Linking genomic and physiological characteristics of psychrophilic Arthrobacter to metagenomic data to explain global environmental distribution

Running title: Polar and alpine Arthrobacter

Liang Shen1,2, Yongqin Liu1,3,*, Michelle A. Allen4, Baiqing Xu1,3, Ninglian Wang3,5, Timothy J. Williams4, Feng Wang1, Yuguang Zhou6, Qing Liu6, Ricardo Cavicchioli4*

1Key Laboratory of Tibetan Environment Changes and Land Surface Processes, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, 100085, China
2College of Life Sciences, Anhui Normal University, Wuhu, 241000, China
3CAS Center for Excellence in Tibetan Plateau Earth Sciences, Beijing, 100085, China
4School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia
5College of Urban and Environmental Science, Northwest University, Xian, 710069, China
6China General Microbiological Culture Collection Center (CGMCC), Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

*Correspondence: yqliu@itpcas.ac.cn
1Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, 100101, China
r.cavicchioli@unsw.edu.au
4School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia
**Abstract**

**Background:** Microorganisms drive critical global biogeochemical cycles and dominate the biomass in Earth’s expansive cold biosphere. Determining the genomic traits that enable psychrophiles to grow in cold environments informs about their physiology and adaptive responses. However, defining important genomic traits of psychrophiles has proven difficult, with the ability to extrapolate genomic knowledge to environmental relevance proving even more difficult.

**Results:** Here we examined the bacterial genus *Arthrobacter*, and assisted by genome sequences of new Tibetan Plateau isolates, defined a clade, Group C, that represents isolates from polar and alpine environments. Group C had a superior ability to grow at -1ºC, and possessed genome G+C content, amino acid composition, predicted protein stability and functional capacities (e.g., sulfur metabolism and mycothiol biosynthesis) that distinguished it from non-polar or alpine Group A-*Arthrobacter*. Interrogation of more than 1,000 metagenomes identified an over-representation of Group C in Canadian permafrost communities from a simulated spring-thaw experiment, indicative of niche adaptation, and an under-representation of Group A in all polar and alpine samples, indicative of a general response to environmental temperature.

**Conclusion:** The findings illustrate a capacity to define genomic markers of specific taxa that potentially have value for environmental monitoring of cold environments, including environmental change arising from anthropogenic impact. More broadly, the study illustrates the challenges involved in extrapolating from genomic and physiological data to an environmental setting.

**Keywords:** Genomics, metagenomics, psychrophiles, polar environment, alpine environment, microbial adaptation
Background

Many biotic and abiotic factors influence the ability of microorganisms to become indigenous members of environmental communities. Certain environmental factors can limit or prevent the growth of microorganisms, while enhancing, or being essential for others, resulting in ecological niches that support specific microbiome structures [1]. This phenomenon is well illustrated by a Winogradsky column where light and oxygen can be seen to exert major influences on the diversity and the dynamic of microorganisms throughout its length. In more recent times, particularly through technological advancement (e.g., metagenomics), the understanding of microbial ecology and the contributions that microorganisms make to the natural world has grown considerably. Appreciation for microorganisms has accrued from discoveries of new biomes capable of supporting microbial colonization, such as the deep subsurface [2]; new examples of life hidden within microbial ‘dark matter’ (e.g. Asgard archaea [3]), and; the dynamic nature of microbial responses, particularly those that provide surprises, such as the major societal upheaval caused by the COVID19 coronavirus. There is a growing realization that microorganisms constitute the life support system of the biosphere and must be properly accounted for when devising strategies to mitigate the impacts of human activity on the natural world [4]. In essence, we are living in a period in history when the need for society to learn about microbial responses to natural and anthropogenic influences is of unprecedented relevance [4-6].

Metagenomic methods have provided a level of insight into microbial communities [7, 8] that could possibly be equated to the advances made by Carl Woese and colleagues when using rRNA sequencing to discover Archaea as the third domain of life. Applied to the cold biosphere, Earth’s single largest biome, metagenomic analyses have catalogued diverse ways in which microbial life has evolved. As an example, Antarctic, marine-derived lake communities have been shown to have evolved independently over their relatively short history of 3,000-5,000 years, adapting not just to low temperature but also to a variety of important environmental factors specific to each lake system (reviewed in REF. [9]).

Metagenomic analyses have also begun to be used to uncover the ways in which communities
in polar environments respond to changing environmental conditions; for example, the effects of the seasonal polar sunlight cycle on Antarctic marine and marine-derived lake communities [10, 11], and the roles that Arctic bacteria play in melting permafrost acting as a CO₂ source and atmospheric CH₄ sink [12, 13].

*Arthrobacter* (Actinobacteria; Micrococcales; Micrococcaceae) are a globally distributed genus of bacteria commonly found in soil, but also in a broad range of environments including water, human skin, and sewage [14-17]. *Arthrobacter* are reported to play important roles in global biogeochemical cycles and decontamination of polluted environments [16, 18]. Their responses to temperature, desiccation, ionizing radiation, oxygen radicals, and a range of chemicals have been described [19-21]. Their growth in the laboratory is characterized by nutritional versatility that translates to an ability to grow aerobically in media utilizing a wide range of carbon and nitrogen sources [15]. Some strains closely related to the type species *A. globiformis* were isolated from a Lapland glacier region [22], and the identification of psychrophilic species has led to the characterization of a number of *Arthrobacter* enzymes for their biotechnological potential (e.g., REF. [23]). Cold environments that *Arthrobacter* have been isolated from include permafrost and glaciers [24-26].

Due to the large scale of the Earth’s cold biosphere and its relevance to global biogeochemical cycles, and the biotechnological potential of psychrophiles and their products, numerous studies have been performed to attempt to define the critical traits of psychrophiles (discussed in REFS [9, 23, 27-33]). In the current study, we sequenced the genomes of 13 *Arthrobacter* strains isolated from the Tibetan Plateau and utilized the existence of more than 100 *Arthrobacter* genomes to assess traits that may explain the presence of the genus in naturally cold environments. After identifying a clade characteristic of polar and alpine environments and determining that representatives had a superior ability to grow at low temperature, we used available metagenome data to assess the environmental relevance of the findings. What we learned illustrated the complexities involved in attempting to extrapolate from genomic and physiological data to an environmental setting. It also revealed possible avenues for utilizing *Arthrobacter* as biomarkers of environmental
Results

Phylogenomics

To increase the number of *Arthrobacter* genomes from polar and alpine (PA) environments, isolates from the Tibetan Plateau were sequenced resulting in high-quality genomes (see Materials and methods) for three glacier and 10 lake isolates (Table S1). The phylogenomic relationships of a total of 210 non-redundant high-quality Micrococcaceae genomes were analyzed by constructing ML and Bayesian trees (Fig. S1). The two trees were congruent and most tree-nodes (194/208) were supported by high bootstrap values (>70 %) (Fig. S1). The *Arthrobacter* lineage formed a cluster with 106 representatives that was clearly separated from other Micrococcaceae genera (Fig. S1). The *Arthrobacter* genomes represented PA isolates (total 31, including the 13 new genomes), with the remainder from a broad range of non-polar or alpine (NPA) environments (Table S1).

The *Arthrobacter* lineage separated into three clades that branched from the root of the *Arthrobacter* tree (Fig. 1). The 31 PA strains were distributed across the tree, although 11 PA strains formed a cluster with three NPA strains in the central clade (Fig. 1a). Within the central clade, 10 strains grouped together (Fig. 1b, blue font) with an $F$ measure of 0.95, defining them as an operationally monophyletic lineage [34]. This was supported by NMDS analysis of amino acid composition (Fig. 1c) and both ANI and AAI distributions (Fig. 1d, e).

The grouping of 10 strains (Fig. 1b, blue font) consisted of nine PA strains, plus *A. psychrolactophilus* B7 which was isolated from Pennsylvania soil following snow melt; the strain was isolated as a source of cold-active enzymes and was capable of growth at 0 °C [35]. The monophyletic lineage of 10 strains was defined as Group C (Fig. 1b, blue font), with all other PA strains as Group B (Fig. 1b, green font), and all strains from NPA environments as Group A (Fig. 1b, red font).

Low temperature growth capacity of Group C
To evaluate the growth temperature response of Group A, B and C, three strains from each group were grown at 25, 5 and -1 °C and growth monitored (OD$_{600}$) over time (Fig. 2). The three Group C strains exhibited a markedly enhanced rate of growth at -1 °C (Fig. 2c) particularly compared to Group A strains, and had a reduced rate of growth at 25 °C compared to some of the Group A and B strains (Fig. 2a, b).

**Genomic characteristics**

The size of the 106 *Arthrobacter* genomes ranged from 3.24 – 5.89 Mbp (Table S1). Between Group A and C, no significant differences occurred in genome size, 16S rRNA and tRNA gene copy number, or coding density (Fig. S2). However, a significant difference was observed in amino acid composition and G+C content (Fig. 3, Fig. S2, Table S1). In Group C, the content of N, K, M, I, S, T, F, Q, W and H was significantly higher, while A, E, G, P, D and R was significantly lower (one-way ANOVA, $p < 0.05$, Fig. S3). The correlation between amino acid composition and G+C content was significantly negative for N, I, M, S, F, K, Q and G+C content ($R^2$ ranged from 0.46 – 0.81; $p < 0.01$), and significantly positive for W, G, D, P, R and A ($R^2$ ranged from 0.39 – 0.77; $p < 0.01$) (Fig. S4).

To evaluate the potential structural relevance of the amino acid compositional differences, temperature-dependent protein stability predictions were made using SCooP, which predicts stability assuming proteins are monomeric and follow a two-state folding transition [36]. From 180 proteins targeted for evaluation (mostly single-copy genes; see Materials and Methods), 86 produced robust stability curves (Fig. 3, Fig. S5, Table S2). Stability was calculated at -1 °C to match the growth temperature at which Group C showed a marked difference in growth ability (Fig. 2). The 86 Group C proteins had significantly higher $\Delta G$ values (Group A, -4.27; Group B, -4.29; Group C, -3.60; $p < 0.01$), with 32 proteins being responsible for the reduced predicted stability (Fig. S5, Table S2). These 32 proteins contained a particularly high representation of the amino acids that were most over-represented in Group C (i.e., N, K and R; Fig. S3). The 32 proteins represented 12 functional categories, primarily metabolism (28 proteins; 9 categories), with four involved in respiration,
stress response, or cell division and cell cycle (Table S2). The marked amino acid
compositional differences, broad representation of functional categories, and high proportion
of proteins with predicted decreases in stability (~ 1/3rd of those tested), demonstrates that
Group C *Arthrobacter* possess broad genomic differences to Group A *Arthrobacter*. If the
decreases in predicted protein stability translate to an increased capacity to perform catalysis
at low temperature, this may contribute to the higher growth rates of Group C at -1°C (Fig. 2).

To further explore the influence of amino acid composition on functional potential, Bray-
Curtis distances of genome-wide amino acid composition were evaluated for proteins
representing 26 functional categories (Fig. 3). The greatest distance was for the category
‘phages, prophages, transposable elements, plasmids’, consistent with previous studies
associating transposable elements with cold adapted microorganisms [37-39]. The functional
potential of Group C was also compared to Group A using enrichment analysis [40]
performed on proteins representing the 26 functional categories. Group C was over-
represented in sulfur metabolism; cofactors, vitamins, prosthetic groups, pigments; protein
metabolism; stress response; cell division and cell cycle (Fig. 3, Fig. S6, Table S5), whereas
Group A, was over-represented in metabolism of aromatic compounds; nitrogen metabolism;
amino acids and derivatives; regulation and cell signaling (Fig. S6, Table S5). The category
‘sulfur metabolism’ also exhibited signatures of amino acid bias (Bray-Curtis distance; Fig.
3), suggesting selection for this functional capacity occurred at the levels of both gene
complement and amino acid composition.

Functional assessments were extended to attempt to identify specific genes unique to
Group C. A number of genes involved in the synthesis of amino acids, vitamins, and
nucleosides were present in all Group C genomes (Fig. 4). The specific genes also tended to
be present in the other four, non-Group C members (Group A and Group B) of the central
clade (Fig. 4), but had low representation in other *Arthrobacter* genomes (Fig. 4). The most
marked feature was a complete mycothiol (MSH) biosynthesis pathway that was present in all
Group C genomes (Fig. 4); MSH is a redox-active thiol, functionally analogous to glutathione
(which is typically absent from Actinobacteria), that maintains intracellular redox balance and
can therefore protect against oxidative damage [41]. Further, MSH potentially serves as a stable reservoir of carbon and sulfur in bacteria [42]. The ability to respond effectively to oxidative damage may be an important trait of microorganisms from cold environments, particularly for facilitating growth at low temperature limits [43-47]. The MSH pathway was also present in the other four, non-Group C members of the central clade, plus one other Group B member (Fig. 4). Therefore, the MSH pathway plus the individual genes involved in the synthesis of amino acids, vitamins, and nucleosides, are characteristic, but not unique to Group C Arthrobacter. If MSH or the other individual genes fulfill roles in facilitating growth at low temperature, the genes may be under stronger positive selection in Group C, while also being retained within Group A and Group B populations (pan genome), but at a significantly lower level.

A total of 48 Group C gene families had significantly higher, and 66 had significantly lower average gene copy number compared to Group A, including four which were absent in Group C genomes (Table S6, \( p < 0.05 \)). The absence of two specific genes is noteworthy: adenosylhomocysteinase, which hydrolyzes S-adenosyl-L-homocysteine (a product of methyl transfer reactions that involve S-adenosyl-L-methionine) to homocysteine and adenosine [48]; and formate--tetrahydrofolate (THF) ligase, which catalyzes the initial recruitment of single carbon units for THF-mediated one-carbon metabolism [49]. The absence of both genes would be expected to disrupt the synthesis of methionine from homocysteine, and instead favor the alternative pathway of synthesizing methionine from cysteine; the latter pathway may be connected to MSH metabolism, in that accumulation of cysteine (a precursor of MSH synthesis) is toxic to cells [42], so surplus cysteine not required for MSH synthesis, or resulting from MSH degradation, could be directed to methionine synthesis (Table S6).

Some of the gene families had particularly high copy numbers per genome (~30 in Group A) with large reductions (~6) in Group C (Table S6); this trend was observed for 3-oxoacyl-[acyl-carrier protein] reductase (FabG), glycerate kinase (GlxK), and alcohol dehydrogenase (Adh). For FabG, this likely reflects a reduced capacity of Group C to catalyze the formation of long-chain fatty acids (Table S6). GlxK is an important catabolic enzyme, in that diverse
substrates are degraded to glycerate, and GlxK links these degradation pathways to central carbon metabolism [42]. Decreased copy numbers of Adh likely indicates decreased capacities to utilize alcohols. Thus, decreases in GlxK and Adh might reflect decreased substrate preferences by these Arthrobacter strains. It was noteworthy that the copy number of cold shock protein (csp) genes was lower in Group C. While csp genes are sometimes equated with an ability to grow in the cold or survive cold shock, these nucleic acid binding proteins can perform diverse roles in cellular function (reviewed in REF. [33]); the findings here reinforce the notion that csp and other ‘stress’ genes are not good molecular markers for identifying psychrophiles [4, 32].

Ecology of Group C Arthrobacter

We hypothesized that if the laboratory-generated growth data (Fig. 2) and genomic traits (Fig. 3) translated to competitiveness in cold environments, Group C Arthrobacter would be over-represented in metagenome data from PA vs NPA environments. The relative abundance of Arthrobacter in environmental samples (publicly available metagenome data) tends to be low, with no metagenome assembled genomes (MAGs) present in the ~8,000 that were constructed from ~1,500 metagenomes [50], and a total of 12 (> 90% completeness) present in 76,831 IMG MAGs (December 2019). To facilitate metagenome analyses, group-specific genes (Table S3) were examined in 797 metagenomes representing PA, temperate and tropical environments (Table S4), with representation shown relative to the Arthrobacter pan genome (Fig. 5a) (see Materials and methods for description of analytical approach).

Group C-specific genes were more highly represented in 11 permafrost metagenomes (Fig. 5b). All of the 11 metagenomes came from a single site: Axel Heiberg Island, Nunavut, Canada [51]. The Axel Heiberg Island study reported 76 metagenomes derived from 1 m cores that were used during a controlled thawing experiment [51]. Most Group C-specific genes were enriched in the 65 cm depth active-layer (7 metagenomes), with one from the 35 cm active-layer and three from the 80 cm permafrost-layer (Table S7). A total of 94% of the variability that exists in hits to Group C for the 797 metagenomes (Fig 5b) was traced to pre-
existing variability in hits to the *Arthrobacter* pan genome, and when this covariance was
removed by ANCOVA analysis, a statistically significant difference in the y-intercepts for the
regression lines ($p < 0.0001$) remained; this confirms the over-representation of Group C-
specific genes in the 11 metagenomes compared to the remaining 786 metagenomes.

To assess whether Group C *Arthrobacter* were generally enriched in permafrost regions,
all other publicly available permafrost metagenomes (361 datasets) were analyzed (Fig. 5c). A
number appeared somewhat enriched in Group C-specific genes (e.g., a metagenome from
Stordalen Mire near Abisko, Sweden; marked by an arrow in Fig. 5c), but as the slopes of the
two regression lines were not parallel, it was not valid to compare the y-intercepts, and hence
the significance of the difference between them could not be evaluated [52].

For Group B-specific genes, no obvious trends separated the PA from NPA metagenomes
(Fig. 5d). However, the distribution of Group A-specific genes clustered according to climate,
with PA metagenomes showing lower Group A content ($y = 0.0161x + 0.1398$, $R^2 = 0.96773$)
compared to temperate and tropical metagenomes ($y = 0.0244x - 1.8835$, $R^2 = 0.96432$) (Fig.
5e). The 11 Axel Heiberg Island metagenomes had a statistically significant under-
representation of Group A-specific genes compared to all other metagenomes (ANCOVA, $p <
0.0001$). This pattern indicates there is selection against Group A *Arthrobacter* in PA
environments, and/or selection for Group A in NPA environments.

To attempt to define variables that may explain the niche adaptation of Group C in the
Axel Heiberg Island permafrost, available abiotic and biotic data were used from the
permafrost study [12, 13]. A range of physicochemical data were available for each of the four
depths (5, 35, 65 and 80 cm) but as the timing of sampling for physicochemical data (0, 4, 6,
8, 11 and 12 weeks) did not align with the timing of sampling for the metagenomes (0, 0.25,
6, 12 and 18 months), the physicochemical data were ultimately not useful for interpreting
Group C distribution. Depth, treatment group and sample core did not explain the variation in
species composition across the sites, and although incubation time had some explanatory
power for the distribution of the entire permafrost study dataset (data not shown), the
metagenomes that were enriched in Group C *Arthrobacter* were widely distributed and did
not cluster together, suggesting the importance of specific microniches in enrichment of these species. Assessment of functional potential of the microbial communities in each of the 76 metagenomes using presence/absence of KO groups also did not identify any significant functional differences (PERMANOVA, \( p > 0.05 \); data not shown).

In contrast, strong taxonomic associations were identified with many members of the microbial community. Analyses were performed to assess taxa that correlated with Group A and Group C, just Group A, and just Group C (Table S8). A total of 107 OTUs positively correlated with Group C *Arthrobacter*, and 63 negatively correlated (Table S8). Of the 107 OTUs positively correlated to Group C, 72 were also positively correlated to Group A above the threshold of 0.5 (the remainder were positively correlated with values 0.359 - 0.499), and no OTUs were positively correlated to Group A that did not also correlate to Group C. The positively correlating OTUs were dominated by both spore- and non-spore-forming members of Actinobacteria and Firmicutes, as well as members of Proteobacteria; the majority of these OTUs were isolated from soil. Negatively correlating bacterial OTUs mainly belonged to marine or lacustrine members of Bacteroidetes, Cyanobacteria, and Proteobacteria, as well as certain eukaryotes (fungi, plants, marine annelid worm).

**Discussion**

Numerous studies have been performed to attempt to define the critical traits of a psychrophile (discussed in REF. [32, 33, 53]). The current study explored genomic characteristics of a lineage with less than 3.5 % difference in 16S rRNA gene identity. The analyses revealed that the *Arthrobacter* lineage contains a clade with members (Group C) possessing a clear capacity to grow faster than their relatives (Group A and B) under laboratory growth conditions at -1 °C (Fig. 2). Genomic characteristics that potentially explain the physiological capacity include an amino acid composition that is predicted to reduce the stability of a large proportion of proteins thereby enhancing enzyme activity at low temperature [52]. Group C genomes are enriched in sulfur metabolism genes, and sulfur is required for the cysteine component of mycothiol. The synthesis of mycothiol may potentially
protect Group C *Arthrobacter* against oxidative damage that may otherwise accumulate as cell division decreases towards the lower temperature limit of growth [43]. Group C also exhibits a relatively high proportion of mobile elements, which is a trait shared with some other cold-adapted microorganisms [37-39]. Collectively, the physiological and genomic traits appear compelling for denoting Group C, a cold-adapted clade of *Arthrobacter*. However, from assessing available metagenome data, we infer that these traits do not translate to a generally enhanced ability to compete in low temperature environments. Other than the 11 specific Axel Heiberg Island permafrost metagenomes, Group C *Arthrobacter* were not highly represented in the other metagenomes from cold environments, including 144 from Arctic peat soil, 22 associated with glaciers, 42 from polar deserts, and importantly, 365 from other permafrost environments. Even at the Axel Heiberg Island site, Group C-specific genes were not highly abundant at 5 and 20 cm depths. Instead, the pattern of abundance of Group C appears to derive not just from low temperature, but from niche-specific conditions. Attempting to identify specific niche conditions is not trivial. For the Axel Heiberg Island study, the permafrost microbial community was reported to be dominated by Actinobacteria and Proteobacteria, with significant increases at depth for Firmicutes and Actinobacteria and significant decreases for Acidobacteria, Proteobacteria and Verrucomicrobia [13]. However, despite these taxonomic differences, we did not identify significant predicted functional differences by depth. When we turned to specifically correlating the abundance of Group C to OTUs from the metagenome data, a large number of OTUs with positive or negative correlations were identified (Table S8). At a broad level, the environmental data of the positively correlating taxa are consistent with Group C associating with other soil bacteria. While this provides scope for investigating specific taxa that may help shape the niche that Group C occupy, determining which taxa are important and the nature of their interactions will require a dedicated effort. For the negatively correlating cohort, they tend to represent isolates from non-soil environments (Table S8) and may therefore represent non-indigenous microorganisms that have been introduced. The permafrost samples were obtained from an “upland polygonal
terrain in proximity to the McGill Arctic Research Station at Expedition Fjord (79°24'57"N, 90°45'46"W)" [12]. The prevalence of negatively correlating OTUs matching to Proteobacteria isolated from sea water may reflect aeolian carriage from Expedition Fjord, which is located ~8 km from the Research Station. As the samples were obtained for a simulated permafrost-thaw experiment [51], the negatively correlating OTUs may also reflect environmental disturbance.

Conclusions

Our study commenced with the analysis of genome sequences of new Group C Arthrobacter isolated from the Tibetan Plateau and progressed through to a rationalization of Group C abundance in global metagenomes. Group C was clearly distinguished from Group A- Arthrobacter by possessing genomic signatures consistent with its representation in PA environments and an ability to grow faster when cultivated at -1°C. Assessment of available metagenome data point to the Group C traits as being more relevant to cold niches rather than competitiveness across global permafrost or cold soil environments. The challenge in being able to define the specific niche parameters enabling Group C Arthrobacter to be relatively competitive illustrates the inherent difficulties associated with defining ‘cause and effect’ for explaining ‘why’ microorganisms reside in the environments in which they are found; that is, the characteristics of the ecological niches that define microbiome structure [1]. Without knowing the specific effectors, the ability to understand and predict responses to environmental changes, is greatly compromised [4, 6, 54, 55]. Establishing long-term data records that include comprehensive metadata associated with monitoring sites, including metadata for each biological sample, will be essential for learning how to link environmental parameters to microbial processes. In a study of sulfate reduction in Arctic marine sediments, growth yield was reasoned to be the most relevant factor for determining competitiveness of sulfate-reducing bacteria in permanently cold marine sediments [53]. These findings illustrate that for cold environments, linking genomic and metagenomic data to measurements of metabolic rates, growth rates and growth yields will undoubtedly help to clarify how specific
microbial processes and associated taxa are influenced by environmental temperature. While the characteristics that define the Group C niche are still to be defined, at sites where Group C *Arthrobacter* are relatively abundant they may have value as a biomarker for monitoring the stability of those locations. Moreover, Group A *Arthrobacter* may serve as a more broadly useful biomarker of soil microbial communities. Group A exhibited high relative abundance across NPA metagenomes and relatively low abundance across PA metagenomes. As the data indicate environmental temperature exerts a broad, strong influence on Group A *Arthrobacter*, we predict that environmental warming will generally increase the relative abundance of Group A. Similar influences of environmental temperature have been described for the marine SAR11 clade, including the predicted displacement of polar specialists by phylotypes from warmer latitudes [56]. Depending on how strongly the environmental factors other than temperature select for Group C in permafrost, the apparent broad influence of temperature on Group A suggests it will displace Group C from the niches in which it is currently relatively competitive.

**Methods**

*Arthrobacter* isolation and genome sequencing

Isolation of *Arthrobacter* strains from glaciers and lakes on the Tibetan Plateau was performed as previously described [25, 26] (Table S1). All strains were deposited in the China General Microbiological Culture Collection Center (CGMCC) with accession numbers: CGMCC 1.16187-1.16198, 1.16223 and 1.16312. Genomic DNA for 16 isolates was extracted using a TIANamp Bacteria DNA Kit (Tiangen, Beijing) following manufacturer’s instructions. Paired-end libraries with an insert size of 500 bp were constructed and sequenced using Illumina Hiseq 2000 platform. Prior to *de novo* sequence assembly, low-quality reads were filtered out using Fastp with default options [57]. Filtered sequencing reads were subjected to assembly using SPAdes v3.11.1 with default options [58]. The assembled genome sequences were deposited in DDBJ/ENA/GenBank under the BioProject PRJNA421662.
Growth temperature response

*Arthrobacter* strains were grown in R2A broth at 25, 5 and −1°C for up to 10 days with optical density measured at 600 nm (OD$_{600}$) using a Microplate Reader (MD, SpectraMax M5). For cultivation at −1°C, flasks were placed in ice produced by an ice maker (TKKY, FM40) with flasks placed in a ~ 4°C refrigerator and ice replaced every 12 h. Growth at 5 and 25°C was performed using a constant-temperature incubator as described previously [59]. All cultures were grown statically, with flasks swirled to resuspend biomass prior to recording OD$_{600}$. Strains used for growth temperature profiles were: Group A: *A. luteolus*, *A. globiformis* and *A. subterraneus*; Group B: *Arthrobacter* sp. 4R501, *Arthrobacter* sp. 9E14 and *Arthrobacter* sp. 08Y14; Group C: *A. alpinus*, *Arthrobacter* sp. A3 and *Arthrobacter* sp. N199823.

Preparation of *Arthrobacter* genomes for analysis

As the taxonomic assignment of the genus *Arthrobacter* is not consistent, in August 2018 all genome sequences with the taxonomy identifier ‘Arthrobacter’ or ‘Micrococcaceae’ were retrieved from GenBank, providing a total of 427 genomes including the 16 new Tibetan Plateau genomes. The completeness of each genome was calculated using CheckM v1.0.7 with default options [60]. Genomes composed of > 300 contigs, with an N50 of < 20 kb, completeness of < 95%, and contamination > 5% were removed. Genomes were dereplicated to remove genomes with an average amino acid identity (AAI) ≥ 99.5%. AAI values were calculated using CompareM with default options (https://github.com/dparks1134/CompareM). Average nucleotide identity (ANI) was calculated using the ANI calculator (http://enve-omics.ce.gatech.edu/ani/). A total of 210 genomes met quality requirements, which included 13 of the 16 new Tibetan Plateau genomes (Table S1). Gene families were clustered using FastOrtho software (--pv_cutoff 1-e5 --pi_cutoff 70 --pmatch_cutoff 70) (http://enews.patricbrc.org/fastortho/) with the cutoff values set according to Parks *et al* [34]. A gene family matrix was produced using custom PERL scripts, and non-functional based
group specific genes were calculated based on this matrix. The annotation of genes was
standardized by annotating all genomes using RAST (Rapid Annotation using Subsystem
Technology) [61] and PROKKA [62].

**Phylogenetic and genomic analyses**

For phylogenomic clustering, *Cellulomonas carbonis* T26 and *C. fimi* ATCC 484 were chosen
as the outgroup as they are close relatives of Micrococcaceae [63], and species that are
closely related to the in-group are more suitable for phylogenetic reconstruction than distantly
related species [64]. A Maximum Likelihood (ML) phylogenomic tree was constructed using
PhyloPhlAn2 with default options [65]. A Bayesian tree was constructed using MPI Mrbayes
v3.2 (preset aamodelpr = mixed, mcmc nchains = 16, ngen = 300,000, and leaving other
parameter values as default) [66]. The F measure (harmonic mean of precision) provides a
metric for determining if taxa are operationally monophyletic (F measure ≥ 0.95) [34]; and
was calculated as $F = 2^\ast ((\text{precision} \ast \text{recall}) / (\text{precision} + \text{recall}))$. Genome wide amino acid
composition was calculated using CompareM with the function aa_usage. The stability curves
of proteins were predicted by SCooP [36] using the PDB (Protein Data Bank) files modeled
by SWISS-MODEL [67]. The stability equations of the same protein from different hosts
were visualized and smoothed using ggplot2 v3.2.1 [68]. The stability curves were analyses
for 180 single-copy genes that were shared by most genomes; a small number of genomes had
multiple copies of genes, and up to three genomes were permitted to have the absence of the
gene in order to account for the use of unclosed genomes (99 of the 106). After retrieval from
SWISS-MODEL of all possible PDB files matching the candidate genes, a total of 17,339
stability equations were generated (Table S2). Ordination and statistical analyses, including
three-dimensional nonmetric multidimensional scaling (3D-NMDS) and gene enrichment
analyses were performed with R v3.3.3 and Origin v9.5. For comparisons between Group A,
B and C Arthrobacter, group-specific genes or functions were defined as being present in
95% of the target group (e.g., Group C) genomes and absent in 95% of each of the other
group(s) (e.g., Group A) genomes. Group-specific genes were identified (Group A, 74
genomes, 16,149 specific genes; Group B, 22 genomes, 4,675 specific genes; Group C, 10
genomes, 969 specific genes; Table S3) and normalized to account for the different number of
genomes used for each group. To account for differences in gene content between strains,
comparisons were calculated relative to the total *Arthrobacter* gene complement from all 106
*Arthrobacter* genomes (referred to as the *Arthrobacter* pan genome). Gene copy number was
calculated as the average number of the gene for each genome in a group (e.g., Group C),
with gene loss or gain calculated from the average copy number for the groups (e.g., Group C
vs Group A). To assess the bias of amino acid composition of different functional classes of
proteins, genes were assigned to functional categories (assigned by RAST) and total amino
acid composition for all proteins from the functional category were compared between groups
(e.g., Group C vs Group A). Similarity was measured by Bray-Curtis distance with larger
Bray-Curtis distances denoting stronger bias, possibly indicative of selection pressure [40].
The functional potential of groups were also compared using enrichment analysis [40].
Briefly, the presence or absence of KEGG Ortholog (KO) groups in genomes and
metagenomes (see Collection and analysis of metagenomes) was assessed [40] and non-
parametric one-way ANOVA was used to identify differentially abundant categories using R
[69].

**Collection and analysis of metagenomes**

Assembled metagenomes were downloaded from IMG (https://genome.jgi.doe.gov/portal/).
Classification into polar and alpine (PA) or non-polar or alpine (NPA) environments were
made using metadata associated with metagenomes, supplemented by Köppen-Geiger climate
classifications (to define temperate and tropical regions) using ArcGIS location data [70].
Analyses were initially performed using 797 metagenomes from environments with mean
ambient annual temperature (MAT) ranging from -24 °C to 28 °C, representing PA (n = 286,
gene count = 244 million), temperate (n = 294, gene count = 291 million), and tropical (n =
217, gene count = 981 million) zones (Table S4). Subsequently, all additional (361) available
(May 2020) assembled permafrost metagenomes were analyzed. Analyses assessed the
relative abundance of each *Arthrobacter* group (A, B and C) using group-specific genes (see
Phylogenetic and genomic analyses) by performing a local alignment search against the
metagenomes using DIAMOND v0.9.24 with the arguments --outfmt 6, --query-cover 70, --id
70, --evalue 1e-5, and leaving others as default [71]. One-way ANCOVA was used to assess
statistical differences between regression lines for groups of metagenomes [72] using the data
import webform for k=2 at http://vassarstats.net/vsancova.html. Correlation analyses were
performed between *Arthrobacter* groups and other members of the microbial community from
76 metagenomes derived from a simulated permafrost thaw experiment [13]. Operational
taxonomic units (OTUs) were assigned from IMG phylodist matches which are based on the
top taxon in the IMG isolate database; < 4% of OTUs had < 35% identity. The raw abundance
of all OTUs was determined, with 956 meeting the criteria of average abundance ≥2, and
detection in at least 56 of the 76 metagenomes. The 956 OTUs were used to construct a
correlation matrix using SparCC [73] implemented in python3 with default parameters (20
iterations). One hundred simulated datasets were created by random shuffling of the original
input with replacement, and their correlation matrices constructed in the same way. The
simulated datasets were used to calculate the one- and two-sided pseudo p-values. The
selected threshold for strong correlations was > | 0.5 |. To assess if depth, incubation time,
treatment group, sample core, or physicochemical data explained the variation in species
composition across permafrost sites, a generalized linear latent variable model was employed
as implemented in the R package gllvm [74]. All custom scripts are available at

Acknowledgements

We thank Qilong Qin, Joshua N. Hamm, Pratibha Panwar, Sten Anslan, Saleh Rahimlou, Ping
Ren, Fei Liu, Weizhi Song and Wei Zhu for their valuable input related to data analyses, and
Maggie Lau, Lyle Whyte and Tullis Onstott for information related to the Axel Heiberg Island
site.
Authors’ contributions

YL, RC, and LS designed the study; YL, BX, and NW collected field samples; YZ and QL provided type strains; LS and FW performed the experiments; LS and MAA analyzed the data; RC, LS, TJW and MAA interpreted the data and wrote the paper.

Funding

This study was financially supported by the National Natural Science Foundation of China (Grant Nos. 91851207 and 41701085), the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA20050101), and the Second Tibetan Plateau Scientific Expedition and Research (STEP) program (Grant No. 2019QZKK0503). The Australian contingent was supported by funding from the Australian Research Council (DP150100244). Liang Shen was funded by the China Scholarship Council (Grant No. 201804910177).

Availability of data and materials

The 16 newly isolated Arthrobacter strains were deposited in the China General Microbiological Culture Collection Center (CGMCC) with accession numbers: CGMCC 1.16187-1.16198, 1.16223 and 1.16312. The assembled genome sequences for newly isolated Arthrobacter strains were deposited in DDBJ/ENA/GenBank under the BioProject PRJNA421662. All custom scripts are available at (https://github.com/environmental-genomes/Arthro).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests
The authors declare that they have no conflict of interest.

References


23. Feller G. Psychrophilic enzymes: from folding to function and biotechnology. Scientifica...


45. Dsouza M, Taylor MW, Turner SJ, Aislabie J. Genome-based comparative analyses of


**Fig. 1 Arthrobacter phylogeny and genome compositional profiling.** (a) Maximum Likelihood Arthrobacter phylogenomic tree. The Arthrobacter portion of ML Micrococcaceae phylogenomic tree (Fig. S1a) is reproduced with each leaf marked as polar and alpine (PA, grey highlight) or non-polar and alpine (NPA). The tree has three major clades with the central clade highlighted (purple box). (b) As for (a) except strain names denoted and font color used to depict Group A (red font; NPA environments), Group B (green font; PA environments clustering with sequences from NPA environments) and Group C (blue font; PA environments that formed an operationally monophyletic lineage with an F measure of 0.95). The specific types of cold environments from where Group C Arthrobacter were isolated are shown to the right of tree. (c) 3D-NMDS plot of genome-wide amino acid composition. (d) Distribution of pairwise ANI. (e) Distribution of pairwise AAI.

**Fig. 2 Growth temperature profiles of Group A, B and C Arthrobacter.** OD$_{600}$ growth
curves for representative *Arthrobacter* of Group A (red symbols and line; *A. luteolus, A. globiformis* and *A. subterraneus*), Group B (green symbols and line; *Arthrobacter* sp. 4R501, *Arthrobacter* sp. 9E14 and *Arthrobacter* sp. 08Y14) and Group C (blue symbols and line; *A. alpinus, Arthrobacter* sp. A3 and *Arthrobacter* sp. N199823) at (a) 25 °C, (b) 5 °C and (c) -1 °C.

**Fig. 3 Overview of genomic characteristics of Group C *Arthrobacter*.** (a) Box plot of G+C content. Group A (red); Group C (blue); boxes represent the interquartile range with horizontal lines showing maximum and minimum values, excluding outliers. Group C had significantly lower G+C content. (b) Scatter plot of amino acid composition. Group A (light red circles); Group C (blue circles); ***, *P < 0.005; *P = 0.05 – 0.01; ns, not significant. The composition of numerous amino acids varied significantly between Group C and Group A *Arthrobacter*. (c) Protein stability predictions calculated using SCooP. Group A (red line); Group C (blue line). The curve is for coenzyme A biosynthesis bifunctional protein, CoaC, and is representative of one of the 32 Group C proteins from a total of 86 which had reduced predicted stability (Table S2, Fig. S5) (d) Box plot of amino acid bias for functional categories. Boxes represent the interquartile range of the Bray-Cruise distances; lines extending from boxes show the maximum and minimum Bray-Cruise distances; dots beyond the lines represent outliers. Biases in amino acid composition (b) were reflected in specific functional categories. (e) Representation of functional categories. Specific functional categories were over- or under-represented in Group C; arrows indicate relative increases (up arrow) or decreases (down arrow) in functional categories in Group C. (f) Representation of specific functions. Specific functional processes defined by genes or pathways were characteristic of Group C (up arrow) or had a restricted capacity in Group C (down arrow) compared to Group A (also see Fig. 4).

**Fig. 4 Arthrobacter genes typifying the functional potential of Group C.** (a) ML *Arthrobacter* phylogenomic tree as for Fig. 1. (b) Heat map of the representation of specific
genes in *Arthrobacter* genomes, highlighting those present in Group C and the central clade. *i*, branched-chain acyl-CoA dehydrogenase; *ii*, enoyl-CoA hydratase; *iii*, biotin repressor; *iv*, hydrolase in cluster with formaldehyde/S-nitrosomycothiol reductase; *v*, mycothiol-dependent formaldehyde dehydrogenase.

**Fig. 5 Metagenome analysis of Group C Arthrobacter.** (a) Depiction of the (mean annual temperature (MAT) of surface air at a height of 2 m (European Centre for Medium-Range Weather Forecasts) relative to latitude. The 797 metagenomes are divided into thermal categories: PA (black squares, 286 metagenome), temperate (grey squares, 294 metagenomes) and tropical (purple squares, 217 metagenomes). (b) Linear regression showing the correlation of the abundance of Group C-specific genes within each of the 797 metagenomes (see panel (a)) relative to the abundance of Group C-specific genes within the *Arthrobacter* pan genome. The 95% prediction interval (dark red band) and 95% confidence interval (light red band) is shown for each regression line (panels b, c and e). The upper cluster contains 11 Axel Heiberg Island permafrost metagenomes. (c) As for panel (b), except with the addition of 361 permafrost metagenomes (total 1158 metagenomes). The Stordalen Mire (Abisko, Sweden) metagenome is shown by an arrow. (d) As for panel (b), except showing Group B-specific genes. (e) As for panel (b), except showing Group A-specific genes present in PA genomes (lower line) and NPA genomes (upper line). The regression line for the 11 Axel Heiberg Island permafrost metagenomes is not shown.
Fig. 1 *Arthrobacter* phylogeny and genome compositional profiling. (a) Maximum Likelihood *Arthrobacter* phylogenomic tree. The *Arthrobacter* portion of ML Micrococcaceae phylogenomic tree (Fig. S1a) is reproduced with each leaf marked as polar and alpine (PA, grey highlight) or non-polar and alpine (NPA). The tree has three major clades with the central clade highlighted (purple box). (b) As for (a) except strain names denoted and font color used to depict Group A (red font; NPA environments), Group B (green font; PA environments clustering with sequences from NPA environments) and Group C (blue font; PA environments that formed an operationally monophyletic lineage with an $F$ measure of 0.95). The specific types of cold environments from where Group C *Arthrobacter* were isolated are shown to the right of tree. (c) 3D-NMDS plot of genome-wide amino acid composition. (d) Distribution of pairwise ANI. (e) Distribution of pairwise AAI.
Fig. 2 Growth temperature profiles of Group A, B and C *Arthrobacter*. OD$_{600}$ growth curves for representative *Arthrobacter* of Group A (red symbols and line; *A. luteolus*, *A. globiformis* and *A. subterraneus*), Group B (green symbols and line; *Arthrobacter* sp. 4R501, *Arthrobacter* sp. 9E14 and *Arthrobacter* sp. 08Y14) and Group C (blue symbols and line; *A. alpinus*, *Arthrobacter* sp. A3 and *Arthrobacter* sp. N199823) at (a) 25 °C, (b) 5 °C and (c) -1 °C.
Fig. 3 Overview of genomic characteristics of Group C *Arthrobacter*. (a) Box plot of G+C content. Group A (red); Group C (blue); boxes represent the interquartile range with horizontal lines showing maximum and minimum values, excluding outliers. Group C had significantly lower G+C content. (b) Scatter plot of amino acid composition. Group A (light red circles); Group C (blue circles); ***, $P < 0.005$; *, $P = 0.05 – 0.01$; ns, not significant. The composition of numerous amino acids varied significantly between Group C and Group A *Arthrobacter*. (c) Protein stability predictions calculated using...
SCoP. Group A (red line); Group C (blue line). The curve is for coenzyme A biosynthesis bifunctional protein, CoaC, and is representative of one of the 32 Group C proteins from a total of 86 which had reduced predicted stability (Table S2, Fig. S5) (d) Box plot of amino acid bias for functional categories. Boxes represent the interquartile range of the Bray-Cruise distances; lines extending from boxes show the maximum and minimum Bray-Cruise distances; dots beyond the lines represent outliers. Biases in amino acid composition (b) were reflected in specific functional categories. (e) Representation of functional categories. Specific functional categories were over- or under-represented in Group C; arrows indicate relative increases (up arrow) or decreases (down arrow) in functional categories in Group C. (f) Representation of specific functions. Specific functional processes defined by genes or pathways were characteristic of Group C (up arrow) or had a restricted capacity in Group C (down arrow) compared to Group A (also see Fig. 4).
Fig. 4 *Arthrobacter* genes typifying the functional potential of Group C. (a) ML *Arthrobacter* phylogenomic tree as for Fig. 1. (b) Heat map of the representation of specific genes in *Arthrobacter* genomes, highlighting those present in Group C and the central clade. i, branched-chain acyl-CoA dehydrogenase; ii, enoyl-CoA hydratase; iii, biotin represser; iv, hydrolase in cluster with formaldehyde/S-nitrosomycothiol reductase; v, mycothiol-dependent formaldehyde dehydrogenase.
Fig. 5 Metagenome analysis of Group C *Arthrobacter*. (a) Depiction of the (mean annual temperature (MAT) of surface air at a height of 2 m (European Centre for Medium-Range Weather Forecasts) relative to latitude. The 797 metagenomes are divided into thermal categories: PA (black squares, 286 metagenome), temperate (grey squares, 294 metagenomes) and tropical (purple squares, 217 metagenomes). (b) Linear regression showing the correlation of the abundance of Group C-specific genes within each of the 797 metagenomes (see panel (a)) relative to the abundance of Group C-specific genes within the *Arthrobacter* pan genome. The 95% prediction interval (dark red band) and 95% confidence interval (light red band) is shown for each regression line (panels b, c and e). The upper cluster contains 11 Axel Heiberg Island permafrost metagenomes. (c) As for panel (b), except with the addition of 361 permafrost metagenomes (total 1158 metagenomes). The Stordalen Mire (Abisko, Sweden) metagenome is shown by an arrow. (d) As for panel (b), except showing Group B-specific genes. (e) As for panel (b), except showing Group A-specific genes present in PA genomes (lower line) and NPA genomes (upper line). The regression line for the 11 Axel Heiberg Island permafrost metagenomes is not shown.