Albanese et al. Microbiome (2021) 9:63
https://doi.org/10.1186/s40168-021-01021-0

Microbiome

RESEARCH Open Access

Pre-Cambrian roots of novel Antarctic cryptoendolithic bacterial lineages

Davide Albanese[1†], Claudia Coleine[2†], Omar Rota-Stabelli[1], Silvano Onofri[2], Susannah G. Tringe[3], Jason E. Stajich[4*], Laura Selbmann[2,5*] and Claudio Donati[1]

Abstract

Background: Cryptoendolithic communities are microbial ecosystems dwelling inside porous rocks that are able to persist at the edge of the biological potential for life in the ice-free areas of the Antarctic desert. These regions include the McMurdo Dry Valleys, often accounted as the closest terrestrial counterpart of the Martian environment and thought to be devoid of life until the discovery of these cryptic life-forms. Despite their interest as a model for the early colonization by living organisms of terrestrial ecosystems and for adaptation to extreme conditions of stress, little is known about the evolution, diversity, and genetic makeup of bacterial species that reside in these environments. Using the Illumina Novaseq platform, we generated the first metagenomes from rocks collected in Continental Antarctica over a distance of about 350 km along an altitudinal transect from 834 up to 3100 m above sea level (a.s.l.).

Results: A total of 497 draft bacterial genome sequences were assembled and clustered into 269 candidate species that lack a representative genome in public databases. Actinobacteria represent the most abundant phylum, followed by Chloroflexi and Proteobacteria. The "Candidatus *Jiangella antarctica*" has been recorded across all samples, suggesting a high adaptation and specialization of this species to the harshest Antarctic desert environment.
The majority of these new species belong to monophyletic bacterial clades that diverged from related taxa in a range from 1.2 billion to 410 Ma and are functionally distinct from known related taxa.

Conclusions: Our findings significantly increase the repertoire of genomic data for several taxa and, to date, represent the first example of bacterial genomes recovered from endolithic communities. Their ancient origin seems to not be related to the geological history of the continent, rather they may represent evolutionary remnants of pristine clades that evolved across the Tonian glaciation. These unique genomic resources will underpin future studies on the structure, evolution, and function of these ecosystems at the edge of life.

Keywords: Antarctica, Extremophiles, Cryptoendolithic communities, Bacteria, Evolution, Adaptation, Metagenomics, MAG, Functionality

* Correspondence: jason.stajich@ucr.edu; selbmann@unitus.it
†The authors Davide Albanese and Claudia Coleine contributed equally to this work.
[4]Department of Microbiology and Plant Pathology and Institute of Integrative Genome Biology, University of California, Watkins Drive 3401, Riverside, Riverside, CA 92507, USA
[2]Department of Ecological and Biological Sciences, University of Tuscia, Largo dell'Università, 01100 Viterbo, Italy
Full list of author information is available at the end of the article

Background

Rocks represent the earliest terrestrial niche for life on Earth when microbes were the only form of life [1, 2]. Porous rocks, in particular, remain the ultimate refuge for life in extreme environments as in the ice-free areas of Antarctica, where complex life-forms became extinct about 60-30 Ma, when the continent reached the South Pole and the Antarctic Circumpolar Current was established. The McMurdo Dry Valleys, covering a surface of approximately 4800 km^2 in Continental Antarctica, are among the most extreme regions on Earth with only minimal resources suitable for supporting life [3, 4]. Specifically, in these desert areas, where soils have been eroded by glaciers and strong winds, life is confined to the endolithic niche that provides microorganisms with thermal buffering, physical stability, protection from abiotic stresses, and access to mineral nutrients, rock moisture and growth surfaces [5, 6]. Indeed, the endolithic environment is a ubiquitous habitat for microorganisms in dryland systems [7], but in the harshest terrestrial climates, characterized by extreme environmental conditions typically incompatible with an active life, it is often the primary or even exclusive refuge for life [8].

Endolithic microbial communities are self-sustaining ecosystems relying on the phototrophic activity of microalgae and cyanobacteria as primary producers which support a diversity of consumers including fungi, bacteria, and archaea [9–11]. In the Antarctic desert areas, the Lichen-Dominated Communities (LDC) are the most complex and successful [5]. Recently, next-generation sequencing studies have brought new insights into their composition, showing that lichens in the Lecanoromycetes and free-living fungi in the Dothideomycetes (Ascomycota) are the dominant eukaryotes, while Actinobacteria and Proteobacteria are the most abundant prokaryotes [12, 13]. Due to their ubiquity in deserts and low taxonomic complexity and biodiversity [14], endoliths are important study systems to understand evolutionary processes in the early history of life, to model how life evolves during the progression of desertification and when the extreme aridity approaches the limits of life, providing also a model for searching life elsewhere in the solar system. However, the understanding of the microbial biodiversity in these communities is limited and our comprehension about their physiology, evolution, and stress responses is still at its infancy [15].

In this study, we performed metagenomic sequencing of eighteen LDC-colonized rock samples collected in Antarctic ice-free areas (Fig. 1a) distributed over a distance of 350 km (Fig. 1b,c) to provide the first survey of the genomic repertoire of bacteria from Antarctic endolithic ecosystems [16]. The metagenomic assemblies generated more than 10 million contigs which were binned into 497 novel bacterial genomes and classified as 269 previously unknown species-level clusters, substantially expanding the sampled genomic diversity within 33 bacterial orders.

While interest is increased in the ecological roles, diversity, conservation, and biotechnological potential of the Antarctic endolithic microbiota, the evolutionary origins are still unexplored. We used molecular clock analysis to characterize the time scale over which these taxa have differentiated from known related species and to test two fundamental hypotheses on the processes that led to the establishment of these communities: (i) Antarctic endolithic bacteria may have evolved from generalist species in response to the climatic changes occurring when Antarctica reached the South Pole and (ii) these taxa could have been selected from pre-existing extremo-tolerant species. Our results give clear evidence of ancient divergence of most Antarctic bacterial clades which date back up to 1.2 billion Ma. This evidence clearly supports the second hypothesis and excludes that they are the result of a recent evolution of genetic traits in response to environmental conditions.

Methods

Sampling area

Victoria Land is a region of Continental Antarctica which fronts the western side of the Ross Sea and the Ross Ice Shelf; this land is positioned between the Polar Plateau and the coast and is exposed to a wide spectrum of climatic extremes, including low and fluctuating temperature, scarce precipitation regimes, and strong winds; the region covers a latitudinal gradient of 8° from Darwin Glacier (78° 00′) to Cape Adare (70° 30′ S) [17]. Ice-free areas dominate the landscape of Southern Victoria Land and the high-altitude locations of Northern Victoria Land, while low-elevation coastal soils of Northern Victoria Land receive considerable marine and biological influence (e.g., sea birds).

Sandstone rocks were collected by L. Selbmann in Victoria Land along a latitudinal transect ranging from 74° 10′ 44.0′′ S 162° 30′ 53.0′′ E (Mt. New Zealand, Northern Victoria Land) to 77° 52′ 28.6′′ S 160° 44′ 22.6′′ E (University Valley, Southern Victoria Land) during the XXXI Italian Antarctic Expedition (Dec. 2015–Jan. 2016). Samples were collected at different conditions namely sun exposure and an altitudinal transect, from 834 to 3100 m a.s.l. to provide a comprehensive overview of endolithic diversity (Fig. 1a–c). Rocks were excised using a geologic hammer and sterile chisel, and rock samples, preserved in sterile plastic bags, transported, and stored at −20 °C in the Culture Collection of Antarctic fungi of the Mycological Section of the Italian

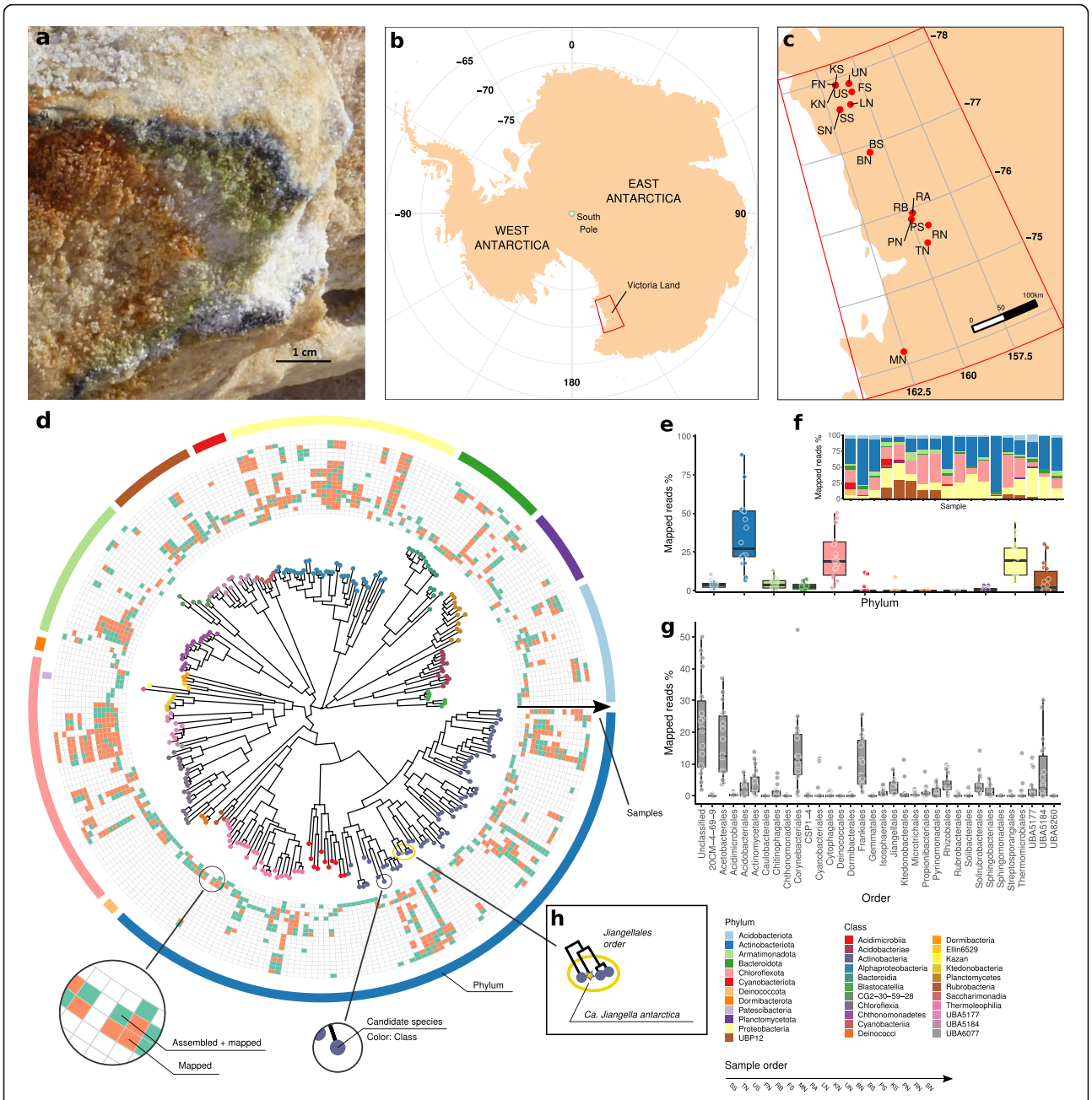

Fig. 1 **a** Cryptoendolithic lichen-dominated community colonizing sandstone at Linnaeus Terrace, McMurdo Dry Valleys, Southern Victoria Land, Continental Antarctica. **b** Map of Antarctica. The area encircled by the red lines represents Victoria Land. **c** Map of the study area showing the location of the sampling sites. Sampled sites are listed in Supplementary Table 4. **d** Phylogenetic tree built from the GTDB-Tk multiple sequence alignment (MSA) of the 269 CBS representatives. Tip points are colored according to the GTDB taxonomic classification at the class level. Phylum-level classification is indicated by the colors in the outer circle. The 18 inner circles represent distribution of each CBS in the samples. Presence inferred by the assembly is indicated in green, while presence inferred only from the alignment of the read to the CBS representative genome is indicated in orange. **e** Percentage of reads that could be mapped to the CBS representatives, grouped by Phylum. **f** Per sample percentage of the reads that could be mapped to the CBS representatives, grouped by Phylum. **g** Same as **e**, at the order taxonomic level. **h** Jiangellales CBS, including the Candidatus *Jiangella antarctica* (yellow star)

Antarctic National Museum (MNA-FCC), until downstream analysis.

DNA extraction, library preparation, and sequencing

DNA was extracted from three samples for each site and then pooled. Metagenomic DNA was extracted from 1 g of crushed rocks using a MoBio Powersoil kit (MOBIO Laboratories, Carlsbad, CA, USA). The quality of the DNA extracted was determined by electrophoresis using a 1.5% agarose gel and with a spectrophotometer (VWR International) and quantified using the Qubit dsDNA HS Assay Kit (Life Technologies, USA).

Shotgun metagenomic libraries were prepared and sequenced at the DOE Joint Genome Institute (JGI) as a part of a Community Science Project (PI: Laura Selbmann; co-PI: Jason E. Stajich) at JGI [16]. Paired-end sequencing libraries were constructed and sequenced as 2×150 bp using the Illumina NovaSeq platform (Illumina Inc, San Diego, CA).

Sequencing reads preparation and assembly

BBDuk (http://sourceforge.net/projects/bbmap/) v38.25 was used to remove contaminants, trim adapters, and low-quality sequences. The procedure removed reads that contained 4 or more "*N*" bases, had an average quality score across the read less than 3, or had a minimum length ≤ 51 bp or 33% of the full read length. Filtered and trimmed paired-end reads were error corrected using BFC [18] r181 with parameters `-1 -s 10g -k 21 -t 10` and orphan reads were removed. Samples were assembled individually with SPAdes [19] 3.12.0 using the parameters `-m 2000 -o spades3 --only-assembler -k 33,55,77,99,127 --meta -t 32`.

Binning

Metagenomic contigs were binned into candidate metagenome-assembled genomes (MAGs) using MetaBAT2 [20] (Metagenome Binning based on Abundance and Tetranucleotide frequency) v2.12.1. Briefly, high-quality reads were mapped on assembled contigs using Bowtie2 [21] v2.3.4.3. Samtools [22] v1.3.1 (htslib v1.3.2) was used to create and sort the BAM files (.bam). The depth of coverage was estimated by applying the `jgi_summarize_bam_contig_depths` tool. Contigs sequences and the depth of coverage estimates were used by MetaBAT2 to recover the candidate MAGs.

Quality assessment and dereplication

Completeness and contamination estimates of bacterial and archaeal MAGs were obtained by CheckM [23]. According to recent guidelines [24], MAGs were classified into "high-quality draft" (HQ) with >90% completeness and <5% contamination and "medium-quality draft" (MQ) with completeness estimates of ≥50% and less than 10% contamination. Candidate bacterial species (CBS) were identified by clustering HQ and MQ MAGs at species level [25] (>95% Average Nucleotide Identity - ANI) using dRep [26] v2.0.0. For each CBS, the MAG with the highest quality score was chosen as representative.

Taxonomic classification

MQ and HQ MAGs were taxonomically classified using the genome taxonomy database toolkit [27, 28] (GTDB-Tk) v0.1.6 and the GTDB release 86, following the recently proposed nomenclature of prokaryotes [29, 30]. GTDB-Tk classifies a query genome combining its placement in the GTDB reference tree (release 86 includes a total of 21,263 genomes in the tree), its RED, and its ANI to reference genomes. Approximately-maximum-likelihood phylogenetic tree from the GTDB protein alignments of the 269 CBS representatives (Fig. 1) and of the orders acetobacterales (Fig. 4) and Frankiales (Fig. S1) were inferred using FastTree [31] v2.1.10 (WAG+CAT model, options `-wag -gamma`) and rooted at midpoint.

Percentage of mapping reads and CBS detection

For each metagenomic sample, high-quality reads were aligned against each CBS representative using Bowtie2 [21] v2.3.4.3 using the parameter `--no-unal`. Samtools v1.3.1 (htslib 1.3.2) was used to create and index the BAM files (.bam). The depth of coverage, the breadth B_n (i.e., the fraction of bases of the CBS representative genome that are covered with depth n), and the number of mapped reads were calculated on the BAM file using pysam (https://github.com/pysam-developers/pysam) v0.15.2 and Python v3.5.3. The fraction of reads mapping on a CBS representative was computed as the number of successfully aligned reads normalized by the total number of reads aligning the entire set of the CBS representatives. Regions with no coverage were identified using BEDtools [32] v2.26.0 with the options `-bga -split`. Variant calling was performed with samtools mpileup and bcftools call [33] (v1.3.1, options `--ploidy 1 -mv`). Tabix [34] v1.3.2 was used to index the output VCF file. The consensus sequence was generated using the command bcftools consensus masking the zero coverage regions previously identified. The ANI between the consensus sequence and the CBS representative (ANI_{CBS}) was estimated using fastANI v1.1. Finally, a CBS was tagged as present in a sample if the breadth of coverage (at depth 2) B_2 was ≥ 0.5 and ANI_{CBS} ≥ 95%. We detected a total of 1094 CBS distributed within the 18 metagenomic samples (see Fig. 1d, S2).

Mash Screen [35] (Mash v. 2.1) was used to validate the presence of CBS in the Antarctic samples. Briefly, we sketched all the CBS representative genomes using a sketch size of 10,000 (replacing the default value of 1000) in order to have a superior representation of the sequences [36]; after that, the metagenomes were independently screened for containment of the CBS using the command mash screen. Given a metagenome, Mash Screen reported the containment score for each CBS (i.e., the estimate of the similarity of the CBS representative to a sequence contained within the metagenome) as a proxy for the average nucleotide identity, its p value, and the CBS median-multiplicity as a proxy for the genome coverage. We found that 1009 out of 1094 (92.2%) detected CBS have been confirmed (containment score >0.95, $p<1.47\times10^{-21}$, see Fig. S5, S6, and Supplementary Table 1). The remaining 85 discoveries have containment scores >0.91, and most of them (75) have a breadth of coverage B_2 between 0.5 and 0.7, which is compatible with the fact that Mash Screen tends to underestimate the identity when the query genome may not be fully represented by the sequencing reads [35].

Divergence estimates

Divergence times were independently estimated on orders containing at least 4 CBS, for a total of 19 analyzed orders. For each order, we built a protein MSA using the 120 GTDB bacterial marker genes including (i) 32 reference sequences from outgroups outside the order, (ii) the GTDB representatives, (iii) the MQ and HQ Antarctic MAGs, and (iv) a set of outgroup in order to reconstruct the first radiation within bacteria as in [37] and using it as a calibration point. The 19 datasets were calibrated with this same prior. We calibrated the crown (divergence) of bacteria using a prior on the root of 3453 million years ago (Ma) and a standard deviation of 60 Ma (values kindly provided by Davide Pisani) and corresponding to the posterior estimate for the crown of the bacteria [37]. Since our taxon sampling replicates the taxon sampling in [37], we could safely apply the previous estimate for the crown of the bacteria to our root (which coincides with the crown of bacteria, as we did not use archaea or eukaryotes outgroups). Markov chain Monte Carlo (MCMC) analyses were performed using BEAST [38] v1.10 for 100 million generations sampling every 1000 generations. Convergence was assessed by using the Effective Sample Sizes (ESS) estimated by Tracer [39] v1.7.1 on posteriors and log-likelihood. In order to maximize the ESS statistics, a burn-in ranging from 50 to 80% of the simulation was used. For computational reasons, we performed model selection using only one dataset (Acidobacteriales) as representative. We compared a relaxed clock (log-normal) versus the strict clock, and a coalescence (constant) versus a speciation (birth-death) demographic model. The most fitting combination of priors (relaxed clock plus coalescence) was found using path sampling and AICM. Amino-acid substitutions were modeled using the LG matrix with amino acid frequencies inferred from the data; among-site rate variation was modeled using a gamma distribution with four discrete categories. All Bayesian posterior annotated Maximum Clade Credibility Trees are reported in Supplementary Data. For each order, the mean age (plus the 95% high posterior densities heights) for the first split of a uniquely Antarctic group (green node) from the known reference sequence from that particular order was plotted. In the case of more than one monophyletic Antarctic group, the age of the second oldest Antarctic group (orange node) was also shown.

Functional annotation

Functional annotation was performed only on HQ CBS representatives of orders containing at least 4 CBS (for a total of 19 orders analyzed). In order to avoid systematic effects due to different annotation methods, both HQ MAGs and GTDB representative genomes (for a total of 3942 genomes) were processed as follows: (i) 16,292,642 translated coding DNA sequences (CDS) were predicted using Prokka [40] v1.13.4 which wraps the software Prodigal [41] and (ii) the CDS were functionally annotated using EggNOG-mapper [42] (option `--database bact`) and the eggNOG Orthologous Groups (OGs) database [43] v4.5.1. The EggNOG database integrates functional annotations collected from several sources, including KEGG functional orthologs [44], COG categories [45], and Gene Ontology (GO) terms.

In order to avoid annotation biases which are intrinsic to reference-based methods, we also clustered the CDS using MMseqs2 release 11-a29379e [46] (parameter --min-seq-id 0.60) generating 3,836,924 protein clusters. The cluster profiles were analyzed using the t-SNE dimensionality reduction (see the "Statistical analysis" section).

Statistical analysis

Downstream analysis was performed using the R environment (https://www.R-project.org/) v3.6.1. T-SNE dimensionality reduction (Jaccard distance) on KO and 60% identity cluster profiles was carried out using the R package "tsne" (https://CRAN.R-project.org/package=tsne) v0.1-3 and the PCoA (Principal Coordinate Analysis) using the function "pcoa()" (default parameters) available in the R library "ape" v5.3. Fisher's exact tests were conducted using the function "fisher.test()" (default parameters) available in the R package "stats" v3.6.1.

Results

Metagenomic assembly identifies novel bacterial species and broadly expands the tree of life

Using shotgun sequencing, we produced more than 10 million contigs that were binned into a total of 1660 metagenome-assembled genomes (MAGs), among which 497 were identified as bacterial and none as archaeal. The bacterial MAGs were partitioned into 263 high quality (HQ) and 234 medium quality (MQ) according to their estimated completeness and contamination (see the "Methods" section). Assembly, completeness and contamination statistics and the taxonomic classification of the 497 bacterial MAGs are given in Supplementary Table 2. Species-level (95% ANI cutoff, see the "Methods" section) dereplication of the MAGs produced a set of 269 clusters—or candidate bacterial species (CBS)—each represented by the MAG of highest quality. The CBS were taxonomically classified using GTDB-Tk [28] (see the "Methods" section). While all CBS could be assigned to a known phylum or class, none could be classified into existing species (Table 1). The most common phylum, both in terms of number and abundance of CBS (estimated by the fraction of mapped reads, Fig. 1e, f, Supplementary Table 3), was Actinobacteria with 101 CBS (median percentage of mapped reads 27.2%, IQR 29.5%), followed by Chloroflexi and Proteobacteria. The newly assembled MAGs increase by more than 50% the number of representative species in the Genome Taxonomy Database [27] (GTDB) for Jiangellales, Frankiales, Thermomicrobiales, Isosphaerales, Solirubrobacterales, and for the order-level UBA5184 UBA lineage [47] (Supplementary Table 4, Fig. 1g).

Distribution of CBS among Antarctic cryptoendolithic communities

We investigated the distribution of CBS across the wide range of sampled environmental conditions (see Supplementary Table 5). Since CBS could be assembled only in samples where they had a relatively high abundance, we complemented the assembly by direct read mapping on assembled MAGs to assess presence in a given sample. Specifically, we considered a species present in a sample either (i) if an assembled genome assigned to the CBS was recovered from that sample or (ii) if the breadth of coverage of the mapped reads on the CBS representative was ≥ 50% and the ANI between the consensus sequence and the CBS was ≥ 0.95. The results of this procedure were in good agreement with the prediction of the Mash Screen algorithm [35] (see the "Methods" section and Supplementary Table 1). We identified a set of 10 CBS that were present in at least 75% (14/18) of the samples (Fig. 1d, S4, Supplementary Table 6), despite the known low sensitivity of shotgun metagenomics for the characterization of biodiversity in environmental samples [48]. This set defined a "core" of conserved species that were taxonomically classified in two phyla (Actinobacteria and Proteobacteria) and two classes, i.e., Actinobacteria and Alphaproteobacteria (Fig. S4). A member of the order Jiangellales (Actinobacteria), that herein we named "Candidatus *Jiangella antarctica*," was present across all samples (average percentage of mapped reads 1.92%, SD 1.93%, estimated median depth of coverage from 2 to 190), Mash Screen containment p value < 1.47×10^{-21}, (Fig. 1d,h, S8, Supplementary Table 1). Extracting and classifying the nearly full-length 16S from the Ca. *Jiangella antarctica* (1,513 bp), we did not found any significant match both in the Ribosomal Database Project [49] (RDP, "unclassified Actinomycetales") and SILVA [50] (identity of the best hit 92.09%), confirming that this species has not been previously reported. We also detected three less ubiquitous species that were related to the Antarctic *Jiangella* (Fig. 1h, Supplementary Table 6). Moreover, we found that, while all samples host at least one representative of the class Chloroflexia, three samples (SS, TN, US) host the majority of CBS from this class (Fig. 1d).

In the overall, we observed a large degree of variability among samples which appeared to host diverse bacterial assemblages. However, the majority of the CBS were detected only in a small fraction of the samples (Fig. 1d, e S4).

Table 1 Number of identified taxa and classified CBS for each taxonomic rank. While 100% of the CBS could be assigned to a known phylum, only 81% were classified at the genus level and none at the species level

Taxonomic rank	# of taxa	# of classified CBS (%)
Phylum	12	269 (100%)
Class	22	269 (100%)
Order	33	226 (84%)
Family	43	212 (82%)
Genus	28	81 (30%)
Species	0	0 (0%)

Antarctic bacteria cluster in ancient monophyletic groups that evolved long before Antarctica separated from Gondwanaland

For each bacterial order with at least 4 CBS (for a total of 19 orders, 377 MAGs, and 200 CBS), we built a phylogenetic tree including both the MQ and HQ MAGs and reference genomes belonging to the same order from the GTDB database (see the "Methods" section). In order to generate homogeneously sized datasets, we selected sequences from the 19 order-specific datasets including all the Antarctic MAGs plus all their immediate reference sister taxa (as defined from the corresponding RAxML [51] phylogenetic tree), plus

reference representatives of other more distant clades distributed within the tree [37]. The size of the datasets ranged between 46 taxa in the Solibacteriales to 189 taxa for the Corynebacteriales, with most datasets comprising between 50 and 100 taxa. Using a molecular clock approach and available divergence estimates for calibrating the trees [37], we inferred the divergence times of the Antarctic clades from the main tree within each bacterial order. Our phylogenetic and clock analyses indicated that the Antarctic MAGs (red branches in Fig. 2b, c and Supplementary Data) are grouped into ancient monophyletic clades. In some cases, all Antarctic samples form a unique clade within a certain bacterial order, as in Jiangellales, Microtrichales, and UBA5184, while in other cases, we observed a large clustering of Antarctic MAGs interleaved by just one or two reference genomes as in Thermomicrobiales, Solirubrobacterales, Ktedonobacterales, and Isosphaerales. In almost all other orders (e.g., Acetobacterales, Acidobacteriales, Actinomycetales, Corynebacterales, Frankiales), two or more unrelated Antarctic clades are revealed. Only in a few orders such as Sphingomonadales and Actinomycetales, Antarctic MAGs did not form distinct clades. Our divergence estimates indicate that the vast majority of the Antarctic clades are old (green and orange estimates in Fig. 2a). The diversification of the oldest Antarctic clades occurred on average circa 800 Ma, with estimates ranging from 1.2 billion to 410 Ma (Supplementary Table 6). While the oldest Cyanobacteriales and Ktedonobacterales Antarctic clades are Silurian to Devonian (before 410 Ma), the oldest Antarctic clades in all other orders are pre-Cambrian, with most of them originated in the Tonian (1000-720 Ma).

Antarctic species encode functions that distinguish them from known references, but are not specific and common to all Antarctic MAGs

To characterize the set of metabolic functions encoded by the genomes of the Antarctic CBS and identify those that distinguish them from known related species, protein-coding sequences (CDS) have been predicted, clustered together with the CDS of GTDB reference genomes (60% identity, see the "Methods" section), and functionally annotated. We found that, for each CBS, the number of protein-coding genes and the fraction of them with homology to known protein families was usually similar to what was found for GTDB reference genomes of the same order (Fig. 3a, Supplementary Table 8). Moreover, the t-SNE analysis on the 3,836,924 protein clusters showed that the protein profiles are distributed in agreement with the taxonomy at the order level, indicating homogeneous metabolic potential within each order, independently of habitat (Fig. 3b). We could

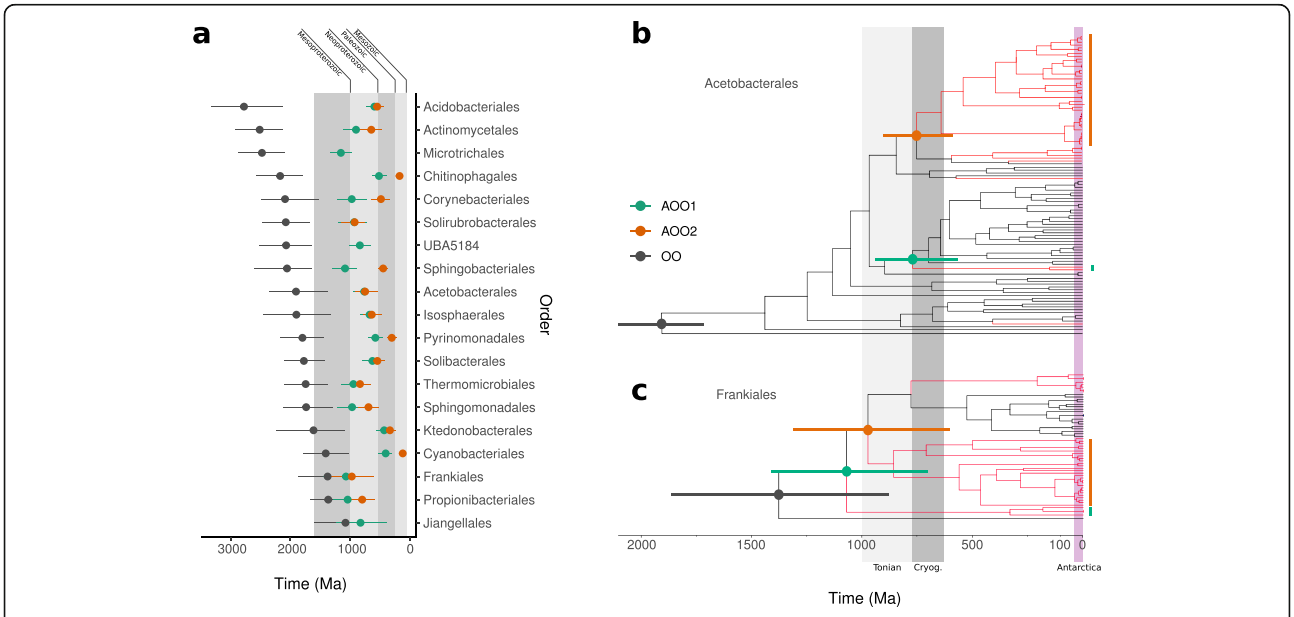

Fig. 2 The neoproterozoic age of Antarctic cryptoendolithic clades. **a** Summary of Bayesian divergence estimates indicate that most of the oldest lineages of Antarctic cryptoendoliths originated during the Neoproterozoic (1000-541 Ma). For each of the orders we show the origin of the order (OO: the split of the order from the closest order), the origin of the oldest uniquely Antarctic clade (AOO1: the split of the Antarctic calde from a non-Antarctic lineage of the same order), and, where present, the origin of the second oldest antarctic clade (AOO2) in each order. **b** Bayesian divergence tree of the Acetobacterales and **c** of the Frankiales with the Antarctic clades highlighted in red. Time-trees for all other orders are in Supplementary Data. The shaded grey areas in **a** identify geological eras, while those in **b** and **c** identify the Tonian and Cryogenian periods of the Neoproterozoic era; pink area in **b** and **c** identify a period compatible with glacial Antarctica. Bars at nodes are uncertainties expressed as 95% of high posterior densities