Title: Spatial scale structure soil bacterial communities across an Arctic landscape

Running Title: Spatial scale structure Arctic soil bacterial communities

Lucie A Malard\textsuperscript{a,1,\#}, Muhammad Zohaib Anwar\textsuperscript{b,2}, Carsten S Jacobsen\textsuperscript{b} and David A Pearce\textsuperscript{a,\#}

Author affiliations:

\textsuperscript{a} Faculty of Health and Life Sciences, Northumbria University, Newcastle-upon-Tyne NE1 8ST, United Kingdom

\textsuperscript{b} Department of Environmental Sciences, Aarhus University, 4000 Roskilde, Denmark

\textsuperscript{1} present address: Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland

\textsuperscript{2} present address: Department of Pathology and Laboratory Medicine, University of British Columbia, 2211 Wesbrook Mall, Vancouver, British Columbia V6T 2B5, Canada

Corresponding Authors: \#Lucie Malard: lucie.malard@unil.ch and \#David Pearce: david.pearce@northumbria.ac.uk
Abstract

Bacterial community composition is largely influenced by environmental factors, and this applies to the Arctic region. However, little is known about the role of spatial factors in structuring such communities. In this study, we evaluated the influence of spatial scale on bacterial community structure across an Arctic landscape. Our results showed that spatial factors accounted for approximately 10% of the variation at the landscape scale, equivalent to observations across the whole Arctic region, suggesting that while the role and magnitude of other processes involved in community structure may vary, the role of dispersal may be stable globally in the region. We assessed dispersal limitation by identifying the spatial autocorrelation distance, standing at approximately 60 m, which would be required in order to obtain fully independent samples and may inform future sampling strategies in the region. Finally, indicator taxa with strong statistical correlations with environment variables were identified. However, we showed that these strong taxa-environment associations may not always be reflected in the geographical distribution of these taxa.

Importance

The significance of this study is threefold. It investigated the influence of spatial scale on the soil bacterial community composition across a typical Arctic landscape and demonstrated that conclusions reached when examining the influence of specific environmental variables on bacterial community composition are dependent upon the spatial scales over which they are investigated. This study identified a dispersal limitation (spatial autocorrelation) distance of approximately 60 m, required to obtain samples with fully independent bacterial communities, and therefore, should serve to inform future sampling strategies in the region.
and potentially elsewhere. The work also showed that strong taxa–environment statistical
associations may not be reflected in the observed landscape distribution of the indicator
taxa.

Introduction

Significant spatial structuring of soil microorganisms has been demonstrated at micro [µm -
mm] (1), plot [cm - to few meters] (1), landscape [hundreds of meters] (2), regional [kms]
(3), national (4, 5), continental (6), and global scales (7-9). Hence, the scale of investigation
is a key parameter to take into account in studies of bacterial biogeography. Martiny et al.
(10) further demonstrated the importance of spatial scale on environmental factors
identified influencing community composition in temperate soils. They found key
environmental drivers differed across spatial scales – ammonia-oxidizing bacterial (AOB)
community composition was dependent on distance, moisture and vegetation cover at the
plot scale; however, at the regional scale, diversity was mainly influenced by water
temperature, air temperature and moisture while nitrate concentration and air temperature
were predominant at the continental scale. Finally, when considering all scales together,
overall key drivers were geographic distance, sediment moisture, air temperature and
vegetation cover. However, most biogeographical studies only investigate communities at
one spatial scale (see Griffiths et al. (4), Tedersoo et al. (7), Bahram et al. (9) for further
examples). The landscape scale (few hundred of meters to few kilometers) is considered
highly relevant for studies of bacterial distribution patterns as it is the scale of human
activities (at which agricultural practices and land management are integrated). Hence, the
majority of studies at that scale investigate human-impacted landscapes (See Bru et al. (3),
Dao (11), Constancias et al. (2), Palta et al. (12) and Neupane et al. (13) for further
examples) with only few studies describing Arctic communities from few meters to 3 km (14-16).

The first aim of this study was to evaluate the influence of the spatial scale on bacterial community structure [Fig. S1] across an Arctic landscape [Fig 1]. Indeed, while the role of environmental parameters such as pH (17, 18), total organic content (TOC) (19), moisture content (20) and C:N ratio (21) on community composition in the Arctic has been demonstrated, much less is known about the influence of spatial parameters (19). However, determining the influence of environmental factors on communities remains an essential step to avoid overestimating the role of the spatial scale. In addition to providing a better understanding of the environmental factors influencing community structure, investigating multiple scales provides better knowledge of the spatial structure, which facilitates the development of sampling strategies where samples are collected beyond the spatial autocorrelation distance and are, therefore, truly independent (22). As autocorrelation distances have been identified from µm to km (22-25), with the potential of nested scales of variability (26), and site to site variation, no standardized protocol exists for soil sampling for metabarcoding studies (27, 28). Therefore, the second aim was to determine the minimum distance required to obtain independent soil samples in the region [Fig. S1], which may inform future sampling strategies in the Arctic. Finally, the last aim was to identify indicator taxa which were closely associated with environmental variables and map their spatial distribution across the landscape [Fig. S1]. Previous studies have attempted to identify indicator taxa that could be used for environmental monitoring (for example Simonin et al. (29) and Yang et al. (30) in rivers or Hermans et al. (31) in soils). As indicator taxa (32) highlight OTUs with strong environmental associations that may respond to ecological change, we expected their distribution to closely follow that of environmental parameters.
Results

Environmental factors

Results showed that all 35 environmental variables had a significant impact on bacterial community structure with approximately 73% of the variance explained by environmental factors [Table 1]. Overall, five key factors (TOC, pH, conductivity, aluminium and arsenic) had the most influence on bacterial community dissimilarity explaining 30% of variation in total. While all other environmental factors individually explained between 0.9% and 2.4% of the variation, the combined soil elemental composition (excluding pH, conductivity and TOC) accounted for 51.5% of the total variation in bacterial community composition.

Variation partitioning

A total of 9 dbMEMs vectors were built using (x,y) geographic coordinates and after forward selection, five dbMEMs were identified as significantly impacting bacterial community diversity and used in subsequent analyses. The variation partitioning analysis differentiated the effect of environmental factors, linear trend and spatial vectors on community composition [Fig. 2]. The environmental fraction X1 explained 73% of the variance [Table S1], equal to the finding by the adonis function and confirming the success of the variation partitioning analysis. Using the adjusted R² values only as they accounted for the number of variables in the model, environmental factors explained 54% of the variance, of which 38% were not spatially structured (fraction [a]). The spatial component (X2 + X3) explained 25.6% of the variation, of which 16.3% could be explained by induced spatial dependence. This was illustrated by fractions [d], [f] and [g], which represented spatially structured
environmental variables where the spatial structure of these environmental variables induced a similar spatial structure in the response data, highlighting the need to evaluate the influence of the environment on communities. The remaining 9.3% of the spatial component represented spatial autocorrelation. The linear trend accounted for 3.8% of the variance (fraction \[b\]) while spatial vectors explained 5.5% of the variation. Fraction \[e\] had a negative \(R^2\) and could be considered null, as prescribed in D. Borcard et al. (33). Each fraction (X1, X2, X3) was tested individually and was significant (ANOVA, \(p < 0.001\)). In total, 62.8% of the bacterial community dissimilarity could be explained by environmental and spatial factors while the remaining 37.2% of the variance could not be explained by the variables measured in this study.

**Spatial scale and autocorrelation**

The distance-decay curve illustrated the increase in community dissimilarity with increasing distance [Fig. 3A]. The power model was better fitted \((R^2 = 0.2261, \ p = 0.005)\) than the linear regression \((R^2 = 0.1844, \ p < 2.2 \times 10^{-16})\). Spatial autocorrelation was visualised on the distance-decay curve [Fig. 3], where geographically close communities were more similar up to 60 m. This was illustrated with the power model on the distance-decay curve, where the blue curve begins to plateau [Fig. 3A]. To further characterise the spatial autocorrelation distance, a Mantel correlogram was used [Fig. 3B] to compute the Mantel statistic between the geographic distance and bacterial community dissimilarity distance (Bray Curtis). The spatial autocorrelation was positive for the first distance class of 21 m, indicating that the bacterial communities were more similar than expected by chance. The second distance class of 63 m displayed no spatial autocorrelation, indicating random distribution beyond 63
m. Other distance classes presented negative autocorrelations indicating that these bacterial communities were more different than expected by chance.

Geography also had some influence on environmental conditions with sites closer together being more similar. The spatial autocorrelation of environmental variables was first visualised in figure 3C, where geographically close sites were geochemically similar within 25 m. However, beyond approximately 25 m, site equally close or far could present similar environmental conditions, as illustrated by the autocorrelation distance [Fig. 3C]. This was also illustrated by the weak linear regression ($R^2 = 0.019, p < 2.2 \times 10^{-16}$) and the best-fitted power model ($R^2 = 0.087, p = 0.005$). Spatial autocorrelation was further tested for each individual variable using the semi-variograms produced prior to kriging. As semi-variograms are specific to each variable, the spatial autocorrelation distances were unique to each parameter. All the semi-variograms produced prior to Kriging indicated positive autocorrelations oscillating between 1 m and 100 m, depending on the variable tested, further illustrating the importance of the scale of investigation [Fig. S2].

**Spatial distribution across the landscape**

Using an ordinary kriging method and after examining the semi-variograms, the spatial distribution of alpha diversity and key environmental variables were mapped across the landscape [Fig. 4]. The bacterial richness, diversity and evenness changed across the landscape [Fig. 4(A, B & C)], and kriged maps illustrated the relationships between diversity, evenness and richness. Overall, low richness indicated low diversity and low evenness, further observed using linear models [Fig. S3]. The kriged maps of alpha diversity and environmental variables showed the strong heterogeneity at the landscape scale with changes from high to low concentrations within just a few meters [Fig. 4(D, E & F)].
The indicator species analysis identified 163 true specialists (statistic >0.98) OTUs associated with 12 environmental variables. Indicator taxa were generally associated with the highest concentration of each element. The phylogenetic tree specific to indicator taxa illustrated the high taxonomic diversity of indicator taxa [Fig. 5], however, figure 6 demonstrated that identified indicator taxa do not necessarily follow environmental gradients as they are expected to. Of the four key factors (excluding pH) influencing bacterial communities [Table 1], only conductivity and arsenic had some indicator taxa associated. Indicators of conductivity (Cond) were restricted to two OTUs associated with high conductivity, both Bacteroidetes classified in the Cytophagales order [Fig. 5]. Peaks of high conductivity were visualised in figure 6A and correlated with peaks in abundance of the two OTUs identified [Fig. 6B, C]. Indicators of arsenic (As) were closely associated with barium (Ba) and were taxonomically diverse, with the majority classified as Actinobacteria, Alphaproteobacteria, Chloroflexi, Halanaerobiales and Firmicutes [Fig. 5]. Arsenic concentration appeared more homogeneous across the landscape [Fig. 6D] with an average concentration = 13 ppm, min = 1.81 ppm, max = 20.51 ppm. These indicator taxa of arsenic were all associated with high concentrations [Fig. 6E, F, G, H & I] and were also associated with high concentrations of barium in the soil. Iron (Fe) and manganese (Mn) are both essential elements of soils. Iron concentration was highly heterogeneous across the landscape, with a strong peak in concentration at one site [Fig. 6J]. This peak was reflected by the presence of unique indicator taxa of which the abundance was closely related to this high concentration [Fig. 6K, L]. Indicators of iron were diverse, with a large number of Proteobacteria (Alpha, Beta, Gamma), Chloroflexi, Bacteroidetes, Cyanobacteria, Planctomycetes and Verrucomicrobia [Fig. 5]. On the other hand, manganese concentration was heterogeneous across the
landscape [Fig. 6M] but unlike other indicator taxa, they were associated with low concentrations in the soil [Fig. 6N, O]. The indicator taxa of manganese were predominantly classified as Proteobacteria [Fig. 5] and were also closely related to low concentrations of niobium (Nb), lead (Pb) and zirconium (Zr), however, they were associated with high concentrations of molybdenum (Mo). Indicator taxa of strontium (Sr) were limited to five OTUs, an unknown Verrucomicrobi, a Ca. Saccharibacterium (TM7), a Deltaproteobacterium and two Alphaproteobacteria while indicators of zinc (Zn) were classified in all almost all phyla [Fig. 5], illustrating the wide array of specialist taxa associated with high concentrations of zinc.

Discussion

Key environmental factors influencing bacterial communities

Total organic carbon, pH and conductivity were identified as the key drivers of bacterial diversity across the Arctic landscape and are also commonly identified in studies across the globe (8, 34-38). While pH was previously identified as the primary driver of bacterial diversity in Arctic soils across the whole Arctic region (19); here, at the landscape scale, TOC was identified as the primary factor influencing bacterial community structure and was tightly linked with soil moisture. Generally, soil organic carbon content increases with increasing precipitation and decreasing temperature (39). In the Arctic tundra, not only precipitation but snowmelt and permafrost thaw have major impacts on soil moisture and hydrology across the landscape (40, 41). In this study, where pH was on average acidoneutral at 6.05 ± 0.36 with very few acidic patches, but organic carbon content was very patchy (6 % - 46 %); the role of TOC in bacterial community structure is perhaps not
surprising. However, it highlights the importance of investigating different spatial scales as drivers at the global scale may not necessarily be the same across the landscape of interest.

Aluminium and arsenic were the fourth and fifth environmental variables accounting for the most variation in bacterial community structure [Table 1]. Aluminium is one of the most abundant metal in the Earth’s crust and microorganisms continuously interact with aluminium in soils (42, 43). While aluminium lacks apparent biological function (42), the aluminium ion (Al\(^{3+}\)) can be toxic to living organisms and is a function of the soil pH; the concentration of toxic Al\(^{3+}\) gradually increases as pH decreases from pH = 6.2 (42, 43). Here, little pH changes but large aluminium concentration variation were observed across the landscape, which were not correlated to each other (linear regression: \(R^2 = 0.00069, p = 0.81\)). The toxicity of Al\(^{3+}\) may be influencing the bacterial community structure, however, the concentration of Al\(^{3+}\) ions was not measured.

Arsenic is ubiquitous in low abundance in the natural environment and recognised as one of the most toxic elements (44, 45). Here, a decrease in diversity and richness was observed with increasing arsenic concentrations, which likely reflects the toxic effect of oxyanions of arsenate on many bacteria, although some can use it as a terminal electron acceptor (44). As with Al\(^{3+}\), the chemical concentration of the various forms of arsenic was not measured and therefore, cannot conclude that the toxicity has an influence on bacterial community structure, although it is a possibility. Indicator taxa associated with high concentrations of arsenic were diverse but dominated by Actinobacteria and Proteobacteria and was in accordance with Dunivin et al. (45) who conducted a global survey of arsenic related genes in soils and identified these phyla as harbouring more arsenic resistance genes.
All other elements measured had some influence on the observed bacterial community [Table 1], from key major elements such as sulphur, calcium, silicon; to key trace elements such as iron, manganese, magnesium, zinc, copper, molybdenum and cadmium; and other elements, toxic or not, such as bromine, yttrium or lead. It should also be noted that while TOC, pH and conductivity had a significant influence on bacterial community composition (21.8%), the soil elemental composition combined explained most of the variation (> 50%). This may serve to highlight the level of complexity of the factors influencing community structure.

**Indicator taxa**

Environmental variables were highly heterogenous across the landscape, which was reflected by the distribution of alpha diversity and indicator taxa. The indicator species analysis determined abundant OTU-environment associations and identified 163 OTUs that could be considered true specialists in relation to 12 environmental variables. These OTUs were generally associated with high concentrations of the variable in question except for those associated with manganese, niobium, lead and zirconium which were representative of low concentrations. As illustrated in the phylogenetic tree [Fig. 5], the diversity of these indicator taxa was high, with numerous representatives of the Proteobacteria, Chloroflexi, Bacteroidetes, Planctomycetes and Verrucomicrobia. The distribution of some indicator taxa, selected for their reported relationship with the associated variable in the literature, was mapped across the landscape to illustrate the association with the elements’ concentration. For arsenic, Clostridium and Clostridia-related (Halanaerobiales) taxa were mapped as they have been identified with some role in arsenic cycling (44, 46) and with arsenic-resistance genes (45). A Gemmatimonadetes and a Candidatus *Parcubacterium*...
clustered closely with the Cyanobacteria) were also mapped, as both have been identified with potential roles in arsenic cycling (46). The distribution of OTUs associated with iron were mapped and included a Cyanobacterium (47, 48) and a Deltaproteobacterium, a class with known taxa involved in iron cycling (47-49). Finally, the OTUs associated with manganese were also associated with other environmental variables and mainly identified as Proteobacteria. A Deltaproteobacterium and the only Chlamydiae identified were mapped, two classes associated with manganese cycling (48). While this analysis showed the strong associations of some OTUs with the measured environmental parameters, it also illustrated the difficulty of using indicator taxa for monitoring purposes due to the large number of associations identified and the high heterogeneity across the landscape. This was clear when the distribution of key indicator taxa was mapped across the landscape and did not clearly follow the distribution of the environmental variable associated. Furthermore, while indicator taxa may be identified, they do not necessarily participate in the associated element cycle. For instance, these OTUs may benefit from high concentration of arsenic due to higher tolerance to toxicity and decreased competition, without having any involvement in arsenic cycling.

**Selection and dispersal structure bacterial communities**

The variation partitioning analysis quantified the importance of both selection (deterministic) and dispersal (stochastic) on bacterial community structure. Environmental variables explained 54% of the total variation, corresponding to selection and 16% were spatially structured, corresponding to the induced spatial dependence. Then, spatial components (trend + dbMEMs) alone explained 10% of the variation, illustrating spatial autocorrelation or dispersal (33). This is the same magnitude of influence as recorded in
Malard et al. (19) investigating biogeographical patterns across the whole Arctic region, suggesting that the magnitude of influence of dispersal of bacterial community structure may be stable in the Arctic.

More specifically, the distance-decay curve of environmental factors showed that edaphic properties were spatially autocorrelated up to approximately 25 m, although this was the overall spatial autocorrelation as each variable autocorrelated within different distances. After that distance, environmental variables were independent, and this was illustrated by the weak slope of the linear regression and the overall variability of edaphic properties. In addition, even highly similar environmental conditions could harbour dissimilar bacterial communities, further illustrating the potential role of dispersal and other processes such as drift or diversification. The distance-decay curve of bacterial communities showed a positive spatial autocorrelation distance at up to 60 m, which was further supported by the Mantel correlogram. For the Arctic region as a whole, an autocorrelation distance within the same order of magnitude, approximating 20 m, was previously identified (19). This limited dispersal range in Arctic soils is in contrast with studies in other regions of the globe. For instance, in a glacier forefield in southern Alaska, this distance was over 600 m (50) while in British soils, it was below 1 km (4). It suggests that Arctic soil bacterial communities only disperse to approximately 60 m and may form rather isolated island communities. Therefore, the scale of sampling is important in these landscapes to capture community variability and therefore, a minimum of 60 m should be maintained between sites to obtain independent samples. Further investigations at other Arctic sites are required to determine whether this applies across the whole Arctic region.
Overall, these results suggest that induced spatial dependence may be an important factor shaping bacterial communities within 25 m, that is, as edaphic properties are very similar, bacterial communities are also similar. Between 25 and 60 m, environmental variability increased and yet, communities remained relatively similar, suggesting that dispersal may be the primary process shaping bacterial communities. Beyond 60 m, the environment was highly heterogeneous, bacterial communities were highly dissimilar and selection was likely the main process structuring communities. While one process may dominate within each distance category, it is still likely the combination of different processes (selection, dispersal, diversification and drift) with different magnitudes still driving community assembly (51).

While 63% of the variation (non-adjusted $R^2 = 81\%$) of bacterial community assemblage could be explained, 37% remained unexplained. Many factors, whether biotic or abiotic could still be influencing bacterial communities. Based on the scale of this study, it is unlikely that most climatic and topographic variables would have much influence on the community structure variation. Instead, other edaphic factors such as total nitrogen or phosphorus content, clay, silt and sand content but also the presence of ice or soil texture may have more impact locally. Furthermore, biotic interactions such as competition and predation within bacterial communities or with other members of the soil biota or higher organisms may have some influence. For instance, grazing is one of the main disturbances to the ecosystem locally, primarily by the Svalbard reindeer and the barnacle goose (52). In addition to impacting the vegetation, they trample over the landscape and fertilise it and therefore, grazing can have significant impacts on the ecosystem and has been shown to decrease microbial respiration and the available carbon (53) while animal faeces increase the available nitrogen and can increase bacterial abundance (54). Human presence may also have some influence as the sampling site was close to another scientific research site with
open-top chambers, few cabins were located in the area, and the coal ‘Mine 7’ was still in
operation, approximately 1.5 km away and 400 m above the sampling site.

Conclusion

In this study, spatial factors accounted for approximately 10 % of the variation in community
composition at the landscape scale, equivalent to observations across the whole Arctic
region, suggesting that while the role and magnitude of other processes involved in
community structure may vary, the role of dispersal may be stable globally in the region.
Furthermore, the identification of different driving environmental factors at different scales
highlights their dependence upon the spatial scales over which they are investigated.
Overall, we suggest that induced spatial dependence may be shaping bacterial communities
within 25 m. Between 25 and 60 m, dispersal may be the primary process shaping bacterial
communities and beyond 60 m, selection is likely the main process structuring communities.
As dispersal may be limited to 60 m, and while further studies should be conducted, we
suggest that soil sampling in the region should be conducted beyond this distance to
capture landscape variability while collecting independent samples. Finally, by mapping the
spatial distribution of indicator taxa across the landscape, we showed that strong taxa-
environment statistical associations may not actually be reflected in the landscape
distribution of these bacterial taxa.

Material and Methods

Sampling site

In July 2017, 44 soil samples were collected in Adventdalen, Svalbard [Fig. 1A] following the
sampling design depicted in figure 1B and characterised by 8 North-South transects of 5
samples each. Samples within each transect were approximately 50 m apart while the distance between transects was approximately 100 m. On transect 6, extra samples were collected 10 m and 1 m apart to investigate smaller scale patterns [Fig. 1B, 1C].

Adventdalen is a broad U-shaped valley open to the West, from which the mouth is located approximately 2 km from Longyearbyen and 6 km from Svalbard Airport. Adventdalen was deglaciated about 10 ka BP (55) and permafrost is estimated to be 100 m thick close to the shore. It is a typical Arctic landscape, in one of the driest areas of Svalbard, with an average of 190 mm of annual precipitation, and mean annual temperature of -6 °C (56). The study site was located approximately 9 km into the valley, 11 km away from Longyearbyen, at 78.17 °N, 16.02 °E. The vegetation is primarily dwarf shrub/grass heath, dominated by Salix spp., mosses, lichens and Graminea spp. (57) [Fig. 1D]. The main disturbances to the site come from grazing, primarily by the Svalbard reindeer (Rangifer tarandus platyrhynchus) and the barnacle goose (Branta leucopsis) (52).

**Sample collection and soil properties**

The coordinates from each site were recorded with a portable GPS. At each location, 50 g of soil in the top 15 cm was collected using ethanol-cleaned trowels and Whirl-Pak bags (Nasco, Fort Atkinson, WI, USA). Plant roots and rocks were removed manually in a class II biological safety cabinet (ESCO, Singapore), samples were homogenised and frozen at -20 °C before transportation to the United Kingdom for analyses. pH and conductivity were measured in the laboratory in a 1:5 freshly thawed soil to water ratio, using a Mettler-Toledo FE20 pH meter (Mettler-Toledo Instruments co., Shanghai, China) and a CMD500 conductivity meter (WPA, Cambridge, UK). Moisture content was measured gravimetrically on soils after drying at 150 °C for 24 h and total organic content (TOC) was measured
gravimetrically by heating previously dried soils to 550 °C for 4 h. To analyse the elemental composition of each sample, 5 g of thawed soil was placed in ceramic crucibles and left to dry at 37 °C for 5 days. Dried samples were crushed to a fine powder using a mortar and pestle, put in a powder sample cup, placed in the XRF spectrometer (X-Lab2000, Spectro, Kleve, Germany) and analysed. Resulting concentrations were adjusted using calibrated values and results were expressed in part per million (ppm).

**DNA extraction and amplicon sequencing**

Soil DNA was extracted in duplicate for each sample using the PowerSoil kit (Qiagen, Hilden, Germany) following the manufacturers’ protocol. 16S rRNA gene libraries were constructed using the universal primers 515F-806R (58) to amplify the V4 region. Amplicons were generated using a high-fidelity Accuprime DNA polymerase (Invitrogen, Carlsbad, CA, USA), were purified using AMPure magnetic bead capture kit (Agencourt, Beckman Coulter, MA, USA) and quantified using a QuantIT PicoGreen fluorometric kit (Invitrogen). The purified amplicons were then pooled in equimolar concentrations using a SequalPrep plate normalization kit (Invitrogen), and the final concentration of the library was determined using a SYBR green quantitative PCR (qPCR) assay. Libraries were mixed with Illumina-generated PhiX control libraries and our own genomic libraries and denatured using fresh NaOH. The resulting amplicons were sequenced on the Illumina MiSeq V2, 500 cycles.

**Bioinformatics processing**

Raw paired-end reads were subjected to adaptor and primer clipping using Cutadapt (59) resulting in 71,207 ± 3,280 reads per sample. Forward and reverse reads were merged using FLASH (60). Reads with over 1.5 total expected errors per read and/or read length less than 245 base pairs were truncated during quality filtration using the Vsearch (61) filtering
module. It resulted in 64,917 ± 4,291 high quality merged reads per sample. Dereplication was performed to identify unique sequences. A two-step chimera detection method was used, first by aligning against ChimeraSlayer Gold database provided with SILVA (62), second by using the *denovo* detection module. An open-reference operational taxonomic unit (OTU) calling was performed on high-quality trimmed sequences at 97% similarity level using the USEARCH (63) algorithm for clustering to generate operational taxonomical units (OTUs). It resulted in (85 DNA samples) a total of 5,436,264 reads (63,956 ± 38,865 reads/sample) assigned against 23,627 OTUs. Unique (chimera filtered) representative sequences were aligned using the Python Nearest Alignment Space Termination (PyNAST) (64) tool with a relaxed neighbour-joining tree built using FastTree (65). OTUs were assigned taxonomy using the Lowest Common Ancestor (LCA) based Classification Resources for Environmental Sequence Tags (CREST) (66) with a minimum classification score of 0.80 against SILVA release 128 as a reference database.

**Data availability**

The sequencing dataset is deposited at the European Nucleotide Archive under the BioProject accession PRJNA564217.

**Statistical analysis**

Alpha diversity (richness, Shannon and Simpson indices) was calculated in QIIME v1.90 (67) on the matrices resulting from multiple rarefactions with the smallest sample size (22316 reads) as maximum depth. Results were collated and averaged to obtain a single representative value for each sample. The OTU table was normalised using cumulative-sum scaling (CSS) (68). The resulting OTU table was input into R for subsequent analyses and the Bray-Curtis dissimilarity distance was calculated using vegan (69).
To evaluate the environmental component, Pearson’s correlation coefficients were calculated using the corrplot package (70) to first identify possible correlations between environmental variables. With these many variables, it was a necessary step to avoid misinterpretation of the results (Katz, 2011). Coefficients over |0.8| indicated strong correlations [Fig. S4] and as such, variables were removed to keep only one representative (Katz, 2011). For example, a high moisture content was correlated with a high TOC content (Pearson’s = 0.88), in this case, moisture was discarded as it is weather-dependent and is expected to be more variable day to day than TOC. Of 48 parameters measured, 35 were independent and considered to be representative. The distribution of the 35 remaining environmental variables was investigated using the moments package (71) to assess the skewness and kurtosis. Skewness evaluates the degree of distribution shift to one side or another and a good distribution should equal 0, while kurtosis evaluates the tail distribution and should also be close to 0 to assume normal distribution. Using diagnostic plots, skewness and kurtosis, the necessary transformations to improve the unimodal distribution of environmental variables were carried (summarised in Table S2) and collinearity was verified again with Pearson’s correlations [Fig. S5]. Transformed environment variables were scaled and a sequential PERMANOVA was conducted using the adonis function implemented in vegan with standard 999 permutations to identify environmental variables significantly associated with the Bray-Curtis community dissimilarity.

To evaluate the spatial component, the geographic locations (x,y) of the sampling sites were transformed to cartesian coordinates using the SoDA package (72) and the Euclidean distance was calculated using vegan. Distance-decay curves were produced using linear regressions of the Euclidean distance of the geographic locations against the Bray-Curtis dissimilarity distance and the Euclidean distance of scaled environmental variables.
The presence of a linear trend (a systematic increase or decrease in the OTU data with (x,y) coordinates) was visualised by the distance-decay curve [Fig. 3A] and tested by RDA and ANOVA, as prescribed in Borcard et al. (33). As a significant linear trend was identified, the OTU table was detrended by linear regression of the (x,y) coordinates. Distance-based Moran’s Eigenvector Maps (dbMEM) were constructed with (x,y) coordinates using the adespatial R package (73). The significance of the spatial vectors (dbMEMs) was assessed using the detrended OTU table and tested with ANOVA. Forward selection was conducted to identify significant dbMEM vectors and the remaining dbMEMs were plotted using RDA.

Variation partitioning analysis (VPA) was used to assess the impact of environmental and spatial factors on community composition (undetrended OTU table) and was conducted using the environmental variables, (x,y) coordinates (linear trend) and significant dbMEM vectors. Individual fractions were tested using RDA and ANOVA, as prescribed in Borcard et al. (33).

To evaluate spatial autocorrelation, the detrended OTU table and the Euclidean distances of cartesian coordinates (x,y) were used to produce a Mantel correlogram with standard 999 permutations using vegan. Semi-variograms were also produced using the autoKrige function of the automap package (Hiemstra and Hiemstra, 2013) to use for geostatistical analyses. Kriging was conducted using the autoKrig and automapPlot functions in the automap package. Environmental variables and alpha diversity measures were interpolated and mapped across the landscape.

Indicator taxa were determined by the Dufrene-Legendre indicator species analysis (32) to identify OTUs that were specifically associated with different environmental variables. The first step was to define categories for each environmental variable (i.e. high conductivity,
medium conductivity and low conductivity). To identify groups statistically rather than subjectively, an automatic cluster approach was employed using the nbclust package (74), which indicated the ideal number of groups (Table S2). Clusters were created using the kmeans function (Table S2) and used with the multipatt function in the indicspecies package with 999 permutations (32). Indicator taxa with a correlation statistic higher than 0.98 were considered true specialists and used for subsequent analyses. The phylogenetic tree of indicator taxa was built using the representative sequences from the identified indicator taxa using FastTree method (65) and visualised using iTOL (75). Indicator taxa distribution was mapped across the landscape by kriging, as previously described and Pearson correlations were calculated between the indicator taxa and the environmental variables of interest.

Acknowledgements

This work was supported by a grant from the European Commission’s Marie Sklowdowska Curie Actions program under project number 675546. The authors also thank Edwin Sia and UNIS for their participation and support in fieldwork. MiSeq sequencing of the 16S rRNA gene was performed by the NU-OMICS sequencing service (Northumbria University).

LAM and DAP conceived and designed the study and sampling design. LAM carried the fieldwork and laboratory work. MZA conducted the bioinformatics processing and LAM conducted the statistical analysis. LAM drafted the manuscript and MZA, DAP and CSJ revised and approved the final version.

Conflict of interest

The authors report no conflict of interests.
References


Siles JA, Margesin R. 2016. Abundance and diversity of bacterial, archaeal, and fungal communities along an altitudinal gradient in alpine forest soils: what are the driving factors? Microbial ecology 72:207-220.


Table 1: The relative influence of environmental factors on bacterial community structure, calculated by PERMANOVA using the adonis function. * 0.05 > p > 0.01, ** 0.01 > p > 0.001, *** p < 0.001.

<table>
<thead>
<tr>
<th>Variable</th>
<th>R²</th>
<th>Pr(&gt;F)</th>
<th>Variable</th>
<th>R²</th>
<th>Pr(&gt;F)</th>
<th>Variable</th>
<th>R²</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOC</td>
<td>0.089</td>
<td>0.001***</td>
<td>Sr</td>
<td>0.018</td>
<td>0.002**</td>
<td>Th</td>
<td>0.013</td>
<td>0.005**</td>
</tr>
<tr>
<td>pH</td>
<td>0.070</td>
<td>0.001***</td>
<td>S</td>
<td>0.016</td>
<td>0.001***</td>
<td>Ag</td>
<td>0.012</td>
<td>0.007**</td>
</tr>
<tr>
<td>Cond</td>
<td>0.059</td>
<td>0.001***</td>
<td>Cu</td>
<td>0.015</td>
<td>0.001***</td>
<td>Mo</td>
<td>0.012</td>
<td>0.013*</td>
</tr>
<tr>
<td>Al</td>
<td>0.041</td>
<td>0.001***</td>
<td>Te</td>
<td>0.015</td>
<td>0.002**</td>
<td>Sb</td>
<td>0.012</td>
<td>0.010**</td>
</tr>
<tr>
<td>As</td>
<td>0.041</td>
<td>0.001***</td>
<td>Ba</td>
<td>0.014</td>
<td>0.003**</td>
<td>Cd</td>
<td>0.011</td>
<td>0.023*</td>
</tr>
<tr>
<td>Br</td>
<td>0.024</td>
<td>0.001***</td>
<td>In</td>
<td>0.014</td>
<td>0.002**</td>
<td>Ta</td>
<td>0.011</td>
<td>0.016*</td>
</tr>
<tr>
<td>La</td>
<td>0.022</td>
<td>0.001***</td>
<td>Nb</td>
<td>0.014</td>
<td>0.004**</td>
<td>Tl</td>
<td>0.011</td>
<td>0.021*</td>
</tr>
<tr>
<td>Y</td>
<td>0.021</td>
<td>0.002**</td>
<td>Nd</td>
<td>0.014</td>
<td>0.008**</td>
<td>Zr</td>
<td>0.011</td>
<td>0.012*</td>
</tr>
<tr>
<td>Ca</td>
<td>0.018</td>
<td>0.003**</td>
<td>Si</td>
<td>0.014</td>
<td>0.004**</td>
<td>Zn</td>
<td>0.010</td>
<td>0.031*</td>
</tr>
<tr>
<td>Cl</td>
<td>0.018</td>
<td>0.001***</td>
<td>Fe</td>
<td>0.013</td>
<td>0.002**</td>
<td>Ge</td>
<td>0.009</td>
<td>0.046*</td>
</tr>
<tr>
<td>Cs</td>
<td>0.018</td>
<td>0.001***</td>
<td>I</td>
<td>0.013</td>
<td>0.006**</td>
<td>Sn</td>
<td>0.009</td>
<td>0.036*</td>
</tr>
<tr>
<td>Pb</td>
<td>0.018</td>
<td>0.001***</td>
<td>Mn</td>
<td>0.013</td>
<td>0.009**</td>
<td>Residuals</td>
<td>0.269</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Figure 1: Map of sampling sites in (A) Svalbard. (B) Sampling design in 8 transects in Adventdalen. (C) Smaller scale samples on transect 6. (D) View of Adventdalen.
Figure 2: Venn diagram illustrating the results of the variation partitioning analysis on the influence of environmental variables and spatial factors on bacterial community composition. Results of each partition can be multiplied by 100 for the percentage of variation explained and are detailed in table S2.
Figure 3: (A) Distance-decay curve illustrating the increase in bacterial community dissimilarity with increasing geographic distance. (B) Mantel correlogram of spatial autocorrelation illustrating the dispersal limitation. Red squares indicate positive significant autocorrelation which was only identified in the first distance class (0-21 m). Beyond 60 m, the autocorrelation was either negative (black squares) or not significant (white squares). (C) Distance-decay curve illustrating increasing environmental variation with increasing geographic distance. The red curve illustrates the linear regression and the blue curve is the power model.
Figure 4: Kriged maps of the spatial distribution across the landscape showing the heterogeneity of (A) Richness, (B) Shannon index, (C) Simpson index, (D) pH, (E) Total organic carbon and (F) Aluminium. The color bar of A, B, C indicates values of alpha diversity while the color bar of environmental variables indicates element concentrations (see units of each variable in table S2, taking into account data transformations).
Figure 5: Phylogenetic tree of indicator taxa associated with environmental variables showing the high phylogenetic diversity. Coloured bands illustrate the taxonomy of each
OTU at the phylum level; labels indicate the taxonomy down to the family level if available.

Coloured points indicate the element associated.

Environmental variables
Trend
Spatial factors (MEMs)

[a] 0.379
[d] 0.045
[b] 0.038

[g] 0.072
[f] 0.046
[e] 0.055
[c] Residuals = 0.372

Values <0 not shown
<table>
<thead>
<tr>
<th>Classification</th>
<th>Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verrucomicrobia</td>
<td>As</td>
</tr>
<tr>
<td>Planctomycetes</td>
<td>Ba</td>
</tr>
<tr>
<td>WA</td>
<td>Pb</td>
</tr>
<tr>
<td>Acidobacteria</td>
<td>Cond</td>
</tr>
<tr>
<td>Elusimicrobia</td>
<td>Sr</td>
</tr>
<tr>
<td>Chlamydiae</td>
<td>Fe</td>
</tr>
<tr>
<td>WS3</td>
<td>Y</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>Mn</td>
</tr>
<tr>
<td>TM7</td>
<td>Zn</td>
</tr>
<tr>
<td>Chlorobi</td>
<td>Mo</td>
</tr>
<tr>
<td>Terrabacteria</td>
<td>Zr</td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>As</td>
</tr>
<tr>
<td>Peregrinibacteria</td>
<td>Ba</td>
</tr>
<tr>
<td>TM6</td>
<td>Pb</td>
</tr>
<tr>
<td>Halanaerobiales</td>
<td>Cond</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>Sr</td>
</tr>
<tr>
<td>Deltaproteobacteria</td>
<td>Y</td>
</tr>
<tr>
<td>Gammaproteobacteria</td>
<td>Mn</td>
</tr>
<tr>
<td>Unclassified</td>
<td>Zn</td>
</tr>
</tbody>
</table>