Rojas-Jimenez Keilor (Orcid ID: 0000-0003-4261-0010)
Grossart Hans-Peter (Orcid ID: 0000-0002-9141-0325)

Variation of bacterial communities along the vertical gradient in Lake Issyk Kul, Kyrgyzstan

Keilor Rojas-Jimenez[1], Alex Araya-Lobo[1], Fabio Quesada-Perez[1], Johana Akerman-Sanchez[1], Brayan Delgado-Duran[1], Lars Ganzert[2,3,4], Peter O. Zavialov[5], Salmor Alymkulov[6], Georgiy Kirillin[7], Hans Peter Grossart[2,8*]

[1]Escuela de Biología, Universidad de Costa Rica, 11501 San José, Costa Rica.

[2]Department of Experimental Limnology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Alte Fischerhuette 2, D-16775 Stechlin, Germany.

[3]GFZ German Research Centre for Geosciences, Helmholtz Centre Potsdam, Section 3.7 Geomicrobiology, Telegrafenberg C-422, 14473, Potsdam, Germany.

[4]University of Goettingen, Experimental Phycology and Culture Collection of Algae (EPSAG), Nikolausberger Weg 18, 37073 Goettingen, Germany.

[5]Shirshov Institute of Oceanology, Moscow, 117997, Russia.

[6]Institute of Physics, National Academy of Sciences of Kyrgyz Republic, Chui Avenue, 265-a, 720071, Bishkek, Kyrgyzstan

[7]Department of Ecohydrology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Müggelseedamm 310, D- 12587 Berlin, Germany.

[8]Institute for Biochemistry and Biology, Potsdam University, Maulbeerallee 2, 14469 Potsdam, Germany.

*Corresponding author
Hans-Peter Grossart, hgrossart@igb-berlin.de

Key words: Microbial ecology, deep oligotrophic lakes, brackish waters, Planctomycetes, Chloroflexi

Summary

We explored the diversity and community composition of bacteria along a vertical gradient in Lake Issyk Kul, Kyrgyzstan, one of the world's largest and deepest brackish lakes. We identified 4904 bacterial amplicon sequence variants (ASVs) based on the 16S rRNA gene analysis and determined significant changes in the composition, responding mainly to depth and salinity. A higher abundance of Proteobacteria and Bacteroidetes was observed in the surface waters and lake tributaries. Cyanobacteria were more abundant in the deep chlorophyll maximum (DCM) from 28.5 to 128 m, while Planctomycetes and Chloroflexi were dominant in the deepest layers, from 128 to 600 m. According to our machine learning analyses, depth and temperature were the most critical environmental factors, with strong effects on Proteobacteria, Planctomycetes, and Chloroflexi, while oxygen was associated with the variations in Cyanobacteria. We also observed that with increasing depth, the alpha diversity values increased. The dominance of Planctomycetes and Chloroflexi in the deepest layers can only be seen in a few lakes of the world. However, the lake is facing increasing anthropogenic and climatic pressure. There is an urgent need to understand better the ecological role and function of these unique deep-water microbial communities.

Introduction

The current knowledge about the distribution of planktonic bacterial communities along the vertical gradient in lake ecosystems is limited, particularly in deep lakes. Of the approximately 117 million lakes with surface areas greater than 0.002 km^2 that exist globally, only 36 have mesopelagic ecosystems with maximum depths greater than 400 m (Verpoorter *et al.*, 2014; Dodds and Whiles, 2020). Of these, only a handful has been studied concerning diversity and composition of their microbial communities, with examples mainly from Lake Baikal in Russia (Glöckner *et al.*, 2000; Bel'kova *et al.*, 2003; Kurilkina *et al.*, 2016), Lake Tanganyika in Africa (De Wever *et al.*, 2005), and Crater Lake in the United States (Urbach *et al.*, 2001).

Water depth is one of the main factors influencing bacterial communities' composition and distribution in both oceans and lakes. Along the vertical gradient, there is a transition from the photic to the aphotic zone, which also includes a decrease in temperature, oxygen, and nutrients, concomitant with an increase in hydrostatic pressure (DeLong *et al.*, 2006; Bryant *et al.*, 2012; Garcia *et al.*, 2013; Mende *et al.*, 2017; Wu *et al.*, 2019). Salinity is another major factor that influences the composition of microbial communities in aquatic

ecosystems in both vertical and horizontal gradients (Herlemann *et al.*, 2011; Newton *et al.*, 2011; Tang *et al.*, 2012, 2015; Rojas-Jimenez *et al.*, 2019).

Lake Issyk Kul is located in the northeast of Kyrgyzstan and has the singularity of being one of the largest and deepest lakes in Central Asia and worldwide. This lake arouses great scientific interest because of its geophysical and ecological characteristics. Being the sixth deepest lake of the world and tenth-largest by volume, it reveals surprisingly high rates (below 11 years) of deep-water renewal (Hofer *et al.*, 2002; Peeters *et al.*, 2003), which maintain the lake water column well-oxygenated down to its deepest parts (Zavialov *et al.*, 2018). Despite the absence of any outflows, the lake water's salinity of 6 g kg^{-1} is relatively low, suggesting the lake to become endorheic relatively recently, possibly due to tectonic processes (Romanovsky, 2002).

Lake Issyk Kul is (ultra)-oligotrophic with highly transparent water, reaching Secchi depths of more than 20 m. The low nutrients content in the surface waters and the high amount of solar radiation penetrating deep into the water column creates a unique environment characterized by a maximum of photosynthesis at depths of 30-40 m. First systematic studies on Lake Issyk Kul date back to the 1920s. It was one of the most studied lakes in the 1960-80s, but research stopped abruptly after the Soviet Union's collapse. During the 1960-80s, the lake ecosystem experienced a substantial impact by human activities, including the introduction of non-native commercial fish species displacing the endemic species, the input of wastewater and pollution by industry and agriculture, with effects on the aquatic ecosystem that remain unknown (Savvaitova and Petr, 1992; Giralt *et al.*, 2004; Baetov, 2005). Yet, almost nothing is known about microbial communities, particularly heterotrophic bacteria at the base of the aquatic food web.

In this work, we used high-throughput sequencing of the V4-V5 region of the 16S rRNA gene to provide a more in-depth insight into the taxonomic composition and vertical distribution of bacterial communities inhabiting Lake Issyk Kul (**Figure 1**). We complemented the analysis with the use of Machine Learning tools, which are helpful to reveal unexpected biological patterns, discover ecological relationships between microorganisms and their environment, and formulate new biological hypotheses (Eraslan *et al.*, 2019; Qu *et al.*, 2019). Considering that this is a unique deep-water brackish lake, exhibiting aerobic conditions throughout the entire depth gradient, we particularly aimed to determine how community composition varies along the depth gradient and which particular

environmental variables may be responsible for those changes. We hypothesize that depth and its related environmental variables such as sunlight accessibility, hydrostatic pressure, temperature, and salinity will lead to the formation of different layers, each one dominated by specific bacterial communities.
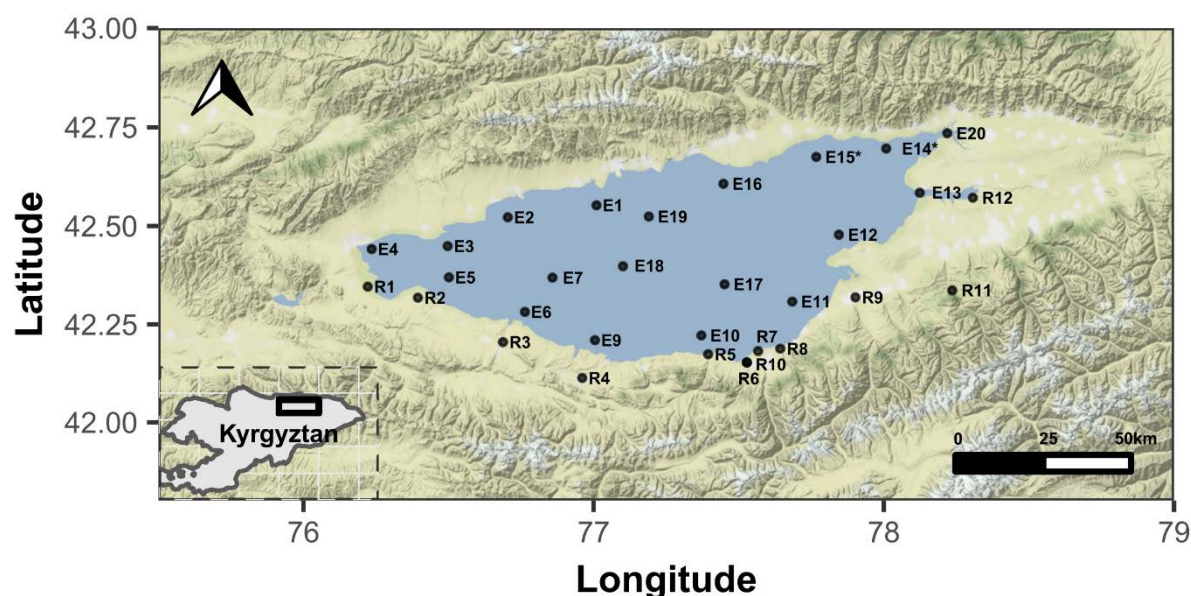
Figure 1. Location of sampling points in Lake Issyk Kul, Kyrgyzstan, and some of its tributary rivers. We set 19 sampling points in the lake (indicated with the letter E) and 12 tributaries (indicated with the letter R). In total, we analyzed 43 water samples from the lake and 12 from the tributaries. The locations marked with an asterisk (E14 and E15) were sampled but were discarded during the data analysis stage due to low sequencing depth.

Results

We observed pronounced changes in the composition of the bacterial communities of Lake Issyk Kul along the vertical gradient, responding mainly to the variables depth and salinity (**Figure 2**). In total, 4904 bacterial ASVs were identified, whereby the phylum Proteobacteria was the most abundant, representing, on average, 28.5% of all sequences. In lower average proportions, other groups were also detected, such as Bacteroidetes (16.8%), Cyanobacteria (14.8%), Planctomycetes (12.85%), Actinobacteria (9.8%), Verrucomicrobia (8.7%), and Chloroflexi (6.7%) (**Suppl. Fig. 2**).
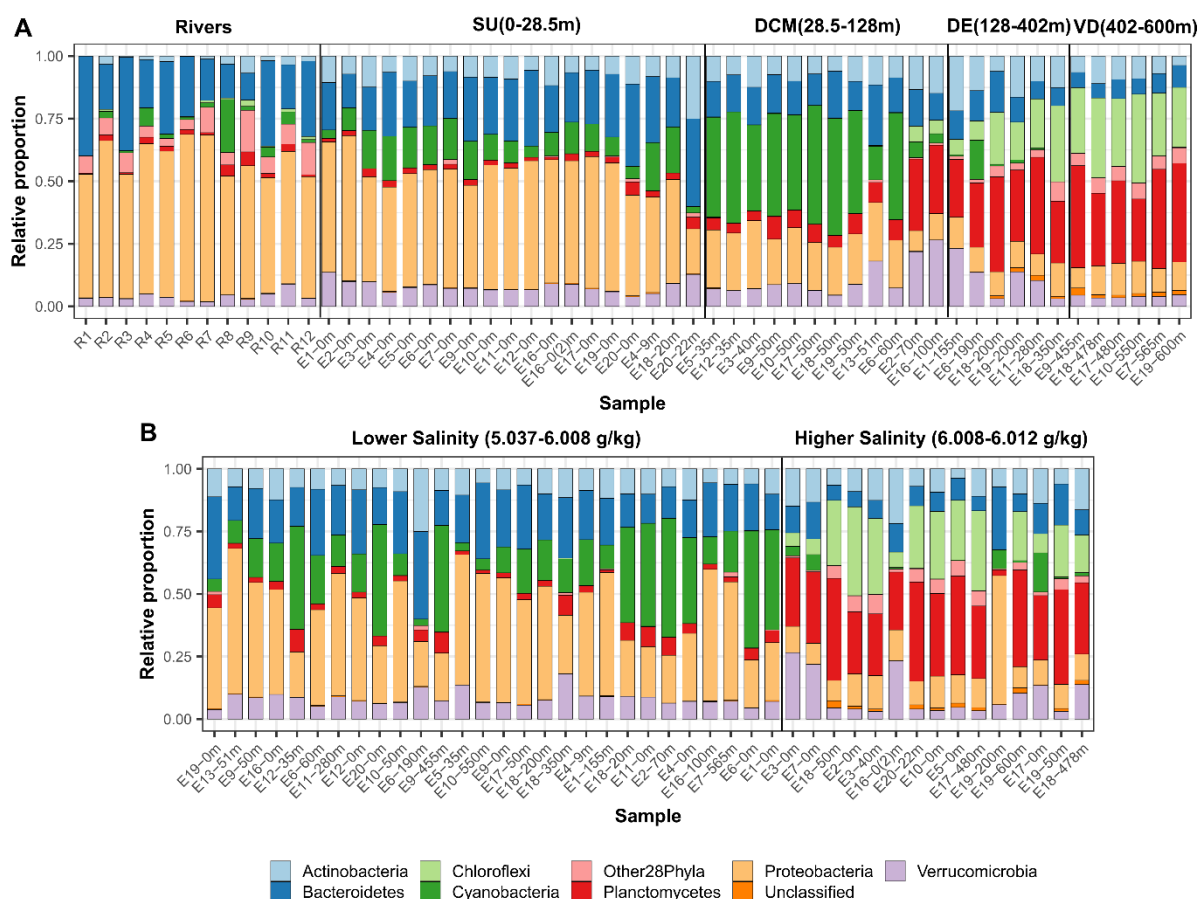
Figure 2. The relative abundance of bacteria, determined at the phylum's taxonomic level, in water samples of Lake Issyk Kul, Kyrgyzstan. Panel **A** shows the results grouped by depth clusters, while panel **B** shows them sorted by salinity. SU: surface, DCM: deep chlorophyll maximum, DE: deep layer, VD: very deep layer.

Differences in bacterial communities according to depth

We used the K-means algorithm to create clusters of samples according to depths, distinguishing four layers: 0 to 28.5 m, 28.5 to 128 m, 128 to 402 m, 402 to 600 m. Another cluster included the tributary rivers. This grouping was consistent with differences observed in their bacterial composition (**Figure 2A**). At the phylum level, a higher abundance of Proteobacteria and Bacteroidetes was found in the lake's surface waters and the river samples. In the deep chlorophyll maximum (DCM, between 28.5 to 128 m), there was a higher abundance of Cyanobacteria, while Planctomycetes and Chloroflexi became more abundant at greater depth, i.e., between 128 to 600 m.

At the ASV level, a strain associated with the genus *Sphingomonas* (Alphaproteobacteria) was the most abundant in the rivers. *Cyanobium* (Cyanobacteria), *Loktanella* (Alphaproteobacteria), and an unknown strain of Burkholderiaceae (recently reclassified within Gammaproteobacteria) were the most abundant in the surface waters of the lake. *Cyanobium* also dominated in the DCM between 28.5 to 128m. Three strains related to families Anaerolineaceae (Chloroflexi) as well as Phycisphaeraceae and Gimesiaceae (Planctomycetes) were the most abundant ASVs in the deep layers (128 to 402 m and 402 to 600 m) although with small differences between each layer.

The observed variation in the composition of bacterial communities along the depth gradient was consistent with the NMDS analysis that showed a clear separation of the clusters (**Figure 3A**). These results were also compatible with the Permanova analysis, showing statistically significant differences between the rivers and the lake's bacterial communities, and even within each of the four depth clusters identified (**Suppl. Table 2**).

Differences in bacterial communities according to salinity

We also observed differences in the composition of bacterial communities, according to salinity (**Figure 2B**). In superficial layers, salinity ranged between 5.037 and 6.008 g/kg, where a higher abundance of Proteobacteria, Cyanobacteria, and Bacteroidetes was observed. The river samples were not included in this analysis because the information of salinity was not recorded. Salinity was more stable in the deeper layers, between 6.008 and 6.012 g/kg, where a higher relative abundance of Planctomycetes and Chloroflexi was observed. These differences in community composition were consistent with the clustering of the NMDS analysis (**Figure 3B**) and the Permanova analysis, showing significant differences in the composition between these two salinity layers (**Suppl. Table 2**). Regarding other environmental parameters (such as oxygen or temperature), we did not find any significant effect on the lake bacterial community composition.
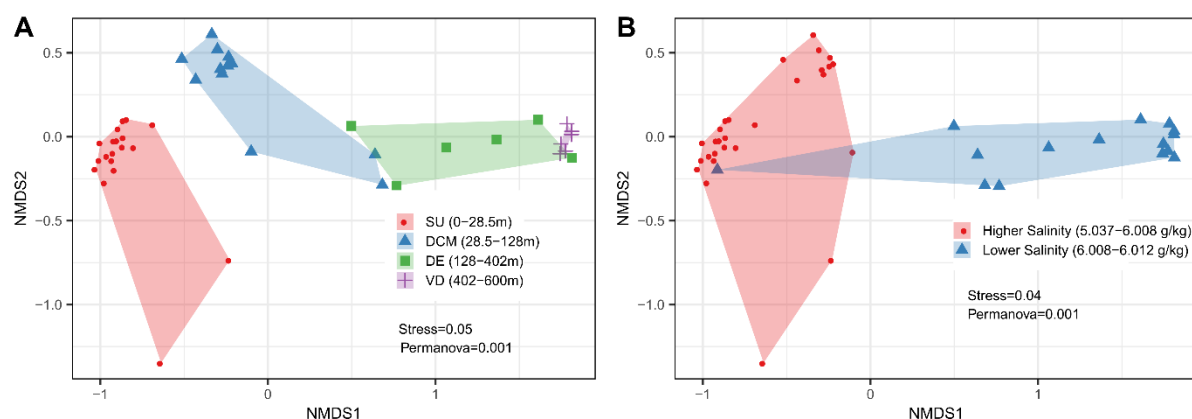
Figure 3. Non-metric multidimensional scaling analyses of the bacterial communities in Lake Issyk Kul. Panel **A** shows the clusters by depth, while panel **B** for salinity. Stress values and p-value of the Permanova are shown. SU: surface, DCM: deep chlorophyll maximum, DE: deep layer, VD: very deep layer.

Effects of environmental variables on alpha diversity values

As a complementary method, we used machine learning (Random Forest algorithm) to identify environmental variables with more weight in the observed variation of the relative abundance of the bacterial phyla (**Figure 4**). According to this analysis, depth and temperature are the environmental factors that have the most substantial effect on the variations in Proteobacteria, Planctomycetes, and Chloroflexi. The higher abundance of Cyanobacteria was associated with high oxygen conditions, while for Actinobacteria and Verrucomicrobia, no specific variables were distinguished as being responsible for the variations in their abundance.
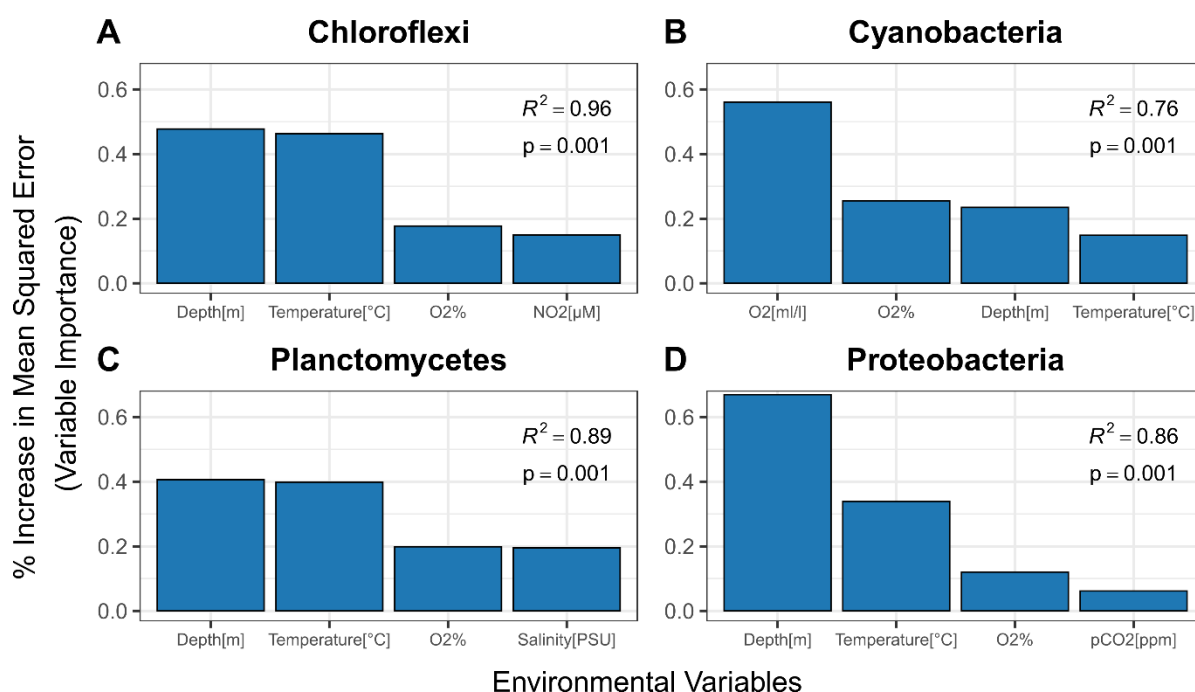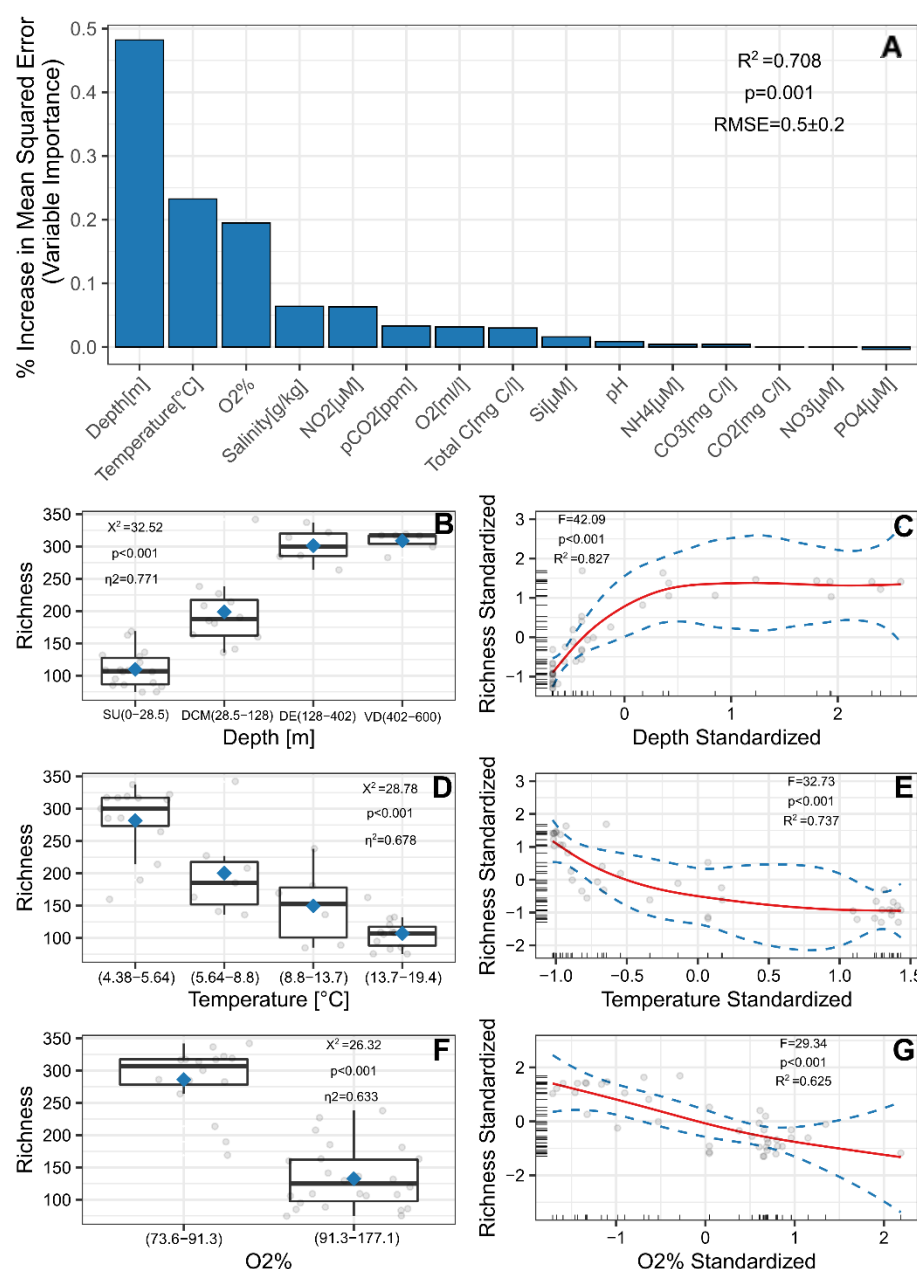
Figure 4. Main environmental factors associated with variations in the abundance of four phyla of bacteria in Lake Issyk Kul. The analyses were performed using Random Forest and considering the most abundant bacterial groups per depth layer. Results are shown for Chloroflexi (**A**), Cyanobacteria (**B**), Planctomycetes (**C**), and Proteobacteria (**D**).

The richness, estimated at the taxonomic level of family, ranged between 75 and 342 per sample. The factor that best explained this variability was depth (Random Forest; p = 0.001, R^2 = 0.708,, **Figure 5A**). We observed an increase in richness with increasing depth, ranging from 110 ± 7 bacterial families in surface waters to 309 ± 6 in the deepest layers (**Figure 5B**). The variation in richness between layers was statistically significant (Kruskal-Wallis, H=35.52, p <0.001), but according to the GAM, there was a stabilization in the number of families at 280 m (GAM; R^2 = 0.827, F = 41.89, p <0.001, **Figure 5C**). At lower temperatures, consistent with greater depths, we also observed an increase in richness (**Figure 5D** and **Figure 5E**).

The percentage of oxygen (O_2 %) was a factor that also explained the variability in the richness (Random Forest; p = 0.001, R^2 = 0.403, **Figure 5F**). Although Lake Issyk Kul presents aerobic conditions along the vertical gradient (**Supplementary Figure 3**), higher richness values were determined in the layers where oxygen was lower (73.6-91.3%). In comparison, values were lower in the superficial and DCM layers, where oxygen content is

higher (91.3-177.1%). Therefore, we observed a slight decrease in richness as the oxygen level increases (GAM; $R^2 = 0.625$, F = 29.34, p <0.001, **Figure 5G**).

Figure 5. Main environmental variables associated with variations in the richness values in Lake Issyk Kul. Analyses were performed at the taxonomic level of family using Random Forest. Panel **A** shows the importance of variables. Panels **B, D and , F** show boxplots of richness estimations according to bins (value intervals) of the leading environmental factors (Depth, Temperature and O2%, respectively). The standardized values considering the GAM

model predicted relationships are shown in panels **C, E and G**. SU: surface, DCM: deep chlorophyll maximum, DE: deep layer, VD: very deep layer.

Discussion

The main finding was that depth plays a decisive role in the variations of bacterial communities in this deep lake. Four layers were distinguished along the depth gradient—each layer with a set of particular environmental conditions and a community structure significantly different from each other. In general, these results are consistent with those obtained in other deep lakes such as Lake Baikal in Russia, Crater Lake in the United States, and Lake Tanganyika in Central Africa, which also show the differentiation of bacterial communities along the depth profile, although each of the bacterial communities in the respective lakes was responding to different physicochemical conditions (Glöckner *et al.*, 2000; Urbach *et al.*, 2001; Bel'kova *et al.*, 2003; De Wever *et al.*, 2005; Kurilkina *et al.*, 2016). However, Lake Issyk Kul showed differences in the composition of the most abundant groups in the different depth layers, which are probably related to this lake's particular conditions, such as having brackish waters and high oxygenation.

The composition of the bacterial communities in the tributaries was significantly different from that found in the lake's different layers, including the surface layer. For example, *Sphingomonas* was the most abundant genus in the rivers while *Loktanella* and an unknown Burkholderiaceae were the most abundant in the lake's surface layer. Also, Cyanobacteria, Planctomycetes, and Verrucomicrobia were almost missing in these mountain rivers characterized by high turbulence and often high sediment load, as previously reported (Wilhelm *et al.*, 2013; Peter and Sommaruga, 2016). Although this work was not designed to determine the role of tributaries in providing seed communities for the lake, the results showed clear differences between these two ecosystems. It will be important to determine further how the lake water layers' unique conditions, including the residence time, the availability of nutrients, and the light regime, exert a filtering effect for bacterial taxa's lateral inputs.

With the increase in depth and slightly more saline conditions, we observed a decrease in the abundance of Proteobacteria, which is consistent with previous studies

showing the effects of these environmental variables on the abundance of this phylum (Simon *et al.*, 1999; Cottrell and Kirchman, 2003; Kirchman *et al.*, 2005; Kan *et al.*, 2008). The DCM was dominated by picoplanktonic unicellular cyanobacteria, particularly of the genus *Cyanobium*, as previously reported in other oligotrophic lakes in Europe (Padisák *et al.*, 2003, 2004; Callieri, 2008; Selmeczy *et al.*, 2016). A DCM at ~50 m depth, which is significantly more profound than in any reported lake, might result from extremely high subsurface radiation in parallel to the high water transparency and low nutrients, which is concurrent with a maximum concentration of oxygen and chlorophyll-a (**Suppl. Fig. 3**).

We show the predominance of Anaerolineaceae (Chloroflexi) and two families of Planctomycetes (Phycisphaeraceae and Gimesiaceae) in the deeper layers of a lake. Other studies have also reported the presence (but not dominance) of these groups in deep waters (Urbach *et al.*, 2001; De Wever *et al.*, 2005; Breuker *et al.*, 2013; Kurilkina *et al.*, 2016; Li *et al.*, 2018). Anaerolineaceae has been described primarily as a group of anaerobic bacteria, but the description of this family was based mainly on mesophilic bacteria from a thermophilic reactor (Yamada *et al.*, 2006). Some members of this group have also been reported in marine sediments (Breuker *et al.*, 2013). Therefore, it is possible, as shown here, that members of this family can most likely be aerobic. This is in accordance of previous studies which have have identified this group (formerly classified within the green non-sulfur bacteria) in aerobic deep waters in Crater Lake and Lake Baikal (Urbach *et al.*, 2001; Kurilkina *et al.*, 2016).

It is necessary to study in more detail the physiological adaptations of microorganisms in the deep layers of inland aquatic ecosystems similar to Issyk Kul (e.g., Lake Titicaca and the Caspian Sea), which withstand aphotic conditions, downward salinity increase, low temperatures, high hydrostatic pressure, but the increased availability of oxygen. Groups inhabiting this ecosystem, such as Anaerolineaceae, should present very particular physiological and ecological characteristics that deserve more attention (McIlroy *et al.*, 2017). It has also been suggested that physiological adaptations to depth could be accompanied by changes at the genomic level, including larger genome size, higher genomic GC content, and proteins with higher nitrogen but lower carbon content (Mende *et al.*, 2017).

Besides, it would be necessary to determine whether a hydrostatic pressure at 600 m in Lake Issyk Kul would be sufficient to trigger some proposed adaptation mechanisms such as increasing the proportion of unsaturated fatty acids in membrane phospholipids (Delong and Yayanos, 1985; DeLong and Yayanos, 1986; Tamburini *et al.*, 2013; Wannicke *et al.*,

2015), systems for active transport of sugars (DeLong and Yayanos, 1987) and an increase in bacterial protein production (Somero, 1992; Wannicke *et al.*, 2015). The occurrence of these above mentioned adaptation mechanisms may be somewhat theoretical; however, it deserves further studies considering that the highest alpha diversity values were found in the ecosystem's deepest part.

In conclusion, in this work, we demonstrate the effect of depth and salinity on the variations in structure of bacterial communities along the vertical gradient in Lake Issyk Kul, which is consistent with previous studies (Koizumi *et al.*, 2003; Bryant *et al.*, 2012; Mestre *et al.*, 2017; Wu *et al.*, 2019). We highlight a high bacterial diversity in the lake's deepest layers, where microorganisms live in brackish waters, aphotic conditions, low temperatures, high hydrostatic pressure, but completely aerobic conditions. Chloroflexi and Planctomycetes' dominance in the deeper layers raises questions about their ecological and functional roles, which should be further explored taking the increasing anthropogenic and climatic pressures into account the lake is facing currently and in the future.

Experimental Procedures

Environmental variables and sampling

The sampling and environmental measurements were carried out in Lake Issyk Kul, a mountain lake with an average altitude of 1,609 meters above sea level and a maximum depth of 702 m. This lake occupies a closed basin of tectonic origin with a total area of 6,280 km² and is considered the second largest mountain lake in the world after Lake Titicaca in Bolivia. The lake expedition was carried out during the cruise of the RV Moltur from June 25th to July 1st, 2017. All sampling stations in the lake and its tributaries are shown in **Figure 1**. In total, 12 tributaries in the Southern and Eastern parts of the lake were sampled in triplicates at the well-mixed water column's surface. Triplicate samples from each tributary (n=12) were pooled, and then a subsample of 0.5-2 l depending on the suspended matter concentration was filtered for DNA extraction (see below). We set 17 sampling points in the lake (one to five different depths each; **Suppl. Table 1**) and, in total, analyzed 43 water samples. Sampling and measurement of physical and chemical variables were performed only from lake water samples using methods previously described (Zavialov *et al.*, 2018). Briefly, water samples were taken using HydroBios 5 L Niskin bottles at all stations at different depth levels from the surface to the bottom. Depth was measured using a digital sounder (model LMS-350, Lowrance). Profiles of temperature, electric conductivity, and fluorescence were

taken with the CTD-probe SeaBird 19plus (Sea-Bird Scientific, USA). Oxygen profiles were taken with the fast-response oxygen optode Rinko-I (JFE Advantech, Japan). Salinity was calculated from temperature, electrical conductivity, and pressure using the empirical formula for the Lake Issyk Kul ionic composition proposed by Peeters et al. (2003). Chemical determinations were performed in the laboratory using standard techniques; Winkler method for oxygen; the potentiometric method for pH; titration method for alkalinity; colorimetric assay method for phosphates, silicates, nitrates, nitrites, and carbon (CO_2, CO_3 (carbon trioxide), pCO_2) in different carbonate equilibria (**Suppl. Fig. 1).**

Molecular analyses

 Samples were immediately filtered on board after retrieval of the Niskin bottle. The water samples were filtered through 0.22 µm Sterivex® filters (EMD Millipore, Germany) connected to a peristaltic pump (EMD Millipore, Germany) to concentrate bacteria and subsequently stored at -20 °C. DNA was extracted from 0.22 µm Sterivex filters using the QIAamp DNA mini kit (Qiagen, Germany) following tissue protocol with some modifications. Briefly, the filters were cut into pieces and put into a 2 mL tube. A mix of zirconium beads and 360 µL of buffer ATL was added and vortexed for 5 min at 3,000 rpm in an Eppendorf MixMate® (Eppendorf, Germany). Proteinase K (>600 mAU/ml, 40 µL) was added and incubated at 57 °C for 1 h. After centrifugation for 1 min at 11,000 rpm, the supernatant was transferred to a new 2 mL tube, and extraction was performed following the manufacturer's protocol. PCR, library preparation, and sequencing were done by LGC Genomics (Berlin, Germany). Briefly, the V4-V5 region was amplified using primers 515F-Y/926R (Parada *et al.*, 2016), followed by library preparation (2×300 bp) and sequencing on a MiSeq Illumina platform. Sequences were quality checked and analyzed using version 1.12 of the DADA2 pipeline (Callahan *et al.*, 2016). This process resulted in an amplicon sequence variant (ASV) table, a higher-resolution analog of the traditional OTU table, which records the number of times each exact amplicon sequence variant was observed in a sample. The taxonomy assignment was performed by comparing sequences against the SILVA reference database v132 (Quast *et al.*, 2013). The NCBI Genbank BLAST tool was also used to provide higher taxonomic resolution to some abundant ASVs poorly classified with SILVA. Global singletons were removed. Samples with low sequencing depth (< 3800 sequences, n = 3) were also discarded from the analysis. After these processes, we obtained 1,790,675 sequences and 4904 bacterial ASVs. The average number of sequences per sample

was 24470 ranging from 3435 to 82992. The sequence data were deposited in GenBank (https://ww w.ncbi.nlm.nih.gov/genbank/) under the SRA accession numbers SAMN14816440- SAMN14816494.

Statistical analysis

 The statistical analyses and their visualization were performed in the R Project statistical program (R-Core-Team, 2019), using the Rstudio interface v3.6.0. To better determine the effects of continuous variables on bacterial populations, we performed the binning of these variables. Binning is a data processing technique where the original values are grouped into intervals. The *n* observations into *k* clusters are done using the k-means algorithm, an unsupervised machine learning method. Each observation belongs to the cluster with the nearest mean. We used package dlookr v0.3.13 (Ryu, 2020) to perform the binning of variable depth and a semisupervised method for variable salinity, based on changes in the relative abundance cyanobacteria. The NDMS and Permanova analyses were generated with package Vegan v2.5-6 (Oksanen *et al.*, 2020). For this purpose, the ASV table was transformed into a table of relative abundances with the function *decostand*. The NDMS was represented in a two-dimensional plot based on a Bray–Curtis similarity matrix. The statistical significance of the effects of the different variables on the bacterial community composition was calculated using the *adonis2* function with 999 permutations. The p-values of the pairwise comparisons were estimated with the *pairwise.adonis* function and ajusted with the Bonferroni method.

We applied random forest models to identify the critical environmental parameters affecting the relative abundance of the seven most abundant bacterial phyla utilizing package randomForest v 4.6-14 (Liaw and Wiener, 2002). The degree of importance of each predictor variable was determined by quantifying the effect of its removal on the model's accuracy across 10001 trees, measured by the percentage increase in mean squared error. The model significance was tested with 1000 permutations using package rfUtilities v 2.1-5 (Evans and Murphy, 2019). We used repeated k-fold cross-validation to estimate the model performance using the package caret v 6.0-86 (Kuhn, 2020). This cross-validation method was selected for its low bias and high reliability (Burman, 1989; Borra and Di Ciaccio, 2010; Kuhn and Johnson, 2013). Briefly, it randomly divides the data into ten blocks of roughly equal size, and the model is trained ten times, each time leaving out one of the subsets from the training,

which is used to compute the prediction error. The whole process is repeated with different random splits, and the predictions are summarized into the root mean squared error (RMSE).

Random forest models were applied to identify the critical environmental factors affecting the community's richness (number of trees=10001). To assess model fit, repeated 10-fold cross-validation and permutation tests were also used. Then, the relationships between these environmental variables and bacterial family richness were explored with generalized additive models (GAM), using packages mgcv v 1.8-31 and mgcViz v 0.1.6 (Wood and Wood, 2015; Fasiolo *et al.*, 2019). Kruskal-Wallis tests by rank and Mann-Whitney tests were used for estimating statistical differences within the intervals of each variable.

Author contributions

Conceptualization: HPG, GK, KRJ. Performing the research: LG, POZ, SA, GK, HPG. Contributing new tools or methods: KRJ, AAL, FQP, JAS, BDD. Analyzing data: KRJ, AAL, FQP, JAS, BDD, LG. Draft: KRJ, HPG, AAL, FQP, JAS, BDD. Review & Editing: all authors. All authors gave final approval for publication.

Competing financial interests: The authors declare no competing financial interests.

References

Baetov, R. (2005) Lake Issyk-Kul. *Manag. Lakes their Basins Sustain. Futur.* 193–204.

Bel'kova, N.L., Parfenova, V. V, Kostornova, T.Y., Denisova, L.Y., and Zaichikov, E.F. (2003) Microbial biodiversity in the water of Lake Baikal. *Microbiology* **72**: 203–213.

Borra, S. and Di Ciaccio, A. (2010) Measuring the prediction error. A comparison of cross-validation,

bootstrap and covariance penalty methods. *Comput. Stat. Data Anal.* **54**: 2976–2989.

Breuker, A., Stadler, S., and Schippers, A. (2013) Microbial community analysis of deeply buried marine sediments of the New Jersey shallow shelf (IODP Expedition 313). *FEMS Microbiol. Ecol.* **85**: 578–592.

Bryant, J.A., Stewart, F.J., Eppley, J.M., and Delong, E.F. (2012) Microbial community phylogenetic and trait diversity declines with depth in a marine oxygen minimum zone. *Ecology* **93**: 1659–1673.

Burman, P. (1989) A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika* **76**: 503.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**: 581–583.

Callieri, C. (2008) Picophytoplankton in Freshwater Ecosystems: The Importance of Small-Sized Phototrophs. *Freshw. Rev.* **1**: 1–28.

Cottrell, M.T. and Kirchman, D.L. (2003) Contribution of major bacterial groups to bacterial biomass production (thymidine and leucine incorporation) in the Delaware estuary. *Limnol. Oceanogr.* **48**: 168–178.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U., et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science (80-.).* **311**: 496–503.

Delong, E.F. and Yayanos, A.A. (1985) Adaptation of the membrane lipids of a deep-sea bacterium to changes in hydrostatic pressure. *Science (80-.).* **228**: 1101–1103.

DeLong, E.F. and Yayanos, A.A. (1986) Biochemical function and ecological significance of novel bacterial lipids in deep-sea procaryotes. *Appl. Environ. Microbiol.* **51**: 730–737.

DeLong, E.F. and Yayanos, A.A. (1987) Properties of the glucose transport system in some deep-sea bacteria. *Appl. Environ. Microbiol.* **53**: 527–532.

Dodds, W.K. and Whiles, M.R. (2020) Chapter 7 - Lakes and Reservoirs: Physiography. In, Dodds,W.K. and Whiles,M.R.B.T.-F.E. (Third E. (eds), *Aquatic Ecology*. Academic Press, pp. 155–187.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F.J. (2019) Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**: 389–403.

Evans, J.S. and Murphy, M.A. (2019) rfUtilities.

Fasiolo, M., Nedellec, R., Goude, Y., and Wood, S.N. (2019) Scalable Visualization Methods for Modern Generalized Additive Models. *J. Comput. Graph. Stat.* **0**: 1–9.

Garcia, S.L., Salka, I., Grossart, H., and Warnecke, F. (2013) Depth-discrete profiles of bacterial communities reveal pronounced spatio-temporal dynamics related to lake stratification. *Environ. Microbiol. Rep.* **5**: 549–555.

Giralt, S., Julià, R., Klerkx, J., Riera, S., Leroy, S., Buchaca, T., et al. (2004) 1,000-year environmental history of Lake Issyk-Kul. In, *Dying and Dead Seas Climatic Versus Anthropic Causes*. Springer, pp. 253–285.

Glöckner, F.O., Zaichikov, E., Belkova, N., Denissova, L., Pernthaler, J., Pernthaler, A., and Amann, R. (2000) Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Appl. Environ. Microbiol.* **66**: 5053–5065.

Herlemann, D.P.R.R., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J.J., and Andersson, A.F. (2011) Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J* **5**: 1571–1579.

Hofer, M., Peeters, F., Aeschbach-Hertig, W., Brennwald, M., Holocher, J., Livingstone, D.M., et al. (2002) Rapid deep-water renewal in Lake Issyk-Kul (Kyrgyzstan) indicated by transient tracers. *Limnol. Oceanogr.* **47**: 1210–1216.

Kan, J., Evans, S.E., Chen, F., and Suzuki, M.T. (2008) Novel estuarine bacterioplankton in rRNA operon libraries from the Chesapeake Bay. *Aquat. Microb. Ecol.* **51**: 55–66.

Kirchman, D.L., Dittel, A.I., Malmstrom, R.R., and Cottrell, M.T. (2005) Biogeography of major bacterial groups in the Delaware Estuary. *Limnol. Oceanogr.* **50**: 1697–1706.

Koizumi, Y., Kojima, H., and Fukui, M. (2003) Characterization of depth-related microbial community structure in lake sediment by denaturing gradient gel electrophoresis of amplified 16S rDNA and reversely transcribed 16S rRNA fragments. *FEMS Microbiol. Ecol.* **46**: 147–157.

Kuhn, M. (2020) caret: Classification and Regression Training.

Kuhn, M. and Johnson, K. (2013) Applied Predictive Modeling Springer New York, New York, NY.

Kurilkina, M.I., Zakharova, Y.R., Galachyants, Y.P., Petrova, D.P., Bukin, Y.S., Domysheva, V.M., et al. (2016) Bacterial community composition in the water column of the deepest freshwater Lake Baikal as determined by next-generation sequencing. *FEMS Microbiol. Ecol.* **92**:.

Li, X., Yuan, Yan, Yuan, Yi, Bi, Z., Liu, X., Huang, Y., et al. (2018) Effects of salinity on the denitrification efficiency and community structure of a combined partial nitritation- anaerobic ammonium oxidation process. *Bioresour. Technol.* **249**: 550–556.

Liaw, A. and Wiener, M. (2002) Classification and Regression by randomForest. *R News* **2**: 18–22.

McIlroy, S.J., Kirkegaard, R.H., Dueholm, M.S., Fernando, E., Karst, S.M., Albertsen, M., and Nielsen, P.H. (2017) Culture-Independent Analyses Reveal Novel Anaerolineaceae as Abundant Primary Fermenters in Anaerobic Digesters Treating Waste Activated Sludge . *Front. Microbiol.* **8**: 1134.

Mende, D.R., Bryant, J.A., Aylward, F.O., Eppley, J.M., Nielsen, T., Karl, D.M., and DeLong, E.F. (2017) Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nat. Microbiol.* **2**: 1367.

Mestre, M., Ferrera, I., Borrull, E., Ortega-Retuerta, E., Mbedi, S., Grossart, H.P., et al. (2017) Spatial

variability of marine bacterial and archaeal communities along the particulate matter continuum. *Mol. Ecol.* **26**: 6827–6840.

Newton, R.J., Jones, S.E., Eiler, A., McMahon, K.D., and Bertilsson, S. (2011) A Guide to the Natural History of Freshwater Lake Bacteria.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2020) vegan: Community Ecology Package (R package version 2.5–6).

Padisák, J., Barbosa, F.A.R., Koschel, R., and Krienitz, L. (2003) Deep layer cyanoprokaryota maxima in temperate and tropical lakes. *Arch. Für Hydrobiol. Beih. Adv. Limnol.* **58**: 175–199.

Padisák, J., Scheffler, W., Koschel, R., and Krienitz, L. (2004) Seasonal patterns and interannual variability of phytoplankton in Lake Stechlin. *Berichte des IGB* **2004**: 105–116.

Parada, A.E., Needham, D.M., and Fuhrman, J.A. (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.* **18**: 1403–1414.

Peeters, F., Finger, D., Hofer, M., Brennwald, M., Livingstone, D.M., and Kipfer, R. (2003) Deep-water renewal in Lake Issyk-Kul driven by differential cooling. *Limnol. Oceanogr.* **48**: 1419–1431.

Peter, H. and Sommaruga, R. (2016) Shifts in diversity and function of lake bacterial communities upon glacier retreat. *ISME J.* **10**: 1545–1554.

Qu, K., Guo, F., Liu, X., Lin, Y., and Zou, Q. (2019) Application of Machine Learning in Microbiology. *Front. Microbiol.* **10**: 827.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**: D590–D596.

R-Core-Team (2019) R: A language and environment for statistical computing.

Rojas-Jimenez, K., Rieck, A., Wurzbacher, C., Jürgens, K., Labrenz, M., and Grossart, H.-P. (2019) A Salinity Threshold Separating Fungal Communities in the Baltic Sea. *Front. Microbiol.* **10**:.

Romanovsky, V. V (2002) Water level variations and water balance of Lake Issyk-Kul. In, *Lake Issyk-Kul: its natural environment*. Springer, pp. 45–57.

Ryu, C. (2020) dlookr: Tools for Data Diagnosis, Exploration, Transformation.

Savvaitova, K. and Petr, T. (1992) Lake Issyk-kul, Kirgizia. *Int. J. Salt Lake Res.* **1**: 21–46.

Selmeczy, G.B., Tapolczai, K., Casper, P., Krienitz, L., and Padisák, J. (2016) Spatial-and niche segregation of DCM-forming cyanobacteria in Lake Stechlin (Germany). *Hydrobiologia* **764**: 229–240.

Simon, M., Glöckner, F.O., and Amann, R. (1999) Different community structure and temperature optima of heterotrophic picoplankton in various regions of the Southern Ocean. *Aquat. Microb. Ecol.* **18**: 275–284.

Somero, G. (1992) Adaptations To High Hydrostatic Pressure. *Annu. Rev. Physiol.* **54**: 557–577.

Tamburini, C., Boutrif, M., Garel, M., Colwell, R.R., and Deming, J.W. (2013) Prokaryotic responses to hydrostatic pressure in the ocean - a review. *Environ. Microbiol.* **15**: 1262–1274.

Tang, X., Xie, G., Shao, K., Dai, J., Chen, Y., Xu, Q., and Gao, G. (2015) Bacterial community composition in oligosaline lake Bosten: Low overlap of betaproteobacteria and bacteroidetes with freshwater ecosystems. *Microbes Environ.* ME14177.

Tang, X., Xie, G., Shao, K., Sai·Bayartu, Chen, Y., and Gao, G. (2012) Influence of Salinity on the Bacterial Community Composition in Lake Bosten, a Large Oligosaline Lake in Arid Northwestern China. *Appl. Environ. Microbiol.* **78**: 4748 LP – 4751.

Urbach, E., Vergin, K.L., Young, L., Morse, A., Larson, G.L., and Giovannoni, S.J. (2001) Unusual bacterioplankton community structure in ultra-oligotrophic Crater Lake. *Limnol. Oceanogr.* **46**: 557–572.

Verpoorter, C., Kutser, T., Seekell, D.A., and Tranvik, L.J. (2014) A global inventory of lakes based on high-resolution satellite imagery. *Geophys. Res. Lett.* **41**: 6396–6402.

Wannicke, N., Frindte, K., Gust, G., Liskow, I., Wacker, A., Meyer, A., and Grossart, H.P. (2015) Measuring bacterial activity and community composition at high hydrostatic pressure using a novel experimental approach: A pilot study. *FEMS Microbiol. Ecol.* **91**:.

De Wever, A., Muylaert, K., Van der Gucht, K., Pirlot, S., Cocquyt, C., Descy, J.-P., et al. (2005) Bacterial community composition in Lake Tanganyika: vertical and horizontal heterogeneity. *Appl. Environ. Microbiol.* **71**: 5029–5037.

Wilhelm, L., Singer, G.A., Fasching, C., Battin, T.J., and Besemer, K. (2013) Microbial biodiversity in glacier-fed streams. *ISME J.* **7**: 1651–1660.

Wood, S. and Wood, M.S. (2015) Package 'mgcv.' *R Packag. version* **1**: 29.

Wu, K., Zhao, W., Wang, Q., Yang, X., Zhu, L., Shen, J., et al. (2019) The Relative Abundance of Benthic Bacterial Phyla Along a Water-Depth Gradient in a Plateau Lake: Physical, Chemical, and Biotic Drivers. *Front. Microbiol.* **10**:.

Yamada, T., Sekiguchi, Y., Hanada, S., Imachi, H., Ohashi, A., Harada, H., and Kamagata, Y. (2006) Anaerolinea thermolimosa sp. nov., Levilinea saccharolytica gen. nov., sp. nov. and Leptolinea tardivitalis gen. nov., sp. nov., novel filamentous anaerobes, and description of the new classes Anaerolineae classis nov. and Caldilineae classis nov. in the . *Int. J. Syst. Evol. Microbiol.* **56**: 1331–1340.

Zavialov, P.O., Izhitskiy, A.S., Kirillin, G.B., Khan, V.M., Konovalov, B. V, Makkaveev, P.N., et al. (2018) New profiling and mooring records help to assess variability of Lake Issyk-Kul and reveal unknown features of its thermohaline structure. *Hydrol. Earth Syst. Sci.* **22**: 6279–6295.