ARTICLE

Climate change negatively impacts dominant microbes in the sediments of a High Arctic lake

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

[1]Department of Biology, University of Ottawa, Ottawa, ON K1N 6N5, Canada

[2]University of Alberta, Department of Biological Sciences, Edmonton, AB T6G 2E9, Canada

[3]Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1N 6N5, Canada

✉: Stéphane Aris-Brosou (sarisbro@uottawa.ca).

Running head: Climate change impacts on High Arctic microbes

Abstract

Temperatures 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 in freshwater sedimentary systems. To address this knowledge gap, we analyzed sediments from Lake Hazen, NU Canada. Here, we exploit the spatial heterogeneity created by varying runoff regimes across the watershed of this uniquely large lake at these latitudes to test how a transition from low to high runoff, used as one proxy for climate change, affects the community structure and functional potential of dominant microbes. Based on metagenomic analyses of lake sediments along these spatial gradients, we show that increasing runoff leads to a decrease in taxonomic and functional diversity of sediment microbes. Our findings are likely to apply to other, smaller, glacierized watersheds typical of polar or high latitude / high altitudes ecosystems; we can predict that such changes will have far reaching consequences on these ecosystems by affecting nutrient biogeochemical cycling, the direction and magnitude of which are yet to be determined.

Keywords: High Arctic, microbial ecology, metagenome assembled genomes (MAGs), high-throughput sequencing

Main

Climate change is amplified in polar regions, where near-surface temperatures have increased almost twice as fast as elsewhere on Earth over the last decade [1, 2, 3]. Climate models predict that temperature will increase in the Arctic by as much as 8°C by 2100 [4]. These changes are already having dramatic consequences on physical [5, 6, 7], biogeochemical [8, 9], and ecological [10, 11] processes across Arctic ecosystems. Yet, while we are starting to understand the effect of thawing permafrost on microbial communities [12, 13, 14], we know very little about how microbes in lentic ecosystems such as lakes respond to environmental changes – even though microbes mediate most global biogeochemical cycles [15, 16]. Furthermore, lakes are broadly considered sentinels of climate change, as they integrate physical, chemical and biological changes happening through their watersheds [17]; however, their microbial community structure and function are relatively understudied, in particular in the Arctic.

To date, much 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 [18, 19, 20, 21, 22]. While these studies are useful to understand the structure of these microbial communities, they provide limited functional insights and can be biased as they often rely on sequence databases where environmental microbes, specifically from the Arctic, may be underrepresented [23, 24]. 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 microbial communities might respond to climate change.

To predict the effect of climate change on microbial functional diversity in Arctic lake sediments, we focused on Lake Hazen, the world's largest High Arctic lake by volume

(82°N) [25]. In this work, we exploited two important properties of Lake Hazen. First, its watershed is already experiencing the effects of climate change, as increasing temperatures there are leading to more glacial melt, permafrost thaw, and increased runoff from the watershed into the lake in warmer years relative to cooler ones citelehnherr2018world. Second, its tributaries are highly heterogeneous, fed by 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) [32].

It is this temporal and spatial heterogeneity in runoff that we used to evaluate the possible consequences of climate change on High Arctic sediment microbial functional diversity, acknowledging that the consequences of increasing temperature are likely slightly more plural and complex. To this effect, we sampled lake sediments along two transects representing low (L transect: samples L1 [shallow] and L2 [deep]) and high (H: samples H1 [shallow] and H2, [deep]) seasonal runoff volume, as well as at a single site that received negligible runoff (C site; Figure 1A). We also collected soil samples (S sites) from three sites in the dried streambeds of the tributaries, on the northern shore between the two transects to assess soil influence on microbial communities present in the sediments. We then resorted to untargeted metagenomics analyses to draw an inventory of dominant microbes, assumed to be the most critical to nutrient cycling and the most relevant to the dynamics of microbial communities. These reconstructed Metagenome Assembled Genomes (MAGs) [26] allowed us to assess the quantitative impact of a change of runoff regime, from low to high, on both the structure of sediment microbial communities and their functional potential. We show that an increase in runoff volume and resultant sedimentation rates, as predicted under climate change scenarios for the region, could lead to a reduced diversity of the dominant microbial community and of their functional potential.

Characterization of the physical and geochemical environments

We first characterized how geochemical properties of the sediments varied along and between the two transects. Sediment samples from these five sites clustered into four distinct geochemical groups (Figure 1B) that reflect spatial variability in glacial runoff, the primary hydrological input to the lake. Indeed, PC1 explained 43% of the total variance (σ^2), and differentiated the L and high H runoff transects, while PC2 (29.9%) separated each transect according to their depth.

Along PC1, higher concentrations of ammonia (NH_3) and sulfate (SO_{4-}^{2-}) in the porewaters, and a greater percentage of calcium carbonate in the sediments, were present in the H transect. However, higher concentrations of dioxygen (O_2), nitrates / nitrites (NO_3^-/NO_2^-), and greater redox potential were present in the L transect and the control (C) sites. Along PC2, sediment organic carbon (OC), and porewater pH and Cl^-, were more determinant when discriminating between the shallow (L1 and H1) and deep (L2 and H2) sites of both transects (Supplementary Figures 4-5). Rather than grouping spatially with the H transect, the C sites were most chemically similar to L1 (Figure 1C, Supplementary Figure 6). The shallow sites were not significantly different from each other in pH or OC concentrations, but were both significantly different from the deeper sites 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 such a clear geochemical separation of the transects along PC1 (43% of σ^2) and significant spatial contrasts (Figure 1C), we had the right context to evaluate the influence of runoff gradients on sediment microbial diversity. We 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 ∼11% of reconstructed 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 ($n = 1$).

However, these MAGs were 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 sediment-laden 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 assess the significance of these shifts at the phylum level, we calculated the relative proportions of each of the reconstructed 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 H 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.

An NMDS ordination allowed us to detect the geochemical features associated with this shift in microbial communities (Supplementary Figure 8). In the sediments, NH_3 concentrations ($P = 0.03$), NO_2^- / NO_3^- 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, and O2 and NO_3^-/NO_2^- concentrations. Sites with the lowest microbial diversity (H sites), contained greater NH_3 and SO_{4-}^{2-} concentrations, and lower redox potential. In addition to gradients shaped by the interplay between microbial metabolism and local geochemical constraints, the physical disturbances asso-

ciated with high sedimentation rates also likely contributed to the homogenization of the microbial community structure; however, we cannot quantify the relative importance of each 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, targeting carbon, nitrogen, and sulfur 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 reconstructed MAGs across the hydrological regimes. Their hierarchical clustering (Supplementary Figures 12-14) 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 marker genes whose product is implicated in carbon and sulfur 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, sulfide, sulfite) 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 [13], there is still a dearth of long-term time series observations. To address this point, we used Lake Hazens spatial geochemical heterogeneity to evaluate the structural and functional response of lake sediment microbial communities to varying runoff, already shown to increase in this warming High Arctic environment [9]. Such an approach can reasonably be interpreted from the lens of a space-for-time design, which assumes that spatial and temporal variations are not only equivalent [27, 28], but also stationary [29]. Whether this latter condition is met cannot be known, but in the absence of any time-series documenting the effect of climate change on lake sediment microbial communities in the High Arctic, the space-for-time design becomes a convenience, if not a necessity [30].

Using metagenomics along two transects experiencing heterogeneous runoff conditions, we presented evidence that climate change, as it drives increasing runoff and sediment loading to glacial lakes, will likely lead to a decrease in both diversity and functional potential of the dominant microbial communities residing in lake sediments. Note that

we specifically focused here on the dominant microbes, that is those for which we could reconstruct the MAGs, in order to (i) have a phylogenetic placement of the corresponding organisms based on a large number of marker genes (Figure 2), rather than partial 16S rRNA gene sequences as usually done [23], and (ii) be able to predict almost complete functional pathways for each of these organisms to test the impact of a change of runoff (Figure 4), rather than inferring function from taxonomic affiliation [23].

Such a decrease in taxonomic and functional diversity may not be unique to Lake Hazen, where rising temperatures have resulted in increasing glacial melt and associated runoff. Although such a pattern was not observed in other regions of the globe where runoff is predicted to decrease [31, 32], our finding are likely to apply to other, smaller, glacierized watersheds typical of high latitudes or altitudes. Indeed, at least in the Arctic, freshwater discharge is broadly expected to increase with increasing temperatures and precipitation loadings [33, 34, 35]. It would thus be immensely valuable to conduct similar studies, replicating where appropriate a similar space-for-time design, at other lakes throughout the world. Additional sampling efforts should carefully consider the spatial heterogeneity of runoff regimes leading to divergent sedimentation rates (Supplementary Table 2), limiting our ability to make temporal predictions.

Despite lacking geochemical measurements for the soil samples, we found that the microbial communities in the sediments at the high runoff sites clustered most frequently with those in the soil sites (Figure 4), highlighting a connection between terrestrial and aquatic sediment communities as a function of the runoff volume, consistent with previous findings [36, 37]. Unsurprisingly, as the soil is likely a source of nutrients (*e.g.*, DOC) and organic and inorganic particles, we would expect increased runoff to the aquatic ecosystems to alter microbial community structure [38]. Some of these structural changes may then alter the functional capacity to metabolize carbon, nitrogen, sulfur compounds

and process toxins such as metals and antibiotics (Supplementary Figure 9). A more experimentally-driven approach, based for instance on *in situ* incubation and geochemical tracers, would have been necessary to quantify such an interplay between microbial metabolism and geochemical features. Yet, as sediments and nutrients are mostly deposited during the summer melt months, it can be expected that lake sediments record microbe-driven seasonal changes in their geochemistry. Indeed, high glacial runoff is known to bring dense, oxygenated river waters with OC directly to the bottom of the lake [32], stimulating aerobic microbial activity. As a result, the geochemistry recorded along the high runoff transect may first reflect a period of greater microbial metabolism, which may actually exceed those in temperate systems [39], eventually followed by low oxygen, low redox, and high NH_3 conditions observed here (Figure 1) as oxygen is depleted and anaerobic metabolisms allowed to proceed.

At a larger temporal scale, a key question that arises from these 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 sediments along the low runoff transect hosts a more diverse microbial community than sediments along the high runoff transect [40]. Presently, climate change is predicted to increase runoff in this High Arctic environment [9], and we found evidence suggesting that the increased runoff homogenizes community structure. This can be expected to disrupt niche differentiation, and hence to reduce the overall and long-term metabolic capacity in lake sediments. It is currently hard to predict the future microbial ecology of these systems. On the one hand, climate change may diminish species diversification, and lead to highly specialized microbial communities adapted to a homogeneous ecological niche characterized by low oxygen, low redox, and high NH_3 concentrations. On the other hand, the seasonal and rapid changes in redox

conditions, predicted to follow the strong but punctual input of oxygen and nutrients during springtime may allow for the development of a short-lived community that eluded our sampling and analysis.

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 dominant 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.

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. McCalley, C. K. *et al.* Methane dynamics regulated by microbial community response to permafrost thaw. *Nature* **514**, 478 (2014).

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

14. Mackelprang, R., Saleska, S. R., Jacobsen, C. S., Jansson, J. K. & Taş, N. Permafrost meta-omics and climate change. *Annual Review of Earth and Planetary Sciences* **44**, 439–462 (2016).

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

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

17. 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).

18. 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).

19. 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).

20. 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).

21. 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).

22. Cavaco, M. A. *et al.* Freshwater microbial community diversity in a rapidly changing high arctic watershed. *FEMS Microbiol Ecol* **95** (2019).

23. 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).

24. Ruuskanen, M. O. *et al.* Microbial genomes retrieved from high arctic lake sediments encode for adaptation to cold and oligotrophic environments. *Limnology and Oceanography* (2019).

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

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. 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).

28. 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).

29. Damgaard, C. A critique of the space-for-time substitution practice in community ecology. *Trends Ecol Evol* **34**, 416–421 (2019).

30. Pickett, S. T. Space-for-time substitution as an alternative to long-term studies. In *Long-term studies in ecology*, 110–135 (Springer, 1989).

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

32. Pierre, K. S. *et al.* Contemporary limnology of the rapidly changing glacierized watershed of the world's largest high arctic lake. *Scientific reports* **9**, 4447 (2019).

33. Peterson, B. J. *et al.* Increasing river discharge to the arctic ocean. *science* **298**, 2171–2173 (2002).

34. Rawlins, M. A. *et al.* Analysis of the arctic system for freshwater cycle intensification: Observations and expectations. *Journal of Climate* **23**, 5715–5737 (2010). Available at http://dx.doi.org/10.1175/2010JCLI3421.1.

35. Bring, A. *et al.* Arctic terrestrial hydrology: A synthesis of processes, regional effects, and research challenges. *Journal of Geophysical Research: Biogeosciences* **121**, 621–649 (2016).

36. 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).

37. 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).

38. 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).

39. 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).

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

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.

Contributions

G.C. and V.S.L. performed sampling, whereas G.C. conducted laboratory analyses. G.C. and S.A.B. performed data analyses. G.C., S.A.B., V.S.L., and A.J.P. designed the study and wrote the manuscript. V.S.L. conducted the microsensor profiles and porewater extractions. G.C., S.A.B., A.J.P., M.R., K.S.P., and V.S.L. reviewed the manuscript.

Ethics declarations / Competing interests

The authors declare no competing interests.

Supplementary information

Supplementary information is available for this paper.

Figure captions

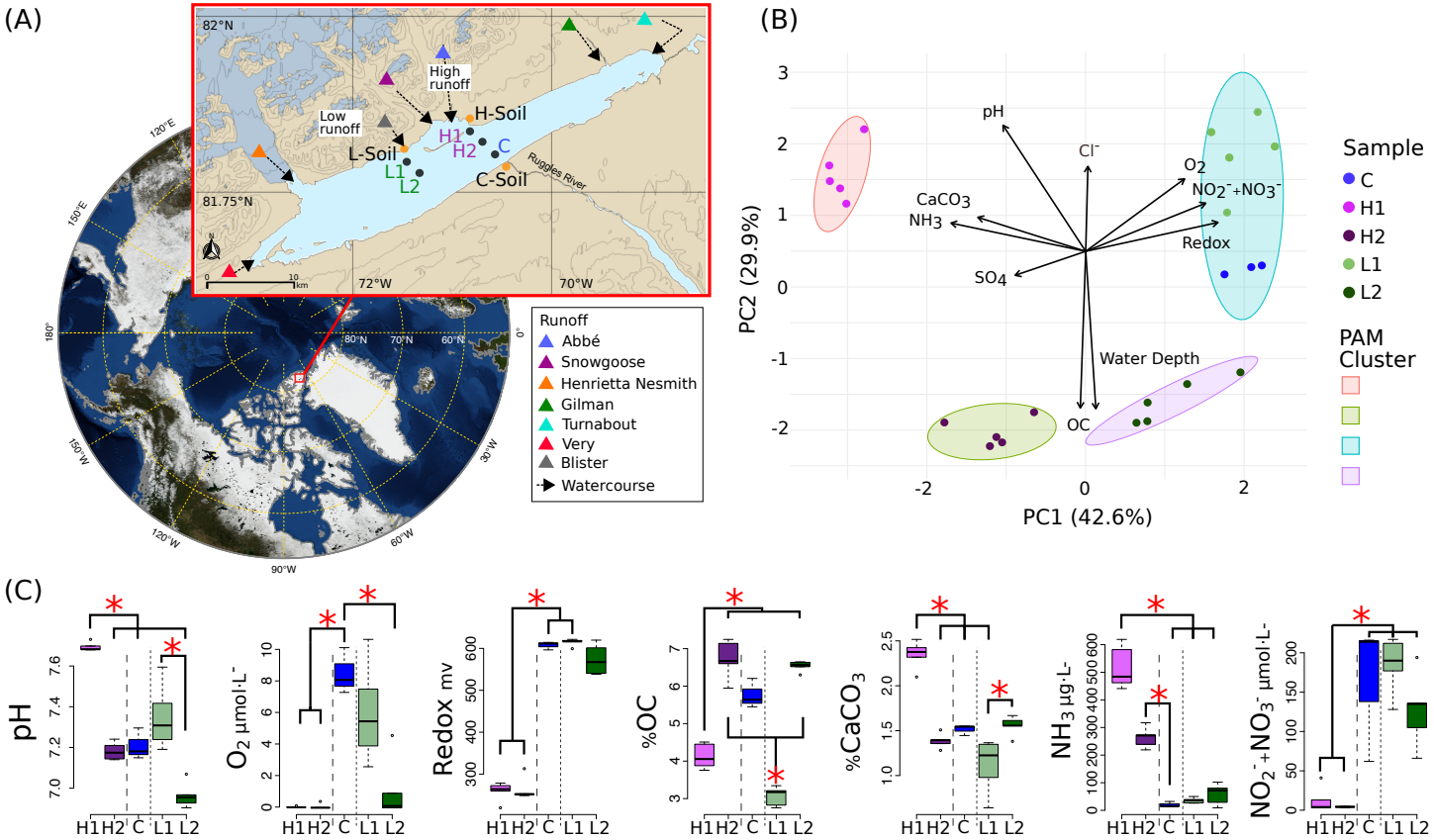


Figure 1. Lake Hazen sampling design and chemical composition. (A) Location of Lake Hazen (red box). Inset map: soil (orange dots) and sediment (black dots) sample sites are separated into hydrological regimes of high (purple), low (green), and negligible/control (blue) runoff. (B) Principal component 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 ($CaCO_3$), sulfate (SO_4^-2) 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.

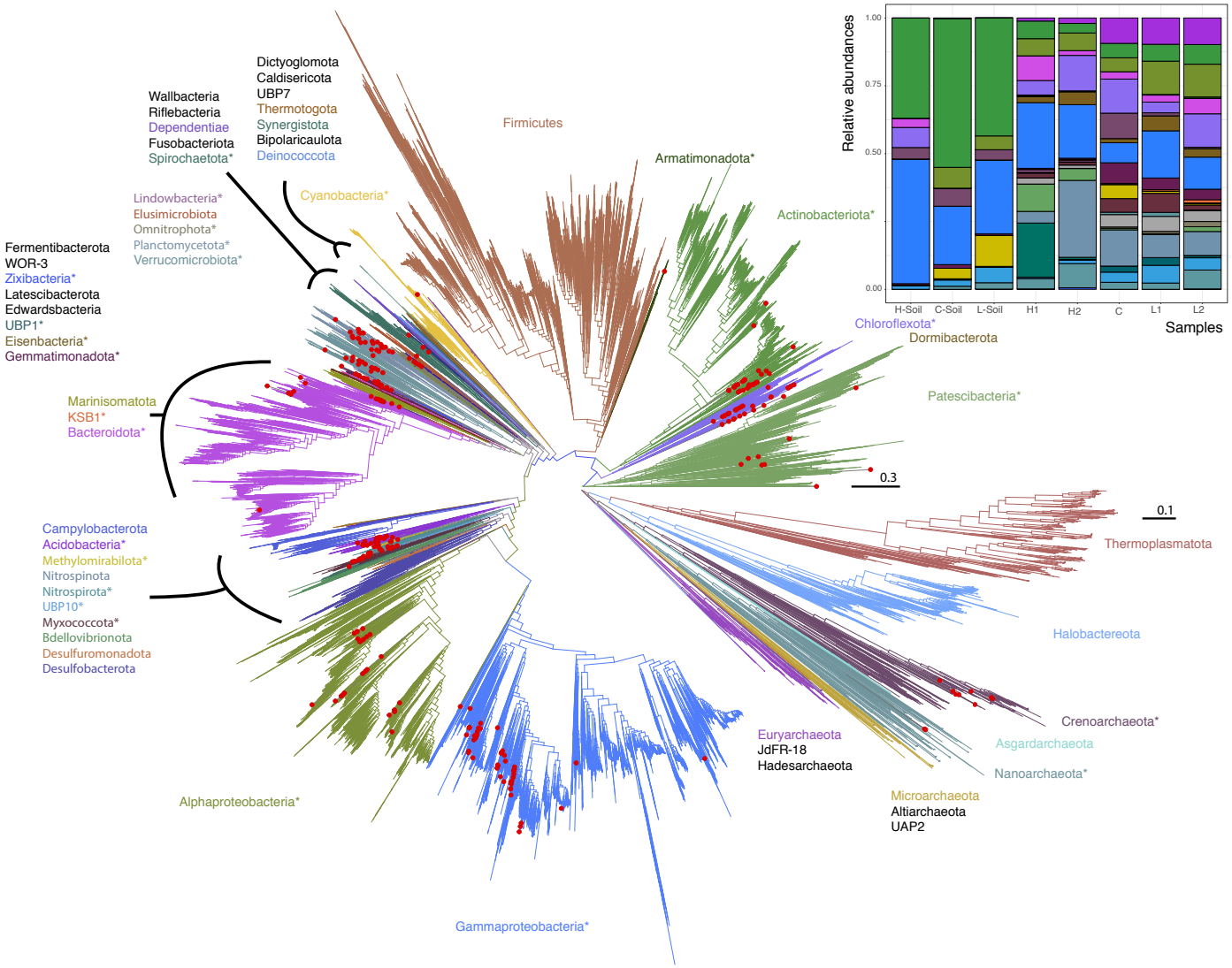


Figure 2. Maximum likelihood phylogenetic trees of Lake Hazen genomes based on 120 concatenated bacteria and 122 concatenated archaea protein-coding genes. Red Dots: Lake Hazen genomes. Asterisks (*) indicate phyla that contain Lake Hazen genomes. Bacteria tree is rooted with Patescibacteria and Archaea 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 normalized to 100%.

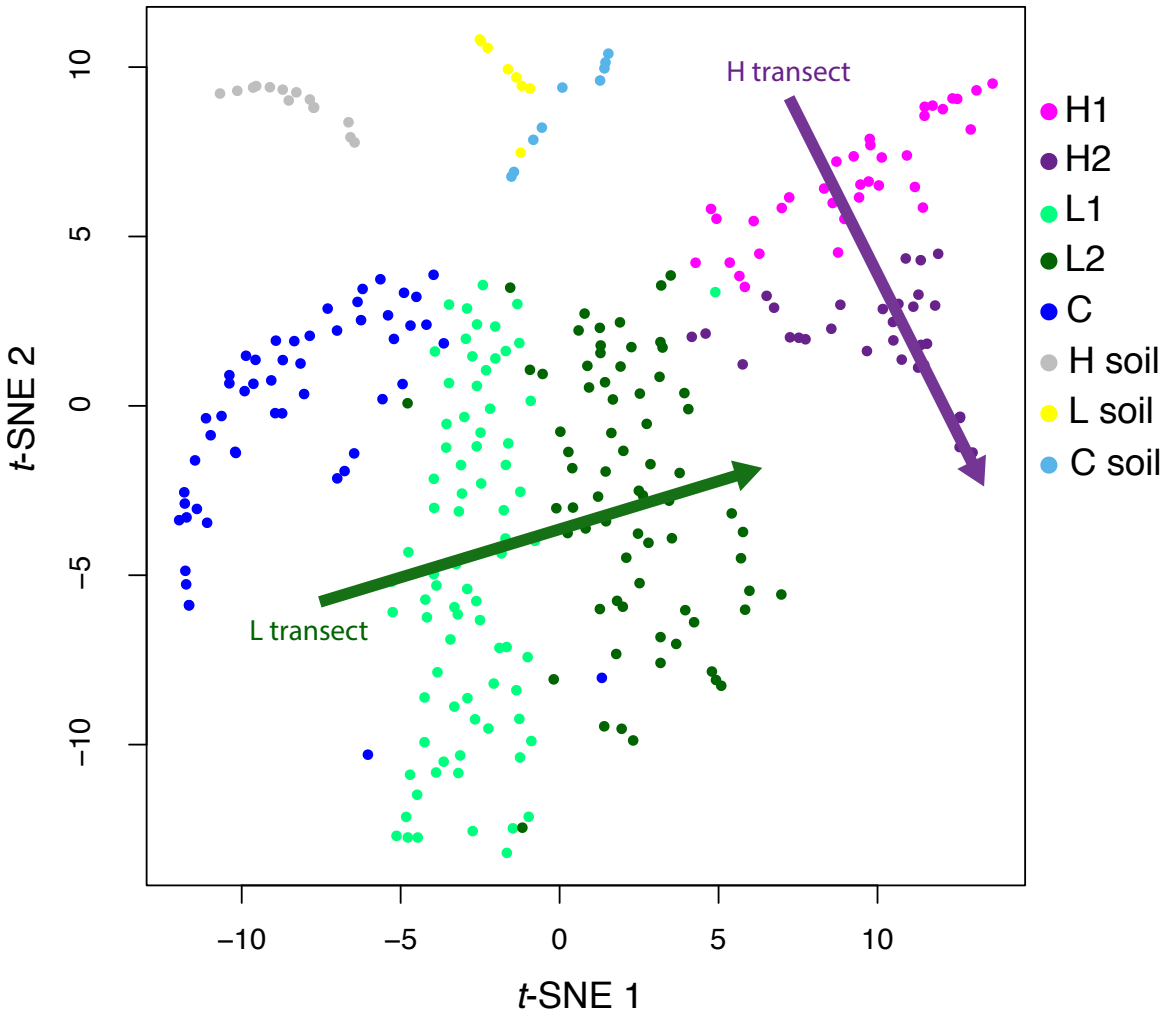


Figure 3. *t*-SNE analysis of genome abundance for each sediment sample. Each of the 300 shown genomes was assigned to the sample where it has the greatest abundance. Shaded arrows display the approximate direction of water flow, from upstream to downstream, for the high (green) and low (purple) transects.

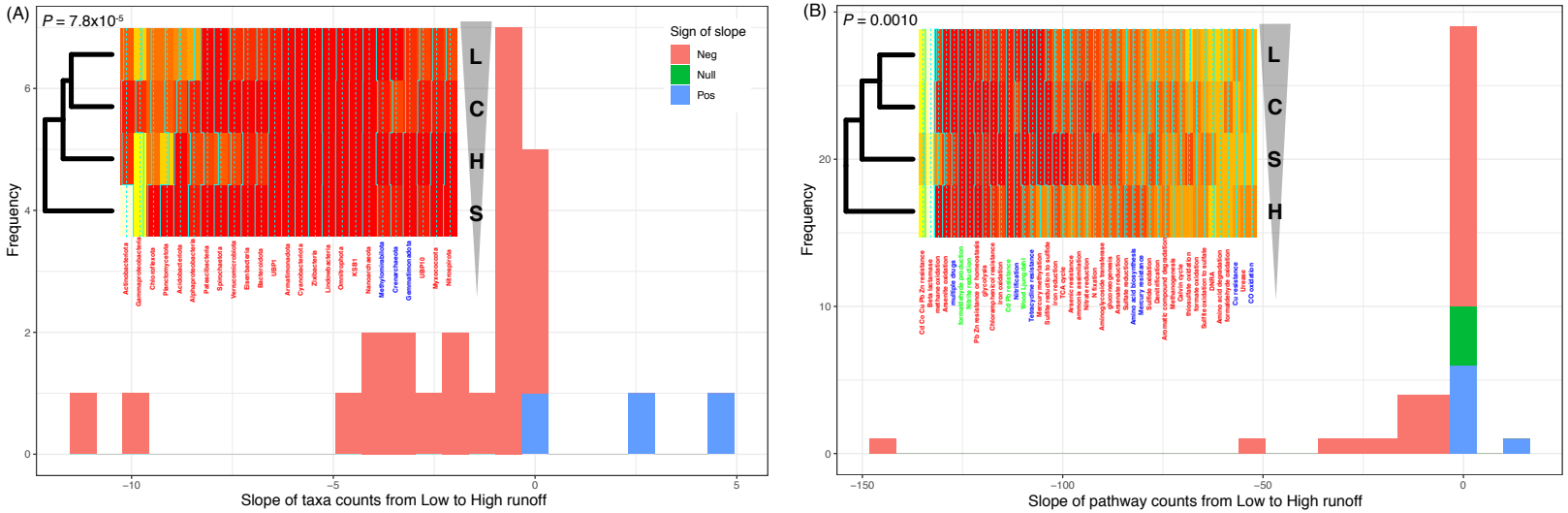


Figure 4. Transition from low to high runoff leads to a decrease in diversity. (A) Distribution of the slopes of taxonomic counts as a function of sites. (B) Distribution of the slopes of pathway counts as a function of sites. In both cases, counts were aggregated by location types (L [Low], C [Control], S [Soil], and H [High] 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.

Online methods

Sample collection and processing

Sediment and soil cores were collected from Lake Hazen (82°N, 71°W: Figure 1A), located within Quttinirpaaq National Park, on northern Ellesmere Island, Nunavut. Sampling took place between May 10 and June 10, 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 two transects, the H1/H2 and L1/L2 sites, respectively. The lake then drains via the Ruggles River along its south-eastern 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 deeper low (L2) and high (H2) runoff sites represented depositional periods of 30 years (1987-2017) and 6 years (2011-2017), respectively (Supplementary Table 3).

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) summarized in Supplementary Table 3. Contamination of samples was minimized by wearing non-powdered latex gloves during sample handling and sterilizing all equipment with 10% bleach and 90% ethanol before sample collection. 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). The soil samples (S sites) were collected from three sites in the dried streambeds of the tributaries, on the northern shore between the two transects. At each site, for both sediments and soil, five cores were sampled, ∼3 m apart for the sediment cores, and approximately ∼1 m apart to account for site heterogeneity.

For sediment core, one of the five cores were used for microprofiling of oxygen (O_2), redox and pH, as well as one core for porewater chemistry and loss on ignition (see [1] for details), and the remaining three cores were combined, prior to their genomic analysis, here again to account for site heterogeneity. For soil samples, three cores per site were collected for DNA analysis, but no additional cores were collected for chemical analyses. As we were mostly interested in the community composition through space, we combined the top 5 cm of sediment and 10 cm of soil for sample extraction and subsequent sequencing. Any remaining length of cores that were used for DNA analysis were discarded. These uppermost layers in the sediment correspond to both the most recent sediment deposition dates [2] and the region of greatest microbial activity [3]. The top of each core was sectioned and placed into Whirlpack bags. These slices were homogenized 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. Soil samples were transferred into falcon tubes, homogenized, and stored as described above for the lake sediment samples.

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 [4, 5]. 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 Glutamine synthetase gene (glnA) as this gene is ubiquitous across microbial life (Supplementary Figure 1 and Supplementary Table 5). Adequate DNA concentrations for sequencing were reached by combining triplicate extractions for a total volume of 45 μl 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 contained no trace of DNA and were not sequenced.

Chemical analyses

Redox potential, pH, and dissolved O_2 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 nitrogen (N_2) before capping. Sediment porewater was extracted following centrifugation at 4,000 rpm. The supernatant was then filtered through 0.45 μm cellulose acetate filters into 15 ml tubes, and were frozen until analysis. Concentrations of nitrates and nitrites ($NO_2^- + NO_3^-$), and ammonia (NH_3), chloride (Cl^-) were measured in the sediment porewater using a Lachat QuickChem 8500 FIA Ion Analyzer, while total dissolved phosphorus (TDP) and SO_{4-}^{2-} were 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 calcium

carbonate ($CaCO_3$) and organic carbon (OC) were estimated through loss on ignition [6].

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 summarized using a Principle Component Analysis (PCA) and projections were clustered using Partitioning Around Medoids [7]. The appropriate number of clusters was determined from silhouettes with the R package hopach [8]. The Dunn test [9] 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) [10] 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) [11] 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 [12] software with a k-mer size of 31 and "meta-large" setting (see Supplementary Table 6 for additional summary statistics). EukRep [13] was used to remove any eukaryotic DNA from the contigs prior to the formation of an Anvio (v5) [14] 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 [15] with default settings. The resulting SAM files were sorted and converted to BAM files using samtools (v0.1.19) [16]. 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 [17]. 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 [18] 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) [19]. Anvio's custom Hidden Markov models were run, along with NCBIs COG [20] annotation to identify protein-coding genes. PFAM [21], TIGRFAM [22], GO terms [23], KEGG enzymes and pathways [24], and Metacyc pathways [25] were predicted with Interproscan (v5) [26]. 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) [27]. 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 [28]. 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) [29]. CheckM was used again to identify contigs that were not contained in any of the 300 high-quality genomes, that is those whose size ranges from 1000–2500 bp. As an attempt to "rescue" these unbinned contigs, an alternative binning algorithm MaxBin (v2.0) [30] 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 marker genes 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 [31]. The marker genes were extracted from each genome by matching Pfam72 (v31) [21] and TIGRFAMs73 (v15.0) [22] annotations from GTDB (v86) [31]. Marker genes from each of the 300 genomes were translated using seqinr [32], 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,

and their concatenated sequences were added those from the reference GTDB sequences. MUSCLE (v3.8.31) [33] was employed to construct the alignment in R (v 3.5.1) [34]. Archaeal sequences were removed from the bacterial alignment on the basis of results from CheckM [27] and independently verified using a custom list of archaea specific marker genes. Alignments were then refined using trimAI [35] and the "-gappyout" parameter. FastTree2 [36], 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 [37, 38, 39]. The archaeal tree was rooted with Euryarchaeota and the bacterial tree was rooted with Patescibacteria using APE [40]. Trees were visualized and colored by phylum with ggtree [41].

Community composition of the MAGs

To determine the relative abundance of each genome in the eight samples, sample-specific genome abundances were normalized 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 [27]. 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 visualized using phyloseq [42] (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 [43].

Metabolic potential of the MAGs

To analyze 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 [44, 45] (Supplementary Data Files on GitHub). Pathways were also predicted using MinPath [46] to map all identified KEGG enzymes to the most parsimonious MetaCyc pathways [25]. 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 ≤ 0.25 (on a $-log_{10}$ scale) was set for a genome to be included in the functional analyses at any site, so that only the most abundant genomes for each site were considered. We aggregated marker genes and pathways by function, summarizing the results by phyla, 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) among sites using a hierarchical clustering; significance was derived from the Approximately Unbiased bootstrap [47] and Fisher's exact test.

Data availability

Scripts and supplemental data files can be accessed from `https://github.com/colbyga/hazen_metagenome_publication`. Raw sequence reads of the shotgun metagenomic data were submitted to the sequence read archive (SRA) under accession no. SRP218124 and under Bioproject PRJNA556841. The geochemical data were submitted to the National Science Foundation's (NSF) Arctic Data Center repository under doi:10.18739/A20R9M41W.

References

1. 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).

2. Pierre, K. S. *et al.* Contemporary limnology of the rapidly changing glacierized watershed of the world's largest high arctic lake. *Scientific reports* **9**, 4447 (2019).

3. 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).

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

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

6. 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).

7. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. *cluster: Cluster Analysis Basics and Extensions* (2019). R package version 2.1.0 — For new features, see the 'Changelog' file (in the package source).

8. 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).

9. Dinno, A. *Dunn's Test of Multiple Comparisons Using Rank Sums* (2017). R package version 1.3.5 — For new features, see the 'Changelog' file (in the package source).

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

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

12. 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).

13. 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).

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

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

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

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

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

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

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

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

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

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

24. 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).

25. 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).

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

27. 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).

28. 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).

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

30. 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).

31. 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).

32. 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.

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

34. 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.

35. 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).

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

37. 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).

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

39. Aris-Brosou, S. & Rodrigue, N. A not-so-long introduction to computational molecular evolution. *Methods Mol Biol* **1910**, 71–117 (2019).

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

41. 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).

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

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

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

45. 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).

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

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

Supplementary Text

Nutrient cycles affected when transitioning from low to high

Overall, markers of carbon and sulphur metabolism significantly decreased when transitioning from the L to H sites, even if nitrogen metabolism was not (Tab. S8). Most carbon pathways, such as carbon fixation through the Calvin-Benson-Bassham (CBB) pathway, as well as the capacity for simple carbon metabolism, were shared across all runoff regimes. In contrast, carbon oxidation and reduction reactions regulated through Wood-Ljungdahl pathway were only observed in the H sites, where sedimentary conditions were anoxic throughout the first 5 cm (Figs. S4, S9). Here, Spirochaetota were likely performing anaerobic respiration and carbon fixation producing acetate as an end product. Methanogenesis pathways were present across all sites, but notably, methane oxidation pathways were absent from high runoff sites, where oxygen is limited.

Greater concentrations of ammonia in the high runoff regimes may suggest that N-containing organic matter was mineralised through ammonification (Fig. S8). In the high runoff regime, there was both an absence of nitrification and a greater presence of markers for ammonia assimilation. Markers for dissimilarity nitrate reduction (DNRA) were present in multiple genomes across all runoff regimes (Fig. S8). In contrast, urease markers were found more abundantly in low runoff regimes, where ammonia concentration was lower (Fig. S5). The functional ability of microbes to cycle sulphur between oxidised and reduced forms was significantly different between the high and low runoff regimes (Tab. S8). In the high runoff regime, Gammaproteobactiera were the only organisms with the metabolic capacity to expansively utilise sulphur, performing sulphide oxidation and thiosulphate reduction. Whereas, sulphate reduction was predominantly found in the soil, control, and low runoff regimes.

Aside from nutrient cycling, we also assessed the capacity of microbial communities to process metals and antibiotics. Metal resistance and cycling was mostly ubiquitous throughout all of the sites, regardless of runoff. Methyl mercury production, identified by the presence of both hgcA and hgcB genes [1], was only implicated in the high runoff sites, in Spirocheatoa and Chloroflexoata. However, genes conferring mercury resistance involved in the conversion of inorganic Hg^{II} to the less toxic Hg^{0} – were evenly distributed throughout the sites. There was a broad presence of metal tolerance that was indicated by genetic determinants related to heavy metal resistance of cadmium, cobalt, copper, lead and zinc (Fig. S9). Furthermore, antimicrobial resistance genes specifically β-lactamases, were ubiquitous across all genomes at all sites. We identified 90 genomes with antibiotic resistance genes (Fig. S**??**). Drug resistance was prevalent throughout all phyla, including one Creanarchaeota (archaeal) genome (Figs. S9, S**??**). Finally, we found that while amino acids were readily synthesised and degraded by most organisms (Fig. S9), the degradation of polycyclic aromatic compounds appeared to be least prevalent in the high runoff sites.

Sites clustered by runoff regime

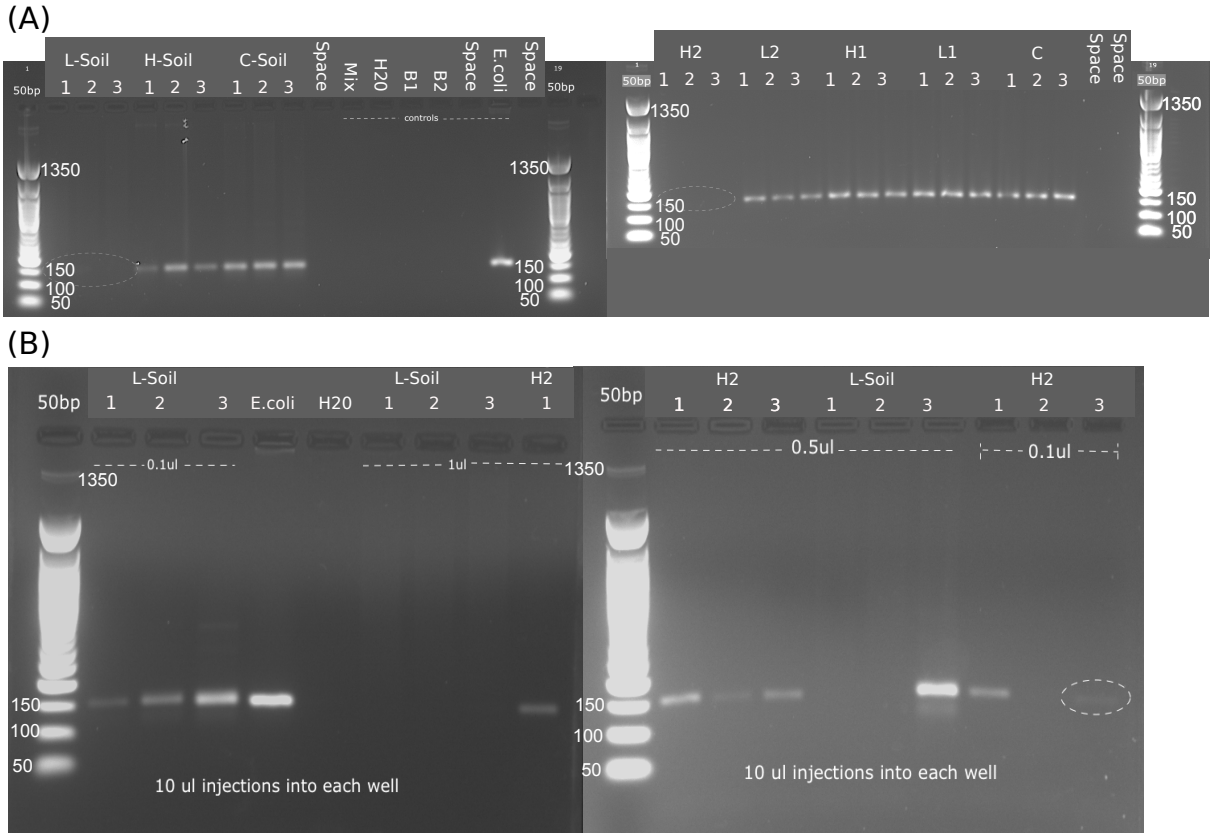
To identify potential drivers of this reduction in diversity when going from the L to the H transects, we bi-clustered genomes by their normalised abundances (on a -log10 scale to reduce skew) and by sample site (Fig. S12), and found that sites clustered following a similar pattern to geochemical features (see Fig. 1B), with H sites grouping separately from L sites (Fig. 3A). The normalised abundances of MAGs showed no strong phylogeographic pattern, in that we did not observe an assemblage of MAGs solely representative of a given site (Fig. S13). In spite of this absence of phylogeographic pattern, the tanglegram suggests that the beta diversity of highly abundant MAGs in the L/C sites was greater than at the H sites (see the distribution of green lines connecting the phylogenetic and clustered trees in Tab. S13). This difference in diversity between samples is further supported by an NDMS ordination (Fig. S14) and a PERMANOVA test on a PCoA ordination (Fig. S14).

References

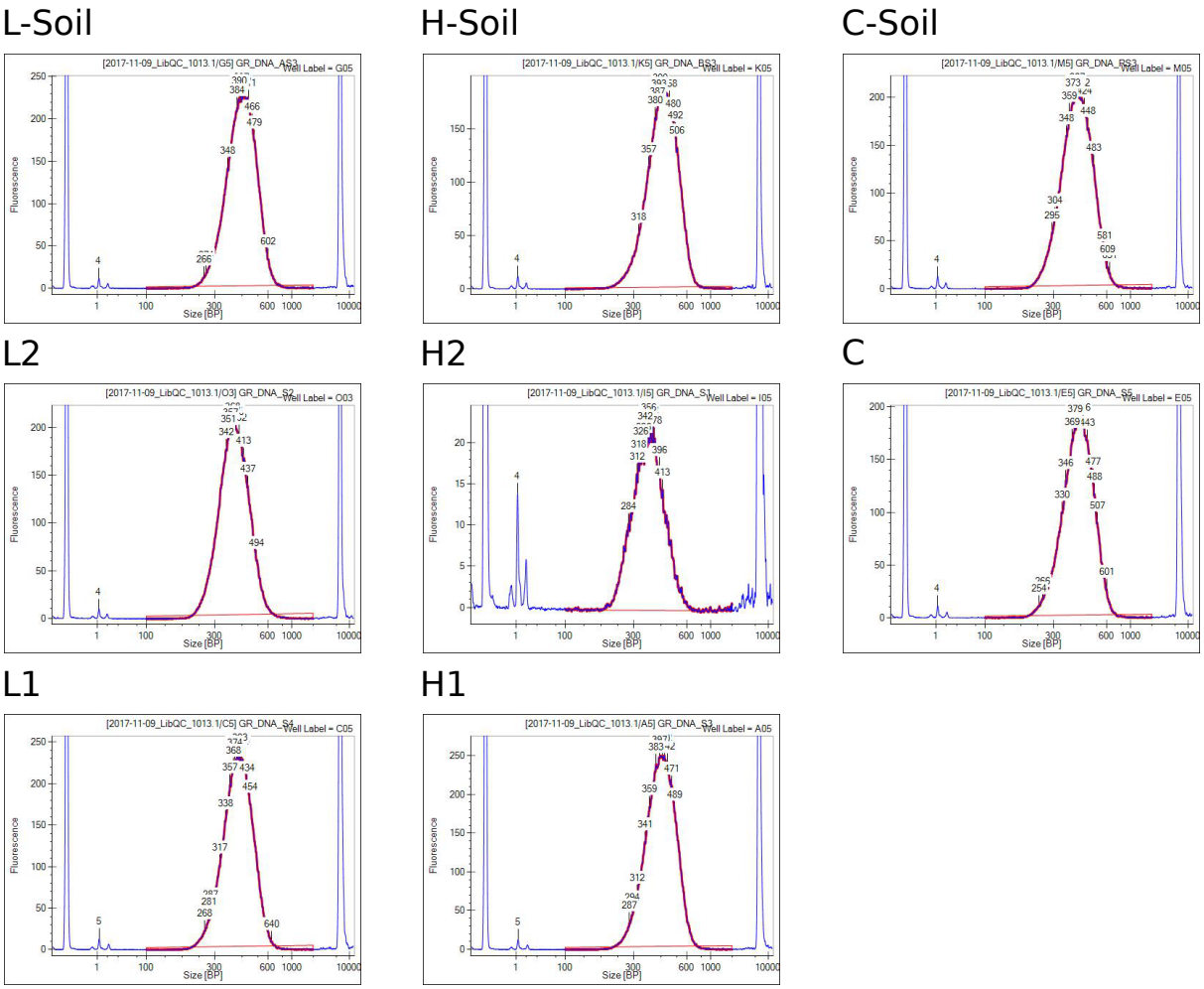
1. Parks JM, Johs A, Podar M, Bridou R, Hurt RA, Smith SD, et al. The genetic basis for bacterial mercury methylation. Science. 2013;339(6125):1332–1335.

2. Pierre KS, Louis VS, Lehnherr I, Schiff S, Muir D, Poulain A, et al. Contemporary limnology of the rapidly changing glacierized watershed of the world's largest High Arctic lake. Scientific reports. 2019;9(1):4447.

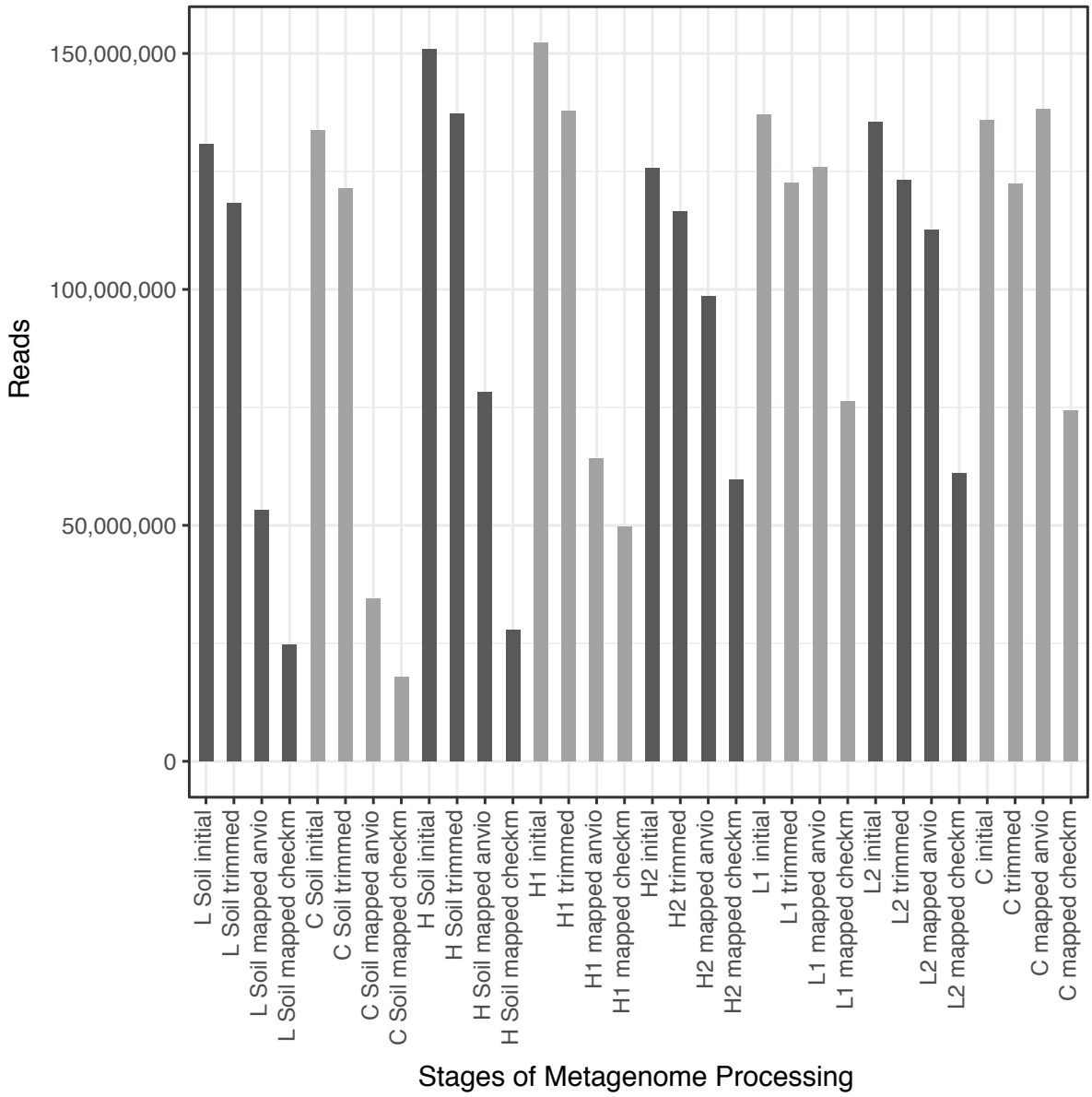
Supplementary Figures



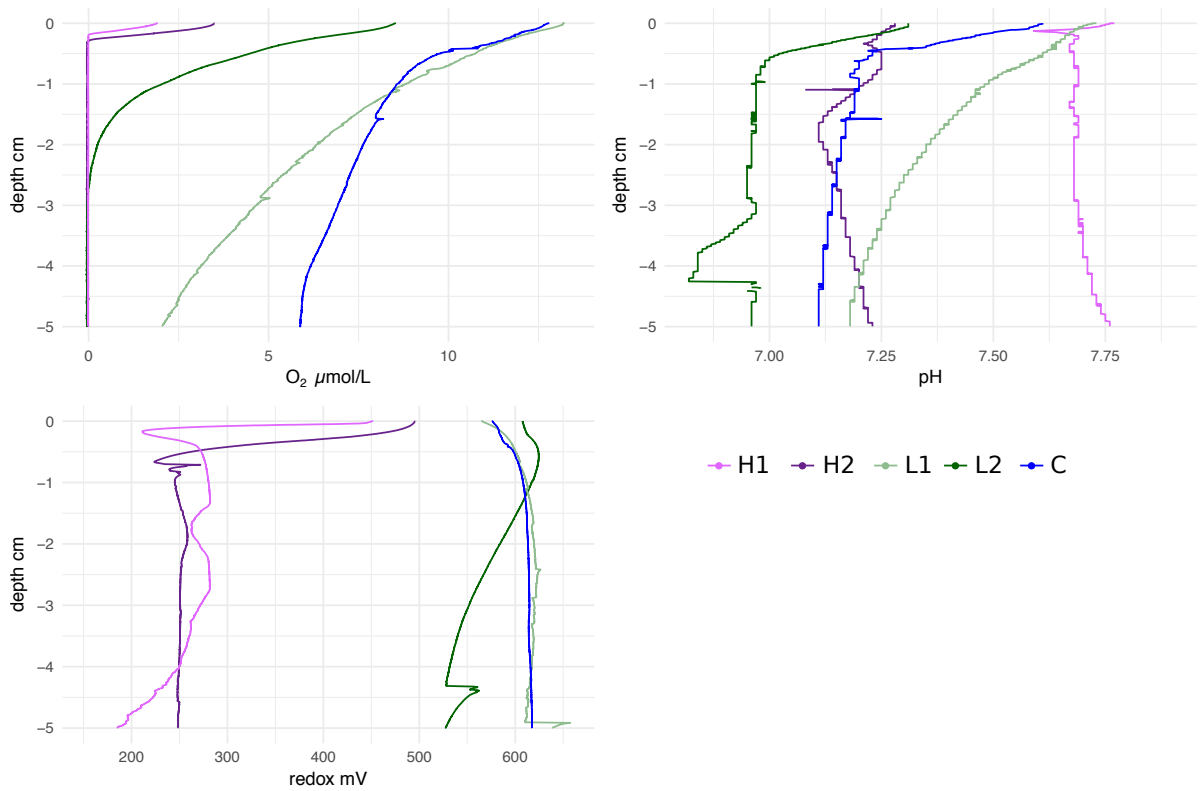
Supplementary Figure 1. Gel electrophoresis images of the *glnA* gene for each sample extracted in triplicate. (A) Initial gel results using 1 μl of DNA for each PCR reaction. Every sample contains *glnA* except L-Soil and H2. (B) Repeated PCR reaction for L-Soil and H2. Diluted DNA concentrations for PCR from 1 μl to 0.1 μl and 0.5 μl.



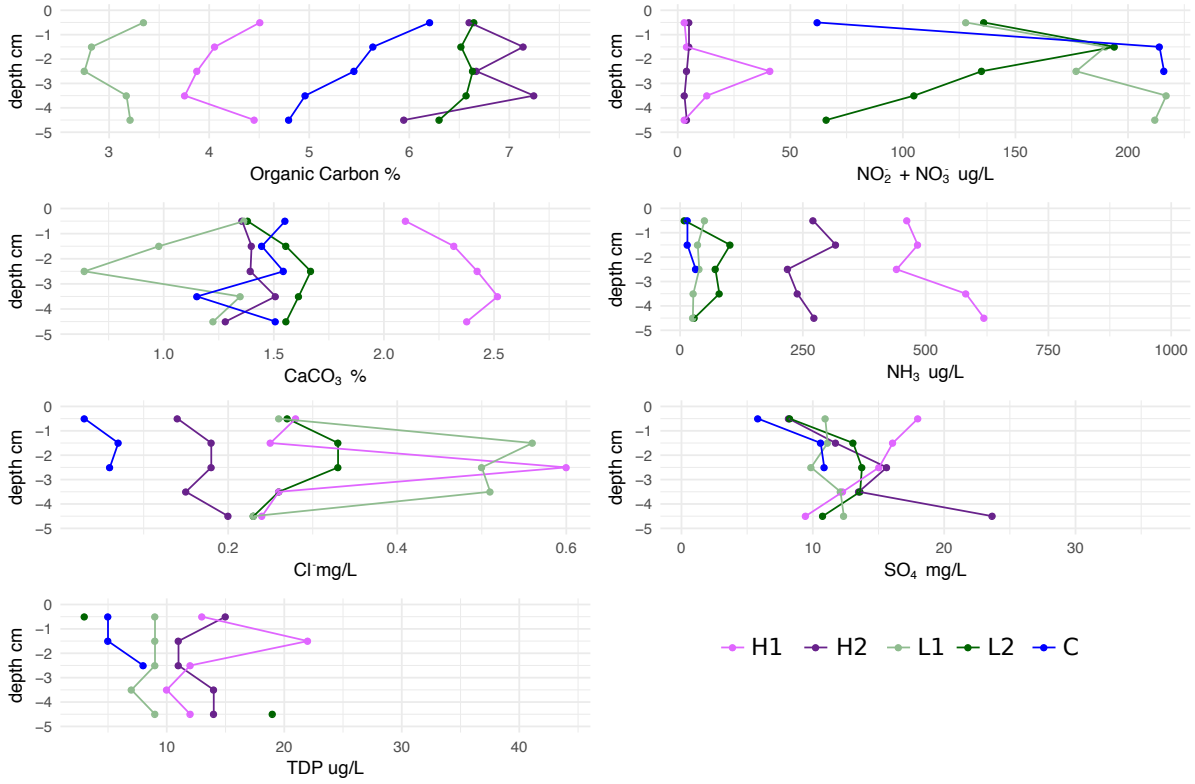
Supplementary Figure 2. Illumina HiSeq DNA sequencing library validation from an Agilent Technology 2100 Bioanalyzer. All libraries were constructed using Illumina TruSeq DNA PCR-Free Library Prep. The library results for each of the eight samples is presented and labelled above the panel: L-Soil, H-Soil, C-Soil, L2, H2, C, L1, H1. Peaks between 300-600 bp indicate the presence of DNA. Peaks below 1 bp and above 1000 bp are internal standards.



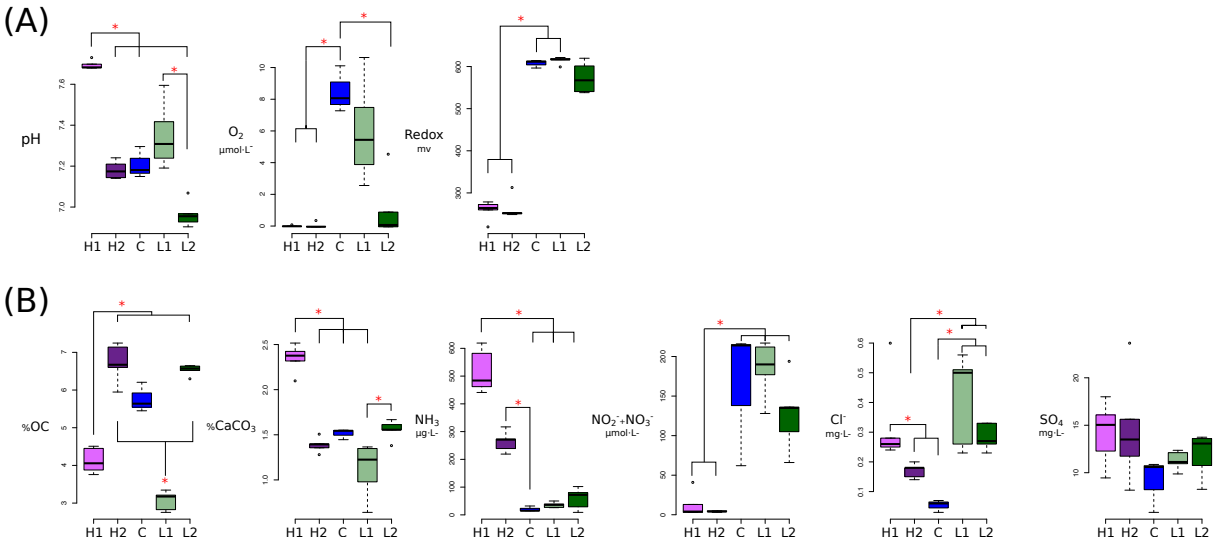
Supplementary Figure 3. Number of reads throughout analysis. Reads are successively decreasing, except for the reads mapped to the Anvio contig database, as each read can map to more than one contig and be counted more than once. The reads mapped using checkM include only reads that map to the 300 high quality reconstructed genomes, opposed to all reconstructed genomes in the Anvio mapping.



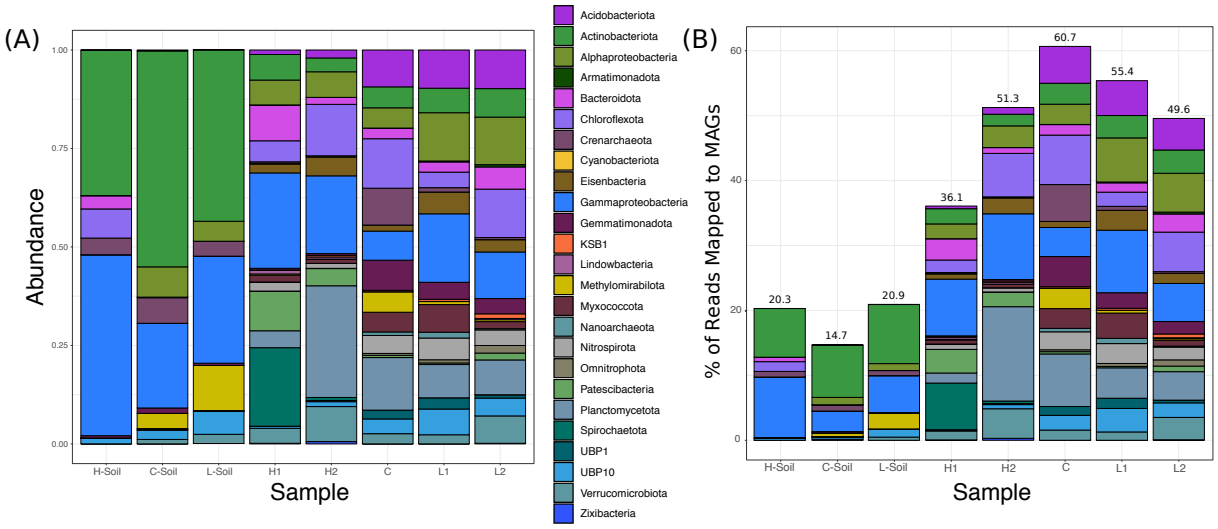
Supplementary Figure 4. Sediment microprobe profiles for oxygen (O_2), pH, and redox measured in 100 μm intervals. Legend (bottom right) indicates the colour of each sediment site in the profiles.



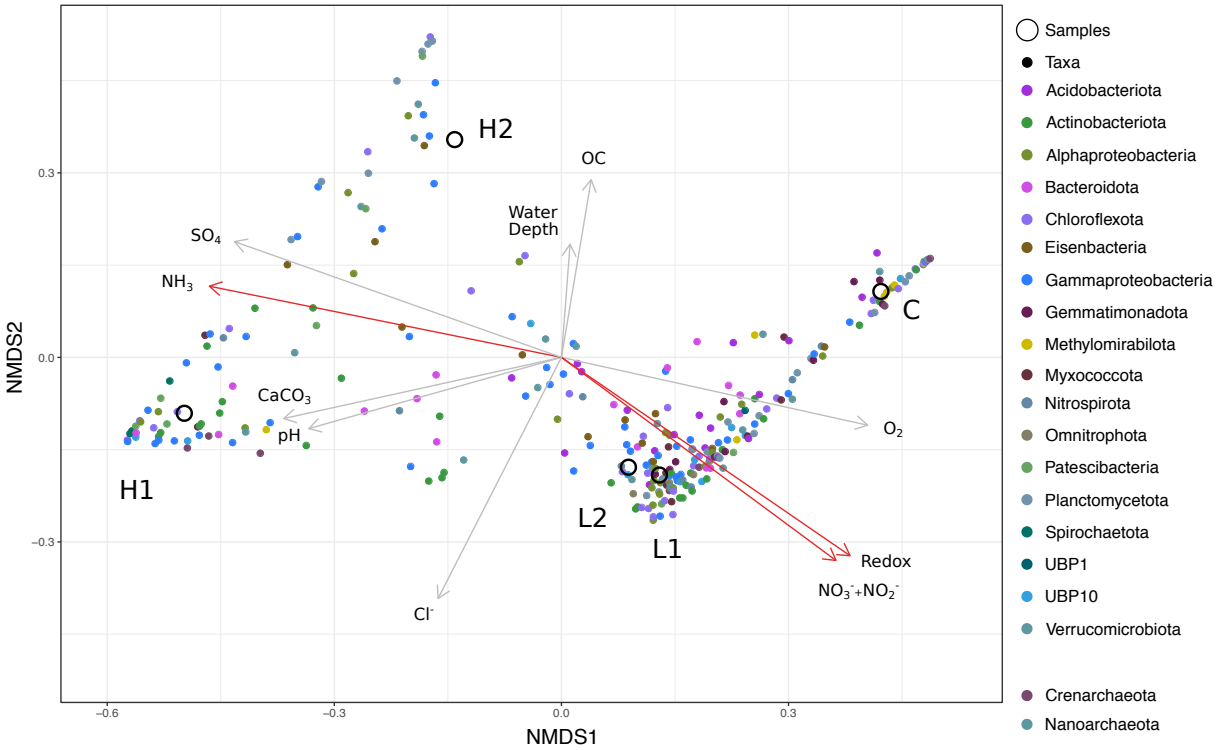
Supplementary Figure 5. Sediment porewater profiles for organic carbon, nitrite and nitrates, calcium carbonate, ammonia, chlorine, sulphate, and total dissolved phosphorus (TDP). TDP was removed when producing PCA and boxplots in Figure 1 because of incomplete measurements in C and L2. Legend (bottom right) indicates the colour of each sediment site in the profiles.



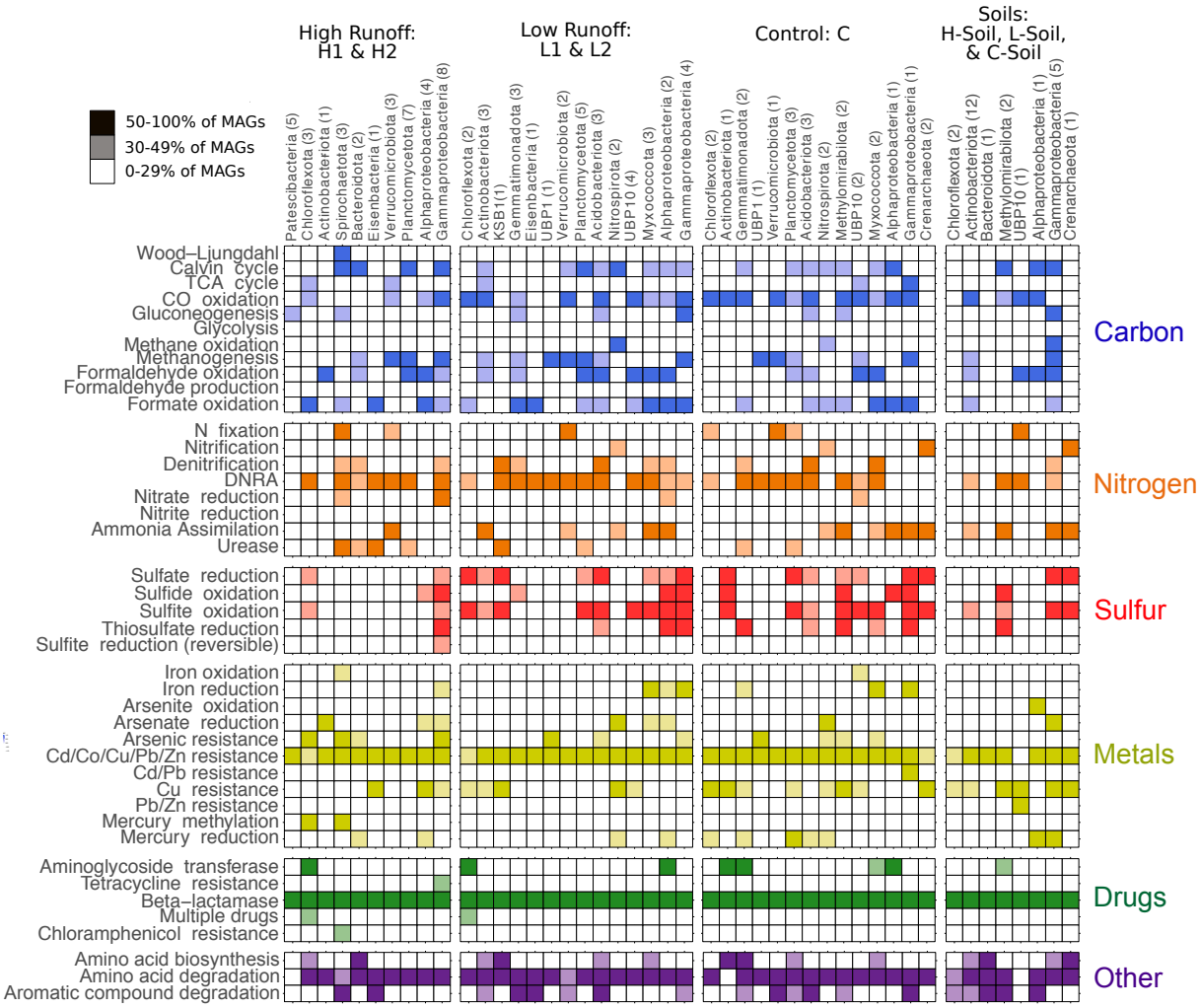
Supplementary Figure 6. Distribution of all chemical features for sediment sites. Includes chlorine and sulphate measurements absent from Figure 1. Branches and asterisks indicate significant differences between sites $P < 0.025$ (Dunn Test). If branches form a dichotomy or trichotomy, the interactions within that group is not significant. (A) Microprobe measurements collected at every 100 μm. (B) Porewater measurements collected from bulk 1 cm intervals.



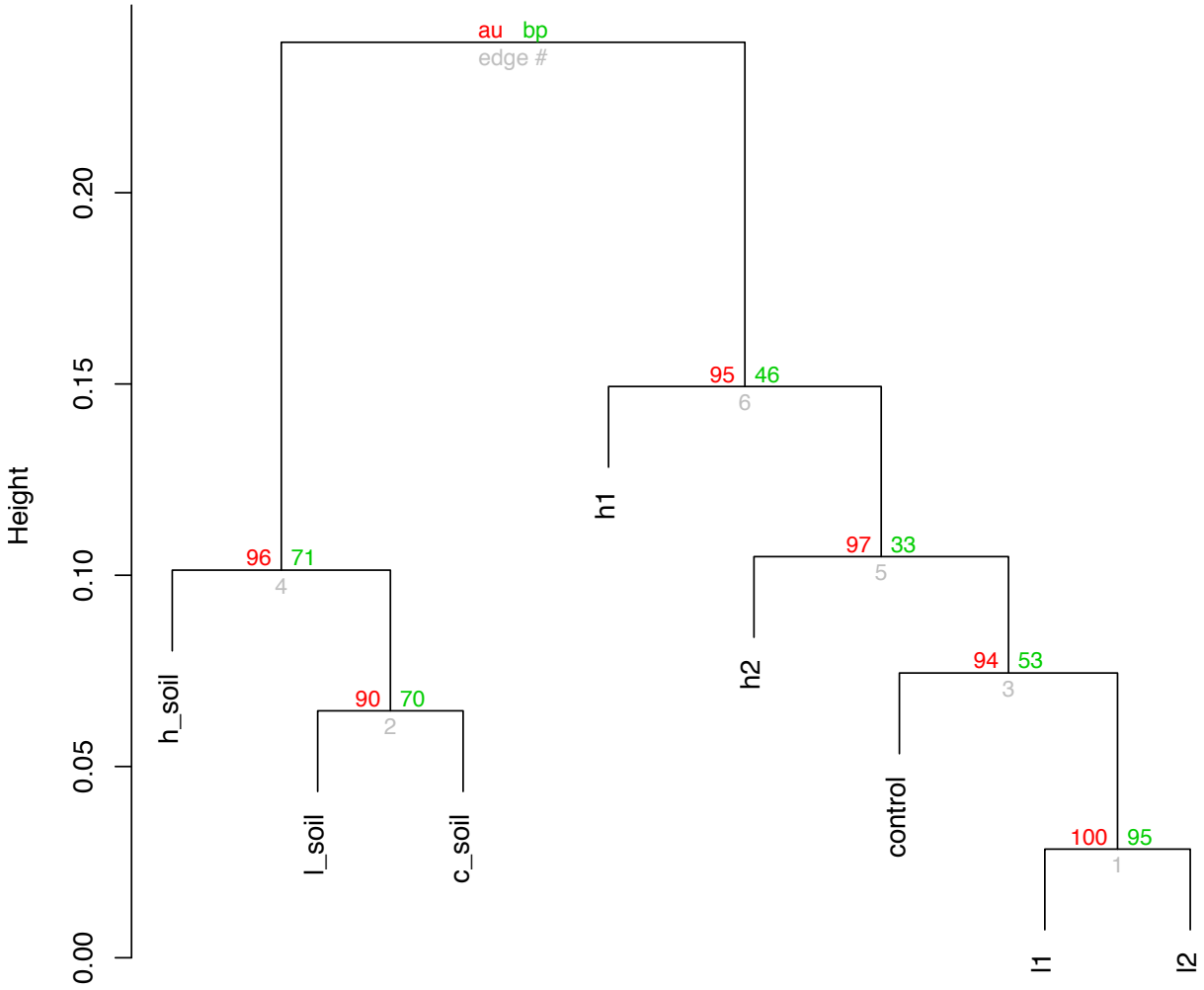
Supplementary Figure 7. Reconstructed genome abundance across sites. (A) Amount of reads mapped to the 300 high quality genomes for each sample normalised to 100%. Only reads that were mapped to genomes are shown and all unmapped reads have been excluded. (B) Amount of reads mapped to the 300 high quality genomes for each sample.



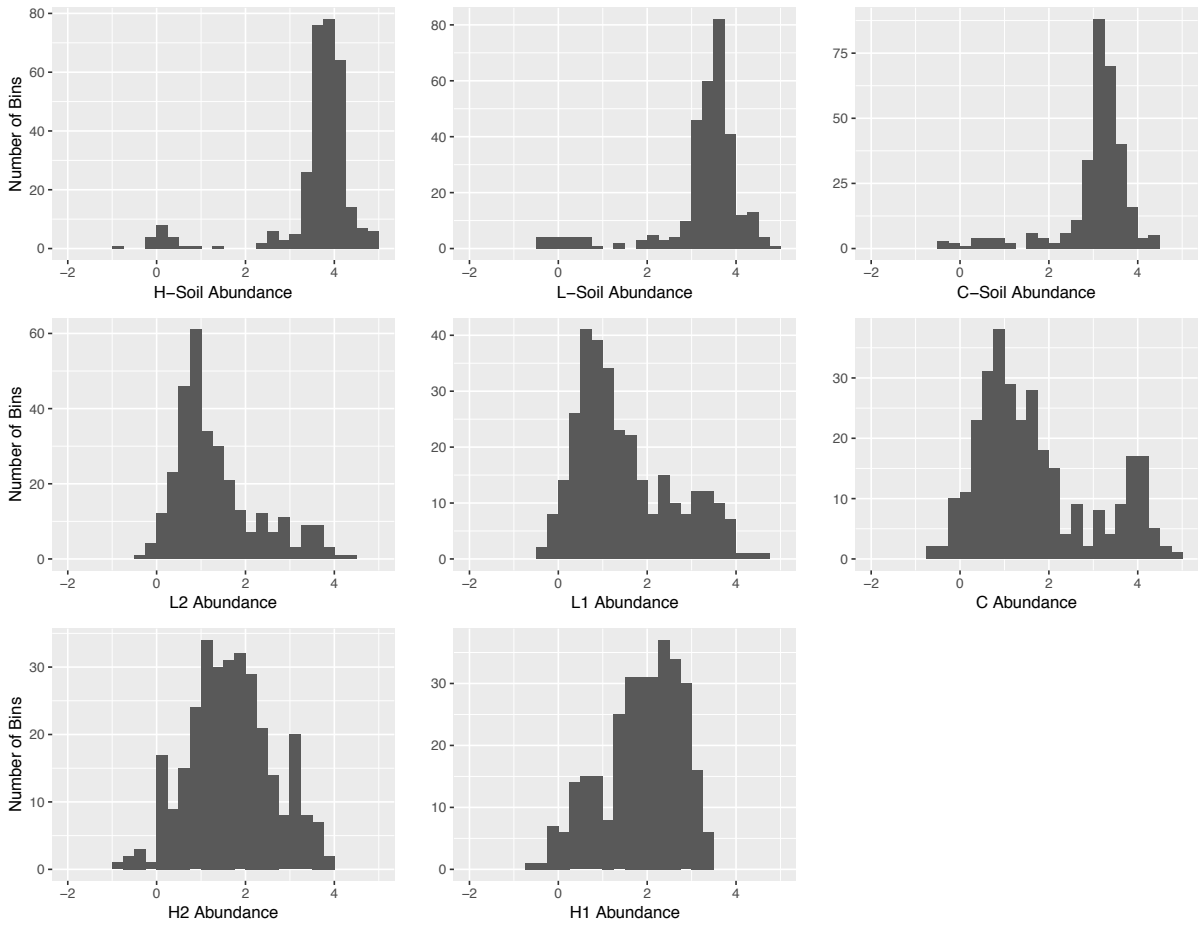
Supplementary Figure 8. NMDS analysis of genomes from 20 most abundant phyla in sediment sites. Physical and chemical vectors are fitted to data. Red vectors are significant (permutation test: $P < 0.05$).



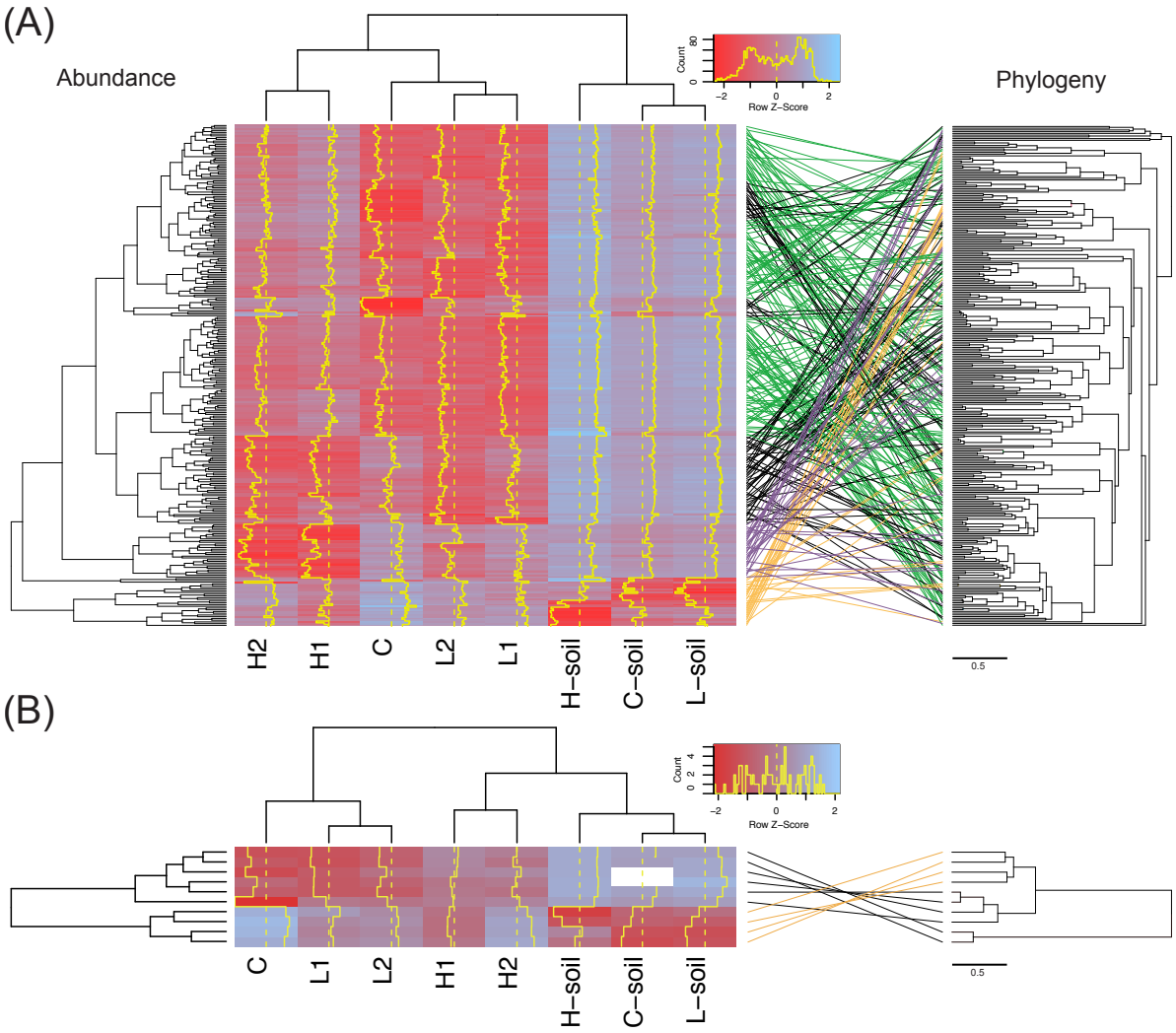
Supplementary Figure 9. Metabolic capacity of genomes separated by hydrological regime. Genomes are only considered to contribute to a site if 0.56% ($-\log_{10} \leq 0.25$) of reads per sample mapped to a genome. Presence of core metabolic genes involved in carbon metabolism, nitrogen metabolism, sulphur metabolism, metal cycling, antibiotic resistance, and other metabolism are shown. Number of genomes for each taxa are shown in parentheses. Blank: genes predicting function are absent or in low abundance. Shaded colours: genes predicting function are present in 30-50% of genomes per phylogenetic group. Dark colours: genes predicting function are present in >50% of genomes per phylogenetic group.



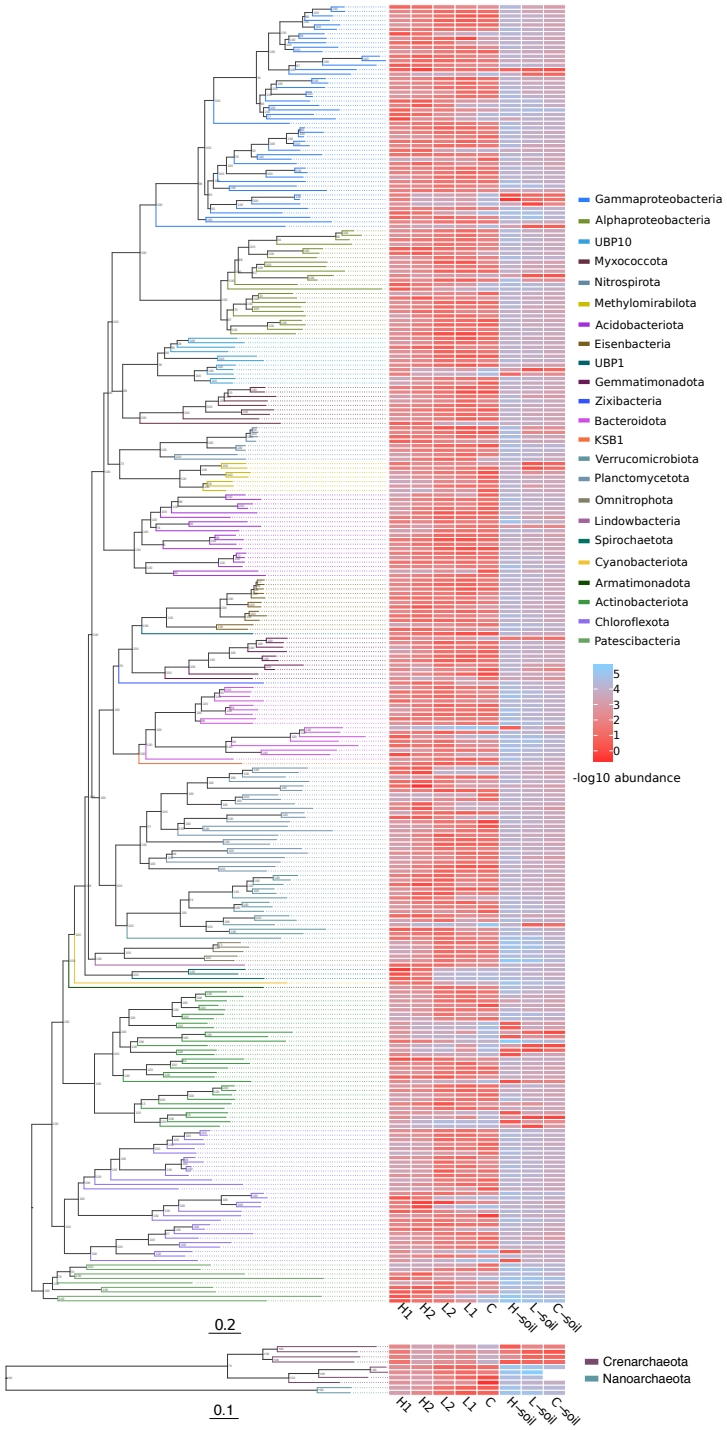
Supplementary Figure 10. Metabolic capacity for carbon, sulphur, and nitrogen cycles clustered with P-values via multiscale bootstrap resampling. AU (approximately unbiased) P-values are shown in red, with any value > 0.95 (significance level 0.05) being significant. BP (bootstrap proportions) values are shown in green. Hierarchical clustering was completed using "correlation" as distance method and "average" as clustering method.



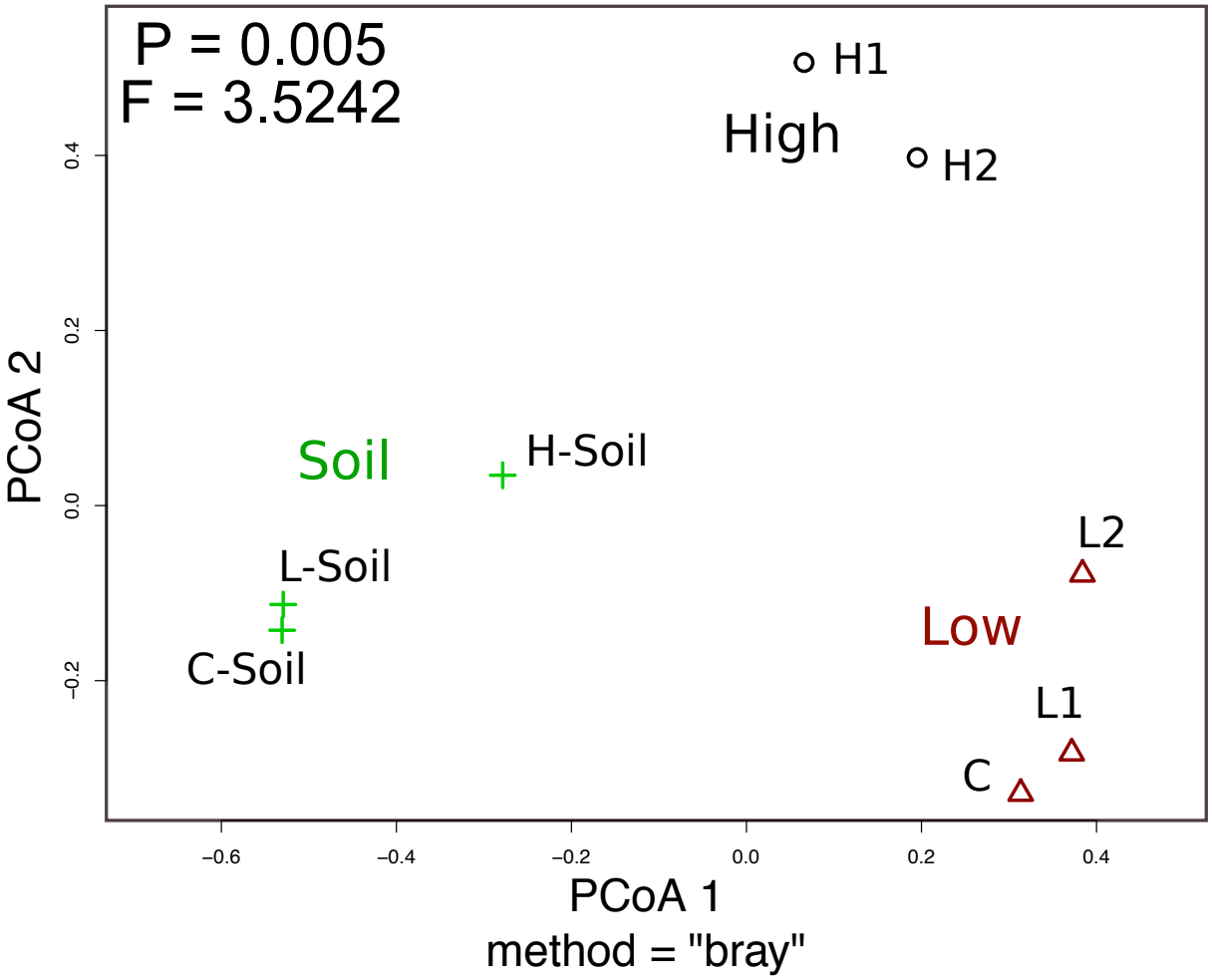
Supplementary Figure 11. Genome abundance per sample ($-\log_{10}$ scale).



Supplementary Figure 12. Genome abundance heatmap, and tanglegram separated by (A) bacterial genomes and (B) archaeal genomes. Left: The heatmap displays genome abundance per site normalised by amount of reads in each sample and transformed to a $-\log_{10}$ scale. Dotted yellow lines represent mean abundance values, and yellow traces represent the raw z-scores above (red) and below (blue) the mean. The abundance values are grouped both by sites (top dendrogram) and genomes (left dendrogram). Right: Tanglegram between dendrogram clustered by similar abundances and phylogenetic tree. Highlights in tanglegram: orange lines are genomes abundant in soil, green lines are genomes abundant in low runoff sediment, purple lines are genomes abundant in high runoff sediment, black lines are genomes shared in multiple environments.



Supplementary Figure 13. Abundance of genomes are presented on a $-\log_{10}$ scale were more negative values (red) are more abundant than positive values (blue). The abundance values correspond by row with the genomes phylogenetic assignment. Support values for phylogenetic tree are shown at each node.



Supplementary Figure 14. Principal coordinate analysis (PCoA) on reconstructed genome abundances for all sites. Grouping the sites into soil, high, and low runoff was proven to be significant (PERMANOVA test: $F = 3.52$, $P = 0.005$).

Supplementary Tables

Supplementary Table 1. Coordinates of Lake Hazen sediment and soil sites with temperature and date at time of sampling.

Sample	Location	Temperature (°C)	Date
H-Soil	81° 84' 840" N; 70° 83' 849" W	-5	June 3, 2017
L-Soil	81° 80' 332" N; 71° 54' 239" W	0	June 7, 2017
C-Soil	81° 79' 382" N; 70° 44' 486" W	-2	June 3, 2017
H1	81° 84' 150" N; 70° 85' 175" W	not measured	May 24, 2017
H2	81° 82' 493" N; 70° 71' 498" W	not measured	May 29, 2017
C	81° 80' 343" N; 70° 50' 447" W	not measured	May 27, 2017
L1	80° 80' 521" N; 70° 52' 699" W	not measured	June 1, 2017
L2	81° 79' 171" N; 71° 46' 926" W	not measured	June 2, 2017

Supplementary Table 2. Sediment deposition dates and rates for the two deep sites, H2 and L2, is based on ^{210}Pb constant rate of supply (CRS) dating model. Analysis was completed on 0.5 cm core intervals. Data adapted from previous study [2].

H2 (Abbe Deepsite)			L2 (Blister Deep Site)		
Interval	**Midpoint CRS date (CRS year)**	**Sedimentation rate (g/cm^2/yr)**	**Interval**	**Midpoint CRS date (CRS year)**	**Sedimentation rate (g/cm^2/yr)**
0-0.5	2017.1	0.349	0-0.5	2016.5	0.0724
0.5-1	2016.6	0.292	0.5-1	2014.0	0.0701
1-1.5	2015.9	0.111	1-1.5	2011.3	0.1163
1.5-2	2014.1	0.671	1.5-2	2009.2	0.1759
2-2.5	2012.7	0.671	2-2.5	2007.3	0.1232
2.5-3	2012.7	0.227	2.5-3	2004.7	0.0909
3-3.5	2012.0	5.563	3-3.5	2001.9	0.0967
3.5-4	2011.3	5.563	3.5-4	1998.8	0.0854
4-4.5	2011.3	5.563	4-4.5	1994.0	0.0513
4.5-5	2011.3	5.563	4.5-5	1987.2	0.0375
5-5.5	2011.3	5.563	5-5.5	1976.7	0.0220
5.5-6	2011.3	5.563	5.5-6	1966.51	0.0515
6-6.5	2011.3	5.563	6-6.5	1957.66	0.0339
6.5-7	2011.3	5.563	6.5-7	1949.00	0.0537
7-7.5	2011.3	5.563	7-7.5	1940.41	0.0451
7.5-8	2011.3	5.563	7.5-8	1931.79	0.0668
8-8.5	2011.3	5.563	8-8.5	1925.07	0.0553
8.5-9	2011.3	0.441	8.5-9	1918.10	0.0553
9-9.5	2010.9	0.224	9-9.5	1911.08	0.0553
9.5-10	2009.8	0.123	9.5-10	1904.42	0.0553

Supplementary Table 3. Glacial runoff to the Lake Hazen Watershed. Lengths of rivers in km are shown in parentheses. Mass balance modelled runoff for years 2015 and 2016. Sampling dates in 2017 were prior to the summer runoff. Data adapted from previous study [2].

	Surface area (km^2)			**Runoff volume (km^3)**	
River	**Catchment**	**Glacier**	**River**	**2015**	**2016**
High Runoff: Abbé (AB)	390	204	7.9 (21)	0.061	0.015
Low Runoff: Blister (BR)	n/a	6	2.5 (11)	0.002	<0.001
Gilman (GL)	992	708	5.1 (22)	0.192	0.08
Henrietta Nesmith (HN)	1274	1041	9.6 (4.6)	0.291	0.075
Snowgoose (SG)	222	87	5.6 (17)	0.026	0.006
Turnabout (TN)	678	259	13.4 (42)	0.082	0.024
Very (VR)	1035	269	32.9 (39)	0.165	0.08
Watershed total	7516	3078	91.2	0.979	0.291

Supplementary Table 4. DNA extraction masses. DNA was extracted in triplicate for each sample and then combined prior to sequencing.

Lake Hazen	Sample Location	Tube ID	Wet Weight (g)				PCR with *glnA*
			1	2	3	Total (grams)	
Sediment	H2: Deephole	S1	0.416	0.455	0.499	1.370	yes (diluted)
	L2: Blister Deep	S2	0.497	0.455	0.389	1.341	yes
	H1: Abbe	S3	0.469	0.514	0.552	1.535	yes
	L1: Blister Shallow	S4	0.429	0.402	0.361	1.192	yes
	C: Ruggles	S5	0.518	0.447	0.331	1.296	yes
Soil	L-Soil: Blister Soil	BS3	0.527	0.535	0.561	1.623	yes (diluted)
	H-Soil: Abbe Soil	AS3	0.443	0.537	0.417	1.397	yes
	C-Soil: Ruggles Soil	RS3	0.447	0.500	0.447	1.394	yes

Supplementary Table 5. DNA fluorescence assay quantification for each sample. Note: H2 sequencing required using full extraction volume of 65 μl to reach an appropriate concentration.

Sample	Volume (μl)	Concentration (ng/μl)	Total DNA (ng)	NanoDrop (concentration ng/μl)
H2	65	0.08	3.36	4.6
L2	42	2.55	107.1	7.6
H1	42	4.33	181.86	10.2
L2	42	17.98	755.16	32.8
C	42	21.2	890.4	36.4
H-Soil	42	38.14	1601.88	65.2
L-Soil	47	69.88	3284.36	149.0
C-Soil	47	74.38	3495.86	110.0

Supplementary Table 6. Contigs were assembled with Megahit and used in Anvio database. Only contigs that were greater than 2500 bp in length were used to form MAGs, however, contigs with a length of 1000 bp or more were used to form the initial database.

	Contigs	**Total Length (bp)**	**Min Contig Length (bp)**	**Max Contig Length (bp)**	**Avg Contig Length (bp)**	**N50 (bp)**	**# of genes prodigal**
Megahit	12026467	8477069127	200	792468	705	756	NA
Anvio	1455655	3414974759	1000	792468	NA	285024	4254625

Supplementary Table 7. Number of bins recovered at each assembly step prior to and after manual refinement with Anvio.

	Anvio-Output (CONCOCT binning)	**CheckM**	**Anvio-Output (manually refined)**	**Final CheckM**
Total Bins	850	850	877	877
Completion, Redundancy				
C: >90, R: <10	52	74	52	72
C: >70, R: <10	180	199	178	198
C: >50, R: <10	321	324	309	300
C: <50, R: any	529	526	568	577
Contamination				
<1	233	261	255	286
<5	544	632	590	676
>10	8	51	7	27
Good bins but contaminated				
C: >50, R: >10	2	27	0	2
C: >70, R: >10	1	6	0	0
C: >90, R: >10	0	2	0	0

Supplementary Table 8. P-values for marker gene and pathway distribution between sites. A Fisher's exact test was used in the place of χ^2 test for small count numbers. P-values of less than 0.05 are considered significant (labelled with an asterisks).

Functional Marker	High vs Low	High vs Control	High vs Soil	Low vs Control	Low vs Soil	Control vs Soil
Carbon Cycle	0.03298 *	0.0004998 *	0.0004998 *	0.07746	0.1384	0.001499 *
Nitrogen Cycle	0.07596	0.001999 *	0.0004998 *	0.7006	0.03448 *	0.1859
sulphur Cycle	0.001999 *	0.002999 *	0.0004998 *	0.4913	0.02999 *	0.3823