Glacier influence shapes the genomic architecture of the downstream aquatic microbiome

Running title: Bacterial genome architecture in GFS

Keywords: phylogenetics, glacier-fed streams, cryosphere, metagenomics, genomic architecture, microbial genomics, Gammaproteobacteria

Massimo Bourquin[1,*], Hannes Peter[1,*], Grégoire Michoud[1], Aileen Geers[1], Susheel Bhanu Busi[2], The Vanishing Glaciers Field Team**, Tom Ian Battin[1]

(1) River Ecosystems Laboratory, Alpine and Polar Environmental Research Center, Ecole Polytechnique Fédérale de Lausanne, EPFL, Lausanne, Switzerland

(2) UK Centre for Ecology and Hydrology, Wallingford, United Kingdom

corresponding authors

Mailing address:

Massimo Bourquin, River Ecosystems Laboratory, Environmental Engineering Institute, École Polytechnique Fédéral de Lausanne, Route des Ronquos 86, 1951 Sion, Switzerland

Hannes Peter, River Ecosystems Laboratory, Environmental Engineering Institute, École Polytechnique Fédéral de Lausanne, Route des Ronquos 86, 1951 Sion, Switzerland

****The Vanishing Glaciers Field Team**: Mike Styllas, Matteo Tolosano, Martina Schön, Vincent de Staercke, Tyler Kohler

Abstract

The factors and processes that shape microbial genomes and determine the success of microbes in different environments have long attracted scientific interest. Here, leveraging 2,855 metagenome-assembled genomes (MAGs) sampled by the *Vanishing Glacier Project* from glacier-fed streams (GFSs), we shed light on the genomic architecture of the benthic microbiome in these harsh ecosystems — now vanishing because of climate change. Owing to glacial influence, the GFS benthic habitat is unstable, notoriously cold and ultra-oligotrophic. Along gradients of glacial influence and concomitant variation in benthic algal biomass across 149 GFSs draining Earth's major mountain ranges, we show how genomes of GFS bacteria vary in terms of size, coding density, gene redundancy, and translational machinery. We develop a novel, phylogeny-rooted analytical framework that allows pinpointing the phylogenetic depth at which patterns in genomic trends occur. These analyses reveal both deep- and shallow- rooting phylogenetic patterns in genomic features associated with key GFS taxa and functional potential relevant to live in these ecosystems. Additionally, we highlight the role of several clades of Gammaproteobacteria in shaping community-level genomic architecture. Our work shows how genome architecture is shaped by selective

environmental constraints in an extreme environment. These insights are important as they reveal putatively important adaptations to the GFS environment which is now changing at rapid pace due to climate change.

Introduction

Bacterial genomes exhibit substantial variation in size and complexity [1, 2] and are shaped by processes including genetic drift [2, 3], selection by environmental constraints such as oligotrophy and symbiotic interactions that facilitate gene loss [1, 4]. Among the factors that shape bacterial genome architecture, environmental drivers related to genome size variation have attracted most attention. For instance, thermophilic microbes thriving in hot springs often possess small genomes [5], whereas psychrophilic microbes in cryospheric environments tend to have larger genomes [6][7, 8]. Increased genome size has also been associated with the need to maintain a broad functional repertoire to cope with fluctuating environmental conditions [9–11]. In addition to habitat characteristics such as temperature and nutrient availability, microbial lifestyle—such as free-living pelagic versus attached forms—have also been associated with genome size variation [12–14]. However, other genomic features, such as variation in guanine-cytosine (GC) content, gene redundancy or the translational machinery have received less attention, particularly for environmental bacteria. Here, we analyse metagenomic and environmental data from glacier-fed streams (GFSs) and investigate relationships between glacial influence and variation in genomic features of the benthic microbiome.

Owing to the direct influence of glaciers [15], GFSs are ultra-oligotrophic, cold and unstable environments, yet they harbour diverse microbial communities [16–18]. In GFSs, bacteria forming biofilms attached to sedimentary surfaces dominate microbial life, where they orchestrate important ecosystem functions [19–21]. These communities are shaped by selective environmental conditions, which is reflected by deterministic community assembly and elevated microdiversity [22, 23]. Yet,

how the environmental characteristics shape the genomic architecture of GFS bacteria remains unknown. In the light of ongoing climate change and glacier recession [24], better understanding genomic trends along environmental factors is however important, as genome architecture underpins the diversity, distribution and metabolic versatility of bacteria [25, 26].

Glacier meltwaters are oligotrophic, which may favor genome streamlining and low GC content, as has been observed in other nutrient-limited environments. GFS streamwater is often turbid due to high loads of fine suspended sediments [15, 27], which reduces light available for primary producers (i.e., benthic algae) and further aggravates resource limitation for heterotrophic bacteria [28]. In line with this, smaller average genome sizes have been reported from GFSs compared tributary streams that are not under glacial influence, albeit draining the same proglacial floodplains [29].

On the other hand, efficient stress response, abundant mobile genetic elements, translational flexibility and genome plasticity have been related to larger genomes of bacteria in cryospheric ecosystems [Margesin & Collins 2019]. In GFSs, rapid variation in flow and sediment loads and low streamwater temperatures may thus promote larger genomes [5]. Moreover, GFS bacteria thrive during windows of opportunity, which mainly arise in spring and autumn when nutrients and light are available and streamflow is moderate [19, 30, 31]. Similarly, bacteria with larger genomes and an expanded functional and regulatory repertoire thrive in in pelagic environments where rapid nutrient influx or depletion can occur [32].

Here, we consider genomic traits prevalent under high glacial influence to reflect selective pressures and thus as being indicative of adaptations to GFS conditions. This perspective is based on the idea that genomic features conferring fitness advantages—such as metabolic efficiency—become prevalent under strong selective constraints, while at the same time being shaped by the need to maintain sufficient functional flexibility to cope with environmental fluctuation [1]. Given the close relationship between genomic traits and evolutionary history, phylogenetic context is critical for interpreting variation in genomic features [33]. To this end, we establish a novel analytical framework

for resolving the phylogenetic signatures of genomic features in GFSs. This null-model based framework allows us to determine the phylogenetic depth at which genomic traits exhibit a significant signal, allowing us to explore how genome variation is structured across phylogenetic scales and to disentangling the contribution of specific clades to community-level genomic patterns. Our work provides new insights into how environmental constraints shape bacterial genome architecture and contribute to the ecological success of specific clades in GFSs. These findings are particularly relevant in the context of climate change, as diminishing glacial influence may alter key selective constraints and potentially threaten microorganisms adapted to the GFS environment.

Material and Methods

Glacier-fed stream sampling and environmental parameters

We sampled benthic biofilms (upper 5 cm of the streambed) from 149 GFSs in the European Alps, Scandinavian Mountains, Himalayas, Pamir and Tian Shan, Ecuadorian Andes, Southwest Greenland, Russian Caucasus, Rwenzori in Africa, and Southern Alps in New Zealand between January 2019 and July 2022. GFSs were sampled in spring or autumn during 'windows of opportunity' when streamflow and streamwater turbidity are relatively low; this sampling strategy facilitates comparability between GFSs. We did not sample GFSs from heavily debris-covered and rock glaciers, and we avoided GFSs downstream of proglacial lakes, with debris flows, or tributaries in the reaches above the sampling sites. At each GFS, we sampled an upstream reach, as close as possible to the glacier snout, and a downstream reach. Within each reach, sandy sediments (250 µm to 3.15 mm size fraction) were collected from three independent patches (approximately 10 m apart). All sampling devices were flame-sterilised in the field. Sediment samples were transferred into sterile cryovials, immediately flash-frozen in liquid nitrogen in the field and subsequently stored at -80°C before and following shipping to Switzerland for DNA extraction and biomass analyses.

For each GFS, the distance to the glacier snout was calculated based on georeferencing (GPSMAPR 66s, GARMIN) of the sampling reach, as well as glacier surface area and glacierized percentage catchment based on satellite imagery (Sentinel-2; Level 2a, March 2019 - July 2022 from *scihub.copernicus.eu*) and a catchment definition derived from the ASTER Global Digital Elevation Model (GDEM) v3. (NASA/Meti/Aist/Japan Spacesystems and US/Japan Aster Science Team, 2019). The glacier index (GI) was calculated as $\frac{\sqrt{Glacier\ area}}{\sqrt{Glacier\ area} + Distance\ to\ the\ glacier}$ according to Jacobsen & Dangles (2011)[34]. Benthic chlorophyll-*a*, a proxy for algal biomass, was extracted from the sediment (90% EtOH) in a hot (78°C) water bath for 10 min and further incubated (24 h, 4°C). After vortexing and centrifugation, chlorophyll-*a* concentration in the supernatant was quantified using a plate reader (BioTek Synergy H1; EX/EM: 436/680) and a spinach chlorophyll-a standard (Sigma Aldrich) and normalised to dry mass (DM) of sediment.

Metagenomics

Metagenomes were sequenced for 149 sediment samples. DNA extraction, purification, library preparation, sequencing and metagenome assembly steps were performed as described elsewhere [19]. Briefly, 5 g of sediments were treated using a phenol:chloroform-based extraction method subsequently followed by an ethanol precipitation step. This protocol yielded on average 50 ng of DNA per sample which was used for library preparation using the NEBNext Ultra II FS library kit, which also included 6 PCR cycles. Sequencing was performed at the Functional Genomics Centre Zurich using a S4 flowcell on a NovaSeq (Illumina).

The metagenomic sequence data was processed using the Integrated Meta-omic Pipeline (IMP) workflow (version 3.0; commit# 9672c874)[35]. Briefly, adapter trimming from reads using *trimmomatic* [36] is followed by an iterative assembly using *MEGAHIT* [37] and *Flye* [38]. To reduce computation time for binning, we removed sequences in the assembly < 1.5 kbp and randomly selected 10% of the pre-processed reads using *seqtk* (v1.3)[39]. For each individual assembly, we

then mapped the selected reads of the 5 spatially closest samples (Euclidean distances of gps coordinates) using *BWA-mem* (v0.7.17). We then used *MetaBAT2* (v2.15)[40], *CONCOCT* (v1.1.0)[41] and *MetaBinner* (v1.4.3)[42] using default parameters to obtain bins {see Code availability}. The quality of bins was assessed with *CheckM2* (v1.0.1)[43], and finally *DASTool* (v1.1.4)[44] was employed to generate a non-redundant set of bins using a score threshold of 0.3.

Bins from all samples (including the ones generated by *IMP3*) with a completeness of more than 50% were then selected for further analyses which accounted for 12,599 bins. We then used *MDMCleaner* (v0.8.3)[45] to reduce contamination of the bins. Finally, after rerunning *CheckM2* on the bins to get final estimates of completeness and contamination, we used *dRep* (v3.2.2)[46] to dereplicate bins using a minimum completeness of 70% and maximum contamination of 10% and an ANI of 99% to obtain 2855 strain-level MAGs. GTDB-Tk (v 2.1)[43, 47] was used to assign taxonomy to MAGs. We further used the concatenated alignment of 120 ubiquitous single-copy proteins created by GTDB-Tk to *de novo* generate a phylogenetic tree using *FastTree2* (v2.1.11)[48] under the WAG model of protein evolution with gamma-distributed rate heterogeneity. Functional annotation of the MAGs was performed with *eggNOG-Mapper* (v2.1.9)[49] after obtaining coding regions (CDS) with *prodigal* (v2.6.3)[50]. The coverage of MAGs was estimated by mapping reads of samples to the genomic contigs using *CoverM* (v0.6.1, available at https://github.com/wwood/CoverM) using the *trimmed_mean* parameter. We normalised the coverage by similarly mapping reads on the recA gene (K03553). For prevalence, presences were defined as abundance above a 10x recA coverage abundance threshold. However, one should keep in mind that metagenome-based analyses can not differentiate between active, dormant or dead cells and that dispersal from upstream habitats may also influence patterns of prevalence in our dataset.

Dimensions of glacial influence and community-weighted mean genomic properties

To identify the main environmental gradients across all GFS samples, Principal Component Analysis (PCA) was performed with the *prcomp* function in R (version 4.3.0), and using a non-redundant set

of key physico-chemical as well as glacier-associated measures (glacier area, glacier coverage, glacier index, streamwater temperature, distance to the glacier, benthic chlorophyll-*a*). Community-weighted means (CWM) of genomic features (i.e. genome size, gene number, tRNA number, GC content, coding density and gene redundancy index) were tested with linear effects against the first two principal components using generalised additive models (GAMs) created with the *bam* function of the *mgcv* R package (v1.9.0)[51]. For this, genomic features were first normalised using completeness and contamination as follows: $value_{normalised} = value * (1/completeness) * (1 - contamination)$. CWM were then obtained by weighing normalised genomic features by MAG relative abundances and averaging across MAGs present in any given sample. To account for large-scale spatial patterns, we used a smoothed spline (bs = 'sos', k = -1) based on latitude and longitude in the GAMs. Detailed results of these GAMs are available in Supplementary Table 1. Significant linear effects ($p<0.01$) were visualised using mean and standard errors of predictions across all GFS in the dataset. All figures were created using the *ggplot2* (version 3.4.3) and *ggpubr* (version 0.6.0) *R* packages (R version 4.3.0)[52, 53].

Abundance-based phylogenetic permutation

To resolve the phylogenetic structure of CWM genomic features, we developed a null-model approach that randomly permutes abundances in a phylogenetic-bin based framework. For 40 values of relative phylogenetic height (h) uniformly distributed between zero and one (i.e. scanning the phylogenetic tree from the root to the tips), we performed phylogenetic agglomeration using the "average" method of the *hclust* R function on the cophenetic distances obtained with the *cophenetic.phylo* function of the *ape* R package (v5.7-1)[54]. Subsequently, for each value of h, abundances were randomly permuted within phylogenetic bins (20 iterations). Finally, GAMs accounting for spatial structure (i.e., including a smoothed spline (bs = 'sos', k = -1) on latitude and longitude as covariate) were created, testing for a linear effect of glacial influence on genomic features. Hence, this approach tests for associations between CWM genomic features and

environmental parameter compared to null-model expectations across phylogenetic depth. This approach further allows identifying the relative depth at which phylogenetic signal in CWM genomic features appear along the gradients of glacier influence. Significant coefficients were assessed by combining p-values of the linear coefficients over the 20 iterations using Stouffer's method in the *poolr* R package (v1.1-1), the mean and the standard deviation of the coefficients were computed to summarise the null-model permutations [55].

Additionally, this approach allowed us to pinpoint phylogenetic clades contributing to the community-level signal at a specific phylogenetic height. To this end, we used a leave-one-cluster-out approach, computing coefficients with and without a given phylogenetic cluster, and comparing the resulting coefficients' distributions. Wilcoxon tests were used to test for difference in coefficient distributions, a median relative effect was computed comparing the median values with and without the target phylogenetic cluster ((value with – value without) / (value with)). MAG taxonomy was used to summarise genera present within these clades. Additionally, to summarise these results at higher taxonomic level (i.e. to identify bacterial classes with disproportionately many MAGs in a phylogenetic cluster), we performed enrichment analyses using Fisher tests (fisher.test function in R). To account for multiple testing, we used the *p.adjust* R function using the Holm method.

Functional potential

To unravel the functional potential associated with increased gene redundancy, we tested for each KO if the number of copies was higher in the MAGs that were part of significant clades compared to all other MAGs. We performed Wilcoxon tests (*wilcox.test* R function) on the log-transformed KO data (half of the minimal non-zero value was added to allow for zeroes in the dataset), and the p-values were adjusted using the *p.adjust* function in R with the 'bonferroni' method. KOs were considered significant if the *p-value* was below 0, and the mean difference above zero. We then compared the KOs for all three relationships using intersects (*intersect* function in R).

We used LASSO regressions to identify functional genes that were associated with clades contributing signal to genomic properties (genome size, gene number, tRNA gene number) in relation to benthic chlorophyll-*a* concentration. For this, log-transformed KO data was used in a LASSO regression to explain the binomial response variable "part of clade" or "not part of clade". The penalisation in this regression type allowed to shrink the coefficient of non-important KOs to keep only KOs with high coefficients. We then compared the KOs for all three relationships using intersects (*intersect* function in R).

Taxonomic summary

CWM genomic features of MAGs classified as *Gammaproteobacteria* were compared to all other MAGs. The taxonomic summary comparing genomic features of MAGs classified as *Gammaproteobacteria* to other taxonomic classes was created using the *dplyr* R package (v1.1.3). Wilcoxon sign rank tests were used to compare the distributions. Relative abundance and prevalence (i.e., the number of occurrences across GFSs) were used as estimates of the 'ecological success' of MAGs. The assumption that abundant and prevalent MAGs in GFSs are ecologically successful is based on previous work , which show that GFS benthic communities assemble deterministically [23] and that benthic communities are distinct from the bacterial community suspended in the streamwater [17]. However, we acknowledge that we present results based on metagenomic dataset, and thus, dormant or inactive cells may be included. GAMs were built using a spline (k=5, bs='ts') for these 'ecological success' covariates, and genome size and coding density were used as response variables. We compared one model with a spline for all MAGs, and one with a different spline for GFS-*Gammaproteobacteria* and all other MAGs (using the 'by' argument in the spline). A Bayes factor analysis was used to compare both models, using the *test_performance* function of the *performance* R package (v0.10.5)[56]. A Bayes factor above 3 was considered significant.

Results and Discussion

Genome characteristics of the GFS microbiome

The GFS environment is directly influenced by glaciers, primarily through the magnitude and variation of meltwater runoff [15, 27]. Runoff determines hydraulic stress, channel stability and sediment loads, while streamwater temperature affects metabolic processes [28, 57]. These physical processes are largely driven by glacier size, which translates into runoff magnitude and variability [57]. Employing PCA on the complete set of measured environmental parameters (complete dataset available as Supplementary Table 2), the first principal component (PC1; 44.6% explained variance) revealed a gradient of benthic chlorophyll-*a* inversely related to glacier area across all studied GFSs (Fig. 1A). This is striking given the overall low chlorophyll-*a* content (median: 0.0056 μg g^{-1} dry mass; IQR: 0.0007-0.0272) and underscores the responsiveness of benthic primary producers to environmental conditions. High runoff and loads of suspended sediments produced by large glaciers abrade benthic algae and attenuate light, thereby inhibiting primary production in GFSs and keeping chlorophyll-*a* concentrations low [58, 59]. Principal Component 2 (21.7% explained variance) depicts a gradient of streamwater temperature related to both distance to the glacier snout and glacier area (as encapsulated by the glacier index) across all GFSs. Indeed, depending on the magnitude of runoff, streamwater warms with increasing distance from the glacier. Taken together, the PCA reveals two main dimensions of glacial influence on GFSs at a global scale, and we will explore them as potential underpinning processes of the genomic landscape of the GFS microbiome.

Weighted by relative abundance of MAGs, bacterial genomes across all GFSs were relatively large in terms of size, had a high number of genes, and showed high GC content (Fig. 1B). These values are bracketed by those reported from other GFSs [29], various cryospheric ecosystems (e.g., permafrost, glacier ice) [6], and psychrophiles [5]. Bacterial genomes generally contain only little non-coding DNA (on average, ORFs account for 87% of genome size [60]). Hence, variation in gene

number and genome size are generally tightly linked [2], a relationship attributed to the importance of effective population size [3]. GC content, coding density and genome size have also been shown to positively correlate in bacteria [5, 61, 62]. However, compared to psychrophilic, mesophilic, and thermophilic bacterial isolates [63], we found a relatively low number of tRNAs, which we mainly attribute to the discrepancy between MAGs and isolates owed to metagenomic assembly and binning [64, 65]. Because translation is energetically expensive, tRNA abundance has been linked to shorter minimal generation time and adaptability to different environmental conditions [66]. The gene redundancy index (i.e., the ratio between the total number of KOs to the number of unique KOs with a genome, median RI ~ 1.4) was lower than previously reported in cryoconite biofilms [67], which we attribute to the dynamic and unpredictable GFS environment that may select for functional plasticity rather than redundancy within a given genome.

To further explore glacier influence on these genomic properties of the GFS microbiome, we implemented GAMs accounting for large-scale spatial variation and isolating linear effects of environmental parameters on genomic properties. GAMs revealed positive associations between benthic chlorophyll-*a* content (correlated with PC1) with average genome size, gene number, and tRNA number, whereas covariates correlating with PC2 (i.e., water temperature, distance from the glacier and glacier index) were associated with the gene redundancy index (Fig. 1C). These findings are in line with previous work suggesting that benthic algae, through the exudation of energy-rich macromolecules, relieve GFS bacteria from energy and carbon limitation [28], ultimately promoting bacteria with larger genomes as glaciers shrink and benthic algal biomass increases [26]. Indeed, metabolic interactions between microbial heterotrophs and algae have been repeatedly reported from stream biofilms [68, 69], which may be particularly important in GFSs largely devoid of allochthonous sources of organic carbon [19, 28]. Furthermore, these analyses revealed increasing numbers of tRNAs with diminishing glacial influence, which essentially follows the observed trends in genome size (Fig. 1C). While tRNAs have been associated with cold adaptation and post-translational modifications in bacteria [70, 71], work on isolates showed that psychrophile genomes

have elevated numbers of tRNAs [63]. Nevertheless, translational efficiency has been shown to be low in organisms that are able to thrive in multiple habitats, and this could potentially explain the low number of tRNAs that we observed [66]. Importantly, our analyses have not revealed any major variation in coding density along any of the glaciological variables tested. In line with expectations [3], this suggests that genome size, number of genes and thus the proportion of non-coding DNA vary concomitantly across environmental gradients in GFS.

Dissecting the phylogenetic signatures of genomic trends along environmental gradients

Variation in community-level genomic properties along environmental gradients can either arise from changes in abundance or the replacement of taxa with different genomic characteristics. Moreover, shared evolutionary histories of microbiome members can shape relationships between genomic properties and environmental constraints [33]. For example, accounting for phylogenetic dependencies, a previous study identified deep phylogenetic signatures in genome size variation of bacteria and archaea [33]. To assess phylogenetic signatures in genomic features, we developed a null model-based approach to first identify the phylogenetic depth at which signal in genomic properties along environmental gradients arise. Using a leave-one-out approach of individual clades at the identified threshold phylogenetic distance, we then find clades that contribute most to this signal. Finally, we investigate the functional potential of these clades in comparison to other community members, to uncover functional traits associated with community-level genomic properties.

We found significant phylogenetic signature exclusively at low depth (i.e., among closely related members) for relationships between the gene redundancy index and streamwater temperature, glacier index and distance to the glacier (below 0.25 relative phylogenetic tree height, corresponding approximately to median genus-level phylogenetic depth; Fig. 2A). This suggests that variation in the gene redundancy index is predominantly structured among closely related taxa.

Using the leave-one-cluster-out approach, we identified 36 clusters (out of a total of 394 clusters at a phylogenetic depth of 0.25, Supplementary Table 3) containing MAGs classified as *ELB16-189* (n=42), *OLB17* (n=16), *CAILRJO1* (n=10), *Palsa-1315* (n=10), *Deinococcus* (n=5) and *Nitrospira_F* (n=1) to drive relationships between the distance to the glacier and the gene redundancy index. This highlights the fine-scaled yet widely distributed phenomenon that GFS taxa possess increased gene redundancy at decreased glacial influence. Interestingly, more clusters were significant for the distance to the glacier (n=36) compared to the glacier index (n=5) and streamwater temperature (n=1). Both, the glacier index and distance to the snout may integrate the longer-term influence of glaciers on the GFS microbiome whereas streamwater temperature fluctuates on timescales of minutes to hours [72].

To unravel which microbial functions exhibit increased redundancy in GFS microbiomes under reduced glacial influence, we compared the number of gene copies per KO in clades with and without significant relationships between redundancy and glacial influence, respectively. We identify a total of 37 KOs with significantly higher copy numbers (Table 1, Wilcoxon rank-sum test, adjusted p-value < 0.01). These include several genes associated with metabolism, including two genes related to sulfur metabolism (*ddhA*, *ddhB*), two genes encoding methane/ammonia monooxygenase subunits (B and C), and carbon-metabolism related genes (*acsE*, *ccsB*, *sucD*, *korD*). This observation aligns with previous findings [26, 28], who reported that declining environmental selection in glacier-fed streams promotes primary production, leading to shifts in microbiome functions, including changes in energy acquisition pathways. Taken together, increased gene redundancy in metabolic pathways with reduced glacial influence, may point towards an adaptive strategy of microbes to cope with environmental changes in GFSs.

Members of Gammaproteobacteria shape the relationship between genomic features and chlorophyll-a

In contrast to gene redundancy, relationships of genome size, gene number, and tRNA number with benthic chlorophyll-*a* concentration arose already at greater phylogenetic depths (approximately 0.6 relative phylogenetic tree height, corresponding to median class-level depth, Fig. 2B). This signal was conserved across the lower range of the phylogenetic tree. Leave-one-cluster-out analysis highlighted the contribution of a single cluster to signal for all three genome properties - comprising all MAGs classified as *Gammaproteobacteria* in our dataset (termed GFS-Gammaproteobacteria, Supplementary Table 4). Additionally, significant changes in genome size and gene number along the benthic chlorophyll-*a* gradient were found for phylogenetic clusters encompassing MAGs classified as *Acidobacteriota*, *Desulfobacterota*, *Myxococcota* and *Nitrospirota*. This finding aligns with previous work on GFS community assembly, which found that homogeneous selection promotes microdiversity among *Gammaproteobacteria* (from the *Burkholderiales* order previously assigned to *Betaproteobacteria*) and *Nitrospira* among a few other taxa [23]. Moreover, these results highlight the importance of chlorophyll-*a* in profoundly shaping the structure of the GFS microbiome. We deem the fact that chlorophyll-*a* concentration, a biological factor, is more important in shaping deep-rooting genomic signatures than physical factors (e.g. temperature) particularly relevant considering the importance of algal-bacterial interactions [19] and pronounced carbon limitation in GFS [28]. This may point to the long-term coherence of these drivers– which are now changing in GFSs due to climate-change induced retreat of glaciers.

Given the abundance and prevalence of GFS-Gammaproteobacteria [18, 21, 23], we next investigated the genomic properties of GFS-Gammaproteobacteria in relation to glacier influence (Fig. 3). Indeed, we found a strong negative relationship between relative abundance of GFS-Gammaproteobacteria and benthic chlorophyll-*a* (Fig. 3A). GFS-Gammaproteobacteria had significantly increased coding density (median difference: 2%), but fewer tRNAs (median difference: 3.05), and a lower gene

redundancy index (median difference: 0.025) compared to all other MAGs in our dataset (Fig 3B). On the other hand, genome size and gene numbers of GFS-Gammaproteobacteria were not significantly different from other MAGs. This contrasts our findings on community-weighted average genomic features and suggests that abundance differences of GFS-Gammaproteobacteria across gradients of glacier influence may contribute to the microbiome-weighted averages.

Next, we examined relationships between genomic properties and prevalence and mean relative abundance of MAGs across our global repository of GFSs. Looking at the distribution of mean relative abundance and prevalence, we find that the GFS-Gammaproteobacteria harbour representatives with high values (Fig. 4 A & D, Wilcoxon tests, adjusted p-values < 0.001, log median difference = 0.54 for both, relative abundance and prevalence). Additionally, we found positive relationships between genome size and MAG prevalence and relative abundance (Fig. 4 C & D, whereas coding density was negatively related to prevalence and abundance (Fig. 4 E and F). Using GAMs and a Bayes factor analysis, we tested whether these relationships differed between GFS-Gammaproteobacteria and other MAGs. A GAM with separate splines for GFS-Gammaproteobacteria and other MAGs was better supported by the data (Bayes factor > 1000 for all comparisons) than a GAM with one spline for all MAGs (Fig. 4). This indicates that GFS-Gammaproteobacteria combine increased coding density with reduced genome size compared to classes that are similarly abundant and prevalent in GFSs. Given the compositional nature of microbiomes, this relationship could in part (at low prevalence) also be driven by the low prevalence of symbiotic *Patescibacteria* that have particularly small and streamlined genomes and seem to show low dispersal capabilities [73].

However, that the signal in genomic properties along the chlorophyll-*a* gradient was conserved across phylogenetic depths indicates that not the entire GFS-Gammaproteobacteria clade but rather specific sub-clades may drive this relationship. Indeed, the leave-one-cluster-out analysis performed at shallower phylogenetic depth (relative phylogenetic depth = 0.2) identified specific clades including

Polaromonas (n_{MAGs} = 25), *Rhodoferax* (n_{MAGs} = 21), JAAFIP01 (n_{MAGs} = 23), *Aquabacterium_A* (n_{MAGs} = 20), and *Rubrivivax* (n_{MAGs} = 27). Other notable taxa included *Novosphingobium* (n_{MAGs} = 42) and the *Patescibacteria* genus *OLB19* (n_{MAGs} = 32). The signal among multiple genus-level clades across the GFS-Gammaproteobacteria suggests that the observed increase in genome size with higher chlorophyll-*a* concentrations (or, inversely, the reduction in genome size under high glacial influence when benthic chlorophyll-*a* concentration is particularly low) may result from either convergent evolution across diverse lineages or an early adaptive expansion within the GFS-*Gammaproteobacteria*. However, further phylogenomic analyses would be needed to be better understand the mechanisms and timescales of these processes. As a first step, we provide comparative pangenome analyses between GFS-Gammaproteobacteria and their sister clade in SI (Supplementary information, section "Pangenome analyses").

We next sought to identify the functional potential associate with the differences in genome size, gene number and tRNA number along gradients of chlorophyll-*a*. We applied lasso regression to pinpoint KOs associated with clades exhibiting significant signal. This approach enabled us to identify functions that were enriched in these clusters, representing candidate drivers of expanded functional potential. While we observed 47 KOs that were shared for genome size and gene number (Table 2), not a single significant clade (and hence KO) was found for the relationship between tRNA number and chlorophyll-*a* at shallower depth. This may be attributable to a generally weaker signal observed for tRNA compared to genome size and gene number (as reflected in the p-values of the GAMs), and potentially also to the tendency for tRNA genes to be underrepresented on MAGs due to metagenomic binning.

Nevertheless, for genome size and gene number, several metabolic pathways—including pyruvate metabolism, the glyoxylate cycle, and nucleotide biosynthesis—were represented by key enzymes such as malate dehydrogenase (decarboxylating), glyoxylate/hydroxypyruvate/2-ketogluconate reductase, and ribonucleotide reductases. Nitrogen and sulfur metabolism were also represented, with

genes like *napB*, *cynT*, and *sqr* suggesting chemolithoautotrophy, typical of oligotrophic glacier-related systems [21, 74, 75]. Notably, genes involved in quorum sensing and secondary metabolite biosynthesis—such as the *aroF/G/H* cluster and *mxcG* (non-ribosomal peptide synthetase)—point to increased microbial interactions and competition at reduced glacial influence, which is compatible with the "greening" of GFSs [26, 28]. Two-component systems (*narL*, *glnG*) and secretion-related proteins (*virB3*) further highlight regulatory complexity linked to environmental responsiveness, potentially a crucial adaptation to the fluctuating environmental conditions of GFSs. Overall, and given the taxonomic diversity and variety of functional adaptations observed, more targeted, taxon-specific analyses will be necessary to gain deeper insights into the ecological strategies of individual lineages.

Conclusions

Evolutionary history and environmental constraints shape the genomic architecture of microbial communities, ultimately with consequences for diversity and function. Here, we developed a phylogeny-rooted analytical framework that unravels signatures of genomic trends in the world's GFSs. The approach allows pinpointing the phylogenetic depth at which these signatures arise and the importance of individual clades at shaping community-level genomic features. We find significant variation in genome size, gene number, tRNA gene numbers, and modulation of genomic redundancy along gradients of glacial influence. Collectively, our findings suggest that the selective constraints in GFSs explain microbiome-level patterns in genome architecture and that changes in genomic features mainly occur via changes in abundance among specific GFS-Gammaproteobacteria clades. We deem these findings critical because the deep phylogenetic rooting of these signatures reflects the long-term and putatively consistent nature of this extreme environment, which is now changing at a rapid pace owing to climate change.

Acknowledgments

We would like to express our deepest gratitude to A. McIntosh and L. Morris in New Zealand, J. Abermann and T. Juul-Pedersen in Greenland, O. Solomina and T. Kuderina Maratovna in Russia, V. Crespo-Pérez and P. Andino Guarderas in Ecuador, J. Yde and S. Leth Jørgensen in Norway, S. Sharma and P. Joshi in Nepal, N. Shaidyldaeva-Myktybekovna and R. Kenzhebaev in Kyrgyzstan, J. Nattabi Kigongo, R. Nalwanga, and C. Masembe in Uganda, M. Gonzaléz and J. Luis Rodriguez in Chile, and C. Kuhle and P. Tomco in Alaska for their logistical support. We extend our appreciation to the many porters and guides in Nepal, Uganda, and Kyrgyzstan, without whom the field campaigns would not have been possible. We also want to thank E. Oppliger for general laboratory support and the Bioscience Core Lab at KAUST for DNA sequencing and three anonymous reviewers for their valuable feedback on the manuscript.

Funding Statement

This research is part of the "Vanishing glaciers" project awarded by the NOMIS foundation to TJB.

Code availability

The code and data used in this study are available on the GitHub repository: https://github.com/Mass23/MAGFS. The MAGs are deposited on NCBI under the BioProject PRJNA781406. Additionally, the code for binning is available on this link: https://github.com/michoug/VanishingGlacierMAGs.

Consortium

The Vanishing Glaciers Field Team

Michael Styllas[1], Martina Schön[1], Matteo Tolosano[1], Vincent de Staercke[1], Hannes Peter[1], Tyler Kohler[2] and Tom J. Battin[1]

1. River Ecosystems Laboratory, Alpine and Polar Environmental Research Center, École Polytechnique Fédérale de Lausanne (EPFL), Sion, Switzerland
2. Department of Ecology, Faculty of Science, Charles University, Prague, Czechia

References

1. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. *ISME J* 2014;**8**:1553–1565. https://doi.org/10.1038/ismej.2014.60

2. Lynch M. Streamlining and Simplification of Microbial Genome Architecture. *Annu Rev Microbiol* 2006;**60**:327–349. https://doi.org/10.1146/annurev.micro.60.080805.142300

3. Bobay L-M, Ochman H. The Evolution of Bacterial Genome Architecture. *Front Genet* 2017;**8**.

4. Coleman ML, Chisholm SW. Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc Natl Acad Sci* 2010;**107**:18634–18639. https://doi.org/10.1073/pnas.1009480107

5. Sabath N, Ferrada E, Barve A, Wagner A. Growth Temperature and Genome Size in Bacteria Are Negatively Correlated, Suggesting Genomic Streamlining During Thermal Adaptation. *Genome Biol Evol* 2013;**5**:966–977. https://doi.org/10.1093/gbe/evt050

6. Bourquin M, Busi SB, Fodelianakis S, Peter H, Washburne A, Kohler TJ, et al. The microbiome of cryospheric ecosystems. *Nat Commun* 2022;**13**:3087. https://doi.org/10.1038/s41467-022-30816-4

7. Dieser M, Smith HJ, Ramaraj T, Foreman CM. Janthinobacterium CG23_2: Comparative Genome Analysis Reveals Enhanced Environmental Sensing and Transcriptional Regulation for Adaptation to Life in an Antarctic Supraglacial Stream. *Microorganisms* 2019;**7**:454. https://doi.org/10.3390/microorganisms7100454

8. Liu Y, Shen L, Zeng Y, Xing T, Xu B, Wang N. Genomic Insights of Cryobacterium Isolated From Ice Core Reveal Genome Dynamics for Adaptation in Glacier. *Front Microbiol* 2020;**11**. https://doi.org/10.3389/fmicb.2020.01530

9. Bentkowski P, Van Oosterhout C, Mock T. A Model of Genome Size Evolution for Prokaryotes in Stable and Fluctuating Environments. *Genome Biol Evol* 2015;**7**:2344–2351. https://doi.org/10.1093/gbe/evv148

10. Konstantinidis KT, Tiedje JM. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci* 2004;**101**:3160–3165. https://doi.org/10.1073/pnas.0308653100

11. Props R, Monsieurs P, Vandamme P, Leys N, Denef VJ, Boon N. Gene Expansion and Positive Selection as Bacterial Adaptations to Oligotrophic Conditions. *mSphere* 2019;**4**:10.1128/mspheredirect.00011-19. https://doi.org/10.1128/mspheredirect.00011-19

12. Rodríguez-Gijón A, Nuy JK, Mehrshad M, Buck M, Schulz F, Woyke T, et al. A Genomic Perspective Across Earth's Microbiomes Reveals That Genome Size in Archaea and Bacteria Is Linked to Ecosystem Type and Trophic Strategy. *Front Microbiol* 2022;**12**.

13. Chuckran PF, Hungate BA, Schwartz E, Dijkstra P. Variation in genomic traits of microbial communities among ecosystems. *FEMS Microbes* 2021;**2**:xtab020. https://doi.org/10.1093/femsmc/xtab020

14. Salcher MM, Schaefle D, Kaspar M, Neuenschwander SM, Ghai R. Evolution in action: habitat transition from sediment to the pelagial leads to genome streamlining in Methylophilaceae. *ISME J* 2019;**13**:2764–2777. https://doi.org/10.1038/s41396-019-0471-3

15. Milner AM, Khamis K, Battin TJ, Brittain JE, Barrand NE, Füreder L, et al. Glacier shrinkage driving global changes in downstream systems. *Proc Natl Acad Sci* 2017;**114**:9770–9778. https://doi.org/10.1073/pnas.1619807114

16. Wilhelm L, Singer GA, Fasching C, Battin TJ, Besemer K. Microbial biodiversity in glacier-fed streams. *ISME J* 2013;**7**:1651–1660. https://doi.org/10.1038/ismej.2013.44

17. Ezzat L, Fodelianakis S, Kohler TJ, Bourquin M, Brandani J, Busi SB, et al. Benthic Biofilms in Glacier-Fed Streams from Scandinavia to the Himalayas Host Distinct Bacterial Communities Compared with the Streamwater. *Appl Environ Microbiol* 2022;**88**:e00421-22. https://doi.org/10.1128/aem.00421-22

18. Ezzat L, Peter H, Bourquin M, Busi SB, Michoud G, Fodelianakis S, et al. Diversity and biogeography of the bacterial microbiome in glacier-fed streams. *Nature* 2025;1–9. https://doi.org/10.1038/s41586-024-08313-z

19. Busi SB, Bourquin M, Fodelianakis S, Michoud G, Kohler TJ, Peter H, et al. Genomic and metabolic adaptations of biofilms to ecological windows of opportunity in glacier-fed streams. *Nat Commun* 2022;**13**:2168.

20. Kohler TJ, Fodelianakis S, Michoud G, Ezzat L, Bourquin M, Peter H, et al. Glacier shrinkage will accelerate downstream decomposition of organic matter and alters microbiome structure and function. *Glob Change Biol* 2022;**28**:3846–3859. https://doi.org/10.1111/gcb.16169

21. Michoud G, Peter H, Busi SB, Bourquin M, Kohler TJ, Geers A, et al. Mapping the metagenomic diversity of the multi-kingdom glacier-fed stream microbiome. *Nat Microbiol* 2025;1–14. https://doi.org/10.1038/s41564-024-01874-9

22. Brandani J, Peter H, Fodelianakis S, Kohler TJ, Bourquin M, Michoud G, et al. Homogeneous Environmental Selection Structures the Bacterial Communities of Benthic Biofilms in Proglacial Floodplain Streams. *Appl Environ Microbiol* 2023;**89**:e02010-22. https://doi.org/10.1128/aem.02010-22

23. Fodelianakis S, Washburne AD, Bourquin M, Pramateftaki P, Kohler TJ, Styllas M, et al. Microdiversity characterizes prevalent phylogenetic clades in the glacier-fed stream microbiome. *ISME J* 2022;**16**:666–675. https://doi.org/10.1038/s41396-021-01106-6

24. Zemp M, Jakob L, Dussaillant I, Nussbaumer SU, Gourmelen N, Dubber S, et al. Community estimate of global glacier mass changes from 2000 to 2023. *Nature* 2025;**639**:382–388. https://doi.org/10.1038/s41586-024-08545-z

25. Ngugi DK, Acinas SG, Sánchez P, Gasol JM, Agusti S, Karl DM, et al. Abiotic selection of microbial genome size in the global ocean. *Nat Commun* 2023;**14**:1384. https://doi.org/10.1038/s41467-023-36988-x

26. Bourquin M, Peter H, Michoud G, Busi SB, Kohler TJ, Robison AL, et al. Predicting climate-change impacts on the global glacier-fed stream microbiome. *Nat Commun* 2025;**16**:1264. https://doi.org/10.1038/s41467-025-56426-4

27. Brown LE, Hannah DM, Milner AM. Vulnerability of alpine stream biodiversity to shrinking glaciers and snowpacks. *Glob Change Biol* 2007;**13**:958–966. https://doi.org/10.1111/j.1365-2486.2007.01341.x

28. Kohler TJ, Bourquin M, Peter H, Yvon-Durocher G, Sinsabaugh RL, Deluigi N, et al. Global emergent responses of stream microbial metabolism to glacier shrinkage. *Nat Geosci* 2024;**17**:309–315. https://doi.org/10.1038/s41561-024-01393-6

29. Michoud G, Kohler TJ, Peter H, Brandani J, Busi SB, Battin TJ. Unexpected functional diversity of stream biofilms within and across proglacial floodplains despite close spatial proximity. *Limnol Oceanogr* 2023;**68**:2183–2194. https://doi.org/10.1002/lno.12415

30. Tolotti M, Brighenti S, Bruno MC, Cerasino L, Pindo M, Tirler W, et al. Ecological "Windows of opportunity" influence biofilm prokaryotic diversity differently in glacial and non-glacial Alpine streams. *Sci Total Environ* 2024;**944**:173826. https://doi.org/10.1016/j.scitotenv.2024.173826

31. Uehlinger U, Robinson CT, Hieber M, Zah R. The physico-chemical habitat template for periphyton in alpine glacial streams under a changing climate. *Hydrobiologia* 2010;**657**:107–121. https://doi.org/10.1007/s10750-009-9963-x

32. Chiriac M-C, Haber M, Salcher MM. Adaptive genetic traits in pelagic freshwater microbes. *Environ Microbiol* 2023;**25**:606–641. https://doi.org/10.1111/1462-2920.16313

33. Martinez-Gutierrez CA, Aylward FO. Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales. *PLOS Genet* 2022;**18**:e1010220. https://doi.org/10.1371/journal.pgen.1010220

34. Jacobsen D, Dangles O. Environmental harshness and global richness patterns in glacier-fed streams. *Glob Ecol Biogeogr* 2012;**21**:647–656. https://doi.org/10.1111/j.1466-8238.2011.00699.x

35. Narayanasamy S, Jarosz Y, Muller EEL, Heintz-Buschart A, Herold M, Kaysen A, et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol* 2016;**17**:260. https://doi.org/10.1186/s13059-016-1116-8

36. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

37. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**:1674–1676. https://doi.org/10.1093/bioinformatics/btv033

38. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 2020;**17**:1103–1110. https://doi.org/10.1038/s41592-020-00971-x

39. Li H. lh3/seqtk. 2023. 2023.

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

41. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. CONCOCT: Clustering cONtigs on COverage and ComposiTion. 2013. arXiv, 2013.

42. Wang Z, Huang P, You R, Sun F, Zhu S. MetaBinner: a high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities. *Genome Biol* 2023;**24**:1. https://doi.org/10.1186/s13059-022-02832-6

43. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023;**20**:1203–1212. https://doi.org/10.1038/s41592-023-01940-w

44. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 2018;**3**:836–843. https://doi.org/10.1038/s41564-018-0171-1

45. Vollmers J, Wiegand S, Lenk F, Kaster A-K. How clear is our current view on microbial dark matter? (Re-)assessing public MAG & SAG datasets with MDMcleaner. *Nucleic Acids Res* 2022;**50**:e76. https://doi.org/10.1093/nar/gkac294

46. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 2017;**11**:2864–2868. https://doi.org/10.1038/ismej.2017.126

47. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2020;**36**:1925–1927. https://doi.org/10.1093/bioinformatics/btz848

48. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* 2010;**5**:e9490. https://doi.org/10.1371/journal.pone.0009490

49. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* 2017;**34**:2115–2122. https://doi.org/10.1093/molbev/msx148

50. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;**11**:119. https://doi.org/10.1186/1471-2105-11-119

51. Wood S. mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. 2023. 2023.

52. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

53. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2023.

54. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 2019;**35**:526–528. https://doi.org/10.1093/bioinformatics/bty633

55. Cinar O, Viechtbauer W. The poolr Package for Combining Independent and Dependent p Values. *J Stat Softw* 2022;**101**:1–42. https://doi.org/10.18637/jss.v101.i01

56. Lüdecke (@strengejacke) D, Makowski (@Dom_Makowski) D, Ben-Shachar (@mattansb) MS, Patil (@patilindrajeets) I, Waggoner P, Wiernik (@bmwiernik) BM, et al. performance: Assessment of Regression Models Performance. 2023. 2023.

57. Zhang T, Li D, East AE, Walling DE, Lane S, Overeem I, et al. Warming-driven erosion and sediment transport in cold regions. *Nat Rev Earth Environ* 2022;**3**:832–851. https://doi.org/10.1038/s43017-022-00362-0

58. Uehlinger U, Robinson CT, Hieber M, Zah R. The physico-chemical habitat template for periphyton in alpine glacial streams under a changing climate. In: Stevenson RJ, Sabater S (eds.), Global Change and River Ecosystems—Implications for Structure, Function and EcosystemServices. Dordrecht: Springer Netherlands, 2010, 107–121.

59. Boix Canadell M, Gómez-Gener L, Ulseth AJ, Clémençon M, Lane SN, Battin TJ. Regimes of primary production and their drivers in Alpine streams. *Freshw Biol* 2021;**66**:1449–1463. https://doi.org/10.1111/fwb.13730

60. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, et al. Insights from 20 years of bacterial genome sequencing. *Funct Integr GENOMICS* 2015;**15**:141–161. https://doi.org/10.1007/s10142-015-0433-4

61. Almpanis A, Swain M, Gatherer D, McEwan N 2018. Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. *Microb Genomics* ;**4**:e000168. https://doi.org/10.1099/mgen.0.000168

62. Bentley SD, Parkhill J. Comparative Genomic Structure of Prokaryotes. *Annu Rev Genet* 2004;**38**:771–791. https://doi.org/10.1146/annurev.genet.38.072902.094318

63. Dutta A, Chaudhuri K. Analysis of tRNA composition and folding in psychrophilic, mesophilic and thermophilic genomes: indications for thermal adaptation. *FEMS Microbiol Lett* 2010;**305**:100–108. https://doi.org/10.1111/j.1574-6968.2010.01922.x

64. Meziti A, Rodriguez-R LM, Hatt JK, Peña-Gonzalez A, Levy K, Konstantinidis KT. The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Appl Environ Microbiol* 2021;**87**:e02593-20. https://doi.org/10.1128/AEM.02593-20

65. Mise K, Iwasaki W. Unexpected absence of ribosomal protein genes from metagenome-assembled genomes. *ISME Commun* 2022;**2**:118. https://doi.org/10.1038/s43705-022-00204-6

66. Arella D, Dilucca M, Giansanti A. Codon usage bias and environmental adaptation in microbial organisms. *Mol Genet Genomics* 2021;**296**:751–762. https://doi.org/10.1007/s00438-021-01771-4

67. Zhang Z, Liu Y, Zhao W, Ji M. Radiation impacts gene redundancy and biofilm regulation of cryoconite microbiomes in Northern Hemisphere glaciers. *Microbiome* 2023;**11**:228. https://doi.org/10.1186/s40168-023-01621-y

68. Battin TJ, Besemer K, Bengtsson MM, Romani AM, Packmann AI. The ecology and biogeochemistry of stream biofilms. *Nat Rev Microbiol* 2016;**14**:251–263. https://doi.org/10.1038/nrmicro.2016.15

69. Haack TK, McFeters GA. Nutritional relationships among microorganisms in an epilithic biofilm community. *Microb Ecol* 1982;**8**:115–126. https://doi.org/10.1007/BF02010445

70. Dalluge JJ, Hamamoto T, Horikoshi K, Morita RY, Stetter KO, McCloskey JA. Posttranscriptional modification of tRNA in psychrophilic bacteria. *J Bacteriol* 1997;**179**:1918. https://doi.org/10.1128/jb.179.6.1918-1923.1997

71. Lorenz C, Lünse CE, Mörl M. tRNA Modifications: Impact on Structure and Thermal Adaptation. *Biomolecules* 2017;**7**. https://doi.org/10.3390/biom7020035

72. Ilg C, Cas[TEL]la E. Patterns of macroinvertebrate traits along three glacial stream continuums. *Freshw Biol* 2006;**51**:840–853. https://doi.org/10.1111/j.1365-2427.2006.01533.x

73. Tian R, Ning D, He Z, Zhang P, Spencer SJ, Gao S, et al. Small and mighty: adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. *Microbiome* 2020;**8**:51. https://doi.org/10.1186/s40168-020-00825-w

74. Wei J, Fontaine L, Valiente N, Dörsch P, Hessen DO, Eiler A. Trajectories of freshwater microbial genomics and greenhouse gas saturation upon glacial retreat. *Nat Commun* 2023;**14**:3234. https://doi.org/10.1038/s41467-023-38806-w

75. Nash MV, Anesio AM, Barker G, Tranter M, Varliero G, Eloe-Fadrosh EA, et al. Metagenomic insights into diazotrophic communities across Arctic glacier forefields. *FEMS Microbiol Ecol* 2018;**94**:fiy114. https://doi.org/10.1093/femsec/fiy114

Figure legends

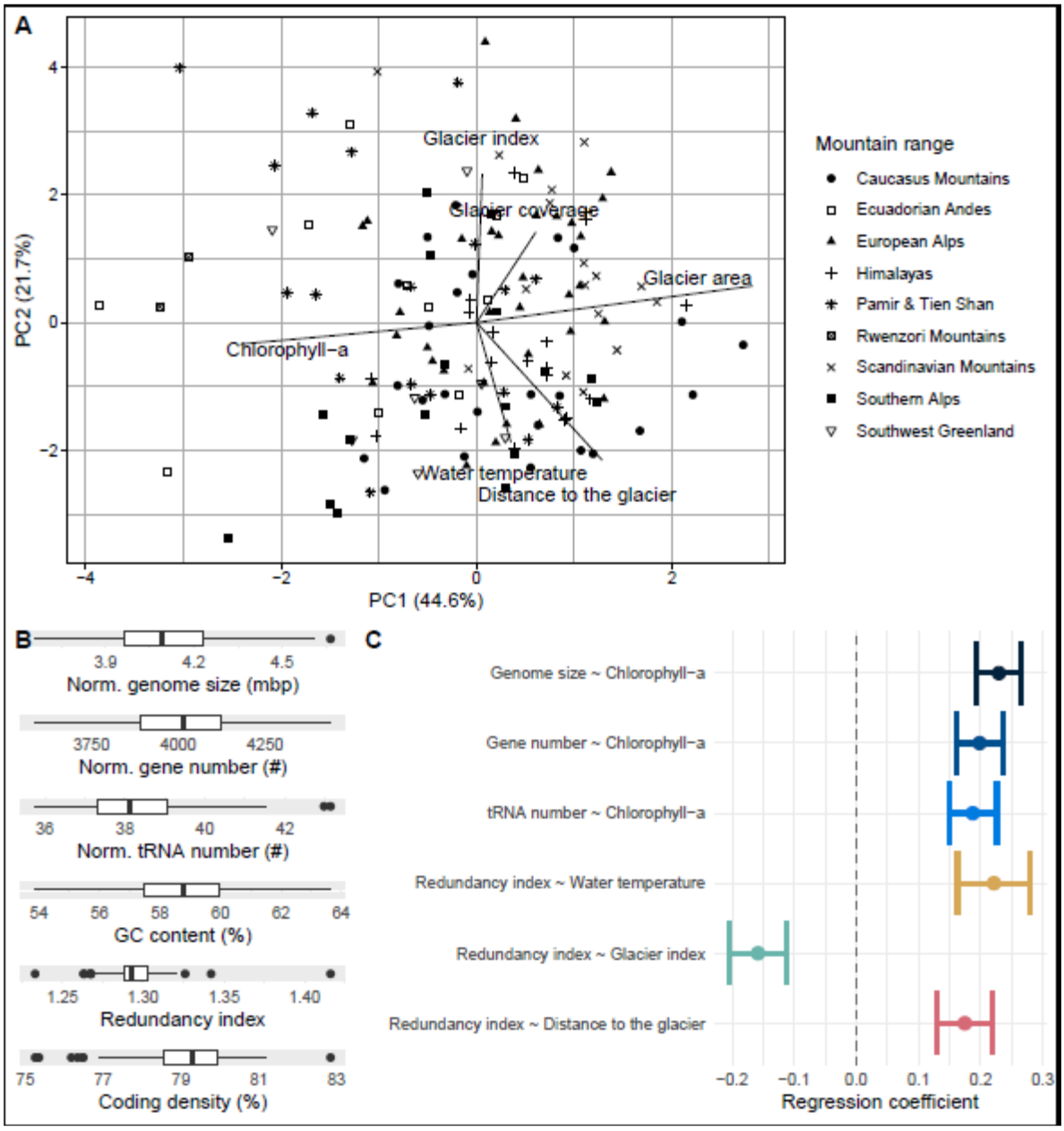

Figure 1. Dimensions of glacial influence and variation in genomic features. (A) The first two dimensions of a principal component analysis (PCA) depict associations among key glacier-associated environmental factors in GFS. Symbols represent mountain ranges; arrows depict scores of environmental variables. (B) Boxplot showing the distribution of community weighted mean genomic features (i.e., weighted with the relative abundance of MAGs) among GFSs. (C) Regression coefficients of genomic features that correlate with glacial covariates in the generalised additive model (GAM) analysis. GAMs considering spatial variations were fitted adding a linear effect for

each pair of genomic features and glaciological variables. Significant relationships after adjusting p-values for multiple testing (Holm's method, $p < 0.05$) are displayed.

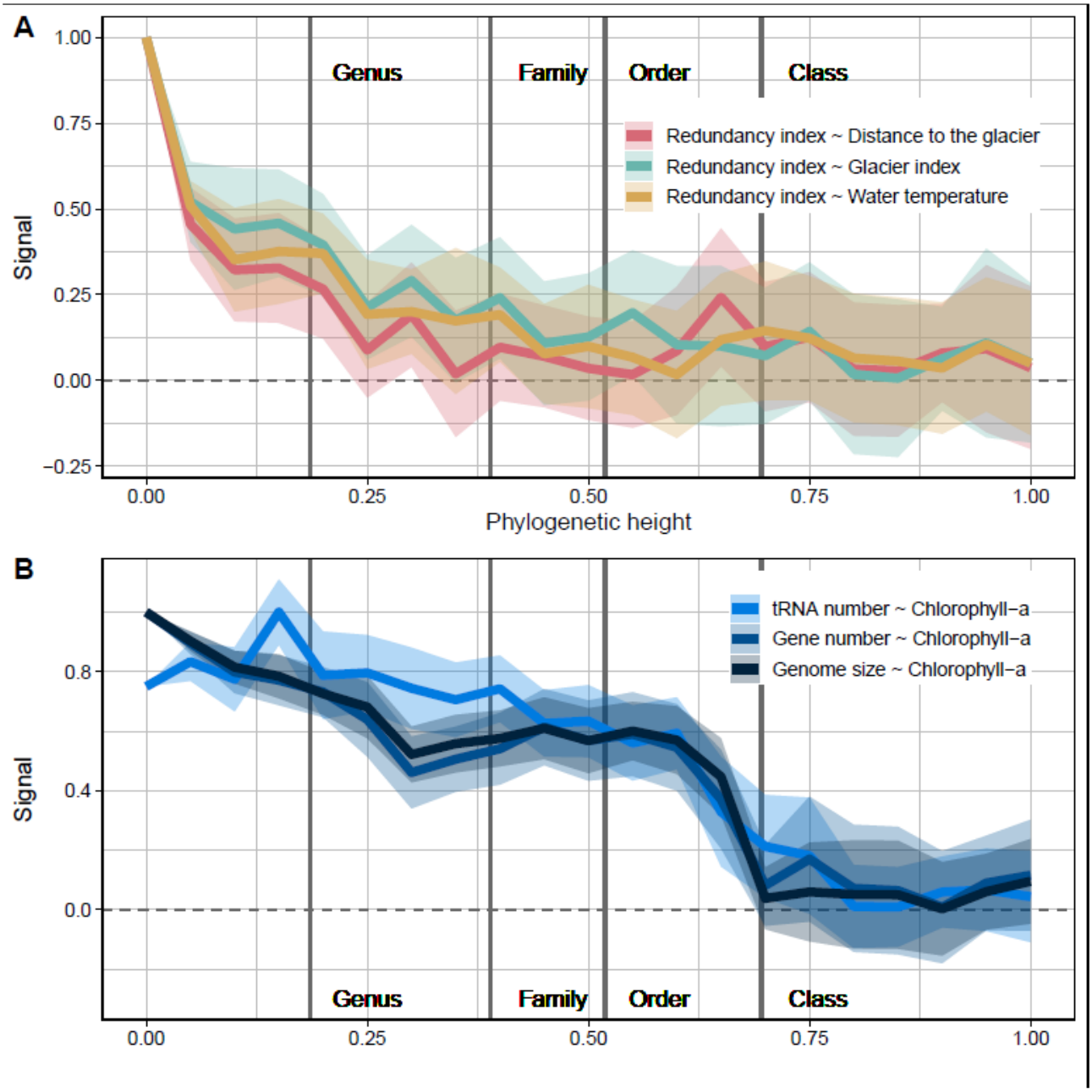

Figure 2. The signal between community-weighted means of genomic features and glacier influence is phylogenetically structured. Line plots displaying the signal in relationships between the gene redundancy index as response variable and the distance to the glacier, the glacier index and the water temperature as covariates (A) and between gene number, genome size and tRNA number and chlorophyll-*a* as covariate (B). The signal was assessed using linear coefficients in the generalised additive models taking spatial variation into account when permuting abundances at various relative phylogenetic heights. Coefficients were normalized by the maximal value for any given glacial covariate-genomic feature pair over the various phylogenetic height values. Shaded areas represent the standard error obtained through 20 null model iterations. Vertical lines indicate median phylogenetic heights for different taxonomic levels and are for visual guidance only.

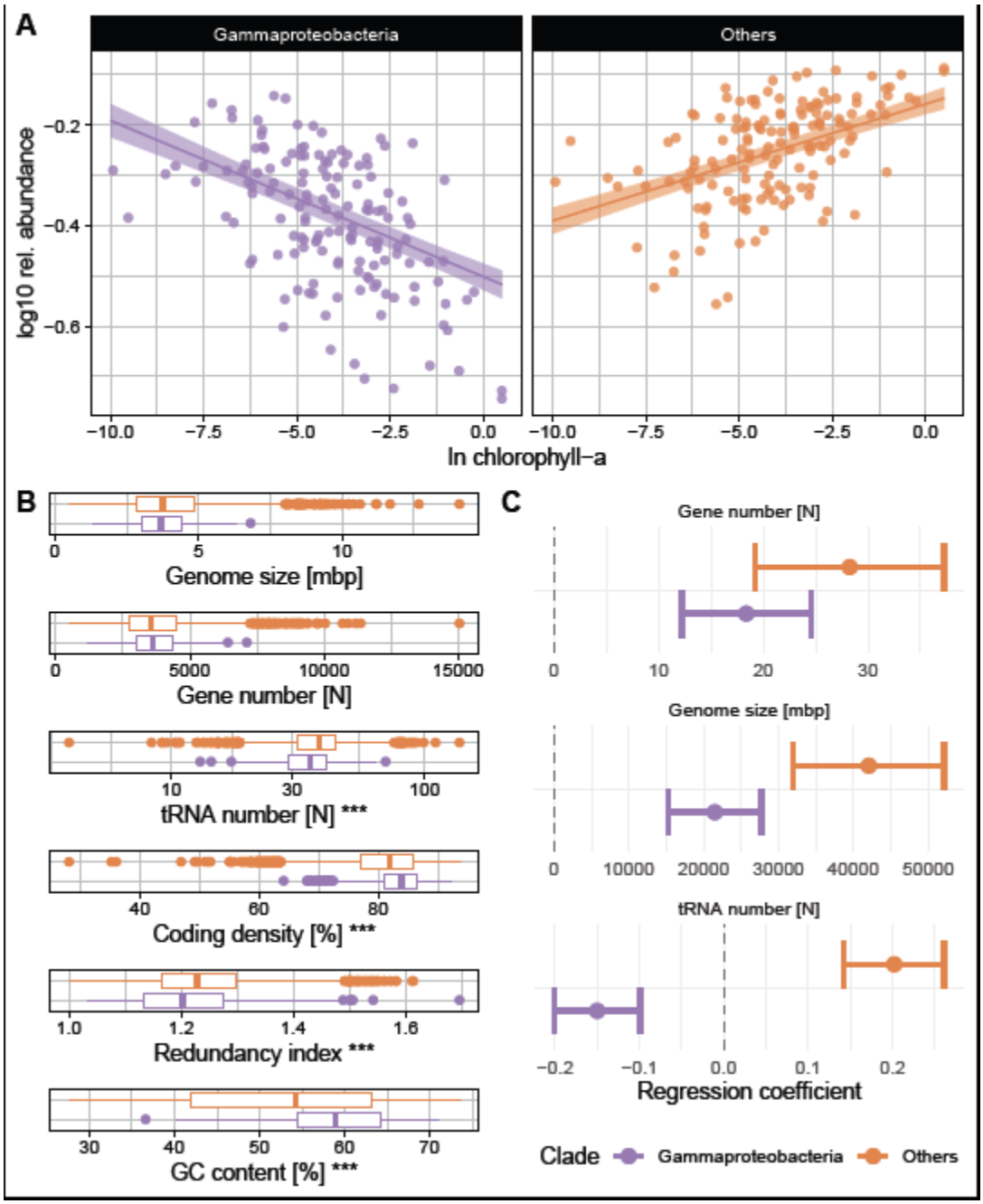

Figure 3. GFS-Gammaproteobacteria drive the variation in genomic features along the gradient of chlorophyll-*a* (A) Scatterplot showing the variation in the relative abundance of GFS-Gammaproteobacteria and all other MAGs along the gradient of benthic chlorophyll-*a* in the world's GFSs. Lines show linear GAM fits accounting for large-scale spatial patterns; shaded areas show prediction intervals. (B) Distributions of genomic features for GFS-Gammaproteobacteria and other MAGs are displayed. Stars denote significance ($p<0.01$) of Wilcoxon signed rank tests comparing the two groups. (C) Linear GAM coefficients representing the variation of genomic feature averages. For all panels, the GFS-Gammaproteobacteria are represented in purple and all other MAGs in orange (see legend).

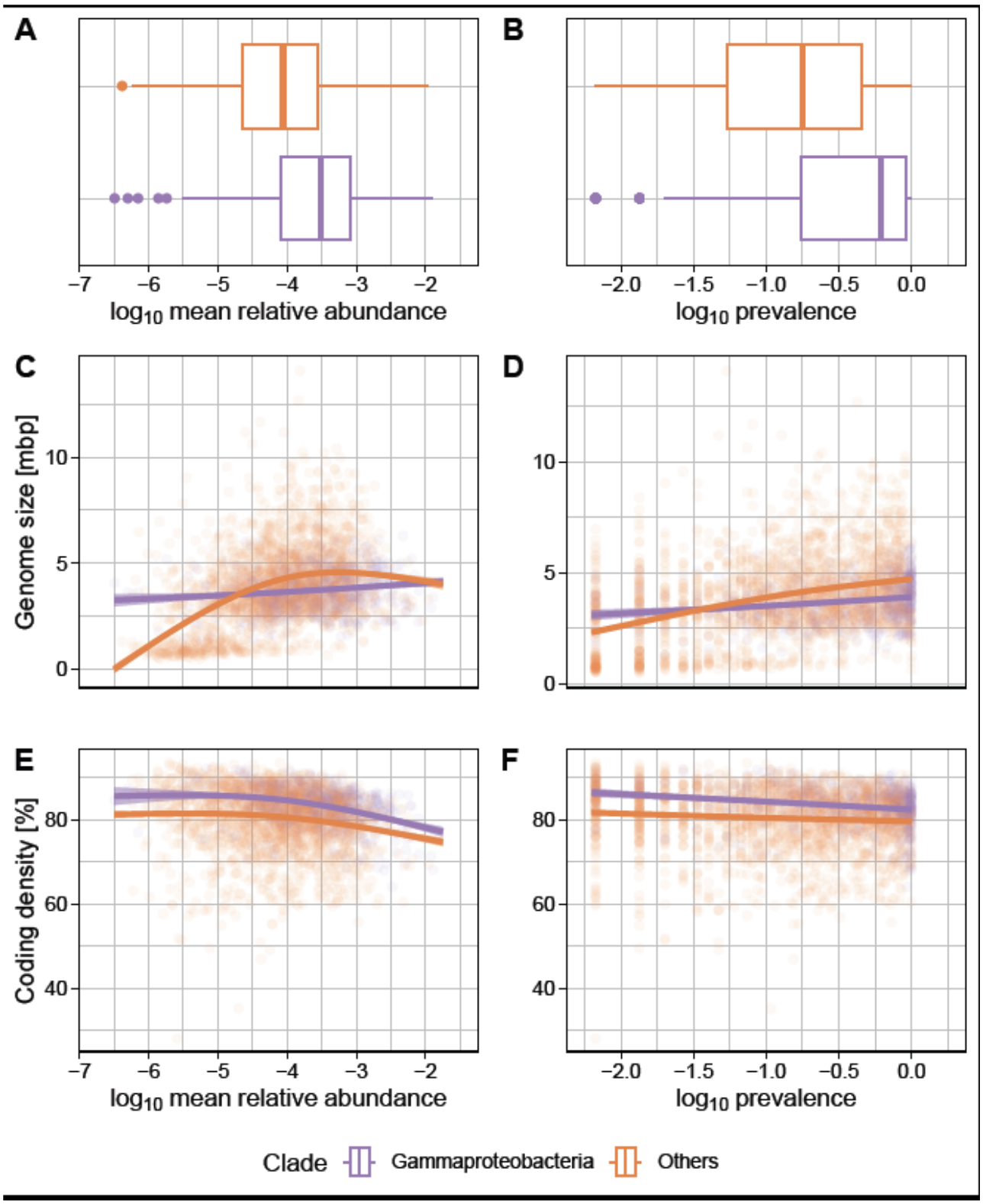

Figure 4. GFS-Gammaproteobacteria are abundant and prevalent. Comparison of relative abundance (A) and prevalence (B) in GFS of MAGs affiliated to *Gammaproteobacteria* (purple) and other classes (orange). Smoothed splines representing GAMs comparing mean abundance (C & E) and prevalence (D & F) with normalised genome size (C & D) and coding density (E & F). Models with separate splines for GFS-Gammaproteobacteria (purple) and all other MAGs (orange) were

better supported (Bayes factor > 1,000) than a combined model. While the difference is driven by high abundance and prevalence of large genomes among other classes (or conversely the absence of small genomes at low abundance and prevalence in GFS-Gammaproteobacteria), GFS-Gammaproteobacteria exhibit increased values for coding density across the entire gradient.

Table 1. KEGG orthologs (KOs) that were significantly ($p < 0.01$) more redundant in the MAGs associated with increase gene redundancy index against all three tested glaciological parameters (distance to the glacier, glacier index, and water temperature). These were tested using Wilcoxon tests, and *p*-values were corrected using the Bonferroni method, only KOs with positive mean differences (i.e. higher redundancy) are displayed. The descriptions and pathways were obtained from the KEGG website (https://www.kegg.jp/entry/).

KO	*Symbol & Description*	*Pathways*
K00003	hom, homoserine dehydrogenase	Glycine, serine and threonine metabolism / Cysteine and methionine metabolism / Lysine biosynthesis / Metabolic pathways / Biosynthesis of secondary metabolites / Microbial metabolism in diverse environments / Biosynthesis of amino acids
K00113	glpC, glycerol-3-phosphate dehydrogenase subunit C	Glycerophospholipid metabolism / Biosynthesis of secondary metabolites
K00176	korD, oorD, 2-oxoglutarate ferredoxin oxidoreductase subunit delta	Citrate cycle (TCA cycle) / Other carbon fixation pathways / Metabolic pathways / Biosynthesis of secondary metabolites / Microbial metabolism in diverse environments / Carbon metabolism / 2-Oxocarboxylic acid metabolism
K00311	ETFDH, electron-transferring-flavoprotein dehydrogenase	
K00373	narJ, narW, nitrate reductase molybdenum cofactor assembly chaperone NarJ/NarW	Two-component system
K00543	ASMT, acetylserotonin O-methyltransferase	Tryptophan metabolism / Metabolic pathways
K00688	PYG, glgP, glycogen phosphorylase	Starch and sucrose metabolism / Metabolic pathways / Biosynthesis of secondary metabolites / Biofilm formation - Escherichia coli
K00979	kdsB, 3-deoxy-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase)	Biosynthesis of various nucleotide sugars / Metabolic pathways / Biosynthesis of nucleotide sugars
K01206	FUCA, alpha-L-fucosidase	Other glycan degradation / Lysosome
K01665	pabB, para-aminobenzoate synthetase component I	Folate biosynthesis / Biosynthesis of cofactors
K01839	deoB, phosphopentomutase	Pentose phosphate pathway / Purine metabolism / Metabolic pathways
K01902	sucD, succinyl-CoA synthetase alpha subunit	Citrate cycle (TCA cycle) / Propanoate metabolism / C5-Branched dibasic acid metabolism / Other carbon fixation pathways / Metabolic pathways / Biosynthesis of secondary metabolites / Microbial metabolism in diverse environments / Carbon metabolism
K02237	comEA, competence protein ComEA	
K03167	top6B, DNA topoisomerase VI subunit B	
K03581	recD, exodeoxyribonuclease V alpha subunit	Homologous recombination
K04477	ycdX, putative hydrolase	
K04767	acuB, acetoin utilization protein AcuB	
K05524	fdxA, ferredoxin	
K05809	raiA, ribosome-associated inhibitor A	
K06998	phzF, trans-2,3-dihydro-3-hydroxyanthranilate isomerase	
K07019	uncharacterized protein	
K07126	uncharacterized protein	
K10700	edbA, ethylbenzene hydroxylase subunit alpha	Ethylbenzene degradation / Metabolic pathways / Microbial metabolism in diverse environments / Degradation of aromatic compounds
K10945	pmoB-amoB, methane/ammonia monooxygenase subunit B	Methane metabolism / Nitrogen metabolism / Metabolic pathways / Microbial metabolism in diverse environments / Carbon metabolism / Nitrogen cycle
K10946	pmoC-amoC, methane/ammonia monooxygenase subunit C	Methane metabolism / Nitrogen metabolism / Metabolic pathways / Microbial metabolism in diverse environments / Carbon metabolism / Nitrogen cycle
K13795	citB, tcuB, citrate/tricarballylate utilization protein	
K15023	acsE, 5-methyltetrahydrofolate corrinoid/iron sulfur protein methyltransferase	Other carbon fixation pathways / Metabolic pathways / Microbial metabolism in diverse environments / Carbon metabolism
K15233	ccsB, citryl-CoA synthetase small subunit	Other carbon fixation pathways / Metabolic pathways / Microbial metabolism in diverse environments / Carbon metabolism

K16130	mcyA, microcystin synthetase protein McyA	Nonribosomal peptide structures
K16964	ddhA, dimethylsulfide dehydrogenase subunit alpha	Sulfur metabolism / Metabolic pathways / Microbial metabolism in diverse environments
K16965	ddhB, dimethylsulfide dehydrogenase subunit beta	Sulfur metabolism / Metabolic pathways / Microbial metabolism in diverse environments
K17048	edbB, ethylbenzene hydroxylase subunit beta	Ethylbenzene degradation / Metabolic pathways / Microbial metabolism in diverse environments / Degradation of aromatic compounds
K17052	serC, clrC, selenate/chlorate reductase subunit gamma	Selenocompound metabolism
K18896	gsmt, glycine/sarcosine N-methyltransferase	Glycine, serine and threonine metabolism / Metabolic pathways
K18897	sdmt, sarcosine/dimethylglycine N-methyltransferase	Glycine, serine and threonine metabolism / Metabolic pathways
K20435	valM, validone 7-phosphate aminotransferase	Acarbose and validamycin biosynthesis / Metabolic pathways / Biosynthesis of secondary metabolites
K21515	aviRa, 23S rRNA (guanine2535-N1)-methyltransferase	

Table 2. KEGG orthologs (KOs) that were identified by the LASSO regression analysis to be enriched in clades driving variations for both genome size and gene numbers, along the gradient of chlorophyll-*a*. The descriptions and pathways were obtained on the KEGG website (https://www.kegg.jp/entry/), human diseases related pathways were not included.

KO	***Symbol & Description***	***Pathways***
K00028	malate dehydrogenase (decarboxylating)	Pyruvate metabolism / Carbon fixation by Calvin cycle / Metabolic pathways / Microbial metabolism in diverse environments / Carbon metabolism
K00090	ghrB, glyoxylate/hydroxypyruvate/2-ketogluconate reductase	Pentose phosphate pathway / Glycine, serine and threonine metabolism / Pyruvate metabolism / Glyoxylate and dicarboxylate metabolism / Metabolic pathways / Biosynthesis of secondary metabolites / Microbial metabolism in diverse environments
K00525	E1.17.4.1A, nrdA, nrdE, ribonucleoside-diphosphate reductase alpha chain	Purine metabolism / Pyrimidine metabolism / Metabolic pathways / Nucleotide metabolism
K00571	E2.1.1.72, site-specific DNA-methyltransferase (adenine-specific)	
K00646	pksF, curC, aprD, corD, malonyl-[acp] decarboxylase	
K01141	sbcB, exoI, exodeoxyribonuclease I	Mismatch repair
K01432	AFMID, arylformamidase	Tryptophan metabolism / Glyoxylate and dicarboxylate metabolism / Metabolic pathways / Biosynthesis of cofactors
K01491	folD, methylenetetrahydrofolate dehydrogenase (NADP+) / methenyltetrahydrofolate cyclohydrolase	One carbon pool by folate / Other carbon fixation pathways / Metabolic pathways / Microbial metabolism in diverse environments / Carbon metabolism / Biosynthesis of cofactors
K01626	E2.5.1.54, aroF, aroG, aroH, 3-deoxy-7-phosphoheptulonate synthase	Phenylalanine, tyrosine and tryptophan biosynthesis / Metabolic pathways / Biosynthesis of secondary metabolites / Biosynthesis of amino acids / Quorum sensing
K01673	cynT, can, carbonic anhydrase	Nitrogen metabolism / Metabolic pathways
K01952	PFAS, purL, phosphoribosylformylglycinamidine synthase	Purine metabolism / Metabolic pathways / Biosynthesis of secondary metabolites
K02083	allC, allantoate deiminase	Purine metabolism / Metabolic pathways / Microbial metabolism in diverse environments
K02568	napB, nitrate reductase (cytochrome), electron transfer subunit	Nitrogen metabolism / Metabolic pathways / Microbial metabolism in diverse environments / Nitrogen cycle
K03169	topB, DNA topoisomerase III	
K03198	virB3, lvhB3, type IV secretion system protein VirB3	Bacterial secretion system
K03442	mscS, small conductance mechanosensitive channel	
K03775	slyD, FKBP-type peptidyl-prolyl cis-trans isomerase SlyD	
K03818	wcaF, putative colanic acid biosynthesis acetyltransferase WcaF	Exopolysaccharide biosynthesis
K03832	tonB, periplasmic protein TonB	
K05962	protein-histidine pros-kinase	
K06192	pqiB, paraquat-inducible protein B	
K07025	putative hydrolase of the HAD superfamily	

K07114	yfbK, Ca-activated chloride channel homolog	
K07343	tfoX, DNA transformation protein and related proteins	
K07684	narL, two-component system, NarL family, nitrate/nitrite response regulator NarL	Two-component system
K07712	glnG, ntrC, two-component system, NtrC family, nitrogen regulation response regulator GlnG	Two-component system
K10012	arnC, pmrF, undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase	Biosynthesis of various nucleotide sugars / Metabolic pathways / Cationic antimicrobial peptide (CAMP) resistance
K10537	araF, L-arabinose transport system substrate-binding protein	ABC transporters
K12055	parA, chromosome partitioning related protein ParA	
K12500	tesC, thioesterase III	
K12601	SKI8, superkiller protein 8	RNA degradation
K12602	WDR61, REC14, SKI8, WD repeat-containing protein 61	RNA degradation
K13117	DHX35, ATP-dependent RNA helicase DDX35	
K14160	imuA, protein ImuA	
K14742	tsaB, tRNA threonylcarbamoyladenosine biosynthesis protein TsaB	
K15653	mxcG, nonribosomal peptide synthetase MxcG	Biosynthesis of siderophore group nonribosomal peptides
K17218	sqr, sulfide:quinone oxidoreductase	Sulfur metabolism / Microbial metabolism in diverse environments
K20036	dmdD, (methylthio)acryloyl-CoA hydratase	Sulfur metabolism / Metabolic pathways / Microbial metabolism in diverse environments
K20534	gtrB, polyisoprenyl-phosphate glycosyltransferase	
K20906	hcmA, 2-hydroxyisobutanoyl-CoA mutase large subunit	
K21211	ncsC1, NDP-hexose 4,6-dehydratase	
K21394	Biosynthesis of enediyne antibiotics / Biosynthesis of secondary metabolites	
K21405	acoR, sigma-54 dependent transcriptional regulator, acetoin dehydrogenase operon transcriptional activator AcoR	
K21739	rclA, probable pyridine nucleotide-disulfide oxidoreductase	
K21843	TTC7, tetratricopeptide repeat protein 7	