. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

71. Alhasawi, A., Costanzi, J., Auger, C., Appanna, N. D. & Appanna, V. D. Metabolic reconfigurations aimed at the detoxification of a multi-metal stress in Pseudomonas fluorescens: Implications for the bioremediation of metal pollutants. *J. Biotechnol.* **200,** 38–43 (2015).

72. Raiger Iustman, L. J. *et al.* Genome sequence analysis of Pseudomonas extremaustralis provides new insights into environmental adaptability and extreme conditions resistance. *Extremophiles* **19,** 207–220 (2015).

73. Uroz, S., Calvaruso, C., Turpault, M.-P. P. & Frey-Klett, P. Mineral weathering by bacteria: ecology, actors and mechanisms. *Trends Microbiol.* **17,** 378–387 (2009).

74. Rosenberg, E. in *The Prokaryotes: Alphaproteobacteria and Betaproteobacteria* 1–1012 (2013). doi:10.1007/978-3-642-30197-1

75. Onstott, T. C. *et al.* Indigenous and contaminant microbes in ultradeep mines. *Environ. Microbiol.* **5,** 1168–1191 (2003).

76. Davidson, M. M. *et al.* Capture of Planktonic Microbial Diversity in Fractures by Long-Term Monitoring of Flowing Boreholes, Evander Basin, South Africa. *Geomicrobiol. J.* **28,** 275–300 (2011).

77. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12,** 87 (2014).

78. Puente-Sánchez, F. *et al.* Viable cyanobacteria in the deep continental subsurface. *PNAS* (2018). doi:10.1073/pnas.1808176115

79. Di Rienzi, S. C. *et al.* The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *Elife* **2,** e01102 (2013).

80. Sogin, M. & Edwards, K. Deep Subsurface Microbiology and the Deep Carbon Observatory. in *DCO Deep Life Workshop* (2010).

81. Davis, N. M., Proctor, D., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *bioRxiv* 221499 (2018). doi:10.1101/221499

82. Direito, S. O. L., Marees, A. & Röling, W. F. M. Sensitive life detection strategies for low-biomass environments: Optimizing extraction of nucleic acids adsorbing to terrestrial and Mars analogue minerals. *FEMS Microbiol. Ecol.* **81,** 111–123 (2012).

83. Karst, S. M. *et al.* Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nat. Biotechnol.* **36,** 190–195 (2018).

84. Martijn, J. *et al.* Amplicon sequencing of the 16S-ITS-23S rRNA operon with long-read technology for improved phylogenetic classification of uncultured prokaryotes. *bioRxiv* 234690 (2017). doi:10.1101/234690

85. Fierer, N. & Cary, C. Don't let microbial samples perish. *Nature* **512,** 253–253 (2014).

86. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7,** 335–336 (2010).

87. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. **26,** 2460–2461 (2010).

88. McMurdie, P. J. & Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8,** e61217 (2013).

89. Wickham, H. & Chang, W. ggplot2: An Implementation of the Grammar of Graphics. (2015).

90. R Development Core Team. R: A language and environment for statistical computing. (2008).

91. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14,** 927–930 (2003).

92. Fuglede, B. & Topsoe, F. Jensen-Shannon divergence and Hilbert space embedding. in *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.* 30–30 (IEEE, 2004). doi:10.1109/ISIT.2004.1365067

Acknowledgements

The work was funded by a National Research Network for Low Carbon Energy and Environment (NRN-LCEE) grant to ACM and AE from the Welsh Government and the Higher Education Funding Council for Wales (Geo-Carb-Cymru). Deep borehole samples from Nevada and California, USA (e.g. Nevares Deep Well 2 and BLM-1) were obtained with help in the field from Alexandra Wheatley, Jim Bruckner, Jenny fisher and Scott Hamilton-Brehm, and technical assistance and funding from the US Department of Energy's Subsurface Biogeochemical Research Program, the Hydrodynamic Group, LLC, the Nye County Nuclear Waste Repository Program Office (NWRPO), the US National Park Service, and Inyo Country, CA. Samples from a mine in Northern Ontario Canada were obtained with funding from the Natural Sciences and Engineering Research Council of Canada and the assistance of Thomas Eckert, and Greg Slater of McMaster University. The Census of Deep Life (CoDL) and Deep Carbon Observatory (DCO) projects are acknowledged for a range of studies used in this analysis, as well as the sequencing team at the Marine Biological Laboratory (MBL). Disclaimer: Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Author Contributions

ARS developed the methodology, collated and analysed the data, and wrote the manuscript. AE and AM conceived the study, supervised AS and helped write the manuscript. Other authors provided data from field sites used in the global meta-analysis. All authors contributed, edited and approved the final manuscript.

Conflict of Interest

We declare no conflict of interest.

Supplementary Information – "A global perspective on microbial diversity in the terrestrial deep subsurface"

Authors:

Soares, A.[1,2,3], Edwards, A.[2,3*], An, D.[4], Bagnoud, A.[5], Barnhart, E[6,7], Bomberg, M.[8], Budwill, K.[9], Caffrey, S.[10], Fields, M.[7], Gralnick., J.[11], Kadnikov, V.[12,13], Momper, L.[14], Osburn, M.[15], Mu, A.[16,17,18,19], Moreau, J.W.[16], Moser, D.[20], Purkamo, L.[8,21,22], Rassner, S. M.[1,3], Sheik, C. S.[23], Sherwood Lollar, B. [24], Toner, B. M.[25], Voordouw, G.[4], Wouters, K.[26], Mitchell, A. C.[1,3*]

Supplementary Text

Depth weakly controls microbial community structure

Life at extreme depths is yet to be analysed at a deep genomic level, but microbial cells have been discovered at depths down to 3,6 km in the terrestrial crust [1,2]. Although cell numbers tend to decrease with depth in both crystalline and sedimentary rock in the continental crust, not much is known regarding large-scale taxonomic trends [2]. No significant correlations were found for the presence of the most abundant clades in the dataset and depth, being Actinobacteria the only major taxonomic group to have a positive, albeit weak, correlation to depth (Pearson's r = 0.42, p<0.01, **Figure S2**). Actinobacteria have already been detected at great depths in both the continental and oceanic crusts and some of its members have further been reported to hold ancestral genes for pyruvate oxidoreductase activity, which could potentially propel microbial growth at higher temperatures [3,4]. Proportions of Beta- and Gammaproteobacteria decreased with depth (Pearson's r = -0.29, Pearson's r = -0.093), and no other major clades were shown to correlate. More data needs to be generated to better investigate which, if any, taxonomic groups prefer deeper terrestrial environments. Biochemical limitations to life such as racemization rates of organic acids with depth and temperature will surely select for adaptable taxa and possibly create a mostly depth-defined gradient of representation of adapted extremophilic clades [5].

99% closed-reference OTU analysis

A 99% similarity closed-reference OTU-picking strategy to further minimize potential contamination showed a ~10-fold decrease in the number of OTUs and read numbers (1,065 OTUs and 70,527 reads were left). Further, this reduced the number of retrieved samples to 93.

Following the previously described filtering steps (hold only OTUs represented by more than 500 sequences and present in at least 10 samples), 335 OTUs (67,151 reads) associated to 14 samples were left. For this reduced dataset, 2 OTUs associated to Chloroplast were found, represented by 115 sequences in total. The very reduced number of samples, OTUs an total reads left in the datasset caused the 99% OTU approach to be discontinued.

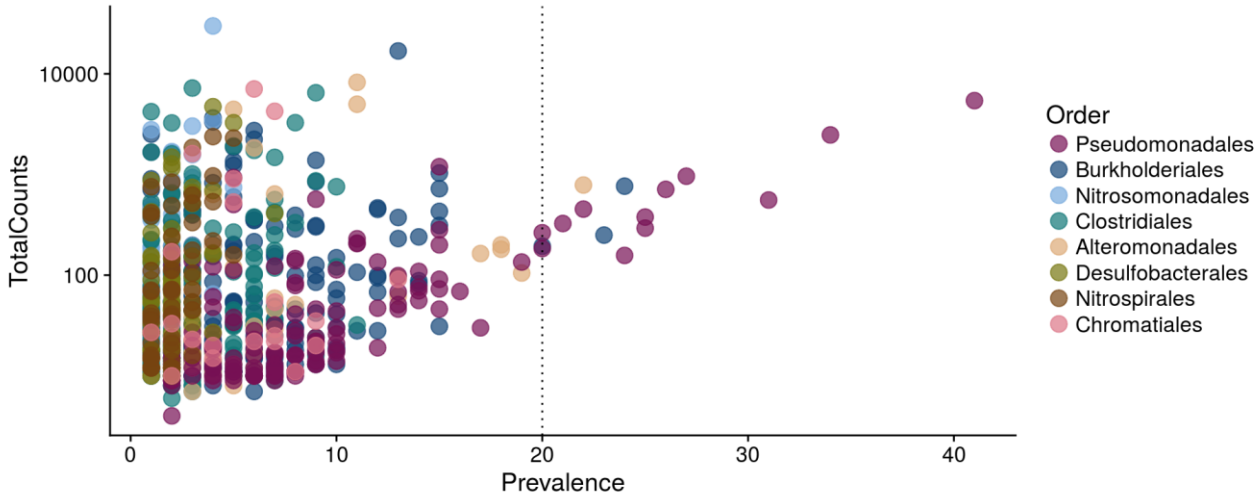
Supplementary Methodology

Phylogeny of Pseudomonas *representative sequences*

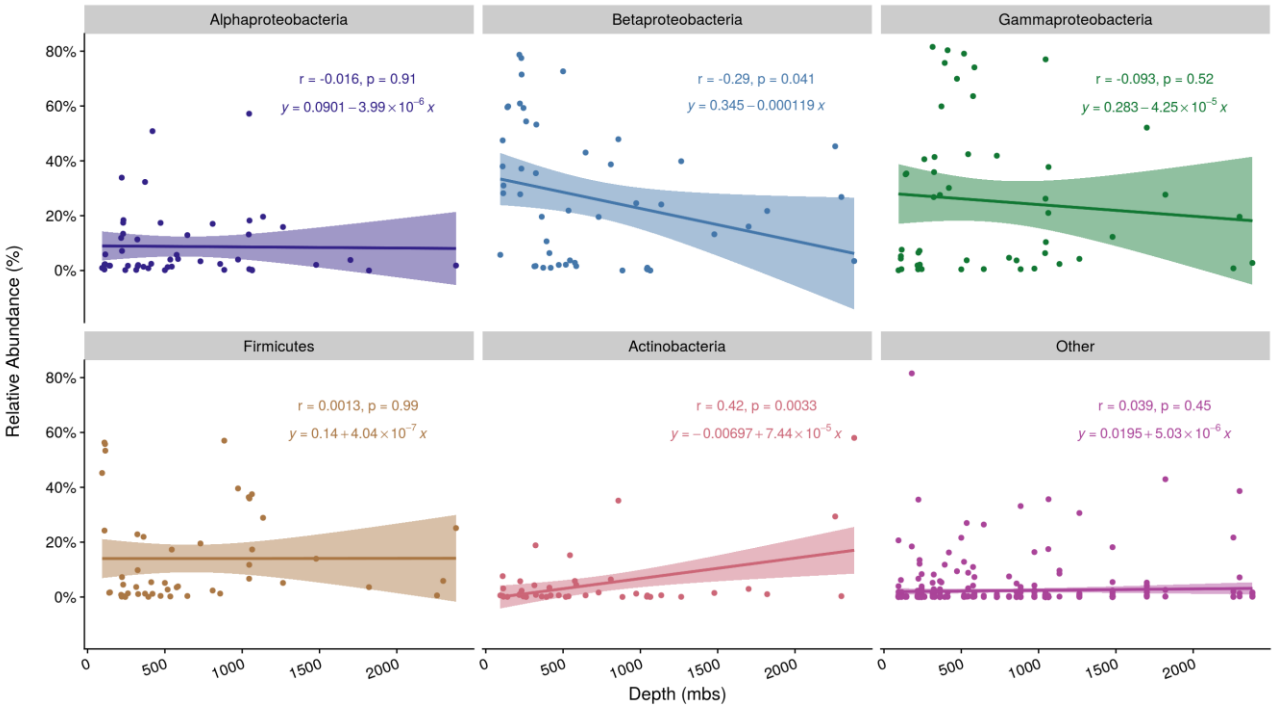
Representative 16S rRNA gene sequences for *Pseudomonas* OTUs in the dataset were isolated by retrieving OTU IDs affiliated to this genus in the final dataset using the *subset_taxa* function within *phyloseq*[6]. The SILVA 123 database was then queried against the list of OTU IDs and the results deposited in a FASTA file. Outgroup sequences of genus *Sphingomonas* (Proteobacteria, Alphaproteobacteria, Sphingomonadales, Sphingomonadaceae) were obtained directly from the SILVA database (https://www.arb-silva.de/) and added along with the retrieved *Pseudomonas* 16S rRNA gene sequences to a final FASTA file.

Using MEGA7[7], an alignment was performed using the MUSCLE[8] algorithm and 8 iterations of the UPGMB (combines Neighbour-Joining[9] and UPGMA - Unweighted Pair Group Method with Arithmetic mean) clustering method. An optimal Neighbour-Joining[9] tree (*cf.* **Supplementary Figure 4**) was then created using 500 bootstrapped replicates and the Maximum Composite Likelihood (MCL)[10] method to calculate evolutionary distances.

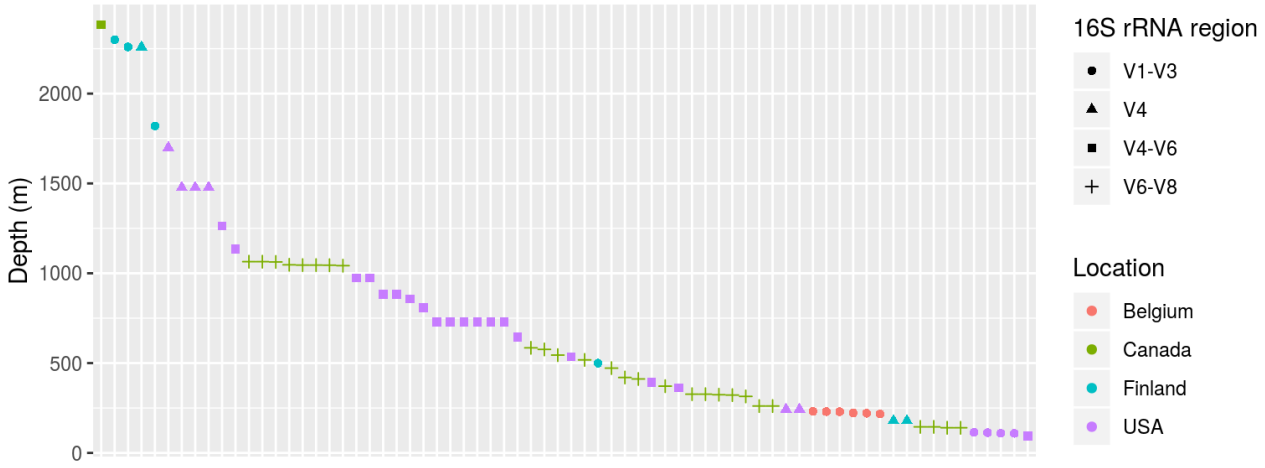
Supplementary Figures



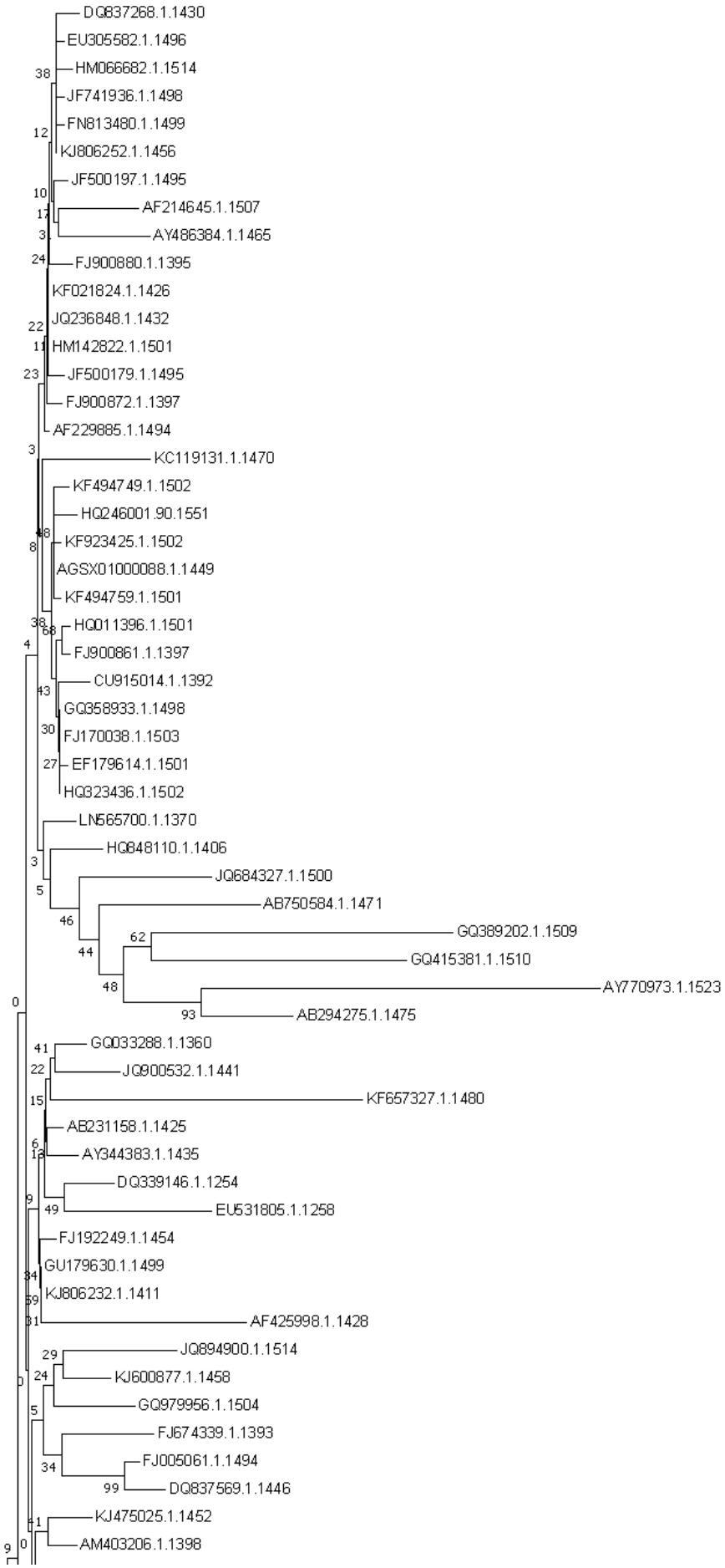
Supplementary Figure 1 - Prevalence (number of samples an OTU is present in, x-axis) of OTUs across the dataset and associated reads (y-axis). Colours depict classification of OTUs at order level. Vertical line crosses 20 samples in the x-axis to highlight OTUs present in 20 or more samples.

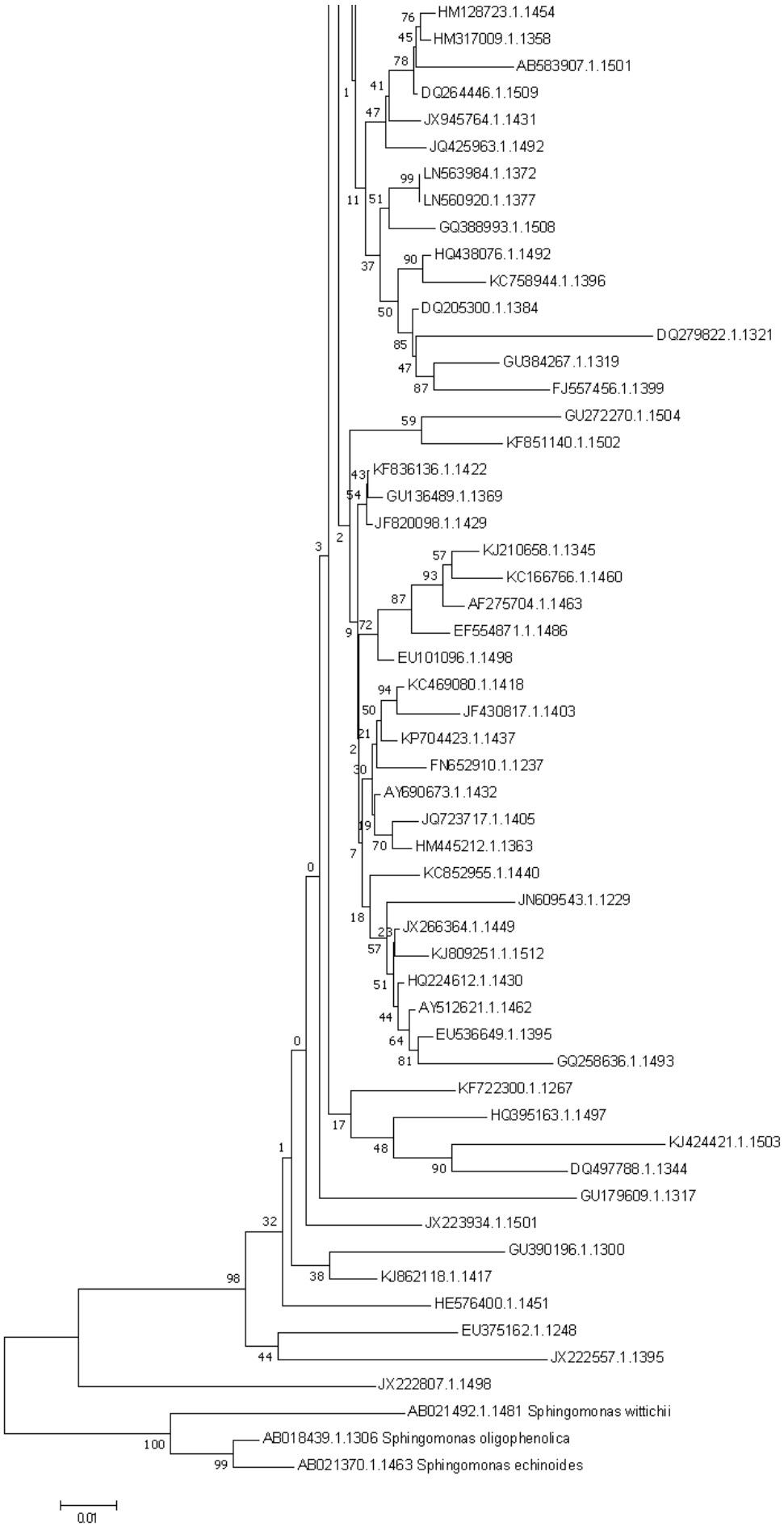


Supplementary Figure 2 - Correlations between relative abundance of OTUs (%, y-axis) associated to the most abundant taxonomic groups across the dataset and depth (meters below surface, x-axis). Regression lines follow the linear model and shading around lines corresponds to the 95% confidence interval. Annotations in plot facets indicate the associated Pearson correlation coefficient, its corresponding p-value and the fitted linear model equation. Each point represents an OTU associated to the taxonomic group in each facet at a certain depth - a same OTU may be depicted more than one time.



Supplementary Figure 3 – Distribution of samples in the final dataset across depth, colourised by general location. Shapes indicate 16S rRNA gene region utilised for that study.





Supplementary Figure 4 – Optimal Neighbor-Joining tree of representative sequences for *Pseudomonas* OTUs in the dataset following MUSCLE aligment. OTU IDs are shown in the end of each branch and outgroup sequences of genus *Sphingomonas* identified by their taxonomic affiliation. The percentage of bootstrap test replicates are shown next to tree branches and scale for MCL evolutionary distances is shown in the bottom.

Supplementary Tables

Supplementary Table 1 - Top 10 OTU network interactions ordered by edge betweenness (number of shortest paths going through an edge - OTU/OTU interactions) values as per the calculated Jaccard distances.

OTU 1 Classification	OTU 2 Classification	Betweenness
Gammaproteobacteria, Chromatiales, Chromatiaceae, Rheinheimera	Gammaproteobacteria, Chromatiales, Chromatiaceae, Rheinheimera	1686
Gammaproteobacteria, Chromatiales, Chromatiaceae, Rheinheimera	Gammaproteobacteria, Pseudomonadales, Moraxellaceae, Acinetobacter	1656.17
Gammaproteobacteria, Chromatiales, Chromatiaceae, Rheinheimera	Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Alishewanella	885.1
Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	Gammaproteobacteria, Chromatiales, Chromatiaceae, Rheinheimera	749
Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	Gammaproteobacteria, Chromatiales, Chromatiaceae, Rheinheimera	618
Betaproteobacteria., Rhodocyclales, Rhodocyclaceae, Azoarcus	Betaproteobacteria, Rhodocyclales, Rhodocyclaceae, uncultured	588
Firmicutes, Clostridia, Clostridiales, Peptostreptococcaceae, Peptoclostridium	Firmicutes, Clostridia, Clostridiales, Peptostreptococcaceae, Intestinibacter	588
Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	Alphaproteobacteria, Rhodobacterales, Rhodobacteraceae, Paracocccus	563.83
Gammaproteobacteria, Chromatiales, Chromatiaceae, Rheinheimera	Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Alishewanella	499.26
Gammaproteobacteria, Pseudomonadales, Moraxellaceae, Acinetobacter	Gammaproteobacteria, Pseudomonadales, Moraxellaceae, Acinetobacter	495

Supplementary Table 2 - SILVA 123 taxonomic affiliations of OTUs present in 20 or more samples. Prevalence is defined as the number of samples an OTU is present in.

Taxa ID	Taxa Classification	Prevalence
JQ236848.1.1432	Proteobacteria, Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	41
HM142822.1.1501	Proteobacteria, Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	34
JX266364.1.1449	Proteobacteria, Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	31
KJ475025.1.1452	Proteobacteria, Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	27
FJ192249.1.1454	Proteobacteria, Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	26
HQ848110.1.1406	Proteobacteria, Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	25
KC852955.1.1440	Proteobacteria, Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	25
HM773515.1.1498	Proteobacteria, Betaproteobacteria, Burkholderiales, Comamonadaceae, uncultured	24
AB231158.1.1425	Proteobacteria, Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	24
JX222276.1.1475	Proteobacteria, Betaproteobacteria, Burkholderiales, Comamonadaceae, Variovorax	23
EU841498.1.1443	Proteobacteria; Gammaproteobacteria; Alteromonadales, Alteromonadaceae, Alishewanella	22
EU305582.1.1496	Proteobacteria, Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	22
FJ900880.1.1395	Proteobacteria, Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Pseudomonas	21

Supplementary Table 3 – Metadata table with complete details for all studies utilized.

SRA Accession	Depth gradient (m)	DNA Extraction method	Sample Type	16S region	Sequencing method	Project
PRJNA262938	243.84-1478.28	Mod. Phenol chloroform	Groundwater	V4	454	
PRJNA268940	94-2383	MoBio UltraClean Microbial DNA isolation kit	Groundwater	V4-V6	454	VAMPS
PRJNA248749	730	MoBio PowerWater	Groundwater	V4-V6	454	VAMPS
PRJNA251746	393-1135	MoBio_PowerMax_Soil	Groundwater	V4-V6	454	VAMPS
PRJNA375701	109-114.7	Mod. MP-Biomedicals FastDNA Spin_Soil [11]	Rock + porewater	V1-V3	454	
NA*	140-1064.4	Foght *et al.* 2004	Rock + porewater	V6-V8	454	HMP
PRJEB1468	217.13-231.83	XS buffer [13]	Groundwater	V1-V3	454	
PRJEB10822	180-2300	MoBio PowerSoil	Groundwater	V4**, V1-V3	454	

* http://hmp.ucalgary.ca/HMP/

* This primer pair corresponds to archaeal 16S rRNA gene primers (A344,A744, *cf.* 10.5194/bg-13-3091-2016 for further details).

Supplementary Table 4 – BLAST hits table for representative sequences associated to OTUs affiliated to genus *Pseudomonas* .

OTU ID	Closest BLAST hit	Associated NCBI Title	Associated Environment	Taxonomy ID	NCBI Accession	Percent similarity (%)	E-value	No. Samples Associated	No. Reads Associated	Observations
AB231158.1.1425	Pseudomonas pseudoalcaligenes gene for 16S rRNA, partial sequence	High nitrogen removal from wastewater with several new aerobic bacteria isolated from diverse ecosystems	Wastewater	330	gi|85539942|dbj|AB231158.1|	100	0	24	157	
AB294275.1.1475	Uncultured bacterium gene for 16S rRNA, partial sequence, clone: YWB06	Molecular characterization of microbial communities in deep coal seam groundwater of northern Japan	Subsurface	77133	gi|126143387|dbj|AB294275.1|	100	0	19	135	
AB583907.1.1501	Uncultured bacterium gene for 16S rRNA, clone: RBC-4B	Microbial community in water-rock-microbe interaction systems in subsurface environments	Subsurface	77133	gi|324604950|dbj|AB583907.1|	100	0	3	137	
AB750584.1.1471	Uncultured bacterium gene for 16S rRNA, partial sequence, clone: 23 hydrate.Cas.16S	Novel integrons and gene cassettes from a Cascadian submarine gas hydrate-bearing core	Subsurface	77133	gi|407969505|dbj|AB750584.1|	100	0	16	69	
AF214645.1.1507	Nitrogen-fixing bacterium MIS 16S ribosomal RNA gene, partial sequence	Molecular characterization of plant associated nitrogen-fixing Bacteria	Subsurface	120486	gi|7542430|gb|AF214645.1|AF214645	100	0	4	20	
AF229885.1.1494	Pseudomonas sp. 3CB6 16S ribosomal RNA gene, partial sequence	Isolation and characterization of diverse halobenzoate-degrading denitrifying bacteria from soils and sediments	Subsurface	126132	gi|9965646|gb|AF229885.1|	100	0	9	22	
AF275704.1.1463	Unidentified Hailaer soda lake bacterium F5 16S ribosomal RNA gene, partial sequence	China: Inner Mongolia	Lake	148462	gi|12275954|gb|AF275704.1|AF275704	100	0	12	47	
AF425998.1.1428	Bacterium UNSW3 16S ribosomal RNA gene, partial sequence	A survey of phylogeny and paralytic shellfish poison production from culturable bacteria associated with a toxic Anabaena circinalis strain	Host-associated	190587	gi|23451685|gb|AF425998.1|	100	0	14	56	
AGSX01000088.1.1449	Pseudomonas stutzeri SDM-LAC contig000089, whole genome shotgun sequence	Genome Sequence of Pseudomonas stutzeri SDM-LAC, a Typical Strain for Studying the Molecular Mechanism of Lactate Utilization	Lab strain	271420	gi|1093258265|emb|LT629970.1|	99.83	0	6	28	
AM403206.1.1398	Pseudomonadaceae bacterium D7-21 partial 16S rRNA gene, isolate D7-21	Diversity of Nitrate-reducing and Denitrifying Bacteria in a Marine Aquaculture Biofilter	Seawater	404904	gi|115334066|emb|AM403206.1|	100	0	8	16	
AY344383.1.1435	Unidentified bacterium clone K2-4-3 16S ribosomal RNA gene, partial sequence	Microbial Communities in the Hawaiian Archipelago: A Microbial Diversity Hotspot	Subsurface	1869227	gi|33391921|gb|AY344383.1|	100	0	6	10	
AY486384.1.1465	Pseudomonas stutzeri strain AU4823 16S ribosomal RNA gene, partial sequence	PCR-based assay for differentiation of Pseudomonas aeruginosa from other Pseudomonas species recovered from cystic fibrosis patients	Host-associated	316	gi|40019083|gb|AY486384.1|	100	0	7	17	
AY512621.1.1462	Pseudomonas veronii strain A1YdBTEX2-5 16S ribosomal RNA gene, partial sequence	Alternative primer sets for PCR detection of genotypes involved in bacterial aerobic BTEX degradation: distribution of the genes in BTEX degrading isolates and in subsurface soils of a BTEX contaminated industrial site	Subsurface	76761	gi|41056888|gb|AY512621.1|	100	0	15	91	
AY690673.1.1432	Sulfitobacter sp. GC07 16S ribosomal RNA gene, partial sequence	The diversity of halotolerant heterotrophic bacteria isolated from rhizosphere soil of salt marshes from southwestern coasts in Korea	Subsurface	290364	gi|51243779|gb|AY690673.1|	100	0	1	33	Latest SILVA database classifies this sequence as genus *Pseudomonas*
AY770973.1.1523	Uncultured bacterium clone W33 16S ribosomal RNA gene, partial sequence	Molecular analysis of the microbial communities of the oilfield	Subsurface	77133	gi|54695050|gb|AY770973.1|	100	0	10	14	
CU915014.1.1392	Pseudomonas stutzeri partial 16S rRNA gene, strain	marine sediment and soils	Subsurface	77133	gi|239913533|emb|CU915014.1|	100	0	2	13	
DQ205300.1.1384	Pseudomonas sp. HI-G1 16S ribosomal RNA gene, partial sequence	Culturable microbial diversity and the impact of tourism in Kartchner Caverns, Arizona	Subsurface	347772	gi|76365589|gb|DQ205300.1|	100	0	2	15	
DQ264446.1.1509	Uncultured bacterium clone BANW452 16S ribosomal RNA gene, partial sequence	High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment	Subsurface	77133	gi|82393910|gb|DQ264446.1|	100	0	5	13	
DQ279822.1.1321	Pseudomonas sp. D14 16S ribosomal RNA gene, partial	Direct Submission, strain D14	Lab strain	358759	gi|82697883|gb|DQ279822.1|	100	0	6	10	
DQ339146.1.1254	Pseudomonas sp. M9J918 16S ribosomal RNA gene, partial sequence	Isolation of glyphosate-resistant bacterial species M9J918 from glyphosate-polluted wheat soil in China	Subsurface	366287	gi|85001608|gb|DQ339146.1|	100	0	9	13	
DQ497788.1.1344	Pseudomonas sp. SGb396 16S ribosomal RNA gene,	Soil, subusrface, bacterial endophytes from Taxus globosa	Subsurface	77133	gi|98975506|gb|DQ497788.1|	100	0	10	44	
DQ837268.1.1430	Pseudomonas stutzeri strain WWvii23 16S ribosomal	wastewater in lagos and ogun states Nigeria	Wastewater	86473	gi|112434217|gb|DQ837268.1|	100	0	7	16	
DQ837569.1.1446	Pseudomonas sp. K4 16S ribosomal RNA gene, partial	Direct Submission, isolate K4	Lab strain	394945	gi|110704228|gb|DQ837569.1|	100	0	9	18	
EF179614.1.1501	Pseudomonas sp. WP02-4-9 16S ribosomal RNA gene,	Isolation from deep-sea sediments	Subsurface	444156	gi|148251203|gb|EF179614.1|	100	0	2	4	
EF554871.1.1486	Pseudomonas sp. AB42 16S ribosomal RNA gene, partial sequence	Adaptative Potential of Alkaliphilic Bacteria towards Chloroaromatic Substrates Assessed by a gfp-Tagged 2,4-D Degradation Plasmid	Lake	443000;443001;443002;1740285	gi|146747229|gb|EF554868.1|	100	0	9	46	
EU101096.1.1498	Uncultured bacterium clone FS0612_B12 16S ribosomal RNA gene, partial sequence	Niche differentiation among sulfur-oxidizing bacterial populations in cave waters	Subsurface	77133	gi|156573149|gb|EU101096.1|	100	0	5	12	
EU305582.1.1496	Uncultured Pseudomonas sp. clone 3-A 16S ribosomal RNA gene, partial sequence	Nitrite removal performance and community structure of nitrite-oxidizing and heterotrophic bacteria suffered with organic Matter	Microcosm	114707	gi|163676410|gb|EU305582.1|	100	0	22	453	
EU375162.1.1248	Uncultured Pseudomonas sp. clone Sc13 16S ribosomal RNA gene, partial sequence	Bacterial communities from shoreline environments (costa da morte, northwestern Spain) affected by the prestige oil spill	Coastal water	114707	gi|166407785|gb|EU375162.1|	100	0	2	17	
EU531805.1.1258	Pseudomonas pseudoalcaligenes 16S ribosomal RNA gene, partial sequence	Bacterial strains isolated from harvested tiger shrimp (Peneaus Monodon)	Host-associated	330	gi|170962995|gb|EU531805.1|	100	0	7	12	
EU536649.1.1395	Uncultured bacterium clone nbt64e03 16S ribosomal	A diversity profile of the human skin microbiota	Host-associated	77133	gi|187964754|gb|EU536649.1|	100	0	5	503	
FJ005061.1.1494	Pseudomonas sp. enrichment culture clone Guo7 16S ribosomal RNA gene, partial sequence	The bioleaching feasibility for Pb/Zn smelting slag and community characteristics of indigenous moderate-thermophilic bacteria	Subsurface	557865	gi|204342383|gb|FJ005061.1|	100	0	7	17	
FJ170038.1.1503	Pseudomonas sp. CF14-10 16S ribosomal RNA gene,	sediments of the South China Sea	Subsurface	562724	gi|206585088|gb|FJ170038.1|	100	0	9	568	
FJ192249.1.1454	Uncultured Pseudomonas sp. clone GI3-S-5-G03 16S ribosomal RNA gene, partial sequence	Comprehensive census of bacteria in clean rooms by using DNA microarray and cloning methods	Urban	114707	gi|209421869|gb|FJ192249.1|	100	0	26	712	

Supplementary Table 4 – BLAST hits table for representative sequences associated to OTUs affiliated to genus *Pseudomonas* .

FJ557456.1.1399	Uncultured bacterium clone ET_H_1d10 16S ribosomal RNA gene, partial sequence	Endotracheal tube biofilm inoculation of oral flora and subsequent colonization of opportunistic pathogens	Host-associated	77133	gi|224569195|gb|FJ557456.1|	100	0	6	11	
FJ674339.1.1393	Uncultured bacterium clone LL141-1D15 16S ribosomal RNA gene, partial sequence	Synecology of the primary and secondary feedlot habitats of Escherichia coli O157:H7	Biological sample	77133	gi|223678636|gb|FJ674339.1|	100	0	7	10	
FJ900861.1.1397	Uncultured bacterium clone C-44 16S ribosomal RNA gene, partial sequence	Comparison of microbial community compositions of injection and production well samples in a long-term water-flooded petroleum Reservoir	Subsurface	77133	gi|229428754|gb|FJ900861.1|	100	0	7	9	
FJ900872.1.1397	Uncultured bacterium clone C-55 16S ribosomal RNA gene, partial sequence	Comparison of microbial community compositions of injection and production well samples in a long-term water-flooded petroleum Reservoir	Subsurface	77133	gi|229428765|gb|FJ900872.1|	100	0	9	14	
FJ900880.1.1395	Uncultured bacterium clone C-62 16S ribosomal RNA gene, partial sequence	Comparison of microbial community compositions of injection and production well samples in a long-term water-flooded petroleum Reservoir	Subsurface	77133	gi|229428773|gb|FJ900880.1|	100	0	21	326	
FN652910.1.1237	Pseudomonas putida partial 16S rRNA gene, isolate P5	Isolation, characterization and screening for antimicrobial activities of Padina pavonica associated bacteria	Host-associated	303	gi|290490824|emb|FN652910.1|	100	0	3	45	
FN813480.1.1499	Pseudomonas stutzeri partial 16S rRNA gene, isolate Gr20	The genetic diversity of culturable nitrogen-fixing bacteria in the rhizosphere of wheat	Subsurface	316	gi|295810396|emb|FN813480.1|	100	0	17	30	
GQ033288.1.1360	Uncultured bacterium clone nbw1028h09c1 16S ribosomal RNA gene, partial sequence	Topographical and temporal diversity of the human skin microbiome	Host-associated	77133	gi|238303670|gb|GQ033288.1|	100	0	4	10	
GQ258636.1.1493	Pseudomonas sp. SR3(2009) 16S ribosomal RNA gene,	Direct Submission, strain SR3	Lab strain	659426	gi|256352433|gb|GQ258636.1|	100	0	5	933	
GQ358933.1.1498	Pseudomonas sp. BSw21506 16S ribosomal RNA gene,	Direct Submission, strain BSw21506	Lab strain	664427	gi|255689465|gb|GQ358933.1|	100	0	12	19	
GQ388993.1.1508	Uncultured bacterium clone J41 16S ribosomal RNA gene, partial sequence	Characterization of Bacterial Community Structure in a Drinking Water Distribution System during an Occurrence of Red Water	Subsurface	77133	gi|256593919|gb|GQ388993.1|	100	0	5	11	
GQ389202.1.1509	Uncultured bacterium clone D75 16S ribosomal RNA gene, partial sequence	Characterization of Bacterial Community Structure in a Drinking Water Distribution System during an Occurrence of Red Water	Subsurface	77133	gi|256594128|gb|GQ389202.1|	100	0	8	16	
GQ415381.1.1510	uncultured bacterium clone DQB-P183 genomic sequence	Microbial diversity in different production waters of Daqing Oilfield	Subsurface	77133	gi|308097127|gb|GQ415381.1|	100	0	10	21	
GQ979956.1.1504	Uncultured bacterium clone SRC_NRB027 16S ribosomal RNA gene, partial sequence	Direct Submission, clone SRC_NRB027	Lab strain	77133	gi|261499583|gb|GQ979956.1|	100	0	15	283	
GU136489.1.1369	Pseudomonas sp. DY1 16S ribosomal RNA gene, partial	Direct Submission, strain DY1	Lab strain	690347	gi|265263160|gb|GU136489.1|	100	0	7	13	
GU179609.1.1317	Uncultured Pseudomonas sp. clone D006011A15 16S ribosomal RNA gene, partial sequence	Microbial diversity profiles of fluids from low-temperature petroleum reservoirs with and without exogenous water perturbation	Subsurface	114707	gi|306491368|gb|GU179609.1|	100	0	2	159	
GU179630.1.1499	Uncultured Pseudomonas sp. clone D118231H09 16S ribosomal RNA gene, partial sequence	Microbial diversity profiles of fluids from low-temperature petroleum reservoirs with and without exogenous water perturbation	Subsurface	114707	gi|306491389|gb|GU179630.1|	100	0	9	17	
GU272270.1.1504	Pseudomonas amygdali pv. lachrymans str. M301315 chromosome, complete genome	Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 Pseudomonas syringae isolates	Subsurface	77133	gi|284025723|gb|GU272270.1|	100	0	3	648	
GU384267.1.1319	Pseudomonas aeruginosa strain SZH16 16S ribosomal RNA gene, partial sequence	In situ degradation of phenol and promotion of plant growth in contaminated environments by a single Pseudomonas aeruginosa strain	Subsurface	287	gi|288189621|gb|GU384267.1|	100	0	4	13	
GU390196.1.1300	Uncultured bacterium clone SLE33F 16S ribosomal RNA gene, partial sequence	Shifts in microbial community structure of granular and liquid biomass in response to changes to infeed and digester design in anaerobic digesters receiving food-processing wastes	Subsurface	77133	gi|312178659|gb|GU390196.1|	100	0	14	70	
HE576400.1.1451	uncultured bacterium partial 16S rRNA gene, clone K16.133 AW	A combined cultivation and cultivation-independent approach shows high bacterial diversity in water-miscible metalworking fluids	Urban	77133	gi|377549755|emb|HE576400.1|	100	0	4	10	
HM066682.1.1514	Uncultured bacterium clone EDW07B006_73 16S ribosomal RNA gene, partial sequence	Microbial diversity and impact on carbonate geochemistry across a changing geochemical gradient in a karst aquifer	Subsurface	77133	gi|313770588|gb|HM066682.1|	100	0	8	16	
HM128723.1.1454	Uncultured bacterium clone SINN704 16S ribosomal RNA gene, partial sequence	Diversity of bacterioplankton in contrasting Tibetan lakes revealed by high-density microarray and clone library analysis	Lake	77133	gi|295879789|gb|HM128723.1|	100	0	20	184	
HM142822.1.1501	Proteobacterium WJQ No.5 16S ribosomal RNA gene,	Direct Submission, strain WJQ No. 5	Lab strain	797062	gi|298256144|gb|HM142822.1|	100	0	34	2472	
HM317009.1.1358	Uncultured bacterium clone ncd315a03c1 16S ribosomal RNA gene, partial sequence	Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis	Host-associated	77133	gi|297010604|gb|HM317009.1|	100	0	4	12	
HM445212.1.1363	Uncultured bacterium clone BL1289fO5 16S ribosomal RNA gene, partial sequence	Comparison of Bacterial Diversity in Azorean and Hawai'ian Lava Cave Microbial Mats	Subsurface	77133	gi|302398051|gb|HM445212.1|	100	0	7	32	
HQ011396.1.1501	Pseudomonas stutzeri strain NCTC10475 genome	NCTC10475	Lab strain	77133	gi|307776705|gb|HQ011396.1|	100	0	8	139	
HQ224612.1.1430	Pseudomonas sp. SGb387 16S ribosomal RNA gene, partial sequence	Phylogenetic relationships of bacterial endophytes from Taxus Globosa	Host-associated	1248114	gi|411172546|gb|JX897952.1|	99.15	0	15	73	
HQ246001.90.1551	UNVERIFIED: Uncultured Pseudomonas sp. 16S ribosomal RNA-like sequence	Proteobacteria dominance in the estuarine belt of river Narmada, India as depicted by molecular phylogenetic approaches	Freshwater	471	gi|226430883|gb|FJ816052.1|	88.52	0	4	19	
HQ323436.1.1502	Pseudomonas stutzeri strain M4 16S ribosomal RNA gene, partial sequence	Diversity of airborne bacteria community in the Mogao Grottoes, Dunhuang, China	Air	316	gi|319658795|gb|HQ323436.1|	100	0	6	10	

Supplementary Table 4 – BLAST hits table for representative sequences associated to OTUs affiliated to genus *Pseudomonas* .

HQ395163.1.1497	Uncultured bacterium clone OTUb51 16S ribosomal RNA gene, partial sequence	Culture-independent and culture-dependent methods reveal diverse bacterial and archaeal communities in a biodegraded Malaysian oil Reservoir	Subsurface	77133	gi|320172034|gb|HQ395163.1|	100	0	4	18	
HQ438076.1.1492	Pseudomonas sp. TeU 16S ribosomal RNA gene, partial sequence	Isolation and characterization of an environmental cadmium- and tellurite-resistant Pseudomonas strain	Subsurface	944290	gi|318063761|gb|HQ438076.1|	100	0	8	11	
HQ848110.1.1406	Pseudomonas argentinensis strain PL-40-1 16S ribosomal RNA gene, partial sequence	Phylogenetic diversity of culturable bacteria isolated from the disused ancient Kiyik River	Freshwater	289370	gi|340025389|gb|HQ848110.1|	100	0	25	378	
JF430817.1.1403	Pseudomonas sp. P56(2011) 16S ribosomal RNA gene, partial sequence	Genetic and functional diversity of fluorescent Pseudomonas from rhizospheric soils of wheat crop	Subsurface	1079824	gi|345132366|gb|JF430817.1|	100	0	14	76	
JF500179.1.1495	Uncultured bacterium clone 1572_50bact 16S ribosomal RNA gene, partial sequence	The deep biosphere in terrestrial sediments in the chesapeake bay area, virginia, USA	Subsurface	77133	gi|343170179|gb|JF500179.1|	100	0	11	230	
JF500197.1.1495	Uncultured bacterium clone 1563_48bact 16S ribosomal RNA gene, partial sequence	The deep biosphere in terrestrial sediments in the chesapeake bay area, virginia, USA	Subsurface	77133	gi|343170197|gb|JF500197.1|	100	0	10	18	
JF741936.1.1498	Uncultured bacterium clone LHJB-8 16S ribosomal RNA gene, partial sequence	Microbial communities in oil reservoirs with different salinities	Subsurface	77133	gi|332712701|gb|JF741936.1|	100	0	3	10	
JF820098.1.1429	Pseudomonas sp. PG-3-1 16S ribosomal RNA gene, partial sequence	Diversity of Culturable Butane-oxidizing Bacteria in Oil and Gas Field Soil	Subsurface	1036156	gi|334690415|gb|JF820098.1|	100	0	8	11	
JN609543.1.1229	Pseudomonas fluorescens strain Cantas14 16S ribosomal RNA gene, partial sequence	The culturable intestinal microbiota of triploid and diploid juvenile Atlantic salmon (Salmo salar) - a comparison of composition and drug resistance	Host-associated	294	gi|358440878|gb|JN609543.1|	100	0	5	38	
JQ236848.1.1432	Pseudomonas stutzeri strain hswx82 16S ribosomal RNA	Direct Submission, strain hswx82	Lab strain	316	gi|377830169|gb|JQ236848.1|	100	0	41	5419	
JQ425963.1.1492	Uncultured bacterium clone CT0C1CA03 16S ribosomal RNA gene, partial sequence	Bacterial diversity in an alkaline saline soil spiked with Anthracene	Lake	77133	gi|385760667|gb|JQ425953.1|	100	0	9	22	
JQ684327.1.1500	Uncultured bacterium clone HWGB-69 16S ribosomal RNA gene, partial sequence	Bacterial and archaeal diversity in permafrost soil from Kunlun Mountains Pass, Tibet Plateau of China	Subsurface	77133	gi|387568220|gb|JQ684327.1|	100	0	3	12	
JQ723717.1.1405	Pseudomonas sp. B-5-1 16S ribosomal RNA gene, partial	beach sand in the west sea, Korea	Coastal water	1172620	gi|384392471|gb|JQ723717.1|	100	0	13	68	
JQ894900.1.1514	Uncultured bacterium clone YPMB-G12 16S ribosomal RNA gene, partial sequence	Microbial diversity of endosymbiont bacteria in the spiraling Whitefly	Host-associated	77133	gi|411181008|gb|JQ894900.1|	100	0	9	21	
JQ900532.1.1441	Pseudomonas mendocina strain B8 16S ribosomal RNA gene, partial sequence	Crude oil degrading bacterial isolates from ecological region of Assam	Subsurface	300	gi|402244732|gb|JQ900532.1|	100	0	2	126	
JX222557.1.1395	Uncultured bacterium clone EMIRGE_OTU_s2b2b_12537 16S ribosomal RNA gene,	Subsurface microbial community response to acetate amendment	Subsurface	77133	gi|395559126|gb|JX222557.1|	100	0	1	12	
JX222807.1.1498	Uncultured bacterium clone EMIRGE_OTU_s3t2d_3949 16S ribosomal RNA gene, partial sequence	Subsurface microbial community response to acetate amendment	Subsurface	77133	gi|395559376|gb|JX222807.1|	100	0	4	9	
JX223934.1.1501	Uncultured bacterium clone EMIRGE_OTU_s6b4a_3133 16S ribosomal RNA gene, partial sequence	Subsurface microbial community response to acetate amendment	Subsurface	77133	gi|395560503|gb|JX223934.1|	100	0	5	10	
JX266364.1.1449	Pseudomonas sp. B2085 16S ribosomal RNA gene, partial sequence	Depth-Related Changes in Community Structure of Culturable Mineral Weathering Bacteria and in Weathering Patterns Caused by Them along Two Contrasting Soil Profiles	Subsurface	1225045	gi|402549839|gb|JX266364.1|	100	0	31	558	
JX945764.1.1431	Pseudomonas sp. LARP66 16S ribosomal RNA gene, partial sequence	Microbial diversity of Ethiopian soda lakes assessed by cultivation Methods	Lake	1266816	gi|428274238|gb|JX945764.1|	100	0	8	11	
KC119131.1.1470	Pseudomonas sp. RCC12 16S ribosomal RNA gene,	Isolation of CO2 fixing bacteria from dehradun caves	Subsurface	1268824	gi|429841843|gb|KC119131.1|	100	0	2	9	
KC166766.1.1460	Uncultured Pseudomonas sp. clone BC061 16S ribosomal RNA gene, partial sequence	Bacterial community and groundwater quality changes in an anaerobic aquifer during groundwater recharge with aerobic recycled water	Subsurface	114707	gi|523453881|gb|KC166766.1|	100	0	5	15	
KC469080.1.1418	Pseudomonas sp. EM174 16S ribosomal RNA gene,	Direct Submission, isolate EM174	Lab strain	1282313	gi|452108584|gb|KC469080.1|	100	0	6	19	
KC758944.1.1396	Uncultured bacterium clone 12ALLV2e09 16S ribosomal RNA gene, partial sequence	Using DNA-stable isotope probing to identify microorganisms involved in mtbe and tba biodegradation	Coastal water	77133	gi|478444906|gb|KC758944.1|	100	0	2	22	
KC852955.1.1440	Pseudomonas cuatrocienegasensis strain LEH6_4A 16S ribosomal RNA gene, partial sequence	Midgut Microbial Community of Culex quinquefasciatus Mosquito Populations from India	Host-associated	77133	gi|523453428|gb|KC852955.1|	100	0	25	293	
KF021824.1.1426	Pseudomonas sp. H-144 16S ribosomal RNA gene,	Cultured diversity of marine bacteria	Seawater	1345863	gi|513045820|gb|KF021824.1|	100	0	2	9	
KF494749.1.1502	Uncultured bacterium clone B24-205 16S ribosomal RNA gene, partial sequence	Vertical changes of the structure of bacterial communities through a permafrost core profile from Qinghai-Tibet Plateau	Subsurface	77133	gi|532529728|gb|KF494749.1|	100	0	2	8	
KF494759.1.1501	Uncultured bacterium clone B9-456 16S ribosomal RNA gene, partial sequence	Vertical changes of the structure of bacterial communities through a permafrost core profile from Qinghai-Tibet Plateau	Subsurface	77133	gi|532529738|gb|KF494759.1|	100	0	5	11	
KF657327.1.1480	Pseudomonas mendocina strain 2E 16S ribosomal RNA	Alkalo Tolerant Bacteria, water	Lake	300	gi|545599206|gb|KF657327.1|	100	0	6	10	
KF722300.1.1267	Uncultured Pseudomonas sp. clone DVBSW_J342 16S ribosomal RNA gene, partial sequence	Response of bacterial community structure to seasonal fluctuation and anthropogenic pollution on coastal water of Alang-Sosiya ship breaking yard, Bhavnagar, India	Coastal water	114707	gi|643039831|gb|KF722300.1|	100	0	5	9	
KF836136.1.1422	Pseudomonas plecoglossicida strain SBADK2 16S ribosomal RNA gene, partial sequence	rhizosphere bacteria for agricultural and environmental use	Subsurface	70775	gi|578003379|gb|KF836136.1|	100	0	15	46	
KF851140.1.1502	Uncultured Pseudomonas sp. clone BJP8S20-c24 16S ribosomal RNA gene, partial sequence	Lithology-Controlled Bacteria Community in an Ammonium-Rich Aquifer-Aquitard System in the Pearl River Delta, China	Subsurface	114707	gi|582054428|gb|KF851140.1|	100	0	10	30	
KF923425.1.1502	Pseudomonas xanthomarina strain 15 16S ribosomal	Culturable bacteria from the Qinghai-Tibet Plateau	Soil	271420	gi|594591039|gb|KF923425.1|	100	0	8	10	
KJ210658.1.1345	Pseudomonas xinjiangensis strain WL-257 16S ribosomal RNA gene, partial sequence	Phylogenetic diversity of eudophytic bacteria from Populus euphratica in 20-year Pupulus jujube forest ecotone of Chakhikh county of Xinjiang	Subsurface	487184	gi|612340510|gb|KJ210658.1|	100	0	8	25	
KJ424421.1.1503	Pseudomonas sp. GW28-5 16S ribosomal RNA gene,	cytotoxic bacteria from Antarctica	Host-associated	77133	gi|601035987|gb|KJ424421.1|	100	0	4	121	

Supplementary Table 4 – BLAST hits table for representative sequences associated to OTUs affiliated to genus *Pseudomonas* .

KJ475025.1.1452	Pseudomonas peli strain IARI-RP26 16S ribosomal RNA gene, partial sequence	Prospecting cold deserts of north western Himalayas for microbial diversity and plant growth promoting attributes	Soil	53406;592361	gi|387285851|gb|JQ795777.1|	100	0	27	962	
KJ600877.1.1458	Uncultured bacterium clone 83A 16S ribosomal RNA gene, partial sequence	Manipulating the banana rhizosphere microbiome for biological control of Panama disease	Subsurface	77133	gi|646117505|gb|KJ600877.1|	100	0	9	24	
KJ806232.1.1411	Pseudomonas alcaligenes strain PBR-49 16S ribosomal RNA gene, partial sequence	Phylogenetic characterization of heterotrophic bacteria isolated from photobioreactor (PBR) cultures of Synechocystis sp. PCC6803	Host-associated	43263	gi|669340641|gb|KJ806232.1|	100	0	10	26	
KJ806252.1.1456	Pseudomonas stutzeri strain PBR-57 16S ribosomal RNA gene, partial sequence	Phylogenetic characterization of heterotrophic bacteria isolated from photobioreactor (PBR) cultures of Synechocystis sp. PCC6803	Host-associated	316	gi|669340661|gb|KJ806252.1|	100	0	7	10	
KJ809251.1.1512	Uncultured bacterium clone F33GN 16S ribosomal RNA gene, partial sequence	Bacteria associated with arbuscular mycorrhizal fungi within roots of plants growing in a soil highly contaminated with aliphatic and aromatic petroleum hydrocarbons	Subsurface	77133	gi|671777739|gb|KJ809251.1|	100	0	2	40	
KJ862118.1.1417	Uncultured Pseudomonas sp. clone 35 16S ribosomal RNA gene, partial sequence	16S rDNA gene clone libraries which isolated from biodesulfurization bioreactor	Microcosm	114707	gi|672443372|gb|KJ862118.1|	100	0	8	79	
KP704423.1.1437	Pseudomonas sp. UYFA113 16S ribosomal RNA gene, partial sequence	Identification and characterization of the part of the bacterial community associated with field-grown tall fescue (Festuca arundinacea) cv. SFRO Don Tom s in Uruguay	Subsurface	1605337	gi|756058359|gb|KP704423.1|	100	0	7	11	
LN560920.1.1377	Uncultured bacterium partial 16S rRNA gene, clone SIGC587_N11D1_16S_B	Leaf-cutter ant refuse dumps are nutrient reservoirs harboring diverse microbial assemblages	Microcosm	77133	gi|697256444|emb|LN563985.1|	100	0	15	200	
LN563984.1.1372	Uncultured bacterium partial 16S rRNA gene, clone SIBS643_N12D0_16S_B	Leaf-cutter ant refuse dumps are nutrient reservoirs harboring diverse microbial assemblages	Microcosm	77133	gi|697256444|emb|LN563985.1|	100	0	6	16	
LN565700.1.1370	Uncultured bacterium partial 16S rRNA gene, clone SIBG551_N12D2_16S_B	Leaf-cutter ant refuse dumps are nutrient reservoirs harboring diverse microbial assemblages	Microcosm	77133	gi|697264995|emb|LN565700.1|	100	0	7	123	

References

1. Borgonie, G. *et al.* Nematoda from the terrestrial deep subsurface of South Africa. *Nature* **474,** 79–82 (2011).
2. McMahon, S. & Parnell, J. Weighing the deep continental biosphere. *FEMS Microbiol. Ecol.* **87,** 113–120 (2014).
3. Lau, M. C. Y. *et al.* Phylogeny and phylogeography of functional genes shared among seven terrestrial subsurface metagenomes reveal N-cycling and microbial evolutionary relationships. *Front. Microbiol.* **5,** 531 (2014).
4. Fang, J., Kato, C., Runko, G. M., Nogi, Y. & Hori, T. Predominance of Viable Spore-Forming Piezophilic Bacteria in High-Pressure Enrichment Coal-Bearing Sediments below the Ocean Floor. *Front. Microbiol.* **8,** (2017).
5. Onstott, T. C. *et al.* Does aspartic acid racemization constrain the depth limit of the subsurface biosphere? *Geobiology* **12,** 1–19 (2014).
6. McMurdie, P. J. & Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8,** e61217 (2013).
7. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33,** 1870–1874 (2016).
8. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–7 (2004).
9. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–25 (1987).
10. Tamura, K., Nei, M. & Kumar, S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci.* **101,** 11030–11035 (2004).
11. Barnhart, E. P. *et al.* Hydrogeochemistry and coal-associated bacterial populations from a methanogenic coal bed. *Int. J. Coal Geol.* **162,** 14–26 (2016).
12. Foght, J. *et al.* Culturable Bacteria in Subglacial Sediments and Ice from Two Southern Hemisphere Glaciers. *Microb. Ecol.* **47,** 329–40 (2004).
13. Leuko, S. *et al.* Lysis efficiency of standard DNA extraction methods for Halococcus spp. in an organic rich environment. *Extremophiles* **12,** 301–308 (2008).