The melting of the cryosphere is among the most conspicuous consequences of climate change, with impacts on microbial life and related biogeochemistry. However, we are missing a systematic understanding of microbiome structure and function across cryospheric ecosystems. Here, we present a global inventory of the microbiome from snow, ice, permafrost soils, and both coastal and freshwater ecosystems under glacier influence. Combining phylogenetic and taxonomic approaches, we find that these cryospheric ecosystems, despite their particularities, share a microbiome with representatives across the bacterial tree of life and apparent signatures of early and constrained radiation. In addition, we use metagenomic analyses to define the genetic repertoire of cryospheric bacteria. Our work provides a reference resource for future studies on climate change microbiology.
Microorganisms dominate the biosphere, maintain ecosystem processes, and play key roles in global biogeochemical cycles. The microbiome of cryospheric ecosystems, the nearly 20% of Earth’s surface where water remains frozen for at least one month of the year, currently figures among the least understood microbiomes on Earth\textsuperscript{2-4}. This is noteworthy given that the cryosphere is melting at an unprecedented pace owing to climate change. Motivated by the exploration of life in a planetary context\textsuperscript{7} and the discovery of new biomolecules for biotechnology\textsuperscript{8}, classical microbiology and (more recently) advances in sequencing technologies have unravelled physiological and molecular processes underpinning the adaptation of cold-adapted bacteria (i.e., psychrophiles) to the cryospheric environment\textsuperscript{9,10}. More specifically, metagenomics has provided new insights into the structure and function of complex microbial communities of some cryospheric ecosystems, such as permafrost soils\textsuperscript{2,11}, leading to a better understanding of the role of these ecosystems in global biogeochemical cycles and their vulnerability to climate change.

However, we are still missing an integrative understanding of the microbiome across the various and often underexplored cryospheric ecosystems on Earth\textsuperscript{3,5,6}. Here we present a global catalogue of microorganisms from various cryospheric ecosystems and at a taxonomic resolution that allows detection of cryosphere-adapted lineages and associated traits. We leverage sequence data from numerous published studies ranging from snow to permafrost ecosystems to shed light on the global cryospheric microbiome. While also illuminating geographical biases and underexplored habitats in the currently available cryospheric data, our study constitutes an important resource for the study of cryospheric life in general and its potential future in a warmer world.

**Results and discussion**

**The dataset.** We curated and explored 695 published 16S rRNA gene samples from cryospheric ecosystems (Methods section and Supplementary Table 7), including polar ice sheets, mountain glaciers and their proglacial lakes, permafrost soils and the coastal ocean under the influence of glacier runoff, and compared these to 3552 published 16S RNA gene samples from non-cryospheric ecosystems, including temperate and tropical lakes and soils (Supplementary Table 7). This approach allowed us to identify and explore features specific to the cryospheric microbiome and compare it to other environmental microbiomes. However, we note a geographical bias towards polar regions in current publicly available repositories, and the paucity of alpine samples specifically highlights the need to further characterise these habitats given that they are among the most endangered cryospheric ecosystems globally. This bias is further compounded by the inconsistent methodologies applied across studies (e.g. primer pairs and sequencers used). To account for potential primer biases, we analysed two 16S rRNA primer pairs (Primer Pair 1, PP1: 341f-785r; Primer Pair 2, PP2: 515f-806r)\textsuperscript{12,13} commonly used in amplicon high-throughput sequencing. In total, this dataset contains 241,502,708 paired sequence reads, resulting in 530,254 and 410,931 amplicon sequence variants (ASVs) for PP1 and PP2, respectively. Moreover, all taxonomic analyses were performed at the genus level, to account for the limitations of 16s rRNA amplicon data. To gain deeper insights into the functional space of the cryospheric microbiome, we compared 34 published metagenomes from cryospheric ecosystems with 56 metagenomes from similar but non-cryospheric ecosystems (Fig. 1A). Given the difficulty of obtaining high-quality metagenomes from cryospheric ecosystems, we restricted our analyses to glacier surfaces, ice-covered lakes, and Antarctic soils. Although our analyses were limited to samples where raw sequence data are available (Methods section), the breadth of habitats covered are representative of the most abundant cryospheric ecosystems, e.g., glacier ice, cryoconites, subglacial lakes and sea ice. On the other hand, several niches such as glacier snow, glacier-fed rivers/streams, and the full-breath of permafrost may not entirely be represented due to data unavailability. We reanalysed all metagenomes using the same bioinformatic pipeline (IMP3; see Methods section) to avoid analytical biases. Overall, the metagenomic analyses from 2,427,818,072 paired reads yielded 41,068,842 gene sequences. Thus, we here present a catalogue representing a snapshot of the functional diversity in the cryospheric microbiome, integrating across diverse habitats. This represents what we believe to be the first global overview of the functional repertoire of the Earth’s cryosphere compared to other ecosystems.

**A cryospheric microbiome.** Given the communal constraints imposed by the harsh environment of cryospheric ecosystems (e.g., low temperature, oligotrophy), we expected them to harbour a specific microbiome. Accordingly, machine-learning classification (logistic regression models, Methods) based on community composition was able to differentiate between cryospheric and non-cryospheric microbiomes with high accuracy (balanced accuracy >0.96, Supplementary Table 1). Both primer pairs consistently yielded a high classification accuracy and especially a high precision. Interestingly, many of the discriminating cryospheric ASVs were spread widely across the bacterial tree of life (Fig. 1A and Supplementary Fig. 1).

The notion that a part of the microbiome is specific to the cryosphere is also strongly supported by phylogenetic analyses of the 16S rRNA gene amplicon dataset. First, we found higher pairwise phylogenetic overlap among cryospheric samples than among cryospheric/non-cryospheric or non-cryospheric samples (Sorensen’s index, Fig. 1C; Wilcoxon test, Holm adj. \(p < 0.001\)). This points towards a phylogenetic diversity that is specific to the cryosphere. Second, when we examined cross-sample nearest taxon distances (\(\beta\)-NTDs), we found that taxa in cryospheric samples have lower \(\beta\)-NTDs in other cryospheric samples than in non-cryospheric samples (Fig. 1D; Wilcoxon test, Holm adj. \(p < 0.001\)). This was less evident for non-cryospheric samples (Supplementary Table 2). Because phylogeny and functional similarity usually correlate at short phylogenetic distances\textsuperscript{14}, this finding suggests higher niche similarity for cryospheric bacteria compared to their non-cryospheric equivalents. This evokes specific selective constraints of cryospheric environments acting on taxa across the entire bacterial tree of life. Interestingly, when we further examined radiation patterns, we found that taxa in a given cryospheric microbial community had on average larger phylogenetic distances (\(\alpha\)-MPD) than their counterparts in a non-cryospheric community (linear model, \(p < 0.001\)). This could suggest early radiation events with subsequent “pruning” of phylogenetic diversity, which would explain the observed patterns\textsuperscript{15}. However, we cannot exclude nor disentangle the action of contemporary evolutionary and assembly processes that can jointly shape community phylogenies. For example, transduction and genome plasticity have repeatedly been linked with cold adaptation in cryospheric bacteria. Moreover, horizontal gene transfer has also been shown to promote the diffusion of cold-adaptation genes\textsuperscript{16}. Nevertheless, given the large number and breadth of included cryospheric ecosystems, we posit that the topologies of the inferred phylogenies are less prone to assembly processes. We rather interpret that the observed patterns are signs of early and constrained radiation in the cryospheric microbiome. Collectively, these results point to similar evolutionary trajectories in cryospheric microbiomes, probably owing to similar environmental conditions across various cryospheric ecosystems, over timescales, relevant for bacterial macroevolution.
The abundance of a given species in an ecosystem generally reflects its fitness and adaptive capacity to the respective environmental conditions. Therefore, we explored patterns of differential abundance (Methods section) and found 589 bacterial genera with higher abundances in cryospheric compared to non-cryospheric samples (Ancom, W statistic > 0.7, CLR mean difference > 0) that hereafter will be referred to as cryospheric genera. These genera were distributed widely across the bacterial tree of life and encompassed 46 different phyla. Despite this wide distribution, we found that 34.8% and 13.4% of the cryospheric genera were affiliated Proteobacteria and Bacteroidota, respectively (Fig. 2B). The relevance of Proteobacteria is in line with the high prevalence of Alpha- and Gammaproteobacteria typically reported in the cryospheric literature. Genera belonging to the Alpha- and Gammaproteobacteria classes displayed the highest differential abundance and included Sphingomonas, Polaromonas, Rhodofex, Brevundimonas, and Acidiphilum (Fig. 2B) — some of them with taxa typically reported to be psychrophiles. Bacteroidota was the second most important phylum of the cryospheric microbiome with Hymenobacter, Ferruginibacter, and Polaribacter (for instance) as dominant genera, all of which are known from permafrost soils and ice ecosystems. Furthermore, as previously reported, the cryospheric genera included members of the Actinobacteria, Chloroflexi and Cyanobacteria phyla, alongside some Firmicutes. The former two are particularly common in supraglacial environments, and Cyanobacteria are important components of
cryoconite microbiores. Our global analyses thus corroborate and extend previous reports on microbiome composition in distinct cryospheric ecosystems. Furthermore, our differential abundance analysis unveiled genera (e.g., *Oryzihumus* or *Pseudolabrys*) that have not been previously associated with the cryosphere (Fig. 2B). More importantly, many of the detected cryospheric genera only have placeholder names due to the lack of cultivated representatives (e.g., CL_500-29_marine_group, hgccl_clade, TRA3-20), underlining unique bacterial groups that are yet to be described. Collectively, these findings unveil an unexpectedly diverse and likely well-adapted microbiome specific to the cryosphere, and supports the notion of the cryosphere as a biome with its distinct association of microorganisms, alongside plants and animals.

**Compositional patterns across cryospheric ecosystems.** We next explored how microbial community composition varies across cryospheric ecosystems. Using similarity analyses, we found that the microbiome composition differed significantly between cryospheric ecosystem types (PERMANOVA, $R^2 = 0.183$, $p < 0.001$; pairwise adonis, $p < 0.001$ for all pairwise comparisons) (Fig. 3A and Supplementary Table 4). Most conspicuous was the segregation of snow/ice and marine communities, bracketing freshwater and terrestrial cryospheric ecosystems. We also found a significant but relatively small effect of the primer pair (PERMANOVA, $r^2 = 0.019$, $p < 0.001$) that could be attributable to primer bias, or inherent differences related to sampling. To further assess these distributions, we explored prevalence patterns to identify a core microbiome suggesting that snow and ice ecosystems including supraglacial habitats (e.g., mountain glaciers, ice sheets, snow and cryoconites) may serve as a source of cold-adapted bacterial diversity, upon losing which the downstream diversity may become threatened as well.

**Functional potential of the cryospheric microbiome.** The adaptive and survival strategies of microorganisms to the extreme environmental conditions of the cryosphere have received substantial attention over the last years. For example, genomic insights from bacterial cultures have revealed mechanisms of thermal adaptation linked to bulk genomic features, such as GC content and genome size. Moreover, genome streamlining has been shown to be a relevant evolutionary force in the cryosphere. Therefore, we analysed the GC content and genome size of 13,414 reference genome sequences from the NCBI Refseq genomes database to investigate shared properties of cryospheric genera, and to provide a framework to contrast future cryospheric metagenomic results. By comparing these reference genomes representing 660 bacterial genera present in our taxonomic classification. Source data are provided as a Source Data file.
difference = 8.8%) compared to the other genera; a pattern also supported by an enrichment in sequences that encode GC-rich amino acids (e.g., Alanine, Arginine, Glycine) (Supplementary Fig. 3A and Supplementary Table 6). Therefore, our findings suggest that cryospheric genera indeed share an elevated GC content31, in line with reports on cold-adapted Synechococcus (SynAce01)32 and Actinobacteria33. We also report that the average genome size of cryospheric genera is closely bracketed by published values for psychrophilic bacteria34.

Next, using a gene-centric approach, we explored the functional space of the cryospheric metagenomes dataset. Out of 17,191 KEGG orthologues (KO), 980 KO were significantly enriched in cryospheric samples. Cryospheric genera and particularly cryospheric core members (e.g., Pseudomonas, Sphingomonas and Novosphingobium) disproportionately accounted for these gene families (Fig. 4A). Our analysis highlighted the relevance of chemolithotrophic pathways (e.g., manganese and iron uptake, sulfur, nitrogen and hydrogen metabolism), complementing earlier reports on these particular functional attributes of cryospheric ecosystems (Fig. 4B)24,35,36. The apparent relevance of chemolithotrophic pathways is likely attributable to a relative scarcity of organic carbon in cryospheric ecosystems. Interestingly, we consistently identified chitinase genes, which are involved in permafrost carbon cycling, but may also be an adaptation to freezing37. Finally, genes involved in adhesion, motility and various secretion systems collectively point to biofilm formation as an important strategy for life in cryospheric ecosystems38, which are often characterised by extended periods of oligotrophy and elevated UV-radiation.

Our cross-ecosystem metagenomic analyses not only shed light on potential functions of the cryospheric microbiome across ecosystems, but also unveiled a large uncharacterised functional space with 43.4% of the protein coding genes in cryospheric samples unannotated to a KEGG orthologous group. While this does not seem unusual for environmental metagenomes in general3, it is notable that we may lose this functional potential as the cryosphere vanishes. In order to shed light on this uncharacterised functional space, we clustered 41,068,842 gene sequences based on a 30% sequence similarity and 80% sequence coverage threshold, subsequently mapping representative sequences of the largest clusters (>29 sequences in at least 2 samples, n = 12,125) to the UniProt TrEMBL database (Fig. 4C).

While the KEGG assigned clusters overall had a high percentage of sequences that matched genes in the UniProt database (Table 1), we found that cryosphere specific sequences show a large decrease in the clusters assigned to multiple KEGG (i.e., ambiguous) and even more in the ones containing exclusively unassigned sequences, compared to non-cryospheric environmental metagenomes. In addition to the low percentage of gene sequences matching UniProt sequences, we found that the cryosphere specific clusters that align to the database show a largely decreased identity with the matching sequence (Supplementary Fig. 4). These findings underline the lack of representation of cryospheric sequences in current gene sequences databases, potentially linked to the specificity of certain taxa to the cryosphere, and/or functions. Finally, the large nucleotide similarity within these clusters (Supplementary Fig. 4) suggests that these are conserved functions of particular importance to...
microbial life under cryospheric constraints, and corroborates the notion of specific lineages of closely related taxa to dominate microbial life in the cryosphere. Aside from being uncharacterised, 170 of the unassigned gene clusters were only detected in cryospheric metagenomes and could thus represent unknown gene families of importance to understanding the adaptation of bacteria to these extreme ecosystems.

Collectively, our insights both at the taxonomic and functional level reveal key microbiome features that are exclusive to cryospheric ecosystems. Although entire taxonomic lineages are not unique to cryospheric ecosystems, it is evident that specific species and potentially strains are novel and adapted to these environments. Similarly, the emergent functional properties shaped by sustained evolutionary forces, suggesting an ancient origin of this biodiversity. While our study highlights key taxonomic groups such as Proteobacteria and Bacteroidota, our findings also disclose the importance of yet-uncultured bacteria and an uncharacterised genetic repertoire. In light of the threatened nature of the cryosphere, targeted efforts to unravel the phylogenetic and genomic underpinnings of bacterial adaptation to cryospheric ecosystems, including prospecting for cold-adapted biomolecules as well as the cultivation of cryospheric bacteria, are urgently required.

Methods

16S rRNA datasets. Two primer pairs typically used in microbial ecology targeting the prokaryotic 16S rRNA were assessed: 341F–785R and 515F–806R. They will be referred to as Primer Pair 1 (PP1) and Primer Pair 2 (PP2), respectively. All articles citing the PP1 and PP2 reference articles were retrieved using Web of Science (All databases, searched on the 7 December 2019, 1727 articles). The first selection based on title and abstract was performed as described herein. Only studies having sequenced environmental samples were kept. Simultaneously, studies not having sequenced environmental samples were removed integrally two studies in the PP1 dataset (all samples discarded as errors). Only sequences assigned to bacterial taxa were filtered as follows: First, Trimmomatic was used to remove low quality reads, truncating the reads at the first instance of a sliding-window (5 bp) having a mean negative value in quality below 15. At this stage, the raw data from each BioProject was imported into qiime2. Denoising was performed with the dada2 plugin using the primers’ quality and having available data were kept; and the corresponding NCBI bioproject entries were imported. At a later stage, a few studies meeting the filtering criteria but not included in the Web of Science search were added.

The raw sequencing (fastq) data were downloaded using the ENA browser (European Nucleotide Archive: www.ebi.ac.uk/ena/browser/). At this stage, only the control samples were downloaded for experimental studies. The read files were filtered as follows: First, Trimmomatic was used to remove low quality reads, truncating the reads at the first instance of a sliding-window (5 bp) having a mean quality below 15. At this stage, the raw data from each BioProject was imported into qiime2. Denoising was performed with the dada2 plugin using the primers’ sequences length for the “p-trim-left-f” and “p-trim-left-f” parameters. This step removed integrally two studies in the PP1 dataset (“negative values in quality” and “all samples discarded”) errors. Only sequences assigned to bacterial taxa were

**Table. 1 Description of the gene sequences clustering approach.**

| Annotation | Category | Number of clusters | Uniprot match (%) |
|------------|----------|--------------------|-------------------|
| KEGG       | Cryosphere | 47                | 61.70             |
|            | Shared    | 1663              | 54.18             |
|            | Non-cryosphere | 2325              | 55.14             |
| Ambiguous  | Cryosphere | 113               | 40.71             |
|            | Shared    | 1056              | 52.65             |
|            | Non-cryosphere | 3105              | 54.17             |
| Unassigned | Cryosphere | 170               | 17.65             |
|            | Shared    | 1524              | 5.18              |
|            | Non-cryosphere | 2122              | 46.94             |

Table summarising the 12125 largest gene sequence clusters present in at least two samples. The annotation refers to the assignment of the genes to one KEGG Orthologous group (KO), multiple KOs or unassigned (Ambiguous) and only unassigned (Unassigned). Distribution of assigned (KEGG), ambiguous and unassigned functional gene clusters highlighting the bias against cryospheric gene clusters. Shared refers to the representatives of both categories of samples contained gene sequences in the cluster. The number of clusters is shown, along with the proportion of clusters having a consensus sequence matching the Uniprot database.
kept, and chloroplast and mitochondrial sequences were also removed. Finally, all samples with less than 5000 reads after this initial filtering were removed. Taxonomy for PP1 and PP2 was performed using the ASV extracted from qiime2 'feature-classifier' plugin and the Silva 138 nr99 database.41 First, reads were extracted from the reference sequences using the extract-reads method. For this, the primer sequences were used for the ‘-p-r-primer’ and ‘-p-f-primer’ arguments. The length of the extracted reads was set to min. 250 and max. 450 for the PP1 dataset and min. 200 and max. 400 for the PP2 dataset. A classifier was then created using the fit-classifier-naive-bayes method with the extracted reads and the reference taxonomy. Finally, this classifier was run on the dataset’s sequences using the ‘classify-sklearn’ method to get the sequences taxonomy.42 To keep only high-quality samples, all samples having <75% of their ASVs assigned to the phylum level, and 50% assigned to the genus level were removed. This filtering resulted in 2508 samples and 350,254 ASVs for PP1 and 1739 samples and 410,931ASVs for PP2. The ASV tables and metadata tables for these datasets can be found on Zenodo, under the file names: ‘Data/PP1_table.tsv’, ‘Data/PP2_table.tsv’ and ‘Metadata/PP1_metadata.tsv’ and ‘Metadata/PP2_metadata.tsv’, respectively.

Metagenomic dataset. To address the functional aspect of identified taxa, accession numbers of studies comprising of the following keywords: metagenomics, whole genome shotgun, and environmental, were queried using NCBI's Eidence39. The results were manually curated to select studies from a broad Geographic distribution, yielding a total of 382 datasets. The selection of metagenomic samples was further restricted to raw fastq data, thus precluding the use of samples from MG-RAST since only the metagenome assembly was provided. At the end, all samples still under embargo in accordance with the Joint Genome Institute (JGI; USA) policy, were excluded. From this collection, samples with fewer than 1 million reads or with a quality of reads less than Q25 were removed for a final collection of 91 samples (Fig. 1A). Paired reads were processed using the Integrated MetaGenome platform and the MG-RAST pipeline. The workflow included pre-processing, primer/adaptor removal and trimming followed by an iterative addition. Additionally, functional annotation of genes based on custom databases was performed. This description covers the entire workflow is setup in a reproducible Snakemake format.39 Briefly, after preprocessing the reads, de novo assembly using MEGAHIT (v1.2.6)40 was performed. As the number of reads was provided, the length of extracted reads was set to min. 250 and max. 450 for the PP1 dataset and min. 200 and max. 400 for the PP2 dataset. A classifier was then created using the fit-classifier-naive-bayes method with the extracted reads and the reference taxonomy. Finally, this classifier was run on the dataset’s sequences using the ‘classify-sklearn’ method to get the sequences taxonomy.42

Differential abundance analysis. Using the Silva Taxonomic information,45 ASV raw counts were aggregated to the genus-level using a custom R script, removing the ASVs not assigned taxonomically to the genus-level. Ancorn v2.1 was used on the count data for the differential abundance analysis, using the default W statistic threshold of 0.75. The ‘zero-curt’ parameter was set to 0.003 to consider all bacterial genera present in at least 21 samples (4279 genera). The pairwise comparison was performed with the pd, rpd and mntf functions of the Picante R package.56 Linear models were used to compare the values of phi-PD, PDP, and theta-PD among samples, taking the logarithm of the species richness and the dataset (PP1 and PP2) as a fixed effect.

NCBI Refseq genomes properties. To assess the genome size and GC content of publicly available prokaryote genomes, a non-redundant list encompassing all the genomes for PP1 and PP2 datasets was compiled. Therefore, the list of genomes (prokaryotes.txt) available on NCBI46 was downloaded on March 15th, 2021 from https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS. The prokaryotes list (prokaryotes.txt) available on NCBI was filtered based on the list of genera found in our dataset, simultaneously retrieving the accession IDs. The accessing IDs were used to download the complete bacterial genome sequences using the ncbi-genome-download python package (https://github.com/kblin/ncbi-genome-download). The genome sizes for the downloaded genomes were additionally retrieved from the prokaryotes.txt metadata file. Subsequently, Prodigal was used to annotate the open-reading frames per genome. Obtaining both the general feature format (gff) files and aminoacid fasta files. These input files were filtered to remove genes with growth time in hours) and their codon usage analyses (CURB) using gcodcon and codRdon (https://github.com/BioinfoHRC/codRdon) R package respectively.26,26 The amino acid enrichment analysis was performed on converting the codon counts to amino acids using the R-package Biostrings using DESeq2 with default parameters.47 For the classification to the closest taxonomic level, the best hits based on bit scores were selected. FeatureCounts was then used to extract the number of reads per functional category.

Logistic regression classification of cryospheric bacterial communities. The Logistic regression implemented in scikit-learn python module (version 0.23.2) was trained on presence-absence ASV tables to classify cryospheric samples.48 To reduce the amount of ASVs considered, the table was reduced based on relative abundance: presence was defined at a 0.005 relative abundance threshold. A stratified 5-fold cross-validation (CV) was ran and the scores were averaged across the CVs. This process was repeated 40 times and the mean and standard deviations are reported for each metric. To ensure reproducibility, the seed was set as 23 for the classifier, and as the iteration number for the stratified cross validation iteration (from 0 to 39). The C parameter controlling L2 penalisation was turned off using the ‘none’ argument and the libsgd was used. ROC curves were plotted using the ‘plot_roc_curve’ function of the scikit-learn python module. Balanced accuracy, precision and recall were computed using the ‘accuracy_score’, ‘precision_score’ and ‘recall_score’ methods, respectively, with sample weights correcting for the sample size of the cryospheric and non-cryospheric dataset (Supplementary Table 1). The best hits and standard deviations of scorings for the classifiers can be found in Table 1. Odds ratios were calculated using the exponent of the logistics models coefficients. The tables containing the ASVs logarithms odds ratios can be found in the Data folder available on Zenodo under the name ‘PP1_Logistic_costs.csv’ and ‘PP2_Logistic_costs.csv’ for PP1 and PP2, respectively.

Phylogenetic analyses. Phylogenetic trees were built using the set of ASVs found in the dataset used for the logistic regression classification. Due to the different S regions targeted, phylogenies for both PP1 and PP2 datasets were constructed separately. The ASVs sequences were aligned using the FFT-NS-2 algorithm implemented in the Mafft aligner with default parameters.48 The alignments were subsequently trimmed and co-aligned using Timdoll with an affine gap penalty of 0.1 and the primer pair (PP1 and PP2) built using IQ-TREE with the GTR model of nucleotide substitution and the ‘fast’ option.48 Phylogenetic tree visualisations were created using the ggtree and ggteeExtra R packages. Only positive coefficients showing enriched presence in cryospheric environments are shown in the phylogenetic barplots (Fig. 1). The number of ASVs with an odds ratio above 1 was shown for taxonomic summaries (Supplementary Fig. 1B, C).

β-diversity phylogenetic metrics (Sorensen’s Index and β-MNNDT) were computed using the ‘phylossor’ and ‘condist’ functions of the Picante R package, using custom functions to compute pairwise comparisons. For each metric, 50 iterations were performed where we calculated the pairwise distances between and within 50 cryospheric, and 50 non-cryospheric samples. This random sub-sampling approach was chosen to reduce computing time. Kruskal–Wallis tests were used to determine whether the distribution was different across groups, and Wilcoxon tests were used to calculate pairwise post-hoc comparisons. Wilcoxon tests implemented in the compare_means function of the ggpubr R package were used, effects sizes (r) were calculated with the wilcox_effsize function implemented in the statix R package. Sample specific calculations of α-PD (and species richness), α-MPD and α-MNTD were computed using the ‘pd’, ‘rpd’ and ‘mntf’ functions of the Picante R package.56 Linear models were used to compare the values of α-PD, α-MPD, and α-MNTD across samples, taking the logarithm of the species richness and the dataset (PP1 and PP2) as a fixed effect.
four ecosystem types (Supplementary Fig. 2B). The core microbiome presence in the different ecosystem types was shown using an upset-plot using the complex-upset R package. The taxonomic tree available in Supplementary Fig. 2A was created using the *Metacoder* R package. The alpha-diversity was calculated using *Shannon’s* index with a custom R function. To test the difference across ecosystems and datasets, the Wald-type statistics implemented in the ‘GFDD’ function of the R ‘GFDD’ package was used (Supplementary Table 5). This test was performed instead of an ANOVA, as the data was not normally distributed. The mean values given by the function were used for the ecosystem comparison.

**KEGG enrichment.** The standard DESeq2 pipeline with default parameters was used on raw KEGG counts for the enrichment analysis, using the default Wald tests. We considered significantly enriched Kegg Orthologs (KOs) with an FDR adjusted p < 0.01 and a log2-fold-change > 1. To unravel the contribution of these gene families to functional pathways, we ran KEGGdecode on the KOs enriched in cryospheric samples, to identify environmental-associated pathways in all samples.

To understand and unravel the origins of the gene families specific to the cryospheric metagenomes, contigs were taxonomically classified following which the specific gene families were mapped to the contigs. We used Kraken2 to taxonomically assign all the contigs present in the metagenomes followed by custom python scripts (provided) to link the genes belonging to enriched KEGG orthologs (KO and the corresponding NCBI taxon IDs). An R script using the NCBI entrez package was used to retrieve the taxonomy based on the taxon ID, and to get the genus-level taxonomy. To link the Silva genus taxonomies with their NCBI counterparts, the grep function included in R allowing partial matches was used to find Silva genera name matching the NCBI genus name. The DEseq2 results, KEGG-decoder output and taxonomy matches can be found in the Data/folder of the Zenodo under the names ‘KEGG_deseg_results.csv’, ‘KEGG_decoder_output.csv’ and ‘KEGG_sign_tax.genera.csv’, respectively.

**Gene clusters and unassigned protein coding sequences.** Predicted gene sequences annotated to the KEGG database and those unassigned were gathered to identify functional homologues within the dataset. We used Kraken2 to taxonomically assign all the contigs present in the metagenomes followed by custom python scripts (provided) to link the genes belonging to enriched KEGG orthologs (KO and the corresponding NCBI taxon ID). An R script using the NCBI entrez package was used to retrieve the taxonomy based on the taxon ID, and to get the genus-level taxonomy. To link the Silva genus taxonomies with their NCBI counterparts, the grep function included in R allowing partial matches was used to find Silva genera name matching the NCBI genus name. The DEseq2 results, KEGG-decoder output and taxonomy matches can be found in the Data/folder of the Zenodo under the names ‘KEGG_deseg_results.csv’, ‘KEGG_decoder_output.csv’ and ‘KEGG_sign_tax.genera.csv’, respectively.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

The data generated in this study have been deposited in Zenodo, under https://doi.org/10.5281/zenodo.6587400. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Code availability**

All scripts used for analyses, along with the conda environments, and additional information is provided in a Github repository archived on Zenodo: https://doi.org/10.5281/zenodo.6587400.

Received: 9 November 2021; Accepted: 20 May 2022; Published online: 02 June 2022

**References**

1. Fountain, A. G. et al. The disappearing cryosphere: impacts and ecosystem responses to rapid cryosphere loss. *BioScience* 62, 405–415 (2012).
2. Jansson, J. K. & Täg, N. The microbial ecology of permafrost. *Nat. Rev. Microbiol.* 12, 414–425 (2014).
3. Nayfach, S. et al. A genomic catalog of Earth’s microbiomes. *Nat. Biotechnol.* 39, 499–509 (2021).
4. Boetius, A., Ainesio, A. M., Deming, J. W., Mikucik, J. A. & Rapp, J. Z. Microbial ecology of the cryosphere: sea ice and glacial habitats. *Nat. Rev. Microbiol.* 13, 677–690 (2015).
5. Thompson, L. R. et al. A communal catalogue reveals Earth’s multiscale microbial diversity. *Nature* 551, 457–463 (2017).
6. Goordial, J. Cryomicrobial ecology: still much to learn about life left out in the cold. *mSystems* 6, e00852–21 (2021).
7. Merino, N. et al. Living at the extremes: extremophiles and the limits of life in a planetary context. *Front. Microbiol.* 10, 780 (2019).
8. Feller, G. & Gerday, C. Psychrophilic enzymes: hot topics in cold adaptation. *Nat. Rev. Microbiol.* 1, 200–208 (2003).
9. Bowman, J. P. Genomics of Psychrophilic Bacteria and Archaea. In *Psychrophiles: From Biodiversity to Biotechnology* (ed. Margesin, R.) 345–387 (Springer International Publishing, 2017).
10. D’Amico et al. Psychrophilic microorganisms: challenges for life. *EMBO Rep.* 7, 385–389 (2006).
11. Tripathi, B. M. et al. Variations in bacterial and archaeal communities along depth profiles of Alaskan soil cores. *Sci. Rep.* 8, 504 (2018).
12. Caporaso, J. G. et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1621–1624 (2012).
13. Klinworth, A. et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 41, e1 (2013).
14. Dini-Andreote, F., Stegen, J. C., Eales, J. Dvan & Salles, J. F. Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession. *Proc. Natl Acad. Sci. USA* 112, E1326–E1332 (2015).
15. Mazel, F. et al. Influence of tree shape and evolutionary time-scale on fungal phylogenetic diversity metrics. *Ecography* 39, 913–920 (2016).
16. Dorrell, R. G. et al. Within-Arctic horizontal gene transfer as a driver of convergent evolution in distantly related microalgae. *bioRxiv* https://doi.org/10.1101/2021.07.31.454568 (2021).
17. Ainesio, A. M. & Laybourn-Parry, J. Glaciers and ice sheets as a biome. *Trends Ecol. Evol.* 27, 219–225 (2012).
18. Sajid, W., Ali, R., Bahadur, A., Ghimire, P. S. & Kang, S. Bacterial diversity and communities structural dynamics in soil and meltwater runoff at the front of a basihui glacier No.1, China. *Microb. Ecol.* 81, 370–384 (2021).
19. Comte, J., Culley, A. L., Lovejoy, C. & Vincent, W. F. Microbial connectivity and sorting in a High Arctic watershed. *ISME J.* 12, 2988–3000 (2018).
20. Fodelianakis, S. et al. Microdiversity characterizes prevalent phylogenetic clades in the glacier-fed stream microbiome. *ISME J.* 16, 666–675 (2022).
21. Collins, R. E., Rocap, G. & Deming, J. W. Persistence of bacterial and archaeal communities in sea ice through an Arctic winter. *Environ. Microbiol.* 12, 1828–1841 (2010).
22. Kim, B.-C. et al. Polaribacter sejongensis sp. nov., isolated from Antarctic soil, and emended descriptions of the genus Polaribacter, Polaribacter butkevichii and Polaribacter ingersii. *Int. J. Syst. Evol. Microbiol.* 63, 4000–4005 (2013).
23. Kohler, T. J. et al. Patterns in microbial assemblages exported from the meltwater of Arctic and Sub-Arctic Glaciers. *Front. Microbiol.* 11, 669 (2020).
24. Christner, B. C., Kvitko, H. B. & Reeve, J. N. Molecular identification of Bacteria and Eukarya inhabiting an Antarctic cryoconite hole. *Extremophiles* 17, 177–183 (2003).
25. Ainesio, A. M., Lutz, S., Chrismas, N. A. M. & Benning, L. G. The microbiome of glaciers and ice sheets. *NPJ Biofilms Microbiomes* 3, 1–11 (2017).
26. Frey, B. et al. Microbial diversity in European alpine permafrost and active layers. *FEMS Microbiol. Ecol.* 92, fiw018 (2016).
27. Tribelli, P. M. & López, N. I. Reporting key features in cold-adapted bacteria. *Life* 8, 8 (2018).
28. Margesin, R. & Collins, T. Microbial ecology of the cryosphere (glacial and permafrost habitats): current knowledge. *Appl. Microbiol. Biotechnol.* 103, 2537–2549 (2019).
29. Wu, H., Zhang, Z., Hu, S. & Yu, J. On the molecular mechanism of GC content variation among eubacterial genomes. *Biol. Direct* 7, 2 (2012).
30. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35, D61–D65 (2007).
31. Almpanis, A., Swain, M., Gatherer, D. & McEwan, N. Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. *Microbiol. Genomics* 4, e00168 (2018).
32. Tang, J., Du, L.-M., Liang, Y.-M. & Daroch, M. Complete genome sequence and comparative analysis of Synechococcus sp. CS-601 (SynAco01), a cold-adapted cyanobacterium from an oligotrophic Antarctic habitat. *Int. J. Mol. Biol.* 20, 152 (2019).
33. Goordial, J. et al. Improved-high-quality draft genome sequence of Rhodococcus sp. JG-3, a eurypsychrophilic Actinobacteria from Antarctic Dry Valley permafrost. *Stand. Genomic Sci.* 10, 61 (2015).

34. Sabath, N., Ferrada, E., Barve, A. & Wagner, A. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic reorganization during thermal adaptation. *Genome Biol.* 5, 866–977 (2013).
35. Boyd, E. S., Skidmore, M., Mitchell, A. C., Bakermans, C. & Peters, J. W. Methanogenesis in subglacial sediments. *Environ. Microbiol. Rep.* 2, 685–692 (2010).
36. Boyd, E. S. et al. Diversity, abundance, and potential activity of nitrifying and nitrite-reducing microbial assemblages in a subglacial ecosystem. *Appl. Environ. Microbiol.* 77, 4778–4787 (2011).
37. Liu, C., Wang, X., Wang, X. & Sun, C. Acclimation of Antarctic Chlamydomonas to the sea-ice environment: a transcriptomic approach. *Extremophiles* 20, 437–450 (2013).

Smith, H. J. et al. Biofilms on glacial surfaces: hotspots for biological activity. *Nat. Biotechnol.* 33, 852–857 (2016).
39. Bolger, A. M., Lohse, M. & Usadel, B. Trimomatic: a flexible trimming tool for Illumina sequence data. *Bioinformatics* 30, 2114–2120 (2014).
40. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857 (2019).
41. Bokulich, N. A. et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2’s q2-feature-classifier plugin. *Microbiome* 6, 9 (2018).
42. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596 (2013).
43. Winter, D. J. *rensieve*: An R package for the NCBI eUtils API. (PeerJ Preprints, 2017).
44. Narayanasamy, S. et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* 17, 260 (2016).
45. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2352–2352 (2012).
46. Li, R. D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assemblies via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676 (2015).
47. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2086–2089 (2014).
48. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. * BMC Bioinform.* 11, 119 (2010).
49. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* 7, e1002195 (2011).
50. Heintz-Buschatz, A. et al. Integrated multi-omics of the human gut microbe in a case study of familial type 1 diabetes. *Nat. Microbiol.* 2, 1–13 (2016).
51. Kanehisa, M. & Goto, S. KGKG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30 (2000).
52. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient and accurate read mapping and counting method for high-throughput sequencing reads. *Bioinformatics* 32, 2858–2869 (2017).
53. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).
54. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780 (2013).
55. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973 (2009).
56. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274 (2015).
57. Yu, G., Lam, T. T.-Y., Zhu, H. & Guan, Y. Two methods for mapping and visualizing associated data on phylogeny using Gttree. *Mol. Biol. Evol.* 35, 3041–3043 (2018).
58. Xu, S. et al. ggtreeExtra: compact visualization of richly annotated phylogenetic data. *Mol. Biol. Evol.* 38, 4039–4042 (2021).
59. Kembel, S. W. et al. Picatree: R tools for integrating phylogenies and ecology. *Bioinformatics* 26, 1463–1464 (2010).
60. Liao, F. Third release of ANCOM. *Zenodo* https://doi.org/10.5281/zenodo.3577802 (2019).
61. Foster, Z. S. L., Sharpont, T. J. & Grinnwald, N. J. Metacoder: an R package for visualization and manipulation of community taxonomic diversity data. *PLoS Comput. Biol.* 13, e1005404 (2017).
62. Weissman, J. L., Hou, S. & Fuhrman, J. A. Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proc. Natl. Acad. Sci. USA* 118, e2016810118 (2021).
63. Elek, A., Kuzman, M. & Vlahoviček, K. coRdon: codon usage analysis and prediction of gene expression. https://github.com/BioinfofHR/coRdon. *R Package Version 1* (2019).
64. Okozañ, J. et al. Package `vegan`. *Community Ecol. Package Version 2*, 1–295 (2013).
65. Martínez Arbizu, P. pairwiseAdonis: Pairwise multilevel comparison using adonis. *R package version 0.4* https://github.com/pmartinezarbizu/pairwiseAdonis (2020).
66. Conway, J. R., Lee, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940 (2017).
67. Shannon, C. E. & Weaver, W. *The Mathematical Theory Of Information*. Vol. 97 (University of Illinois Press, Urbana, 1949).
68. Love, M. L., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550 (2014).
69. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J.* 12, 1861–1866 (2018).
70. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257 (2019).
71. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* 40, D136–D143 (2012).
72. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028 (2017).
73. Bairoch, A. et al. The universal protein resource (UniProt). *Nucleic Acids Res.* 33, D154–D159 (2005).
74. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948 (2007).

**Acknowledgements**

This work was supported by The NOMIS Foundation (Vanishing Glaciers) to T.J.B.; S.B.B. was supported by a Swiss National Science Foundation (CRSII2_180241) grant to T.J.B. and P.W. We extend our gratitude to Laura de Nies, Patrick May, Cedric Laczym, and Valentina Galata for their advice on metagenomic analyses. Computational work was carried out using the HPC facilities of the University of Luxembourg.

**Author contributions**

M.B.: conceptualisation, methodology, data curation and writing – original draft preparation, visualisation; S.B.B.: conceptualisation, methodology, investigation, data curation and writing – original draft preparation; S.B.B.: conceptualisation, methodology, investigation, data curation and writing; H.P.: conceptualisation, methoddology, data curation and writing; T.K.: data curation and writing; L.E.: methodology; G.M.: methodology; P.W.: conceptualisation and supervision; T.J.B.: conceptualisation, writing – original draft preparation, data curation, supervision and funding acquisition.

**Competing interests**

The authors declare no competing interests.

**Additional information**

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-30816-4.

**Correspondence** and requests for materials should be addressed to Massimo Bourquin or Tom J. Battin.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.