Climate change lowers diversity and functional potential of microbes in Canada's high Arctic

Graham A. Colby[1], Matti O. Ruuskanen[1], Kyra A. St. Pierre[2], Vincent L. St. Louis[2], Alexandre J. Poulain[1], and Stéphane Aris-Brosou[1, 3, *]

[1]University of Ottawa, Department of Biology, 30 Marie Curie Pvt., Ottawa, ON K1N 6N5, Canada
[2]University of Alberta, Department of Biological Sciences, Edmonton, AB T6G 2E9, Canada
[3]University of Ottawa, Department of Mathematics and Statistics, Ottawa, ON K1N 6N5, Canada
[*]sarisbro@uottawa.ca

ABSTRACT

Temperatures and precipitations in the Arctic are expected to increase dramatically over the next century, yet little is known about how microbial communities and their underlying metabolic processes will be affected by these environmental changes. To address this knowledge gap, we adopted a unique space-for-time design to analyse sediments sampled from Lake Hazen, NU Canada. Over the past decade, in this world's largest high Arctic lake by volume, climate change has enhanced glacial melt, resulting in increased annual runoff from the watershed. We exploit the spatial heterogeneity created by varying runoff regimes, conducting metagenomic analyses of lake sediments along these spatial gradients to study how a transition from low to high runoff, used as a proxy for environmental change, affects microbial community structure and functional potential. Here we show that increasing runoff leads to a decrease in taxonomic and functional diversity. While our data suggest that rising Arctic temperatures will negatively impact microbial diversity and nutrient cycling in Arctic lakes, it is still unclear how these losses at the microbial community level will affect biogeochemical cycles, or whether this will lead to feedback loops of uncertain direction and magnitude.

Introduction

Climate change is amplified in polar regions, where near-surface warming has been almost twice that of the global average over the last decade[1–3]. The future is not much brighter, as climate models predict temperature and precipitation increases in the Arctic by as much as 8°C and 40%, respectively, by 2100[4]. These changes are already having dramatic consequences on physical[5–7], biogeochemical[8,9], and ecological[10,11] processes across Arctic ecosystems. Yet, we know very little about how complex Arctic microbial communities respond to environmental changes – even though microbes mediate most global biogeochemical cycles[12,13]. Furthermore, lakes are broadly considered sentinels of climate change, as they integrate physical, chemical and biological changes happening through their watersheds[14]; however, their microbial community structure and function are relatively understudied, in particular in the Arctic.

To date, most of the research performed on microbial communities in Arctic lakes has been limited to studies that were mostly based on partial 16S rRNA gene sequencing[15–18]. While these studies are useful to understand the structure of these microbial communities, they are often biased because they rely on sequence databases

where environmental microbes, specifically from the Arctic, may be underrepresented[19]. More critically, being circumscribed both in space and in time, previous studies only offer snapshots of microbial communities and hence, have a limited power to predict how these communities might respond to climate change.

To address these shortcomings, we focused on Lake Hazen, located within Quttinirpaaq National Park, on northern Ellesmere Island, Nunavut, in the High Arctic (82°N). This lake has the distinctive advantage of being the world's largest High Arctic lake by volume[20], a feature that we harness in a unique manner to predict the effects of climate change on its microbial communities. Instead of adopting a longitudinal design, where the same sites are studied repeatedly over long periods of time[21], we exploited the fact that Lake Hazen presents two unique properties. First, its watershed is already experiencing the effects of climate change, as increasing temperatures in Lake Hazen are leading to more glacial melt, permafrost thaw, and increased runoff from the watershed into the lake[9] (unlike other regions of the globe where runoff is predicted to decrease[22,23]). Second, its tributaries are highly heterogeneous, with eleven glaciers ranging from 6 to 1041 km^2 in surface area, and annual runoff volume approximately scaling with their size (from <0.001 to 0.080 km^3 in 2016)[23]. It is this spatial heterogeneity in runoff that we use here to mimic the temporal effects that climate change may have on High Arctic sediment microbial communities. This space-for-time design[24,25] allowed us to study the effects of rising temperatures, by contrasting low and high runoff regimes to predict how climate change might affect microbial communities in this lake[6]. To this effect, we sampled lake sediments along two transects of contrasted runoff regimes (space), allowing us to analyse the potential impact of climate change (time) on microbial communities. We also sampled soil located in the dried streambeds of the tributaries to assess the potential influence of transported soil communities to lake sediments. To garner an unbiased microbial inventory, we resorted to untargeted metagenomics analyses, which allowed us to reconstruct 300 metagenome assembled genomes (MAGs)[26], and assess the quantitative impact of a change of runoff regime, from low to high, on both the structure of microbial communities and their functional potential. Our results show that a change from low to high runoff, as predicted under climate change scenarios for the region, could lead to a reduced diversity of the microbial community and of its functional potential.

Materials and methods

Sample collection and processing

Sediment and soil cores were collected from Lake Hazen (82°N, 71°W: Figure 1A) between the 10th of May and the 10th of June 2017, when the lake was still completely ice-covered (Supplementary Table 1). Within the watershed, runoff flows from the outlet glaciers along the northwestern shoreline through poorly consolidated river valleys,

depositing sediments at the bottom of Lake Hazen along our two transects, the H1/H2 and L1/L2 sites, respectively. The lake then drains via the Ruggles River along its southeastern shoreline (C sites). The surrounding glacial rivers deliver different amounts of sediments, nutrients and organic carbon unevenly to the lake as a consequence of heterogeneous sedimentation rates (Supplementary Table 2). More specifically, the top 5 cm of sediments from the deep low (L2) and high (H2) runoff sites represented depositional periods of 30 years (1987-2017) and 6 years (2011-2017), respectively (Table sedimentation rates), something to keep in mind when interpreting the results.

Both soil and sediment samples were collected along two transects and can be separated into three hydrological regimes by seasonal runoff volume: low (L transect), high (H transect), and negligible runoff (C sites; Supplementary Table 3). Contamination of samples was minimised by wearing non-powdered latex gloves during sample handling and sterilising all equipment with 10% bleach and 90% ethanol before sample collection. We were most interested in the community composition through space, hence we combined the top 5 cm of sediment and 10 cm of soil for sample extraction and subsequent sequencing. These uppermost layers in the sediment correspond to both the most recent sediment deposition dates[23] and the region of greatest microbial activity[27].

Sediment cores approximately 30 cm in length were collected with an UWITEC (Mondsee, Austria) gravity corer from five locations: C (overlying water depth: 50 m) far from the direct influence of glacial inflows serving as a control site; L1 (water depth: 50 m) and L2 (water depth: 251 m), at variable distances from a small glacial inflow (Blister Creek, <0.001 km^3 in summer 2016); and, H1 (water depth: 21 m) and H2 (water depth: 253 m), located adjacent to several larger glacial inflows (*i.e.*, the Abbé River, 0.015 km^3 and Snow Goose, 0.006 km^3 in 2016). Sediment cores were collected in triplicate per site, approximately 3 m apart. These triplicate cores were ultimately combined to account for site heterogeneity prior to genomic analysis. The top 5 cm of each core was sectioned and placed into Whirlpack bags. These slices were homogenised manually inside of the bags and stored in a -20°C freezer until shipment back to the University of Ottawa where they were then stored at -80°C. Two additional sediment cores per site were collected for O_2, redox and pH microprofiling, as well as analyses of porewater chemistry and loss on ignition (for extended methods see[18]).

In addition to lake sediments, soil samples were collected from three sites in the dried streambeds of the tributaries, on the northern shore between the two transects. Following the same procedure for soil cores (L-Soil, H-Soil, and C-Soil), samples were collected in triplicate approximately 1m apart to account for site heterogeneity, and then combined prior to genomic analysis. After snow and surface vegetation was removed, soil cores approximately 10 cm long were collected using a stainless-steel soil corer. Soil samples were transferred into falcon tubes, homogenised, and stored as described above for the lake sediment samples.

For each sample, DNA from the three replicate sediment cores was extracted and combined. Samples were thawed overnight and 250-400 mg (wet weight; Supplementary Table 4) were then washed in a sterile salt buffer (10 mM EDTA, 50 mM Tris-HCl, 50 mM Na_2 HPO_4 $7H_2O$ at pH 8.0) to remove PCR inhibitors[28,29]. All sample handling was conducted in a stainless-steel laminar flow hood (HEPA 100) treated with UVC radiation and bleach before use. DNA extractions were performed using the DNeasy PowerSoil Kit (MO BIO Laboratories Inc, Carlsbad, CA, USA), following the kit guidelines, except that the final elution volume was 30 μl instead of 100 μl. DNA integrity was validated with a NanoDrop Spectrometer and PCR combined with electrophoresis of the *glnA* gene (Supplementary Figure 1 and Supplementary Table 5). Adequate DNA concentrations for sequencing were reached by combining triplicate extractions for a total volume of 45 μM and a concentration $\geq$ 50 ng/μl (Supplementary Table 4). Positive and negative controls were used to verify the integrity of the PCR amplification. Two kit extraction blanks were absent of DNA and were not sequenced.

Chemical analyses

Redox potential, pH, and dissolved O2 concentration profiles were measured at 100 μM intervals in the field within an hour of collection, using Unisense (Aarhus, Denmark) microsensors connected to a Unisense Field Multimeter. Cores used for porewater chemistry analysis were sectioned in 1 cm intervals into 50 mL falcon tubes, followed by flushing of any headspace with ultra-high-purity N_2 before capping. Sediment porewater was extracted following centrifugation at 4,000 rpm, after which the supernatant was filtered through 0.45 μm cellulose acetate filters into 15 ml tubes, which were then frozen until analysis. Concentrations of $NO_2^- + NO_3^-$, and NH_3, Cl^- were measured in the sediment porewater using a Lachat QuickChem 8500 FIA Ion Analyzer, while total dissolved phosphorus (TDP) and SO_4^{2-} was measured in the sediment porewater using an ion chromatograph at the Biogeochemical Analytical Service Laboratory (Department of Biological Sciences, University of Alberta). However, TDP was removed from data analysis because insufficient porewater was collected to measure TDP at site C. The centrifuged sediments were retained and percentages of $CaCO_3$ and organic carbon (OC) were estimated through loss on ignition[30].

The chemical features of the top 5 cm of the sediment cores were derived from measurements performed at 1 cm intervals throughout the cores. The geochemical properties of each sediment site were summarised using a Principle Component Analysis (PCA) and projections were clustered using Partitioning Around Medoids. The appropriate number of clusters was determined from silhouettes with the hopach[31] R package. The Dunn test was used to compare samples, controlling for multiple comparisons with the Benjamini-Hochberg adjustment.

Sequencing and data processing

Metagenomic libraries were prepared and sequenced by Genome Quebec on an Illumina HiSeq 2500 platform (Illumina, San Diego, CA, USA; Supplementary Figure 2) on a paired-end 125 bp configuration using Illumina TruSeq LT adapters (read 1: AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC, and read 2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT). The DNA from the eight sites (five sediments, three soils) was sequenced, generating over 150 GB of data. Read count summaries were tracked throughout each step of the pipeline for quality control (Supplementary Figure 3). Low quality reads, adapters, unpaired reads, and low quality bases at the ends of reads were removed to generate quality controlled reads with Trimmomatic (v0.36)[32] using the following arguments: phred33, ILLUMINACLIP:TruSeq3-PE-2.fa:3:26:10, LEADING:3 TRAILING:3, SLIDINGWINDOW:4:20, MINLEN:36, CROP:120, HEADCROP:20, AVGQUAL:20. FASTQC (v0.11.8)[33] was then used to confirm that the Illumina adapters were removed and that trimmed sequence lengths were at least 90 bp in length with a Phred score of at least 33.

Reconstruction of environmental genomes and annotation

To reconstruct environmental genomes, metagenomic quality-controlled reads from all samples were coassembled using Megahit[34] software with a k-mer size of 31 and "meta-large" setting (see Supplementary Table 6 for additional summary statistics). EukRep[35] was used to remove any eukaryotic DNA from the contigs prior to the formation of an Anvio (v5)[36] contig database. The contig database was generated by removing contigs under 1000 bp, and gene prediction was performed in the Anvio environment. Sequence coverage information was determined for each assembled scaffold by mapping reads from each sample to the assembled contig database using Bowtie2[37] with default settings. The resulting SAM files were sorted and converted to BAM files using samtools (v0.1.19)[38]. Each BAM file was prepared for Anvio using the "anvi-init-bam" and contig database generated using "anvi-gen-contigs-database". The contig database and BAM mapping files were further used as input for "anvi-profile", which generated individual sample profiles for each contig over the minimum length of 2500 bp. These profiles were then combined using "anvi-merge" and summary statistics for abundance and coverage were generated with "anvi-summarise." Automated binning was performed using CONCOCT[39]. Scaffolds were binned on the basis of GC content and differential coverage abundance patterns across all eight samples. Manual refinement was done using Anvio's refine option (Supplementary Table 7). Kaiju[40] was used to classify taxonomy of the assembled contigs with "anvi-import-taxonomy-for-genes" and aided in the manual refinement process. Open reading frames were predicted with Prodigal (v2.6.3)[41]. Anvio's custom Hidden Markov models were run, along with NCBIs COG[42] annotation to identify protein-coding genes. PFAM[43], TIGRFAM[44], GO terms[45], KEGG enzymes and pathways[46],

and Metacyc pathways[47] were predicted with Interproscan (v5)[48]. These annotations were then combined with the Anvio database with "anvi-import-functions".

Genome completeness and contamination were evaluated on the presence of a core set of genes using CheckM (v1.0.5) "lineage_wf" (Supplementary Table 7)[49]. Only genomes that were at least 50% complete and with less than 10% contamination were further analysed – meeting the MIMAG standard for medium or high-quality genomes[26]. All recovered genomes were used to calculate an average amino acid identity across all genomes using compareM (v0.0.23, function "aai_wf"; https://github.com/dparks1134/CompareM)[50]. CheckM was used again to identify contigs that were not contained in any of the 300 high-quality genomes. As an attempt to "rescue" these unbinned contigs, an alternative binning algorithm MaxBin (v2.0)[51] was employed. An additional 481 genomes were recovered, but were not included in further analysis as only 21 genomes were of average completion >65% (Supplementary Data 1: https://github.com/colbyga/hazen_metagenome_publication/blob/master/Supplemental_Data_1_maxbin2_unbinned_contigs_summary.csv).

Phylogenetic placement of the MAGs

Phylogenetic analyses were performed using two different sets of markers from the Genome Taxonomy Database (GTDB): one for bacteria (120 marker genes) and one for archaea (122 marker genes), as previously been used to assign taxonomy to MAGs[52]. The marker genes were extracted from each genome by matching Pfam72 (v31)[43] and TIGRFAMs73 (v15.0)[44] annotations from GTDB (v86)[52]. Marker genes from each of the 300 genomes were translated using seqinr[53], selecting the genetic code that returned no in-frame stop codon. As some genomes had multiple copies of a marker gene, duplicated copies were filtered out by keeping the most complete sequence. Marker genes that were missing from some genomes were replaced by indel (gap) characters. The markers from each bin were combined with the individual GTDB trimmed multiple sequence alignments. MUSCLE (v3.8.31)[54] was used to align the genomes to the GTDB reference sequences in R (v 3.5.1)[55]. Archaeal sequences were removed from the bacterial alignment on the basis of results from CheckM[49] and independently verified using a custom list of archaea specific marker genes. Alignments were then refined using trimAI[56] and the "-gappyout" parameter. FastTree2[57], recompiled with double precision to resolve short branch lengths, was used to infer maximum likelihood phylogenetic trees from protein sequence alignments under the WAG $+\Gamma$ model[58,59]. The archaeal tree was rooted with Euryarchaeota and the bacterial tree was rooted with Patescibacteria using APE[60]. Trees were visualised and coloured by phylum with ggtree[61].

Community composition of the MAGs

To determine the relative abundance of each genome in the eight samples, sample-specific genome abundances were normalised by sequencing depth [(reads mapped to a genome)/(total number of reads mapped)], making comparisons across samples possible. Genome abundances were generated using the CheckM "profile" function[49]. To determine the average abundance of major taxonomic groups across sites (determined by the phylogenetic analysis described above), the abundances for genomes from the same taxonomic group were summed and visualised using phyloseq[62] (usually at the phylum level, unless otherwise stated). These same abundance values were the basis for a community composition analysis. The *t*-SNE plots were constructed by assigning each genome to a site based on where it was most abundant using Rtsne[63].

Metabolic potential of the MAGs

To analyse functional marker genes in the metagenomes, we used a custom database of reference proteins sequences (COG, PFAM, TIGRFAM, KEGG) based on the marker genes used in other studies[64,65] (Supplementary Data Files on GitHub). Pathways were also predicted using MinPath[66] to map all identified KEGG enzymes to the most parsimonious MetaCyc pathways[47]. As these MAGs were incomplete, some genes in pathways may be absent. MinPath presented only parsimonious pathways represented by multiple genes. As most genomes were present even at low abundances across all sites, a cut-off value of $-log_{10} \leq 0.25$ was set for a genome to be considered functionally relevant at any site. The most abundant genomes for each site were used as the basis of further examination. We aggregated marker genes and pathways by function and phylogenetic groups by phylum, except for Proteobacteria that was separated by class. We further grouped all taxa together at each site to test for significant differences in major nutrient cycling processes (carbon, nitrogen, and sulphur) between sites using a hierarchical clustering with *P*-values derived from the Approximately Unbiased bootstrap[67] and Fisher's exact test.

Results

Validity of the geochemical context for the space-for-time design

As a space-for-time design requires spatial contrasts, we first characterised how geochemical properties of the sediments varied along and between our two transects (Figure 1A). Samples from these eight sites clustered into four distinct geochemical groups (Figure 1B) that most likely reflect glacial runoff, the primary hydrological input to the lake. Indeed, PC1 explained 43% of the total variance (σ^2), and differentiated the L and H transects, while PC2 (29.9%) separated each transect according to their depth.

Along PC1, higher concentrations of ammonia, sulphate, and a greater percentage of calcium carbonate were

present in the H transect. However, higher concentrations of O_2, NO_3^-/NO_2^-, and greater redox potential were present in the L transect and the C sites. The lower concentration of dissolved O_2 in the H transect, compared to the L transect, likely reflects a greater activity of chemotrophic aerobes consuming the greater amounts of organic carbon present at this site[23].

Along PC2, organic carbon, pH, and Cl- were more determinant when discriminating between the shallow and deep sites of both transects (Supplementary Figures 4-5). Rather than grouping spatially with the high H transect, the C sites were most chemically similar to L1 (Figure 1C, Supplementary Figure 6). The shallow sites (L1 and H1) were not significantly different from each other in pH and OC concentrations, but were both significantly different from the deeper sites (L2 and H2) suggesting that although most chemical features were similar within each transect, some features might still be influenced by their spatial proximity to the shoreline or depth of the overlying water column (Figure 1C).

Contrasting low *vs.* high runoff transects revealed a decrease in biodiversity

With the existence of such a clear geochemical separation of our transects along PC1 (43% of σ^2) and of significant spatial contrast (Figure 1C), we had the proper context for a space-for-time design, and hence could contrast the two transects in terms of their microbial diversity. For this, we co-assembled a total of 300 (290 bacterial and 10 archaeal) MAGs that were >50% complete and with <10% contamination (Supplementary Tables 6-7). By constructing phylogenetic trees for Bacteria and Archaea, we noted that while most major phyla were represented in the MAGs, no Firmicutes and only a small number of Archaea were identified (Figure 2). In contrast, Gammaproteobacteria ($n = 50$), Actinobacteria ($n = 31$), Alphaprobacteria ($n = 24$), Chloroflexoata ($n = 30$), Planctomycetota ($n = 24$), and Acidobacteriota ($n = 19$) were the most commonly recovered taxa across the entire watershed. Uncultured phyla comprised $\sim$11% of our MAGs, including representatives from multiple taxa: Eisenbacteria ($n = 12$), Patescibacteria ($n = 9$), Omnitrophica ($n = 5$), KSB1 ($n = 1$), Armatimonadota ($n = 1$), Lindowbacteria ($n = 1$), USBP1 ($n = 1$), UBP10 ($n = 1$), and Zixibacteria ($n = 1$).

These MAGs were however not evenly distributed across all sites (Figure 2, inset; Supplementary Figure 7). To quantify this uneven distribution, we determined the site where each genome was most abundant. Based solely on this information, we performed an unsupervised clustering (t-SNE), and found that the directions defined by water flowing from the shallow to the deep site within each transect in the projection space were almost orthogonal between transects (see arrows in Figure 3). This orthogonality suggests that transitioning from the L to the H transect could lead to a dramatic shift in microbial communities.

To understand what these shifts meant at the phylum level, we calculated the relative proportions of each of our

300 MAGs at each site, and tallied these numbers by phylum, over the 43 phyla represented in our data. We did this along each transect – essentially pooling sites H1/H2 together to represent the H transect, and doing the same for sites L1/L2 (the L transect), while keeping proportions for the S and C sites separate. Hierarchical clustering on this table of MAGs proportions by phyla *vs.* sites showed a divergence from the L to H transects (following the (((L,C),H),S) clustering pattern; Figure 4A, inset), confirming the clear contrast between the two transects in terms of taxa proportions (see Figure 3). To test if these taxa proportions tended to increase or decrease when transitioning from L to S along the (((L,C),H),S) clustering pattern, we fitted linear models (ANOVA) regressing the proportions of each of the 43 phyla against sites, ordered as per their hierarchical clustering (L→C→H→S). Essentially, we regressed a single data point for each of the four classes (L, C, H, and S), so that P-values could not be obtained, but slope could be estimated (Figure 4A). Strikingly, most of these slopes were negative (binomial test: $P = 7.8 \times 10^{-8}$), demonstrating a significant decrease in diversity at the phylum level as one goes from low to high runoff regimes in this space-for-time design.

An NMDS ordination allowed us to detect the geochemical features associated with this shift in microbial communities (Supplementary Figure 8). In the sediments, ammonia concentrations ($P = 0.03$), nitrate and nitrite concentrations ($P = 0.03$), and redox potential ($P = 0.03$) were significant in determining the distribution of MAGs (permutation test: $P < 0.05$). We further observed that the sites with the greatest diversity (L/C sites) were also those with the greatest redox potential, oxygen concentrations, and nitrate/nitrite concentrations. Sites with the lowest diversity (H sites), contained greater ammonia and sulphate concentrations, and lower redox potential. In addition to gradients shaped by the interplay between microbial metabolism and local geochemical constraints, the physical disturbances associated with high sedimentation rates also likely contributed to the homogenisation of the microbial community structure; however, we cannot quantify the relative importance of these processes here.

Contrasting low *vs.* high runoff transects also revealed a loss of functional potential

To assess the functional implications of this decrease of biodiversity, we assigned metabolic functions and pathways to proteins in each MAG. We focused on genes and pathways involved in key elements, focusing on carbon, nitrogen, and sulphur cycling (Supplementary Figures 9-10). Only the most abundant genomes per site were reported within each phylum (Supplementary Figure 11), allowing us to compute the proportions of functions and pathways in each of the 43 phyla present in our MAGs across the hydrological regimes. Their hierarchical clustering (Supplementary Figures 12-13) led to a picture consistent with the ones derived from both geochemical (Figure 1) and taxonomic abundances (Figure 4A). Indeed, the two transects were again clearly separated (clustering pattern (((L,C),S),H); Figure 4B, inset), and fitting linear models regressing function/pathway proportions against sites showed that,

again, most of these slopes were negative (binomial test: P = 0.0010). Forcing the same site ordering as for the taxonomic abundances (L→C→H→S as in Figure 4A, inset) led to similar results (binomial test: $P = 7.8 \times 10^{-5}$), demonstrating a significant decrease in metabolic diversity when going from the L to the H transect.

More specifically, we found that markers of carbon and sulphur metabolisms significantly decreased when going from the L to H, while nitrogen metabolism was unaffected (Supplementary Table 8; see Supplementary Text for details). When considering the individual functions present or absent across the transects, we noted that most oxidative pathways (CO, methane, formaldehyde, sulphide, sulphite) appeared less common in the H transect (Supplementary Figure 9), corresponding to lower oxygen concentrations and constraints on aerobic metabolism. Furthermore, while most carbon fixation processes were shared between the two transects, carbon oxidation and reduction reactions regulated through Wood-Ljungdahl pathway were only observed in the H transect, where sedimentary conditions were anoxic throughout the first 5cm (Supplementary Figures 4-5), consistent with a more reductive environment.

Discussion

Even if Arctic microbial communities are changing rapidly[68], there is still a dearth of long-term times series observations. In this study, we adopted a space-for-time design that was made possible by the unique features of Lake Hazen watershed, as its spatial geochemical heterogeneity allowed us to evaluate the structural and functional response of lake sediment microbial communities to the consequences of climate change in the High Arctic. Using metagenomics along two transects facing heterogeneous runoff conditions as a proxy for such consequences, we presented evidence that climate change will likely lead to a decrease in both microbial diversity and functional potential of lake sediment communities.

Despite lacking geochemical measurements for the soil samples, we found that the high runoff sites clustered most frequently with the soil sites (Figure 3), suggesting that the microbial ecology of Lake Hazen sediments was connected to the terrestrial environment as a function of the runoff volume, consistent with previous findings[69,70]. Unsurprisingly as the soil is likely a source of nutrients (*e.g.*, DOC) and organic and inorganic particles, we would expect that increased runoff to the aquatic ecosystems to consequently alter microbial community structure[71]. Some of these structural changes may alter the functional capacity to metabolise carbon, nitrogen, sulphur compounds and process toxins such as metals and antibiotics (Supplementary Figure 9).

However, structural differences across sites did not always translate into direct changes in functional potential, suggesting some functional redundancy (Supplementary Figure 9). Specifically, we observed core sulphur metabolic markers remained unchanged across sites (Supplementary Table 8). Yet, when we examine the trends from low to

high runoff regimes (L→C→H→S), we observed that the abundance of metabolic processes significantly decreased (Figure 4). Many of these metabolic processes were partitioned with only a few taxa possessing marker genes encoding for critical biogeochemical pathways. Therefore, should these taxa and associated genes be lost (below abundance cut-off), these metabolic processes would also likely be lost. Furthermore, rapidly changing environmental conditions may disrupt steps in syntrophic or cooperative metabolic pathways that are distributed over several individuals within a community. As many metabolic pathways are distributed across taxa[72,73], changes in the microbial abundance may have broader consequences for community metabolism and geochemical cycling.

Our study was not designed to directly quantify the interplay between microbial metabolism and the geochemical features that we measured; indeed, a more experimentally-driven approach, based on *in situ* incubation and geochemical tracers, would have been necessary. As a consequence, one complication in interpreting the dynamics or causality of these results is that high glacial runoff is known to, at least initially, alleviate low oxygen conditions at the bottom of the lake as it brings dense, oxygenated river waters with organic matter directly to the bottom of the lake[23]. As a result, some of the geochemical conditions along our high runoff transect may be secondary to greater microbial metabolism, which may actually exceed those in temperate systems[74]. Such increased metabolic activity can eventually lead to low oxygen, low redox, and high ammonia solely due to greater microbial activity and high respiration rates (Figure 1). Another limitation of our study is that the spatial heterogeneity of runoff regimes also leads to divergent sedimentation rates, in particular at the deepest ends of our transects, limiting our ability to make further temporal predictions (Supplementary Table 2).

At a larger temporal scale, a key question that arises from our results is how changes in hydrological regimes will alter the evolutionary dynamics of microbial communities in lake sediments. Niche differentiation, where the coexistence of ecological opportunities can facilitate species diversification, may explain why the low runoff transect hosts a more diverse microbial community[75]. Presently, climate change is predicted to increase runoff in this environment[9], and we found evidence suggesting that the increased runoff homogenises community structure. This can be expected to disrupt niche differentiation, and hence to reduce the overall and long-term metabolic capacity in lake sediments. While not within the scope of this study, we predict that climate change may slow down species diversification, and lead to highly specialized microbial communities adapted to a homogeneous ecological niche characterised by low oxygen, low redox, and high ammonia concentrations.

The rapid changes that affect Lake Hazen's watershed in response to climate warming were already known to directly alter its hydrological regime. Here we further provide evidence that a combination of increasing runoff and changing geochemical conditions are associated with the reduced diversity and metabolic potential of its microbial

communities. While longitudinal studies are needed to confirm these patterns, it is still unclear how such losses in biodiversity and metabolic potential in Arctic ecosystems will impact key biogeochemical cycles, potentially creating feedback loops of uncertain direction and magnitude.

Acknowledgements This study was made possible through a collaborative effort undertaken by Igor Lehnherr, Stephanie Varty, Victoria Wisniewski (University of Toronto, Mississauga), Charles Talbot (Environment and Climate Change Canada), and Maria Cavaco (University of Alberta). We thank Linda Bonen, Marina Cvetkovska, Manon Ragonnet and Alex Wong for comments and discussions. Funding support was provided by the Natural Science and Engineering Research Council of Canada (VSL, AJP, SAB), ArcticNet Network Centre of Excellence (VSL, AJP), and the Polar Continental Shelf Program (VSL) in Resolute, Nunavut, which provided logistical and financial support.

References

1. Overpeck, J. *et al.* Arctic environmental change of the last four centuries. *Science* **278**, 1251–1256 (1997).
2. Serreze, M. C. & Francis, J. A. The Arctic amplification debate. *Climatic change* **76**, 241–264 (2006).
3. Screen, J. A. & Simmonds, I. The central role of diminishing sea ice in recent Arctic temperature amplification. *Nature* **464**, 1334 (2010).
4. IPCC. *Summary for Policymakers*, book section SPM, 1–30 (Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013). Available at www.climatechange2013.org.
5. Laudon, H. *et al.* Save northern high-latitude catchments. *Nature Geoscience* **10**, 324 (2017).
6. Bliss, A., Hock, R. & Radić, V. Global response of glacier runoff to twenty-first century climate change. *Journal of Geophysical Research: Earth Surface* **119**, 717–730 (2014).
7. O'Reilly, C. M. *et al.* Rapid and highly variable warming of lake surface waters around the globe. *Geophysical Research Letters* **42**, 10–773 (2015).
8. Frey, K. E. & McClelland, J. W. Impacts of permafrost degradation on Arctic river biogeochemistry. *Hydrological Processes: An International Journal* **23**, 169–182 (2009).
9. Lehnherr, I. *et al.* The world's largest High Arctic lake responds rapidly to climate warming. *Nature Communications* **9**, 1290 (2018).
10. Smol, J. P. *et al.* Climate-driven regime shifts in the biological communities of Arctic lakes. *Proceedings of the National Academy of Sciences* **102**, 4397–4402 (2005).

11. Wrona, F. J. *et al.* Transitions in Arctic ecosystems: Ecological implications of a changing hydrological regime. *Journal of Geophysical Research: Biogeosciences* **121**, 650–674 (2016).

12. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).

13. Fuhrman, J. A. Microbial community structure and its functional implications. *Nature* **459**, 193 (2009).

14. Williamson, C. E., Saros, J. E., Vincent, W. F. & Smol, J. P. Lakes and reservoirs as sentinels, integrators, and regulators of climate change. *Limnology and Oceanography* **54**, 2273–2282 (2009).

15. Stoeva, M. K. *et al.* Microbial community structure in lake and wetland sediments from a High Arctic polar desert revealed by targeted transcriptomics. *PLoS One* **9**, e89531 (2014).

16. Thaler, M., Vincent, W. F., Lionard, M., Hamilton, A. K. & Lovejoy, C. Microbial community structure and interannual change in the last epishelf lake ecosystem in the north polar region. *Frontiers in Marine Science* **3**, 275 (2017).

17. Mohit, V., Culley, A., Lovejoy, C., Bouchard, F. & Vincent, W. F. Hidden biofilms in a far northern lake and implications for the changing Arctic. *npj Biofilms and Microbiomes* **3**, 17 (2017).

18. Ruuskanen, M. O., St. Pierre, K. A., St. Louis, V. L., Aris-Brosou, S. & Poulain, A. J. Physicochemical drivers of microbial community structure in sediments of Lake Hazen, Nunavut, Canada. *Frontiers in Microbiology* **9**, 1138 (2018).

19. Ruuskanen, M. O., St Pierre, K. A., St Louis, V. L., Aris-Brosou, S. & Poulain, A. J. Physicochemical drivers of microbial community structure in sediments of lake hazen, nunavut, canada. *Front Microbiol* **9**, 1138 (2018).

20. Köck, G. *et al.* Bathymetry and sediment geochemistry of Lake Hazen (Quttinirpaaq National Park, Ellesmere Island, Nunavut). *Arctic* 56–66 (2012).

21. Caruana, E. J., Roman, M., Hernández-Sánchez, J. & Solli, P. Longitudinal studies. *Journal of thoracic disease* **7**, E537 (2015).

22. Huss, M. & Hock, R. Global-scale hydrological response to future glacier mass loss. *Nature Climate Change* **8**, 135 (2018).

23. St. Pierre, K. *et al.* Contemporary limnology of the rapidly changing glacierized watershed of the world's largest High Arctic lake. *Scientific reports* (2019).

24. Blois, J. L., Williams, J. W., Fitzpatrick, M. C., Jackson, S. T. & Ferrier, S. Space can substitute for time in predicting climate-change effects on biodiversity. *Proceedings of the National Academy of Sciences* **110**, 9374–9379 (2013).

25. Lester, R. E., Close, P. G., Barton, J. L., Pope, A. J. & Brown, S. C. Predicting the likely response of data-poor ecosystems to climate change using space-for-time substitution across domains. *Global change biology* **20**, 3471–3481 (2014).

26. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature biotechnology* **35**, 725 (2017).

27. Haglund, A.-L., Lantz, P., Törnblom, E. & Tranvik, L. Depth distribution of active bacteria and bacterial activity in lake sediment. *FEMS Microbiology Ecology* **46**, 31–38 (2003).

28. Zhou, J., Bruns, M. A. & Tiedje, J. M. DNA recovery from soils of diverse composition. *Applied and environmental microbiology* **62**, 316–322 (1996).

29. Poulain, A. J. *et al.* Microbial DNA records historical delivery of anthropogenic mercury. *The ISME journal* **9**, 2541 (2015).

30. Heiri, O., Lotter, A. F. & Lemcke, G. Loss on ignition as a method for estimating organic and carbonate content in sediments: reproducibility and comparability of results. *Journal of paleolimnology* **25**, 101–110 (2001).

31. van der Laan, M. J. & Pollard, K. S. Hybrid clustering of gene expression data with visualization and the bootstrap. *Journal of Statistical Planning and Infererence* **117**, 275–303 (2003).

32. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

33. Andrews, S. *et al.* Fastqc: a quality control tool for high throughput sequence data (2010).

34. 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* **31**, 1674–1676 (2015).

35. West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res* **28**, 569–580 (2018).

36. Eren, A. M. *et al.* Anvio: an advanced analysis and visualization platform for omics data. *PeerJ* **3**, 1319 (2015).

37. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357 (2012).

38. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

39. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nature methods* **11**, 1144 (2014).

40. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications* **7**, 11257 (2016).

41. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* **11**, 119 (2010).

42. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4**, 41 (2003).

43. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic acids research* **44**, D279–D285 (2015).

44. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic acids research* **31**, 371–373 (2003).

45. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25 (2000).

46. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic acids research* **44**, D457–D462 (2015).

47. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research* **36**, D623–D631 (2007).

48. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

49. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**, 1043–1055 (2015).

50. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature microbiology* **2**, 1533 (2017).

51. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2015).

52. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology* **36**, 996–1004 (2018).

53. Charif, D. & Lobry, J. SeqinR 1.0.2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In Bastolla, U., Porto, M., Roman, H. & Vendruscolo, M. (eds.) *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, 207–232 (Springer Verlag, New York, 2007). ISBN : 978-3-540-35305-8.

54. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792–1797 (2004).

55. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2008). Available at http://www.R-project.org. ISBN 3-900051-07-0.

56. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

57. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2: approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490 (2010).

58. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution* **18**, 691–699 (2001).

59. Aris-Brosou, S. & Rodrigue, N. The essentials of computational molecular evolution. In *Evolutionary Genomics*, 111–152 (Springer, 2012).

60. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

61. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**, 28–36 (2017).

62. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one* **8**, e61217 (2013).

63. Krijthe, J., van der Maaten, L. & Krijthe, M. J. Package 'Rtsne' (2018).

64. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications* **7**, 13219 (2016).

65. Dombrowski, N., Teske, A. P. & Baker, B. J. Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nature Communications* **9**, 4999 (2018).

66. Ye, Y. & Doak, T. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Computational Biology* **5**, e1000465 (2009).

67. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540–1542 (2006).

68. Hultman, J. *et al.* Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* **521**, 208 (2015).

69. Comte, J., Culley, A. I., Lovejoy, C. & Vincent, W. F. Microbial connectivity and sorting in a High Arctic watershed. *The ISME journal* **12**, 2988–3000 (2018).

70. Ruiz-González, C., Niño-García, J. P. & del Giorgio, P. A. Terrestrial origin of bacterial communities in complex boreal freshwater networks. *Ecology letters* **18**, 1198–1206 (2015).

71. Le, H. T. *et al.* Responses of aquatic bacteria to terrestrial runoff: effects on community structure and key taxonomic groups. *Frontiers in microbiology* **7**, 889 (2016).

72. Morris, J. J., Lenski, R. E. & Zinser, E. R. The black queen hypothesis: evolution of dependencies through adaptive gene loss. *MBio* **3**, e00036–12 (2012).

73. Hug, L. A. *et al.* Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* **1**, 22 (2013).

74. Probst, A. J. *et al.* Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. *Nature microbiology* **3**, 328 (2018).

75. Cordero, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Reviews Microbiology* **12**, 263 (2014).

Figure captions

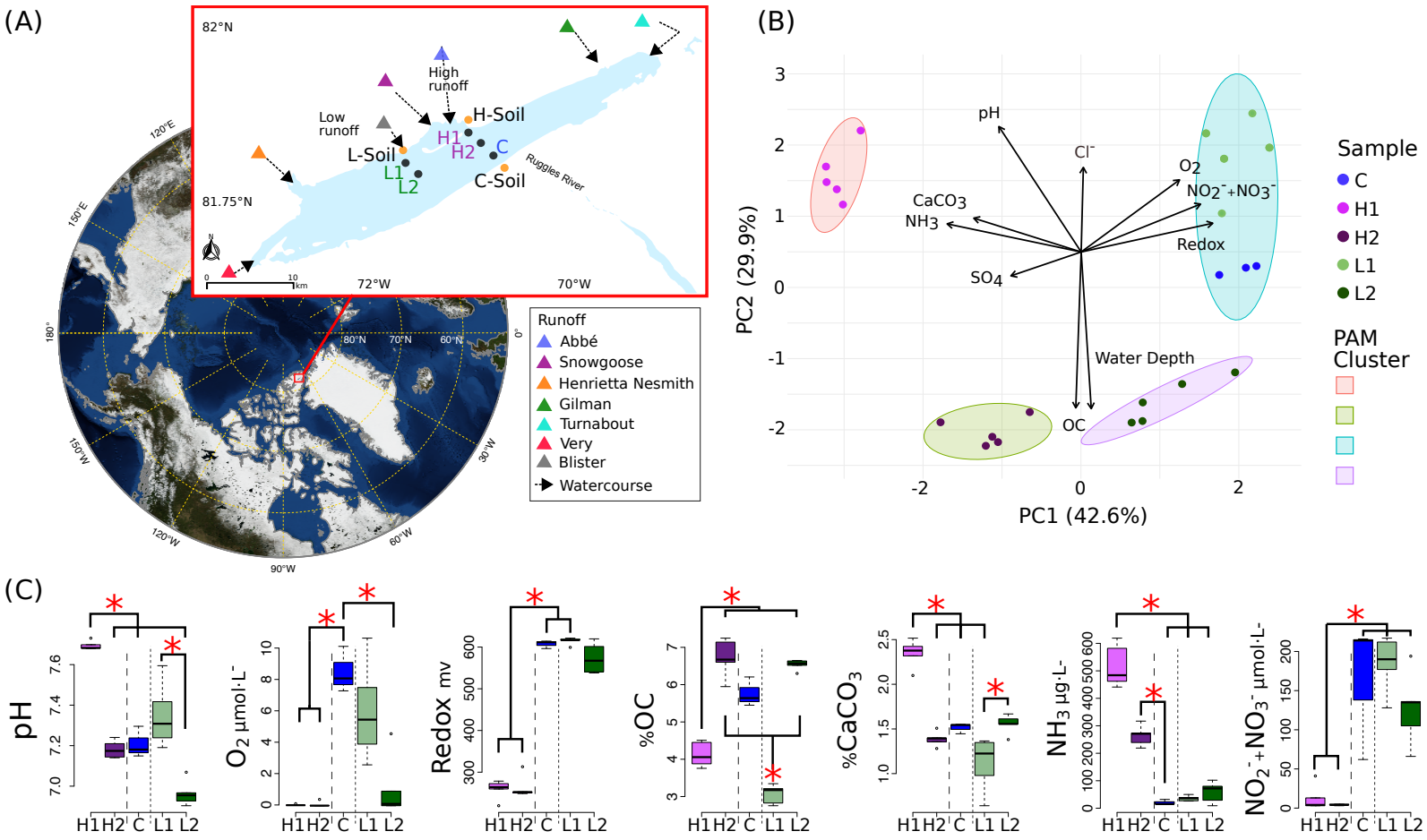

Figure 1. Lake Hazen sampling design and chemical composition. (A) Basemap The location of Lake Hazen (red box). Inset map: soil (orange dots) and sediment (black dots) sample sites are separated into hydrological regimes of high (purple), low (green), and negligible/control (blue) runoff. (B) Principal component analysis (PCA) showing the differences in physical and chemical composition of the sediment sites. Vectors display pH, dissolved dioxygen (O_2), redox potential, nitrates and nitrites concentration ($NO_2^- + NO_3^-$), water depth, percent organic carbon (OC), percent calcium carbonate ($CaCO_3$), sulphate concentration (SO_4), and ammonia concentration (NH_3). Individual points represent the mean values using 1 cm intervals measured in the top 5 cm. Partitioning around medoids was used to identify clusters. (C) Distribution of chemical features for sediment sites. Branches and asterisks indicate significant differences between sites $P < 0.025$ (Dunn Test). If branch tips form a dichotomy or trichotomy, the interactions within that group is not significant. Long dashes separate high runoff sites and dotted line separates low runoff sites. There was insufficient data to include soil sites in B and C.

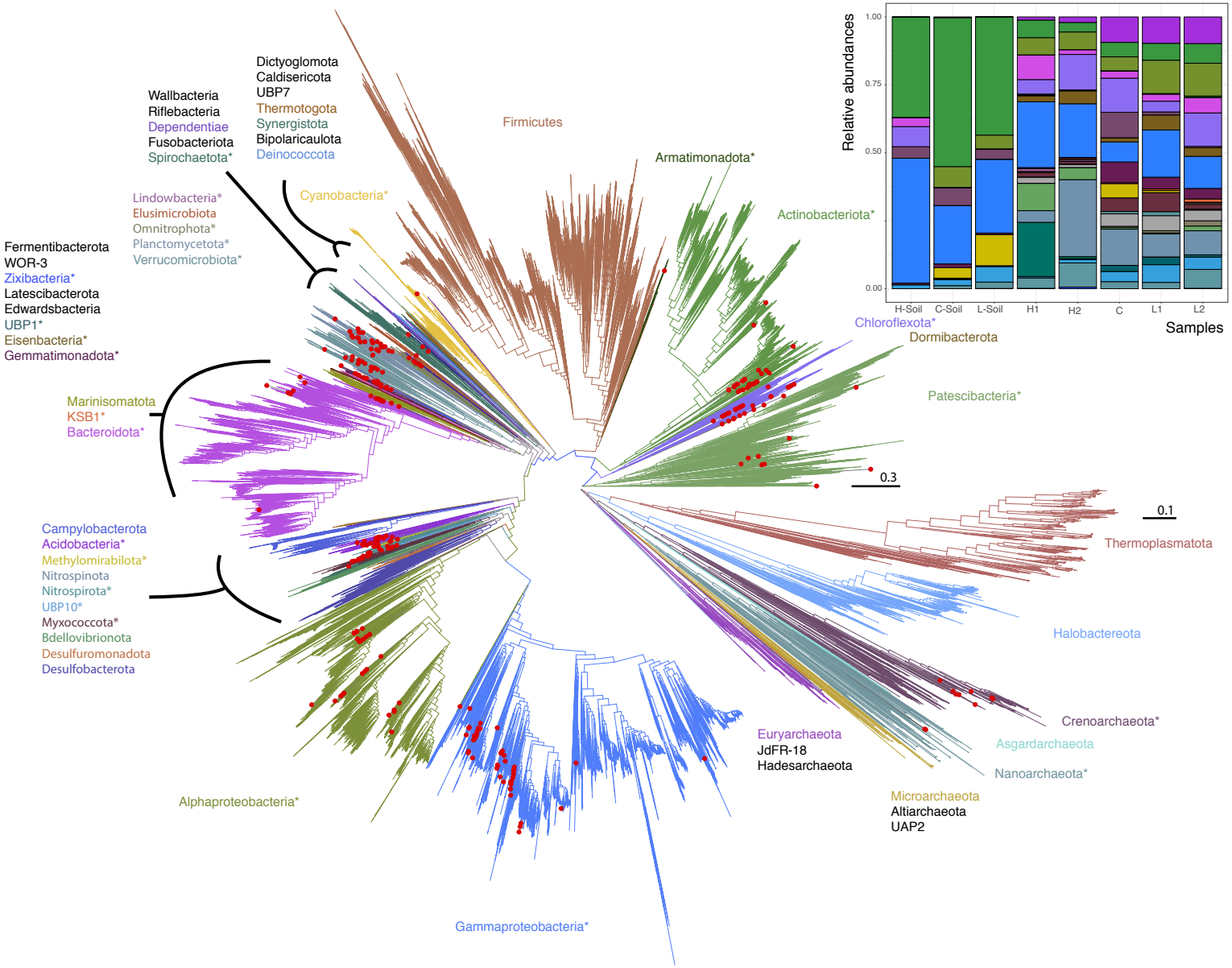

Figure 2. Maximum likelihood phylogenetic trees of Lake Hazen genomes based on 120 concatenated bacteria and 122 concatenated archaea protein-coding genes. Red Dots: Lake Hazen genomes. Asterisks (*) indicate phyla that contain Lake Hazen genomes. Bacteria tree is rooted with Patescibacteria and Archaea tree is rooted with Euryarchaeota. See GitHub account for full taxonomy tree files and for original tree files (Supplemental Data File 2 and 3). Inset shows MAG abundance across sites, in the 300 high quality genomes for each sample normalised to 100%.

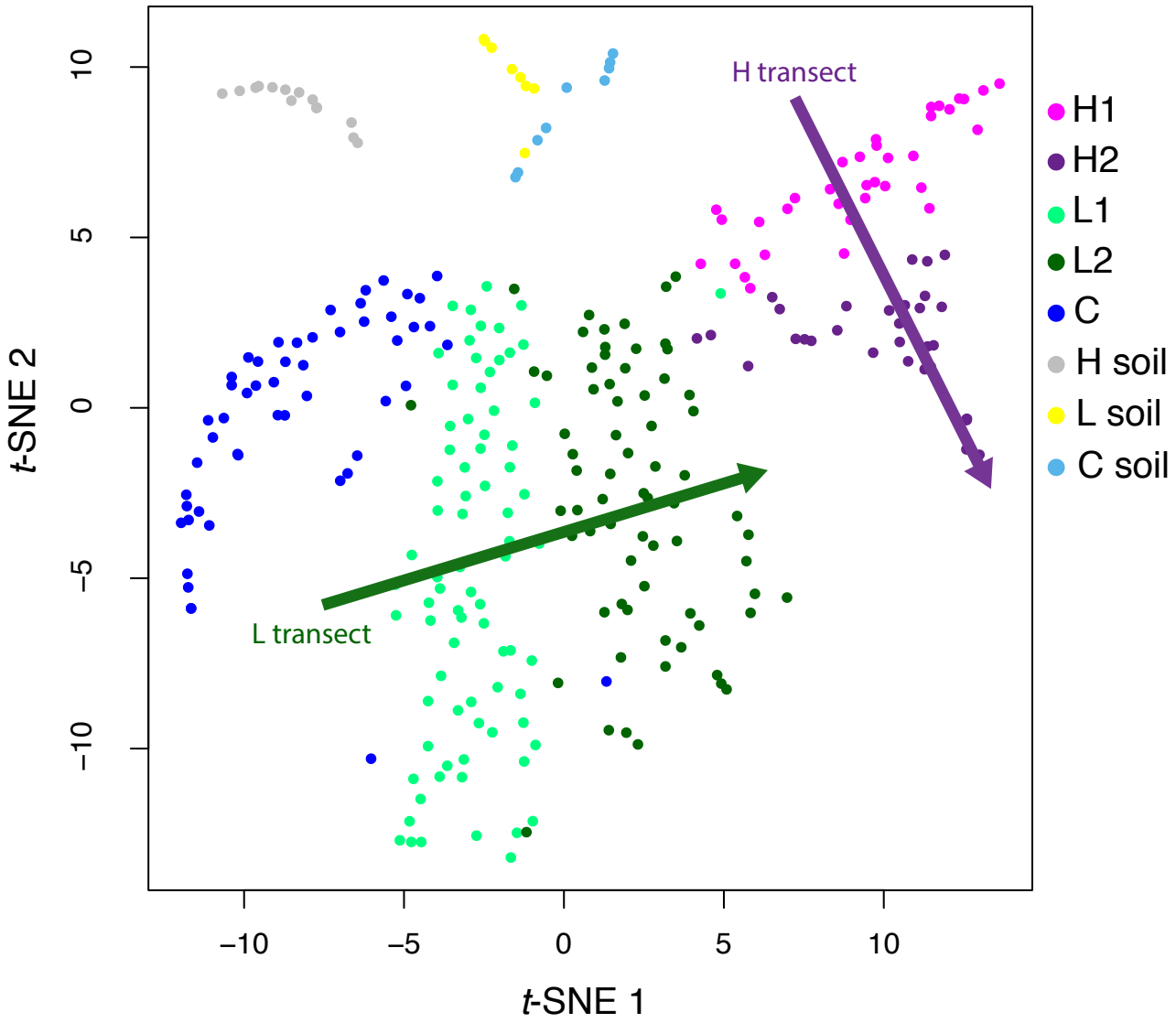

Figure 3. *t*-SNE analysis of genome abundance for each sediment sample. Each of the 300 shown genomes was assigned to the sample where it has the greatest abundance. Shaded arrows display direction of water flow between the high (green) and low (purple) transects.

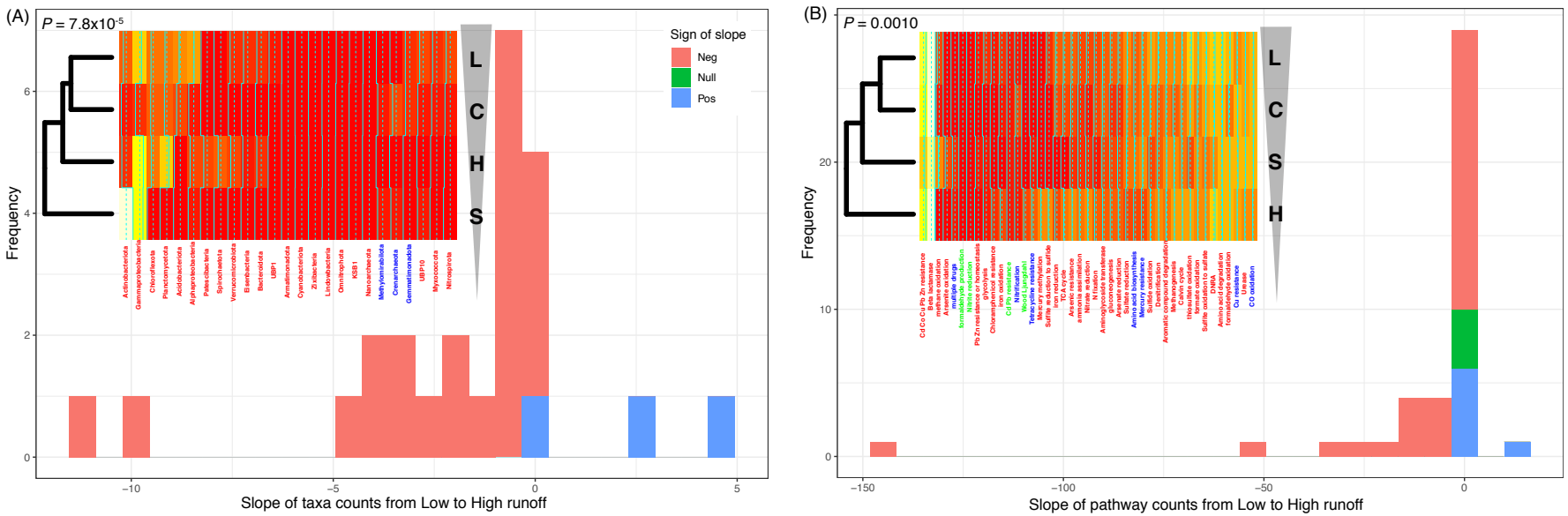

Figure 4. Transition from low to high runoff leads to a decrease in diversity. (A) Distribution of the slopes of taxonomic counts as a function of sites. (B) Distribution of the slopes of pathway counts as a function of sites. In both cases, counts were aggregated by location types (L, C, S, and H sites), and linear models (ANOVA) were fitted to estimate the slope of each regression. Insets: heatmap representations of count tables; leftmost dendrograms show how the location types cluster, transitioning from L to H runoffs (vertical triangle pointing down). *P*-values: one-sided binomial test for enrichment in negative slopes.