Using machine learning to associate bacterial taxa with functional groups through flow cytometry, 16S rRNA gene sequencing, and productivity data

Peter Rubbens[1*], Marian L. Schmidt[2*#], Ruben Props[3], Bopaiah A. Biddanda[4], Nico Boon[3], Willem Waegeman[1], Vincent J. Denef[2]

[*]Peter Rubbens and Marian L. Schmidt contributed equally to this work.

[#]Corresponding author: marschmi@umich.edu

[1]KERMIT, Department of Data Analysis and Mathematical Modelling, Ghent University, Coupure Links 653, B-9000, Gent, Belgium; [2]Department of Ecology and Evolutionary Biology, University of Michigan, 830 North University Ave., Ann Arbor, MI 48109, USA; [3]CMET, Center for Microbial Ecology and Technology, Ghent University, Coupure Links 653, B-9000, Gent, Belgium; [4]Annis Water Resources Institute, Grand Valley State University, 740 West Shoreline Drive, Muskegon, MI 49441, USA

Abstract

High- (HNA) and low-nucleic acid (LNA) bacteria are two separated flow cytometry (FCM) groups that are ubiquitous across aquatic systems. HNA cell density often correlates strongly with heterotrophic production. However, the taxonomic composition of bacterial taxa within HNA and LNA groups remains mostly unresolved. Here, we associated freshwater bacterial taxa with HNA and LNA groups by integrating FCM and 16S rRNA gene sequencing using a machine learning-based variable selection approach. There was a strong association between bacterial heterotrophic production and HNA cell abundances ($R^2 = 0.65$), but not with more abundant LNA cells, suggesting that the smaller pool of HNA bacteria may play a disproportionately large role in the freshwater carbon flux. Variables selected by the models were able to predict HNA and LNA cell abundances at all taxonomic levels, with highest accuracy at the OTU level. There was high system specificity as the selected OTUs were mostly unique to each lake ecosystem and some OTUs were selected for both groups or were rare. Our approach allows for the association of OTUs with FCM functional groups and thus the identification of putative indicators of heterotrophic activity in aquatic systems, an approach that can be generalized to other ecosystems and functioning of interest.

Introduction

A key goal in the field of microbial ecology is to understand the relationship between microbial diversity and ecosystem functioning. However, it is challenging to associate bacterial taxa to specific ecosystem processes. Marker gene surveys have shown that natural bacterial communities are extremely diverse, however, the presence of a taxon does not imply their activity. Taxa present in these surveys may have low metabolic potential, be dormant, or have recently died [1, 2]. Therefore, new methodologies which integrate different data types are needed to associate bacterial taxa with ecosystem functions in order to ultimately model and predict them [3].

One such advance is the use of flow cytometry (FCM), which has been used extensively to study aquatic microbial communities [4–6]. This single-cell technology partitions individual microbial cells into phenotypic groups based on their observable optical characteristics. Most commonly, cells are stained with a nucleic acid stain (*e.g.* SYBR Green I) and upon analysis assigned to either a low nucleic acid (LNA) or a high nucleic acid (HNA) group [7–10]. HNA cells differ from LNA cells in both a considerable increase in fluorescence due to cellular nucleic acid content and scatter intensity due to cell morphology. The HNA group is thought to correspond to the 'active' fraction, whereas the LNA population has been considered as the 'dormant' or 'inactive' group of a microbial community [4, 11–13]. This is based on positive linear relationships between HNA abundance and (a) bacterial heterotrophic production (BP) [8, 12–15], (b) bacterial activity measured using the dye 5-cyano-2,3-ditolyl tetrazolium chloride [16, 17], and (c) phytoplankton abundance [18]. Additionally, growth rates are higher for HNA than

LNA cells [11, 14, 19] and HNA cells accrue cell damage significantly faster than the LNA cells under temperature [20] and chemical oxidant stress [21].

One main research question that still remains is whether HNA and LNA groups are composed of unique taxa or if they are different physiological states of the same taxa. Bouvier et al. [9] proposed four possible scenarios: (1) bacteria start their life cycle in the HNA group and move to the LNA group upon death or inactivity; (2) cells in the HNA group originate from LNA cells undergoing cell division; (3) HNA and LNA consist of different non-overlapping taxa; (4) bacteria switch between groups from time to time in addition to having part of the community that is unique to each fraction. The view that HNA cells are more active is in line with scenario 1 and 2. On the other hand, several studies have found distinct groups with little taxonomic overlap and proposed scenario 3 [22, 23] or 3 and 4 [24]. In this case, HNA and LNA groups have been associated with different life strategies in bacterioplankton communities, such as large cell size (HNA) versus small cell size (LNA) [13, 23], genome size [15] and ploidy [22]. By combining FCM with taxonomic identification of bacterial communities, one can associate individual taxa with population dynamics and functioning.

In this study, we developed a novel approach to associate the dynamics of individual taxa with those of the LNA and HNA groups in freshwater lakes by using a machine learning variable selection strategy. We applied two variable selection methods, the Randomized Lasso [25] and the Boruta algorithm [26] to associate individual taxa with HNA and LNA cell abundances. This approach allowed us to associate specific taxa to FCM functional groups, and via the observed HNA-productivity relationship, to functioning. In addition, this approach enabled us to test the

77 influence of rare taxa on these two groups as recent research has found that rare taxa may have a
78 strong impact on community structure and functioning [27, 28]. To validate the RL-based
79 association with the HNA and/or LNA group, we correlated taxon abundances with specific
80 regions in the FCM fingerprint without prior knowledge of the HNA/LNA group. Furthermore,
81 we tested for phylogenetic conservation of HNA and LNA functional groups and for the
82 association between the selected taxa and productivity. The combination of FCM and 16S rRNA
83 gene sequencing allows for the inference and assessment of the taxonomic structure of HNA and
84 LNA groups, therefore advancing our ability to link bacterial taxa to their functionality in nature.
85 This knowledge will help identify the taxa that drive carbon fluxes in freshwater ecosystems,
86 which are disproportionately large relative to the global freshwater surface area [29].

87 **Results**

88 In this study, we developed a machine learning variable selection strategy to integrate FCM and
89 16S rRNA gene sequencing with the aim of inferring the bacterial drivers of functional groups in
90 freshwater lake systems. We studied a set of oligo- to eutrophic small inland lakes, a short
91 residence time mesotrophic freshwater estuary lake (Muskegon Lake), and a large oligotrophic
92 Great Lake (Lake Michigan), all located in Michigan, USA. We showed that abundance variation
93 of these FCM functional groups is predicted by a small subset of all taxa that are present in the
94 environment. Selected taxa were mostly FCM groups and lake system specific, and across
95 systems, association with HNA or LNA was not phylogenetically conserved. The relationship
96 between selected taxa and productivity measurements was assessed for one of the lake systems
97 (Muskegon Lake), thereby showing that HNA cells (and their putative bacterial taxa) likely turn
98 over faster and disproportionately contribute to the freshwater carbon flux.
99

Study lakes are dominated by LNA cells

The inland lakes (6.3 x 10^6 cells/mL) and Muskegon Lake (6.0 x 10^6 cell/mL) had significantly higher total cell abundances than Lake Michigan (1.7 x 10^6 cell/mL; p = 2.7 x 10^{-14}). Across all lakes, the mean proportion of HNA cell counts (HNAcc) to total cell counts was much lower (29-33%) compared to the mean proportion of LNA cell counts (LNAcc; 67-71%). Through ordinary least squares regression, there was a strong correlation between HNAcc and LNAcc across all data (R^2 = 0.45, P = 2 x 10^{-24}; **Figure 1A**), however, only Lake Michigan (R^2 = 0.59, P = 5 x 10^{-11}) and Muskegon Lake (R^2 = 0.44, P= 2 x 10^{-9}) had significant correlations when the three ecosystems were considered separately.

HNA cell counts and heterotrophic bacterial production are strongly correlated

For mesotrophic Muskegon Lake, there was a strong correlation between total bacterial heterotrophic production and HNAcc (R^2 = 0.65, p = 1e-05; **Figure 1B**), no correlation between BP and LNAcc (R^2 = 0.005, p = 0.31; **Figure 1C**), and a weak correlation between heterotrophic production and total cell counts (R^2 = 0.18, p = 0.03; **Figure 1D**). There was a positive (HNA) and negative (LNA) correlation between the fraction of HNA or LNA to total cells and productivity, however, the relationship was weak and not significant (R^2 = 0.14, p = 0.057).

Association of OTUs to functional groups by Randomized Lasso regression

The relevance of specific OTUs for predicting freshwater FCM functional group abundance was assessed using the Randomized Lasso (RL) approach, which assigns a score between 0 (unimportant) to 1 (highly important) to each taxon in function of the target variable: HNAcc or LNAcc. This score can be interpreted as the probability that an OTU will be included in the

Lasso model to predict HNA or LNA cell abundances. Variations of HNAcc and LNAcc were modelled in function of relative changes of OTUs. To address the negative correlation bias intrinsic to compositional data, compositions were first transformed using a centered log-ratio (CLR) transformation.

The RL score was used to implement a recursive variable elimination scheme. Specifically, we iteratively removed the lowest-ranked OTUs based on the RL score (*i.e.* OTUs were ranked according to the score from high to low) and the Lasso was fitted to the data to predict HNAcc and LNAcc based on the corresponding subset of OTUs. The performance was expressed in terms of the R^2_{CV}, the R^2 between predicted and true values of HNAcc and LNAcc of samples that were held-out using a leave-one-group-out cross-validation scheme, in which samples were grouped according to year and location of measurement. If R^2_{CV} equals 1, predictions were equal to the true values, a value of 0 is equivalent to random guessing.

There was taxonomic dependency for both HNAcc and LNAcc across lake systems (**Figure 2**). R^2_{CV} increased when lower-ranked OTUs were removed (moving from right to left on **Figure 2**), which was gradual for the inland lakes (**Figure 2A**) and Muskegon Lake (**Figure 2C**) but was abrupt for Lake Michigan (**Figure 2B**). The number of taxa that resulted in the highest R^2_{CV} contained less than a quarter of the total amount of taxa that were present (*see solid (HNA) and dotted (LNA) lines in* **Figure 2**), being 10.2% HNA and 15.3% LNA for the inland lakes, 4.0% HNA and 3.0% LNA for Lake Michigan, and 25.0% for both HNA and LNA in Muskegon Lake. This behavior was consistent for each lake system and FCM population. The Lake Michigan results differed the most from other lake systems, having the lowest R^2_{CV}, a sharp increase in R^2_{CV}

instead of gradual, and a considerably lower minimal amount of OTUs (13 for HNAcc, 10 for LNAcc). No relationship could be established between rankings of variable selection methods and the relative abundance of individual OTUs (**Figure S1**). Multiple taxa with low average abundance were included in the minimal set of predictive variables, whereas few highly abundant OTUs were included. HNAcc and LNAcc could be predicted with equivalent performance to relative HNA and LNA proportions, yet the increase between initial and optimal performance was bigger (**Figure S2**). The final predictive performance was lower when compositional data was not transformed using the CLR-transformation (**Figure S3**).

Identification on different taxonomic levels: OTUs outperform all other taxonomic levels

To assess whether HNA and LNA groups were taxonomically conserved, compositional data was analyzed on all possible taxonomic levels for Muskegon Lake (**Figure 3**), using the same strategy as outlined in previous paragraph. The resulting R^2_{CV} values were considerably higher than zero on all taxonomic levels, meaning that at all levels individual taxonomic changes can be related to changes in HNAcc and LNAcc. Even though the OTU level resulted in the best prediction of HNAcc and LNAcc (**Figure 3**), each individual OTU has a lower RL score compared to other taxonomic levels, which on average became lower as the taxonomic level decreased (**Figure S4**). The fraction of variables (taxa) that could be removed to reach the maximum R^2_{CV} decreased as the taxonomic level became less resolved.

Validation of OTU selection results with the Boruta algorithm

The OTU results were validated with an additional variable selection strategy, called the Boruta algorithm. This approach allowed the further generalization of the findings presented above. In

addition, it connects with Random Forest results from other studies, which have been described recently in microbiome studies of other systems *(see [30] and [31])*. The Boruta algorithm selects relevant variables based on statistical hypothesis testing between the variable importance of an original variable and the importance of the most important permuted variable (*see materials and methods*), as retrieved from multiple Random Forest models. Selected variables are ranked as '1', tentative variables as '2', and all other variables get lower ranks, depending on the stage in which they were eliminated. The Boruta algorithm was applied for all three lake systems at the OTU-level, selected OTUs are visualized in **Figure S5**. The fraction of selected OTUs was always smaller than 1% across lake systems and functional groups (**Figure S6)**. The top scored OTU according to the RL was also selected according to the Boruta algorithm for HNAcc for all lake systems; for LNAcc both methods only agreed for Lake Michigan (**Table 1**). OTU060 (Proteobacteria;Sphingomonadales;alfIV_unclassified) was the only OTU selected in function of LNAcc across all lake systems, whereas no OTUs were selected across lake systems for HNAcc. As Random Forest regressions are the base method of the Boruta algorithm, we compared the predictive power of Boruta selected OTUs to those of all OTUs using Random Forest regression. For all lake systems and functional group performance increased when only selected OTUs were included in the model (**Table S1)**. Lasso predictions, in which OTUs were selected according to the RL, were better as opposed to Random Forest predictions in which OTUs were selected according to the Boruta algorithm (**Figure S7**). The fraction of selected OTUs according to the Boruta algorithm was lower than the optimal amount of OTUs according to the RL.

In this way, a number of findings could be generalized independent of a specific method: 1) Selected OTUs were mostly lake systems specific, 2) a small fraction of OTUs was needed to predict changes in community composition, 3) selected OTUs are often rare and do not show a relationship with abundance and 4) top RL-ranked HNA OTUs were also selected according to the Boruta algorithm, suggesting to inspect more closely the phylogeny of these taxa.

HNA- and LNA-associated OTUs differed across lake systems

Selected OTUs were mostly assigned to either the HNA or LNA groups and there was limited correspondence across lake systems between the selected OTUs (**Figure 4**). In Muskegon Lake, OTU173 (Bacteroidetes;Flavobacteriales;bacII-A) was selected as the major HNA-associated taxon while OTU29 (Bacteroidetes;Cytophagales;bacIII-B) had the highest RL score for LNA OTUs. In Lake Michigan, OTU25 (Bacteroidetes;Cytophagales;bacIII-A), was selected as the major HNA-associated taxon while OTU168 (Alphaproteobacteria:Rhizobiales:alfVII) was selected as a major LNA-associated taxon. For the inland lakes, OTU369 (Alphaproteobacterial;Rhodospirillales;alfVIII) was the major HNA-associated OTU while the OTU555 (Deltaproteobacteria;Bdellovibrionaceae;OM27) was the major LNA-associated taxon. Many more OTUs were selected in Muskegon Lake (197 OTUs; compared to 134 OTUs from the Inland Lakes and 21 OTUs from Lake Michigan) and these OTUs were often associated with both HNA and LNA groups.

RL scores were correlated for HNAcc and LNA within each lake system (Inland $r = 0.25$, $P < 0.001$; Michigan $r = 0.59$, $P < 0.001$, Muskegon $r = 0.59$, $P < 0.001$). Only OTUs that were present in all three freshwater environments were considered to calculate correlations between

lake systems (190 in total, **Figure S8**). RL scores were lake ecosystem specific, with only a significant similarity between the Inland lakes and Muskegon lake using the RL for HNAcc ($r = 0.21$, $P = 0.0042$). Note that the correlation within a lake system therefore differs from previously reported values (as not all OTUs were considered), yet differences were small and results were comparable. The Boruta algorithm selected mostly OTUs which were unique both for the lake system and functional population (**Figure S5**).

Selected HNA and LNA OTUs do not have a phylogenetic signal

While many of the 258 OTUs selected by the RL were one of a few members of their phylum (*e.g.* Firmicutes; Epsilonproteobacteria; OTU717 in Lentisphaerae; OTU267 in Omnitrophica; etc), the Bacteroidetes (60 OTUs), Betaproteobacteria (36 OTUs), Alphaproteobacteria (22 OTUs), and Verrucomicrobia (21 OTUs) were a total of 54% of the selected OTUs (**Figure 5**). Of these top four phyla, the majority of their membership were within the LNA group (41-52% of selected OTUs), with the minority of OTUs within the HNA group (14-30% of selected OTUs), and a quarter to a third of the OTUs were selected as members of both the LNA and HNA groups (23-36% of selected OTUs).

To evaluate how much phylogenetic history explains whether a selected taxon was associated with the HNA and/or LNA group(s), we calculated the phylogenetic signal, which is a measure of the dependence among species' trait values on their phylogenetic history [32]. If the phylogenetic signal is very strong, taxa belonging to similar phylogenetic groups (*e.g.* a Phylum) will share the same trait (*i.e.* association with HNAcc or LNAcc). Alternatively, if the phylogenetic signal is weak, taxa within a similar phylogenetic group will have different traits.

For the most part, Pagel's lambda was used [33] to test for phylogenetic signal where lambda varies between 0 and 1. A lambda value of 1 indicates complete phylogenetic patterning whereas a lambda of 0 indicates no phylogenetic patterning and leads to a tree collapsing into a single polytomy. There was no phylogenetic signal with FCM functional group used as a discrete character (*i.e.* HNA, LNA, or Both). As a continuous character using the RL scores for HNA (**Figure S9**), there was also no phylogenetic signal (lambda = 0.16; P = 1). There was a significant LNA signal (p = 0.003), however, the lambda value was 0.66, suggesting weak phylogenetic structuring in the LNA group. However, this significant result in the LNA was not replicated with other measures of phylogenetic signal (Blomberg's K (HNA: p = 0.63; LNA: p = 0.54), and Moran's I (HNA: p = 0.88; LNA: p = 0.12)) indicating that there is likely no phylogenetic signal in the taxa that drive the dynamics in either the HNA or the LNA group.

Flow cytometry fingerprints confirm associated taxa and reveal complex relationships between taxonomy and flow cytometric fingerprints

To confirm the association of the final selected OTUs with the HNA and LNA groups, we calculated the correlation between the density of individual regions (i.e. "bins") in the flow cytometry data with the relative abundances of the OTUs. The Kendall rank correlation coefficient between OTU abundances and counts in the flow cytometry fingerprint was calculated for each of the top HNA OTUs selected by the RL within each of the three systems. The correlation coefficient was visualized for each bin in the flow cytometry fingerprint (**Figure 6**). As these values denote correlations, they do not indicate actual presence. OTU25 correlated with almost the entire HNA region, whereas OTU173 was limited to the lower part of the HNA region. In contrast, OTU369 was positively correlated to both the LNA and HNA regions of the

cytometric fingerprint, highlighting results from **Figure 4** where OTU369 was selected in function of both HNA and an LNA. The threshold that was used to define HNAcc and LNAcc lies very close to the actual corresponding regions.

Proteobacteria and rare taxa correlate with productivity measurements

The Kendall rank correlation coefficient was calculated between CLR-transformed abundances of individual OTUs and productivity measurements. OTU481 was significantly correlated after correction for multiple hypothesis testing using the Benjamini-Hochberg procedure ($P < 0.001$, $P\_adj = 0.016$). This OTU had however a low RL score (0.022) and was not selected according to the Boruta algorithm. Of the top 10 OTUs according to the RL, three still had significant P-values (OTU614: $P = 0.0064$; OTU412, $P = 0.044$; OTU487, $P = 0.014$). Some OTUs that had a high RL score also had a positive response to productivity measurements (**Figure S10)**. At the phylum level, only Proteobacteria were significantly correlated to productivity measurements after Benjamini-Hochberg correction ($P < 0.001$, $P\_adj = 0.010$).

Discussion

Our study introduces a novel computational workflow to investigate relationships between microbial diversity and ecosystem functioning. Specifically, we aimed to study the ecology of flow cytometric functional groups (i.e. HNA and LNA) by associating their dynamics with those of bacterial taxa (i.e. OTUs). We simultaneously collected flow cytometry and 16S rRNA gene sequencing data from three types of freshwater lake systems in the Great Lakes region, and bacterial heterotrophic productivity from one lake ecosystem, and used a machine learning based variable selection strategy, known as the Randomized Lasso, to associate one with another. Our results showed that (1) there was a strong correlation between bacterial heterotrophic

283 productivity and HNA cell abundances, (2) HNA and LNA cell abundances were best predicted
284 by a small subset of OTUs that were unique to each lake type, (3) some OTUs were included in
285 the best model for both HNA and LNA abundance, (4) there was no phylogenetic conservation
286 of HNA and LNA group association and (5) freshwater FCM fingerprints display more complex
287 patterns related to OTUs and productivity than compared to the traditional dichotomy of HNA
288 and LNA. While HNA and LNA groups are universal across aquatic ecosystems, our data
289 suggest that some bacterial taxa contribute to both HNA and LNA groups and that the taxa
290 driving HNA and LNA abundance are unique to each lake system.
291
292 Although high-nucleic acid cell counts (HNAcc) and low-nucleic acid cell counts (LNAcc) were
293 correlated with each other, only the association between bacterial heterotrophic production (BP)
294 and HNAcc was strong and significant. This correlation between BP and HNA is higher than
295 previously reported values, though previous reports have focused on the proportion of HNA
296 rather than absolute cell abundances with the majority of data collected from marine systems.
297 For example, Bouvier et al. [9] found a correlation between the fraction of HNA cells and BP
298 within a large dataset of 640 samples across various freshwater to marine samples ($r = 0.49$),
299 whereas a study off the coast of the Antarctic Peninsula found a moderate correlation ($R^2 = 0.36$;
300 [15]). Another study in the Bay of Biscay also found this association ($R^2 = 0.16$; [13]), however,
301 the authors attributed this difference to be related to cell size and not due to the activity of HNA.
302 Notably, these studies were predominantly testing the association of marine HNA and the reason
303 for the stronger correlation in our study may be due to the nature of the freshwater samples. As
304 such, future studies in freshwater environments should test this hypothesis, which is especially
305 important for understanding the broader influence that HNA bacteria may have in the context of

306 the disproportionately large role that freshwater systems play as hotspots in the global carbon
307 cycle [29]. Finally, as our correlations with proportional HNA abundance also indicated less
308 strong correlations than with absolute HNAcc, we suggest absolute HNAcc should be used to
309 best predict heterotrophic bacterial production with FCM data.
310
311 The use of machine learning methods, such as the Lasso and Random Forest, are becoming more
312 common in microbiome literature as these approaches are able to deal with multi-dimensional
313 data and test the predictive power of a combined set of variables ([34–36]. Although the Lasso
314 already uses an intrinsic variable selection strategy, it has been noted that the Lasso method is
315 not suited for compositional data because the regression coefficients have an unclear
316 interpretation, and single variables may be selected when correlated to other variables [37].
317 When performing variable selection with Random Forests, traditional variable importance
318 measures such as the mean decrease in accuracy can be biased towards correlated variables [38].
319 Our approach included algorithms which extended on these traditional machine learning
320 algorithms, i.e. the Randomized Lasso or Boruta algorithm [25, 26]. These methods make use of
321 resampling and randomization which allow to either assign a probability of selection (RL) or
322 statistically decide which OTU to select (Boruta). Both the RL and Boruta algorithm have been
323 applied to microbiome studies before. Examples for RL include the selection of genera in the gut
324 microbiome relation to BMI [34] or the selection of OTUs from the oral microbiome in function
325 of salivary pH and lysozyme activity [39]. The Boruta algorithm has been applied to select
326 relevant genera, for example in the gut microbiome in relation to multiple sclerosis [31] or in
327 function of different diets during pregnancy of primates [30]. Moreover, the Boruta algorithm
328 has been recently proposed as one of the top-performing variable selection methods that make

use of Random Forests [40]. The ability of our approach to identify unique sets of OTUs predictive of HNAcc and LNAcc despite the correlation between HNAcc and LNAcc (**Figure 1A**) illustrates the power of the machine learning based-variable selection methods. However, there is still room for improvement when attempting to integrate these different types of data. For example, 16S rRNA gene sequencing still faces the hurdles of DNA extraction [41] and 16S copy number bias [42]. Moreover, detection limits are different for FCM (expressed in the number of cells) and 16S rRNA gene sequencing (expressed in the number of gene counts or relative abundance), which create data that may be different in resolution. Future work may focus on developing ways around these shortcomings to further improve the integration of FCM with 16S rRNA gene sequencing.

In our study, only a minority of OTUs was needed to predict specific flow cytometric group abundances. While each OTU individually had low predictive power, the selected group of OTUs was generally a strong predictor of HNAcc and LNAcc. In addition, the selected OTUs were often rare and thus no relationship could be established between the RL score and the abundance of an OTU (**Figure S3**). These results are in line with recent findings of Herren & McMahon [28], who reported that a minority of low abundance taxa explained temporal compositional changes of microbial communities. The selection of different sets of HNA and LNA OTUs across the three freshwater systems indicates that different taxa underlie the universally observed HNA and LNA functional groups across aquatic systems. This is in line with strong species sorting in lake systems [43, 44], shaping community composition through diverging environmental conditions between the lake systems presented here [45]. This high system specificity also explains the low RL scores for individual OTUs, as the spatial dynamics

of an OTU diverged strongly across systems. (For example, an OTU that has an RL score of 0.5 implies that on average it will only be chosen one out of two times in a Lasso model).

Based on the high correlation of BP with HNAcc and low correlation with BP and LNAcc, the high proportion of LNA cells across all lake systems might indicate that the majority of cells in the bacterial community are dormant or have very low activity. This agreest with previous research showing that up to 40% [46] or even 64-95% [47] of cells in freshwater systems to be inactive or dormant. In fact, up to 60-80% of the OTUs in freshwater lakes have been reported to be dormant [48]. Based on variable environmental conditions sampled across our dataset, some of the taxa that are predominantly dormant in one sample may contribute to activity in another sample. If this differing contribution to activity also covaries with a taxon's abundance, these taxa may be considered to be 'conditionally rare taxa' [49] and previously 1-2% of freshwater lake OTUs have been reported to be conditionally rare [27]. It has also been shown that marine heterotrophic bacteria can survive for at least 8 months (maximum tested length) in a starved state [50]. These factors may explain why some OTUs were included in both the HNAcc and LNAcc models and is in line with scenario 1 from Bouvier et al [9] (*i.e.* the transitioning of cells from active growth to death or inactivity). Alternatively, the same OTU may occur in both HNA and LNA groups due to phenotypic plasticity. Phenotypic plasticity has been shown for bacterial morphology and size, for example during predation and carbon starvation [51]. The fact that HNA and LNA groups have been suggested to correspond to cells of differing size, with HNA harboring larger cell sizes [10, 23], is in line with this hypothesis. Finally, the OTU level grouping of bacterial taxa can disguise genomic and phenotypic heterogeneity [52–55], which may be an explanation for inconsistent associations between OTUs and FCM functional groups.

While all taxonomic levels resulted in a model with predictive power, the best model was at the most resolved taxonomy (*i.e.* OTU) indicating that it is unlikely that OTUs within the HNA and LNA groups are phylogenetically conserved. Indeed, when analyzing the data at an OTU level, very little phylogenetic conservation was found between selected OTUs for HNA and LNA groups. This is in contrast to a recent study that found a clear signal at the phylum level [23]. Proctor et al. [23] showed separate bacterial clusters between HNA and LNA groups across different aquatic systems. However, this was not the case for lake water samples. It is notable that Proctor et al. [23] separated HNA and LNA cells based on cell size (where HNA cells were >0.4 um and LNA cells were 0.2-0.4 um, based on 50-90% removal of HNA cells after filtering), while our study separated these FCM functional groups on the basis of fluorescence intensity alone. Moreover, our study assessed associations between OTUs and population dynamics, while Proctor et al. [23] assessed actual presence.

The Boruta algorithm and RL scores agreed on the top-ranked HNA OTU for all lake systems, which motivates further investigation of the ecology of these OTUs. While little information on the identities of HNA and LNA freshwater lake bacterial taxa exists, several studies identified Bacteroidetes among the most prominent HNA taxa and is in line with our findings. Vila-Costa et al. [24] found that the HNA group was dominated by Bacteroidetes in summer samples from the Mediterranean Sea, Read et al. [17] showed that HNA abundances correlated with Bacteroidetes, and Schattenhofer et al. [22] reported that the Bacteroidetes accounted for the majority of HNA cells in the North Atlantic Ocean. In Muskegon Lake, OTU173 was the dominant HNA taxon and is a member of the Order *Flavobacteriales* (bacII-A). The bacII group

398 is a very abundant freshwater bacterial group and has been associated with senescence and
399 decline of an intense algal bloom [56]. BacII-A has also made up ~10% of the total microbial
400 community during cyanobacterial blooms, reaching its maximum density immediately following
401 the bloom [57]. In Lake Michigan, OTU25, a member of the Bacteroidetes Order *Cytophagales*
402 known as bacIII-A, was the top HNA OTU. However, much less is known about this specific
403 group of Bacteroidetes. Though, the bacII-A/bacIII-A group has been strongly associated with
404 more heterotrophically productive headwater sites (compared to higher order streams) from the
405 River Thames, showing a negative correlation in rivers with dendritic distance from the
406 headwaters, indicating that these taxa may contribute more to productivity [17]. In the inland
407 lakes, OTU369 was the major HNA taxon and is associated with the Alphaproteobacteria Order
408 Rhodospirillales (alfVIII), which to our knowledge is a group with very little information
409 available in the literature. In contrast to our findings of Bacteroidetes and Alphaproteobacterial
410 HNA selected OTUs, Tada & Suzuki [58] found that the major HNA taxon from an oceanic algal
411 culture was from the Betaproteobacteria whereas LNA OTUs were within the Actinobacteria
412 phylum.
413

414 **Conclusions**

415 Our results indicate that there are taxonomic differences between HNA and LNA groups in
416 freshwater lake systems, though these are lake system specific. This result may be due to taxa
417 switching between these groups, potentially due to genomic or phenotypic plasticity. The
418 difference between selected taxa is larger between lake systems as opposed to differences
419 between HNA and LNA groups, which were not conserved phylogenetically. Thus, our results
420 correspond most with research presented by Vila-Costa et al. [24], in which a taxonomic division

421 was found between HNA and LNA groups, yet this was not rigid and followed seasonal trends.
422 Overall, our results motivate scenario 4 proposed by Bouvier et al. [9], where HNA and LNA
423 exhibit a different taxonomy, but this taxonomy changes over time and space and may overlap.
424 With this study, we show that different types of microbial ecological data can be integrated with
425 machine learning to learn about the composition and functioning of bacterial populations in
426 aquatic systems. Future studies on HNA and LNA bacterial groups should use genome-resolved
427 metagenomics, metatranscriptomics, or single-cell genomics to decipher whether the traits that
428 underpin the association of a taxon with a FCM group are related to genomic or phenotypic
429 plasticity.
430

431 **Materials and Methods**

432 ***Data collection and DNA extraction, sequencing and processing***

433 In this study, we used a total of 173 samples collected from three types of lake systems described
434 previously [45], including: (1) 49 samples from Lake Michigan (2013 & 2015), (2) 62 samples
435 from Muskegon Lake (2013-2015; one of Lake Michigan's estuaries), and (3) 62 samples from
436 twelve inland lakes in Southeastern Michigan (2014-2015). For more details on sampling, please
437 see **Figure 1** and the *Field Sampling, DNA extraction, and DNA sequencing and processing*
438 sections within Chiang et al. [45]. In all cases, water for microbial biomass samples were
439 collected and poured through a 210 μm and 20 μm bleach sterilized nitex mesh and sequential in-
440 line filtration was performed using 47 mm polycarbonate in-line filter holders (Pall Corporation,
441 Ann Arbor, MI, USA) and an E/S portable peristaltic pump with an easy-load L/S pump head
442 (Masterflex®, Cole Parmer Instrument Company, Vernon Hills, IL, USA) to filter first through a
443 3 μm isopore polycarbonate (TSTP, 47 mm diameter, Millipore, Billerica, MA, USA) and

444 second through a 0.22 µm Express Plus polyethersulfone membrane filters (47 mm diameter,
445 Millipore, MA, USA). The current study only utilized the 3 - 0.22 µm fraction for analyses.
446
447 DNA extractions and sequencing were performed as described in Chiang et al. [45]. Fastq files
448 were submitted to NCBI sequence read archive under BioProject accession number
449 PRJNA412984 and PRJNA414423. We analyzed the sequence data using MOTHUR V.1.38.0
450 (seed = 777; [59] based on the MiSeq standard operating procedure and put together at the
451 following link: https://github.com/rprops/Mothur_oligo_batch. A combination of the Silva
452 Database (release 123; [60]) and the freshwater TaxAss 16S rRNA database and pipeline [61]
453 was used for classification of operational taxonomic units (OTUs).
454
455 For the taxonomic analysis, each of the three lake datasets were analyzed separately and treated
456 with an OTU abundance threshold cutoff of at least 5 sequences in 10% of the samples in the
457 dataset (similar strategy to [62]). For comparison of taxonomic abundances across samples, each
458 of the three datasets were then rarefied to an even sequencing depth, which was 4,491 sequences
459 for Muskegon Lake samples, 5,724 sequences for the Lake Michigan samples, and 9,037
460 sequences for the inland lake samples. Next, the relative abundance at the OTU level was
461 calculated using the *transform_sample_counts()* function in the phyloseq R package [63] by
462 taking the count value and dividing it by the sequencing depth of the sample. For all other
463 taxonomic levels, the taxonomy was merged at certain taxonomic ranks using the *tax_glom()*
464 function in phyloseq [63] and the relative abundance was re-calculated.
465
466 ***Heterotrophic bacterial production measurements***

Muskegon Lake samples from 2014 and 2015 were processed for heterotrophic bacterial production using the [^{3}H] leucine incorporation into bacterial protein in the dark method [64, 65]. At the end of the incubation with [^{3}H]-leucine, cold trichloroacetic acid-extracted samples were filtered onto 0.2 μm filters that represented the leucine incorporation by the bacterial community. Measured leucine incorporation during the incubation was converted to bacterial carbon production rate using a standard theoretical conversion factor of 2.3 kg C per mole of leucine [65].

Flow cytometry, measuring HNA and LNA

In the field, a total of 1 mL of 20 μm filtered lake water were fixed with 5 μL of glutaraldehyde (20% vol/vol stock), incubated for 10 minutes on the bench (covered with aluminum foil to protect from light degradation), and then flash frozen in liquid nitrogen to later be stored in -80°C freezer until later processing with a flow cytometer. Flow cytometry procedures followed the protocol laid out in Props et al. [66], which also uses the samples presented in the current study. Samples were stained with SYBR Green I and measured in triplicate. The lowest number of cells collected after denoising was 2342. HNA and LNA groups were selected using the fixed gates introduced in Prest et al. [67] and plotted in **Figure S11**. Cell counts were determined per HNA and LNA group and averaged over the three replicates (giving rise to HNAcc and LNAcc).

Data analysis

Processed data and analysis code for the following analyses can be found on the GitHub page for this project at https://deneflab.github.io/HNA_LNA_productivity/.

490 ***HNA-LNA and HNA-Productivity Statistics and Regressions***

491 We tested the difference in absolute number of cells within HNA and LNA functional groups
492 across running analysis of variance with a post-hoc Tukey HSD test (*aov()* and *TukeyHSD();*
493 *stats* R package; [68]). In addition, we tested the association of HNA and LNA to each other and
494 with productivity by running ordinary least squares regression with the *lm()* (*stats* R package;
495 [68]).

496

497 ***Ranking correlation***

498 Ranking correlation between variables was calculated using the Kendall rank correlation
499 coefficient, using the *kendalltau()* function in Scipy (v1.0.0) or *cor()* in R (v3.2). The 'tau-b'
500 implementation was used, which is able to deal with ties. Values range from -1 (strong
501 disagreement) to 1 (strong agreement). The same statistic was used to assess the similarity
502 between rankings of variable selection methods.

503

504 ***Centered-log ratio transform***

505 First, following guidelines from Paliy & Shanker, Gloor et al. and Quinn et al.[69–71], relative
506 abundances of OTUs were transformed using a centered log-ratio (CLR) transformation before
507 variable selection was applied. This means that the relative abundance $x_i$of a taxa was
508 transformed according to the geometric mean of that sample, in which there are p taxa present:

509 $$x_i' = \log(x_i/(\prod_{j=1}^{p} x_j)^{1/p})$$.

510 Zero values were replaced by $\delta = 1/p^2$. This was done using the scikit-bio package
511 (www.scikit-bio.org, v0.4.1).

Lasso & stability selection

Scores were assigned to taxa based on an extension of the Lasso estimator, which is called *stability selection* [25]. In the case of *n*samples, the Lasso estimator fits the following regression model:

$$\hat{\beta}^{\lambda} = \text{argmin}_{\beta \in \mathbb{R}^p} ||y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad ,$$

in which X denotes the abundance table, y the target to predict, which either is HNA cell abundances (HNAcc) or LNA cell abundances (LNAcc), and λ is a regularization parameter which controls the complexity of the model and prevents overfitting. The Lasso performs an intrinsic form of variable selection, as the weights of certain variables will be put to zero.

Stability selection, when applied to the Lasso, is in essence an extension of the Lasso regression. It implements two types of randomizations to assign a score to the variables, and is therefore also called the *Randomized Lasso* (RL). The resulting RL score can be seen as the probability that a certain variable will be included in a Lasso regression model (*i.e.,* its weight will be non-zero when fitted). When performing stability selection, the Lasso is fitted to B different subsamples of the data of fraction $n/2$, denoted as X' and corresponding y'. A second randomization is added by introducing a weakness parameter α. In each model, the penalty λ changes to a randomly chosen value in the set $[\lambda, \lambda/\alpha]$, which means that a higher penalty will be assigned to a random subset of the total amount of variables. The Randomized Lasso therefore becomes:

$$\hat{\beta}^{\lambda} = \text{argmin}_{\beta \in \mathbb{R}^p} ||y' - X'\beta||_2^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{w_j} \quad ,$$

where w_j is a random variable which is either α or 1. Next, the Randomized Lasso score (RL score) is determined by counting the number of times the weight of a variable was non-zero for each of the B models and divided by B. Meinshausen and Bühlmann show that, under stringent conditions, the number of falsely selected variables is controlled for the Randomized Lasso when the RL score is higher than 0.5. If λ is varied, one can determine the stability path, which is the relationship between π and λ for every variable. For our implementation, $B = 500$, $\alpha = 0.5$ and the highest score was selected in the stability path for which λ ranged from 10^{-3} until 10^{3}, logarithmically divided in 100 intervals. The *RandomizedLasso()* function from the scikit-learn machine learning library was used [72], v0.19.1).

Random Forests & Boruta

The Boruta algorithm is a *wrapper* algorithm that makes use of Random Forests as a base classification or regression method in order to select all relevant variables in function of a response variable [26]. Similar to stability selection, the method uses an additional form of randomness in order to perform variable selection. Random Forests are fitted to the data multiple times. To remove the correlation to the response variable, each variable gets per iteration a so-called *shadow variable*, which is a permuted copy of the original variable. Next, the Random Forest algorithm is run with the extended set of variables, after which variable importances are calculated for both original and shadow variables. The shadow variable that has the highest importance score is used as reference, and every variable with significantly lower importance, as determined by a Bonferroni corrected t-test, is removed. Likewise, variables containing an importance score that is significantly higher are included in the final list of selected variables. This procedure can be repeated until all original variables are either discarded or included in the

final set; variables that remain get the label 'tentative' (i.e., after all repetitions it is still not possible to either select or discard a certain variable). We used the boruta_py package to implement the Boruta algorithm (https://github.com/scikit-learn-contrib/boruta_py). Random Forests were implemented using *RandomForestRegressor()* function from scikit-learn [72], v0.19.1). Random Forests were run with 200 trees, the number of variables considered at every split of a decision tree was $p/3$ and the minimal number of samples per leaf was set to five. The latter were based on default values for Random Forests in a regression setting [73]. The Boruta algorithm was run for 300 iterations, variables were selected or discarded at $P < .05$ after performing Bonferroni correction.

Recursive variable elimination

Scores of the Randomized Lasso were evaluated using a recursive variable elimination strategy [74]. Variables were ranked according to the RL score. Next, the lowest-ranked variables were eliminated from the dataset, after which the Lasso was applied to predict HNAcc and LNAcc respectively. This process was repeated until only the highest-scored taxa remained. In this way, performance of the Randomized Lasso was assessed from a minimal-optimal evaluation perspective [75]. In other words, the lowest amount of variables that resulted in the highest predictive performance was determined.

Performance evaluation

In order to account for the spatiotemporal structure of the data, a blocked cross-validation scheme was implemented [76]. Samples were grouped according the site and year that they were collected. This results in 5, 10 and 16 distinctive groups for the Michigan, Muskegon and Inland

579 lake systems respectively. Predictive models were optimized in function of the R^2 between
580 predicted and true values of held-out groups using a leave-one-group-out cross-validation
581 scheme with the *LeaveOneGroupOut()* function. This results in a cross-validated R^2_{CV} value. For
582 the Lasso, λ was determined using the lassoCV() function, with setting eps=10^{-4} and
583 n_alphas=400. The Random Forest object was optimized using a grid search where max_features
584 was chosen in the interval $[1, \sqrt{p}, 2\sqrt{p}, ..., p]$ (all variables) or $[1, ..., p]$ (Boruta selected variables)
585 and min_samples_leaf in the interval $[1, ..., 5]$, using the *GridSearchCV()* function. The number
586 of decision trees (n_trees) was set to 200. All functions are part of scikit-learn ([72]; v0.19.1)
587
588 ***Stability of the Randomized Lasso***
589 Similarity of RL scores between lake systems and functional groups was quantified using the
590 Pearson correlation. This was done using the *pearsonr()* function in Scipy (v1.0.0).
591
592 ***Patterns of HNA and LNA OTUs across ecosystems and phylogeny***
593 To visualize patterns of selected HNA and LNA OTUs across the three ecosystems, a heatmap
594 was created with the RL scores of each OTU from the Randomized Lasso regression that were
595 higher than specified threshold values. The heatmap was created with the *heatmap.2()* function
596 (*gplots* R package) using the euclidean distances of the RL scores and a complete linkage
597 hierarchical clustering algorithm (**Figure 4**).
598
599 ***Correlations between taxa and productivity measurements***
600 Kendall tau ranking correlations between productivity measurements and individual abundances
601 were calculated on the phylum and OTU level using the *kendalltau()* function from Scipy

(v1.0.0). P-values were corrected using Benjamini-Hochberg correction, reported as P_adj. This was done using the *multitest()* function from the Python module Statsmodels ([77]; v0.5.0).

Phylogenetic tree construction and signal calculation

We calculated the best performing maximum likelihood tree using the GTR-CAT model (-gtr -fastest) model of nucleotide substitution with fasttree (version 2.1.9 No SSE3; [78]). Phylogenetic signal with both discrete (*i.e.* HNA, LNA, or both) and continuous traits (*i.e.* the RL score) using the newick tree from FastTree was then used to model phylogenetic signal using Pagel's lambda (discrete trait: fitDiscrete() from the geiger R package [79]; continuous trait: phylosig() from the phytools R [80]), Blomberg's K (phylosig() function from the phytools R package [80]), and Moran's I (abouheif.moran() function from the adephylo R package [81]).

Acknowledgements

PR was supported by Ghent University (BOFSTA2015000501) and MLS was supported by the National Science Foundation Graduate Research Fellowship Program (Grant No. DGE 1256260). Part of the computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government department EWI. Flow cytometry analysis was supported through a Geconcerteerde Onderzoeksactie (GOA) from Ghent University (BOF15/GOA/006).

References

1. Lennon JT, Jones SE. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nat Rev Microbiol* 2011; **9**: 119–30.

2. Carini P, Marsden PJ, Leff JW, Morgan EE, Strickland MS, Fierer N. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat Microbiol* 2016; **2**: 16242.

3. Widder S, Allen RJ, Pfeiffer T, Curtis TP, Wiuf C, Sloan WT, et al. Challenges in microbial ecology: Building predictive understanding of community function and dynamics. *ISME J* 2016; **10**: 2557–2568.

4. Gasol JM, Del Giorgio PA. Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. *Sci Mar* 2000; **64**: 197–224.

5. Vives-Rego J, Lebaron P, Caron Nebe-von. Current and future applications of flow cytometry in aquatic microbiology. *FEMS Microbiol Rev* 2000; **24**: 429–448.

6. Wang Y, Hammes F, De Roy K, Verstraete W, Boon N. Past, present and future applications of flow cytometry in aquatic microbiology. *Trends Biotechnol* 2010; **28**: 416–424.

7. Gasol JM, Zweifel UL, Peters F, Jed A, Zweifel ULI, Fuhrman JEDA. Significance of Size and Nucleic Acid Content Heterogeneity as Measured by Flow Cytometry in Natural Planktonic Bacteria Significance of Size and Nucleic Acid Content Heterogeneity as Measured by Flow Cytometry in Natural Planktonic Bacteria. *Appl Environ Microbiol* 1999; **65**: 4475–4483.

8. Lebaron P, Servais P, Agogué H, Courties C, Joux F. Does the High Nucleic Acid Content of Individual Bacterial Cells Allow Us to Discriminate between Active Cells and Inactive Cells in Aquatic Systems? *Appl Environ Microbiol* 2001; **67**: 1775–1782.

9. Bouvier T, Del Giorgio PA, Gasol JM. A comparative study of the cytometric characteristics of High and Low nucleic-acid bacterioplankton cells from different aquatic ecosystems. *Environ Microbiol* 2007; **9**: 2050–2066.

10. Wang Y, Hammes F, Boon N, Chami M, Egli T. Isolation and characterization of low nucleic acid (LNA)-content bacteria. *ISME J* 2009; **3**: 889–902.

11. Lebaron P, Servais P, Baudoux a.-C, Bourrain M, Courties C, Parthuisot N. Variations of bacterial-activity with cell size and nucleic acid content assessed by flow cytometry. *Aquat Microb Ecol* 2002; **28**: 131–140.

12. Servais P, Casamayor EO, Courties C, Catala P, Parthuisot N, Lebaron P. Activity and diversity of bacterial cells with high and low nucleic acid content. *Aquat Microb Ecol* 2003; **33**: 41–51.

13. Morán X, Bode A, Suárez L, Nogueira E. Assessing the relevance of nucleic acid content as an indicator of marine bacterial activity. *Aquat Microb Ecol* 2007; **46**: 141–152.

14. Servais P, Courties C, Lebaron P, Troussellier M. Coupling bacterial activity measurements with cell sorting by flow cytometry. *Microb Ecol* 1999; **38**: 180–189.

15. Bowman JS, Amaral-zettler LA, Rich JJ, Luria CM, Ducklow HW. Bacterial community segmentation facilitates the prediction of ecosystem function along the coast of the western Antarctic Peninsula. *ISME J* 2017; **11**: 1460–1471.

16. Morán XAG, Ducklow HW, Erickson M. Single-cell physiological structure and growth rates of heterotrophic bacteria in a temperate estuary (Waquoit Bay, Massachusetts). *Limnol Oceanogr* 2011; **56**: 37–48.

17. Read DS, Gweon HS, Bowes MJ, Newbold LK, Field D, Bailey MJ, et al. Catchment-scale biogeography of riverine bacterioplankton. *ISME J* 2015; **9**: 516–526.

18. Sherr EB, Sherr BF, Longnecker K. Distribution of bacterial abundance and cell-specific nucleic acid content in the Northeast Pacific Ocean. *Deep Res Part I Oceanogr Res Pap* 2006; **53**: 713–725.

19. Jochem FJ, Lavrentyev PJ, First MR. Growth and grazing rates of bacteria groups with different apparent DNA content in the Gulf of Mexico. *Mar Biol* 2004; **145**: 1213–1225.

20. Arnoldini M, Heck T, Blanco-Fernández A, Hammes F. Monitoring of Dynamic Microbiological Processes Using Real-Time Flow Cytometry. *PLoS One* 2013; **8**: e80117.

21. Ramseier MK, von Gunten U, Freihofer P, Hammes F. Kinetics of membrane damage to high (HNA) and low (LNA) nucleic acid bacterial clusters in drinking water by ozone, chlorine, chlorine dioxide, monochloramine, ferrate(VI), and permanganate. *Water Res* 2011; **45**: 1490–1500.

22. Schattenhofer M, Wulf J, Kostadinov I, Glöckner FO, Zubkov M V., Fuchs BM. Phylogenetic characterisation of picoplanktonic populations with high and low nucleic acid content in the North Atlantic Ocean. *Syst Appl Microbiol* 2011; **34**: 470–475.

23. Proctor CR, Besmer MD, Langenegger T, Beck K, Walser J-C, Ackermann M, et al. Phylogenetic clustering of small low nucleic acid-content bacteria across diverse freshwater ecosystems. *ISME J* 2018.

24. Vila-Costa M, Gasol JM, Sharma S, Moran MA. Community analysis of high- and low-nucleic acid-containing bacteria in NW Mediterranean coastal waters using 16S rDNA pyrosequencing. *Environ Microbiol* 2012; **14**: 1390–1402.

25. Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc Ser B Stat Methodol* 2010.

26. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *J Stat Softw* 2010; **36**: 1–13.

692 27. Shade A, Jones SE, Caporaso JG, Handelsman J, Knight R, Fierer N, et al. Conditionally
693 rare taxa disproportionately contribute to temporal changes in microbial diversity. *MBio*
694 2014; **5**: e01371-14.

695 28. Herren CM, McMahon KD. Keystone taxa predict compositional change in microbial
696 communities. *Environ Microbiol* 2018; 1–34.

697 29. Biddanda BA. Global Significance of the Changing Freshwater Carbon Cycle Emerging
698 Role of Freshwater in the Global Theater. *Eos (Washington DC)* 2017; **98**: 1–5.

699 30. Ma J, Prince AL, Bader D, Hu M, Ganu R, Baquero K, et al. High-fat maternal diet during
700 pregnancy persistently alters the offspring microbiome in a primate model. *Nat Commun*
701 2014; **5**: 1–11.

702 31. Chen J, Chia N, Kalari KR, Yao JZ, Novotna M, Soldan MMP, et al. Multiple sclerosis
703 patients have a distinct gut microbiota compared to healthy controls. *Sci Rep* 2016; **6**: 1–
704 10.

705 32. Revell LJ, Harmon LJ, Collar DC. Phylogenetic signal, evolutionary process, and rate.
706 *Syst Biol* 2008; **57**: 591–601.

707 33. Pagel M. Inferring the historical patterns of biological evolution. *Nature* 1999; **401**: 877–
708 884.

709 34. Lin W, Shi P, Feng R, Li H. Variable selection in regression with compositional
710 covariates. *Biometrika* 2014; **101**: 785–797.

711 35. Baxter NT, Zackular JP, Chen GY, Schloss PD. Structure of the gut microbiome
712 following colonization with human feces determines colonic tumor burden. *Microbiome*
713 2014; **2**: 1–11.

714 36. Schubert AM, Rogers M a M, Ring C, Mogle J, Petrosino JP, Young VB, et al.
715 Microbiome Data Distinguish Patients with Clostridium difficile Infection and Non- C .
716 difficile -Associated Diarrhea from Healthy. *MBio* 2014; **5**: 1–9.

717 37. Li H. Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis.
718 *Annu Rev Stat Its Appl* 2015; **2**: 73–94.

719 38. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable
720 importance for random forests. *BMC Bioinformatics* 2008; **9**: 307.

721 39. Zaura E, Brandt BW, Prodan A, Teixeira De Mattos MJ, Imangaliyev S, Kool J, et al. On
722 the ecosystemic network of saliva in healthy young adults. *ISME J* 2017; **11**: 1218–1231.

723 40. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for
724 random forests and omics data sets. *Brief Bioinform* 2017; 1–12.

725 41. McCarthy A, Chiang E, Schmidt ML, Denef VJ. RNA Preservation Agents and Nucleic

Acid Extraction Method Bias Perceived Bacterial Community Composition. *PLoS One* 2015; **10**: e0121659.

42. Louca S, Doebeli M, Parfrey LW. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome* 2018; **6**: 1–12.

43. Van der Gucht K, Cottenie K, Muylaert K, Vloemans N, Cousin S, Declerck S, et al. The power of species sorting: local factors drive bacterial community composition over a wide range of spatial scales. *Proc Natl Acad Sci U S A* 2007; **104**: 20404–20409.

44. Adams HE, Crump BC, Kling GW. Metacommunity dynamics of bacteria in an arctic lake: The impact of species sorting and mass effects on bacterial production and biogeography. *Front Microbiol* 2014; **5**: 1–10.

45. Chiang E, Schmidt ML, Berry MA, Biddanda BA, Burtner A, Johengen TH, et al. Verrucomicrobia are prevalent in north- temperate freshwater lakes and display class-level preferences between lake habitats. *PLoS One* 2018; **13**: 1–20.

46. Jones SE, Lennon JT. Dormancy contributes to the maintenance of microbial diversity. *Proc Natl Acad Sci* 2010; **107**: 5881–5886.

47. Zimmerman R, Iturriaga R, Becker-Birck J. Simultaneous determination of the total number of aquatic bacteria and the number thereof involved in respiration. *Appl Environ Microbiol* 1978; **36**: 926–935.

48. Aanderud ZT, Vert JC, Lennon JT, Magnusson TW, Breakwell DP, Harker AR. Bacterial dormancy is more prevalent in freshwater than hypersaline lakes. *Front Microbiol* 2016; **7**: 1–13.

49. Jia X, Dini-Andreote F, Falcão Salles J. Community Assembly Processes of the Microbial Rare Biosphere. *Trends Microbiol* 2018; **xx**: 1–10.

50. Amy PS, Morita RY. Starvation-survival patterns of sixteen freshly isolated open-ocean bacteria. *Appl Environ Microbiol* 1983; **45**: 1109–1115.

51. Corno G, Jürgens K. Direct and indirect effects of protist predation on population size structure of a bacterial strain with high phenotypic plasticity. *Appl Environ Microbiol* 2006; **72**: 78–86.

52. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, et al. Genomic Islands and the Ecology and Evolution of Prochlorococcus. *Science (80-)* 2006; **311**: 1768–1770.

53. Hunt DE, David L a, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science (80-)* 2008; **320**: 1081–1085.

54. Denef VJ, Kalnejais LH, Mueller RS, Wilmes P, Baker BJ, Thomas BC, et al.

Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proc Natl Acad Sci U S A* 2010; **107**: 2383–2390.

55. Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol* 2014; **22**: 235–247.

56. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. A guide to the natural history of freshwater lake bacteria. *Microbiology and molecular biology reviews* . 2011.

57. Woodhouse JN, Kinsela AS, Collins RN, Bowling LC, Honeyman GL, Holliday JK, et al. Microbial communities reflect temporal changes in cyanobacterial composition in a shallow ephemeral freshwater lake. *ISME J* 2016; **10**: 1337–1351.

58. Tada Y, Suzuki K. Changes in the community structure of free-living heterotrophic bacteria in the open tropical Pacific Ocean in response to microalgal lysate-derived dissolved organic matter. *FEMS Microbiol Ecol* 2016; **92**: 1–13.

59. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009; **75**: 7537–7541.

60. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* 2013; **41**: 590–596.

61. Rohwer RR, Hamilton JJ, Newton RJ, McMahon KD. TaxAss: Leveraging Custom Databases Achieves Fine-Scale Taxonomic Resolution. *bioRxiv* 2017; 214288.

62. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J* 2016; **10**: 1669–1681.

63. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* 2013; **8**: e61217.

64. Kirchman D, K'nees E, Hodson R. Leucine incorporation and its potential as a measure of protein synthesis by bacteria in natural aquatic systems. *Appl Environ Microbiol* 1985; **49**: 599–607.

65. Simon M, Azam F. Protein content and protein synthesis rates of planktonic marine bacteria. *Mar Ecol Prog Ser* 1989; **51**: 201–213.

66. Props R, Schmidt ML, Heyse J, Vanderploeg HA, Boon N, Denef VJ. Flow cytometric monitoring of bacterioplankton phenotypic diversity predicts high population-specific feeding rates by invasive dreissenid mussels. *Environ Microbiol* 2017; **00**.

67. Prest EI, Hammes F, Kötzsch S, van Loosdrecht MCM, Vrouwenvelder JS. Monitoring

796 microbiological changes in drinking water systems using a fast and reproducible flow
797 cytometric method. *Water Res* 2013; **47**: 7131–7142.

798 68. R Core Team. R: A Language and Environment for Statistical Computing. 2018. Vienna,
799 Austria.

800 69. Paliy O, Shankar V. Application of multivariate statistical techniques in microbial
801 ecology. *Mol Ecol* 2016; **25**: 1032–1057.

802 70. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are
803 compositional: And this is not optional. *Front Microbiol* 2017; **8**: 1–6.

804 71. Quinn TP, Erb I, Richardson MF, Crowley TM. Understanding sequencing data as
805 compositions: an outlook and review. *Bioinformatics* 2018; 1–9.

806 72. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn:
807 Machine Learning in Python. *J Mach Learn Res* 2011; **12**: 2825–2830.

808 73. Probst P, Wright M, Boulesteix A-L. Hyperparameters and Tuning Strategies for Random
809 Forest. *arXiv* 2018; preprint.

810 74. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using
811 Support Vector Machines. *Mach Learn* 2002; **46**: 389–422.

812 75. Nilsson R, Peña JM, Björkegren J, Tegnér J. Consistent Feature Selection for Pattern
813 Recognition in Polynomial Time. *J Mach Learn Res* 2007; **8**: 589–612.

814 76. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-
815 validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.
816 *Ecography (Cop)* 2017; **40**: 913–929.

817 77. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python.
818 *Proc 9th Python Sci Conf* 2010; 57–61.

819 78. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees
820 for large alignments. *PLoS One* 2010; **5**.

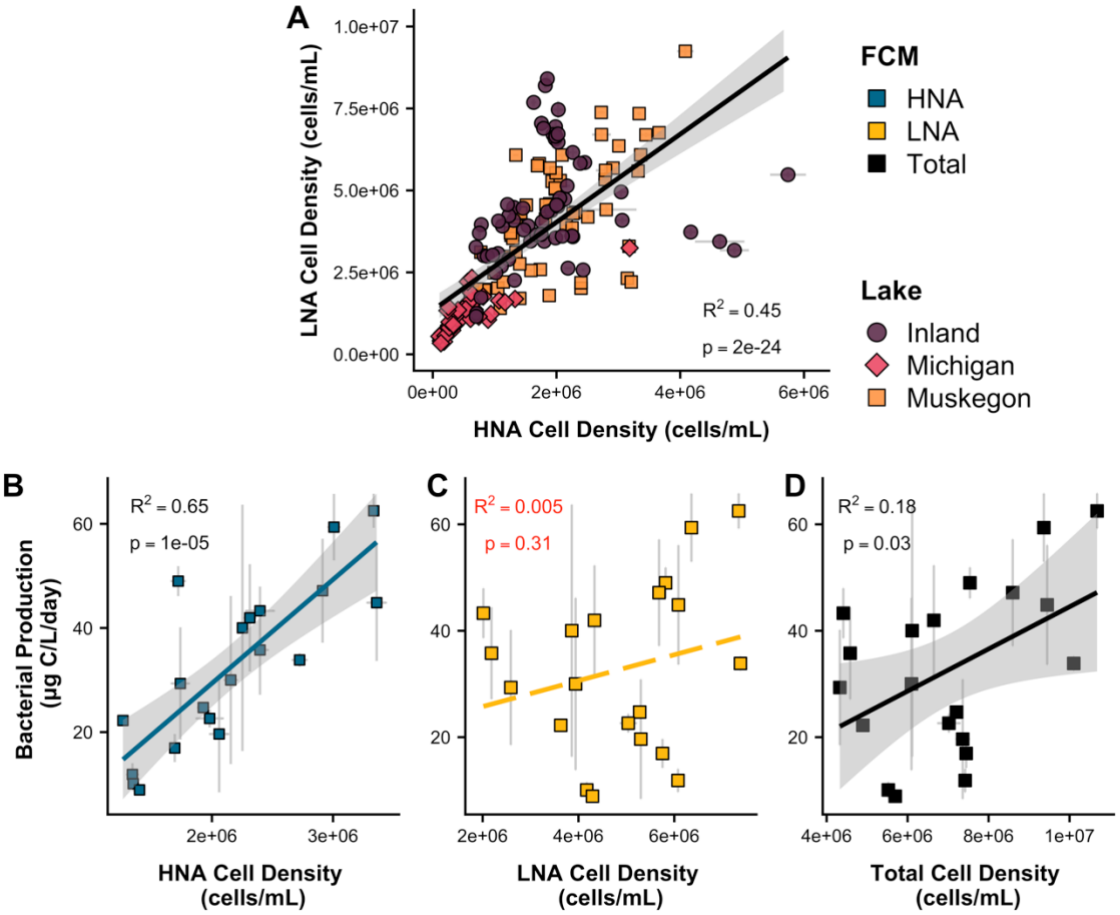
821 79. Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. GEIGER: Investigating
822 evolutionary radiations. *Bioinformatics* 2008; **24**: 129–131.

823 80. Revell LJ. phytools: An R package for phylogenetic comparative biology (and other
824 things). *Methods Ecol Evol* 2012; **3**: 217–223.

825 81. Jombart T, Balloux F, Dray S. adephylo: New tools for investigating the phylogenetic
826 signal in biological traits. *Bioinformatics* 2010; **26**: 1907–1909.

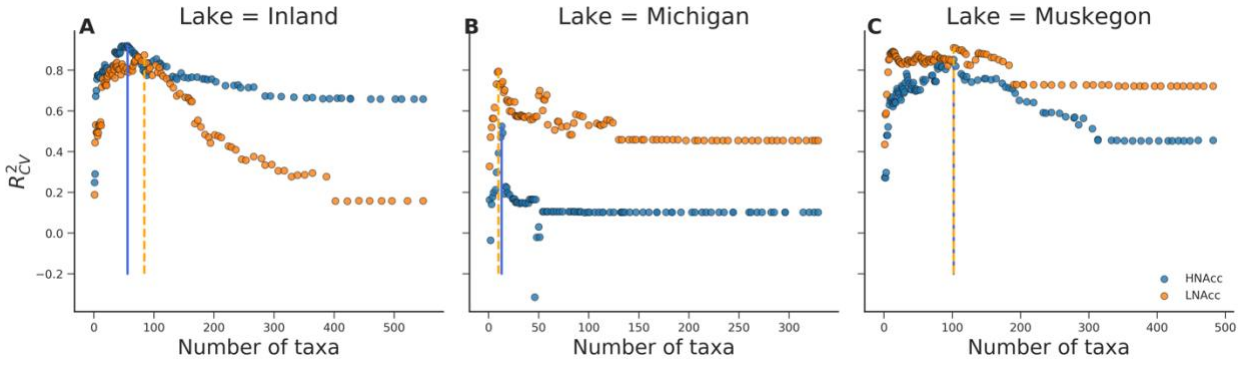
827

828 **Figure 1: (A)** Correlation between HNA cell counts and LNA cell counts across the three
829 freshwater lake ecosystems. **(B-D)** Muskegon Lake bacterial heterotrophic production and its
830 correlation with **(B)** HNA cell counts, **(C)** LNA cell counts, and **(D)** total cell counts. The grey
831 area in plots A, B, and D represents the 95% confidence intervals.

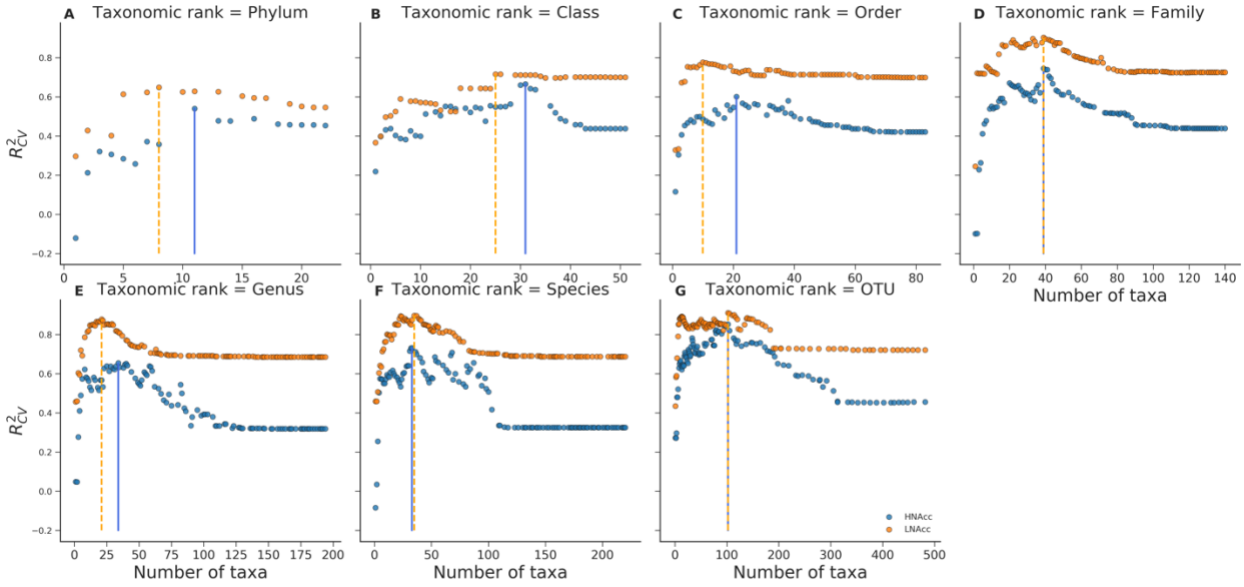


832

Figure 2: R^2_{CV} in function of the number of OTUs, which were iteratively removed based on the RL score and evaluated using the Lasso at every step. The solid (HNA) and dashed (LNA) vertical lines corresponds to the threshold (i.e., number of OTUs) which resulted in a maximal R^2_{CV}. **(A)** Inland system ($R^2_{CV,max} = 0.92$), HNAcc; **(B)** Lake Michigan ($R^2_{CV,max} = 0.53$), HNAcc; **(C)** Muskegon lake, HNAcc ($R^2_{CV,max} = 0.85$); **(D)** Inland system, LNAcc ($R^2_{CV,max} = 0.87$); **(E)** Lake Michigan, LNAcc ($R^2_{CV,max} = 0.79$); **(F)** Muskegon lake, LNAcc ($R^2_{CV,max} = 0.91$).

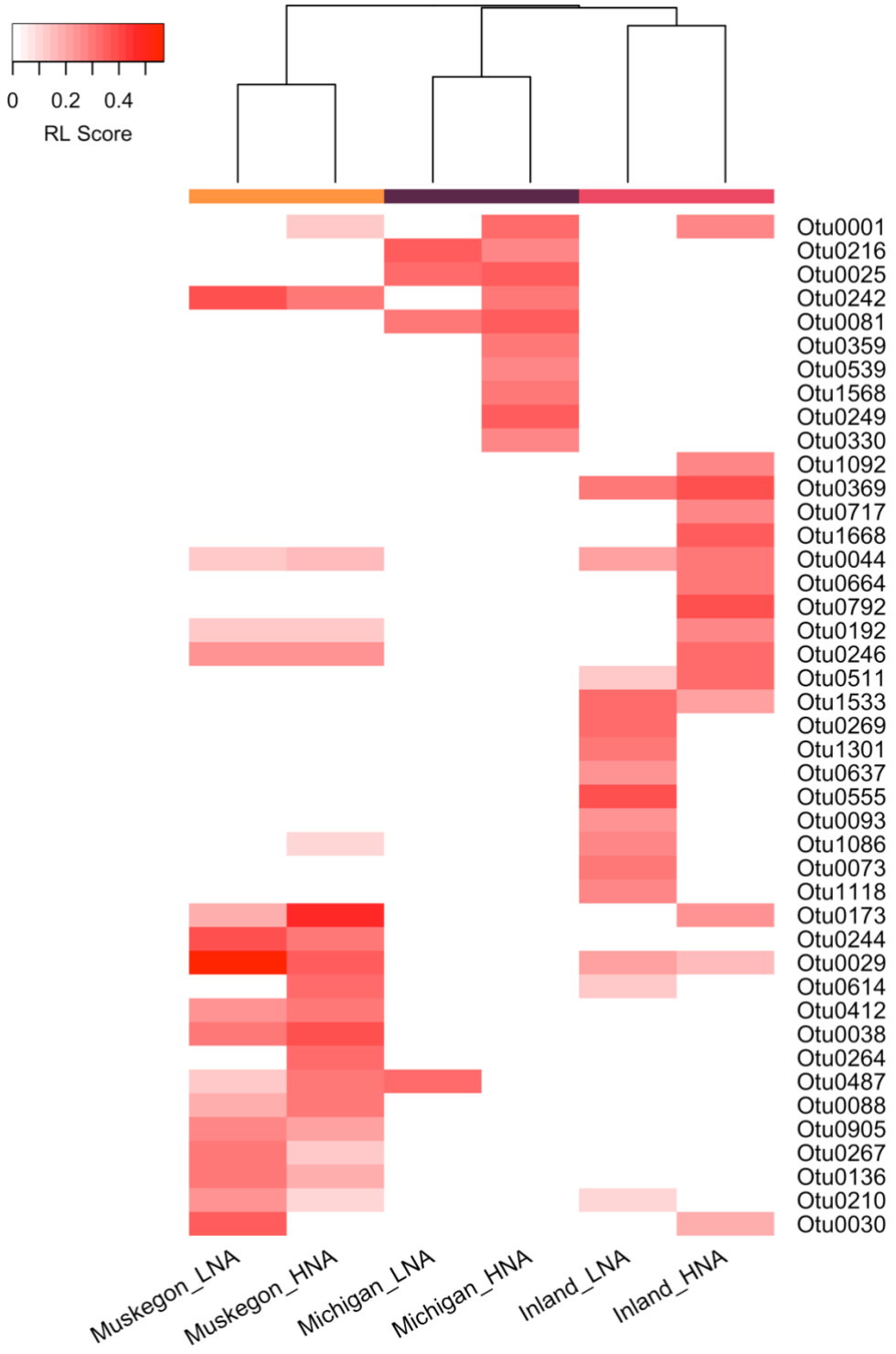


841 **Figure 3:** Evaluation of HNAcc and LNAcc predictions using the Lasso at all taxonomic levels
842 for the Muskegon lake system, expressed in terms of R^2_{CV}, using different subsets of taxonomic
843 variables. Subsets were determined by iteratively eliminating the lowest-ranked taxonomic
844 variables based on the RL score.



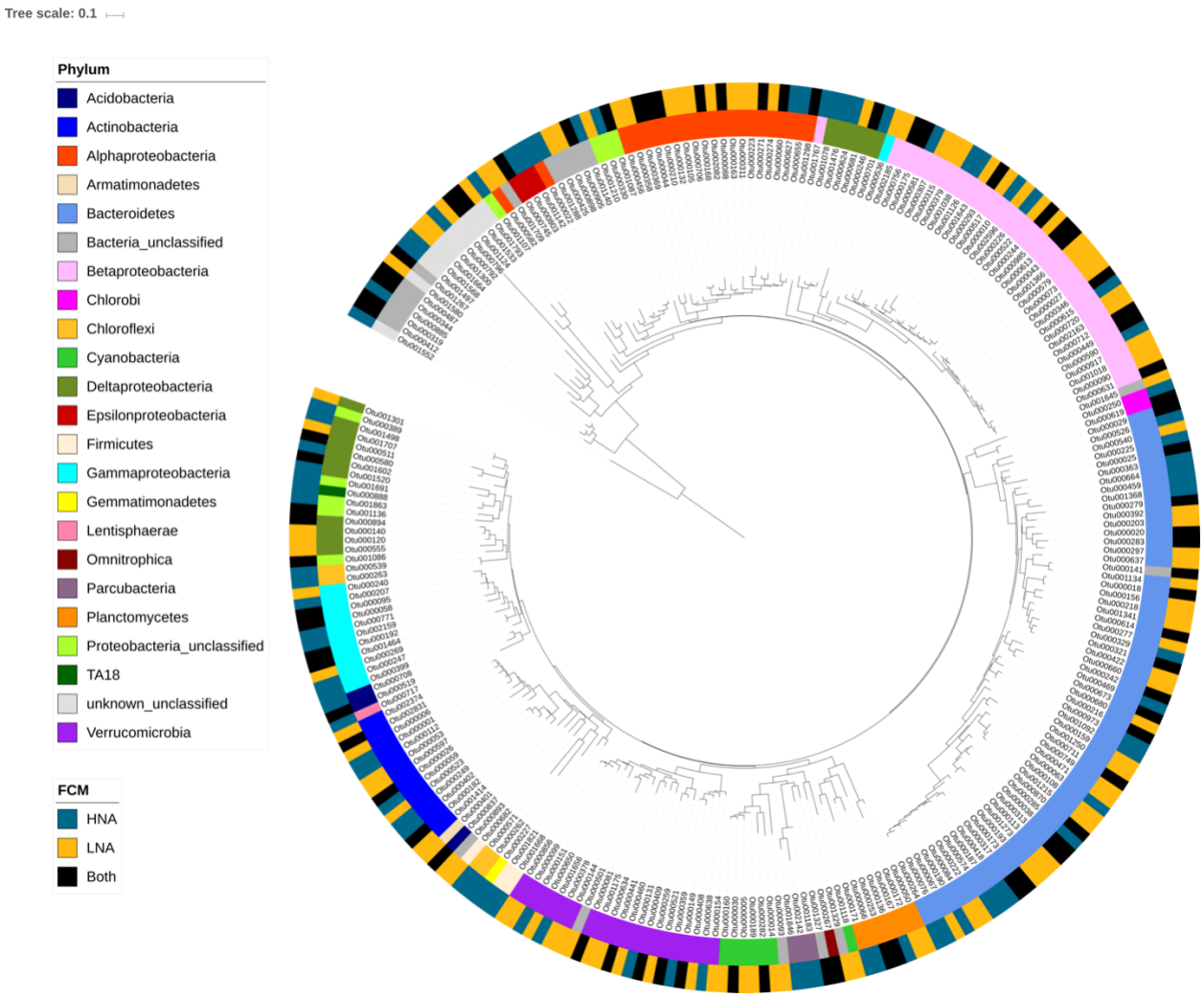
845

846 **Figure 4:** Hierarchical clustering of the RL score for the top 10 selected OTUs within each lake
847 system and FCM functional groups with the selected OTU (rows) across HNA and LNA groups
848 within the three lake systems (columns).

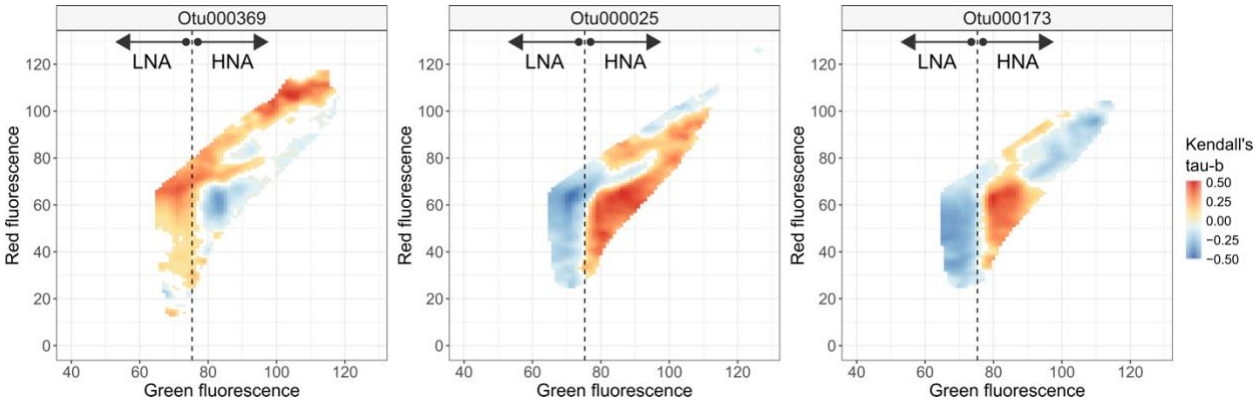


849

Figure 5: Phylogenetic tree with all HNA and LNA selected OTUs from each of the three lake systems with their phylum level taxonomic classification and association with HNA, LNA or to both groups based on the RL score threshold values.



854 **Figure 6:** Correlation (Kendall's tau-b) between the relative abundances of the top three OTUs
855 selected by the RL and the densities in the cytometric fingerprint. The fluorescence threshold
856 used to define HNA and LNA populations is indicated by the dotted line.



857

858 **Table 1:** Top scored OTUs according to the RL per functional population and lake ecosystem.
859 Selection according to the Boruta algorithm is given in addition to the RL score. Descriptive
860 statistics by means of the Kendall rank correlation coefficient (KRCC) have been added with
861 level of significance in function of the HNA/LNA population. Full taxonomy of the OTUs is
862 given in **Table S2**.
863

Lake system	Functional group	OTU	RL score	Boruta selected	Kendall's tau (HNA)	P-value (HNA)	Kendall's tau (LNA)	P-value (LNA)
Inland	HNA	OTU369	0.382	yes	-0.43	<0.001	-0.28	0.0012
	LNA	OTU555	0.384	no	0.089	N.S.	0.22	0.011
Michigan	HNA	OTU025	0.362	yes	0.46	<0.001	0.41	<0.001
	LNA	OTU168	0.428	yes	0.26	0.0092	0.4	<0.001
Muskegon	HNA	OTU173	0.462	yes	0.5	<0.001	0.2	0.019
	LNA	OTU029	0.568	no	0.26	0.0029	0.49	<0.001

Tree scale: 0.1

Phylum

- Acidobacteria
- Actinobacteria
- Alphaproteobacteria
- Armatimonadetes
- Bacteroidetes
- Bacteria_unclassified
- Betaproteobacteria
- Chlorobi
- Chloroflexi
- Cyanobacteria
- Deltaproteobacteria
- Epsilonproteobacteria
- Firmicutes
- Gammaproteobacteria
- Gemmatimonadetes
- Lentisphaerae
- Omnitrophica
- Parcubacteria
- Planctomycetes
- Proteobacteria_unclassified
- TA18
- unknown_unclassified
- Verrucomicrobia

FCM

- HNA
- LNA
- Both