Environmental drivers define contrasting microbial habitats, diversity and functional redundancy in Lake Baikal, Siberia

Paul Wilburn[1,2,3], Kirill Shchapov[4], Edward C. Theriot[5], Elena Litchman[1,2,3]

[1]W.K. Kellogg Biological Station, Michigan State University, USA
[2]Department of Integrative Biology, Michigan State University, USA
[3]Ecology, Evolutionary Biology, and Behavior Graduate Program, Michigan State University, USA
[4]Large Lakes Observatory, University of Minnesota, Duluth, USA
[5]University of Texas at Austin, USA

ABSTRACT (250 words)

Understanding how microbial communities respond to environmental change requires knowledge of the main drivers of their community structure, diversity and potential resilience. For many rapidly changing ecosystems this information is still not available. Lake Baikal in Siberia is the most ancient, deep, voluminous, and biologically diverse lake in the world, with 20 percent of global unfrozen fresh water, that is undergoing rapid warming. Little is known about its bacterioplankton communities and their drivers. In the first extensive survey of Baikal's microbial communities, we show that temperature, stratification, nutrients, and dissolved oxygen, and not the geographic distance, define major microbial habitats and microbial community similarity. Communities in ML and DW exhibited contrasting patterns of richness, diversity and evenness and comprised different cohesive modules in the whole Baikal OTU co-occurrence network. The network exhibited small-world properties that may make it resistant to perturbations but sensitive to changes in the abundances of central, most connected OTUs. Functional redundancy, often associated with higher resilience, was low in cold, open water communities and increased with temperature in the mixed layer and with depth in the deep-water samples. Our results suggest that bacterial communities of open waters in Lake Baikal may be

more sensitive to warming and other anthropogenic stressors than the littoral communities and may reorganize significantly in the changing climate. A better understanding of factors structuring bacterial communities in ecosystems that undergo rapid changes, including Lake Baikal and other northern lakes, will allow us to better predict the overall ecosystem responses to anthropogenic stressors.

SIGNIFICANCE (120 words)

This study is the first to identify distinct bacterial assemblages in Lake Baikal during summer stratification and link them to specific environments. Multiple linear regression and model averaging show that community diversity is driven by richness in the mixed layer (ML) and evenness in the deep waters (DW), and we identify environmental covariates that explain these contrasting trends. Network analyses reveal assemblages specific to ML and DW, where phylogeny reflects preferred environments. We then use PICRUSt to predict metagenomes and reveal that redundancy is lowest in the coolest areas of ML, potentially placing them at greatest risk of microbial functional diversity loss. This is a significant step towards understanding and addressing challenges faced by microbial communities in a changing world.

INTRODUCTION

The ecological importance of microorganisms in aquatic systems has been recognized at least since the appearance of "ooze" in Lindeman's trophic energy transfer diagram (1). Their central place in material and energy fluxes is now recognized for nearly all nutrient cycles, with greater relative importance of prokaryotic organisms in more oligotrophic systems (2). The advent of next-generation sequencing, combined with environmental monitoring, enabled new discoveries

52 of microbial diversity and function in various aquatic habitats. However, the environmental
53 drivers of microbial community diversity, community structure, function, and stability remain
54 poorly characterized in many aquatic ecosystems, including the world's most ancient (25 My)
55 lake Baikal – a UNESCO heritage site and known hotspot for endemism of its biota. Baikal is the
56 world's deepest (1643 m) and most voluminous lake, holding about 20% of world's surface
57 unfrozen freshwater (3). Of the approximately 2600 plant and animal species in the lake, two-
58 thirds are endemic, including the dominant primary producers, grazers, benthic and pelagic fish
59 and the top predator – world's only freshwater seal (3, 4).

60 Hampton *et al.* (4) showed that water temperatures have risen by 1.2 °C over 60 years of
61 high-resolution time series, contributing to an increase in numbers and kinds of non-endemic
62 zooplankton and algal species, with potential consequences for nutrient cycling, food web
63 structure (3) and microbial communities. Moreover, the Lake Baikal region is predicted to warm
64 by 3-4°C in the next century (5), with ongoing changes likely to continue and even accelerate.
65 Because the biota of Lake Baikal, including microbial communities, is adapted to cold
66 temperatures, it may be especially vulnerable to warming. Additionally, other changing
67 environmental factors may also alter the lake's microbial communities.

68 The vulnerability of communities to environmental change depends in part on their
69 functional redundancy (FR), linked to the number of species that perform a similar function (6).
70 High number of functionally similar species, i.e., high FR, is usually thought of as a mechanism
71 that preserves community functions during disturbance. High bacterial diversity, rapid generation
72 times, and the mobility of bacterial genes through genome rearrangement and horizontal gene
73 transfer enables relatively quick functional loss or gain within a phylogenetic lineage, and,
74 conversely, sharing among phylogenetic lineages. Indeed, prokaryotic functional genes are

75 promiscuous when it comes to species boundaries and are often found in multiple taxa adapted to
76 similar environmental conditions (7), resulting in functional redundancy. The degree to which
77 Lake Baikal microbial communities exhibit such redundancy, and how it may vary over the
78 multiple environmental and geographic gradients is unknown.

79 Here we present the first comprehensive survey and analysis of microbial plankton in Baikal,
80 spanning all three basins, from open waters to the shallow bays, multiple depths, including the
81 surface layer and depths below 300 m. We reveal major community composition trends that
82 correlate with continuous environmental gradients in a spatial context. We then identify strongest
83 environmental covariates to community composition in a multivariate framework. We use co-
84 occurrence network analyses to reveal clusters of OTUs with contrasting environmental
85 associations and identify clades and taxa most responsible for maintaining the structure of each
86 network cluster. Finally, we use the functional prediction tool PICRUSt to estimate functional
87 redundancy and show that it correlates with temperature and is lowest in open Baikal and in
88 communities resident in its hypolimnion.

89

90 **MATERIALS AND METHODS**

91 Our survey was guided by the recorded natural history of Baikal (3, 8, 9) that divides the lake
92 into eight distinct regions (Fig. 1a, S1). Among them, are the open Baikal basins, Chivirkuy Bay,
93 Proval Bay, and Selenga river plume. Chivirkuy Bay was sampled extensively to capture
94 transition from the shallow innermost bay (9 m depth) to open waters. Proval Bay and Selenga
95 plume stations represented the two most eutrophic areas in Baikal. In total, we collected samples
96 from 24 stations, of which 10 were sampled at various depths for a total of 46 samples.
97 Temperature and dissolved oxygen profiles were measured with a YSI Instruments sonde (YSI,

98 Inc., Yellow Springs, OH, USA). Whole water was taken using a Van Dorn sampler (Wildco,
99 Inc., Yulee, Florida, USA), and 5 L were filtered onto 3 μm and 0.22 μm mixed nitrocellulose
100 acetate membranes (EMD Millipore, Billerica, MA, USA) and stored at −20°C in RNAlater
101 (Life Technologies, Grand Island, NY, USA) to capture particle-attached (3 μm filter) and free-
102 living (0.22 μm filter) fractions of microorganisms. Genomic DNA was extracted using the Mo-
103 Bio PowerSoil Kit (Mo-Bio Laboratories, Carlsbad, CA), following manufacturer's protocol.
104 Then, the V4 region of the 16S rRNA gene was sequenced on a MiSeq platform (250PE), as
105 described previously (10). Raw sequences and metadata needed to reproduce our results are
106 available in GenBank under BioProject PRJNA530071.
107 Sequences were processed generally following the closed-reference mothur pipeline.
108 Statistical analyses were performed in the R (3.2.2) environment, unless noted otherwise.
109 Multiple regression search and model selection were done with the glmulti package (11).
110 Distance matrices, ordinations, and correlations with environmental variables we calculated with
111 vegan (12) and phyloseq (13) packages. We constructed the co-occurrence network using
112 sparCC (14), following recommended best practices by Berry and Widder (15). Next, we
113 identified network modules with the optimum modularity algorithm in the iGraph package (16).
114 We followed the WGCNA package (17) to generate eigenvectors for each module and correlate
115 the first eigenvector (first principal component, PC1) with environmental variables to generate
116 plots and a heatmap (Fig. 5). For the functional redundancy measure, the per sample number of
117 OTUs that could be characterized by the genome prediction software PICRUSt (18) was divided
118 by the gene content, estimated using PICRUSt software, with the nearest sequenced taxon index
119 (NSTI) cutoff at 0.15, as recommended by the package authors, to select accurate gene copy
120 number predictions. Next, we validated the predictions using an in-house shotgun dataset

121 available for a subset of samples. For genes that had greater than 10X mean shotgun coverage
122 (4012 genes), PICRUSt predictions had a significant correlation ($r^2 = 0.35$, $p=0.017$) with the
123 metagenomic community characterization, which was consistent with values reported in the
124 original publication (18). Detailed explanation of bioinformatics and statistical analyses is in
125 the Supplement.
126

127 **RESULTS AND DISCUSSION**

128 **Taxonomic diversity.** At a depth of 28059 sequences per sample, we detected 38457 OTUs in
129 the combined 3.0 μm and the 0.22 μm fractions. Non-singletons (6099 in 3.0 μm and 3346 in
130 0.22 μm) were classified into 61 phyla, dominated by typical freshwater Actinobacteria,
131 Bacteroidetes, Cyanobacteria, Proteobacteria and Verrucomicrobia (Fig. 1).
132 Microbial community composition notably differed from previous studies of Lake Baikal,
133 which explored pelagic communities on Sanger (19, 20) and Roche 454 (21, 22) platforms. In the
134 most recent effort, Kurilkina et al. (22) sampled one station depth profile in September and June
135 to reveal the main taxonomic groups as Proteobacteria, Actinobacteria, Bacteroidetes,
136 Firmicutes, Chloroflexi, Acidobacteria, and Cyanobacteria in both seasons, and Caldiserica in
137 September only. Our study did not detect Caldiserica and revealed only a very minor presence of
138 Chloroflexi, while showing a substantial Verrucomicrobia presence (absent in Kurilkina *et al.*) in
139 every sampled region (Fig. 1).
140 Our study revealed compositional differences between the 0.22 μm (free-living) and 3 μm
141 (mostly particle-attached) fractions. Actinobacteria were significantly enriched on the 0.22 μm
142 fraction (Table S1), similar to results from lakes in Michigan, USA (23). Proteobacteria were
143 only marginally more prevalent in the 0.22 μm fraction. We also found enrichment of

Bacteroidetes and Cyanobacteria on the 3 µm fraction, the latter owing to filamentous taxa. Indeed, the top three most differentially abundant taxa identified with permutation-based analyses (24) were all classified as the Nostocales genus *Dolichospermum* and contributed up to 90% of Cyanobacteria in the 3 µm fraction.

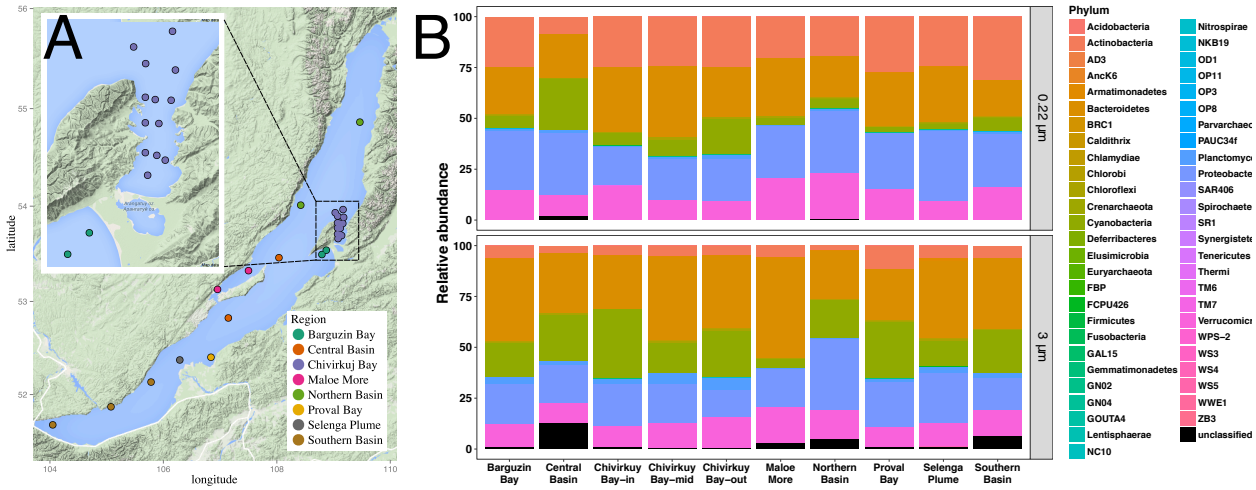


Figure 1: (A) Baikal sampling locations. Stations in outside of Chivyrkuy Bay, Proval Bay, and Selenga Plume were sampled at multiple depths. (B) Taxonomic composition of the free-living (0.22 µm, top) and particle-attached (3 µm, bottom) size fractions across surface samples. While both fractions were mostly dominated by common freshwater phyla, Actinobacteria were enriched in the 3 µm fraction.

Free-living (0.22 µm) surface samples in the open Baikal were dominated by betaproteobacterium *Limnohabitans* sp. Littoral zones were more variable, represented by multiple OTUs classified as *Limnohabitans*, *Synechococcus* and *Actinobacteria* acI. Surface samples of the shallow Proval Bay and the Selenga river plume represented the extreme end of the eutrophic gradient and were dominated by Actinobacteria acI and Verrucomicrobia *Chthoniobacter*. The latter is reportedly incapable of growth on amino acids or organic acids other than pyruvate, suggesting the taxon is likely involved in the breakdown of partially oxidized organic carbon. Together with Bacteroidetes *Sediminibacterium* and *Cytophagia*, and Betaproteobacteria *Limonhabitans* and *Polynucleobacter*, the six OTUs make up ~46% of each

164 Proval Bay and Selenga River plume's relative abundance. The deepest samples in this study,
165 collected at 500 m and 300 m, were dominated by Actinobacteria acI and acIV clades as well as
166 ammonia oxidizing Group 1a Crenarchaeota *Nitrosopumilus*, commonly found in oligotrophic
167 marine environments.
168
169 **Community Richness and Diversity.** Free-living community diversity showed opposing trends
170 in the upper mixed layer (ML) and deeper waters (DW). Overall, richness and evenness strongly
171 correlated with temperature in the ML and with depth in DW. For each sample, we calculated the
172 effective number of species (ENS), a measure of diversity (25–27). Hill (25) identified ENS as
173 the number of equally-common species that yields a given value of a diversity index, such as the
174 Shannon H'. Jost *et al.* (26) have argued that because ENS scales linearly with richness of
175 equally-common species, it is the preferred metric for quantitative analyses (see Supplementary
176 Methods). We show that the ENS increased with depth in DW (Fig. 2A) and was driven by an
177 increase in evenness (Fig. 2C) with no accompanying trend in OTU richness (Fig. 2B). In the
178 ML, ENS showed a modest increase at the surface in samples collected at 0 m (Fig. 2B). Unlike
179 in DW, higher ML diversity was generated by higher OTU richness (Fig. 2B), while evenness
180 remained unaffected. Mixed layer diversity was positively correlated with temperature (Fig. 2D),
181 along with OTU richness (Fig. 2E). Deep water diversity and richness were marked by high
182 variability with respect to temperature, and richness showed a marginally significant increase in
183 samples from cooler water. Evenness was not directly correlated with temperature in either layer
184 (Fig. 2F). Furthermore, in mixed layer, richness had a greater response (slope) to temperature
185 than diversity did, suggesting that samples at higher temperatures and shallower depths, while
186 supporting the greatest richness, were dominated by a few successful OTUs. This could be due to

variable and high resource conditions leading to coexistence of more taxa, including the persistence of rare taxa.

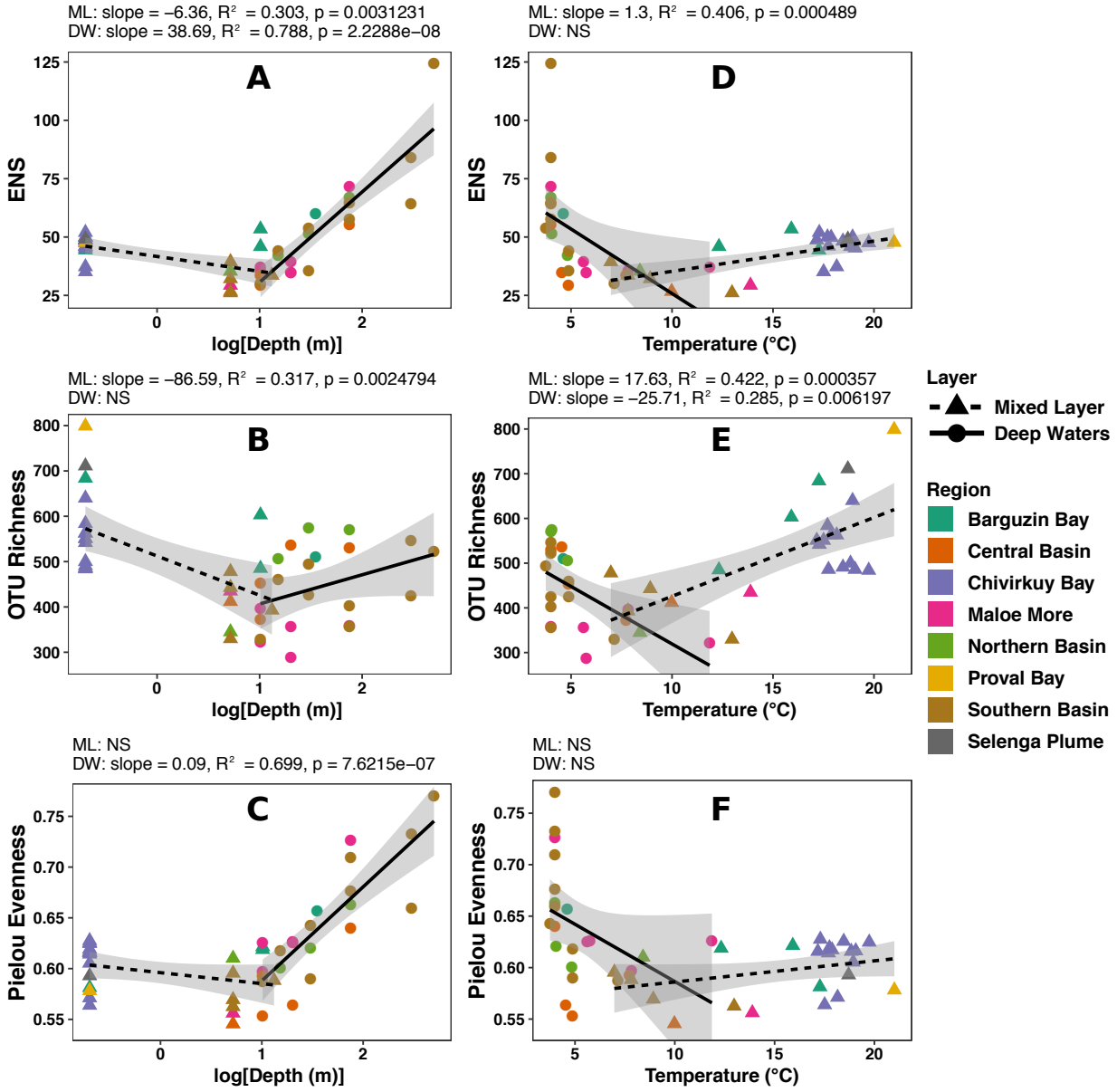


Figure 2: Diversity and evenness trends across depth and temperature in the upper mixed layer and deep waters. Effective Number of Species (ENS) was driven by richness in mixed layer and by evenness in deep waters.

Deep waters offer the opposite story. While richness had a significant but weak relationship with depth, ENS and Pielou evenness were very strongly positively correlated with depth

196 (R^2=0.79, p=2.2x10^{-8}; R^2=0.70, p=7.6x10^{-7}), indicating that a substantial increase in diversity in
197 the deep hypolimnion was not driven by richness but by an even community structure. These
198 trends reveal an intriguing picture: warmer shallower part of ML and deeper parts of DW both
199 support more diverse communities. But ML diversity, while rich in OTUs, is highly uneven,
200 whereas the deep water communities achieve diversity by increasing evenness. The trends were
201 similar for the 3 μm size fraction (Fig. S7). To better resolve underlying drivers that underpin
202 these contrasting community features, we modeled ENS diversity, richness, and evenness with
203 measured environmental variables using iterative multiple regression and model averaging (11).
204 Diversity trends in the ML were different, compared with deep waters. Diversity in the mixed
205 layer correlated the most with temperature and total dissolved silica (TDS) (Fig 3A). These
206 covariates were also responsible for driving the OTU richness, while no measured environmental
207 variable was a reliable predictor of ML evenness (Fig. 3B). Depth was not an important predictor
208 for any ML community feature (Fig. 3C). These results highlighted the importance of
209 temperature and silica in driving the ML diversity, by elevating OTU richness. In the deep
210 waters, total phosphorus (TP) and depth correlated with ENS (Fig. 3D). Furthermore, only TP
211 was a significant predictor for the deep water OTU richness, while only depth was for evenness
212 (Fig. 3E, F). Temperature, when controlling for other covariates, was not an important
213 community feature predictor in deep water. The importance of temperature in ML and depth in
214 DW was not unexpected, considering that the surveyed ML spanned hundreds of kilometers
215 across all three Baikal's basins and multiple bays, where temperature ranged from a median
216 9.5°C in open waters to 21°C in Proval Bay (Fig. S1), while DW was characterized by a large
217 range of depths from 10 to 500 m and a relatively constant temperature environment.
218 Additionally, nutrients appeared to associate specifically with OTU richness. Dissolved silica,

219 combined with light availability in ML, is known to promote diatom growth. Diatom exudates,
220 acting as resources, could create additional niches, enabling coexistence of a greater number of
221 OTUs. In deep waters, TP could promote OTU richness by supporting higher phosphorus
222 demands often attributed to fast-growing copiotrophic organisms (28). Logue and colleagues
223 (29) found TP alone to be significantly correlated with OTU richness in a survey of Swedish
224 lakes. Our results suggest that eutrophication leads primarily to an increase in the overall number
225 of OTUs, but also to dominance of those with opportunistic lifestyles.
226
227 **Multivariate trends.** Ordination of the Bray-Curtis dissimilarity matrix of the 0.22 μm samples
228 revealed clear grouping, which we designated into four significant clusters (Fig. 4A,
229 permANOVA $p<0.001$). The first cluster (eutrophic) comprised samples from shallow and warm
230 areas, specifically inner Chivyrkuy and Proval bays. The second cluster (transition) included
231 samples from the Chivyrkuy Bay to open lake Baikal gradient, the main tributary Selenga River
232 plume, and the two samples collected in Barguzin Bay – the deepest and most open of the three
233 examined bays. The third cluster (open) contained the bulk of our samples collected in open
234 waters that largely separated in ordination space along a depth gradient. The last cluster (deep)
235 comprised the three deep water samples from 300 and 500 m. *Post hoc* pairwise comparisons of
236 cluster centroids revealed significant differences between each cluster pair (FDR corrected
237 $p<0.05$). Thus, the clusters broadly captured the different habitats of lake Baikal, separated by
238 depth and the transitions between open lake waters and bays. Given the confounding effects of
239 abiotic forcing and spatial autocorrelation, we used reciprocal causal modeling (30, 31) to test
240 whether selection by abiotic environment or spatial dispersal/distance better explained
241 community composition trends in mixed layer.

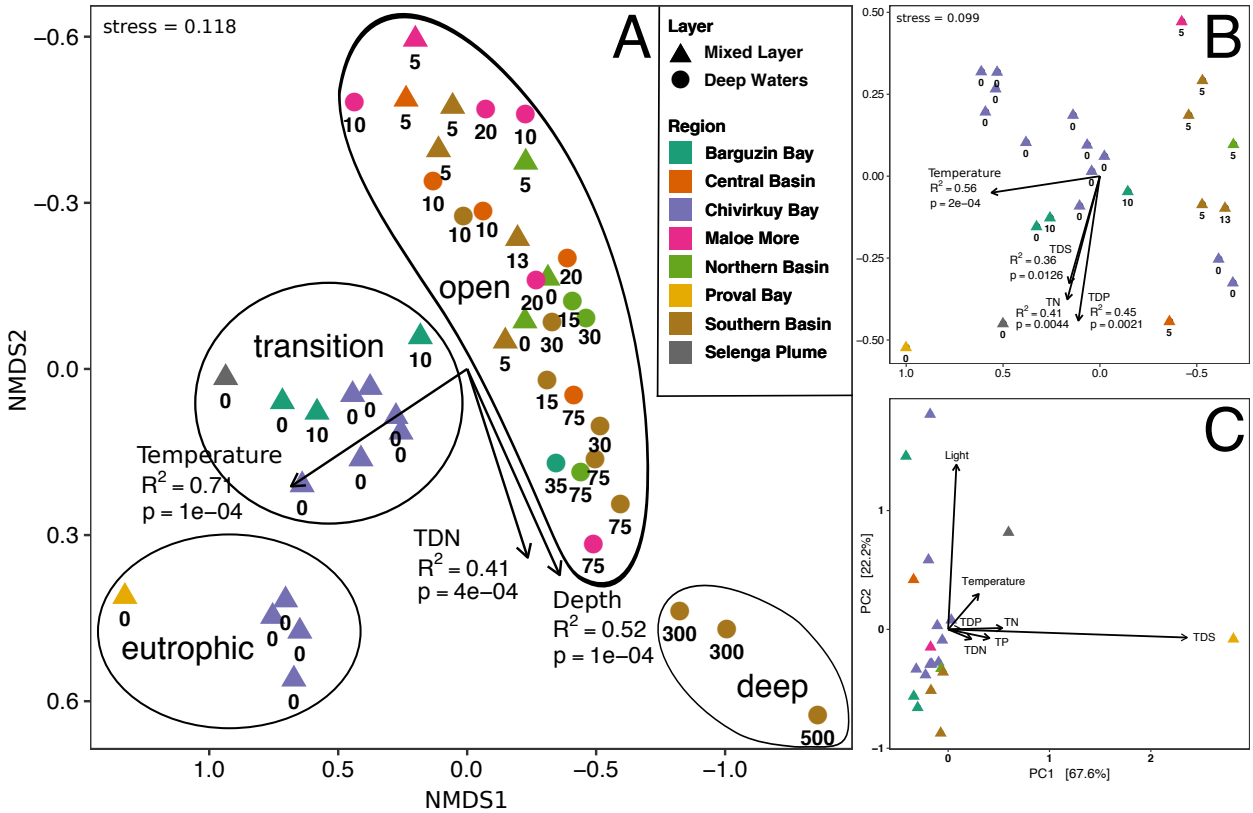


Figure 4: Ordination of the Bray-Curtis dissimilarity matrix for all 0.22 µm fraction samples (A) and the mixed layer only samples (B). (C) Principal component analysis of the mixed layer samples in environmental space using all measured covariates. In all panels, colors reflect major recognized regions of lake Baikal (15, 22). In (A) and (B), numbers below bubbles indicate sample depth (m), and arrows show correlation of environmental covariates with the layout of sample points in ordination space. Each arrow length = R^2; p-values were obtained using permutations. In (A), ellipses were drawn to aid visualization.

Environmental conditions, and not the distance, played a dominant role in structuring free-living mixed layer communities. Reciprocal causal effects models test the hypotheses for significant correlation between the community dissimilarity matrix and each of the geographic distance and environmental distance matrices, while controlling for the other. We found that, when controlling for environment, geographic distance had no correlation with community dissimilarity. However, when controlling for geographic distance, community structure did show a significant correlation with the environment (R^2=0.33, $p<1x10^{-6}$). Our results are consistent

258 with the majority of other studies in freshwater (32, 33) and marine systems (34), which usually
259 note stronger effects of environment on species composition, compared with dispersal.
260 Temperature, depth, and total dissolved nitrogen (TDN) were the strongest environmental
261 covariates with multivariate dissimilarity trends among free-living communities (Fig. 4). The
262 temperature vector (Fig 4A, R^2=0.71, p=1x10^{-4}) indicated the direction of greatest temperature
263 variability along the open waters to bays gradient. We also confirmed presence of the depth
264 covariate among samples collected in the open waters (R^2=0.52, p=1x10^{-4}). Interestingly, open
265 water community dissimilarities also revealed correlation with TDN in approximately the same
266 direction as depth. Because depth can confound the effects of various environmental factors, we
267 further considered an ordination of samples just from the ML (Fig. 4B). Among the ML samples,
268 temperature was still the strongest predictor of community dissimilarity (R^2=0.56, p=2x10^{-4}), and
269 TDS was also significant (R^2=0.36, p=0.012). Notably, these were also the only two model-
270 averaged important predictors of OTU richness and ENS in ML (Fig. 2A, B). However, TN and
271 TDP were also significant predictors of community dissimilarity in ML (Fig. 4B), although they
272 did not show an association with alpha diversity metrics (Fig. 2A, B, C). As expected, depth was
273 a not a significant predictor of community dissimilarity in the ML. Lastly, we investigated
274 whether the same environmental variables that predicted differences in microbial community
275 structure also accounted for major abiotic differences between sampled sites. For this, we
276 constructed a PCA of mixed layer samples using all measured environmental factors (Fig. 4C).
277 The first two principal components captured 89.8% of the variation. TDS and light had the
278 highest scores of all environmental variables. Furthermore, they were almost parallel to PC1 and
279 PC2, respectively, suggesting that the two were responsible for explaining most of the measured
280 abiotic differences between sampled ML sites. While TDS was a significant environmental

281 covariate in ordination of biotic community structure, light was not (Fig. 4B). Indeed, light was
282 also not a significant predictor of biotic alpha diversity metrics (Fig 2). In an opposing example,
283 TDP was the least important predictor of abiotic differences in ML (Fig. 4C); however, it was the
284 second most significant predictor of biotic dissimilarity in ML (Fig. 4C). These features
285 highlight that changes in microbial community structure were not correlated with simply the
286 most variable environmental covariates, and that significant associations of particular
287 environmental factors warrant attention in further studies.
288
289 **OTU co-occurrence networks.** To gain more insight into bacterial community structure in Lake
290 Baikal and its dependence on the abiotic drivers and to identify OTUs that tend to co-occur, we
291 constructed an OTU co-occurrence network. Co-occurrence networks enable data-guided
292 identification of OTU assemblages and correlation of their abundance with continuous
293 environmental variables. In contrast to ANOVA-based permutation procedures, networks do not
294 carry a user bias of arbitrarily defining sample categories, which, even if chosen wisely, result in
295 information loss.
296 Our network captured the bulk of Baikal's OTUs. We constructed a co-occurrence network
297 for OTUs present in at least 80% of samples (105 nodes and 819 edges, Fig. 5; see
298 Supplementary Methods). Importantly, although 105 OTUs appeared to be a large reduction
299 from the total >38,000 detected in Baikal, the network OTUs amounted to over >81% of
300 cumulative relative abundance of every sample in >85% of samples. For the remainder of the
301 samples, coverage averaged 63% ± 17%. Environments that were better observed had more
302 OTUs in the networks. Lower coverage environments included the notable outliers collected at
303 500 m, 300 m in Southern Basin and at the surface of the eutrophic Proval Bay and Selenga

River plume. Thus, with network analyses we used non-categorized continuous data to directly capture the vast majority of microbial community and environmental covariate data, revealing the dominant groups of co-occurring OTUs and their association with major environmental drivers.

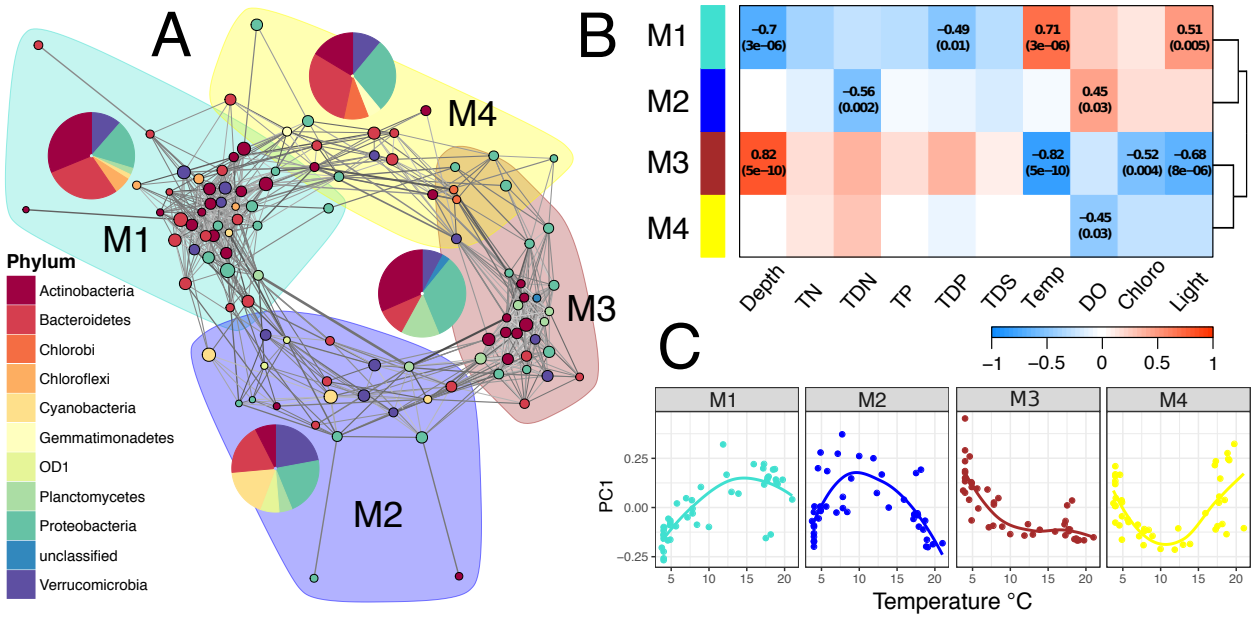


Figure 5: (A) Co-occurrence network of OTUs across all samples. Edge grayscale hue reflects pairwise correlation strength (darker edges show stronger correlations), and thickness indicates edge betweenness score. Modules were defined using maximum modularity optimization. For each module, the first principal component (eigenvector, PC1) of just that module's constituent OTUs abundance matrix was used to summarize the dominant abundance trends across sampled sites. (B) PC1 trends are summarized in a heatmap, where numbers are Spearman correlation coefficients with p-values in parentheses below. Empty cells indicate non-significant results. (C) Example PC1 trends are shown with respect to temperature.

The resulting network exhibited the "small world" properties. Small world networks are characterized by high connectivity between neighboring nodes and low connectivity between distant nodes (35), creating clusters or modules of consistently co-occurring OTUs. Within each module, OTUs with many connections (central nodes) are thought to reflect the processes that bring together module members. Small world networks may be more robust to perturbations but

324 changes to the abundances of the central, well-connected OTUs may disproportionately affect
325 the whole modules (36).
326 Using the optimal modularity approach (16, 37), we identified four modules of OTUs that
327 tend to co-occur across sampled sites (Fig. 5A). Strong cohesion (see Supplementary Methods)
328 within Module 1 (M1; clustering coefficient, CC_{M1}=0.68; Table S1) and Module 3 (M3;
329 CC_{M3}=0.82) suggested common drivers for the consistently co-occurring OTUs. In contrast, low
330 cohesion within Module 2 (M2; CC_{M2}=0.54) and Module 4 (M4; CC_{M4}=0.58) pointed to a looser
331 internal structure.
332 Module eigenvectors (17) or M1 and M3 revealed opposing monotonic relationships with
333 respect to environmental covariates (detailed introduction in Supplementary Methods). M1
334 showed a negative monotonic relationship with depth (Spearman ρ=-0.70, p=$3x10^{-6}$) and a
335 positive relationship with temperature (ρ=0.71, p=$3x10^{-6}$). In contrast, M3 had a strong positive
336 relationship with depth (ρ=0.82, p=$5x10^{-10}$) and a negative relationship with temperature (ρ=-
337 0.82, p=$5x10^{-10}$). As expected, the direction of the opposing trends for the two clusters was
338 reversed for light, owing to its inverse relationship with depth (Fig. 5B). Thus, we found that a
339 large fraction of Baikal's planktonic prokaryotes can be classified into one of the two cohesive
340 clusters: mixed layer "warm" M1 and deeper waters "cold" M3 cluster.
341 Modules M2 and M4 (Fig. 5a) showed inverse opposing unimodal (non-monotonic) trends
342 with depth, temperature and light (Fig. 5B, C). Weaker clustering within these modules
343 suggested less cohesion, possibly due to the inclusion of taxa with weaker habitat preferences or
344 terrestrial or riverine dispersal. M2 showed highest cumulative relative abundance at an
345 intermediate temperature of about 10°C, with a strongly positive relationship with oxygen. In
346 contrast, M4 showed a negative relationship with oxygen, abundance peaks at low and high

347 temperature extremes and minimum abundance at approximately 11°C. Specifically, M4 was
348 high in abundance in the immediate surface (0 m) and deep water, but with a sharp drop in
349 abundance at 5 m depth.
350
351 **Phylogenetic signal in modules.** Membership in the modules had a weakly significant
352 phylogenetic signal (Fig. S11). This, combined with different habitat preferences among the
353 modules, suggests that different taxonomic groups displayed preferential environmental
354 associations, thus supporting the importance of niche differences and phylogenetic niche
355 conservatism.
356 M1 and M4 had a high percentage of *Bacteroidetes*, known as opportunistic degraders of
357 high molecular weight organic matter, such as proteins and carbohydrates, with genomes
358 containing numerous carbohydrate-active enzymes covering a large spectrum of substrates from
359 plant, algal and animal origin (38). In aquatic systems, *Bacteroidetes* have been noted to follow
360 pulses of organic matter inputs and cyanobacterial blooms (39). This further casts M1 as a warm
361 water ML module and M4 as the loose product of sediment and terrestrial input.
362 M1 is the only module to contain members (three OTUs) of the Chloroflexi phylum.
363 Although Chloroflexi have been found in diverse environments, including oxygenated (its type
364 genus) and anoxic (40) hot springs, the CL500-11 clade – a subclass of the deep-ocean SAR202
365 clade – has been suggested as characteristic of oxygenated hypolimnia in deep lakes, including
366 Crater Lake (41), Lake Biwa (42) and the Laurentian Great Lakes (43). However, the three
367 Chloroflexi OTUs detected in our Lake Baikal network were from the Chloroflexi class (two
368 OTUs) and Roseiflexales order (one OTU), and, as part of M1, appear to prefer warmer and
369 shallower environments in the lake.

370 The cold deep-water cluster M3 had the highest relative abundance of OTUs in unclassified
371 phyla. While the module also contained characteristically terrestrial Gammaproteobacteria
372 *Pseudomonas* and *Acinetobacter*, their low connectivity suggested they are unlikely to play a
373 central role in M3 processes.
374 M2 is the only module with Cyanobacteria; the module's three most abundant OTUs –
375 indeed, second and fourth most abundant in the entire >38,000 OTU dataset – classified as
376 *Synechococcus*. Cumulative relative abundance of *Synechococcus* OTUs peaked at ~15 m,
377 reflecting the approximate location of deep chlorophyll maximum at stratified stations. To
378 further understand internal structure of the four modules, we focused on central OTUs that
379 played important roles in creating the module structure.
380
381 **Central OTUs in each module reflect module ecology.** The meaning of a node (in our case
382 OTU) position in the network is the subject of much discussion. A central OTU has a significant
383 positive correlation with a large number of that cluster's OTUs. We offer one abiotic and one
384 biotic interpretation, resulting from two explanations for the presence of network clusters in the
385 first place. First, clusters of OTUs that correlate with important environmental covariates, like
386 temperature, could reflect abiotic habitat filtering. In this case, OTUs in modules with higher
387 clustering coefficients would more consistently reflect environmental conditions associated with
388 that module. For example, every OTU in M3, which has the highest clustering coefficient (Table
389 S1), shows clear negative relationship with temperature (Fig. S13). A second explanation
390 requires an assumption that co-occurrences reflect biotic species interactions. Then, a central
391 OTU could be hypothesized to directly increase the abundance of other OTUs, e.g., by
392 synthesizing or otherwise making available a limiting resource. It is important to point out that

393 the two explanations are not mutually exclusive and could both be parts of ecological
394 mechanisms that give rise to observed clusters of co-occurring microbial species.
395 The most globally central (most connected) OTUs are likely to belong to the most populous
396 module with the greatest number of connections (M1). The most central OTU in the network is
397 OTU137, classified as an autotrophic methylotroph LD19, which was previously reported as a
398 summer-fall bloomer in Lakes Michigan and Muskegon, MI (44). Although not very abundant in
399 our dataset, its high centrality suggests it is indicative of an ecological process that is at least
400 partially responsible for supporting other OTUs in the ML.
401 OTU001 (*Limnohabitans*) – also in M1 – has a high overall abundance and the network
402 position that straddles the balance between centrality in M1 and connectedness to M2. Its high
403 relative abundance and ubiquity across the lake suggest important roles in lake ecology.
404 Connectedness to two modules could reflect this ubiquity and further indicate an influence of
405 OTU001 on OTUs that occupy both niches. More broadly, *Limnohabitans* genus has many
406 ecotypes. Its fine-scale phylogeny and functioning, shown experimentally with isolates (45),
407 revealed diverse lifestyles. In fact, three OTUs out of 105 in our network were classified as
408 *Limnohabitans*, and one of the other OTUs (OTU104) was in the hypolimnetic M3. Separation of
409 different *Limnohabitans* OTUs into M1 and M3 with opposing spatial occurrences and
410 environmental preferences suggests they are indeed ecotypes (46) with distinct functions. Future
411 studies can experimentally assess *Limnohabitans* ecotype responses to temperature and other
412 environmental drivers.
413 The most connected OTU in the hypolimnetic M3 was Planctomycetes OTU022, classified
414 into the Phycisphaerales order. Little is known about the ecological role of Planctomycetes in
415 aquatic environments (39). The order is known for its distinct visual appearance, and a recent

416 report of a 3D reconstruction of Phycisphaerales cellular membrane revealed characteristic deep
417 invaginations in the cellular envelope (47). It is possible the increased surface area of
418 Phycisphaerales becomes useful in the oligotrophic hypolimnion of Baikal. In contrast to the
419 most central but not abundant OTU in M1, OTU022 is also the most abundant in M3, which is
420 the most tightly clustered module in the network (Table S1). These data suggest Phycisphaerales
421 may substantially contribute to nutrient cycling below the thermocline with direct impacts for the
422 rest of M3.
423
424 **Functional Redundancy**, calculated as the ratio of the number of OTUs in the sample which
425 genomes could be characterized using PICRUSt to the number of unique KEGG orthologs (KOs)
426 in each sample, increased with temperature in the mixed layer (Fig. 6) and was highest in bays.
427 This may be because littoral (nearshore) zones are expected to exhibit more variable
428 environments, supporting co-existence of functionally redundant species. Functional redundancy,
429 where more than one species perform a similar function, enables greater functional stability of
430 ecosystems in the face of perturbations, including environmental change (48). The positive
431 relationship with temperature suggests that, because of the lower functional redundancy,
432 microbial communities in cold surface waters of Lake Baikal may be more vulnerable to the loss
433 of functionality following disturbances. In deep waters, functional redundancy increased with
434 depth (Fig. 6). While our PICRUSt-inferred estimates of functional redundancy are based on an
435 indirect and incomplete genomic characterization of OTUs in samples, the OTU coverage was
436 adequately uniform and did not have major taxonomic gaps, which allowed comparing different
437 samples. Moreover, the PICRUSt estimates agreed with the metagenomic data for the samples
438 that were shotgun sequenced. Our results warrant further in-depth analyses of microbial

communities using techniques like isotope probing of phylogenetic microarrays, that track resource utilization, to reveal potential trade-offs between resource partitioning in variable environments and streamlining in constant conditions.

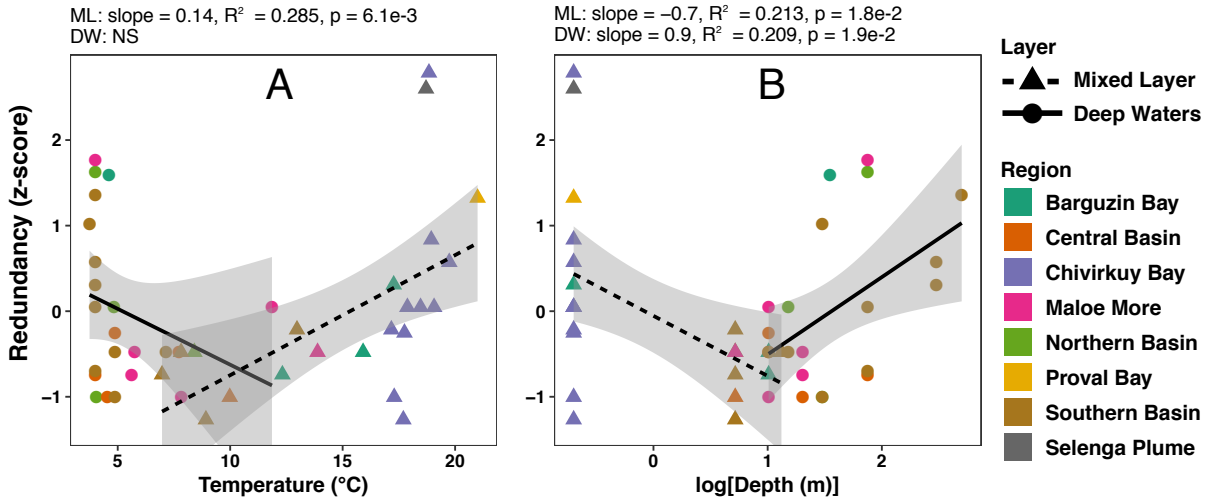


Figure 6: Functional redundancy (number of OTUs/number of KOs) across sampled temperature (A) and depth (B) gradients. The number of OTUs are the number of OTUs that had the PICRUSt-inferred genomes and KOs were determined using the PICRUSt tool.

In summary, our first extensive survey of Lake Baikal's bacterioplankton revealed that temperature and nutrients are the major drivers of the microbial community composition, diversity and functional redundancy, so that the anthropogenic changes in these factors would likely significantly alter planktonic microbial communities. The OTU co-occurrence network analysis identified two major clusters, associated with the upper mixed layer and deep waters, with a detectable phylogenetic signal in each cluster composition. The "small world" properties of the network suggest that the communities may be resilient to perturbations but the changes in abundances of the central OTUs may result in the network rearrangement. The lower functional redundancy in cold, low nutrient open water communities compared to warmer, higher nutrient waters may result in their higher vulnerability to changing environment.

ACKNOWLEDGMENTS. We thank Ted Ozersky for discussions, Lev Yampolsky for help with sampling plan design, and Pam Woodruff and Allyson Hutchens for assistance in the laboratory with nutrient concentration measurements. We are grateful to the crew of the R.V. *Treskov* of the Limnological Museum in Listvyanka and field technicians Elena Pislegina and Alexander Pislegin at the Institute of Experimental Biology, Irkutsk State University for facilitating fieldwork. We thank Ashley Shade and Jim Tiedje for comments on the manuscript. This study was funded by National Science Foundation (NSF) Dimensions of Biodiversity grant DEB-1136710.

REFERENCES

1. Lindeman RL. 1942. The Tophic-Dynamic Aspect of Ecology. Ecology 23:399–417.
2. Cotner JB, Biddanda BA. 2002. Small players, large role: Microbial influence on biogeochemical processes in pelagic aquatic ecosystems. Ecosystems 5:105–121.
3. Moore MVM, Hampton SES, Izmest'eva LLR, Silow E a., Peshkova E V., Pavlov BK. 2009. Climate Change and the World's "Sacred Sea"—Lake Baikal, Siberia. Bioscience 59:405–417.
4. Hampton SE, Izmest'Eva LR, Moore M V., Katz SL, Dennis B, Silow E a. 2008. Sixty years of environmental change in the world's largest freshwater lake - Lake Baikal, Siberia. Glob Chang Biol 14:1947–1958.
5. Team CW, Pachauri RK, Meyer LA. 2014. IPCC, 2014: climate change 2014: synthesis report. Contribution of Working Groups I. II III to Fifth Assess Rep Intergov panel Clim Chang IPCC, Geneva, Switz 151.
6. Holtzendorff J, Partensky F, Mella D, Lennon JF, Hess WR, Garczarek L. 2008. Genome streamlining results in loss of robustness of the circadian clock in the marine cyanobacterium Prochlorococcus marinus PCC 9511. J Biol Rhythms 23:187–199.
7. Martiny AC, Coleman ML, Chisholm SW. 2006. Phosphate acquisition genes in Prochlorococcus ecotypes: Evidence for genome-wide adaptation. Proc Natl Acad Sci 103:12552–12557.
8. Kozhov M. 1963. Lake Baikal and Its Life. Springer Netherlands, Dordrecht.
9. Kozhova OM, Izmest'eva LR. 1998. Lake Baikal : evolution and biodiversityBiology of inland waters. Backhuys Publishers.
10. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a

dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. Appl Environ Microbiol 79:5112–5120.

11. Calcagno V, Mazancourt C De. 2010. glmulti : An R Package for Easy Automated Model Selection with (Generalized) Linear Models. J Stat Softw 34:1–29.

12. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Henry M, Stevens H, Szoecs E, Wagner H. 2018. vegan: Community Ecology Package. R package version 2.5-1.

13. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217.

14. Friedman J, Alm EJ. 2012. Inferring correlation networks from genomic survey data. PLoS Comput Biol 8:e1002687.

15. Berry D, Widder S. 2014. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Front Microbiol 5:219.

16. Csárdi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal Complex Syst 1695:1695.

17. Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.

18. Langille MGIM, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes J a J, Clemente JJC, Burkepile DE, Vega Thurber RRL, Knight R, Beiko RGR, Huttenhower C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31:814–21.

19. Bel'kova NL, Parfenova V V., Kostornova TY, Denisova LY, Zaichikov EF. 2003. Microbial biodiversity in the water of lake baikal. Microbiology 72:203–212.

20. Denisova LY, Bel'kova NL, Tulokhonov II, Zaĭchikov EF. 1999. Diversity of bacteria at various depths in the southern part of lake Baikal as detected by 16S rRNA sequencing. Microbiology 68:475–483.

21. Parfenova VV, Gladkikh AS, Belykh OI. 2013. Comparative analysis of biodiversity in the planktonic and biofilm bacterial communities in Lake Baikal. Microbiology 82:91–101.

22. Kurilkina MI, Zakharova YR, Galachyants YP, Petrova DP, Bukin YS, Domysheva VM, Blinov V V., Likhoshway Y V. 2016. Bacterial community composition in the water column of the deepest freshwater Lake Baikal as determined by next-generation sequencing. FEMS Microbiol Ecol 92:1–13.

23. Schmidt ML, White JD, Denef VJ. 2016. Phylogenetic conservation of freshwater lake habitat preference varies between abundant bacterioplankton phyla. Environ Microbiol 18:1212–1226.

24. Cáceres M De, Legendre P. 2009. Associations between species and groups of sites:\nindices and statistical inference. Ecology 90:3566–3574.

25. Hill M. 1973. Diversity and evenness: a unifying notation and its consequences. Ecology 54:427–432.

26. Jost L. 2006. Entropy and diversity. Oikos 113:363–375.

27. Leinster T, Cobbold CA. 2012. Measuring diveristy: the importance of species similarity. Ecology 93:477–489.

28. Klausmeier C, Litchman E, Daufresne T, Levin S. 2004. Optimal nitrogen-to-phosphorus stoichiometry of phytoplankton. Nature 429:171–174.

29. Logue JB, Langenheder S, Andersson AF, Bertilsson S, Drakare S, Lanzén A, Lindström

ES. 2012. Freshwater bacterioplankton richness in oligotrophic lakes depends on nutrient availability rather than on species-area relationships. ISME J 6:1127–36.

30. Cushman SA, Landguth EL. 2010. Spurious correlations and inference in landscape genetics. Mol Ecol 19:3592–3602.

31. Cushman SA, Wasserman TN, Landguth EL, Shirk AJ. 2013. Re-evaluating causal modeling with mantel tests in landscape genetics. Diversity 5:51–72.

32. Logue JB, Lindström ES. 2010. Species sorting affects bacterioplankton community composition as determined by 16S rDNA and 16S rRNA fingerprints. ISME J 4:729–738.

33. Lindström ES, Forslund M, Algesten G, Bergström A-K. 2006. External control of bacterial community structure in lakes. Limnol Oceanogr 51:339–342.

34. Sjöstedt J, Martiny JBH, Munk P, Riemann L. 2014. Abundance of broad bacterial taxa in the sargasso sea explained by environmental conditions but not water mass. Appl Environ Microbiol 80:2786–2795.

35. Watts DJ, Strogatz SH. 1998. Collectivedynamics of "small-world" networks. Nature 393:440–442.

36. Comte J, Lovejoy C, Crevecoeur S, Vincent WF. 2015. Co-occurrence patterns in aquatic bacterial communities. BGD Biogeosciences Discuss 12:10233–10269.

37. Brandes U, Delling D, Gaetler M. 2008. On Modularity Clustering. IEEE Trans Knowl Data Eng 20:172–188.

38. Thomas F, Hehemann JH, Rebuffet E, Czjzek M, Michel G. 2011. Environmental and gut Bacteroidetes: The food connection. Front Microbiol 2:1–16.

39. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S, R. J. Newton, S. E. Jones, A. Eiler, K. D. McMahon, S. Bertilsson. 2011. A guide to the natural history of freshwater

lake bacteria.Microbiology and molecular biology reviews : MMBR.

40. Overmann J. 2008. Green Nonsulfur Bacteria. Life Sci.

41. Urbach E, Vergin KL, Larson GL, Giovannoni SJ. 2007. Bacterioplankton communities of Crater Lake, OR: Dynamic changes with euphotic zone food web structure and stable deep water populations. Hydrobiologia 574:161–177.

42. Okazaki Y, Hodoki Y, Nakano SI. 2013. Seasonal dominance of CL500-11 bacterioplankton (phylum Chloroflexi) in the oxygenated hypolimnion of Lake Biwa, Japan. FEMS Microbiol Ecol 83:82–92.

43. Denef VJ, Mueller RS, Chiang E, Liebig JR, Vanderploeg HA. 2015. Chloroflexi CL500-11 populations that predominate keep lake hypolimnion bacterioplankton rely on nitrogen-rich DOM metabolism and C1 compound oxidation. Appl Environ Microbiol.

44. Fujimoto M, Cavaletto J, Liebig JR, McCarthy A, Vanderploeg HA, Denef VJ. 2016. Spatiotemporal distribution of bacterioplankton functional groups along a freshwater estuary to pelagic gradient in Lake Michigan. J Great Lakes Res 42:1036–1048.

45. Salcher MM. 2014. Same same but different: Ecological niche partitioning of planktonic freshwater prokaryotes. J Limnol 73:74–87.

46. Fuhrman JA. 2009. Microbial community structure and its functional implications. Nature 459:193–199.

47. Santarella-Mellwig R, Pruggnaller S, Roos N, Mattaj IW, Devos DP. 2013. Three-Dimensional Reconstruction of Bacteria with a Complex Endomembrane System. PLoS Biol 11.

48. Walker BH. 1992. Biodiversity and Ecological Redundancy. Conserv Biol 6:18–23.

SUPPLEMENTARY MATERIALS

MATERIALS AND METHODS

Sampling and environmental contextual data

The survey was guided by the recorded natural history of Baikal (1, 2) that divides the lake into eight distinct regions. We collected samples from 24 spatial locations, where 10 of those locations were sampled at various depths for a grand total of 46 samples across the lake (Fig. S1). The cruise took place on board the R.V. Treskov on August 3-17, 2013.

Temperature and dissolved oxygen profiles were measured with a YSI Instruments sonde (YSI, Inc., Mod4 Springs, OH, USA). Light profiles were measured using Walz model US- SQS/L model Li-185B light probe (Heinz Walz GmbH, Germany), connected to a Li-Cor Quantum photometer (LI-COR Biosciences, Lincoln, NE, USA).

Whole water was taken using a Van Dorn closing bottle (Wildco, Inc., Yulee, Florida, USA). From each sample, (a) 1L was filtered onto a GF/F glass fiber filter (GE Healthcare Bio-Sciences, Pittsburgh, PA, USA) for Chl a measurements and frozen at −20°C; (b) 5 L were sequentially filtered onto 3 µm and 0.22 µm mixed nitrocellulose acetate membranes (EMD Millipore, Billerica, MA, USA) and stored at −20°C in RNAlater (Life Technologies, Grand Island, NY, USA) to capture particle-attached and free-living fractions of microorganisms (3); (c) 50 mL were frozen at −20°C for total nutrient analysis; and (e) 15 mL were filtered through a 0.22 µm mixed nitrocellulose acetate membrane (EMD Millipore) and frozen at −20°C for dissolved nutrient analysis.

Samples were transported to the USA in insulated containers cooled with liquid nitrogen. In the USA, samples for molecular analyses were stored at −80°C and samples for nutrient analyses were stored at −20°C.

Temperature and light profile modeling

Temperature values at collection sites were estimated from YSI instruments sonde profiles. YSI sonde profiled temperature at each station down to between 40 m to 50 m. Five temperature values per second were recorded on the downcast (lowering approximately 1 m per second) and used for modeling. For each station, high-order polynomial functions were fit to the data, and point estimates were used to infer temperature values at exact depths used for water sample collection (Fig. S2). For sites samples collected at 75 m, 300 m and 500 m, temperature was assigned a value of 4°C, based on literature values (4).

Light values at collection sites were estimated from profiles collected using the Li-Cor Quantum photometer (LI-COR Biosciences, Lincoln, NE, USA). Profiles were measured down to site bottom or maximum 20 m depth. The light extinction function $I_D = I_0e^{-kD}$, where I_D = light intensity at depth, I_0 = light intensity at the surface, k = extinction coefficient (0.035 for pure water) and D = depth, was fit to the to the data to estimate light values at collection sites (Fig. S3).

Nutrient Measurements

Samples for nutrient analysis were thawed and digested with potassium persulfate at 120°C for 30 min. Total and dissolved nitrogen were measured colorimetrically on a Shimadzu UV-240-PC spectrophotometer (Shimadzu Scientific Instruments, Columbia, Maryland, USA) at 224 nm using the 2nd derivative method (5). Total and dissolved phosphorus was measured using the orthophosphate method on a Lachat Instruments Quick Chem 8500 autoanalyzer (Lachat Instruments, Loveland, Colorado, USA). Chl *a* filters were extracted with ethanol, and Chl *a* was determined fluorimetrically (6).

DNA extraction and amplicon sequencing

Genomic DNA was extracted using the Mo-Bio PowerSoil Kit (Mo-Bio Laboratories, Carlsbad, CA), following manufacturer's protocol. Then, the V4 region of the 16S rRNA gene was sequenced using dual-index primers as described previously (7). After PCR amplification, the products were normalized and pooled. The pool was loaded on an Illumina MiSeq v2 flow cell and sequenced with a standard 500 cycle reagent kit for paired-end 250 bp reads (PE250). Base calls were done with Real Time Analysis software v1.18.54. Output of RTA was demultiplexed and converted to FastQ with Illumina Bcl2Fastq v1.8.4.

Amplicon sequence processing

The FastQ output files were processed using mothur, following general MiSeq protocol and the options below (7, 8). Sequences were aligned to a full SILVA v.119 database. Chimeras were removed with UCHIME in mothur environment. The remaining sequences were classified with a naïve Bayesian RDP classifier (9) and the Greengenes (August 2013 release) database. Sequences classified as Mitochondria, Chloroplasts and Eukaryota were removed, resulting in 197,738 unique sequences that were clustered into OTUs at 97% similarity. Consensus taxonomy for each OTU was determined following mothur protocol. Coverage for sequenced samples varied between 24348 and 76838 (Fig. S5, left), and was rarefied to the lowest coverage sample i.e. 24348 reads (Fig. S5, right).

Multiple regression

We used the brute force approach to multiple linear regression (exhaustive search, followed by model selection) to determine the relationship of community richness, diversity, and evenness with measured environmental variables. Richness was taken as the number of OTUs in each sample. This comparison was possible because our sampling effort was rarefied in mothur (see above). The effective number of species (ENS), represented diversity and was calculated using the Shannon diversity index H' (10, 11), as shown in Eq. 1.

$$ENS = e^{H'} \qquad \text{Eq. 1}$$

It was important to use ENS because, unlike diversity indices, such as Shannon and Simpson, ENS has a linear response to change in community diversity (Fig. S6) – a necessary property of a response variable in quantitative modeling (10, 11). Statistical modeling and model selection were performed with the glmulti package in R (12).

Multivariate statistical analyses

<u>Distance matrix</u>. Ordination of the OTU abundance matrix was performed in the R environment (version 3.2.2) using vegan (13) and phyloseq (14) packages. First, we compared different techniques for calculating community dissimilarities. We used the Mantel test to reveal correlations between the Bray-Curtis, Jaccard, unweighted Unifrac, and weighted Unifrac distance matrices. Results showed greatest differences between presence-absence distance metrics and abundance-weighted distance metrics. However, within each of the two categories, the metrics were approximately interchangeable. Thus, we decided to use the Bray-Curtis metric for better compatibility with other studies. The high correlation

of weighted Unifrac matrix indicated that phylogenetic information did not largely affect distance matrix results.

Test for dispersion (geographic distance) was done with reciprocal causal modeling using partial Mantel tests. The use of reciprocal partial mantel tests for reciprocal causal modeling is advocated in the literature by Cushman (15, 16), where he criticized simple Mantel tests for their Type I error rates, and proposed reciprocal partial mantel as a solution (15), later expanding to a more sophisticated "relative support" technique - also based on partial Mantel tests (16). Rousset criticized all Mantel-based options in favor of mixed effects regression methods (17) because they are not robust when there is autocorrelation in the environment matrix. In a summary review, Cushman conceded that LME is indeed the best method for decoupling the effect of spatial dispersal and selection; however, that doesn't mean reciprocal partial mantel tests are inappropriate, especially if autocorrelation in data is weak (18). Cushman pointed out that reciprocal partial mantel tests performed almost as well as mixed linear models if conclusions were based on R^2 effect sizes, and not just p-values. Fortunately, *our data does not have significant spatial autocorrelation with respect to environment* (Moran's I > 0.05). Thus, we proceeded using Cushman's method for its simplicity.

Geographic distance between sampled sites was calculated with the geosphere R package using the Vincenty Ellipsoid model. Mantel test was performed using the vegan R package. Because the geographic distance matrix was not normally distributed (Fig. S16), we used the rank-based Kendall test statistic option in the Mantel tests.

Ordination was done in the phyloseq package (14) and modified for visualization with a custom R script. NMDS plots are freely rotatable and scalable, and Fig. 4 panels a and b were thus adjusted for greater visual clarity.

Correlation with environmental variables was done with the envfit function in the vegan package, which correlates continuous environmental values with separation of points in ordination space. Significance was calculated by bootstrapping using 9999 permutations.

Network construction and analyses

Co-occurrence matrix and network construction. The OTU co-occurrence matrix was calculated for OTUs that were present in 80% of all samples (109 OTUs) with sparCC software (19), as recommended in best practices for co-occurrence network construction by Berry and Widder (20). Significance values for correlations were bootstrapped, as described by sparCC authors, and then corrected for multiple testing using FDR manually in R with a custom script. Positive co-occurrences with adjusted $p < 10^{-5}$ were used for downstream analyses. The network was constructed, displayed and analyzed using the iGraph package (21) in the R environment. Two pairs of OTUs that were only connected to each other but not the rest of the network were removed, resulting in 105 OTUs in all subsequent analyses.

Comparison with simulated networks. We compared basic network statistics with null distributions of two types of simulated networks. First, we created 10,000 random Erdös-Rényi networks, which were undirected, without loops and used the gnm model with the same number of nodes (105) and edges (814) as our Baikal network. We also created 10,000 small-world Watts-Strogatz networks with 105 nodes and replacement probability p=0.05. Average path length and the clustering coefficient (transitivity) of the lake Baikal network were compared to distributions of simulated networks (Fig. S13). Two-tailed *p*-values for

138 Baikal network statistics were calculated as the number of simulated observations greater
139 than the absolute values of the Baikal network divided by the total number of simulations.
140 The network exhibited small world properties. Both overall CS (0.618) and the average
141 path length (AP=2.62) were much higher than random Erdös-Rényi simulations (CC
142 mean=0.150, p=0, n=10,000; AP mean=1.93, p=0, n=10,000; Fig. S13a) and higher than
143 small-world Watz-Strogatz simulations (CC mean=0.514; p=0, n=10,000; AP mean=2.27,
144 p=0, n=10,000; Fig. S13b). The basic network statistics are summarized in Table S1.
145 Small-world properties place greater topological importance on central nodes (nodes
146 with many connections) and to a lesser extent bottlenecks in maintaining the network
147 structure. Hubs can be defined as nodes with high eigenvector centrality, and bottlenecks as
148 high betweenness nodes. Hubs are well-connected, in many cases to other well-connected
149 nodes and are therefore considered topologically more central. In an ecological co-
150 occurrence network, they have been hypothesized to mediate processes important to their
151 neighbors (22). For example, a central OTU may produce a common good limiting resource,
152 like a vitamin.
153 Bottleneck nodes are positioned along a high number of shortest paths between pairs of
154 other nodes. This is called high betweenness. Fig. 4 displays betweenness values for each
155 edge as proportional to its width. In a small world network with few bridges between
156 modules, bottlenecks are the only way one module can interact with another. Therefore,
157 bottlenecks have been hypothesized to mediate feedbacks between ecologically meaningful
158 OTU assemblages (22). These features emphasize the strength of network methods to detect
159 candidate keystone taxa not necessarily based on their large abundance but on how they
160 affect other network players.
161 Community (module) detection. Network modules were identified using the optimal
162 modularity algorithm (23) in the iGraph package. Optimal modularity calculates the
163 arrangement and membership of clusters that gives the highest modularity score, as defined
164 by Newman *et al* (24) and adopted by the iGraph authors.
165 Environmental trends of modules. To summarize OTU abundance trends across sampled
166 sites, we used the first principal components of OTU abundance matrices (eigenOTUs (25))
167 for OTU members of each of the four modules. The resulting four eigenOTUs were used to
168 assess correlation of modules with environmental factors, such as temperature, nutrients,
169 oxygen, and chlorophyll levels. For example, module M1 has 39 member OTUs. They are
170 together because, by definition, they have similar occurrence patterns across samples. But
171 how can we summarize all OTU abundances in one vector? One option is to sum their relative
172 abundances in each sample and get per sample cumulative values. However, this option is
173 heavily biased towards the more abundant OTUs. Indeed, the difference between the most
174 and least abundant OTUs in M1 is approximately two orders of magnitude. A better option is
175 to use the widely practiced principal component analysis (PCA), which weighs OTUs by their
176 power to explain variation in other OTUs. In PCA, the first principal component (PC1) is the
177 best summary of OTU abundances in each sample. Thus, we ran an independent PCA for OTU
178 members of each of the four modules and used PC1 for each module as a summary of its
179 abundance across samples. The PC1s were correlated with measured environmental
180 variables (Fig 5b, c).
181 Phylogenetic signal calculation. We first constructed a phylogenetic tree of OTUs by
182 computing pairwise DNA distances using the K80 model and constructing a tree with the
183 bionj algorithm (26). Phylogenetic signal was detected using a modified phylo.signal.disc R

184 script developed by Enrico Rezende (Universidad Autònoma de Barcelona). This function was
185 used in several studies to reveal a phylogenetic signal in discrete traits (27–32). It works with a
186 tree of network OTUs and treats module assignment of each OTU as a character state with
187 random transitions. The strongest possible phylogenetic signal in the OTU tree would be
188 indicated by *num* - 1 transitions, where *num* is the number of modules. To quantitatively
189 assess the phylogenetic signal, the script first uses maximum parsimony to generate a null
190 model by calculating the minimum number of character transitions needed to create
191 permuted character states of the OTUs for 10,000 simulations. Then, the actual number of
192 transitions in the tree of Baikal OTUs with known module membership is compared with a
193 distribution of transitions obtained from the permuted simulations. A p-value for a
194 phylogenetic signal is calculated from comparing the true number of transitions with a null
195 distribution of transitions obtained from the permuted simulations (Fig. S13).
196
197 *Gene content and functional redundancy calculations*
198 Gene content estimation. Gene content per sample was estimated using the PICRUSt (33)
199 python package. PICRUSt functional inference works in three steps. It uses a reference
200 tree(34), constructed from the entire Greengenes database (35) of 408,315 curated full-
201 length 16s sequences, and a table of 6910 KEGG ortholog (KO) abundances for 2,590 known
202 genomes that have identifiers in the Greengenes tree. First, PICRUSt prunes the full reference
203 tree to known genomes and uses phylogenetic modeling to reconstruct ancestral states. The
204 result is estimated copy numbers for each KO in each ancestor in the pruned tree. Second,
205 PICRUSt infers gene content for the tips of the entire Greengenes reference tree. This
206 prediction is generated by an average of the contents of extant and inferred ancestral
207 genomes, both weighted exponentially by the reciprocal of phylogenetic distance, in line with
208 previous research that reported an inverse exponential relationship between 16s
209 phylogenetic distance and gene content conservation (36). In the last step, user-supplied
210 OTUs are matched to identifiers in the reference tree, their abundances are normalized by
211 predicted 16s copy numbers, and a table containing copy numbers of their KOs is created.
212 Prediction of the sample-wide genome content (metagenome) is a simple addition of KO
213 copy numbers for all OTUs in the sample, weighted by their relative abundances. Our
214 preliminary comparison of the PICRUSt-inferred metagenomes with the sequenced
215 metagenomes from a subset of our samples (33) reveals a good agreement (Wilburn et al. in
216 prep.), verifying the reliability of the PICRUSt approach (Fig. S13).
217 PICRUSt offers a quality control metric NSTI (nearest sequenced taxon index) for the per-
218 OTU accuracy of gene content prediction. For example, because OTUs are clustered at 97%
219 sequence similarity, an OTU with NSTI < 0.03 is identical to a known sequenced genome.
220 Since phylogenetic distance from known genomes is the best indicator of predicted genome
221 quality, we worked with PICRUSt results based on OTUs with NSTI < 0.15, as per PICRUSt
222 authors recommendation (33).
223 In Baikal dataset, OTUs compatible with the PICRUSt pipeline were identified within
224 mothur using the make.biom command as described in the MiSeq SOP. Of the PICRUSt-
225 compatible OTUs, 486 non-singleton OTUs with NSTI < 0.15 were used to predict
226 metagenomic content of samples for redundancy analysis.
227 Functional redundancy was calculated by dividing the number of non-zero OTUs by the
228 number of unique non-zero KOs in each sample. This calculation ensures that each KO
229 present in the sample is counted only once and does not depend on the number of gene

230 copies. With more OTUs contributing the same KOs, the ratio (functional redundancy) goes
231 up.
232
233
234 **SUPPLEMENTARY RESULTS**
235 *Temperature, light and nutrient profiles*
236 Most stations were thermally stratified, while others exhibited weaker stratification
237 because of the storm at the end of the cruise (Fig. S2). At the surface, temperature ranged
238 from 7.0°C in northernmost SB to 21°C in Proval Bay with a median at 17.4°C. In the open
239 waters, the median temperature was 9.5°C, consistent with expected values (1).
240 Light profiles with models are shown in Fig. S3, top. Light extinction coefficients (Fig. S3,
241 bottom) were greatest at the single station in the eutrophic Proval Bay, high in the Selenga
242 River plume and lowest in Maloe More. Central Baikal stations, sampled along the gradient
243 from shallow bay to open lake conditions, showed variable values.
244 Nutrient profiles are shown in Fig. S4. Total nitrogen (TN) ranged from 3.0 μM in
245 southern Maloe More Straight to 25.2 μM in Proval Bay with both an overall and open waters
246 medians at 10.0 μM. TP ranged from 0.16 μM in northernmost South Baikal to 1.4 μM in
247 Proval Bay with lakewide and open waters medians at 0.35 and 0.38 μM. Molar TN/TP ratios
248 varied from 5.6 in southern Maloe More to 56.9 in southern Central Baikal; median TN/TP
249 was 26.1. Chl a ranged from 0.66 μg L1 in mid-Chivirkuy Bay to 10.5 μg L^{-1} in Proval Bay with
250 the median at 2.0 μg μg L^{-1}. Nutrients increase, but with mild trends at depths above 80 m
251 (not significant) with a visible increase at the deeper stations at 300 and 500 m. While
252 relationships with TN, TP, DN, DP, DS and chlorophyll were all significant using all samples,
253 only TN and chlorophyll remained significant with two high nutrient outlier stations (Proval
254 Bay and Selenga Shallow) removed. The remaining positive relationship illustrated an
255 expected correlation between nutrient load and primary productivity.
256
257 *Amplicon sequence processing*
258 see methods
259
260 *Multiple regression*
261 Trends on the particle-attached 3um fraction were similar to the free-living 0.22 um fraction
262 (Fig. S6, S7)
263
264 *Multivariate statistical analyses*
265 Dissimilarity matrices based on phylogenetic distance metrics (unweighted and
266 weighted unifrac) correlated very highly with phylogeny-free Jaccard and Bray-Curtis
267 matrices (Fig. S8) and were concluded not to add significant additional information to
268 modeling the multivariate community structure.
269 Ordination of the 3um fraction showed similar results to the 0.22um fraction (Fig. S9).
270
271 *Network construction and analyses*
272 see methods
273

SUPPLEMENTARY FIGURES

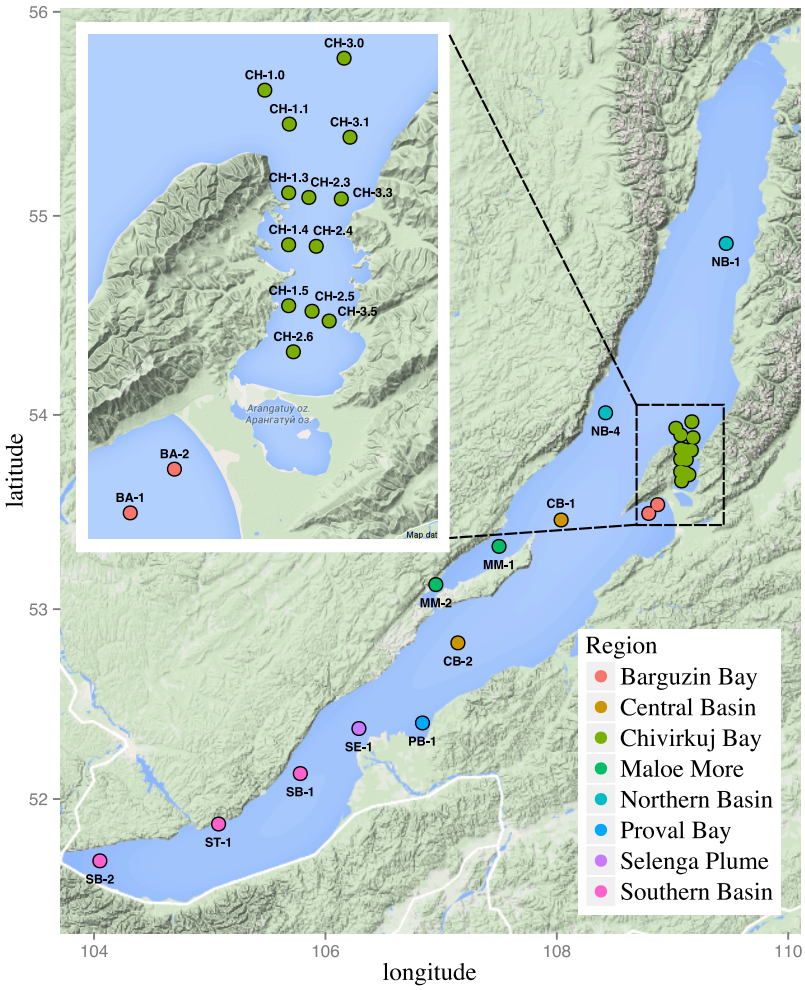


Fig. S1: Sampling sites at Lake Baikal. Colors reflect major recognized regions of lake Baikal (1, 4).

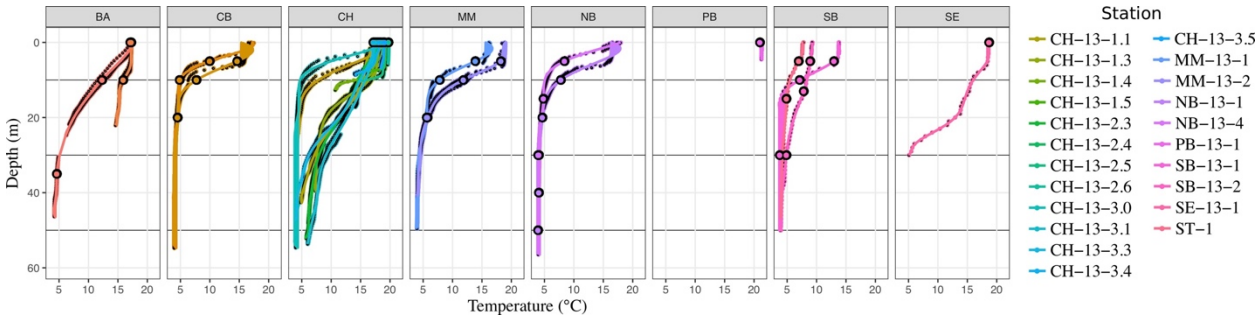


Figure S2: Temperature profiles and location of collected samples in the water column.

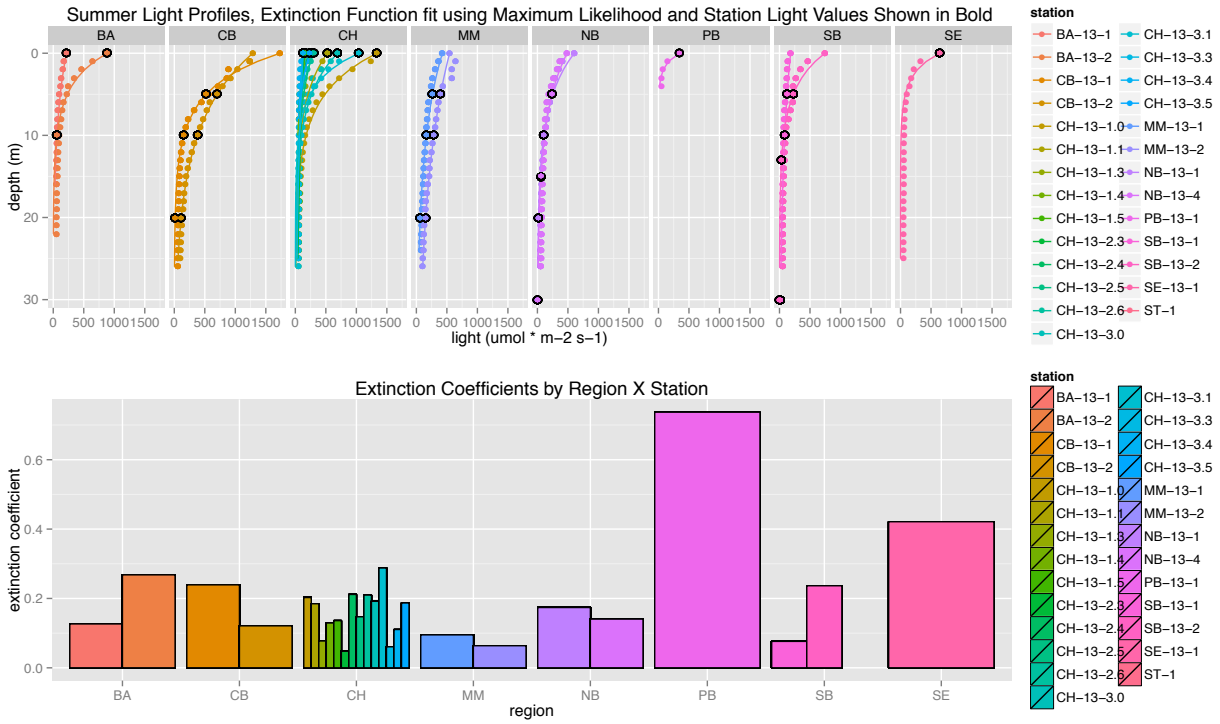


Figure S3: Light profiles, fitted models and extinction coefficients.

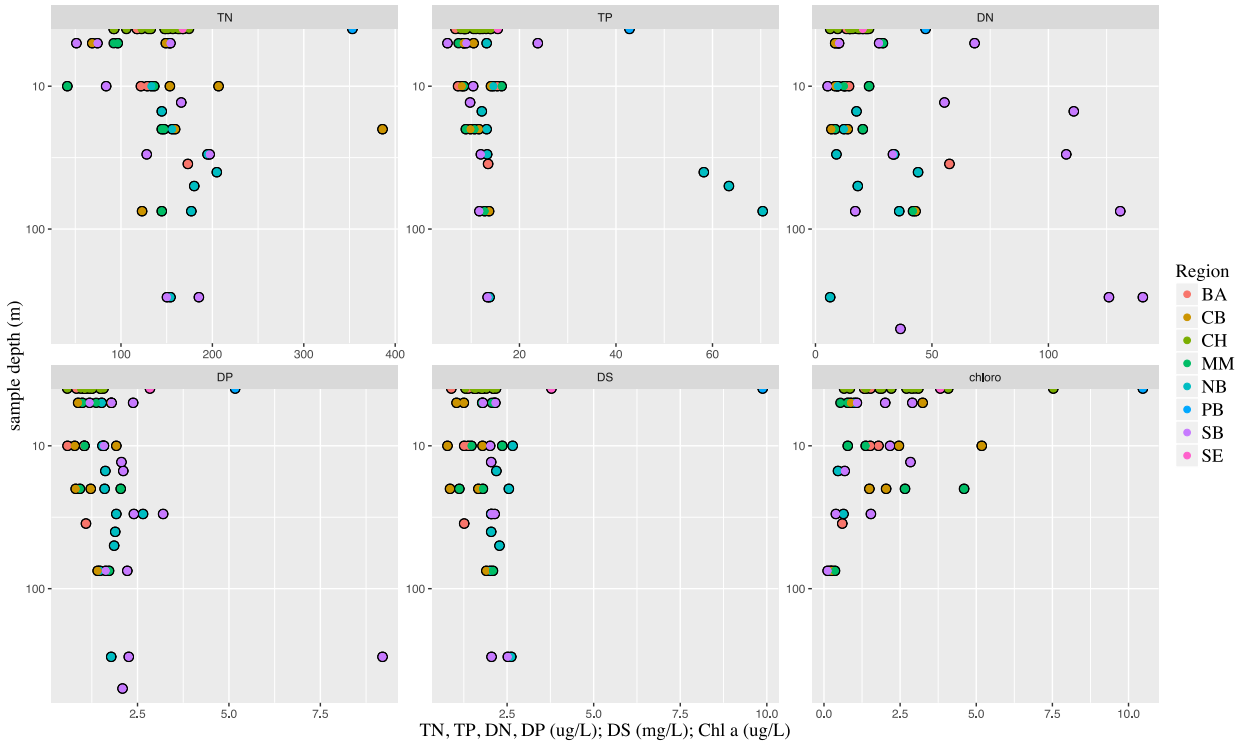


Figure S4: Nutrients show non-significant increase with depth.

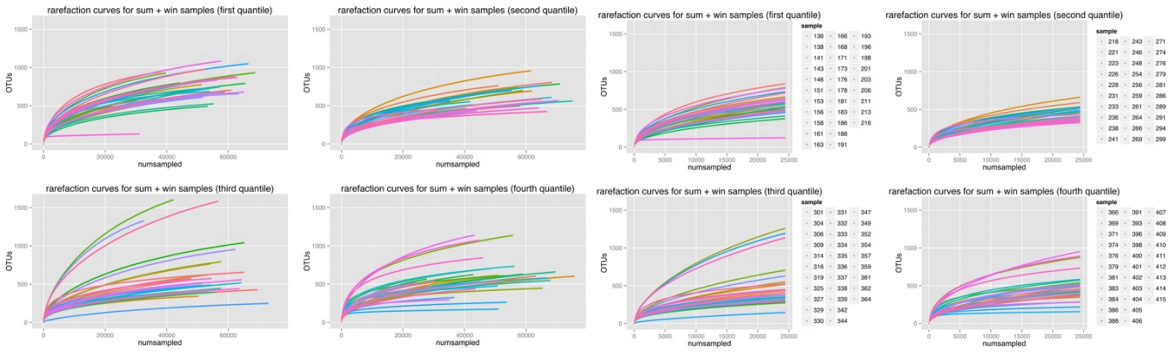


Figure S5: Rarefaction curves showing sampling depth per sample (left) and rarefied samples (right)

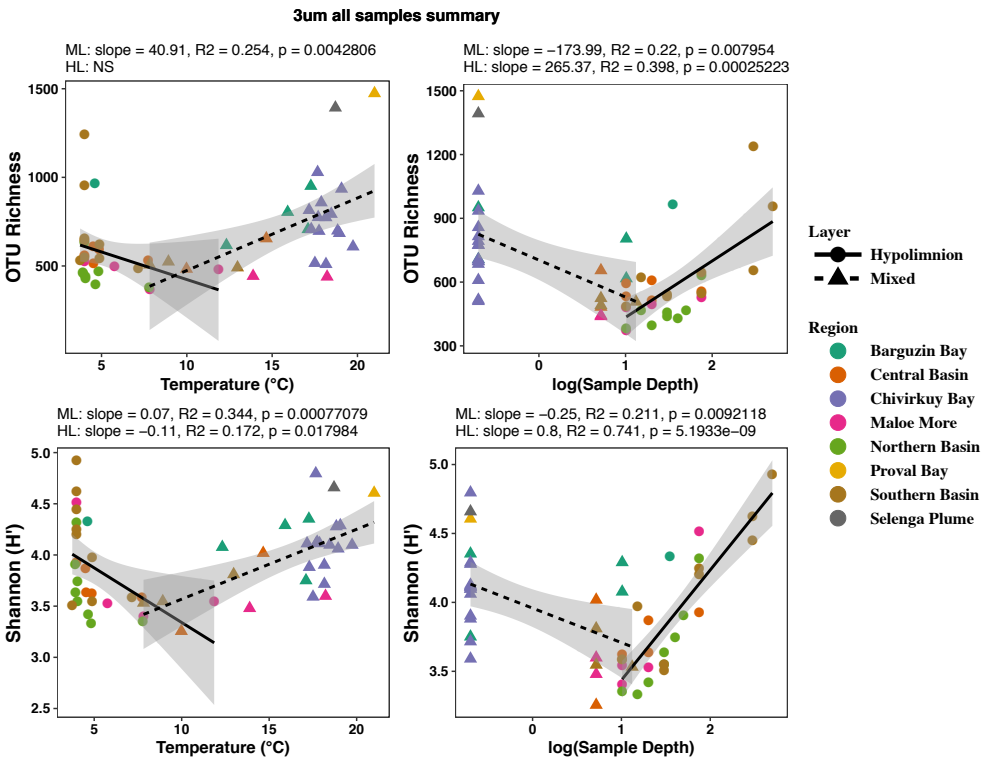


Figure S7: OTU richness and Shannon diversity on the 3 um fraction size, showing the same trends as the 0.22 um fraction.

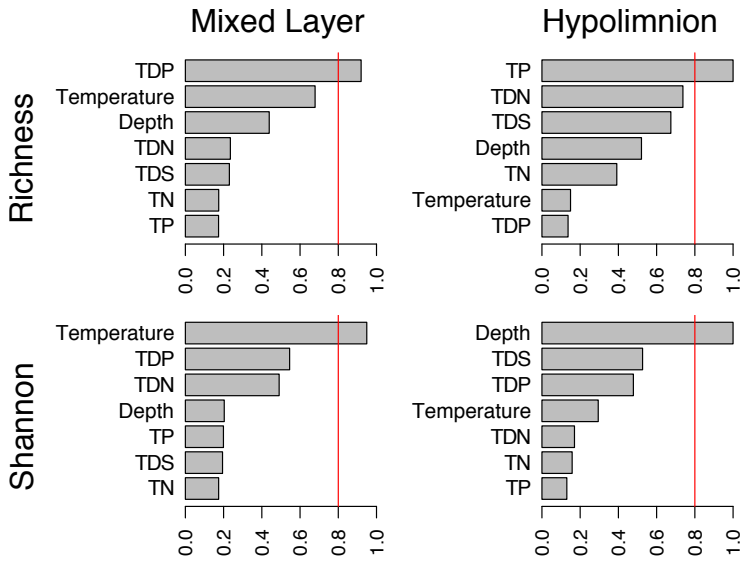


Figure S8: Model-averaged importance of environmental predictors for OTU richness (top) and Shannon diversity (bottom) in the mixed layer and the hypolimnion – on the 3 μm size fraction.

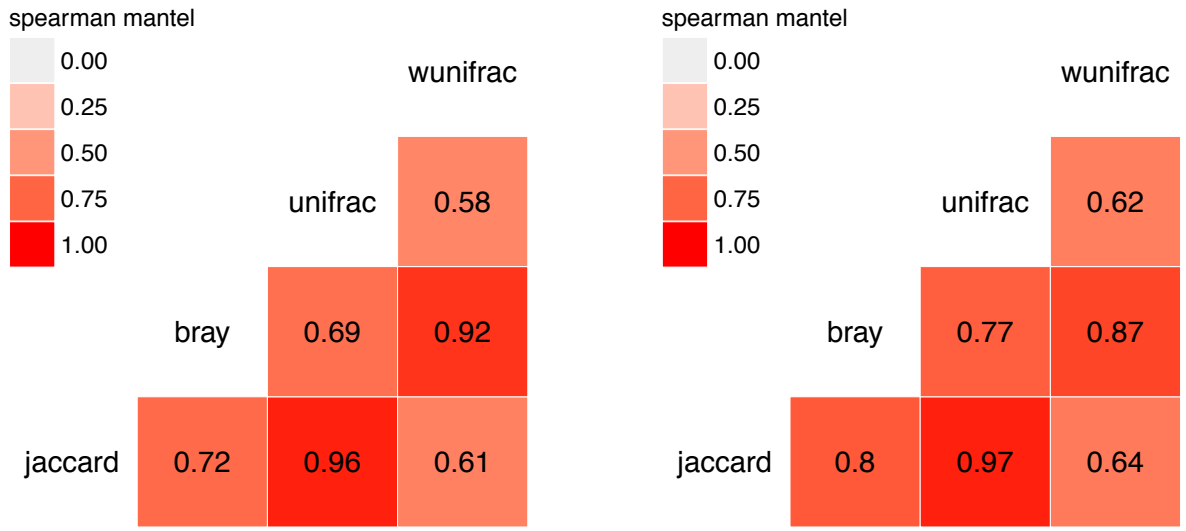


Figure S9: Pairwise Mantel test correlations between distance matrices based on phylogeny-free and phylogenetically-informed metrics for the 0.22 um (left) and 3um (right) size fractions. Unweighted distance calculators treat all OTUs equally, while weighted versions emphasize differences among the more abundant OTUs.

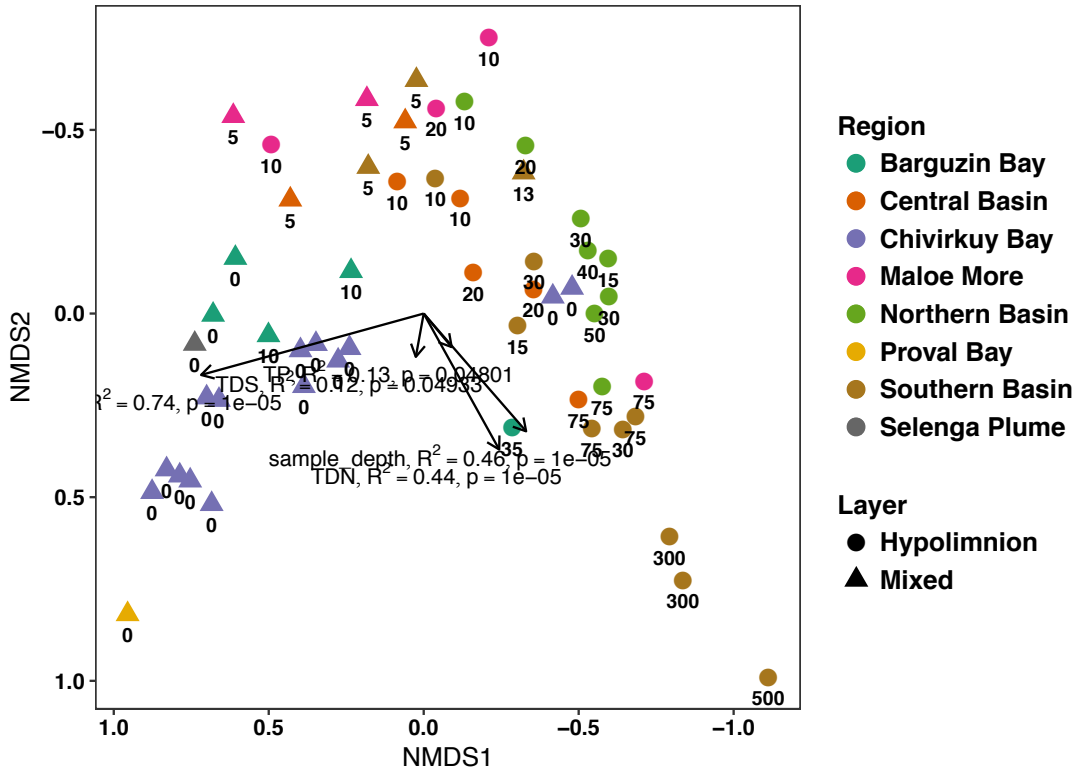
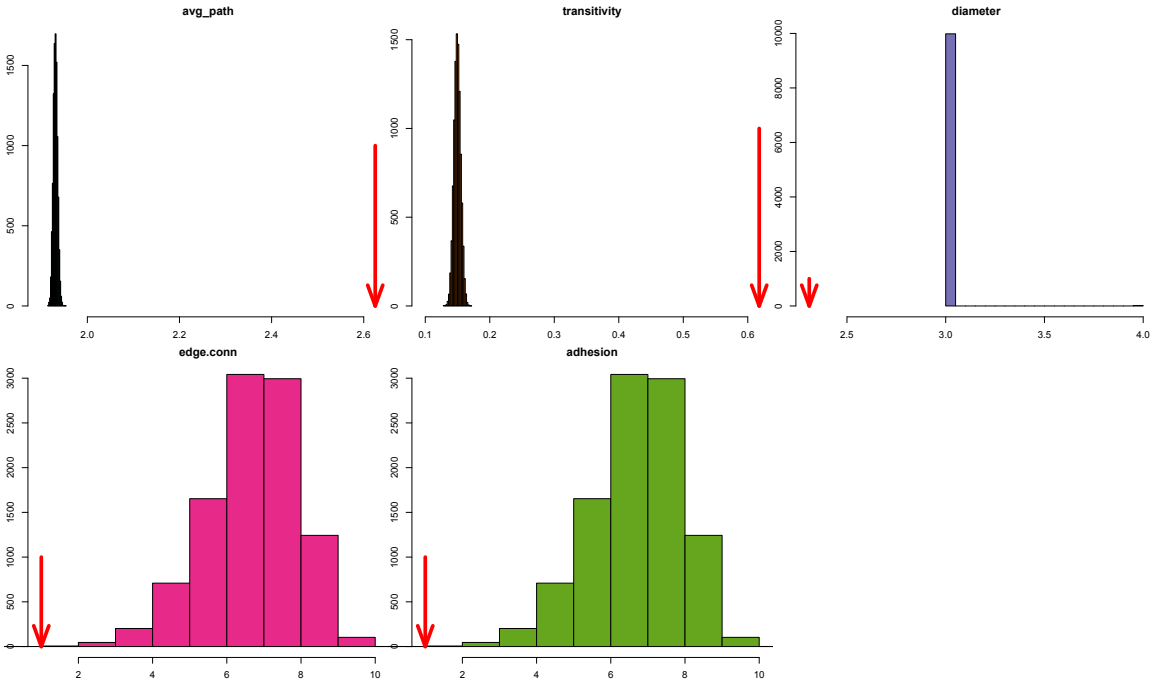


Figure S10: NMDS ordination of the particle-attached 3um fraction



B

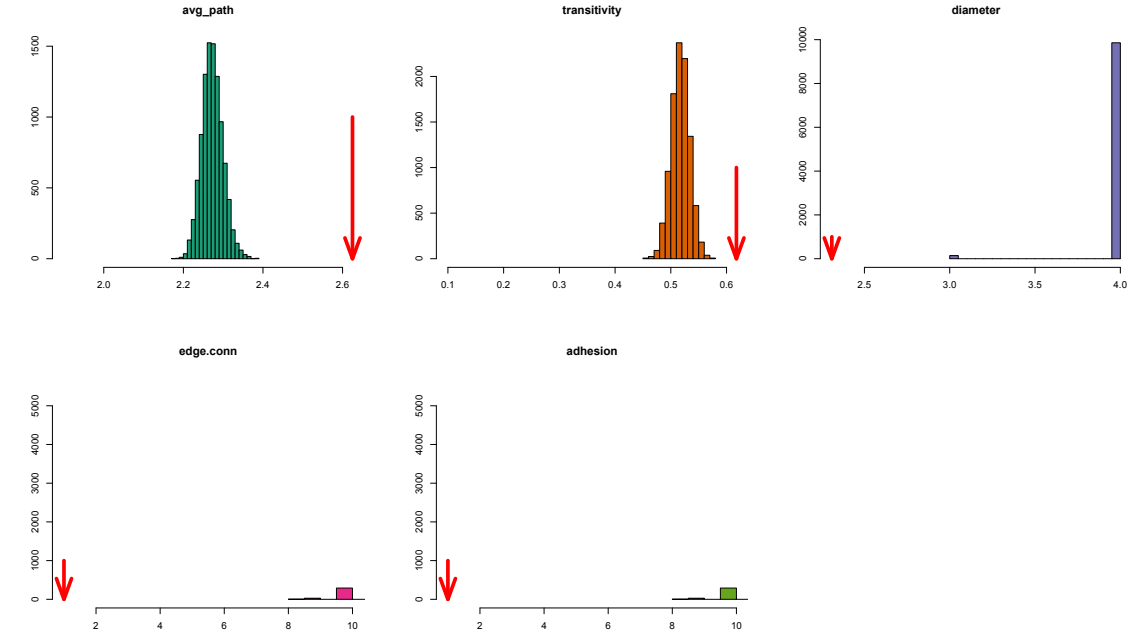


Figure S11: Network statistics comparison with simulated Erdos-Renyi (A) and Watts-Strogatz (B), run for 10,000 simulations each. Red arrows indicate statistics for Baikal networks.

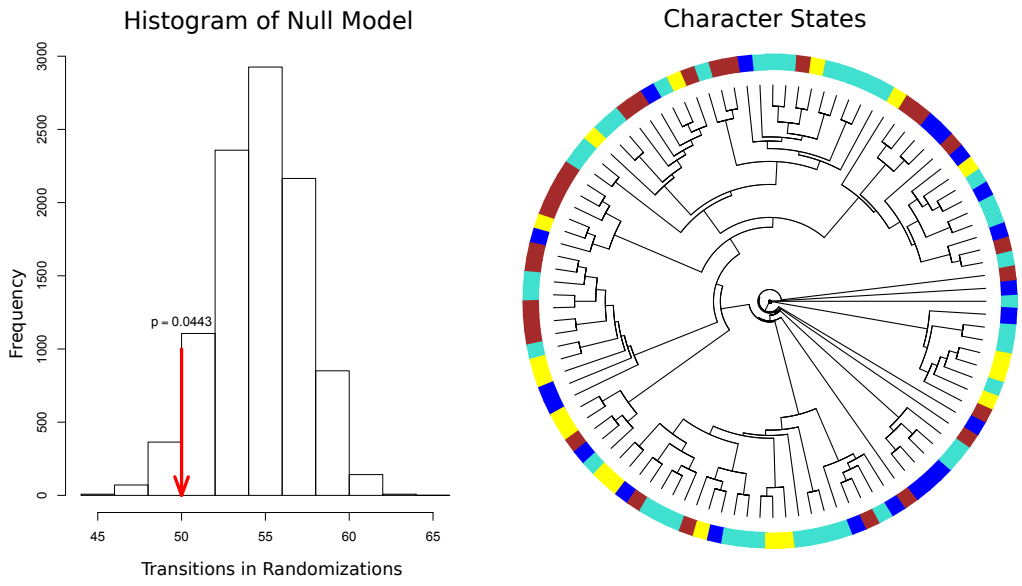


Figure S12: Phylogenetic signal for co-occurrence network modules as discrete character states for individual OTUs. Phylogenetic distribution was non-random across modules, possibly reflecting their ecological roles.

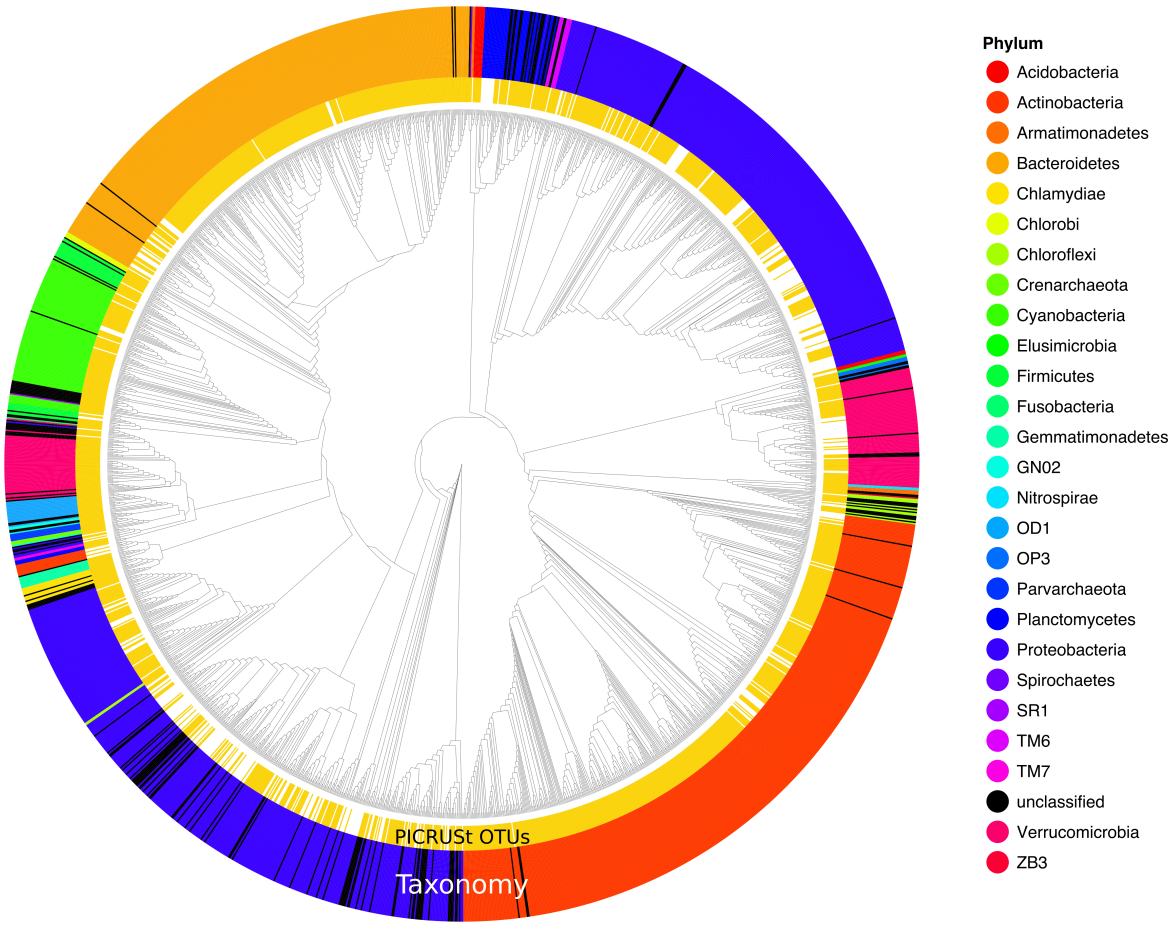


Figure S13: trade-off between eigen centrality and betweenness in the Baikal small world co-occurrence network.

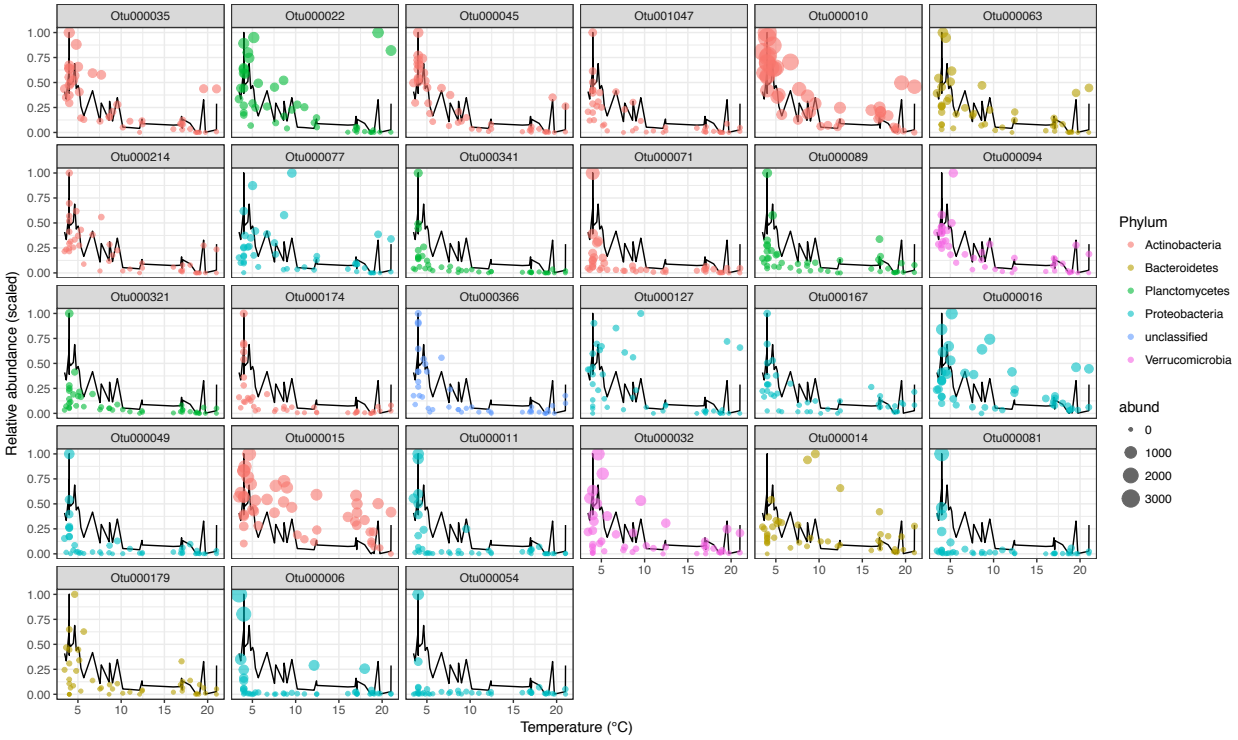


Figure S14: OTU abundance (scaled to 0-1 range) in the **M1** (ML module) are shown in bubbles. The PC1 trend for M1 (scaled to 0-1 range) is shown as a black line in each panel. OTU panels are arranged in order of decreasing centrality (connectedness). Bubble sizes indicate actual relative abundance values of the OTUs to communicate which OTUs were generally more or less abundant in Baikal.

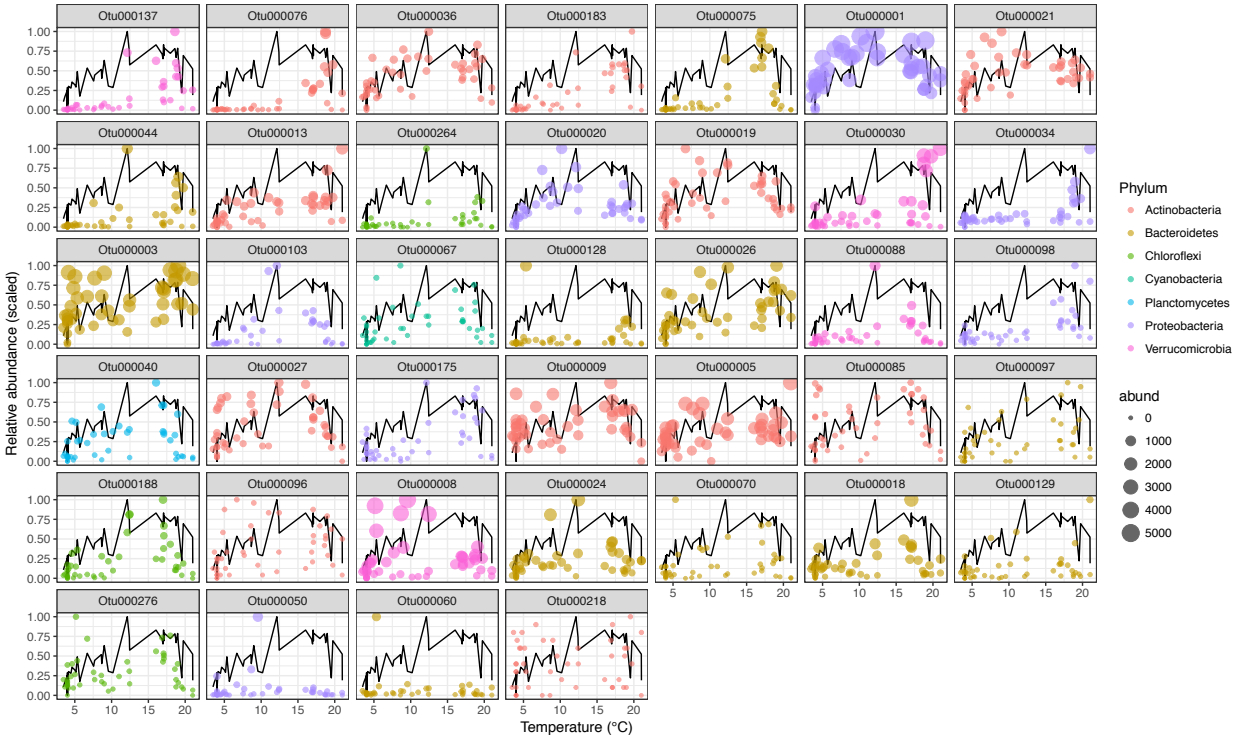


Figure S15: OTU abundance (scaled to 0-1 range) in the **M3** (DW module) are shown in bubbles. The PC1 trend for M1 (scaled to 0-1 range) is shown as a black line in each panel. OTU panels are arranged in order of decreasing centrality (connectedness). Bubble sizes indicate actual relative abundance values of the OTUs to communicate which OTUs were generally more or less abundant in Baikal.

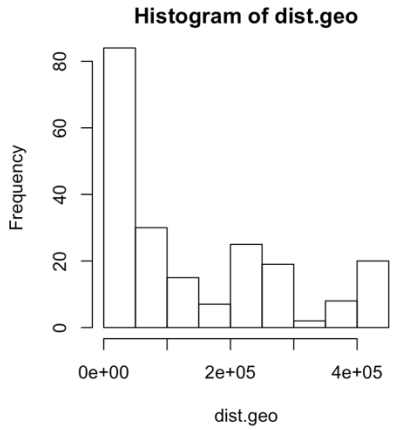


Figure S16: The non-normal frequency distribution of the geographic distance matrix.

Table S1: statistics for the four detected modules in the Baikal co-occurrence network. M1 and M3 have the highest transitivity (clustering coefficient) values.

Module	Average Path	Transitivity
1	1.626	0.684
2	1.797	0.540
3	1.410	0.823
4	1.795	0.583

REFERENCES

1. Kozhova OM, Izmest'eva LR. 1998. Lake Baikal : evolution and biodiversityBiology of inland waters. Backhuys Publishers.
2. Moore MVM, Hampton SES, Izmest'eva LLR, Silow E a., Peshkova E V., Pavlov BK. 2009. Climate Change and the World's "Sacred Sea"—Lake Baikal, Siberia. Bioscience 59:405–417.
3. Lincoln S a., Wai B, Eppley JM, Church MJ, Summons RE, DeLong EF. 2014. Reply to Schouten: Planktonic Euryarchaeota are a significant source of archaeal tetraether lipids in the ocean. Proc Natl Acad Sci 111:1409439111-.
4. Kozhov M. 1963. Lake Baikal and Its Life. Springer Netherlands, Dordrecht.
5. Crumpton W, Isenhart T, Mitchell P. 1992. Nitrate and organic N analyses with second-derivative spectroscopy. Limnol Ocean 37:907–9.
6. Welschmeyer N. 1994. Fluorometric analysis of chlorophyll a in the presence of chlorophyll b and pheopigments. Limnol Oceanogr 39:1985–1992.
7. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. Appl Environ Microbiol 79:5112–5120.
8. Schloss P, Westcott S. MiSeq SOP.
9. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–7.
10. Jost L. 2006. Entropy and diversity. Oikos 113:363–375.
11. Leinster T, Cobbold CA. 2012. Measuring diveristy: the importance of species similarity. Ecology 93:477–489.
12. Calcagno V, Mazancourt C De. 2010. glmulti : An R Package for Easy Automated Model Selection with (Generalized) Linear Models. J Stat Softw 34:1–29.
13. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2016. vegan: Community Ecology Package.
14. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217.
15. Cushman SA, Landguth EL. 2010. Spurious correlations and inference in landscape genetics. Mol Ecol 19:3592–3602.
16. Cushman SA, Wasserman TN, Landguth EL, Shirk AJ. 2013. Re-evaluating causal modeling with mantel tests in landscape genetics. Diversity 5:51–72.
17. Guillot G, Rousset F. 2013. Dismantling the Mantel tests. Methods Ecol Evol 4:336–344.
18. Shirk AJ, Landguth EL, Cushman SA. 2017. A comparison of regression methods for model selection in individual-based landscape genetic analysis. Mol Ecol Resour 55–67.
19. Friedman J, Alm EJ. 2012. Inferring correlation networks from genomic survey data. PLoS Comput Biol 8:e1002687.
20. Berry D, Widder S. 2014. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Front Microbiol 5:219.

21. Csárdi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal Complex Syst 1695:1695.
22. Fuhrman JA. 2009. Microbial community structure and its functional implications. Nature 459:193–199.
23. Brandes U, Delling D, Gaetler M. 2008. On Modularity Clustering. IEEE Trans Knowl Data Eng 20:172–188.
24. Newman M, Girvan M. 2004. Finding and evaluating community structure in networks. Phys Rev E 69:1–16.
25. Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.
26. Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685.
27. Nichols E, Uriarte M, Bunker DE, Favila ME, Slade EM, Vulinec K, Larsen T, Vaz-de-Mello FZ, Louzada J, Naeem S, Spector SH. 2013. Trait-dependent response of dung beetle populations to tropical forest conversion at local and regional scales. Ecology 94:180–189.
28. Montesinos-Navarro A, Segarra-Moragues JG, Valiente-Banuet A, Verdú M. 2012. Plant facilitation occurs between species differing in their associated arbuscular mycorrhizal fungi. New Phytol 196:835–44.
29. Bauer U, Clemente CJ, Renner T, Federle W. 2012. Form follows function: morphological diversification and alternative trapping strategies in carnivorous Nepenthes pitcher plants. J Evol Biol 25:90–102.
30. Moro MF, Silva IA, de Araújo FS, Nic Lughadha E, Meagher TR, Martins FR. 2015. The role of edaphic environment and climate in structuring phylogenetic pattern in seasonally dry tropical plant communities. PLoS One 10:e0119166.
31. Valiente-Banuet A, Verdú M. 2007. Facilitation can increase the phylogenetic diversity of plant communities. Ecol Lett 10:1029–36.
32. Verdú M, Pausas JG. 2007. Fire drives phylogenetic clustering in Mediterranean Basin woody plant communities. J Ecol 95:1316–1323.
33. Langille MGIM, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes J a J, Clemente JJC, Burkepile DE, Vega Thurber RRL, Knight R, Beiko RGR, Huttenhower C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31:814–21.
34. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610–8.
35. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072.
36. Zaneveld JR, Lozupone C, Gordon JI, Knight R. 2010. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. Nucleic Acids Res 38:3869–3879.