The evolution of biotic and abiotic realized niches within freshwater *Synechococcus*

Authors: Nicolas Tromas[1#], Mathieu Castelli[2#], Zofia E. Taranu[3], Juliana S. M. Pimentel[4], Daniel A. Pereira[4], Romane Marcoz[1], Alessandra Giani[4] and B. Jesse Shapiro[1*]

1- Département de sciences biologiques, Université de Montréal, 90 Vincent-d'Indy, Montréal, QC, Canada, Montréal, QC H2V 2S9, Canada

2- mathieu.castelli@gmail.com

3- Department of Biology, University of Ottawa, Gendron Hall, 30 Marie Curie, Ottawa

4- Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

*Corresponding authors: B. Jesse Shapiro. Phone: 514-343-6033. E-mail: jesse.shapiro@umontreal.ca; Nicolas Tromas. Phone 514-343-3188. E-mail: nicolas.tromas@umontreal.ca.

# These authors contributed equally to this work

Originality-Significance Statement

We address a fundamental question in ecology and evolution: how do niche preferences change over evolutionary time? Using time-series analysis of 16S rRNA gene amplicon sequencing data, we develop a new approach to highlight the importance of biotic factors in defining realized niches, and show how niche preferences change "clock-like" within the genus *Synechococcus.* Ours is also one of few studies on the ecology of freshwater *Synechococcus,* adding significantly to our knowledge about this abundant and widespread lineage of Cyanobacteria.

Summary

Understanding how ecological traits have changed over evolutionary time is a fundamental question in biology. When closely-related organisms share similar environmental preferences, "habitat filtering" is expected to determine how communities are assembled. Yet in practice, it is challenging to assess the impact of habitat filtering, due to our inability to measure all relevant abiotic variables and distinguish the impact of biotic versus abiotic factors. Here we explored the co-occurrence patterns of freshwater cyanobacteria at the sub-genus level to investigate whether closely-related taxa share similar niches, and to what extent these niches were defined by abiotic or biotic variables. We used deep 16S rRNA gene amplicon sequencing and measured several environmental parameters in water samples collected over time and space in Furnas Reservoir, Brazil. We found that closely-related *Synechococcus* did not have similar preferences for abiotic niche dimensions. However, closely-related *Synechococcus* did tend to co-occur with one another, and also with similar surrounding microbial communities. These results suggest that biotic factors may be stronger niche determinants than abiotic factors. Alternatively, cryptic abiotic drivers may determine niche and community structure, but biotic factors provide the most informative measure of niche similarity.

Introduction

A bacterial community is a group of potentially interacting organisms that coexist at a particular place and time (Magurran, 2003). Environmental selective pressures are a strong force shaping microbial community assembly (Martiny et al, 2015). We know, for example, that certain abiotic factors explain a large portion of the variation in microbial community composition (e.g. the effect of pH on soil bacterial communities; Fierer and Jackson, 2006). Therefore, associations between microbial traits - generally defined as a phenotypic response to a specific environmental condition - and abiotic niches could help explain community assembly rules (Green et al., 2008; Burke et al., 2011). However, trait-based approaches to understand communities may be challenging, as important abiotic variables may go unmeasured. Even less is known about biotic interactions, despite their importance in determining community composition, diversity, and dynamics (Needham et al., 2016).

Abiotic factors are generally thought to determine an organism's fundamental niche (where it is theoretically capable of living), whereas biotic factors determine its realized niche (where it actually lives in nature; Hutchinson, 1957). Species (or taxonomic) co-occurrence networks are often used to infer niche similarity among organisms that tend to co-occur in nature over space and time, and microbial co-occurrence networks can easily be constructed from 16S rRNA gene amplicon sequencing surveys (Friedman and Alm, 2012; Röttjers and Faust, 2018; Tromas et al., 2018). However, co-occurrence can be driven by both abiotic and biotic factors, which are hard to disentangle in practice (Kraft et al., 2015). Regardless, organisms sharing a similar niche are expected to be associated with similar surrounding communities (Cohan and Koeppel, 2008; Faust et al.2012; Pascual-García et al., 2014).

A fundamental question spanning ecology and evolution is how ecological traits change over evolutionary time. For example, some traits (such as bacteriophage host range) evolve rapidly at the tips of a phylogenetic tree, whereas other traits (such as salinity preference) are deeply conserved (Martiny *et al.* 2015). When closely related organisms share similar ecological preferences, so-called "habitat filtering" or "environmental" filtering is expected to result in phylogenetic clustering, meaning that a community tends to contain more closely related organisms than expected by a random draw from the phylogeny (Webb et al., 2002; Horner-Devine & Bohannan, 2006; Martiny et al., 2015). In contrast, if close relatives evolve different traits to avoid competitive exclusion, this will result in phylogenetic overdispersion (*i.e.* a community composed of more distant relatives than expected by chance). However, the relative importance of these two processes in shaping microbial communities is still widely debated, and difficult to distinguish (Koeppel and Wu, 2014; Cadotte and Tucker, 2017). We have previously shown that within the cyanobacterial genus *Dolichospermum* (Tromas et al., 2018), the relationships between phylogenetic distance and ecological similarity varies by trait, suggesting that it might be necessary to analyze each niche dimension or trait separately (Martiny *et al*., 2015).

In this study, we explored the co-occurrence patterns of freshwater cyanobacteria at the sub-genus level, to investigate if closely related taxa have more similar niches, and to what extent these niches can be quantified by abiotic or biotic variables. We focused on *Synechococcus*, the most abundant cyanobacterial genus in Furnas Reservoir (Brazil) at the time of sampling (2006-2008). *Synechococcus* is among the most abundant organisms living in oceans and lakes (Stockner et al. 2000; Scanlan, 2003). The phylogenetic coherence of the genus has been

100 questioned by a recent study showing it to be polyphyletic (Coutinho et al 2016). *Synechococcus*
101 is physiologically highly plastic, ubiquitous, and able to acclimate to different environmental
102 conditions (Callieri 1996; Vörös et al. 1998; Callieri et al., 2011). Previous studies have shown
103 that different *Synechococcus* strains could co-exist in the same site but respond differently to
104 environmental changes, suggesting niche partitioning (Ferris et al., 2003; Allewalt et al., 2006;
105 Becker et al., 2007; Becraft et al., 2011; Callieri et al. 2012). Recently, Zheng et al. (2018),
106 observed a geographical pattern of heterotrophic bacteria associated with different marine
107 *Synechococcus* strains, indicating that strains living in the same area tend to be associated with
108 similar communities. It remains, however, unclear whether the *Synechococcus* genotype or the
109 environment are the main drivers of *Synechococcus* interactions with the surrounding microbial
110 community.
111
112 Using deep 16S rRNA gene amplicon sequencing of 86 water samples collected in time series
113 across nine locations in the Furnas Reservoir, we tracked genetic diversity within the
114 *Synechococcus* genus, along with the surrounding microbial community, and measured several
115 abiotic variables. We found that that closely-related *Synechococcus* tended to co-occur with one
116 another and also with similar surrounding microbial communities. Such phylogenetic clustering
117 indicates that overall realized niche similarity tends to evolve "clock-like" in the *Synechococcus*
118 lineage. However, closely-related *Synechococcus* did not have similar abiotic niche preferences
119 (with the exception of total phosphorus). These results suggest that biotic factors may be stronger
120 niche determinants than abiotic factors. Alternatively, cryptic abiotic drivers may determine niche
121 and community structure, but biotic factors provide the most informative measure of niche
122 similarity.
123

Materials and methods

Sampling, environmental data measurements

A total of 90 water samples were collected from September 2006 to April 2008 at nine stations in Furnas Reservoir (Minas Gerais, Brazil). Furnas is a large reservoir (1440 km^2), located in southeastern Brazil (20o40'S; 46o19'W) and formed by the damming of two main rivers (Rio Grande and Rio Sapucaí), which divide the reservoir in two separated branches (Figure S1). Sampling stations S1, S4, S6, S9 are from a relatively pristine branch of the reservoir, whereas S12, S14, S18 and S20 are impacted by human activities. Temperature profiles were measured in the water column by aid of a multi-parameter probe (556 YSI, USA). Water samples were collected from the euphotic zone (determined by Secchi disc depth) by a Van Dorn sampler. Samples were stored in bottles that had been acid-washed and rinsed with deionized water. A portion of each water sample was immediately filtered through glass fiber filters (Whatmann GF/F, 0.7 μm pore size). The exact filtered volumes were recorded and filters were kept frozen until further analyses (chlorophyll, DNA and microcystin). For dissolved nutrient analyses, 200 mL of filtered water samples were stored at −20 °C. For total phosphorus, samples were frozen with no previous filtration. Nutrient analyses were performed using spectrophotometric methods according to APHA (2005). All nutrient analyses (Nitrate, Nitrite and total phosphorus (TP) were performed on three replicates. For phytoplankton analyses, samples were Lugol-preserved for subsequent cyanobacteria identification and quantification.

DNA extraction, purification and sequencing

DNA was extracted from frozen filters according to Kurmayer et al. (2003), with few modifications. Briefly, filters were treated with a sucrose buffer (25% w/v sucrose, 50 Mm Tris-HCl, 100 Mm EDTA, pH 8) on ice for 2 h and with addition of lysozyme (5mg/mL, 1h, 37°C).

Proteinase K (100 µg/mL) in sodium dodecyl sulfate (2% v/v) was added and filters were incubated overnight at 55°C. A phenol:chloroform:isoamyl alcohol solution (25:24:1, v/v) was used for protein precipitation and DNA isolation. The DNA was cleaned in 100% ethanol and pellets rinsed with 70% ethanol. The DNA was resuspended in TE (10 mM Tris- HCl, pH 8, and 1 mM EDTA). The DNA extract was quantified by a spectrophotometer, at 260 nm and 280 nm, and its quality checked in 1% (w/v) agarose gel, stained with ethidium bromide.

Sequence analysis

We followed the same protocol described in Tromas et al., (2017) to generate a library of V4 region amplicons. We performed one sequencing run using MiSeq reagent Kit V2 (Illumina, San Diego, CA, USA) on a MiSeq instrument (Illumina). A total of 4,476,747 sequences of the 16S rRNA gene V4 region were obtained from 90 lake samples, two negative controls, and two mock community samples. We obtained a median of 37,682 sequences per sample. Using a similar approach as described in Tromas *et al*., (2018), we processed the sequences with SmileTrain (https://github.com/almlab/SmileTrain/wiki; Preheim et al., 2013) for read quality filtering, primer removal, chimera filtering, and merging using USEARCH (version 7.0.1090, http://www.drive5.com/usearch/, default parameter) (Edgar, 2010), Mothur (version 1.33.3) (Schloss *et al*., 2009), and Biopython (version 2.7). Minimum Entropy Decomposition (MED) was then applied to the filtered and merged reads to partition sequence reads into MED nodes (Eren *et al*., 2015). MED was performed using the following parameters: –M noise filter set to 1000, resulting in ~17.5% of reads filtered and 466 MED nodes representing the whole bacterial community. Samples with less than 1000 reads were removed, yielding a final dataset of 86 reservoir samples. Finally, we assigned taxonomy to MED nodes using the assign_taxonomy.py QIIME script (default parameters), and a combination of GreenGenes and a freshwater-specific

173 database (Freshwater database 2016 August 18 release; Newton *et al*., 2011), using TaxAss
174 (https://github.com/McMahonLab/TaxAss, installation date: September 13th 2016; Rohwer *et al*.,
175 2017).
176
177 ***Spatio-temporal analysis***
178 We used multivariate regression tree analyses (Breiman *et al*. 1984; De'ath 2002) to
179 investigate if spatio-temporal variables could explain genetic variation within *Synechococcus*. We
180 used two different temporal predictors: year and month and one spatial predictor: station. The
181 analysis was performed using the function *mvpart*() and *rpart.pca*() from the R mvpart package
182 (Therneau and Atkinson, 1997; De'ath, 2007). Prior to analysis, the *Synechococcus* MED nodes
183 table was Hellinger transformed to downweight the effect of double-zeros (Rao, 1995).
184
185 ***Genetic distance between Synechococcus nodes***
186 We measured the genetic distance between *Synechococcus* MED nodes using the software
187 MEGA (Kumar et al., 2016; version 7.0.18) with the p-distance (the proportion of nucleotide
188 sites at which two sequences differ), calculated by dividing the number of sites with nucleotide
189 differences by the total number of sites compared (excluding sites with gaps).
190
191 ***Co-response to abiotic factors***
192 As described in Tromas *et al.* (2018), we used a Latent Variable Model (LVM)
193 framework (boral package in R; Hui 2015, Warton *et al.* 2015) to explore how *Synechococcus*
194 nodes co-responded to abiotic gradients and used these co-responses as indicators of niche
195 similarity. That is, for each abiotic factor, we ran separate LVMs, regressing the bacterial
196 community as a function (both linear and non-linear) of the given factor. This component of the

LVM thus defined the taxon-specific environmental responses. To then identify remaining patterns of co-occurrence after accounting for all measured environmental variables, we fit a global LVM which included all abiotic factors and two latent variables (sensu Letten *et al.* 2015 and Warton *et al.* 2015). To visualize patterns of co-occurrence arising from the different environmental factors, we calculated two types of correlation matrices. The first, a co-response correlation matrix, was constructed by calculating, for any two nodes, the correlation between their fitted values. This correlation matrix thus represented the correlation between nodes that can be attributed to a shared or diverging environmental responses. A significant positive correlations between the fitted response of any two taxa to an environmental variable represented a co-response (i.e., as one taxa increases in response to an environmental variable, the other likewise increases), whereas a significant negative correlations between the fitted response of two taxa represented some degree of niche separation (i.e., as one taxa increases in response to an environmental variable, the other decreases).

To account for the correlation between nodes that may be attributable to biotic processes or missing environmental covariates, a second type of correlation matrix, a residual correlation matrix, was calculated using the latent variable coefficients of the global LVM. Since the latent variables are the output of an ordination of the residuals, we expected weak residual correlations induced by the latent variables if the environmental variables were the dominant force structuring patterns of species co-occurrence (*i.e. t*he model explained most of the variance, with little left unexplained in the residuals). Conversely, if any unmeasured environmental factors and/or biotic processes are equally, or more, important than measured environmental factors, we expected strong correlations based on the latent variables (*i.e. t*he model was poor and much of the observed variability remained in the residuals).

To test how *Synechococcus* niche similarity varied with genetic distance, we examined the relationship between the co-response of *Synechococcus* taxa to environmental parameters and their genetic distance (i.e., plotting the correlation coefficient of the LVM co-responses vs. genetic distances) (R_script1). However, given the large number of environmental factors driving phytoplankton community dynamics (Hutchinson, 1961), we expected some degree of co-limitation whereby the response of a taxon to an environmental variable (and consequently the degree of co-response among any two taxa) would be limited by other environmental variables. In such cases, the taxon's response and rate of change would have an upper limit (set by all measured environmental factors) but may not reach this limit if other, unmeasured factors are co-limiting (Cade and Noon 2003). As an increasing number of unmeasured factors become limiting at some sample location, or time point, the relationship between the response and the measured factor becomes increasingly heterogeneous or wedge shaped. When such heterogeneous variances are observed (wedge shape biplot), it suggests that there is not a single slope coefficient that characterizes the relationship, and that focusing solely the coefficient fit by ordinary least squares regression (mean response) may underestimate the true rate of change. Thus, to accurately examine co-limitations among measured and unmeasured factors, we applied quantile regressions using 'quantreg' R package (Koenker, 2015) (R_script1), which, instead of fitting a regression to the mean response (ordinary least squares regression), fits regression curves to other quantiles of the response variable (Cade and Noon 2003).

Co-occurrence with biotic factors

A caveat of the above LVM-quantile regressions is that although it identifies whether some relationships are limited by unmeasured abiotic and/or biotic factors, it could not tease apart the relative importance of each type of factor (everything is instead lumped as an unknown). In

order to determine the relative contribution of biotic processes, we thus quantified node co-occurrence patterns among *Synechococcus* and the remaining community. In particular, we explored the relationship between *Synechococcus* nodes and the surrounding community by measuring: (1) co-occurrences between *Synechococcus* and other taxa, and (2) paired differences in co-occurrence among *Synechococcus* nodes and a specific taxon. For the former, we calculated co-occurrences among taxa using SparCC (Friedman and Alm, 2012), including 20 iterations to estimate the median correlation of each pair of MED nodes, 500 bootstraps to assess the statistical significance and centered log ratio (CLR) transform to correct for compositionality. Correlations were then filtered using a false discovery rate threshold ($Q < 0.05$). For the latter, we further calculated the absolute difference of Sparcc correlation (r) between a *Synechococcus* node (X_i) and a specific taxon T, such that $|\Delta r| = |\ Corr(X_1, T) - Corr(X_2, T)\ |$ where Corr is defined here as the SparCC correlation between X_i and T.

For each non-*Synechococcus* taxon, we then estimated the relationship between $|\Delta r|$ and the distance between the given *Synechococcus* nodes (X_1 and X_2). A positive correlation would indicate that closely related *Synechococcus* nodes have a lower $|\Delta r|$ than more distant nodes, *i.e.* closely related nodes would have more similar correlations with potentially interacting community members. To reduce potential bias or noise due to a small sample size of *Synechococcus* pairs, we selected all non-*Synechococcus* nodes that were significantly co-occurrent with at least seven different *Synechococcus*. This cutoff was chosen to ensure at least 20 unique pairs of *Synechococcus* for each non-*Synechococcus* taxon. A total of 373 non-*Synechococcus* taxa were selected using this cutoff, and the correlation between $|\Delta r|$ and genetic distance was non-significant (Spearman correlation, $P<0.05$) for 120 of them. Finally, we performed a permutation test to quantify any bias in the method due to data structure. For each of

the significant non-*Synechococcus* taxa, we estimated the rate of false positive correlations by sampling values of *Synechococcus* genetic distance with replacement, while randomizing the association between genetic distance and $|\Delta r|$. We performed 1000 permutations for each significant non-*Synechococcus* taxon using the script R_script2. We then recorded the proportion (p) of permutations yielding a larger correlation than observed, after adding a pseudocount of 1 to both the numerator (the number of permutations yielding a larger correlation than observed) and the denominator (the total number of permutations).

Phylogenetic analysis

We first verified that *Synechococcus* group was monophyletic by building a phylogenetic tree with FastTree (version 2.1.8, Price et al., 2009), using all MED node sequences aligned with MAFFT (v7.154b; Katoh and Standley, 2013). We then specifically aligned *Synechococcus* MED node sequences using muscle (in Mega, version 7.0.26; Kumar et al., 2016*)* and built phylogenetic trees using PhyML (version 3.0; Guindon et al., 2010). We used the ALDEx2 R package (version: 1.5.0; Fernandes *et al*., 2014) and the aldex() function to identify MED nodes associated with the pristine or human-impacted branches of Furnas Reservoir. To do so, we used Welsh's t-test and 128 Monte Carlo samples. ALDEx2 uses the CLR transformation to avoid compositional effects. A Q-value below 0.05 after Benjamini-Hochberg correction was taken as evidence for association. We also used a second association method, DESeq2 implemented in the "MicrobiomeAnalyst" web-based tool (Dhariwal et al., 2017).

Statistical analysis

All statistical tests were performed in R.

Results

Abiotic factors shape genetic variation within Synechococcus

Furnas Reservoir is divided into two branches: one, which is heavily human-impacted (Figure S1, sampling stations 12-20) and the other, less impacted (stations 1-9). *Synechococcus* dominated the bacterial community in both branches during the sampling period (2006-2008) (Figure S2) and formed a monophyletic group (Figure S3, S4). We first asked whether particular *Synechococcus* strains (which we used interchangeably with MED nodes) were associated with different branches of the reservoir, different time periods, or different abiotic factors. We applied a multivariate regression tree analysis (Breiman et al. 1984; De'ath 2002) to test whether the composition of *Synechococcus* strains changed over time and space. We found that the year and month of sampling did not explain much variance (R^2=0.02 and R^2=0.03, respectively), nor did the station (R^2=0.06) or branch of the reservoir (R^2=0.02). In contrast, Chl *a*, Nitrite, TP, Depth, TC, PTC, and water temperature all explained a significant proportion of the variance (Figure S5; 40% of variance explained), and depth, temperature and nitrite alone explained most of the variability (Figure S6; 16% unique, and an additional 6% co-explained with Chl *a*).

Closely related Synechoccus do not share similar preferences for measured abiotic niches

Having established that abiotic factors, but not temporal or spatial variation, explain a significant amount of variation in the relative abundances of *Synechococcus* strains, we asked whether closely related strains tended to share similar abiotic niches. Here we defined abiotic niche similarity by estimating the co-responses of pairs of *Synechococcus* strains to each abiotic factor using LVM-quantile regression (Methods). For most abiotic factors, we did not observe any significant relationship between co-responses and genetic distance, with the exception of total

phosphorus for which the relationship was negative (Figure 1, S7). Similarly, the degree of abiotic niche separation (*i.e.* significant negative correlations between the fitted response of any two *Synechococcus* to an abiotic factor) was largely independent of genetic distance (Figure S8, Figure S9).

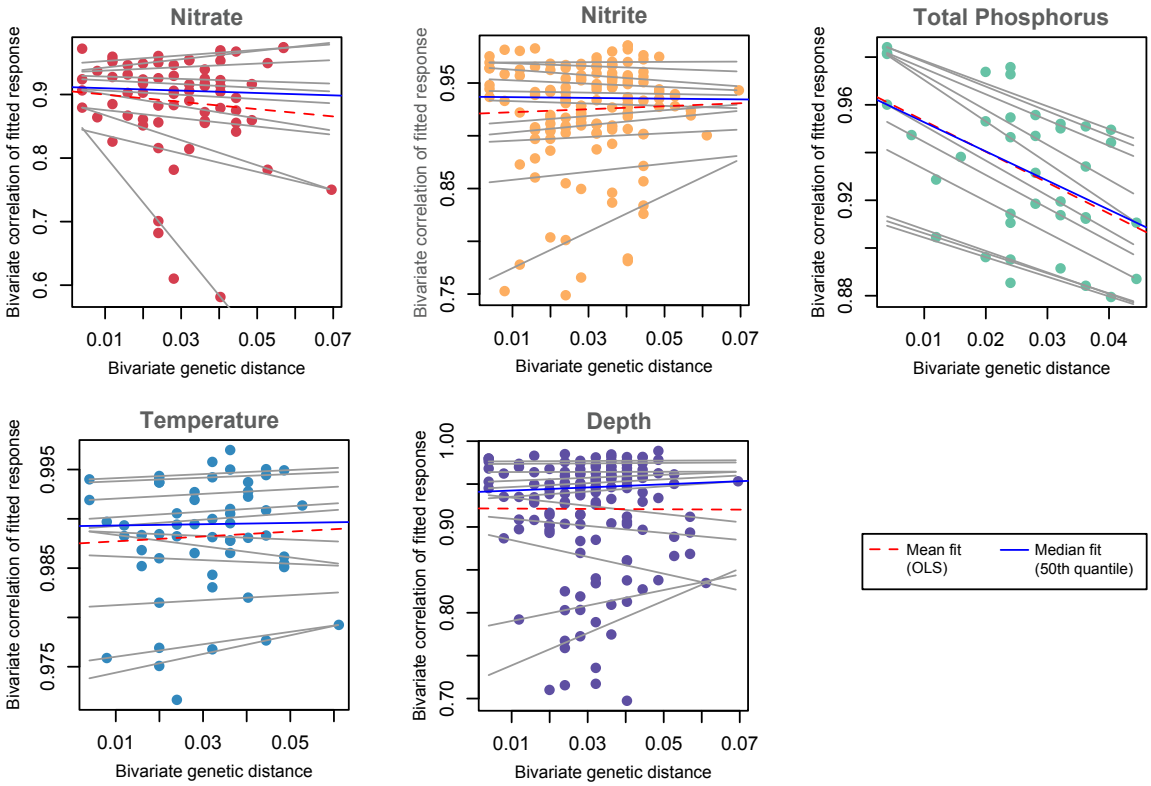


Figure 1. Relationship between *Synechococcus* niche similarity and genetic distance. Niche similarity is defined as a positive LVM co-response (y-axis). Shown are the scatterplots of LVM co-responses versus each variables superimposed with the lines for 5th, 10th, 15th, 25th, 35th, 45th, 55th, 65th, 75th, 85th and 90th quantile regression fits (grey lines, for each co-response quantile), the median fit (50th quantile; blue line), and the least squares estimate or mean fit (dashed red line).

To take into account the correlation between nodes that may be attributable to biotic processes or missing environmental covariates, we used the latent variables of the global LVM to examine

how the correlation among node residuals varied with genetic distance. For the positive correlations (Figure 2A, B), the correlation coefficients tended to decrease with genetic distance. This result suggests a negative relationship between the residuals (*i.e.* unmeasured abiotic variables and/or biotic processes) and phylogenetic distance. For the negative residual correlations (Figure 2C, D), there were no apparent relationships with the genetic distance. Therefore, there remain significant co-responses that were not explained by the abiotic factors measured in this study. Overall, most of the measured abiotic environmental variables were not related to *Synechococcus* genetic distance. However, preferences for unmeasured biotic or abiotic factors do tend to change with genetic distance, such that more closely-related *Synechococcus* have similar realized niche preferences.

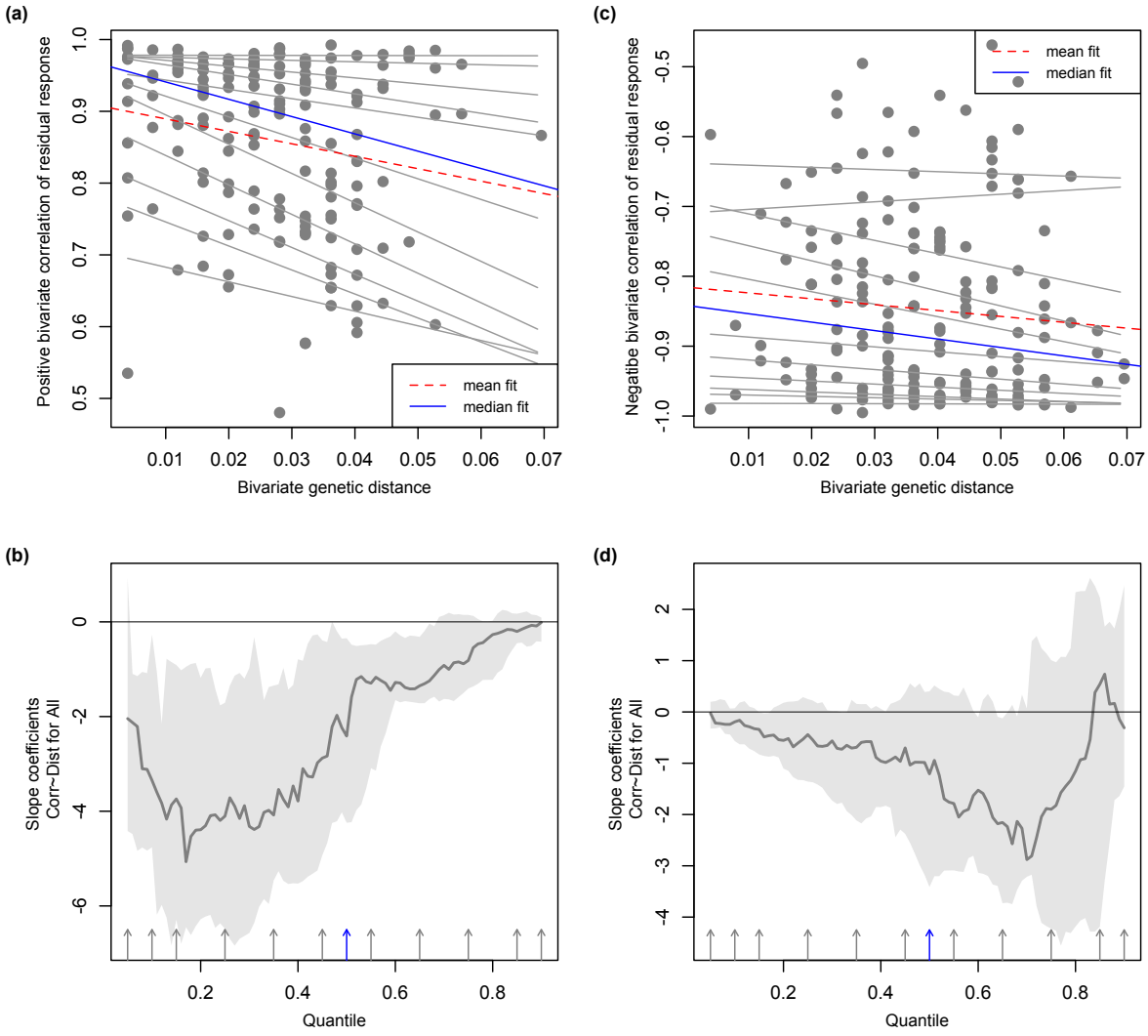
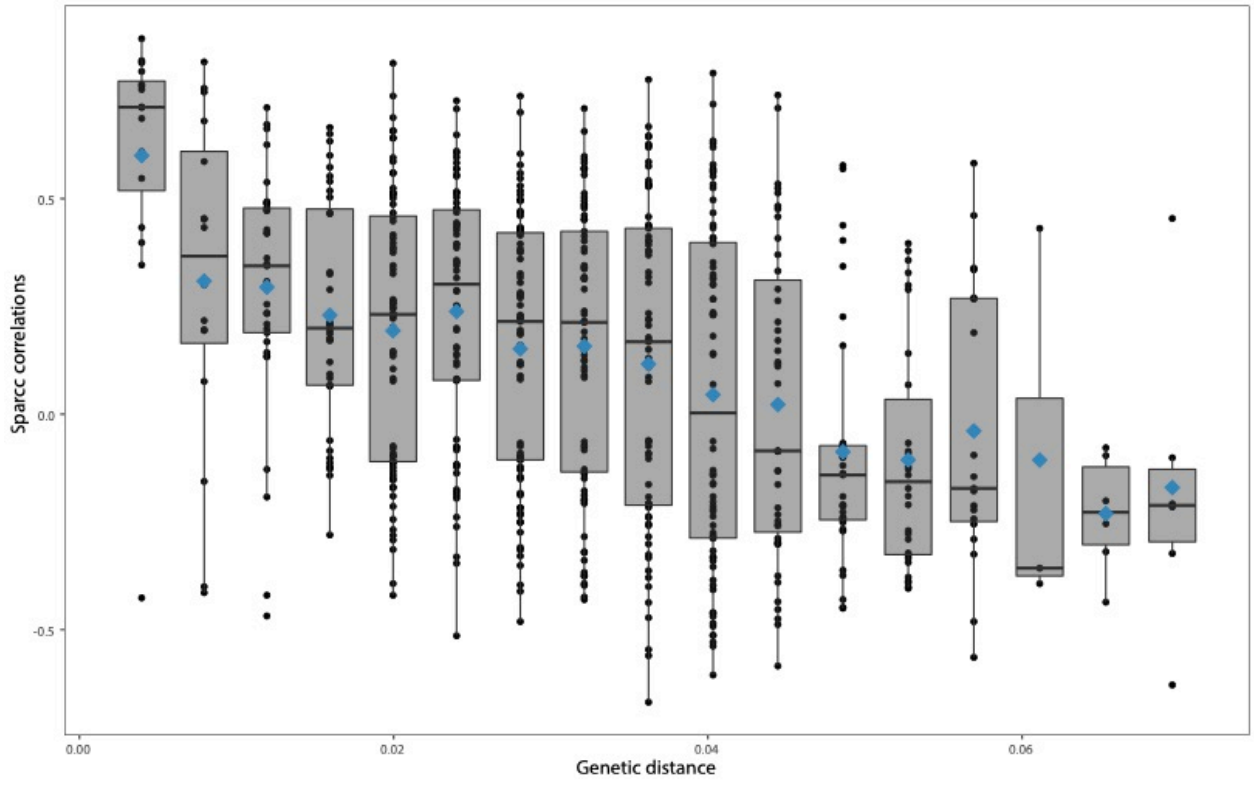


Figure 2. Relationship between *Synechococcus* co-responses (LVM residuals) and genetic distance. Positive co-responses are shown in panels **A** and **B**; negative co-resposes in panels **C** and **D.** *(*A, C*)* Scatterplots of LVM co-responses versus LVM residuals superimposed with the lines for 5th, 10th, 15th, 25th, 35th, 45th, 55th, 65th, 75th, 85th and 90th quantile regression fits (grey lines), the median fit (50thquantile; blue line), and the least squares estimate or mean fit (dashed red line). (B, D) Panels show the quantile slope estimates (dark grey line) and corresponding confidence intervals (light grey bands) of the response models across all quantiles (arrows, with the median shown in blue). Significant quantile slopes occur when confidence intervals do not overlap zero.

Closely related Synechococcus share similar biotic interactions

356 The observed tendency for closely-related *Synechococcus* to share similar realized niches (Figure
357 2A) could be explained by shared preferences for either unmeasured abiotic factors or biotic
358 interactions. We used co-occurrence network analysis to determine the role of biotic interactions.
359 First, we asked if closely related *Synechococcus* tend to co-occur across samples. We observed a
360 negative relationship between node co-occurrence and pairwise genetic distance (Figure 3),
361 indicating that genetically similar *Synechococcus* nodes are indeed more likely to co-occur, and
362 thus to have similar realized niches. This pattern was significant [linear regression, $F(1,488) = 108.6$, $P < 0.001$, adjusted $R^2 = 0.18$].
363
364



365

366 **Figure 3**. **Negative relationship between pairwise SparCC correlation coefficients and**
367 **pairwise genetic distance between *Synechococcus* nodes.** Only significant SparCC correlations
368 ($Q < 0.05$) were included. Pairwise genetic distance was computed as percent nucleotide identity
369 (p-distance). Blue diamonds represent the mean SparCC correlation for each distance. Boxplots

show the median (horizontal line), the 25th and 75th percentile (enclosed in box) and 95% confidence intervals (whiskers).

We further investigated whether more closely-related *Synechococcus* nodes have more similar associations with surrounding community members. To do so, we performed a pairwise analysis to examine the relationship between *Synechococcus* genetic distance and $|\Delta r|$, a measure of the similarity of *Synechococcus* co-occurrence with non-*Synechococcus* taxa (Methods). A total of 373 non-*Synechococcus* taxa were analysed and the correlation between $|\Delta r|$ and genetic distance was non-significant for 120 of them. The remaining 253 significant non-*Synechococcus* have consistently positive relationships between co-occurrence and genetic distance, *i.e.* closely related *Synechococcus* nodes have a higher chance of co-occurring with similar non-*Synechococcus* taxa (Figure 4). This result is robust to data structure, as determined by permuting the associations between genetic distance and $|\Delta r|$ (Figure S10; Methods).

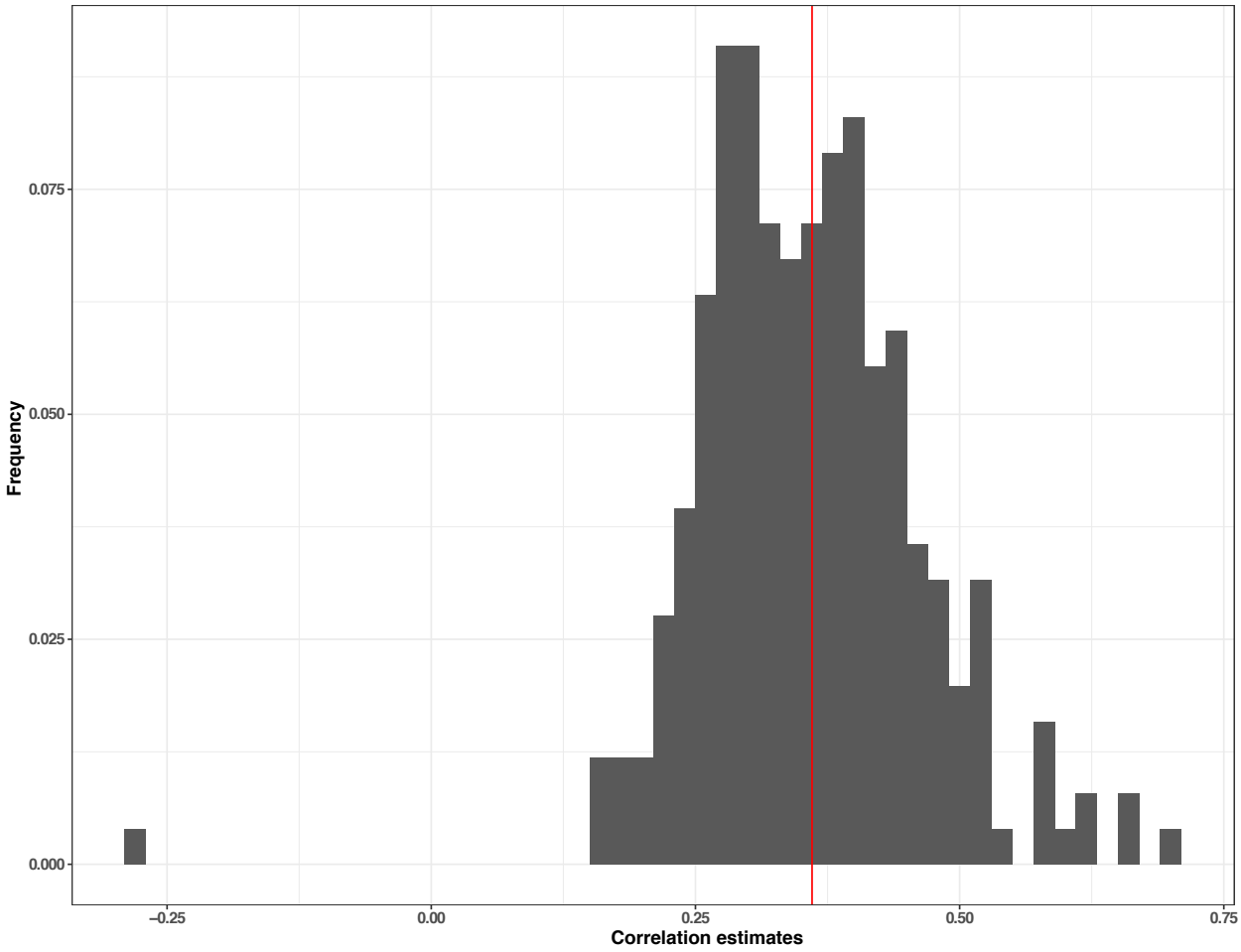


Figure 4. Closely related *Synechococcus* nodes co-occur with similar surrounding community. The histogram shows the distribution of Spearman correlations between the genetic distance of *Synechococcus* nodes and their association with similar surrounding communities. Spearman correlations were calculated between the genetic distance of *Synechococcus* pairs and the absolute difference of Sparcc correlation (r) of a *Synechococcus* node with a specific taxon T, such that $|\Delta r| = |\ Corr(X_1, T) - Corr(X_2, T)\ |$ where Corr is defined here as the SparCC correlation score of X_i (*Synechococcus* node) and T (non-*Synechococcus* node). We then estimated the correlation of $|\Delta r|$ and the genetic distance between nodes X_1 and X_2. The mean Spearman correlation estimate was 0.373 (red line) and 99.6% of correlations are positive.

Finally, we selected the non-*Synechococcus* taxa with the highest correlation scores between |Δr| and *Synechococcus* distance genetic. Five taxa had correlation estimates higher than 0.6, including three Proteobacteria, one Verrucomicrobia, and one Planctomycetes (Table S1). Previous studies have shown that members of the *Comamonadaceae* family and *Spirobacillales* order (members of Proteobacteria; Table S1) were associated with *Synechococcus* (Guedes et al., 2018) or more broadly with algae (Fullbright et al., 2019). Moreover, the *Gallionella* genus was previously observed colonizing filamentous algae (Mori et al., 2015).

Overall, these results show that the co-occurring surrounding community is, on average, more similar for closely related *Synechococcus*. Members of this surrounding community could interact directly with *Synechococcus,* or share similarity in their abiotic niche.

Discussion

In this study, we investigated *Synechococcus* niche preferences at relatively fine (sub-genus) taxonomic resolution using a 16S gene sequence variants approach previously used to study other bacterial lineages (Koeppel and Wu, 2014; Tromas et al., 2018). We found that abiotic factors, such as reservoir depth and nitrite concentration, greatly affected the relative abundance of *Synechococcus* strains (Figure S5, S6). However, the measured abiotic factors were not related to genetic distance, indicating that these traits are probably loosely conserved and may evolve on shorter time scales, being thus consistent with phylogenetic overdispersion. An exception was total phosphorus preference, an abiotic niche dimension that was correlated with genetic distance, consistent with phylogenetic clustering. Phosphorus use may therefore evolve '"clock-like" along the *Synechococcus* phylogeny. This contrasts with the idea that organic

422 phosphate niches are fast-evolving and relatively uncorrelated with phylogeny, perhaps due to
423 horizontal transfer of relevant genes (Coleman & Chisholm 2010; Martiny et al., 2015).
424 Therefore, how phosphorus-related traits evolve along phylogenies likely depend on the lineage
425 considered (*e.g. Synechococcus* vs. *Prochlorococcus* or *Pelagibacter*) and evolutionary time
426 scales. We note that only 77 of the 780 *Synechococcus* strain pairs showed a significant co-
427 response to phosphorus (Figure 1), indicating that the effect is driven by a small number of
428 closely-related strains with similar co-responses, and cannot be generalized to all *Synechococcus*.
429 This result might be explained if *Synechococcus* represents a polyphyletic group (Coutinho et al.,
430 2016), composed of different clades with different phosphorus niches. However, in our sample,
431 *Synechococcus* is monophyletic (Figure S3, S4), excluding this explanation. Overall, our results
432 demonstrate how abiotic factors can shape community composition within a genus, but measured
433 abiotic niches generally do not evolve along the genus-level phylogeny.
434
435 Habitat filtering was originally defined to describe how abiotic factors select for genetically
436 similar (closely related) organisms (Cadotte and Tucker 2017). Yet in practice, it is difficult to
437 disentangle the contributions of biotic and abiotic factors, and thus to satisfactorily define habitat
438 filtering (Kraft et al., 2015). In this study, we therefore estimated the influence of both
439 unmeasured abiotic variables and biotic factors, and found a negative association with genetic
440 distance (Figure 2). This suggests that closely related *Synechococcus* tend to share similar
441 realized niches. As a result, closely related *Synechococcus* tend to co-occur (Figure 3), as
442 observed in studies using genomic (rather than 16S) similarity (Kamneva, 2017). This result is
443 consistent with habitat filtering, broadly defined to include both biotic and abiotic factors.
444
445 We suggest that biotic niches are the major drivers of habitat filtering and the prime determinants

446 of realized niches in *Synechococcus*. This is because closely-related *Synechococcus* tend to co-
447 occur with more similar surrounding communities (Figure 4), whereas they do not tend to share
448 abiotic niche preferences (Figure 1). Of course, it is possible that unmeasured abiotic factors are
449 driving the observed biotic effects. However, we measured abiotic factors such as temperature
450 and nutrients, which are important in structuring microbial communities, and none of these
451 factors are sufficient to account for realized niches or their distribution across the *Synechococcus*
452 phylogeny. Thus, as previously suggested (Cohan and Koeppel, 2008), biotic factors (quantified
453 here with 16S amplicon sequencing) provide more information about niches than abiotic factors,
454 which are difficult to measure comprehensively. For example, the presence of certain taxa could
455 indicate the importance of an unmeasured abiotic factor (*e.g* Acidophiles or Alkaliphiles as
456 indicators of pH). In this case, considering biotic factors allowed us to uncover a signal of habitat
457 filtering within *Synechococcus* that would have been obscured by only considering commonly
458 measured abiotic factors.
459

References

Allewalt, J.P., Bateson, M.M., Revsbech, N.P., Slack, K., and Ward, D.M. (2006) Effect of Temperature and Light on Growth of and Photosynthesis by Synechococcus Isolates Typical of Those Predominating in the Octopus Spring Microbial Mat Community of Yellowstone National Park. *Appl Environ Microbiol* **72**: 544–550.

Becker, S., Richl, P., and Ernst, A. (2007) Seasonal and habitat-related distribution pattern of Synechococcus genotypes in Lake Constance. *FEMS Microbiol Ecol* **62**: 64–77.

Becraft, E.D., Cohan, F.M., Kühl, M., Jensen, S.I., and Ward, D.M. (2011) Fine-Scale Distribution Patterns of Synechococcus Ecological Diversity in Microbial Mats of Mushroom Spring, Yellowstone National Park. *Appl Environ Microbiol* **77**: 7689–7697.

Breiman L, Friedman JH, Olshen RA, Stone CJ. (1984) Classification and Regression Trees. Wadsworth International Group, Belmont, CA, USA.

Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S., and Thomas, T. (2011) Bacterial community assembly based on functional genes rather than species. *Proc Natl Acad Sci USA* **108**: 14288–14293.

Cade, B.S. and Noon, B.R. (2003) A Gentle Introduction to Quantile Regression for Ecologists. *Frontiers in Ecology and the Environment* **1**: 412–420.

Cadotte, M.W. and Tucker, C.M. (2017) Should Environmental Filtering be Abandoned? *Trends in Ecology & Evolution* **32**: 429–437.

Callieri, C. (1996) Extinction coefficient of red, green and blue light and its influence on picocyanobacterial types in lakes at different trophic levels. *Memorie dell'Istituto Italiano di Idrobiologia* 54: 135-142.

Callieri, C., Caravati, E., Corno, G., and Bertoni, R. (2012) Picocyanobacterial community structure and space-time dynamics in the subalpine Lake Maggiore (N. Italy). *Journal of Limnology* **71**: e9–e9.

Callieri, C., Lami, A., and Bertoni, R. (2011) Microcolony Formation by Single-Cell Synechococcus Strains as a Fast Response to UV Radiation. *Appl Environ Microbiol* **77**: 7533–7540.

Cohan, F.M. and Koeppel, A.F. (2008) The origins of ecological diversity in prokaryotes. *Curr Biol* **18**: R1024-1034.

Coleman, M. L., and Chisholm, S. W. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc. Natl. Acad. Sci. U. S. A.* **107** : 18634–18639.

Coutinho, F., Tschoeke, D.A., Thompson, F., and Thompson, C. (2016) Comparative genomics

of Synechococcus and proposal of the new genus Parasynechococcus. *PeerJ* **4**: e1522.

De'ath G. (2007) mvpart: Multivariate partitioning. R package version 1.6-2.

Dhariwal, A., Chong, J., Habib, S., King, I.L., Agellon, L.B., and Xia, J. (2017) MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res* **45**: W180–W188.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Eren, A.M., Morrison, H.G., Lescault, P.J., Reveillaud, J., Vineis, J.H., and Sogin, M.L. (2015) Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* **9**: 968–979.

Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012) Microbial Co-occurrence Relationships in the Human Microbiome. *PLOS Computational Biology* **8**: e1002606.

Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., and Gloor, G.B. (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**: 15.

Ferris, M.J., Kühl, M., Wieland, A., and Ward, D.M. (2003) Cyanobacterial Ecotypes in Different Optical Microenvironments of a 68°C Hot Spring Mat Community Revealed by 16S-23S rRNA Internal Transcribed Spacer Region Variation. *Appl Environ Microbiol* **69**: 2893–2898.

Fierer, N. and Jackson, R.B. (2006) The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* **103**: 626–631.

Friedman, J. and Alm, E.J. (2012) Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology* **8**: e1002687.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

Green, J.L., Bohannan, B.J.M., and Whitaker, R.J. (2008) Microbial Biogeography: From Taxonomy to Traits. *Science* **320**: 1039–1043.

Horner-Devine, M.C. and Bohannan, B.J.M. (2006) Phylogenetic Clustering and Overdispersion in Bacterial Communities. *Ecology* **87**: S100–S108.

Hui, F.K.C., Taskinen, S., Pledger, S., Foster, S.D., and Warton, D.I. (2015) Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution* **6**: 399–411.

Hutchinson, G.E. (1957) Concluding Remarks. *Cold Spring Harb Symp Quant Biol* **22**: 415–427.

Hutchinson, G.E. (1961) The Paradox of the Plankton. *The American Naturalist* **95**: 137–145.

Kamneva, O.K. (2017) Genome composition and phylogeny of microbes predict their co-occurrence in the environment. *PLOS Computational Biology* **13**: e1005366.

Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30**: 772–780.

Koeppel, A.F. and Wu, M. (2014) Species matter: the role of competition in the assembly of congeneric bacteria. *ISME J* **8**: 531–540.

Koenker, R. (2015). Package 'quantreg'

Kraft, N. J. B., Adler, P. B., Godoy, O., James, E. C., Fuller, S., and Levine, J. M. (2015). Community assembly, coexistence and the environmental filtering metaphor. Funct. Ecol. **29**, 592–599.

Kumar, S., Stecher, G., and Tamura, K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**: 1870–1874.

Kurmayer, R., Christiansen, G., and Chorus, I. (2003) The Abundance of Microcystin-Producing Genotypes Correlates Positively with Colony Size in Microcystis sp. and Determines Its Microcystin Net Production in Lake Wannsee. *Appl Environ Microbiol* **69**: 787–795.

Letten, A.D., Keith, D.A., Tozer, M.G., and Hui, F.K.C. (2015) Fine-scale hydrological niche differentiation through the lens of multi-species co-occurrence models. *Journal of Ecology* ***103***: 1264–1275.

Magurran, A.E. and Henderson, P.A. (2003) Explaining the excess of rare species in natural species abundance distributions. *Nature* **422**: 714.

Martiny, J.B.H., Jones, S.E., Lennon, J.T., and Martiny, A.C. (2015) Microbiomes in light of traits: A phylogenetic perspective. *Science* **350**: aac9323.

Needham, D. M., Fuhrman, J. A. (2016) Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature Microbiology*, **1**: 16005.

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. *Microbiol Mol Biol Rev* **75**: 14–49.

Pascual-García, A., Tamames, J., and Bastolla, U. (2014) Bacteria dialog with Santa Rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions? *BMC Microbiol* **14**: 284.

Preheim, S.P., Perrotta, A.R., Martin-Platero, A.M., Gupta, A., and Alm, E.J. (2013) Distribution-based clustering: using ecology to refine the operational taxonomic unit. *Appl Environ Microbiol* **79**: 6593–6603.

Price MN, Dehal PS, Arkin AP. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**:1641–1650.

Rohwer, R.R., Hamilton, J.J., Newton, R.J., and McMahon, K.D. (2018) TaxAss: Leveraging a Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution. *mSphere* **3**: e00327-18.

Röttjers, L. and Faust, K. (2018) From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiol Rev* **42**: 761–780.

Scanlan, D.J. (2003) Physiological diversity and niche adaptation in marine Synechococcus. In, *Advances in Microbial Physiology*. Academic Press, pp. 1–64.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., et al. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl Environ Microbiol* **75**: 7537–7541.

Stockner, J., Callieri, C., and Cronberg, G. (2002) Picoplankton and Other Non-Bloom-Forming Cyanobacteria in Lakes. In, Whitton,B.A. and Potts,M. (eds), *The Ecology of Cyanobacteria: Their Diversity in Time and Space*. Dordrecht: Springer Netherlands, pp. 195–231.

Tromas, N., Taranu, Z.E., Martin, B.D., Willis, A., Fortin, N., Greer, C.W., and Shapiro, B.J. (2018) Niche Separation Increases With Genetic Distance Among Bloom-Forming Cyanobacteria. *Front Microbiol* **9** : 438

Vörös, L., Callieri, C., Balogh, K.V., and Bertoni, R. (1998) Freshwater picocyanobacteria along a trophic gradient and light quality range. *Hydrobiologia* **369**: 117–125.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., and Hui, F.K.C. (2015) So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution* **30**: 766–779.

Webb, C.O., Ackerly, D.D., McPeek, M.A., and Donoghue, M.J. (2002) Phylogenies and Community Ecology. *Annual Review of Ecology and Systematics* **33**: 475–505.

Zheng, Q., Wang, Y., Xie, R., Lang, A.S., Liu, Y., Lu, J., et al. (2018) Dynamics of Heterotrophic Bacterial Assemblages within Synechococcus Cultures. *Appl Environ Microbiol* **84**: e01517-17.

Data availability

Raw sequence data have been deposited NCBI GenBank under BioProject number PRJNA544938.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments

NT is funded by a project from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 656647. BJS was supported by a Canada Research Chair and NSERC Discovery Grant. Sampling and samples processing were supported by a grant from Furnas Centrais Hidroelétrica S.A. (Brazil) to AG. JSMP and DAP received a scholarship from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). A sabbatical visit of AG to the Université de Montréal AG was supported by a CAPES Senior Fellowship.