https://doi.org/10.5194/essd-2023-395
Preprint. Discussion started: 1 December 2023

Open Access Earth System Science Data Discussions

A database of glacier microbiomes for the Three Poles

Yongqin Liu[1,2,3]*, Songnian Hu[3,4], Tao Yu[1,3,4], Yingfeng Luo[3,4], Zhihao Zhang[2,3], Yuying Chen[2,3], Shunchao Guo[4,5,6], Qinglan Sun[4,5,6], Guomei Fan[4,5,6], Linhuan Wu[4,5,6], Juncai Ma[4,5,6], Keshao Liu[2], Pengfei Liu[1], Junzhi Liu[1], Mukan Ji[1]*

[1]Center for Pan-third Pole Environment, Lanzhou University, Lanzhou, China
[2]State Key Laboratory of Tibetan Plateau Earth System, Resources and Environment (TPESRE), Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
[5]Microbial Resource and Big Data Center, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
[6]Chinese National Microbiology Data Center (NMDC), Beijing, China

Correspondence to: Yongqin Liu (yql@lzu.edu.cn), Mukan Ji (jimk@lzu.edu.cn)

Abstract. Glaciers cover 10% of Earth's land area and are a pool of carbon and nitrogen for downstream ecosystems. Microbes, including bacteria, fungi, algae, and other microeukaryotes, are the primary inhabitants of glacier ecosystems and are key drivers of carbon and nitrogen transformation. Here, we present a dataset on supraglacial bacteria and archaea (referred to as microorganisms hereafter) communities across Antarctic, Arctic, and Tibetan glaciers. The dataset comprises 815 amplicon sequencing data, 952 cultured bacterial genomes data, and 208 metagenome data, covering ice, snow, and cryoconite habitats. The dataset contains 67,224 amplicon sequencing phylotypes, with a higher microbial diversity in the Tibetan glaciers than in the Antarctic and Arctic glaciers, which were respectively enriched with Gammaproteobacteria, Bacteroidota, and Alphaproteobacteria. Additionally, 2,517 potential pathogens were identified, accounting for 1.9% of the total microorganisms identified. Snow and ice exhibited a higher relative abundance of pathogens than cryoconite, which could be attributed to the similar adaptation mechanisms for microbial survival in aerosol and immune evasion. The dataset contains 62,595,715 unique genes and 4,327 microbial genomes, a 34% expansion from previous publications. Genes were annotated for those associated with carbohydrate-active enzymes, nitrogen cycling, methane cycling, antimicrobial resistance, and microbial virulence, revealing the dynamic microbial functions in glacial habitats. This comprehensive dataset provides standardized microbial diversity, taxonomy, community structure, and genetic functions of glacial microbiomes. The data can be leveraged to elucidate ecological principles governing the distribution of microorganisms, to gain insights into the key functional genes for supraglacial microbiomes, to build mechanistic models, and to identify any potential biohazards for policymakers to make informed decisions regarding climate change. The dataset is available at the National Tibetan Plateau Data Center (https://doi.org/10.11888/Cryos.tpdc.300830, Liu et al., 2023)

1 Introduction

Glaciers cover 10% of Earth's land area (Cauvy-Fraunié and Dangles, 2019) and are mainly distributed in the Antarctic, Arctic, and Tibetan Plateau (the Three Poles) (Qiu et al., 2008). Glaciers store approximately three-quarters of Earth's freshwater (Boetius et al., 2015) and are also a pool of carbon and nitrogen. The six Pg of carbon stored in global glaciers is readily released into downstream ecosystems with glacier runoff (Hood et al., 2015), influencing key elemental cycling in downstream ecosystems. Before carbons and nitrogen are released, they undergo extensive biological transformation (Guo et al., 2022), primarily microbial-driven. Microbes, including bacteria, fungi, algae, and other microeukaryotes are the main habitant of glacier ecosystems (Cauvy-Fraunié and Dangles, 2019). These microorganisms employ strategies to survive the glacial conditions, such as strong UV radiation, low temperature, and low carbon and nitrogen nutrients (Ciccazzo et al., 2016). As the key driver of carbon and nitrogen transformation in glacier ecosystems, knowledge of microbial biogeography and functions can greatly enhance our understanding of the biogeochemical cycling in glacial ecosystems and aid in predicting the impact of climate change.

The glacier as a habitat is not homogeneous and is divided into supraglacial, englacial, and subglacial ecosystems. Of these, the microorganisms in supraglacial ecosystem are the most active, due to its exposure to external environment and ambient temperature. Supraglacial ecosystems can be further separated into the snow, ice, and cryoconite hole (cylindrical depressions formed by the preferential melting of dark debris into the surface, typically comprises surface water and cryoconite at the bottom) (Cook et al., 2016), each of which has distinct microbial composition (Anesio and Laybourn-Parry, 2012). Algae and Cyanobacteria are the primary producers in supraglacial ecosystems, with other heterotrophic microorganisms participating in the transformation and degradation of endogenous and exogenous nutrients. Active metabolism is reported in glacial ecosystems; for instance, cryoconite is a source of methane but a sink of carbon dioxide, with a rate of 4.60 μmol $m^{-2}d^{-1}$ and −1.77 μmol $m^{-2}d^{-1}$, respectively (Zhang et al., 2021). Furthermore, organisms with photosynthesis, nitrification, and denitrification functions are also widespread in glacier cryoconite (Cameron et al., 2012; Stibal et al., 2020).

It was estimated that the mean microbial abundance in glacier surface meltwater is 10^4 cells mL^{-1} (Stevens et al., 2022), this quantity may further increase with the enhanced glacier retreat. Some of these naturally occurring microorganisms are known as emerging contaminants, which are not commonly monitored in the environment but have the potential to enter the environment and cause known or suspected adverse ecological and/or human health effects (Taheran et al., 2018). A previous study cultivated hemolytic bacteria from Spitsbergen glacier meltwater with potential pathogenicity (Mogrovejo-Arias et al., 2020). Other emerging contaminants in glaciers such as antibiotic resistance genes and microbial virulence factors have also received increased attention (Mao et al., 2023)

Here, we present a glacier dataset on supraglacial bacteria and archaea (referred to as microorganisms hereafter) for the Three Poles. This dataset includes amplicon sequencing data from 815 samples, 952 cultured bacterial genomes, as well as shotgun metagenomic sequencing from 208 samples. From an ecological perspective, this dataset with standardized microbial diversity, taxonomy, and community structure can improve understanding of the ecological principles governing the distribution of microorganisms across glaciers, as well as their partitioning across the various habitats in the supraglacial ecosystem. From a geochemical cycling perspective, the database can provide insights into the key functional genes for supraglacial microbiomes, which can be used to better comprehend carbon and nitrogen cycling and allow the building of a model to anticipate glacial carbon and nitrogen dynamics in the future. The dataset archives glacial-specific microorganisms and unique genes in digital form, thus representing an invaluable alternative method for preserving biodiversity. Additionally, the dataset can be employed to identify any potential biohazards (pathogens and emerging contaminants) of glaciers and evaluate the impact of glacier melting on downstream ecosystems from a biosafety perspective, thereby assisting policymakers in making informed decisions regarding climate change.

2 Materials and methods

2.1 Data acquisition

Amplicon data: 485 amplicon sequencing results are first time released in the present work, while the rest were downloaded from NCBI Short Read Archive based on keyword search by terms “Antarctic”, “Arctic”, or “Tibetan Plateau”. After careful manual curation, any samples that are not from ice, snow, or cryoconite-related habitats (cryoconite and cryoconite water) were removed (**Table S1**).

Metagenome data: All articles containing the keyword of “glacier metagenome” were retrieved using the Web of Science (searched on the 1st of December 2022). Only studies having sequenced ice, snow, or cryoconite samples with raw sequence data uploaded on NCBI Short Read Archive were kept. Additionally, a few metagenome data without published articles were added from IMG/M database and NMDC database based on keyword search by terms “ice”, “snow” and “cryoconite”. In addition to metagenomes from the Antarctic, Arctic, and Tibetan Plateau, metagenomes from the Andes and Alps were also downloaded. (**Table S2**).

Cultivated bacterial genome data: 883 isolate genome data of Tibetan Plateau glaciers were obtained from the TG2G dataset (Liu et al., 2022) and other 69 isolate genome data of Non-Tibetan Plateau glaciers were downloaded from the NCBI Genome database based on keyword search by terms “Antarctic”, and “Arctic”. After careful manual curation, only samples that from ice, snow, and cryoconite habitats were kept (**Table S3**).

2.2 Amplicon sequencing data processing

Sequencing data were processed using the USEARCH v11 pipeline (Edgar, 2010). Paired-end reads were merged and quality screened with a max expected errors threshold of 0.5, while single-end read was directly quality screen with the same threshold. Sequences from all samples were merged and aligned against SILVA reference alignment (release 128), then were trimmed to common start and end positions, so that all samples are directly comparable. Phylotypes were clustered with 97%

identity and chimeric sequences were identified and removed. The phylotype representative sequences were taxonomically classified using the Bayesian classifier against the Silva database (release 128) (Quast et al., 2012), and then eukaryotic, mitochondrial, and chloroplast sequences were removed. After the phylotype table was constructed, samples were randomly sub-sampled at an equal depth of 10,014. Samples with lower sequencing reads than this threshold and those missing metadata were removed, this ended with 815 samples with a total of 67,224 microbial phylotypes (bacteria and archaea) for downstream analysis. The presence of potential pathogens was identified by comparing the 16S gene sequences against the bacterial pathogens database (Wardeh et al., 2015) using BLAST (Mcginnis and Madden, 2004) with the thresholds of 97% identity and 100 % coverage.

The Shannon diversity, richness (number of phylotypes), and evenness indices were calculated from the rarefied phylotype table using Primer-E V6 (Clarke and Warwick, 2006). The alpha diversity indices (richness, evenness, and Shannon diversity) and the relative abundance of dominant taxonomy lineages were compared using Kruskal-Wallis one-way ANOVA by region (Antarctic, Arctic, and Tibetan Plateau) and habitats (snow, ice, cryoconite, and cryoconite water), multiple testing was performed based on the Dunn's post-hoc test using FSA package in the R environment (Ogle et al., 2022).

The community structure variations were visualized using an NMDS ordination plot based on the Hellinger-transformed Bray-Curtis distance matrix. Permutational analysis of variance (PERMANOVA) was used to test the significance of community differences in samples by region and habitat (Anderson, 2017) using Primer-E V6 (Clarke and Warwick, 2006) with 999 permutations.

Core phylotypes were defined as occurring in more than 55% of the samples in each habitat-region pair. If a phylotype was identified as a core phylotype for all habitats of a region, then it was designated as the core phylotype for the region. This classification was modified from Delgado-Baquerizo et al. (2018), so that the dominant phylotype designation is less affected by the unbalanced samples for each habit-region pair.

2.3 Metagenome data processing

Metagenome data processing has been described previously (Liu et al., 2022). Briefly, it includes raw data quality filtering, assembly, open reading frames prediction, and genome binning. Gene open reading frames (ORFs) for the metagenomic assemblies were predicted using Prodigal (Hyatt et al., 2010), and dereplicated by clustering at 80% aligned region with 95% nucleotide identity using MMseqs2 (Steinegger and Söding, 2017) with parameters: easy-linclust -e 0.001, --min-seq-id 0.95, -c 0.80.

Metagenomic assemblies were binned using MetaBAT 2 (v2.12.1) (Kang et al., 2019), MaxBin 2 (v2.2.7) (Wu et al., 2016), and VAMB (v2.0.1) (Nissen et al., 2021) respectively. The resulting bins (or MAGs) were then refined using RefineM (v0.0.20) (Parks et al., 2017) by removing contigs with divergent GC content, coverage, or tetranucleotide signatures. Then only MAGs meeting the medium and higher quality of MIMAG (Bowers et al., 2017a) were retained (completeness > 50%, contamination <10%). These MAG together with the downloaded isolate genomes were dereplicated using the thresholds of 30% aligned fraction and a genome-wide average nucleotide identity (ANI) threshold of 95%, they were then taxonomically annotated using the Genome Taxonomy Database Toolkit (GTDB-Tk, v0.3.2) (Chaumeil et al., 2019) against the GTDB release R06-RS202.

2.4 Gene function annotation

The functions of the dereplicated genes were also annotated using eggNOG-mapper (Huerta-Cepas et al., 2017) and the eggNOG Orthologous Groups (OGs) database (v5.0) (Huerta-Cepas et al., 2019). This includes the KEGG functional orthologs (Kanehisa et al., 2017), the carbohydrate-active enzymes database (CAZy) (Levasseur et al., 2013), and the COG categories (Tatusov et al., 2003). Antibiotic resistance genes (ARGs) were annotated against the Comprehensive Antibiotic Resistance Database (CARD) (Jia et al., 2017) and Resistance Gene Identifier (RGI v3.1.4) (Alcock et al., 2020) with the loose model (--include_loose). Virulence factors were annotated by aligning gene sequences against the Virulence Factors Database (VFDB 2019) (Liu et al., 2019) with DIAMOND blastp (Buchfink et al., 2021) (e-value threshold of 1e-5).

3 Results

3.1 Amplicon-based dataset

A total of 815 glacier-related samples were retained after quality filtering (**Fig. 1a** and **Table S1**), comprising 517 from cryoconite-related habitats (cryoconite and cryoconite water), 184 from snow, and 114 from ice. Spatially, 69.7% of all samples (n = 568) were from Tibetan glaciers, 24% (n = 196) were from Antarctic glaciers, while those from Arctic glaciers were slightly under-represented (6.3%, n = 51).

3.1.1 Microbial diversity

The amplicon sequencing dataset comprised 67,224 phylotypes, Spatially, Tibetan glaciers exhibited a significantly higher microbial richness than Arctic and Antarctic glaciers (Kruskal-Wallis One-way ANOVA Dunn's post-hoc analysis, $P < 0.05$, **Figs. 1b** and **Table S4**), but not significant difference detected between Arctic and Antarctic glaciers ($P = 1.00$), In comparison, the richness was similar among different habitats ($P = 0.738$**, Fig. 1c**). In Tibetan glaciers, microbial richness in cryoconite was significantly greater than those in snow and ice (**Fig. S1**). This pattern, however, was distinct from that observed in Arctic glaciers, wherein cryoconite had lower richness compared to snow, which is consistent with a report previously (Franzetti et al., 2017).

Spatially, the evenness was significantly higher in Antarctic glaciers across all samples (both $P < 0.001$, **Fig. S2a)**. Across different habitats, the microbiome of cryoconite displayed significantly higher evenness in comparison to cryoconite water, snow, and ice ($P < 0.001$, **Table S4**, and **Fig. S2b**). The higher evenness in cryoconite was observed in both Tibetan and Arctic glaciers, but not in the Antarctic glaciers (**Fig. S3**).

https://doi.org/10.5194/essd-2023-395
Preprint. Discussion started: 1 December 2023

3.1.2 Microbial taxonomy composition in the glaciers of the Three Poles

We identified 50 bacterial and archaeal phyla across the dataset. The glacier microbiomes across the three poles had similar microbial taxonomy composition, dominated by Gammaproteobacteria, Bacteroidota, Alphaproteobacteria, Cyanobacteria, Actinobacteria, and Firmicutes (**Fig. 1d** and **Tables S5**). Cryoconite was enriched with Cyanobacteria, but had lower relative abundances of Gammaproteobacteria and Alphaproteobacteria; conversely, cryoconite water was enriched with Bacteroidota (**Table S6**). Snow samples had a high relative abundance of Crenarchaeota, suggesting a potential for nitrification capacity, which is consistent with the wet deposition (snowfall) being a major source of nitrogen for glacier ecosystems (Telling et al., 2011). Spatially, Tibetan glaciers had a significantly lower relative abundance of Actinobacteriota and Cyanobacteria (**Fig. 1e**), but were enriched with Gammaproteobacteria. In comparison, Antarctic and Arctic glaciers showed enrichment of Bacteroidota and Alphaproteobacteria, respectively (**Tables S5** and **S7**).

3.1.3 Bacteria community structure in the glaciers of the Three Poles

NMDS ordination plot revealed distinct microbial communities among various habitats (**Fig. 2a**, PERMANOVA, $P < 0.001$). Analysis of Bray-Curtis similarity revealed that snow and ice microbiomes were more similar (average Bray-Curtis similarity 18.9%) compared with that among other habitats (**Fig. S4**). Additionally, cryoconite water was more similar to ice (14.3%) than to either snow (11.5%) or cryoconite (13.1%), contrary to the native instinct that microorganisms in cryoconite and cryoconite water should be more similar due to their close connection. Additionally, PERMANOVA analysis showed that the microbial community structures across Antarctic, Arctic, and Tibetan glaciers were distinct ($P < 0.001$), with a similar R^2 value ($R^2 = 0.117$) compared to the habitat effect ($R^2 = 0.110$), suggesting that spatially location (dispersal limitation) and habitat (environmental filtering) play comparable roles in shaping glacier microbiomes.

We identified the three most abundant phylotypes for each habitat-region pair (**Fig. 2b**). The results exhibited strong spatial effects, with all habitats from the Tibetan glaciers clustering together. Additionally, the cryoconite-related habitats from the Antarctic and Arctic also clustered together, despite the two regions being geographically distinct.

We identified ubiquitous phylotypes for each region-habitat pair (i.e., identified in more than 55% of samples). There were five phylotypes identified as ubiquitous in all region-habitat pairs (**Table S8**), affiliated with Gammaproteobacteria (*Comamonadaceae*) or Actinobacteria (*Microbacteriaceae*). These phylotypes accounted for 5.0% of microbial communities by relative abundance across all samples on average, suggesting wide dispersal capacity and can adapt to various glacier-related habitats. In addition, 161, 13, and 52 phylotypes were identified as regional core phylotypes that were ubiquitously distributed in Antarctic, Arctic, and Tibetan glaciers, respectively. These phylotypes accounted for 22%, 26%, and 4% of the Antarctic, Arctic, and Tibetan glaciers' microbial community by relative abundance on average, respectively. Furthermore, 31, 41, and 44 phylotypes were identified as habitat core phylotypes for cryoconite, cryoconite water, and snow, accounting for 17%, 27%, and 22% of microbial communities on average, respectively. These region- and habitat-specific core microbes may be useful in elucidating the distribution of key microbiomes in glacier ecosystems.

3.1.4 Potential pathogens

We compared the amplicon data with a curated pathogen database and identified 2,517 potential pathogen phylotypes at a sequence identity threshold of 97% and a complete sequence alignment coverage (**Table S9**). *Acinetobacter baumannii* or *A. junii* were the most abundant, accounting for 1.9% of the total sequences recovered. These pathogens may be capable of causing diseases in humans, rodents, insects, and plants (Wardeh et al., 2015). Significant differences in the relative abundance of potential pathogens were detected across the habitats ($P < 0.001$) (**Fig. 3**). Specifically, snow and ice exhibited a higher relative abundance of potential pathogens than cryoconite water and cryoconite. There were snow and ice samples exhibited an extremely high relative abundance of potential pathogens. We propose that this could be explained by the similar selection mechanisms for long-distance dispersal survival and host-immune evasion. For example, *Staphylococcus* can express wall teichoic acid and lipoteichoic acid that assists in host infection and immune evasion (Xia et al., 2010), while the latter can also enhance bacterial cryo-survival in Gram-positive bacteria (Percy and Gründling, 2014). Spatially, the relative abundance of potential pathogens in Antarctic and Arctic glaciers were similar (13.5% and 7.8%, respectively), both of which were significantly lower than that in Tibetan glaciers (18.2%, $P < 0.05$).

3.2 Metagenome- and genome-dataset

We acquired 208 glacier metagenome data (**Table S2**) and 952 genomes from bacteria isolated from glacial environments (**Table S3**). After quality filtering and assembly, 62,595,715 unique Open Reading Frames (ORFs) were obtained. Of these dereplicated ORFs, 47.8% (29,947,128) were functional annotations using eggNOG.

3.2.1 Overall features glacier metagenome-assembled genomes

After binning, the dataset generated 3,375 metagenome-assembled genomes of medium quality (Genome completion ≥ 50%, contamination < 10%) and higher (Bowers et al., 2017b). After combining the genomes of cultivate glacier bacteria, this expanded the total genome number to 4,327 from the previously published 3,246 (Liu et al., 2022) (**Table S10**), an 34% increase. The median genome size was 3.46 Mb ranging between 0.42 Mb and 10.49 Mb; the GC% was 60%, ranging from 30% to 60%. These genomes were clustered into 1,610 genome OTUs (gOTUs) at 95% ANI, which were taxonomically affiliated with 33 phyla, 80 classes, 159 orders, 282 families, and 689 genera. Notably, 87% of the gOTUs were unable to be classified at the species level (**Fig. S5**), likely representing novel species. The glaciers of the Tibetan Plateau hosted the highest number of gOTUs (n=1447), followed by the Arctic (n=757), Antarctic (n=221), Alps (n=188), and South American glaciers (n=230) (**Fig. S6**). Only 2.5% of the gOTUs were observed in all regions, most (49.0%) were regional-specific.

3.2.2 Key functional genes

Carbohydrate-active enzymes: The dataset contains 1,052,745 genes encoding carbohydrate-active enzymes (CAZY, **Fig. 4a**), i.e., those enzymes involved in the metabolism of glycoconjugates, oligosaccharides, and polysaccharides (Zerillo et al., 2013). Genes associated with carbohydrate hydrolysis and biosynthesis were the most abundant, accounting for 48.1% and 47.4%, respectively. In contrast, those genes associated with non-hydrolytic cleavage of glycosidic bonds, hydrolysis of carbohydrate esters, and assisting in degrading biomass substrates were relatively scarce, accounting for 0.9%, 3.4%, and 0.2% of the predicted CAZY, respectively. This indicates that the glacier microbiome is competent in a diverse range of carbon transformation processes, mediating the delivery of carbon to downstream ecosystems.

Nitrogen cycling: The dataset contained 136,424 unique genes associated with nitrogen cycling, most of which (99.4%) were associated to nitrate reduction and/or denitrification pathways (**Fig. 4b**). These genes included the *nirB* gene responsible for the nitrite reduction to ammonia in the assimilatory nitrate reduction pathway, the *narB* and *nirA* genes responsible for sequential nitrate reduction to ammonia in dissimilatory nitrate reduction pathways, and the *nirK* gene responsible for nitrite reduction to nitric oxide in denitrification pathway. This suggests that microbial-driven nitrate reduction is widespread in glacial habitats for both nitrogen assimilation and energy supply, highlighting their potential roles in NO_x formation. In comparison, genes involved in nitrogen fixation (*nifH*, *nifK*, *nifD*, and *anfG*) and nitrification (*hao*) were relatively rare, with only 638 and 292 unique genes identified, respectively, suggesting that microorganisms capable of these high-energy demand processes only account for a small fraction of the glacial microbiome, which is consistent with the low nitrogen fixation rates reported in glacier related-habitats (Telling et al., 2011).

Methane cycling: The dataset contained 154 methane cycling-related genes. Of these, 93 were the soluble form of methane oxidase (*mmoX*), accounting for 61% of the total methane-cycling genes identified (**Fig. 4c**). More *mmoX* genes were identified from cryoconite (34.4% of the total methane-cycling genes identified) than from ice (20.8%) or snow (5.2%). Conversely, genes associated with the particulate form of methane oxidation (*pmoA*, *pmoB*, and *pmoC*) were more frequently identified from ice (21.4%) than from cryoconite (5.8%) and snow (3.2%). The different partition of soluble- and particulate-form methane oxidizers likely reflects their distinct environmental selection process. Only six unique methanogenesis-related genes (*mcrA*) were identified, almost exclusively in cryoconite metagenomes. This is consistent with the cryoconite as a methane source in the literature (Zhang et al., 2021).

Antimicrobial resistance genes: Using thresholds of 80% identity and 80% sequence coverage, we identified 960 ORFs that exhibited high sequence similarity to 224 antibiotic resistance genes (ARG). Of these identified ARGs, *MexF*, beta-lactamase, and *mexK* were the most abundant, accounting for 8.1%, 4.2%, and 3.7% of the ARGs identified, respectively (**Table S11**). The predominant antibiotic resistance mechanisms were antibiotic efflux and antibiotic inactivation, accounting for 44% and 41% of the total ARGs identified, respectively. These ARGs were predicted to confer resistance against 30 different antibiotics, with penam, tetracycline, and macrolide being the most commonly encountered resistant targets. Additionally, 54% of the identified ARGs provided multiple drug resistance, with the *OprM*, *CpxR*, and *tolC* genes conferring resistance to 16, 15, and 15 types of antibiotics, respectively. ARGs were identified in 566 genomes (13% of total genomes obtained). This low proportion of ARG-bearing genomes suggests that the glacier habitats are only weakly affected by antibiotic contamination. Of the genomes containing ARGs, 48.6% and 34.1% were affiliated with Proteobacteria and Firmicutes, respectively (**Table S12**, **Fig. 4d**). However, the resistance mechanisms exhibited by these two bacterial phyla were markedly distinct, with antibiotic efflux (*MexF*) and antibiotic target alternation (*vanZf*)/inactivation (*FosB*) being the most common mechanisms for Proteobacteria and Firmicutes, respectively. Most of the genomes (n=224) carried only a single ARG, while seven genomes possessed more than ten ARG genes, with *Pseudomonas aeruginosa* genomes hosting up to 48 ARGs.

Virulence factors: Using thresholds of 80% identity and 80% coverage, the dataset contains 66,017 virulence factors, accounting for 0.11% of the total ORFs identified (**Table S13**). Virulence factors were predominately associated with adherence, motility, and immune modulation functions, while those associated with toxin production accounted for only 0.49% (**Fig. 4e**). We did not detect any toxin genes from the genomes obtained using the same thresholds, with only those associated regulation, effector delivery systems, and metabolic factors being identified from Proteobacteria, Actinobacteria, and Deinococcota genomes. Nevertheless, 878 potential toxin genes were identified from the genomes if the criteria were loosened, with sequence identified ranging from 20.1% to 67.8%, these genes may represent novel toxins without references in the dataset, or non-toxin genes homologues to known toxin genes. These candidate toxin genes were most abundantly identified in Gammaproteobacteria (**Table S14**), followed by Bacteroides (15.9%) and Alphaproteobacteria (10.0%).

4 Data availability

The data introduced here is the first step in archiving global glacier microbial data. For this purpose, the data is deposited into the Global Glacier Genome and Gene Database (4GDB, https://tp.lzu.edu.cn/4gdb/) and the National Tibetan Plateau Data Center (https://doi.org/10.11888/Cryos.tpdc.300830, Liu et al., 2023), which provides a comprehensive solution for glacier microbial studies, featuring amplicon sequencing phylotype table, representative sequences, taxonomic annotations, metagenomic raw sequences, assembled contigs, annotated gene sequences, sequences of metagenome-assembled genomes, and the growth characteristics of cultivated microorganisms, into a user-friendly website. The 4GDB website is mainly structured into three sections, comprising amplicon sequencing, metagenome/genome sequences, and function prediction. The user-friendly web interface allows data filtering based on sample type, sample location, habitat type, gene type, and taxonomy, enabling seamless download of the filtered results. In conclusion, 4GDB (Reviewer link: https://nmdc.cn/4gdb/downloadtemp) provides an open-access genome- and gene-orientated resource platform that is regularly updated to include newly published and in-house generated sequence data.

Author contributions

YL conceptualized the paper, SH, YL, TY, ZZ, YC, KL, PL, JL, and MJ analyzed the data, TY, SG, QS, GF, LW, and JM developed the website, MJ and YL prepared the manuscript with contributions from all authors.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgements

This study was supported by the Second Tibetan Plateau Scientific Expedition and Research Program (STEP) 2019QZKK0503 (YL) and 2021QZKK0100 (PL, YL, and TY); the National Key Research and Development Plans 2021YFC2300904 (MJ); the National Natural Science Foundation of China U21A20176 and 92251304 (YL), and the strategic development program lzujbky-2021-sp66 (MJ).

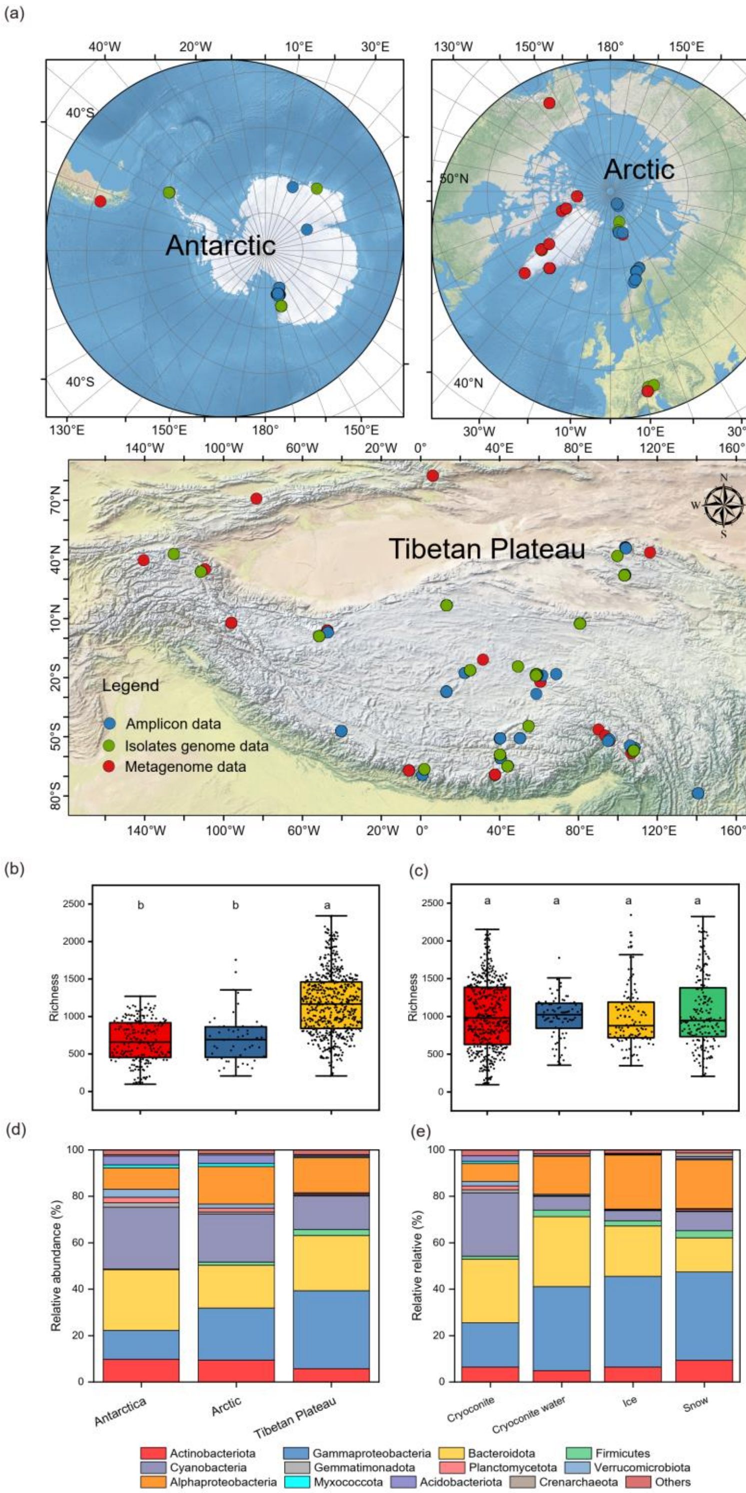

(a)
Antarctic
Arctic
Tibetan Plateau
Legend
Amplicon data
Isolates genome data
Metagenome data
40°W
20°W
10°E
30°E
40°S
130°E
150°E
180°
150°E
130°W
150°W
180°
150°E
130°E
50°N
40°N
30°W
10°W
10°E
30°E
140°W
100°W
60°W
20°W
0°
40°E
80°E
120°E
160°E
70°N
40°N
10°N
20°S
50°S
80°S
N
E
S
W
(b)
Richness
b
b
a
0
500
1000
1500
2000
2500
(c)
Richness
a
a
a
a
(d)
Relative abundance (%)
0
20
40
60
80
100
Antarctica
Arctic
Tibetan Plateau
(e)
Relative relative (%)
Cryoconite
Cryoconite water
Ice
Snow
Actinobacteriota
Cyanobacteria
Alphaproteobacteria
Gammaproteobacteria
Gemmatimonadota
Myxococcota
Bacteroidota
Planctomycetota
Acidobacteriota
Firmicutes
Verrucomicrobiota
Crenarchaeota
Others

Fig. 1 The location of glacier samples across the Antarctica, Arctic, and Tibetan glaciers, as well as their diversity indices and community taxonomic compositions.

290

a: The location of the glacier samples retained; b: Microbial richness comparison by region; c: microbial richness comparison by habitat; d: Microbial taxonomic compositions by region; and e: Microbial taxonomic compositions by habitat. Microorganisms includes bacteria and archaea. Significance is based on Kruskal-Wallis one-way ANOVA, multiple testing was performed based on the Dunn’s post-hoc test.

295

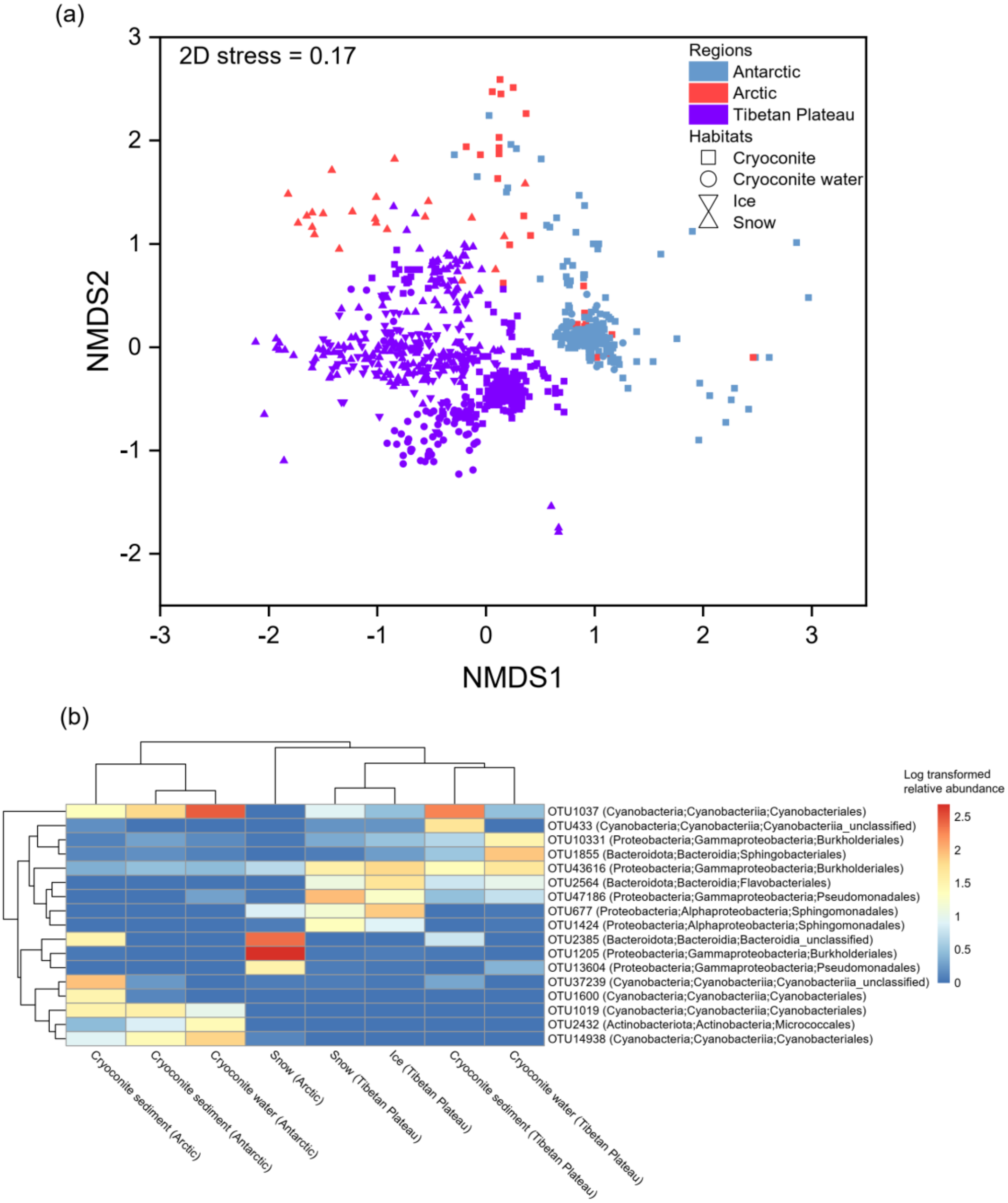


Fig. 2 The community structure of glacier microbiomes across the Antarctica, Arctic, and Tibetan Plateau.

a: Microbial community structure differences visualized using the non-metric multidimensional scaling ordination plot; b: The heatmap highlights the distribution pattern of dominant phylotypes for each habitat-region pair.

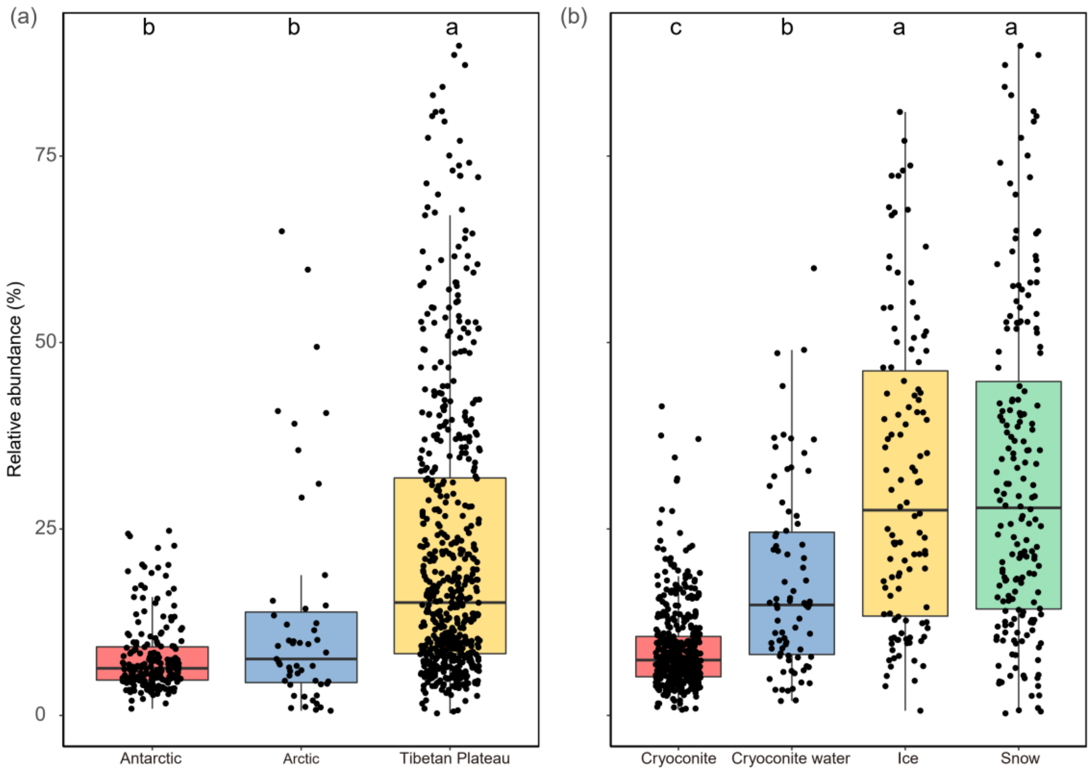
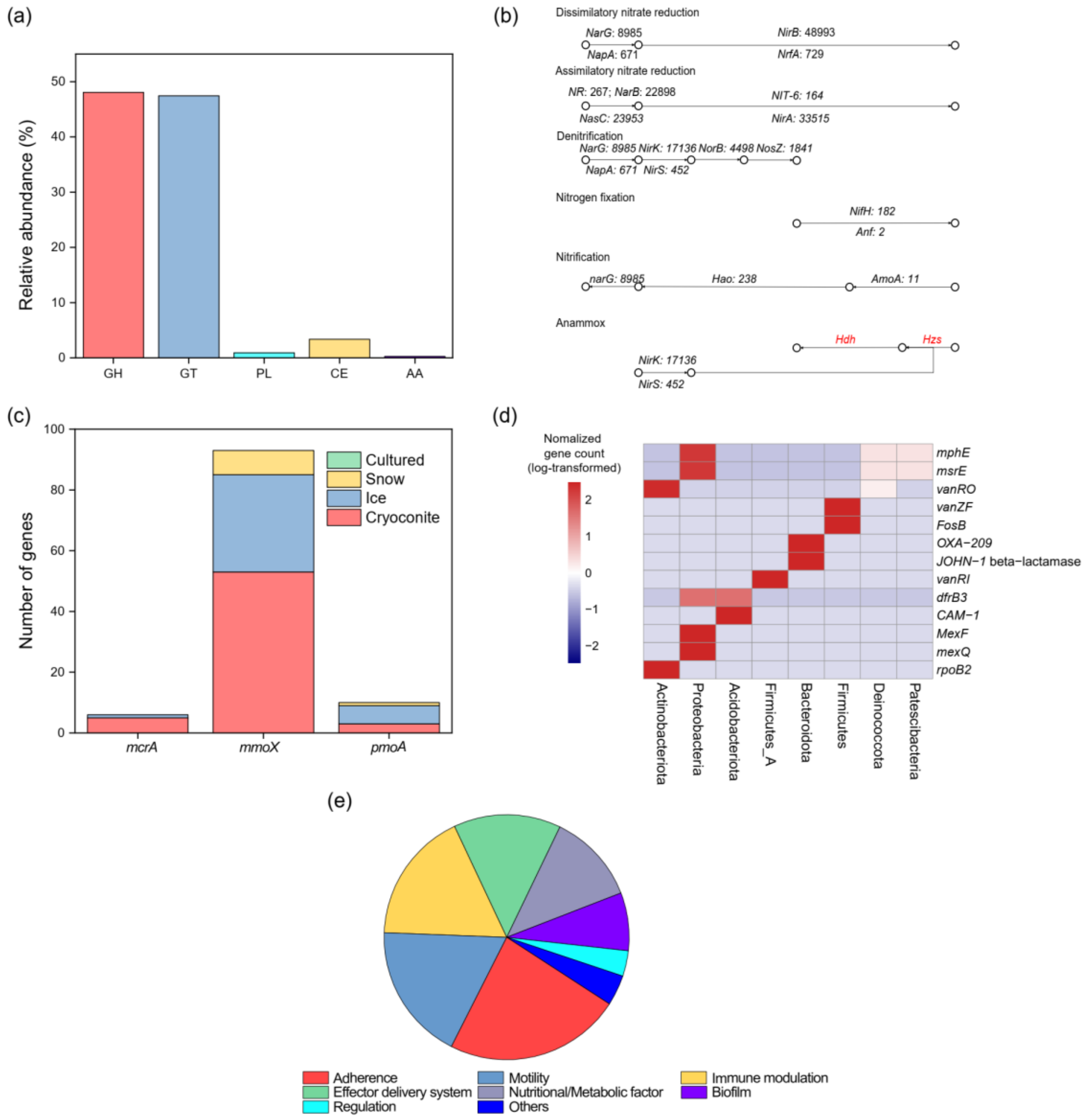


Fig. 3 The relative abundance of potential pathogens identified in the glacier microbiomes across the Antarctica, Artic, and Tibetan Plateau.

305

a: Relative abundance comparison by habitat; and b: Relative abundance comparison by region. Significance is based on Kruskal-Wallis one-way ANOVA, multiple testing was performed based on the Dunn's post-hoc test.

https://doi.org/10.5194/essd-2023-395
Preprint. Discussion started: 1 December 2023



310 **Fig. 4 Features of the functional genes across the glacier metagenomes from the Antarctica, Arctic, and Tibetan Plateau.**

a: Genes associated with carbohydrate-active enzymes (GH: Glycoside hydrolases; GT: glycosyl transferases; PL: Polysaccharide lyases; CE: Carbohydrate esterases; AA: Auxiliary activities); b: nitrogen-cycling (The numbers indicate the number of genes identified, *Hdh* and *Hzs* gene of the anaerobic ammonium oxidation pathway are not identified in the
315 glacier metagenomes); c: methane cycling (*mcrA* gene is responsible for methanogenesis; *mmoX* gene is the soluble form methane oxidation gene, while *pmoA* is the particulate form methane oxidation gene, both of which are associated with methane oxidation); d: genes associated with antibiotic resistance; and f: genes associated with virulence factors..

References

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H. K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., Hernandez-Koutoucheva, A., Sharma, A. N., Bordeleau, E., Pawlowski, A. C., Zubyk, H. L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G. L., Beiko, R. G., Brinkman, F. S. L., Hsiao, W. W. L., Domselaar, G. V., and McArthur, A. G.: CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database, Nucleic Acids Res., 48, D517-D525, https://doi.org/10.1093/nar/gkz935, 2020.

Anderson, M. J.: Permutational Multivariate Analysis of Variance (PERMANOVA), Wiley StatsRef: Statistics Reference Online, 2017.

Anesio, A. M. and Laybourn-Parry, J.: Glaciers and ice sheets as a biome, Trends Ecol. Evol., 27, 219-225, https://doi.org/10.1016/j.tree.2011.09.012, 2012.

Boetius, A., Anesio, A. M., Deming, J. W., Mikucki, J. A., and Rapp, J. Z.: Microbial ecology of the cryosphere: sea ice and glacial habitats, Nat. Rev. Microbiol., 13, 677-690, https://doi.org/10.1038/nrmicro3522, 2015.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., Weinstock, G. M., Garrity, G. M., Dodsworth, J. A., Yooseph, S., Sutton, G., Glöckner, F. O., Gilbert, J. A., Nelson, W. C., Hallam, S. J., Jungbluth, S. P., Ettema, T. J. G., Tighe, S., Konstantinidis, K. T., Liu, W. T., Baker, B. J., Rattei, T., Eisen, J. A., Hedlund, B., McMahon, K. D., Fierer, N., Knight, R., Finn, R., Cochrane, G., Karsch-Mizrachi, I., Tyson, G. W., Rinke, C., Lapidus, A., Meyer, F., Yilmaz, P., Parks, D. H., Eren, A. M., Schriml, L., Banfield, J. F., Hugenholtz, P., and Woyke, T.: Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea, Nat. Biotechnol., 35, 725-731, https://doi.org/10.1038/nbt.3893, 2017a.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., Weinstock, G. M., Garrity, G. M., Dodsworth, J. A., Yooseph, S., Sutton, G., Glöckner, F. O., Gilbert, J. A., Nelson, W. C., Hallam, S. J., Jungbluth, S. P., Ettema, T. J. G., Tighe, S., Konstantinidis, K. T., Liu, W.-T., Baker, B. J., Rattei, T., Eisen, J. A., Hedlund, B., McMahon, K. D., Fierer, N., Knight, R., Finn, R., Cochrane, G., Karsch-Mizrachi, I., Tyson, G. W., Rinke, C., Kyrpides, N. C., Schriml, L., Garrity, G. M., Hugenholtz, P., Sutton, G., Yilmaz, P., Meyer, F., Glöckner, F. O., Gilbert, J. A., Knight, R., Finn, R., Cochrane, G., Karsch-Mizrachi, I., Lapidus, A., Meyer, F., Yilmaz, P., Parks, D. H., Murat Eren, A., Schriml, L., Banfield, J. F., Hugenholtz, P., Woyke, T., and The Genome Standards, C.: Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea, Nat. Biotechnol., 35, 725-731, https://doi.org/10.1038/nbt.3893, 2017b.

Buchfink, B., Reuter, K., and Drost, H. G.: Sensitive protein alignments at tree-of-life scale using DIAMOND, Nat. Methods, 18, 366-368, https://doi.org/10.1038/s41592-021-01101-x, 2021.

Cameron, K. A., Hodson, A. J., and Osborn, A. M.: Carbon and nitrogen biogeochemical cycling potentials of supraglacial cryoconite communities, Polar Biol., 35, 1375-1393, https://doi.org/10.1007/s00300-012-1178-3, 2012.

Cauvy-Fraunié, S. and Dangles, O.: A global synthesis of biodiversity responses to glacier retreat, Nat. Ecol. Evol., 3, 1675-1685, https://doi.org/10.1038/s41559-019-1042-8, 2019.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., and Parks, D. H.: GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database, Bioinformatics, 36, 1925-1927, https://doi.org/10.1093/bioinformatics/btz848, 2019.

Ciccazzo, S., Esposito, A., Borruso, L., and Brusetti, L.: Microbial communities and primary succession in high altitude mountain environments, Ann. Microbiol., 66, 43-60, https://doi.org/10.1007/s13213-015-1130-1, 2016.

Clarke, K. R. and Warwick, R. M.: PRIMER v6: user manual/tutorial, 2, PRIMER-E, Plymouth, 2006.

Cook, J., Edwards, A., Takeuchi, N., and Irvine-Fynn, T.: Cryoconite: The dark biological secret of the cryosphere, Prog. Phys. Geogr., 40, 66-111, https://doi.org/10.1177/0309133315616574, 2016.

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-Gonzalez, A., Eldridge, D. J., Bardgett, R. D., Maestre, F. T., Singh, B. K., and Fierer, N.: A global atlas of the dominant bacteria found in soil, Science, 359, 320-325, https://doi.org/10.1126/science.aap9516, 2018.

Edgar, R. C.: Search and clustering orders of magnitude faster than BLAST, Bioinformatics, 26, 2460-2461, https://doi.org/10.1093/bioinformatics/btq461, 2010.

Franzetti, A., Navarra, F., Tagliaferri, I., Gandolfi, I., Bestetti, G., Minora, U., Azzoni, R. S., Diolaiuti, G., Smiraglia, C., and Ambrosini, R.: Potential sources of bacteria colonizing the cryoconite of an Alpine glacier, Plos One, 12, 0174786, https://doi.org/10.1371/journal.pone.0174786, 2017.

Guo, B. X., Liu, Y. Q., Liu, K. S., Shi, Q., He, C., Cai, R. H., and Jiao, N. Z.: Different dissolved organic matter composition between central and southern glaciers on the Tibetan Plateau, Ecol. Indic., 139, 108888, https://doi.org/10.1016/j.ecolind.2022.108888, 2022.

Hood, E., Battin, T. J., Fellman, J., O'Neel, S., and Spencer, R. G. M.: Storage and release of organic carbon from glaciers and ice sheets, Nat. Geosci., 8, 91-96, https://doi.org/10.1038/ngeo2331, 2015.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., and Bork, P.: Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper, Mol. Biol. Evol., 34, 2115-2122, https://doi.org/10.1093/molbev/msx148, 2017.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., and Bork, P.: eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses, Nucleic Acids Res., 47, D309-D314, https://doi.org/10.1093/nar/gky1085, 2019.

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J.: Prodigal: prokaryotic gene recognition and translation initiation site identification, BMC Bioinform., 11, 119, https://doi.org/10.1186/1471-2105-11-119, 2010.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., Lago, B. A., Dave, B. M., Pereira, S., Sharma, A. N., Doshi, S., Courtot, M., Lo, R., Williams, L. E., Frye, J. G., Elsayegh, T., Sardar, D., Westman, E. L., Pawlowski, A. C., Johnson, T. A., Brinkman, F. S., Wright, G. D., and McArthur, A. G.: CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database, Nucleic Acids Res., 45, D566-D573, https://doi.org/10.1093/nar/gkw1004, 2017.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K.: KEGG: new perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Res., 45, D353-D361, https://doi.org/10.1093/nar/gkw1092, 2017.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z.: MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies, PeerJ, 7, e7359, https://doi.org/10.7717/peerj.7359, 2019.

Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M., and Henrissat, B.: Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes, Biotechnol. Biofuels, 6, 41, https://doi.org/10.1186/1754-6834-6-41, 2013.

Liu, B., Zheng, D., Jin, Q., Chen, L., and Yang, J.: VFDB 2019: a comparative pathogenomic platform with an interactive web interface, Nucleic Acids Res., 47, D687-D692, https://doi.org/10.1093/nar/gky1080, 2019.

Liu, Y., Ji, M., Yu, T., Zaugg, J., Anesio, A. M., Zhang, Z., Hu, S., Hugenholtz, P., Liu, K., and Liu, P.: A genome and gene catalog of glacier microbiomes, Nat. Biotechnol., 40, 1341–1348, https://doi.org/10.1038/s41587-022-01367-2, 2022.

Liu, Y., Hu, S., Yu, T., Luo, Y., Zhang, Z., Chen, Y., Guo, S., S, Q., Fan, G., Wu, L., Ma, J., Liu, K., Liu, P., Liu, J., Ji, M.: A database of glacier microbiomes for the Three Poles [data set], https://doi.org/10.11888/Cryos.tpdc.300830.

Mao, G., Ji, M., Jiao, N., Su, J., Zhang Z., Liu, K., Chen, Y. and Liu Y.: Monsoon affects the distribution of antibiotic resistome in Tibetan glaciers, Environ. Pollut., 317, 120809, https://doi.org/10.1016/j.envpol.2022.120809, 2023.

McGinnis, S. and Madden, T. L.: BLAST: at the core of a powerful and diverse set of sequence analysis tools, Nucleic Acids Res., 32, W20-W25, https://doi.org/10.1093/nar/gkh435, 2004.

Mogrovejo-Arias, D. C., Brill, F. H. H., and Wagner, D.: Potentially pathogenic bacteria isolated from diverse habitats in Spitsbergen, Svalbard, Environ. Earth Sci., 79, 109, https://doi.org/10.1007/s12665-020-8853-4, 2020.

Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., Jensen, L. J., Nielsen, H. B., Petersen, T. N., Winther, O., and Rasmussen, S.: Improved metagenome binning and assembly using deep variational autoencoders, Nat. Biotechnol., 39, 555-560, https://doi.org/10.1038/s41587-020-00777-4, 2021.

Ogle, D. H., Doll, J. C., Wheeler, P., and Dinno, A.: FSA: Fisheries Stock Analysis [code], 2022.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., and Tyson, G. W.: Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life, Nat. Microbiol., 2, 1533-1542, https://doi.org/10.1038/s41564-017-0012-7, 2017.

Percy, M. G. and Gründling, A.: Lipoteichoic acid synthesis and function in gram-positive bacteria, Annu. Rev. Microbiol., 68, 81-100, https://doi.org/10.1146/annurev-micro-091213-112949, 2014.

Qiu, J.:China: The thrid pole, Nature, 454, 393-396, https://doi.org/10.1038/454393a, 2008.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O.: The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, Nucleic Acids Res., 41, D590-D596, https://doi.org/10.1093/nar/gks1219, 2012.

Steinegger, M. and Söding, J.: MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets, Nat. Biotechnol., 35, 1026-1028, https://doi.org/10.1038/nbt.3988, 2017.

Stevens, I. T., Irvine-Fynn, T. D. L., Edwards, A., Mitchell, A. C., Cook, J. M., Porter, P. R., Holt, T. O., Huss, M., Fettweis, X., Moorman, B. J., Sattler, B., and Hodson, A. J.: Spatially consistent microbial biomass and future cellular carbon release from melting Northern Hemisphere glacier surfaces, Commun. Earth Environ., 3, 275, https://doi.org/10.1038/s43247-022-00609-0, 2022.

Stibal, M., Bradley, J. A., Edwards, A., Hotaling, S., Zawierucha, K., Rosvold, J., Lutz, S., Cameron, K. A., Mikucki, J. A., Kohler, T. J., Sabacka, M., and Anesio, A. M.: Glacial ecosystems are essential to understanding biodiversity responses to glacier retreat, Nat. Ecol. Evol., 4, 686-687, https://doi.org/10.1038/s41559-020-1163-0, 2020.

Taheran, M., Naghdi, M., Brar, S. K., Verma, M., and Surampalli, R. Y.: Emerging contaminants: Here today, there tomorrow!, Environ. Nanotechnol. Monit. Manag., 10, 122-126, https://doi.org/10.1016/j.enmm.2018.05.010, 2018.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A.: The COG database: an updated version includes eukaryotes, BMC Bioinform., 4, 41, https://doi.org/10.1186/1471-2105-4-41, 2003.

Telling, J., Anesio, A. M., Tranter, M., Irvine-Fynn, T., Hodson, A., Butler, C., and Wadham, J.: Nitrogen fixation on Arctic glaciers, Svalbard, J. Geophys. Res. Biogeosci., 116, G03039, https://doi.org/10.1029/2010JG001632, 2011.

Wardeh, M., Risley, C., McIntyre, M. K., Setzkorn, C., and Baylis, M.: Database of host-pathogen and related species interactions, and their global distribution, Sci. Data, 2, 150049, https://doi.org/10.1038/sdata.2015.49, 2015.

Wu, Y. W., Simmons, B. A., and Singer, S. W.: MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets, Bioinformatics, 32, 605-607, https://doi.org/10.1093/bioinformatics/btv638, 2016.

Xia, G., Kohler, T., and Peschel, A.: The wall teichoic acid and lipoteichoic acid polymers of *Staphylococcus aureus*, Int. J. Med. Microbiol., 300, 148-154, https://doi.org/10.1016/j.ijmm.2009.10.001, 2010.

Zerillo, M. M., Adhikari, B. N., Hamilton, J. P., Buell, C. R., Levesque, C. A., and Tisserat, N.: Carbohydrate-Active Enzymes in Pythium and their role in plant cell wall and storage polysaccharide degradation, Plos One, 8, 0072572, https://doi.org/10.1371/journal.pone.0072572, 2013.

Zhang, Y. L., Kang, S. C., Wei, D., Luo, X., Wang, Z. Z., and Gao, T. G.: Sink or source? Methane and carbon dioxide emissions from cryoconite holes, subglacial sediments, and proglacial river runoff during intensive glacier melting on the Tibetan Plateau, Fundam. Res., 1, 232-239, https://doi.org/10.1016/j.fmre.2021.04.005, 2021.