A global perspective on bacterial diversity in the terrestrial deep subsurface

A. Soares1,2,3†, A. Edwards2,4*, D. An5, A. Bagnoun6, J. Bradley7, E. Barnhart8,9, M. Bomberg10, K. Budwill11, S. M. Caffrey12, M. Fields13, J. Gralnick3, V. Kadnikov14, L. Mompier15, M. Osburn16, A. Mu17,18,19, J. W. Moreau20, D. Moser21, L. Purkamo10,22,23, S. M. Rassner1,4, C. S. Sheik24, B. Sherwood Lollar25, B. M. Toner26, G. Voordouw5, K. Wouters27 and A. C. Mitchell1,4,*

Abstract

While recent efforts to catalogue Earth’s microbial diversity have focused upon surface and marine habitats, 12–20% of Earth’s biomass is suggested to exist in the terrestrial deep subsurface, compared to ~1.8% in the deep subsea floor. Metagenomic studies of the terrestrial deep subsurface have yielded a trove of divergent and functionally important microbes from a range of localities. However, a wider perspective of microbial diversity and its relationship to environmental conditions within the terrestrial deep subsurface is still required. Our meta-analysis reveals that terrestrial deep subsurface microbiota are dominated by Betaproteobacteria, Gammaproteobacteria and Firmicutes, probably as a function of the diverse metabolic strategies of these taxa. Evidence was also found for a common small consortium of prevalent Betaproteobacteria and Gammaproteobacteria operational taxonomic units across the localities. This implies a core terrestrial deep subsurface community, irrespective of aquifer lithology, depth and other variables, that may play an important role in colonizing 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.

Received 26 January 2022; Accepted 07 March 2022; Published 17 January 2023

Author affiliations: 1Department of Geography and Earth Sciences (DGES), Aberystwyth University (AU), Aberystwyth, UK; 2Institute of Biology, Environmental and Rural Sciences (IBERS), AU, Aberystwyth, UK; 3Department of Plant and Microbial Biology, University of Minnesota, Minneapolis, MN, USA; 4Interdisciplinary Centre for Environmental Microbiology (iCEM), AU, Aberystwyth, UK; 5Department of Biological Sciences, University of Calgary, Calgary, Canada; 6Institut de Génie Thermique (IGT), Haute École d’Ingénierie et de Gestion du Canton de Vaud (HEIG-VD), Yverdon-les-Bains, Switzerland; 7School of Geography, Queen Mary University of London, London, UK; 8U.S. Geological Survey (USGS), USA, Reston, VA, USA; 9Center for Biofilm Engineering (CBE), Montana State University, Bozeman, MT, USA; 10VT Technical Research Centre of Finland, Finland; 11Alberta Innovates, Canada; 12University of Toronto, Toronto, Canada; 13Department of Microbiology & Immunology, MSU, Bozeman, MT, USA; 14Institute of Bioengineering, Research Center of Biotechnology, Russian Academy of Sciences, Russia; 15Department of Earth, Atmospheric and Planetary Sciences (DEAPS), The Massachusetts Institute of Technology (MIT), Cambridge, MA, USA; 16Department of Earth and Planetary Sciences, Northwestern University, Evanston, IL, USA; 17Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Australia; 18Doherty Applied Microbial Genomics, Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Australia; 19Microbiological Diagnostic Unit Public Health Laboratory, Department of Microbiology and Immunology, University of Melbourne, Melbourne, Australia; 20School of Earth Sciences, The University of Melbourne, Parkville, Australia; 21Division of Hydrologic Sciences, Desert Research Institute (DRI), Las Vegas, NV, USA; 22School of Earth and Environmental Sciences (SEES), University of St. Andrews, St. Andrews, UK; 23Geological Survey of Finland (GTK), Finland; 24Large Lakes Observatory, University of Minnesota, Duluth, MN, USA; 25Department of Earth Sciences, University of Toronto, Toronto, Canada; 26Department of Soil, Water & Climate, University of Minnesota, Minneapolis/Saint Paul, MN, USA; 27Institute for Environment, Health and Safety (EHS), Belgian Nuclear Research Centre SCK•CEN, Mol, Belgium.

†Present address: Group for Aquatic Microbial Ecology (GAME), University of Duisburg-Essen, Campus Essen – Environmental Microbiology and Biotechnology, Universitätsstr. 5, 45141 Essen, Germany.

*Correspondence: A. C. Mitchell, nem@aber.ac.uk; A. Edwards, aye@aber.ac.uk

Keywords: deep subsurface; 16S rRNA gene; meta-analysis; bacterial.

Abbreviations: ANOVA, analysis of variance; BIOM, biological observation matrix; CR, closed-reference; CTAB, hexadecyltrimethylammonium bromide; EPS, extracellular polymeric substances; ITS, internal transcribed spacer; MPa, megapascal; OTU, operational taxonomic unit; OTU, operational taxonomic unit; PERMANOVA, permutational multivariate analysis of variance; PgC, petagrams of carbon; qPCR, quantitative polymerase chain reaction; rRNA, ribosomal RNA; SRA-NCBI, sequence read archive database of the National Center for Biotechnology Information.

OPEN ACCESS

This is an open-access article distributed under the terms of the Creative Commons Attribution License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author’s institution.

Five supplementary figures and four supplementary tables are available with the online version of this article.

001172 © 2023 The Authors

DOI 10.1099/mic.0.001172
DATA SUMMARY

The 16S rRNA gene sequencing data utilized in the present study are available on NCBI under the following project accessions: PRJNA262938, PRJNA268940, PRJNA248749, PRJNA251746, PRJNA375701, PRJEB1468 and PRJEB10822. The code used for the processing and data analysis of the datasets is available at: https://github.com/GeoMicroSoares/mads_scripts.

INTRODUCTION

Understanding the distribution of microbial diversity is pivotal for advancing our knowledge of deep subsurface global biogeochemical cycles [1, 2]. 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, ca. 0.5 billion years ago [3]. 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) [4]. These values halved estimates from efforts earlier that year but maintained the notion that the terrestrial deep subsurface holds ca. 5-fold more bacterial and archaeal biomass than the deep marine subsurface [4, 5]. Further, it is expected that 20–80% of the possible 2–6×10^29 prokaryotic cells present in the terrestrial subterranean biome exist as biofilms and play crucial roles in global biogeochemical cycles [4, 6, 7].

Cataloguing microbial diversity and functionality in the terrestrial deep subsurface has mostly been achieved by means of marker gene and metagenome sequencing from aquifers associated with coals, sandstones, carbonates and clays, as well as deep igneous and metamorphic rocks [8–19]. Only recently has the first comprehensive database of 16S rRNA gene-based studies targeting terrestrial subsurface environments been compiled [4]. 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 [4]. While highlighting Firmicutes and Proteobacterial dominance in the bacterial component of the terrestrial deep subsurface, no further taxonomic insights emerged. Genus-level identification remains an important niche necessary for understanding community composition, inferred metabolism and hence microbial contributions of distinct community members to biogeochemical cycling in the deep subsurface [15, 20–22]. Indeed, such genus-specific traits have been demonstrated to be critical for understanding crucial biological functions in other microbiomes, and genus-specific functions of relevance for deep subsurface biogeochemistry are clear [23–25].

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 [16, 17, 19, 26–29]. Recent studies of deep terrestrial subsurface microbial communities further suggest that these are metabolically active, often associated with novel uncultured phyla, and potentially directly involved in carbon and sulphur cycling [30–36]. Concomitant advancements in subsurface drilling, molecular methods and computational techniques have aided exploration of the subsurface biosphere, but serious challenges remain, mostly related to deciphering sample contamination by drilling methods, community interactions with reactive casing materials and sample transportation to laboratories for processing [37, 38]. The logistical challenges inherent in accessing and recovering in situ samples from hundreds to thousands of metres below the 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 bioinformatics scripts to generate insights into the microbial community structure and controls upon bacterial and archaeal 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, varying 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 for the purposes of this initial examination as the zone at least 100 m from the surface [7, 40, 41].

METHODS

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 their metadata and information on sequencing platform used – i.e. only samples derived from 454 pyrosequencing and Illumina sequencing were considered. Due to a lack of public availability for Illumina datasets targeting environments of interest, only 454 pyrosequencing datasets were retained. Analysis of related literature resulted in the detection of other deposited studies that 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://hpc.nih.gov/apps/sratoolkit.html)
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 customized pipeline was created in *bash* language making use of *python* scripts developed for QIIME v1.9.1, to facilitate bioinformatic analyses in this study (see https://github.com/GeoMicroSoares/mads_scripts for scripts) [42]. Briefly, demultiplexed FASTQ files were processed to create an operational taxonomic unit (OTU) table. Quality control steps involved trimming, quality-filtering and chimera checking by means of USEARCH 6.1 [43]. Sequence data that passed quality control were then subjected to closed-reference (CR) OTU-picking on a per-study basis using UCLUST and reverse strand matching against the silva v123 taxonomic references (https://www.arb-silva.de/documentation/release-123/) [43]. CR 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 that recorded in the database, cross-study comparisons of bacterial 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 from the merging of all classification outputs. A single BIOM (Biological Observation Matrix) file was generated using QIIME's *merge_ou_tables.py* script. The BIOM file was then filtered to exclude samples represented by fewer than two OTUs using *filter_samples_from_ou_table.py*, as well as OTUs represented by one sequence (singleton OTUs) by using *filter_otus_from_ou_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_ou_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 [44–46]. 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 the phylum level. For the metadata category-directed analyses, the function *merge_samples (phyloseq)* created averaged OTU tables, which permitted testing of hypotheses for whether geography or depth had significant impacts on bacterial 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) [47, 48]. All figures were plotted via the *ggplot2* R package (https://github.com/tidyverse/ggplot2), except for the UpsetR plot in Fig. S4, which was plotted with the package *UpsetR* (https://github.com/hms-dbmi/UpSetR).

**RESULTS**

A total of 233 publicly available subsurface samples targeting multiple 16S rRNA gene hypervariable regions originating in nine countries were originally downloaded from the NCBI SRA database. These accounted for 24632035 chimera-checked sequences [11, 27, 49–53], which underwent silva 123-aided CR OTU-picking. The discovery of 46 OTUs classified as Chloroplast (Cyanobacteria) and phototrophic members of the phyla Chloroflexi and Chlorobi as well as orders Rhodospirillales and Chromatiales (Alpha- and Gammaproteobacteria classes, respectively) justified the use of additional stricter contamination-aware filtering (see Methodology, Table S1 for differences in numbers of reads between methods).

The final dataset consisted of 70 samples and 2207 OTUs (513929 sequences). Seventeen aquifers were included that were associated with either sedimentary- or crystalline-host rocks, from depths spanning 94–2300 m below the land surface, targeting mostly groundwater across five countries (Table S2). Nine DNA extraction techniques were used in these studies, ranging from standard and modified kit protocols (e.g. MOBIO PowerSoil) to phenol–chloroform and CTAB/NaCl-based methods [50, 51, 54–57]. Six different primer pair amplified regions of the 16S rRNA gene with 454 pyrosequencing technology were used to generate the datasets (see Fig. S1). Metadata variables that were unavailable for all samples in the dataset were excluded from the statistical analyses. All studies followed aseptic sample handling protocols and included DNA extraction and PCR controls (for further information see Methods sections of the papers enumerated in Table 1) as per recommended guidelines for the subsurface microbiology community [38, 58].

Among a total of 45 detected bacterial phyla, Proteobacteria were seen to dominate most deep subsurface community profiles in this dataset (Fig. 1). The most abundant proteobacterial classes (Alpha-, Beta-, Delta-, Gammaproteobacteria) represented 57.2% of the total number of reads. Betaproteobacteria, chiefly represented by the order Burkholderiales, accounted for 26.1% of all reads in the dataset. The order Burkholderiales was the main component of some host-rocks, accounting for up to 59.5 and 92.7% of host-rock-level relative abundance profiles for biotite-gneiss and chlorite-sericite-schist (see Fig. S2 for standard deviations of Fig. 1) and co-dominated others. Gammaproteobacteria and Clostridia (Firmicutes) were key components of other
Table 1. Metadata table for the studies utilized in this meta-analysis (see Table S3, Fig. S3 for more details).

NA, Not available. The dataset unavailable through SRA is available in JGI GOLD Study ID Gs0047444.

<table>
<thead>
<tr>
<th>SRA accession</th>
<th>DOI</th>
<th>Year(s) of sampling</th>
<th>Final no. of samples</th>
<th>Location</th>
<th>Depth gradient (m)</th>
<th>Host rock</th>
<th>Final no. of sequences</th>
<th>Final no. of OTUs</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRJNA262938</td>
<td>10.3389/fmicb.2014.00610</td>
<td>2013/2014</td>
<td>6</td>
<td>South Dakota, USA</td>
<td>243.84–1478.28</td>
<td>Sulphide-rich schists</td>
<td>170,364</td>
<td>1367</td>
<td>[50]</td>
</tr>
<tr>
<td>PRJNA268940</td>
<td>na</td>
<td>2007–2011</td>
<td>8</td>
<td>California, USA, Ontario, Canada</td>
<td>94–2383</td>
<td>Dolomite, tuff, rhyolite, tuff breccia</td>
<td>68,071</td>
<td>741</td>
<td>na</td>
</tr>
<tr>
<td>PRJNA248749</td>
<td>na</td>
<td>2011</td>
<td>6</td>
<td>Minnesota, USA</td>
<td>750</td>
<td>Haematite iron formation</td>
<td>69,757</td>
<td>511</td>
<td>na</td>
</tr>
<tr>
<td>PRJNA251746</td>
<td>na</td>
<td>2009</td>
<td>6</td>
<td>Washington, USA</td>
<td>393–1135</td>
<td>Basalt</td>
<td>30,121</td>
<td>613</td>
<td>na</td>
</tr>
<tr>
<td>PRJNA475701</td>
<td>10.1016/j.coal.2016.05.001</td>
<td>2013</td>
<td>4</td>
<td>Montana, USA</td>
<td>109–114.7</td>
<td>Sub-bituminous coal, shale, sandstone, siltstone</td>
<td>35,926</td>
<td>154</td>
<td>[51]</td>
</tr>
<tr>
<td>PRJEB1468</td>
<td>10.1128/AEM.01737-15</td>
<td>2009</td>
<td>27</td>
<td>Alberta, Canada</td>
<td>140–1064.4</td>
<td>Sub-bituminous coal, volatile bituminous coal</td>
<td>100,618</td>
<td>5110</td>
<td>[52]</td>
</tr>
<tr>
<td>PRJEB10822</td>
<td>10.1111/j.1574-6941.12171</td>
<td>na</td>
<td>6</td>
<td>Mol, Belgium</td>
<td>217–232</td>
<td>Kaolinite/Halite and smectite</td>
<td>47,123</td>
<td>467</td>
<td>[27]</td>
</tr>
</tbody>
</table>
profiles. *Clostridia* and other *Firmicutes* accounted for large fractions of sedimentary host-rocks (dolomite, siltstone and shale) and a haematite iron formation. Finally, *Actinobacteria* was the most abundant taxonomic group in rhyolite-tuff-breccia.

Analysis of prevalence across the dataset revealed that seven OTUs, all affiliated with the genus *Pseudomonas*, were present in more than 25 and up to 41 samples, accounting for 18 149 reads (3.5% of the total reads, see Fig. 2, Table S3). Other bacterial orders, namely *Burkholderiales*, *Alteromonadales* and *Clostridiales* (*Betaproteobacteria*, *Gammaproteobacteria*, *Clostridia*) were also highly prevalent throughout. Network analysis (Table 2) highlighted a *Pseudomonas* OTU highly connected to other OTUs in the dataset. Furthermore, BLAST results indicated that recovered sequences for OTUs affiliated with this genus were generally
Table 2. Top 10 most central OTUs (Proteobacterial classes highlighted) 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 corresponding closeness centrality (scores of shortest paths to and from an OTU to all the remaining OTUs in a network) and degree (number of directly connected edges, or OTUs) values

<table>
<thead>
<tr>
<th>OTU ID</th>
<th>OTU classification</th>
<th>Centrality</th>
<th>Closeness</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF554871.1.1486</td>
<td><strong>Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas</strong></td>
<td>1.0000000</td>
<td>2.13e-05</td>
<td>38</td>
</tr>
<tr>
<td>HH792658.1.1492</td>
<td><strong>Proteobacteria; Betaproteobacteria; Rhodocyclus; Rhodocyclus; Thauera</strong></td>
<td>0.9753542</td>
<td>2.13e-05</td>
<td>36</td>
</tr>
<tr>
<td>HQ681977.1.1496</td>
<td><strong>Proteobacteria; Betaproteobacteria; Burkholderiales; Comamonadaceae; Diaphorobacter</strong></td>
<td>0.9445053</td>
<td>2.13e-05</td>
<td>34</td>
</tr>
<tr>
<td>KF465077.1.1336</td>
<td><strong>Proteobacteria; Betaproteobacteria; Burkholderiales; Comamonadaceae; Acidovorax</strong></td>
<td>0.8887751</td>
<td>2.13e-05</td>
<td>30</td>
</tr>
<tr>
<td>JQ072853.1.1348</td>
<td><strong>Proteobacteria; Betaproteobacteria; Rhodocyclus; Rhodocyclus; Thauera</strong></td>
<td>0.8808455</td>
<td>2.13e-05</td>
<td>30</td>
</tr>
<tr>
<td>KM200734.1.1449</td>
<td><strong>Proteobacteria; Alphaproteobacteria; Rhizobiales; Rhizobiales; Rhizobium</strong></td>
<td>0.8716886</td>
<td>2.13e-05</td>
<td>31</td>
</tr>
<tr>
<td>KC758926.1.1392</td>
<td><strong>Proteobacteria; Betaproteobacteria; Burkholderiales; Comamonadaceae; Acidovorax</strong></td>
<td>0.8662805</td>
<td>2.13e-05</td>
<td>29</td>
</tr>
<tr>
<td>FJ032194.1.1456</td>
<td><strong>Proteobacteria; Betaproteobacteria; Burkholderiales; Comamonadaceae; Rhodofex</strong></td>
<td>0.8662805</td>
<td>2.13e-05</td>
<td>29</td>
</tr>
<tr>
<td>EU771645.1.1366</td>
<td><strong>Firmicutes; Bacilli; Bacillales; Planococcaceae; Planomicrobium</strong></td>
<td>0.8476970</td>
<td>2.13e-05</td>
<td>30</td>
</tr>
<tr>
<td>JN245782.1.1433</td>
<td><strong>Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacterales; Defluviimonas</strong></td>
<td>0.8356655</td>
<td>2.13e-05</td>
<td>29</td>
</tr>
</tbody>
</table>

associated with marine and terrestrial soil and sediments (see Fig. S3, Table S4) [59]. 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. The genus *Thauera* (Betaproteobacteria, Rhodocyclus), represented by a single OTU, was the second most central to the dataset. While relative abundance patterns across the dataset (Fig. 1) indicate that lithology could influence microbial community composition and structure, sample sizes for each host-rock in the final dataset were insufficient to provide robust statistical support of that hypothesis. Despite this, host-rocks (10 out of 15) presented, on average, more unique OTUs than they shared with other host-rocks (Fig. S4). In particular, in sulphide-rich schists, 73% of the OTUs were, on average, unique to the host-rock. Sub-bituminous and volatile bituminous coals shared a total of 143 OTUs; this was the strongest interaction between host-rocks. In the dataset, no significant correlations were found for the presence of the most abundant clades in the dataset and depth, *Actinobacteria* being the only major taxonomic group to have a positive, albeit weak, correlation with depth (Pearson’s *r* = 0.42, *P* < 0.01, Fig. S5). Proportions of *Beta*- and *Gammaproteobacteria* generally decreased with depth (Pearson’s *r* = −0.29 and −0.093, respectively), but no other major clades were shown to correlate.

Ordination of the final dataset further suggests 50.6% of Jensen–Shannon distances were significantly explained by aquifer lithology (ADONIS/PERMANOVA, F-statistic = 4.65, *P* < 0.001, adjusted Bonferroni correction *P* < 0.001). Other environmental features such as absolute 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 bacterial community structure, respectively (ADONIS/PERMANOVA, F-statistic = 3.95, 3.57, *P* < 0.001, adjusted Bonferroni correction *P* < 0.001). Finally, no evidence was found for DNA extraction or 16S rRNA gene region significantly affecting bacterial community structure in this meta-analysis (ADONIS/PERMANOVA, F-statistic = 3.85, 3.23, *P* < 0.01, adjusted Bonferroni correction *P* < 0.001).

**DISCUSSION**

The deep biosphere is an active, diverse biome still largely under-investigated in terms of the Earth’s biogeochemistry [12, 60–62]. In this study, publicly available 16S rRNA gene data revealed a prevalence of *Betaproteobacteria* and *Gammaproteobacteria* in the deep biosphere that may be explained by the diverse metabolic capabilities of taxa within these clades. The families *Gallionellaceae*, *Pseudomonadaceae*, *Rhodocyclusaceae* and *Hydrogeniphillaceae* within *Betaproteobacteria* and *Gammaproteobacteria* are suggested to play critical roles in deep subsurface iron, nitrogen, sulphur and carbon cycling across the world [50, 61, 63]. The relative abundance of the order *Burkholderiales* (Betaproteobacteria) in surficial soils has previously been correlated (*R*² = 0.92, ANOVA *P* < 0.005) with mineral dissolution rates, while the genus *Pseudomonas* (*Gammaproteobacteria*) is widely known to play a key role in hydrocarbon degradation, denitrification and coal solubilization in different locations [64–66]. 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 [67–70].

The metabolic plasticity of the orders *Pseudomonadales* and *Burkholderiales* has been demonstrated and may be a catalyst for their apparent centrality across the terrestrial deep subsurface microbiomes analysed in this study [71–74]. These bacterial orders may represent important keystone taxa in microbial consortia responsible for providing critical 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 play a central role in establishing conditions for microbial colonization in many terrestrial subsurface environments. The genus *Pseudomonas* and possibly several members of *Burkholderiales* may therefore comprise an important component of the global core terrestrial deep subsurface bacterial community [11]. Geographically comprehensive RNA-based approaches should in the future investigate the potential roles of the genus *Pseudomonas* and order *Burkholderiales* in this biome.

The class *Clostridia* was found to be prevalent across the dataset and to dominate in sedimentary host-rocks (dolomite, silstone and shale) in this study. This class includes anaerobic hydrogen-driven sulphate reducers also known to sporulate and metabolize a wide range of organic carbon compounds [77]. Previously, members of *Clostridia* have also been identified as dominant components in extremely deep subsurface ecosystems beneath South Africa, Siberia and California (USA) from metasubalitc and metasedimentary lithologies [78]. Adaptation to extreme environments in this class has been associated with diverse metabolic capabilities that include sporulation ability and capacity for CO$_2$- or sulphur-based autotrophic H$_2$-dependent growth [21, 79]. In this study, network analysis and prevalence values suggested roles of putative importance for the classes *Betaproteobacteria*, *Gammaproteobacteria* and *Clostridia* in the deep terrestrial subsurface (Fig. 2, Table S3). Their maintained presence in this biome across strikingly dissimilar host-rocks and depth, among others, could be indicative of higher metabolic plasticity, providing physiological advantages over other members of microbial communities.

Lithotrophic microbial metabolisms and mineralogy-driven microbial colonization of relatively inert lithologies have previously been demonstrated with low abundance but more reactive minerals within rock matrices often cited as key controls on community structure [80–83]. Limiting factors for life in the terrestrial deep subsurface such as pressure and temperature are more closely correlated with depth. Growth of bacterial isolates from the deep subsurface has been documented at up to 48 MPa and 50 °C and has been associated with production of extracellular polymeric substances (EPS) [84, 85]. However, robust conclusions on the effects of lithology or depth on the structure and composition of microbial communities across Earth’s crust have presented a widespread challenge for science, as in this study due to the small and varied sample sizes resulting from the contamination-aware filtering process and the limited number of comparable lithology types. Large-scale evidence for the roles of eukaryotes, bacteria, archaea and viruses in the deep terrestrial subsurface and the environmental controls over their occurrence in this biome is still lacking. We recommend a field-level research strategy to gain insights into these aspects of life within Earth’s crust. Larger scale collation of data from samples collected and processed using unified, reproducible workflows will be cognizant of significant potential for contamination and ultimately allow robust insights on wide-ranging microbial metabolic processes in the terrestrial subsurface.

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 – https://deepcarbon.net/tag/census-deep-life), where physical, chemical and biological, but also in silico bioinformatics strategies to prevent erroneous conclusions have been highlighted [38, 86, 87]. This study also follows frequency-based OTU filtration techniques similar to those recommended previously [38] and designed to remove possible contaminants introduced during sampling or during the various steps related to sample processing [38]. The pre-emptive quality control steps hereby undertaken support a non-contaminant origin for the taxa analysed in this dataset. As such, the predominance of typically contaminant taxa affiliated, for example, to the genus *Pseudomonas* is accepted as a representative trend in reflecting the microbial ecology of the terrestrial deep subsurface.

Standardizing sampling, DNA extraction, sequencing and bioinformatics methods and strategies across the subsurface research community would help further reduce methodology-based variations. This would more efficiently permit re-analyses after collection, where methodological variations would be controlled, and robust wide-ranging overarching conclusions would more easily be achieved. Despite this, host-rock matrices and local geochemical conditions often pose unique challenges that require particular protocol adjustments [88]. 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 [89, 90]. 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 [91]. Finally, the ecology of domains Eukarya and Archaea across the terrestrial deep subsurface remains generally under-characterized and requires future attention. This study presents an important first step towards characterizing bacterial community structure and composition in the terrestrial deep biosphere.

A global-scale meta-analysis addressing the available 16S rRNA gene-based studies of the deep terrestrial subsurface revealed a dominance of *Betaproteobacteria*, *Gammaproteobacteria* and *Firmicutes* across this biome. Evidence for a core terrestrial deep subsurface microbiome population was recognized through the prevalence and centrality of the genus *Pseudomonas* (*Gammaproteobacteria*) and several other genera affiliated with the class *Betaproteobacteria*. The adaptable metabolic capabilities associated with the above-mentioned taxa may be critical for colonizing the deep subsurface and sustaining communities therein. The terrestrial deep subsurface is a hard-to-reach, complex ecosystem crucial to global biogeochemical cycles. Efforts by multiple teams of investigators to sequence subsurface ecosystems over the last decade were hereby consolidated to characterize the 12–20%
of global biomass this biome represents [5]. The strict contamination-aware filtering process applied whittled down the publicly available datasets representing terrestrial subsurface bacterial diversity to just 70 samples from two continents, indicating the need for systematic exploration of biodiversity within this major component of the biosphere. As a first step, this study consolidates a global-scale understanding of taxonomic trends underpinning a major component of terrestrial deep subsurface microbial ecology and biogeochemistry.

Funding information
The work was funded by a National Research Network for Low Carbon Energy and Environment (NRR-LCEE) grant to A.C.M. and A.E. from the Welsh Government and the Higher Education Funding Council for Wales (Geo-Carb-Cymru). 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 (SBR), 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
A.R.S. developed the methodology, collated and analysed the data, and wrote the manuscript. A.E. and A.M. conceived the study, supervised A.S. 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.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References


46. R Development Core Team. (n.d.) R: A language and environment for statistical computing.


68. Yagi JM, Sims D, Brettin T, Bruce D, Madsen EL. The genome of Polaronomas napththalenivorans strain CJ2, isolated from coal tar-contaminated sediment, reveals physiological and metabolic versatility and evolution through extensive horizontal gene transfer. Environ Microbiol;2009;11:2253–2270.


