Comments from the reviewer

General assessment

This manuscript presents a valuable contribution by assembling a large-scale, multi-source dataset of glacier microbiomes across the Three Poles and beyond. The integration of 16S rRNA amplicon data, metagenome-assembled genomes (MAGs), and cultured isolates is a strength, and the public database (4GDB) offers potential for long-term impact. However, several issues should be addressed before the manuscript can be considered for publication.

1. Definition and scope of "microbiomes"

The term "microbiomes" is generally understood to include a broad range of microbial life, including bacteria, archaea, microeukaryotes (e.g., algae and fungi), and viruses. While the manuscript refers to some of these groups in the introduction, the actual dataset and analyses are limited entirely to prokaryotes. There are no sequences or taxonomic assignments for eukaryotes or viruses.

I recommend clarifying this taxonomic scope explicitly in both the abstract and introduction. The current title may give the impression of broader taxonomic coverage than is actually presented, and I suggest revising it to reflect the prokaryote-specific focus more accurately.

Response

Thank you for the comments. We have specified that the dataset is "on supraglacial bacterial and archaeal (referred to as prokaryotic hereafter) communities across the Antarctic, Arctic, Tibetan Plateau, and other alpine regions". To further clarify the scope of our database, we changed the title of the manuscript to "A database of glacier prokaryotic genomes and genes for the Three Poles" and made other modifications listed below.

Amended manuscript

Abstract

Microbes, including bacteria, fungi, algae, and other microeukaryotes, are the primary inhabitants of glacier ecosystems and are key drivers of carbon and nitrogen transformation. Among them prokaryotes (including bacteria and archaea) are the most diverse and abundant.

The dataset contains 64,510 prokaryotic amplicon sequencing phylotypes, with a higher prokaryotic diversity in the Tibetan glaciers than in the Antarctic and Arctic glaciers…

The data can be leveraged to elucidate ecological principles governing the distribution of prokaryotes, to gain insights into the key functional genes for supraglacial microbiomes.

Introduction

From an ecological perspective, this dataset with standardized prokaryotic diversity, taxonomy, and community structure can improve understanding of the ecological principles governing the distribution of microorganisms across glaciers.

2. Functional gene analysis

The manuscript compiles an impressive array of functional gene annotations across thousands of MAGs, including pathways related to carbon degradation (CAZy), nitrogen cycling, methane metabolism, and antimicrobial resistance. However, the presentation remains largely descriptive.

While the authors provide a detailed catalog of gene abundances and distributions, the data are mostly reported in isolation, with limited interpretation in ecological or biogeochemical context. For example, the distinction between mmoX and pmoA gene distributions in cryoconite versus ice is noted, but the possible drivers—such as redox conditions, organic content, or microbial niche structure—are not explored. Similarly, CAZy profiles are summarized without considering variation across glacier types or habitats, and nitrogen-related genes are not discussed in terms of environmental gradients or host taxa.

Given the scope of the dataset, even basic exploratory analyses (e.g., ordination, correlation with metadata) could yield valuable insights. As it stands, this section reads more like a gene inventory than a functional synthesis.

Response

In response to your suggestion regarding expanding the ecological and biogeochemical interpretation of the gene distributions, we have incorporated new analyses and discussions to better link gene profiles with habitat differences and potential environmental drivers. Specifically, we added comparisons of CAZy gene distributions across different glacier types and habitats, explored the variations in nitrogen cycling genes in relation to habitat-specific microbial taxa and conditions, and discussed the environmental significance of methane monooxygenase gene distributions in cryoconite versus ice.

However, we would like to kindly note that the primary aim of our manuscript is to provide a comprehensive, original dataset and a functional gene catalogue across glacier-associated habitats. Earth System Science Data focuses on the dissemination and sharing of data resources that can support further Earth system research. While we agree that more detailed mechanistic or process-based ecological analyses would enrich interpretations, such in-depth studies are beyond the intended data reporting and synthesis scope of the manuscript.

Amended manuscript

For the comment regarding that CAZy profiles are summarized without considering variation across glacier types or habitats

The dataset contains 1,082,125 genes encoding carbohydrate-active enzymes (CAZymes, **Fig. 5a**), i.e., those enzymes involved in the metabolism of glycoconjugates, oligosaccharides, and polysaccharides (Zerillo et al., 2013). Genes associated with carbohydrate hydrolysis (GH) and biosynthesis (GT) were the most abundant, accounting for 45.2% and 44.4%, respectively. In contrast, those genes associated with non-hydrolytic cleavage of glycosidic bonds (PL), hydrolysis of carbohydrate esters (CEW), and assisting in degrading biomass substrates (AA) were relatively scarce, accounting for 0.8%, 3.1%, 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. We further examined the distribution of CAZyme genes across different glaciers. At the CAZY Family level, GT2, GT4, GH13, and GT51 are the most diverse across all habitats (**Fig. 5b**). PERMANOVA tests revealed significant composition differences by both habitat (pseudo-F = 10.014, $P < 0.001$) and regions (pseudo-F = 5.038, $P = 0.002$), with habitat exhibiting a greater influence on CAZY composition. Specifically, the PCA plot revealed a greater number of genes classified as GH5 and GH9 in cryoconites (**Fig. 5c**). Furthermore, 30 GHs and 20 GTs exhibited significantly higher contribution in cryoconites than in ice or snow (**Fig. 5d**). Comparatively, 15 GHs and 27 GTs exhibited significantly higher contribution in ice or snow than in cryoconites. Thus, cryoconites exhibited higher capacity in the catabolism of organic carbon, whereas snow and ice are dominated by anabolism.

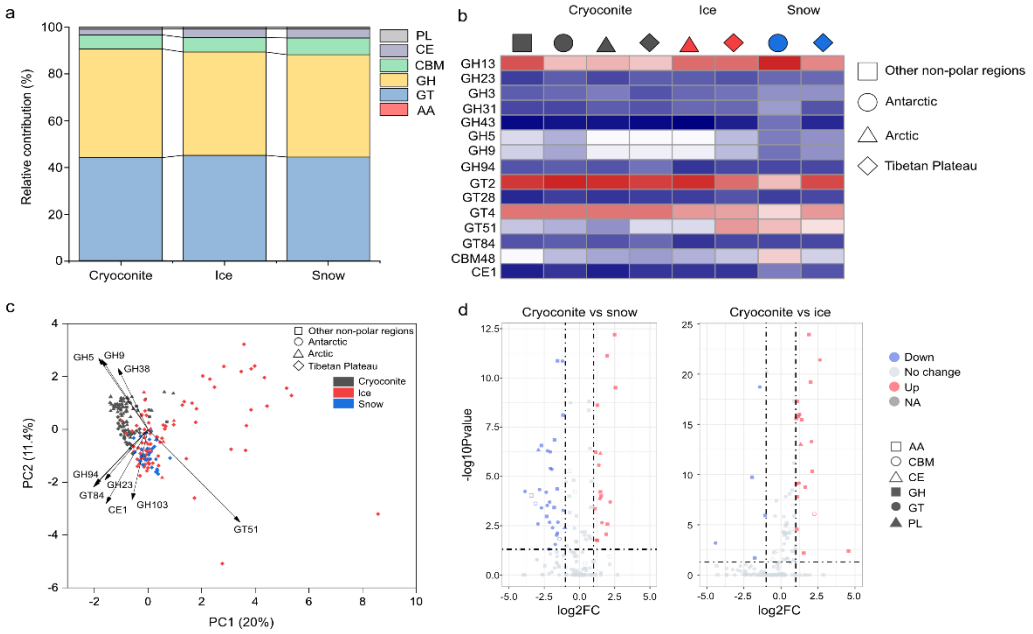

Fig. 5 The distribution of genes associated with carbohydrate-active enzymes (CAZymes) in glaciers.

a: The relative contributions of different genes associated with CAZymes in each habitat (GH: Glycoside hydrolases; GT: glycosyl transferases; PL: Polysaccharide lyases; CE: Carbohydrate esterases; AA: Auxiliary activities); b: The most diverse CAZymers in each habitat-region combination. c: PCA plots shows the distribution of CAZymes across habitats and regions; d: the selective enrichment of CAZymes between cryoconite and snow and between cryoconite and ice.

For the comment regarding that Nitrogen-related genes are not discussed in terms of environmental gradients or host taxa.

The differences in gene distributions are apparent among different habitats. Specifically, the genes associated with nitrogen fixation and denitrification are mainly present in cryoconites. The former could be explained by the high abundance of Cyanobacteria (**Fig. 6**), while the latter could be attributed to its anaerobic conditions. Thus, the

cryoconite could be a potential source of N_2O. Comparatively, *amoA* (in nitrification) is mainly identified from ice, while *norB* and *nosZ* (in denitrification) are also abundant. This reflects that ice could harbor greater functional diversity than snow, with the capacity for nitrogen transformation and influencing the nutrient discharged to downstream ecosystems.

For the comment regarding the Distinction between mmoX and pmoA gene distributions in cryoconite versus ice

The two forms of methane monooxygenase (sMMO and pMMO) have distinct enzymatic characteristics and are expressed in different growth conditions. Specifically, sMMO is expressed under low copper conditions (≤ 0.9 nmol of Cu/mg of cell protein), while pMMO is expressed under relatively high copper/biomass ratios (Zhang et al., 2017). Furthermore, sMMO has a broader substrate range, can oxidize methane, short chain alkane, alkene, and aromatic compounds, while pMMO can only oxidize alkanes of less than five carbons (Trotsenko and Murrell, 2008). Despite being not empirically measured, the cryoconite is expected to contain a higher concentration of copper than the ice, as the former is precipitated dust. Thus, the higher diversity of sMMO in the cryoconite could be associated with its ability to oxidize diverse alkanes, which may be present in soils. In comparison, high-affinity pMMO could oxidize methane at atmospheric levels, which may support microbial communities in the oligotrophic ice habitat. 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).

3. Amplicon analysis

The alpha and beta diversity analyses are competently executed, and the dataset is extensive. However, the manuscript does not explain **how differences in sequencing protocols**, **depth**, or **metadata completeness** across studies were handled.

Some discussion of **data harmonization strategies**, **rarefaction**, or the **potential for batch effects** would be helpful—particularly because inter-study comparisons are a central part of the analysis.

Response

Data homogenization is important for comparison among different datasets. We further discussed the harmonization strategy used, assessed the sequencing depth, and evaluated the influence of batch effect and primer usage. Specifically, we performed Kruskal Kruskal-Wallis one-way ANOVA test on all alpha diversity indices by the projects (for batch effects), the primers used, the amplified regions, and the sequencing platforms. The results showed that most indices are significantly different by these factors. However, this could be due to the different study areas employed. Thus, we repeated the comparison among different habitats and regions using data that was generated using the same primer set. The result patterns are consistent with our analysis using the entire dataset. Thus, the selection does not impact the conclusions. Similarly, we performed PERMANVOA analysis on the community structure, with consistent result patterns being observed, indicating that the different primers and studies do not affect the validity of the results.

We have added additional results and discussions to clarify this and commented on the completeness of metadata and sequencing depth.

Amended manuscript

For the discussions on the harmonization strategy

A variety of primers were used by these projects, amplifying the hypervariable regions V3V4, V4, and V4V5 regions. These data were harmonized by retaining only the V4 region (sequence trimming). Surprisingly, four bioprojects that used primers 783F and 1046R (V5V6 region) were also retained. We speculate that incorrect primers may have been provided in the NCBI. Nevertheless, we provided necessary information on the primer and regions amplified, users may choose to use the entire dataset or only those using the same primer or same amplification region.

Comments on metadata availability

The retained datasets are originated from 66 bioprojects, eight of which missed sequencing platform information and two of which missed primer information (Table S2). Most of these bioprojects do not have environmental metadata, therefore are not included in the dataset.

Assessments on sequencing depth

The Good's coverage index provides estimation for the number of singletons in a sample, reflects the coverage of the sequencing. The values of the index were 0.98±0.02 and 0.96±0.02 for the datasets without and with subsampling, respectively (Table S7). This indicates that majority of the OTUs were identified.

Evaluation on the batch effect and primer usage on alpha diversity indices

We further assessed the influence of the region amplified on the validity of the results. For each region-habitat pair, the alpha diversity indices were significantly different by these factors to a certain extent. However, those from Arctic basal ice, non-polar glacier ice, and Tibetan Plateau supraglacial meltwater were less affected (**Table S9**). Nevertheless, the influence may be explained by the different sampling locations, which have distinct microbial composition. We further tested the validity of the diversity comparison results using the data that were generated using the same primer set (**Table S10**). The influence of primer selection on prokaryotic diversity was inconsistent. For instance, primers targeting the V4 region resulted in a higher richness in supraglacial ice than primers targeting the V3V4 region in other alpine glaciers. In contrast, the primers targeting the V3V4 region had a higher in the Arctic. Such inconsistency in microbial community assessment by different primers and platforms has been reported previously (Fredriksson et al., 2013; Tremblay et al., 2015). Thus, the homogenization method may not fully overcome the bias caused by primer selection. Nevertheless, we provided necessary information on the primer and regions amplified, user of the dataset may choose to use the entire dataset or only those using the same primer for analysis.

4. Database usability and FAIR principles

The 4GDB appears to be a useful and well-organized resource. However, the manuscript provides little detail regarding its long-term maintenance or accessibility features.

It remains unclear how often the database will be updated, whether APIs or bulk downloads are available, or under what license the data can be reused. A short section outlining how 4GDB aligns with FAIR principles—especially regarding reuse and interoperability—would enhance transparency and user trust.

Response

The website is constructed under the OSI-approved CC BY 4.0 Open Source license (https://creativecommons.org/licenses/by/4.0/), all data can be accessed and reused freely, without any restrictions, for both academic and commercial purposes. Regarding the FAIR (Findable, Accessible, Interoperable and Reusable) principles. We permit bulk download at https://nmdc.cn/4gdb/download, which provide compressed files for the genomes of the recovered MAGs, predicted genes from the MAGs, and the OTU tables for the 16S rRNA gene amplicon sequencing. These data can be downloaded using FTP client such as FileZilla. For the raw sequencing files, they are stored in NCBI and EMBL-EBI, we do not provide direct download link for these files, but listed the accession number to redirect the user to the correct data. For predicted genes from the metagenome, we provide blast search option, so that the user can find the genes of interest and download them individually. Due to the large size, we don't provide bulk download for these sequences. The website will be maintained by the author teams, at the stage, we will try to update at least once per year.

Amended manuscript

The website is constructed Under the OSI-approved CC BY 4.0 Open Source license (https://creativecommons.org/licenses/by/4.0/), all data can be accessed and reused freely, without any restrictions, for both academic and commercial purposes.

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 (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.