Article Type : Original Article

Genome Evolution and Host Microbiome Shifts Correspond with Intraspecific Niche Divergence within Harmful Algal Bloom-Forming *Microcystis aeruginosa.*

Short Title: Niche differentiation in *Microcystis aeruginosa*

Sara L. Jackrel[1*], Jeffrey D. White[2,3], Jacob T. Evans[1], Kyle Buffin[1], Kristen Hayden[1], Orlando Sarnelle[3], Vincent J. Denef[1]

[1]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109 USA
[2]Department of Biology, Framingham State University, Framingham, MA 01701, USA
[3]Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI 48824, USA
* Corresponding author

Abstract: Intraspecific niche divergence is an important driver of species range, population abundance, and impacts on ecosystem functions. Genetic changes are the primary focus when studying intraspecific divergence; however, the role of ecological interactions, particularly host-microbiome symbioses, is receiving increased attention. The relative importance of these evolutionary and ecological mechanisms has seen only limited evaluation. To address this question, we used *Microcystis aeruginosa*, the globally-distributed cyanobacterium that dominates freshwater harmful algal blooms. These blooms have been increasing in occurrence and intensity worldwide, causing major economic and ecological damages. We evaluated 46 isolates of *M. aeruginosa* and their microbiomes, collected from 14 lakes in Michigan, USA that vary over 20-fold in phosphorus levels, the primary limiting nutrient in freshwater systems. Genomes of *M. aeruginosa* diverged along this phosphorus gradient in genomic architecture and protein functions. Fitness in low-phosphorus lakes corresponded with additional shifts within *M. aeruginosa* including genome-wide reductions in nitrogen use, an expansion of phosphorus assimilation genes, and an alternative life history strategy of non-clonal colony formation. In addition to host shifts, despite culturing in common garden conditions, host-microbiomes diverged along the gradient in taxonomy, but converged in function with evidence of metabolic interdependence between the host and its microbiome. Divergence corresponded with a physiological tradeoff between fitness in low-phosphorus environments and growth rate in phosphorus-rich conditions. Co-occurrence of genotypes adapted to different nutrient environments

in phosphorus-rich lakes may have critical implications for understanding how *M. aeruginosa* blooms persist after initial nutrient depletion. Ultimately, we demonstrate that the intertwined effects of genome evolution, host life history strategy, and ecological interactions between a host and its microbiome correspond with an intraspecific niche shift with important implications for whole ecosystem function.

Keywords: genome evolution; adaptation; intraspecific variation; host-microbiome; nutrient limitation; cyanobacterial harmful algal blooms

Introduction:

Variation within a species has been shown to rival the effects of among-species variation in regulating community structure (Crutsinger et al., 2006), trophic interactions (Post et al., 2008; Chislock et al. 2013), and nutrient cycling through ecosystems (Bassar et al., 2010). Therefore, predicting how these community- and ecosystem-level processes vary over time and space requires a mechanistic understanding of intraspecific divergence. Variation within a species that leads to niche divergence can occur as the result of evolutionary change, phenotypic plasticity, and even through altered ecological interactions with symbiotic partners (Lamichhaney et al., 2015; Lau & Lennon, 2012; Pfennig et al., 2010). As evolutionary and ecological mechanisms can occur over different time scales and their effects can have varying degrees of permanence, understanding the relative roles of these mechanisms towards shaping and maintaining intraspecific variation can elucidate what factors mediate stability of important ecosystem processes.

Genomic drivers of niche divergence between closely related organisms has been documented in several systems (Johnson et al. 2006; Lamichhaney et al., 2015). While selection upon standing genetic variation within a population can occur relatively rapidly (Barrett & Schluter, 2008), new mutations or lateral transferred genes rarely improve fitness (Drake et al., 1998; Vox et al., 2015), although examples of selective sweeps mediated by niche-determining genes acquired by lateral gene transfer have been documented (Shapiro and Polz, 2014). In contrast to evolutionary change, ecological mechanisms provide an alternative that may permit organisms to acclimate to divergent environmental conditions more rapidly. For example, the focal organism may take advantage of beneficial functionalities that pre-exist in other organisms. As such, microbial symbionts can contribute towards essential functions for the survival of their host in novel environments, such as those bacteria that facilitated the transition of the marine algae into freshwater environments (Dittami et al., 2016) and soil microbes that improved plant fitness under drought stress (Lau & Lennon, 2012).

To further our understanding of the mechanisms by which intraspecific niche divergence occurs, we simultaneously investigated the roles of 1) genomic changes within the host and, 2) shifts in

community membership and functionality of the host microbiome. We used the cyanobacterial phytoplankter *Microcystis aeruginosa* as our model to probe the roles of these two factors in driving fitness across a nutrient gradient from highly eutrophic, phosphorus-replete lakes to oligotrophic, low-phosphorus-lakes. This cyanobacterium is a useful model system as we can infer fitness of each of our *M. aeruginosa* isolates in their respective environments by collecting actively growing colonies, which are frequently composed of 10^4 to 10^5 cells (Costas et al., 2008). Understanding drivers of *M. aeruginosa* fitness across a nutrient gradient has major ecological, socioeconomic, and human health implications. As a dominant, bloom-forming cyanobacterium in nutrient-enriched freshwater systems world-wide, *M. aeruginosa* has produced concentrated levels of hepatotoxic microcystins that have caused mass wildlife mortalities (Masango et al. 2010) and drinking water crises, such as the 2007 bloom in Lake Taihu (Qin et al. 2010) and the 2014 bloom in Lake Erie (Steffen et al., 2017). Models predict that blooms of record-breaking intensity will become increasingly common due to cultural eutrophication and increased temperatures driven by climate change (Michalak et al., 2013).

As a consequence, eutrophic environments have been the focus of genome sequencing effort for *M. aeruginosa*. Their genomes are noted as having an unusually high percentage of long DNA repeats, insertions, transposable elements, and lateral gene transfers (Franguel et al., 2008; Kaneko et al., 2007; Meyer et al., 2017). Sequenced *M. aeruginosa* genomes from eutrophic habitats are highly variable, but this variation does not show clear biogeographic patterns (Meyer et al., 2017). In contrast, we previously identified a monophyletic ecotype of *M. aeruginosa* occurring in oligotrophic inland lakes in Michigan (Berry et al., 2017). Berry et al. (2017) constructed a phylogeny of *M. aeruginosa* isolates from inland lakes of Michigan using five housekeeping genes but did not investigate any further differences in the genomes or physiology of *M. aeruginosa* or host-associated microbiomes. Here, using this same dataset, we investigated the underlying mechanisms that may increase fitness of different ecotypes across a phosphorus gradient. We also evaluated genomes of heterotrophic bacteria residing in the *M. aeruginosa* phycosphere, defined as the nutrient-rich microenvironment immediately surrounding a phytoplankton cell where metabolites between the host and associated-bacteria are most readily exchanged (Seymour et al. 2017). The phycosphere of *M. aeruginosa* has previously been described as harboring a diversity of α- and β-proteobacteria, as well as bacteroidetes (Cai et al., 2014; Louati et al., 2015).

To test for evolutionary mechanisms promoting fitness across a lake phosphorus gradient, we probed for genomic signatures of streamlining in *M. aeruginosa* as a means for improved efficiency in nutrient-limited environments (Giovannoni et al., 2014). We considered strategies that would improve assimilation and efficiency of both phosphorus and nitrogen because oligotrophic lakes are typically co-limited (Sterner, 2008). We also evaluated cyanobacterial genomes for functional shifts across the

phosphorus gradient, including gene loss and gain, as well as indicators of positive selection. In addition, by evaluating population-level genomic heterogeneity within our host genome assemblies, we were able to infer life history strategies of *M. aeruginosa* colony formation as clonal (by cell division from a single cell) or non-clonal (by cell adhesion) (Xiao et al., 2018). To test for ecological mechanisms promoting fitness across the phosphorus gradient, we assessed whether changes in community membership and functionality of the host microbiome were a function of the phosphorus environment, and examined the functional roles played by the ubiquitously-associated bacterium, *Phycosocius bacilliformis* (Tanabe et al., 2015). This bacterium has been previously detected in 35 of 39 blooms of *M. aeruginosa* sampled in Asia and Africa (Tanabe et al., 2015), and was found in association with each of our 46 isolates of *M. aeruginosa*. Lastly, we evaluated the interdependence between the functionality of the host genome and that of its microbiome. We demonstrated that intraspecific niche divergence in *M. aeruginosa* corresponds with a combination of altered life history strategy, genome evolution, and ecological mechanisms. Further, evolutionary change, such as genome-wide shifts and novel gene acquisitions in the host, interacts with ecological shifts in the host microbiome that combine to ultimately correspond with improved host fitness. Additionally, from a methodological perspective, we used our data to demonstrate the need for cautious interpretation of apparent gene loss and reduced genome size when using metagenome-assembled genomes.

Materials and Methods:

Isolate Collection and Maintenance:

We collected colonies of *M. aeruginosa* from 14 inland lakes located throughout southern Michigan, USA in July 2011, August 2011, and August 2013 (see Table S1 and Fig. S2 for coordinates of each lake and a map of lake locations). Lakes were selected to span a large nutrient gradient, from oligotrophic to hyper-eutrophic, as determined by total phosphorus concentration (TP, a widely used index of lake productivity). Our TP concentrations spanned over an order of magnitude from ~8 – 200 µg/L, which encompasses the range of TP documented in over 82% of lakes in the Northeastern United States (Soranno et al. 2017). While we also measured lake NH_4^+ and NO_3^-, neither nitrogen measure corresponded with TP (correlations not significant, both $p > 0.10$, see Table S1 for TP, NH_4^+ and NO_3^- measures per lake). We collected water from the mixed layer of each lake via two pooled casts of an integrating tube sampler (12 m x 2.5 cm inside diameter). A subset of the water sample was stored for measurement of lake total phosphorus using the standard molybdenum-blue colorimetric technique and long path length spectrophotometry following persulfate digestion of organic matter (Menzel & Corwin,

1965; Murphy & Riley, 1962). We used standard thresholds in TP for assigning trophic status with an oligotrophic-mesotrophic boundary of 10 µg/L and a mesotrophic-eutrophic boundary of 30 µg/L (Wetzel, 2001). To confirm trophic status of these lakes was consistent over time, each lake was sampled at least three times during multiple years, except for Lake Lansing which was sampled twice. Longer-term data sets are also reported for Gull Lake, MSU Lake 2, and Little Long Lake (see Table S1 for mean, min, and max TP, years sampled, and number of observations). To isolate *M. aeruginosa* from water samples, we used a Leica MS5 dissecting scope at 16X to pipette individual colonies. All colonies isolated were distinctive in shape, rather than amorphous masses or loose aggregations of cells. To retain only closely-associated phycosphere bacteria inhabiting the mucilage of the *M. aeruginosa* colony, we washed individual colonies by pipetting each sequentially through a series of six well plates each containing sterile 0.5x WC-S growth medium. However, we note that this washing step might not have eliminated all free-living bacteria. We then transferred colonies into 20 mL tubes of sterile 0.5x WC-S growth medium, which typically has a higher successful establishment rate of ~80% for inland lake *M. aeruginosa* compared to other mediums such as BG-11 (White et al., 2011; Wilson et al., 2005). We then maintained all successfully established isolates in 200 mL batch cultures of 0.5x WC-S medium, incubated isolates at 23°C under a 12:12 h light:dark cycle of 80 µmol m^{-2} s^{-1}, and on a monthly basis, transferred an inoculum of each culture to fresh, sterile, 0.5x WC-S medium.

16S rRNA gene and Metagenomic Sequencing:

On November 14 -18, 2014, we trapped subsamples of each culture on 0.45 µm nitrocellulose filters, froze filters immediately, and stored at -80°C until extraction. This pore size allowed smaller free-living bacteria to pass through the filter. However, we caution that while the washing of colonies at the start of cultivation and filtration steps should have reduced inclusion of free-living bacteria in sequencing, these steps would not eliminate them entirely. We then thawed and incubated filters in 100 µL of Qiagen ATL tissue lysis buffer, 300 µL Qiagen AL lysis buffer, and 30 µL proteinase K for 1 h at 56°C on a rotisserie at maximum speed. We vortexed cells for 10 minutes to lyse, homogenized lysates with a Qiashredder column, and purified DNA from the filtrate using a DNeasy Blood and Tissue kit (Qiagen, Hilden, Germany).

We surveyed the phycosphere bacterial community from the extracted DNA of each culture by generating PCR amplicon of the V4 region of the 16S rRNA gene using 515f/806r primers (Bergmann et al., 2011). DNA amplicon were sequenced on a 2x250 Illumina MiSeq v2 run at the University of Michigan Medical School. Data were generated using RTA v1.17.28 and MCS v2.2.0 software. We also generated metagenomic data of the *M. aeruginosa* host and associated phycosphere bacteria on an

Illumina HiSeq 100 cycle 2 x 100 nt PE sequencing run at the University of Michigan Sequencing Core. Libraries were generated with a 500 nt insert size using an automated Apollo 324 library preparation system (Wafergen Biosystems, Fremont, CA). We aimed to obtain approximately equal coverage of *M. aeruginosa* across all metagenome samples by adjusting the proportions of each individual library to the pooled-libraries sample based on the relative abundance estimates of *M. aeruginosa* from our 16S amplicon data. Raw sequencing data files are available under SRA accession number PRJNA351875.

Sequencing Analyses:

We analyzed 16S rRNA gene survey data, including quality control of raw reads, read alignment, taxonomy assignment, and OTU clustering at 97% sequence similarity, using the mothur v1.34.3 standard operating procedure (accessed March 13, 2016 at) (Schloss et al., 2009, Kozich et al. 2013). We completed taxonomy assignment of sequences using the TaxAss pipeline, which classifies sequences to a smaller database of freshwater taxa (Newton et al., 2011) and the larger SILVA database (Quast et al., 2012; Wang et al., 2007). Based on methods recommended by McMurdie and Homes (2014), read depth was normalized to depth of the smallest sample (n = 10,090 reads, see Table S2 for original and scaled read depths for each sample) using custom scripts that can be found here: . Raw metagenome reads were trimmed of adapters using Scythe and quality trimmed using Sickle with default parameters (Joshi & Fass, 2011). Sequence quality was assessed before and after quality filtering using FastQC. We ran these quality control steps with a composite bash script that can be found here: https://github.com/Geo-omics/scripts/blob/master/wrappers/Assembly/qc.sh. Sequencing reads were then assembled using idba-ud with the following parameters (--mink 50, --maxk 92, --step 4 or 6, --min_contig 500) (Anantharaman et al., 2014). Metagenomic assemblies were first visualized with ESOM (Dick et al., 2009). We identified bins for both the target organism, *M. aeruginosa*, and an abundant phycosphere bacterium, *Phycosocius bacilliformis*. We used the default protocol for ESOM (Emergent Self-Organizing Maps), which is a binning approach that takes advantage of taxon-specific genomic signatures that arise due to genome-specific biases in codon usage. All sequences from our 46 metagenomic assemblies were trimmed into sequences of 10 kb in length and then imported into ESOM for assessment of tetranucleotide frequency of each contig. Each contig was plotted as a dot on the map using an unsupervised clustering algorithm to minimize distances between contigs sharing similar tetranucleotide frequency. Clusters of sequences were then manually selected and extracted as a bin. We identified high quality *P. bacilliformis* genomes using drep (Olm et al., 2017), compared similarity among these genomes using compareM (), and then retained only those genotypes at least 0.70% divergent from all other genotypes for further analysis. We uploaded 1) complete metagenome assemblies, 2) isolated *M. aeruginosa* bins, and 3) *Phycosocius bacilliformis*

bins, all with a 4kb contig length cutoff, into the Joint Genome Institute Integrated Microbial Genomes database. We passed these sequencing data through standard analysis pipelines for assigning protein families (Huntemann et al., 2015). Analyses of these *M. aeruginosa* bins indicated our contig length cutoff was likely too stringent, because core photosynthesis genes, among others, often resided on contigs of approximately 3kb in length. We therefore repeated binning of *M. aeruginosa* at a lower 2 kb length threshold using VizBin (Laczny et al., 2015). Coverage of sequences within this bin was then calculated using bwa (Li & Durbin, 2009). Histograms of contig frequency versus coverage were then used to visualize coverage distributions for each sample. We then discarded all contigs below the main coverage distribution, which varied depending on sample.

Microcystis aeruginosa genomes at this 2 kb length cutoff were then re-annotated using our custom pipeline. We quantified single nucleotide polymorphisms, inserts, and deletions within each genome using samtools (Li et al., 2009). We also identified paralogs as any reciprocal hits within a genome below an e-value of 0.10 using the diamond protein alignment software (Buchfink et al., 2014). We identified sigma factors as genes assigned to any of 27 different protein families that contained the keyword 'sigma' within either the pfam name or pfam summary. We report sigma factors as a percentage of all genes in the genome. Given the tendency for some genes to occur on even smaller contigs, especially in the low nutrient group of genomes, we added all called genes on scaffolds below 2 kb that were assigned as *M. aeruginosa* using the USEARCH based Phylogenetic Distribution tool in the JGI IMG Standard Operating Procedure (Huntemann et al., 2015). We used these final bins with contigs of all lengths for determining GC content, genome size, percent completeness and percent contamination using checkM (Parks et al., 2015). We determined which genes are considered 'core' to the cyanobacterial phylum using checkM, which estimates genome completeness by referencing sets of marker genes that are specific to the inferred lineage of a genome within a reference genome phylogeny. We also used these final bins with contigs of all lengths for determining pairwise average nucleotide identity (ANI) between genomes using PYANI (Pritchard et al. 2016), where the boundary for prokaryotic species is generally accepted as ~ 95 – 96% ANI (Richter and Rossello-Mora, 2009). We used our custom pipeline to assign protein families to each gene residing on these < 2 kb contigs and added these protein family assignments to our primary annotation of all *M. aeruginosa* 2 kb+ bins. We used this approach to be conservative when determining if any gene functions were completely absent from a genome. However, because it becomes more challenging to call genes on such short fragments, we report all gene percentage data, including % coding DNA, % sigma factors, and % paralogs, using the 2 kb+ bins. However, we provide a comparison of all genome metrics with and without scaffolds below 2 kb in supplementary materials.

We constructed a multilocus sequence typing phylogeny using our 46 isolates collected from inland lakes in Michigan, as well as additional *M. aeruginosa* genotypes that had been collected from multiple locations across six continents. In addition to our 46 isolates first reported in Berry et al. (2017), genomes obtained from NCBI included 12 genomes referenced in Humbert et al. (2013) (), 8 genomes referenced in Meyer et al., (2017) and a *Synechococcus* outgroup (CP000097.1). We used gene sequences from five housekeeping genes (*pgi*, *gltX*, *ftsZ*, *glnA*, *gyrB*) that were obtained from *M. aeruginosa* strain NIES483. As in our previous analysis (Berry et al., 2017), we searched for gene orthologs in the metagenomic data of each of the Michigan inland lake isolate and the genomes obtained from NCBI using a custom ruby script available on this project's github page (https://github.com/DenefLab/microcystis-oligotypes). Extracted gene sequences were concatenated and aligned with MUSCLE using default parameters (Edgar et al., 2004). A phylogeny was constructed using RAxML v8.2.8 with a *Synechococcus* outgroup (Berry et al., 2017; Stamatakis, 2006), and a Newick tree was visualized using FigTree v1.4.3 software (Rambaut, 2012).

For a subset of isolates collected from a relatively uniform environment (either Gull or Wintergreen Lakes on August 8, 2013), we measured the amount of genome variation among isolates with a pangenome analysis. We estimated the size of the core genome shared across isolates versus isolate-level variation by finding homologous gene clusters using the GET_HOMOLOGUES software (Contreras-Moreira & Vinuesa, 2013). We used only a consensus subset of gene clusters identified by both OrthoMCL Markov Cluster and COGtriangle algorithms.

To detect positive selection in the two main branches that divide all 46 inland lake isolates of *M. aeruginosa*, we calculate genome-wide synonymous to non-synonymous substitution rate ratios with the posigene software package (Sahm et al., 2017). For this analysis, we supplied a smaller-scale phylogenetic tree with only eight genomes representative of the oligotrophic and eutrophic/mesotrophic branches from Berry et al. (2017) (K13-10, G13-05, G13-09, LL11-07, LG13-11, F13-15, K13-06, K13-05). We ran analyses separately for each branch and identified orthologs against the largest genome within each respective branch as the anchor species (i.e., G13-05 and K13-10).

Growth Assays:

We aimed to detect whether growth rate of *M. aeruginosa* varies across populations according to phylogenetic group and a total phosphorus gradient. To assess variation across phylogenetic groups, we measured growth rates of 19 strains from our culture collection that had been isolated in 2011. To further validate our results of different growth rates across a total phosphorus gradient, we added an additional 12

strains that were assayed by Wilson et al. (2006). In total, these growth rate data represent 31 strains originating from 22 lakes throughout lower Michigan.

Common-garden growth assays to detect variation among phylogenetic groups of *M. aeruginosa* were conducted with the general design as follows. Fresh 20 mL cultures of isolates were initiated 7 days prior to an assay to ensure that *M. aeruginosa* was exponentially growing. One colony from each of 19 strains of *Microcystis* were then inoculated via pipette (1 μL) into randomized, separate wells containing 0.5 mL sterile 0.5× WC-S medium within 8-well chambered slides (Nunc Lab-Tek II Chamber Slide System) (Wilson et al., 2010). Once inoculated (day 0), colonies were photographed every 2 days for 6 days at 100× using a light microscope (Nikon Eclipse E600) interfaced with a digital camera (Diagnostic Instruments). See supplementary Fig. S3 for a sequence of digital micrographs depicting growth of a *M. aeruginosa* colony during a 6-day growth assay. Measurements, added to the images with computer software (Spot Advanced, Diagnostic Instruments), were made of colony surface area and depth (the straight line length perpendicular to the greatest linear dimension); colony volume (μm^3) was determined as the product of surface area and depth (Wilson et al. 2010). Growth rate was determined as the slope of the linear regression of natural logarithm-transformed colony volumes over time.

Since coloniality is a characteristic trait of *M. aeruginosa* in nature (Wehr & Sheath, 2003), all growth assays were performed using colonial isolates that had been in culture for less than ~1.5 years. This contrasts with many previous lab studies of *M. aeruginosa* that have utilized older, single-celled culture collection strains. Furthermore, since all isolates employed in the experiments were the same age and were recently isolated, concerns arising from evolution in culture were minimized (Burkholder & Glibert, 2009; Demott & Mckinney, 2015; Lakeman et al., 2009). Assaying growth of individual *M. aeruginosa* colonies is necessary and advantageous because, unlike batch culture assays, this permits controlling for colony size and inoculation density effects on growth rate, since small colonies grow faster than large colonies (Wilson et al., 2010). To further minimize the confounding effects of initial colony size and shape on growth rate, round colonies of approximately the same equivalent diameter were selected for each isolate to the fullest extent possible using an ocular micrometer.

Statistical Analyses:

We tested whether the phylogenetic grouping and lake of origin of *M. aeruginosa* isolates was predictive of growth rate, genome size, genome completeness and contamination, GC content, number and length of contigs, percentage of coding DNA, and number of sigma factors, paralogs, and SNPs with linear-mixed effects models using the lmer {lme4} function in R. To account for multiple *M. aeruginosa* isolates originating from the same lake, we used random effects terms for lake and collection date. All

percentage data was arcsine-square-root transformed to meet model assumptions. We also tested for a correlation between growth rate and total phosphorus concentration using linear regression. To test whether there was a correlation within the LL/LG group between genome completeness and percentage of polymorphic sites, we used robust linear regression (rlm {MASS} function in R) because it remains robust despite the presence of outliers or highly influential data points.

We next measured functional similarities among isolates of *M. aeruginosa* depending on lake origin and phylogenetic grouping. We calculated isolate dissimilarity using a Bray-Curtis distance metric on a matrix consisting of the number of genes within each of 1,820 different protein families. We excluded protein families that had zero variance across our 46 isolates. We tested whether isolates differed significantly in genome function by lake origin and phylogenetic grouping using analysis of variance of distance matrices with adonis {vegan}, which is a version of PERMANOVA that can accept categorical and continuous variables. We visualized clustering among isolates using principal coordinate analysis with the pcoa {ape} function in R. We also repeated these dissimilarity analyses with the bacterial phycosphere community, using both a Bray-Curtis distance metric of OTU community composition and a functional distance metric using protein families as done with host isolates. However, within the protein family phycosphere distance matrix, many pfams were not shared across samples (i.e., high beta-diversity or turnover). Such data structure often results in a strong arch or horseshoe shape in an ordination that is indicative of non-independent axes and is known as the Guttman effect. To facilitate data interpretation, we minimized the arching effect by using the stepacross flexible shortest path correction with a 'toolong' parameter of 0.75 {vegan} (Smith, 2017). Further, we then used these Bray-Curtis distances from the host and phycosphere data matrices to test whether host genomes that were more similar in function harbored phycosphere communities that were more similar in function. We assessed this association with a linear mixed-effects model using the lmer function in R and lake as the random effects term.

To determine whether certain protein families or Kegg Orthology terms within the host genome or phycosphere metagenome were associated with different phylogenetic groupings of *M. aeruginosa*, we ran analysis of variance models with a Benjamini-Hochberg false discovery rate correction in STAMP (Benjamini & Hochberg, 1995; Parks et al., 2014). To account for multiple samples collected within the same lake, rather than input gene counts for each host genome or phycosphere metagenome into these models, we instead averaged gene counts for each protein family across all genomes/metagenomes that both belonged to the same phylogenetic grouping and were collected from the same lake.

Results:

Structural Genome Variation within *M. aeruginosa*

Previously, we found two distinct phylogenetic groups of *M. aeruginosa* isolates from inland lakes of Michigan using a multi-locus sequence typing analysis (Berry et al., 2017). Here, we expand this analysis to include all publically available sequences of *M. aeruginosa* originating from multiple locations spanning six continents. We observed that our phylogenetic clustering of the 46 inland lake isolates corresponded with the trophic status of lake of origin, as determined by lake total phosphorus (Table S1). We note that although phosphorus and nitrogen can be co-limiting in lakes, we found no evidence that our phylogenetic clustering corresponded with either NH_4^+ or NO_3^- concentrations (Table S1), and therefore use the classic, phosphorus-based definition for assigning freshwater trophic status (see methods). We refer to these isolates in three categories that correspond to the trophic status of the lake of origin on the one hand, and this phylogenetic clustering on the other. Nineteen of the 20 NCBI genomes clustered with 17 isolates that originated from eutrophic and mesotrophic Michigan lakes, which we refer to as the 'High Phosphorus Lake/High Phosphorus Genotype' (HL/HG) group (Fig. S1). A second group contained an additional 29 isolates originating from oligotrophic, mesotrophic, and eutrophic Michigan lakes. Here, we subdivided this group into 18 isolates collected from oligotrophic lakes, which we refer to as the 'Low Phosphorus Lake/Low Phosphorus Genotype' (LL/LG) group, and into 11 isolates collected from eutrophic and mesotrophic lakes, which we refer to as the 'High Phosphorus Lake/Low Phosphorus Genotype' (HL/LG) group. Genomes within the LL/LG and HL/LG groups were less variable compared to other genomes within their respective groups (pairwise average nucleotide identity of strains within the group, $\mu = 98.5 \pm 0.035$ s.e. and 98.4 ± 0.082, respectively), while genomes within the HL/HG group were notably more variable (95.6 ± 0.09). Based on average nucleotide identity of genomes belonging to different groups, the LL/LG and HL/LG groups were most similar (98.1 ± 0.026) while HL/HG was more distant to LL/LG (95.2 ± 0.030) and HL/LG (95.3 ± 0.037).

Metagenomic assemblies of *M. aeruginosa* from the LL/LG group, and to a lesser extent, HL/LG, were more fragmented than the HL/HG group (Fig. S4; $F_{2,46} = 59.0$, $p < 0.001$; # contigs in LL/LG genomes: $\mu = 632 \pm 72$ s.d., HL/LG: $\mu = 548 \pm 31$ s.d., HL/HG: $\mu = 318 \pm 77$ s.d.). As assembly fragmentation is often caused by increased genomic heterogeneity, we determined the number of polymorphic sites in all assemblies. LL/LG isolates contained a higher percentage of polymorphic sites in their genomes compared to the other groups (Fig. 1A; $F_{2,46} = 15.4$, $p < 0.001$), and within these LL/LG isolates, higher percentages of polymorphic sites correlated with both lower estimated genome completeness (Fig. 1B; robust linear regression: $F_{1,16} = 5.7$, $p = 0.029$) and a greater number of fragments ($F_{1,16} = 14.1$, $p = 0.0018$, $R^2 = 0.43$). LL/LG assemblies also contained a higher proportion of low sequence

read coverage contigs (Fig. 1C, and Fig. S5). As further evidence that these low coverage contigs were the result of exceptionally high heterogeneity causing separate contigs to be generated, we found that genomes with many low coverage contigs had a greater median Reference:Alternate allele ratio at polymorphic sites located throughout the genome. This indicates that high heterogeneity within a sequence caused the assembly to divide divergent groups of reads into two separate contigs, which would decrease occurrence of within-contig heterogeneity (Fig. 1D, Fig. S6). Further, in the 7 LL/LG genomes with many low coverage contigs (shown in Fig. S6), these contigs shared 100% average nucleotide identity to contigs of higher coverage within the same genome. As higher fragmentation leads to smaller average contig size, the disparate levels of heterogeneity and assembly fragmentation between phylogenetic groups could lead to inaccurate comparisons in structural or functional genome variation across groups. Therefore, in addition to contigs 2 kb in length and longer included in our initial binning analysis, we added any contigs that were less than 2 kb in length but annotated as *M. aeruginosa*. This approach had minimal effects for the HL/HG or HL/LG groups, but added numerous genes that had initially appeared absent to the LL/LG group (Fig. S4, compares genome metrics with and without these shorter fragments; Fig. S7 lists core genes occurring on short fragments in LL/LG genomes). Unless noted otherwise, all subsequent genome descriptions use these modified bins that included *M. aeruginosa*-annotated fragments less than 2 kb.

Genomes from the LL/LG diverged subtly, though significantly, from the HL/HG group in several characteristics commonly assessed to detect genome streamlining. Further, the HL/LG group fell at intermediate levels in these streamlining characteristics, where HL/LG was more similar to LL/LG for certain characteristics and more similar to HL/HG for others. Genomes across groups were similar in size ($p = 0.13$), however LL/LG genomes were less complete than those in HL/HG (Fig. 2; $F_{2,46} = 17.40$, $p < 0.001$; LL/LG: $\mu = 96.77\% \pm 1.49$ s.d. versus HL/HG: $\mu = 99.27\% \pm 0.56$ s.d., and at intermediate completeness HL/LG: $\mu = 98.91 \pm 0.59$ s.d.). Eleven genes considered core to the cyanobacterial phylum, which includes the *Microcystis* genus, were absent from multiple LL/LG and HL/LG isolates (Fig. S7). LL/LG isolates frequently lacked core genes for the enzyme 3-dehydroquinate dehydratase in the shikimate pathway, which blocks the biosynthesis of the aromatic amino acids (i.e., tryptophan, tyrosine, and phenylalanine), and the enzyme acetolactate synthase, which inhibits synthesis of the branch-chain amino acids (i.e., valine, leucine, and isoleucine). We confirmed these core genes were absent on even very small contigs (i.e., < 2 kb) by searching scaffolds below 2 kb in length (see Fig. S7 for a comparison of core genes inferred absent before and after inclusion of these shorter fragments). Additionally, LL/LG and HL/LG isolates also had greater nitrogen-use efficiency via a slightly lower GC content (Fig. 2; $F_{2,46} = 7.71$, $p = 0.021$; LL/LG genomes: $\mu = 42.48\% \pm 0.049$ s.d., HL/LG genomes: $\mu = 42.51\% \pm 0.075$ s.d., compared to HL/HG genomes: $\mu = 42.71 \pm 0.15$). LL/LG and HL/LG genomes also contained a higher

percentage of coding versus non-coding DNA, while HL/HG was notably more variable (Fig. 2; $F_{2,46}$ = 5.00, p = 0.082; LL/LG genomes: μ = 81.40% ± 0.17 s.d. and HL/LG genomes: μ = 81.47% ± 0.26 s.d. compared to HL/HG genomes: μ = 81.13% ± 0.69 s.d.). Streamlined genomes often contain fewer paralogs, however we found similar percentages across groups (Fig. 2; $F_{2,46}$ = 4.81, p = 0.09). Lastly, HL/LG genomes contained significantly more sigma factors than HL/HG genomes (Fig. 2; $F_{2,46}$ = 8.78, p = 0.012; HL/LG genomes: μ = 0.291% ± 0.01 s.d. compared to HL/HG genomes: μ = 0.266% ± 0.03 s.d., while LL/LG genomes were not significantly different from either group: μ = 0.281% ± 0.02 s.d.). Note that metrics requiring gene identification, including % coding DNA, sigma factors, and paralogs, were inferred using only the 2 kb+ bins due to low accuracy of gene calling on shorter fragments.

Functional Genome Variation within *M. aeruginosa*

Isolates of *M. aeruginosa* differed in their genome-wide predicted gene functions based on protein families across the ten different lakes from which strains had originated (Note: four lakes were omitted as only one isolate per lake was sequences; Fig. 3A; PCoA on a Bray-Curtis dissimilarity of genes categorized by protein family; adonis $F_{9,40}$ = 8.56, p < 0.01, R^2 = 0.311). Further, the three phylogenetic groups of *M. aeruginosa* were functionally distinct (Fig. 3B; adonis $F_{2,45}$ = 9.01, p < 0.01, R^2 = 0.295). LL/LG and HL/LG genomes were more similar to each other than either was to HL/HG genomes. Congruent with our phylogenetic results, HL/HG showed more variability among isolates (Fig. 3B). Group clustering was driven in part by the frequent absence of 11 core genes from LL/LG and HL/LG genomes (Fig. S7), where core genes were determined via checkM (Parks et al., 2015). Gene abundances within 671 protein families also varied significantly across phylogenetic groups (Fig. S8; FDR-corrected p-values < 0.05). Focusing our analysis on nutrient uptake and metabolism, we found that 16 LL/LG isolates, originating from 3 of 4 oligotrophic lakes, contained the alkaline phosphatase phoA (pfam00245). In contrast, none of the 28 isolates originating from phosphorus-rich lakes contained the alkaline phosphatase phoA. Further, while all isolates contained at least one gene within pfam05787 annotated as alkaline phosphatase phoX, all 29 LL/LG and HL/LG genomes contained a second gene within this protein family, while only 1 of the 17 HL/HG isolates contained a second phoX gene. Also pertinent given that phosphorus-limitation in oligotrophic lakes often co-occurs with nitrogen limitation, we found that LL/LG and HL/LG genomes also contain additional genes for nitrate and nitrite transport (K15576-8; Fig. S8). Additionally, a nifU-like protein involved in iron binding and FeS cluster formation occurred in 8 of 18 LL/LG isolates across three different oligotrophic lakes but none occurred in the 27 HL/HG or HL/LG isolates. Further, considering the importance of colonial growth, we highlight the occurrence of several

genes regulating cell-cell recognition and adhesion in only LL/LG isolates. These genes include NeuB or sialic acid synthase, sialidases, an FRG1-like domain involved in underwater adhesion, and a gene in pfam03865 that is involved in secretion of adhesins.

Further, each phylogenetic group had different genes under positive selection. Genes under positive selection for the phylogenetic branch containing LL/LG isolates included a hybrid sensor histidine kinase/response regulator, chromosome segregation protein SMC, DNA-directed RNA polymerase subunit, NADPH-dependent glutamate synthase, and cell division protein ZipN/Ftn2/Arc6 (all FDR-corrected p-values < 0.05). In contrast, on the phosphorus-rich branch with a common ancestor of K13-05 and K13-06, we found evidence of positive selection for a gene annotated to encode a Fe-S cluster assembly protein SufB, a domain of unknown function (DUF748), pyruvate phosphate dikinase PEP/pyruvate-binding protein, and a biosynthetic arginine decarboxylase (all FDR-correct p-values < 0.05).

Taxonomic and Functional Variation in the *M. aeruginosa* Phycosphere

Given clear divergence among isolates of *M. aeruginosa* across a phosphorus gradient, we asked whether associated bacteria in the phycosphere of *M. aeruginosa* also varied across this gradient. First, we found community-level divergence among the phycospheres associated with the three phylogenetic groups of hosts (Fig. 4A; adonis $F_{2,49} = 2.4$, $p = 0.001$, $R^2 = 0.092$). Strongest separation was between the phycospheres of LL/LG hosts and that of both HL/HG and HL/LG hosts. We provide a taxonomic description of the core microbiome of *M. aeruginosa* in Fig. S9. Among HL/HG and HL/LG hosts, *Phycosocius bacilliformis* was the most abundant taxon in the phycosphere, comprising on average, 19.6% of the community (Table S3). In contrast, among LL/LG hosts, Caulobacterales and Cytophagaceae OTUs were more abundant, comprising 18.1 and 18.2% of the community respectively (Table S3).

Despite taxonomic differences among the phycosphere communities associated with hosts belonging to each phylogenetic group, we found no significant differences in protein functionality (Fig. 4B; adonis not significant). Phycosphere functionality also did not differ by lake of origin (Fig. S9C; adonis not significant). In general, the phycospheres associated with different phylogenetic groups contained few unique functions (Fig. S10). Phycospheres of LL/LG hosts showed increased FeS cluster binding, as well as increased biosynthetic capacity for fatty acids, serine, threonine, histidine, and ubiquinone (Fig. S10). Nif genes indicative of nitrogen-fixation were also more commonly associated with the phycospheres of HL/LG (81%: 9 of 11 isolates) and LL/LG (50%: 9 of 18) than HL/HG (35%: 6 of 17 isolates). However, 17 of these phycosphere communities contained only the nifA gene, which is the key

459 transcriptional regulator of nif genes, but may also regulate genes not involved in nitrogen fixation
460 (Nienaber et al., 2000).
461 We chose *P. bacilliformis* for more in-depth investigation of function because this taxon occurred
462 in all 46 phycospheres and was found, on average, at the greatest relative abundance via 16S marker gene
463 surveys (Fig. S9). We detected substantial genome variation within *P. bacilliformis*. In contrast to a lab
464 contaminant that would be identical in all cultures, we identified 7 genotypes that were each at least 96%
465 complete and at least 0.70% divergent from all other genotypes (Table S4). To shed light on the
466 predominance of *P. bacilliformis* in the *M. aeruginosa* phycosphere, we compared the metabolic
467 capacities of each of 7 *M. aeruginosa* hosts with each of their associated genotypes of *P. bacilliformis*. We
468 found that while no hosts harbored genes for the synthesis of threonine or serine, each *P. bacilliformis*
469 genome showed the capacity to biosynthesize each of these amino acids via L-serine synthesis from 3-
470 phosphoglycerate, L-threonine synthesis from L-homoserine, and L-homoserine synthesis from L-
471 aspartate. Additionally, only one *M. aeruginosa* genome showed evidence of asparagine synthesis, while
472 all *P. bacilliformis* genomes contained pathways for tRNA-dependent asparagine synthesis. Four genomes
473 of *P. bacilliformis* were also indicative of galactose utilization via the Leloir Pathway (Table S4).
474
475 Functional Interdependence between *M. aeruginosa* and the Phycosphere
476
477 In addition to assessing host and host microbiome function independently, we aimed to determine
478 whether genome variation within a host provides further predictive power beyond the lake environment
479 and time in predicting the functional capacity of the *M. aeruginosa* phycosphere. We selected 8 isolates
480 from oligotrophic Gull Lake and 5 isolates from eutrophic Wintergreen Lake collected on August 8, 2013.
481 This approach retained substantial genetic variation among host isolates on which to test our hypothesis.
482 Among Gull isolates, 46.5% of homologous gene clusters were shared across all 8 isolates and 59.6% of
483 clusters were shared across 7 isolates. Similarly, among Wintergreen isolates, 50.9% of homologous gene
484 clusters were shared across all 5 isolates and 68.6% of clusters were shared across 4 isolates. Functional
485 divergence between hosts showed a weak, positive correlation with functional divergence between
486 phycosphere communities (Fig. 4C: linear mixed-effects model, marginal $R^2 = 0.093$, $p = 0.060$ or
487 similarly, best fitting model with polynomial term, marginal $R^2 = 0.14$, $p = 0.065$). We show results from
488 the eutrophic and oligotrophic lake with the greatest number of isolates collected on a single date because
489 we found that there was a sharp reduction in detectability of a significant relationship as sample size
490 decreased. For example, among Gull Lake 2013 isolates, we tested all subsetted sample-sizes of 4, 5, 6,

and 7 isolates and found that the probability of finding a significant result was 15.7, 25, 60.7, and 100% respectively.

Physiological Variation within *M. aeruginosa*

One colony (source lake TP = 16.6 μg L^{-1}) exhibited negative growth during the experiment and was omitted from analysis, leaving n = 18 for the experiment. We found that the observed genetic differences in *M. aeruginosa* corresponded with physiological differences among isolates, with maximum intrinsic growth rate being significantly and positively correlated with the total phosphorus concentration (TP, an index of trophic status) of the source lake (linear regression: $n = 18$, $p = 0.030$, $R^2 = 0.26$). Observed growth rates ranged nearly 7-fold from 0.08 d^{-1} (source lake TP = 7.9 μg L^{-1}) to 0.55 d^{-1} (source lake TP = 196.1 μg L^{-1}). More specifically, LL/LG and HL/LG isolates tended to grow more slowly at saturating resource levels than HL/HG isolates (Fig. 5; linear mixed-effects regression: $F_{2,17} = 2.8$, $p = 0.053$, $R^2 = 0.27$; LL/LG isolates: $\mu_{max} = 0.162 \pm 0.038$ s.e., HL/LG isolates: $\mu_{max} = 0.198 \pm 0.012$, and HL/HG isolates: $\mu_{max} = 0.327 \pm 0.061$). This correlation between μ_{max} and lake TP remained evident when adding strains from an additional 11 Michigan lakes reported by Wilson et al. (2006) (see analysis in Fig. S11). In contrast to the evident correlation with lake TP, growth rate did not correspond with either NH_4^+ or NO_3^- concentrations of the source lake (linear regressions: $n = 18$, $p = 0.41$, $R^2 = 0.036$ and $n = 18$, $p = 0.70$, $R^2 = 0.011$ respectively). Further, the HL/HG group was notably variable in growth rate. Possible drivers of this variability are unclear. The four strains exhibiting slower growth rates originated from four different lakes. Further, strains most closely related to each other according to the phylogeny shown in Fig. 2 fell on opposite sides of the growth gradient (μ_{max} of W11-03 = 0.115 versus W11_06 = 0.489, and L211-101 = 0.218 versus L211-11 = 0.419).

Discussion:

Biodiversity at the within-species scale can have broad consequences for community structure and ecosystem function (Bolnick et al., 2011; Crutsinger et al., 2006; Whitham et al., 2006). Understanding factors that correspond with intraspecific variation across niche space is therefore key to understanding the maintenance of vital functions across environmental gradients. We show that intraspecific niche divergence can culminate from varied ecological and evolutionary factors. In *M. aeruginosa*, these contributing factors include genome evolution and frequently, an altered life history strategy of clonal vs. non-clonal colony formation. We also showed shifts in the host microbiome due to environmental

selection, as well as correlations between host genome variation and microbiome function. Independent, and perhaps synergistic, effects of each of these factors play important roles in the fitness of *M. aeruginosa* across a phosphorus gradient of freshwater lakes in Michigan. These results shed light on the potential importance of intraspecific variation in regulating dynamics of cyanobacterial harmful algal blooms.

Our study broadens understanding of this environmentally important cyanobacterium from genotypes of *M. aeruginosa* that are adapted to phosphorus-rich environments (Meyer et al., 2017) to genotypes of *M. aeruginosa* that evidently correspond with survival and colony formation of *Microcystis* in low-phosphorus environments. Architecture of these genomes is indicative of selective pressure towards more streamlined cellular efficiency. Although genome size is the most notable feature of streamlining theory, more broadly, streamlining refers to selection in resource-poor environments towards reductions in resource use, cell and genome size, and cell complexity (Giovannoni et al. 2014). Compared to well described examples of streamlining in marine cyanobacteria (i.e. *Prochlorococcus*, Rocap et al., 2013), our results in *M. aeruginosa* are much smaller in effect size. Subtle shifts in our dataset are not surprising considering we surveyed intraspecific differences across a relatively small geographic region.

Genomes from the LL/LG and HL/LG groups lacked potentially important functions compared to the HL/HG group. LL/LG and HL/LG genomes frequently lacked 11 core cyanobacterial genes, however the database that identifies 'core' genes consists mostly of *M. aeruginosa* isolates collected from eutrophic environments. LL/LG and HL/LG genomes frequently lacked cobS, which is necessary for synthesis of cobalamin, Vitamin B12. Frequent absence of mutT, involved in preventing DNA mutations, in LL/LG and HL/LG isolates suggests a higher mutational rate may be beneficial for adaptation to oligotrophic environments (Denamur & Matic, 2006). However, because mutT prevents AT to GC transversions (Yanofsky et al., 1966), it is unclear how LL/LG and HL/LG genomes lacking mutT would maintain a lower GC content than HL/HG isolates. Future study is warranted into whether subsequent or concurrent genetic changes occurred with the loss of mutT to facilitate increased mutational rate while maintaining low GC content.

In addition to gene loss and reduced genome complexity, acquisition of new gene functions may contribute to survival under phosphorus-limited environments. Most notable, we found that LL/LG isolates contained alkaline phosphatases that were absent in HL/HG and HL/LG genomes. Of three known alkaline phosphatase protein families (PhoA, PhoD, and PhoX (Luo et al., 2009)), only PhoX has been previously described in *M. aeruginosa* (Harke et al., 2012), and is also common among bacteria of the oligotrophic open ocean (Kathuria & Martiny, 2011; Luo et al., 2009; Sebastian & Ammerman, 2009). Two factors that may be facilitating LL/LG survival in low-P environments are that all LL/LG and HL/LG

isolates gained a second PhoX gene, and that LL/LG isolates gained an additional gene annotated as PhoA (pfam00245). Considering that PhoA requires activation by zinc and magnesium ions, while PhoX requires bioavailable calcium, different alkaline phosphatases may be more advantageous for *M. aeruginosa* inhabiting different environments.

We also had two notable findings in regards to histidine kinases and response regulators, which enable cells to detect and respond to changes in its environment such as nutrients and light. A gene annotated to histidine kinase KdpD was found to occur in significantly less abundance in the LL/LG group than the HL/HG or HL/LG groups, which is aligned with findings of a substantial loss of response regulation in the streamlined cyanobacterium *Prochlorococcus* (Mary and Vaulot, 2003). On the other hand, a gene annotated to histidine kinase SsrA was under positive selection only in the LL/LG group, which is aligned with the finding that certain histidine kinases are essential for survival of the cyanobacterium *Synechococcus* under extreme nutrient limitation (Schwarz and Grossman, 1998).

We also inferred differences in life history traits among phylogenetic groups. We infer that in contrast to clonal colony formation in HL/HG and HL/LG isolates, LL/LG *M. aeruginosa* colonies more frequently (though not exclusively) initially form through non-clonal cellular adhesion. Colony formation can occur either gradually via clonal growth or rapidly via cell adhesion, and both modes of colony formation can be induced by abiotic and biotic stressors including predation (Xiao et al., 2018). In light of lower growth rates among LL/LG isolates, cell adhesion may be compensate for slower clonal colony growth. The increased frequency of polymorphic sites observed in LL/LG genomes are beyond what could be explained by *de novo* mutations during lab culturing (~ 1,200 generations between field collection and sequencing (Baldia et al., 2007)). We also found similar heterogeneity between isolates collected in 2011 and 2013 despite differences in culturing time. Further supporting an increased occurrence of genomic heterogeneity among LL/LG colonies, we found increased frequencies of genes in LL/LG genomes that may facilitate non-clonal colony formation: (a) NeuB synthesizes sialic acids, which are a component of cyanobacterial extracellular polymeric substances (Zippel & Neu, 2011; Strom et al., 2017), and play important roles in cellular recognition and adhesion (Gunawan et al., 2005), (b) Sialidases may facilitate cellular adhesion by uncovering carbohydrate receptors that are recognized by bacterial adhesins (Vimr, 1994), (c) pfam06229 (FRG1-like domain) that contains a *Hydra spp.* gene linked to this freshwater cnidarian's ability to adhere to underwater surfaces (Rodrigues et al., 2016), and (d) a bacterial adhesin (pfam03865: Haemolysin secretion/activation protein ShlB/FhaC/HecB) (Moslavac et al., 2005) that facilitates adhesion to other cells in pathogenic and symbiotic interactions (Hooper & Gordon, 2001). Genes regulating cell recognition and adhesion should be less important for clonal colonies, where daughter cells remain attached after binary fission (Kessel & Eloff, 1975). We note that all colonies that

we isolated, including the evidently non-clonal LL/LG colonies, were of the distinctive, tightly-packed morphology typical of natural *M. aeruginosa* colonies. In contrast, experimentally induced non-clonal colonies are amorphous, loose aggregations of cells that do not resemble *M. aeruginosa* morphologies common in nature. The non-amorphous shapes of the non-clonal LL/LG colonies suggests that growth occurred mostly by cell division but may have included some level of aggregation only at earlier stages of colony development. Lastly, the colonies that we isolated may have been descendants from the initial colonies that formed via some level of cell aggregation.

We also note that the HL/HG group of *M. aeruginosa* was more highly variable than the LL/LG group, which corresponds to the variation in nutrient levels observed in high versus low nutrient lakes. HL/HG isolates were more variable according to our MLST phylogeny as well as the genome wide metrics, GC content and percentage of coding DNA. Greater genome variation in the HL/HG group also translated into greater variance in function, according to protein family annotation, as well as physiology (μmax). Low-nutrient lakes generally always have low concentrations of bioavailable nutrients, particularly in the summer epilimnion of deep, stratified lakes where co-limitation by both N and P may be especially common. In contrast, high-nutrient lakes have higher average concentrations than low-nutrient lakes, but also have much greater variance, including occasional low concentrations following periods of intense phytoplankton growth (Sarnelle, 1992). The lowest nutrient levels tend to occur in the epilimnion during late summer, where and when *M. aeruginosa* tends to reach peak abundance (Sarnelle, 1992).

In contrast to clear genomic divergence of *M. aeruginosa* across the phosphorus gradient, we found subtle changes in the phycosphere. Functional convergence despite taxonomic divergence as we see in this study has also been noted in other study systems, including the human gut microbiome (Turnbaugh et al. 2009), however there are also several limitations to our approach that may explain this result. First, although the functional component of our study analyzes pfam profiles based on near complete genomes of phycosphere bacteria, sequence variation of proteins belonging to the same pfam, either present in different species or in different strains of the same species may alter phycosphere functionality in ways not reflected by our current analysis. In addition, the phycosphere of *M. aeruginosa* may be similar in genome functions across the trophic gradient, but may express different genes when associated with hosts belonging to different phylogenetic groups. Analyses of these differences are beyond the scope of this study, but are worth further study. Second, to generate adequate quantities of DNA for sequencing, and to enable physiological characterization, all isolates were cultured under common garden, laboratory-based conditions. Phycosphere community composition likely changed during this transition from the natural environment. Yet, as a host genotype effect on overall taxonomic and specific functional gene content remained by both phylogenetic group and lake of origin, these changes were constrained by the limits

imposed by the natural community associated with each individual colony from which the cultures were started. Future studies repeating our analysis on sequence data generated directly from individual colonies collected from the environment will clarify the extent to which laboratory culturing impacted the host genotypic signal.

Despite these limitations, phycosphere community composition remained representative of heterotrophic bacteria associated with blooms of *M. aeruginosa*. For example, two of the three most abundant heterotrophic taxa across our isolate collection, *P. bacilliformis* and Cytophagaceae (Table S3), are strongly associated with blooms of *M. aeruginosa* (Tanabe et al., 2015; Berry et al., 2016). We had hypothesized that the phycosphere could facilitate survival of a streamlined host with atypical nutrient requirements caused by gene loss, e.g. the loss of key genes for amino acid biosynthesis in LL/LG isolates. Our hypothesis is based on the Black Queen Hypothesis which proposed that streamlined bacteria may compensate for gene loses through increased community connectivity (Morris et al., 2012). In contrast, we found convergence in phycosphere gene functions across the nutrient gradient. This convergence in phycosphere function may be driven in part to providing uniform culturing conditions with standardized nutrients and vitamins. In contrast to function, taxonomic composition of the phycosphere varied across the gradient. Considering a common garden likely caused some degree of convergence among phycospheres, this magnitude of taxonomic divergence likely underestimates divergence under natural environmental conditions. Functional convergence despite taxonomic divergence suggests that hosts may select for certain essential functions among the available pool of heterotrophic bacteria, which themselves are strongly shaped by lake environmental conditions (Crump et al., 2007). This host-mediated selection of phycosphere function was also apparent in how functional similarity of the host phycosphere was weakly predicted by the functional similarity of the host. Although this was a weak prediction that accounted for less than 10% of the total variance among phycosphere functionality, these results correspond with findings that intraspecific plant variation can have small but significant influences on the rhizosphere of maize and *Arabidopsis* genotypes (Lundberg et al., 2012; Peiffer et al., 2013). Different species of phytoplankton hosts have been shown to harbor distinct phycosphere communities (Jasti et al. 2005; Eigemann et al. 2013), but the relative effects of intra- versus interspecific-level variation on the phycosphere has not been directly studied.

One notable functional contribution of LL/LG and HL/LG *M. aeruginosa* phycospheres that may explain survival across a nutrient gradient was an increased occurrence of nitrogen-fixation genes. However, much of this pattern was driven by the increased occurrence of only a single nif gene, nifA, so it is unclear what role these phycosphere taxa may contribute towards nitrogen assimilation. Additionally, genome investigation into *Phycosocius bacilliformis*, which we found in the phycosphere of each of our

M. aeruginosa isolates, suggests that complementary amino acid biosynthesis may be one component of this symbiotic interaction. Previously identified in surveys of *M. aeruginosa* blooms, *P. bacilliformis* can increase growth of colonial green algae (Tanabe et al., 2015). Several genotypes of *P. bacilliformis* appear to derive a significant source of energy from galactose, which is the primary component of the polysaccharide-based mucilage that binds cells of colonial *M. aeruginosa* (Plude et al., 1991; Rohrlack et al., 1999).

From a methodological perspective, our case study emphasizes the need for cautious interpretation of metagenome-assembled genomes. Comparing genome traits and functions across phylogenetic groups was challenging due to fundamental differences in genome architecture and assembly. For example, repetitive elements are a well known technical complication in genome assembly (Treangen & Salzberg, 2012). Our LL/LG assemblies were the most fragmented due to a larger number and/or problematic locations of repetitive elements, as well as a higher incidence of colony heterogeneity. *Microcystis aeruginosa* is noted for an unusually high percentage of mobile repetitive elements (Kaneko et al., 2007). Such varied genomic architecture can lead to biased inferences. For example, when we initially binned only contigs over 4kb in length, because LL/LG genomes contained a sizable portion of contigs under 4kb in length, we incorrectly inferred that LL/LG genomes were smaller in size and lacked many more gene functions. In-depth investigation of these patterns revealed that many genes seemingly missing from LL/LG genomes were merely on contigs shorter than 4kb. This lead us to the cautious approach of binning 2 kb fragments and adding even shorter contigs that were taxonomically classified as *M. aeruginosa*.

Genome data of *M. aeruginosa*, which dominates freshwater harmful algal blooms worldwide, has previously existed only for isolates that originated from phosphorus-rich environments. Our findings that *M. aeruginosa* inhabiting oligotrophic environments differ in genome structure, function, and life history compared to isolates derived from eutrophic environments, has important implications for understanding the dynamics of harmful algal blooms and their response to ongoing global change. Particularly notable, we found that divergence among *M. aeruginosa* isolates results in a growth tradeoff. HL/HG isolates retain the ability for rapid growth when resources are high (i.e., a greater μ_{MAX} that matches rate estimates reported by Wilson et al. 2006 & 2010, Reynolds 2006, Seip and Reynolds 1995 for colonial *Microcystis*). LL/LG and HL/LG isolates have acquired the ability to subsist when resources are low but at the cost of an ability to increase growth when resources become more readily available. Similar growth tradeoffs have been observed in streamlined bacteria (Giovannoni et al., 2005). This tradeoff is regarded as a key property of oligotrophs versus copiotrophs, which instead have growth rates more responsive to nutrient flux (Koch, 2001). Further, this growth trade-off is an important future direction to consider in the context of harmful algal blooms. These blooms can last several months, during which time bloom development

drives down available nutrients in the water column to low levels (Heisler et al., 2008; Sarnelle, 1992), thus constructing a niche for low-nutrient adapted genotypes. Our findings are especially notable considering these distinct architectures occurred within a species of cyanobacterium within a single lake. Co-occurrence of such isolates adapted to thrive in different micro-environments may have critical implications for the temporal variability and spatial extent of toxic cyanobacterial blooms. For example, the HL/LG-type isolates residing in eutrophic and mesotrophic lakes may be key players in extending the duration of blooms after HL/HG-type isolates have depleted phosphorus to levels that would otherwise lead to a recovery period or dominance of other non-toxic phytoplankton. Such physiological adaptation to low-phosphorus conditions also helps explain the recent expansion of *M. aeruginosa* in oligotrophic lakes invaded by dreissenid mussels, lakes which are otherwise an uncharacteristic habitat for this cyanobacterium (Raikow et al., 2004; Knoll et al. 2008; Sarnelle et al., 2010).

Overall, evolutionary divergence of *M. aeruginosa* corresponds with maintenance of high fitness across a wide phosphorus gradient. Evolutionary changes included direct effects on the host genome that increased nutrient-use efficiency and nutrient assimilation. However, genomic changes within the organism do not operate independently from its ecology. Changes within *M. aeruginosa* genomes may have facilitated changes in the behavioral ecology of the cyanobacterium by acquiring gene functions that have seemingly enabled an altered life history strategy of non-clonal colony formation. Host genome changes further correspond with changes in the symbiotic and/or commensal ecological interactions between the host and the host microbiome, in which function of the host is linked with function of the host microbiome. These findings demonstrate the intricate and non-independence of ecological and evolutionary processes that may facilitate intraspecific niche divergence.

Acknowledgments:

This project was supported by funding from NOAA distributed through the Cooperative Institute for Great Lakes Research (NA17OAR4320152) and the National Science Foundation (Division of Environmental Biology-1737680) to VJD, a Dow Sustainability Postdoctoral Fellowship to SLJ, the National Science Foundation (Division of Environmental Biology-0841864, Division of Environmental Biology-0841944) to OS, and the Gull Lake Quality Organization and the Robert C. Ball and Betty A. Ball Fisheries and Wildlife Fellowship at Michigan State University to JDW.

References:

Anantharaman, K., Duhaime, M. B., Breier, J. A., Wendt, K. A., Toner, B. M. & Dick, G. J. (2014). Sulfur oxidation genes in diverse deep-sea viruses. *Science*, *344*, 757-760.

Baldia, S. F., Evangelista, A. D., Aralar, E.V., & Santiago, A. E. (2007). Nitrogen and phosphorus utilization in the cyanobacterium *Microcystis aeruginosa* isolated from Laguna de Bay, Philippines. *Journal of Applied Phycology, 19,* 607-613.

Barrett, R. D. H, & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution 23,* 38-44.

Bassar, R. D., Marshall, M. C., López-Sepulcre, A., Zandonà, E., Auer, S. K., Travis, J., Pringle, C. M., Flecker, A. S., Thomas, S. A., Fraser, D. F., & Reznick, D. N. (2010). Local adaptation in Trinidadian guppies alters ecosystem processes. *Proceedings of the National Academy of Sciences, 107,* 3616-3621.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B, 57,* 289-300.

Bergmann, G. T., Bates, S. T., Eilers, K. G., Lauber, C. L., Caporaso, J. G., Walters, W. A., Knight, R., & Fierer, N.. (2011). The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biology and Biochemistry, 43,* 1450-1455.

Berry, M. A., White, J. D., Davis, T. W., Jain, S., Johengen, T. H., Dick, G. J., Sarnelle, O., & Denef, V. J. (2017). Are oligotypes meaningful ecological and phylogenetic units? A case study of *Microcystis* in freshwater lakes. *Frontiers in Microbiology, 8,* 1-7.

Berry, M. A., Davis, T. W., Cory, R. M., Duhaime, M. B., Johengen, T. H., Kling, G. W., Marino, J. A., Den Uyl, P. A., Gossiaux, D., Dick, G. J., & Denef, V. J. (2017). Cyanobacterial harmful algal blooms are a biological disturbance to western Lake Erie bacterial communities. *Environmental Microbiology, 19,* 1149 – 1162.

Bolnick, D. I., Amarasekare, P., Araújo, M. S., Bürger, R., Levine, J. M., Novak, M., Rudolf, V. H. M., Schreiber, S. J., Urban, M. C., & Vasseur, D. A. 2011. Why intraspecific trait variation matters in community ecology. *Trends in Ecology and Evolution, 26,* 183-192.

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods, 12,* 59-60.

Burkholder, J. M., & Gilbert, P. M. (2009). The importance of intraspecific variability in harmful algae- preface to a collection of topical papers. *Harmful Algae, 8,* 744-745.

Cai, H., Jiang, H., Krumholz, L. R., & Yang, Z. (2014). Bacterial community composition of size-fractioned aggregates within the phycosphere of cyanobacterial blooms in a eutrophic freshwater lake. *Plos ONE, 9,* e102879.

Chislock, M. F., O. Sarnelle, B. Olsen, E. Doster, & A. E. Wilson. 2013. Large effects of consumer offense on ecosystem structure and function. Ecology 94: 2375-2380

Contreras-Moreira, B., Vinuesa, P. (2013). GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pan-genome analysis. *Applied and Environmental Microbiology, 79,* 7696-7701.

Costas, E., Lopez-Rodas, V., Javier Toro, F., & Flores-Moya, A. (2008). The number of cells in colonies of the cyanobacterium *Microcystis aeruginosa* satisfies Benford's Law. *Aquatic Botany, 89,* 341-343.

Crump, B. C., Adams, H. E., Hobbie, J. E., & Kling, G. W. (2007). Biogeography of bacterioplankton in lakes and streams of anarctic tundra catchment. *Ecology, 88,* 1365-1378.

Crutsinger, G. M., Collins, M. D., Fordyce, J. A., Gompert, Z., Nice, C. C, & Sanders, N. J. (2006). Plant genotypic diversity predicts community structure and governs an ecosystem process. *Science, 313,* 966-968.

DeMott, W. R., & Mckinney, E. N. (2015). Use it or lose it? Loss of grazing defenses during laboratory culture of the digestion-resistant green alga *Oocystis*. *Journal of Plankton Research, 37,* 399-408.

Denamur, E., & Matic, I. (2006). Evolution of mutation rates in bacteria. *Molecular Microbiology, 60,* 820-827.

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., & Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biology, 10,* R85.

Dittami, S. M., Duboscq-Bidot, L., Perennou, M., Gobet, A., Corre, E., Boyen, C., & Tonon, T. (2016). Host–microbe interactions as a driver of acclimation to salinity gradients in brown algal cultures. *ISME Journal, 10,* 51-63.

Drake, J. W., Charlesworth, B., Charlesworth, D., & Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics, 148,* 1667-1686.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32,* 1792-1797.

Eigemann, F., Hilt, S., Salka, I. & Grossart, H.-P. (2013). Bacterial community composition associated with freshwater algae: Species specificity versus dependency on environmental conditions and source community. *FEMS Microbiology Ecology, 32,* 650 - 663.

Franguel, L., Quillardet, P., Castets, A., Humbert, J., Matthijs, H. C. P., Cortez, D., Tolonen, A., Zhang, C., Gribaldo, S., Kehr, J., Zilliges, Y., Ziemert, N., Becker, S., Talla, E., Latifi, A., Billault, A.,

Lepelletier, A., Dittmann, E., Bouchier, C., & Tandeau de Marsac, N. (2008). Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics, 9,* 274.

Giovannoni, S. J., Thrash, J. C., & Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *ISME Journal, 8,* 1553-1565.

Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptiasa, D., Bibbs, L. Eads, J., Richardson, T. H., Noordewier, M., & Rappé, M. S. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science, 309,* 1242-1245.

Gunawan, J., Simard, D., Gilbert, M., Lovering, A. L., Wakarchuk, W. W., Tanner, M. E., & Strynadka, N. C. J. (2005). Structural and mechanistic analysis of sialic acid synthase NeuB from *Neisseria meningitidis* in complex with Mn2+, phosphoenolpyruvate, and N-acetylmannosaminitol. *Journal of Biological Chemistry, 280,* 3555-3563.

Harke, M. J., Berry, D. L., Ammerman, J. W., & Gobler, C. J. (2012). Molecular response of the bloom-forming cyanobacterium, *Microcystis aeruginosa*, to phosphorus limitation. *Microbial Ecology, 63,* 188-198.

Heisler, J., Glibert, P. M., Burkholder, J. M., Anderson, D. M., Cochlan, W., Dennison, W.C., Dortch, Q., Gobler, C. J., Heil, C. A., Humphries, E., Lewitus, A., Magnien, R., Marshall, H. G., Sellner, K., Stockwell, D. A., Stoecker, D. K., & Suddleson, M. (2008). Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae*, *8*, 3-13.

Hooper, L. V., & Gordon, J. I. (2001). Glycans as legislators of host–microbial interactions: Spanning the spectrum from symbiosis to pathogenicity. *Glycobiology, 11,* 1-10.

Humbert, J. F., Barbe, V., Latifi, A., Gugger, M., Calteau, A., Coursin, T., Lajus, A., Castelli, V., Oztas, S., Samson, G. & Longin, C. (2013). A tribute to disorder in the genome of the bloom-forming freshwater cyanobacterium *Microcystis aeruginosa*. *PLoS One*, *8*, p.e70747.

Huntemann, M., Ivanova, N. N., Mavromatis, K., Tripp, H. J., Paez-Espino, D., Palaniappan, K., Szeto, E., Pillay, M., Chen, I. M., Pati, A., Nielsen, T., Markowitz, V. M., & Kyrpides, N. C. (2015). The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Standards in Genomic Sciences, 10,* 86.

Jackrel, S. L., White, J. D., Evans, J. T., Buffin, K., Hayden, K., Sarnelle, O. & Denef, V. J. (2019). Microcystis cultures isolated from Michigan inland lakes genome sequencing, assembly, and targeted locus. *NCBI SRA*.

Jasti, S., Sieracki, M. E., Poulton, N. J., Giewat, M. W., & Rooney-Varga, J. N. (2005). Phylogenetic diversity and specificity of bacteria closely a*ssociated with Alexandrium* spp. and other phytoplankton. *Applied and Environmental Microbiology, 71,* 3483-3494.

Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. S., & Chisholm, S. W. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science, 311,* 1737-1740.

Joshi, N. A., & Fass, J. N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. V.1.33.

Kaneko, T., Nakajima, N., Okamoto, S., Suzuki, I., Tanabe, Y., & Tamaoki, M., Nakamura, Y., Kasai, F., Watanabe, A., Kawashima, K., Kishida, Y., Ono, A., Shimizu, Y., Takahashi, C., Minami, C., Fujishiro, T., Kohara, M., Nakazaki, N., Nakayama, S., Yamada, M., Tabata, & Watanabe, M. M. (2007). Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Research, 14,* 247-256.

Kathuria, S., & Martiny, A. C. (2011). Prevalence of a calcium-based alkaline phosphatase associated with the marine cyanobacterium *Prochlorococcus* and other ocean bacteria. *Environmental Microbiology, 13,* 74-83.

Kessel, M., & Eloff, J. N. (1975). The ultrastructure and development of the colonial sheath of *Microcystis marginata*. *Archives of microbiology*, *106*, 209-214.

Knoll, L. B., Sarnelle, O. Hamilton, S. K., Scheele, C. E. H., Wilson, A. E., Rose, J. B., & Morgan, M. R. (2008). Invasive zebra mussels (*Dreissena polymorpha*) increase cyanobacterial toxin concentrations in low-nutrient lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*, 448-455.

Koch, A. L. (2001). Oligotrophs versus copiotrophs. *Bioessays, 23*, 657-661.

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K, & Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology,79,* 5112-5120.

Laczny, C. C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H. H., Cordonado, S., van der Maaten, L., Vlassis, N & Wilmes P. (2015). VizBin-an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome, 3,*1.

Lakeman, M. B., von Dassow, P., & Cattolico, R. A. (2009). The strain concept in phytoplankton ecology. *Harmful Algae, 8,* 746-758.

Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., Promerova, M., Rubin, C., Wang, C., Zamani, N., Grant, B. R., Grant, P. R., Webster, M. T., & L. Andersson. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature, 518,* 371.

Lau, J.A., & Lennon, J. T. (2012). Rapid responses of soil microorganisms improve plant fitness in novel environments. *Proceedings of the National Academy of Sciences, 109,* 14058-14062.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25,* 2078-2079.

Louati, I., Pascault, N., Debroas, D., Bernard, D., Humbert, J. F., & Leloup, J. (2015). Structural diversity of bacterial communities associated with bloom-forming freshwater cyanobacteria differs according to the cyanobacterial genus. *Plos ONE, 10,* e0140614.

Lundberg, D. S., Lebeis, S. L., Herrera Paredes, S., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrekston, A., Kunin, V., Glavina del Rio, T., Edgar, R. C., Eickhorst, T., Ley, R. E., Hugenholtz, P., Green Tringe, S., & Dangl J. L. (2012). Defining the core *Arabidopsis thaliana* root microbiome. *Nature, 488,* 86-90.

Luo, H., Benner, R., Long, R. A., & Hu, J. (2009). Subcellular localization of marine bacterial alkaline phosphatases. *Proceedings of the National Academy of Sciences, 106,* 21219-21223.

Masango, M. G., Myrburgh, J. G., Labuschagne, L., Govender, D., Bengis, R. G., & Naicker, D. (2010). Assessment of *Microcystis* bloom toxicity associated with wildlife mortality in the Kruger National Park, South Africa. *Journal of Wildlife Diseases, 46,* 95-102.

McMurdie, P. J., & Homes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PloS Computational Biology*, 10, e1003531.

Menzel, D. W., & Corwin, N. (1965). The measurement of total phosphorus in seawater based on the liberation of organically bound fractions by persulfate oxidation 1. *Limnology and Oceanography, 10,* 280-282.

Meyer, K. A., Davis, T. W., Watson, S. B., Denef, V. J., Berry, M. A., & Dick, G. J. (2017). Genome sequences of lower Great Lakes *Microcystis* sp. reveal strain-specific genes that are present and expressed in western Lake Erie blooms. *PloS ONE, 12,* e0183859.

Michalak, A. M., Anderson, E. J., Beletsky, D., Boland, S., Bosch, N.S., Bridgeman, T.B., Chaffin, J.D., Cho, K., Confesor, R., Daloğlu, I., DePinto, J.V., et al. (2013). Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. *Proceedings of the National Academy of Sciences, 110,* 6448-6452.

Morris, J. J., Lenski, R. E., & Zinser, E. R. (2012). The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio*, *3*, e00036-12.

Moslavac, S., Mirus, O., Bredemeier, R., Soll, J., von Haeseler, A., & Schleiff, E. (2005). Conserved pore-forming regions in polypeptide-transporting proteins. *The FEBS Journal, 272,* 1367-1378.

Murphy, J., & Riley, J. P. (1962). A modified single solution method for the determination of phosphate in natural waters. *Analytica Chimica Acta, 27,* 31-36.

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., & Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews, 75,* 14–49.

Nienaber, A., Huber, A., Göttfert, M., Hennecke, H., & Fischer, H. M. (2000). Three new NifA-regulated genes in the *Bradyrhizobium japonicum* symbiotic gene region discovered by competitive DNA-RNA hybridization. *Journal of Bacteriology, 182,* 1472-1480.

Olm, M. R., Brown, C. T., Brooks, B., & Banfield, J. F. (2017). dRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME Journal, 11,* 2864-2868.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research, 25,* 1-13.

Parks, D. H., Tyson, G. W., Hugenholtz, P., & Beiko, R. G. (2014). STAMP: Statistical analysis of taxonomic and functional profiles. *Bioinformatics, 30,* 3123-3124.

Peiffer, J. A., Spor, A., Koren, O., Jin, Z., Green Tringe, S., Dangl, J. L., Buckler, E. S., & Ley, R. E. (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. Proceedings of the National Academy of Sciences, 110, 6548-6553.

Pfennig, D. W., Wund, M. A., Snell-Rood, E. C., Cruickshank, T., Schlichting, C. D., & Moczek, A.P. 2010. Phenotypic plasticity impacts on diversification and speciation. *Trends in Ecology & Evolution 25,* 459-467.

Plude, J. L., Parker, D. L., Schommer, O. J., Timmerman, R. J., Hagstrom, S. A., Joers, J. M., & Hnasko, R. (1991). Chemical characterization of polysaccharide from the slime layer of the cyanobacterium *Microcystis flos-aquae* C3-40. *Applied and Environmental Microbiology, 57,* 1696-1700.

Post, D. M., Palkovacs, E. P., Schielke, E. G., & Dodson, S. I. (2008). Intraspecific variation in a predator affects community structure and cascading trophic interactions. *Ecology, 89,* 2019-2032.

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., & Toth, I. K. (2016). Genomics and taxonomy in diagnostics for food security: Soft-rotting enterobacterial plant pathogens. *Analytical Methods, 8,* 12 – 24.

915 Qin, B., Zhu, G., Gao, G., Zhang, Y., Li, W., Paerl, H. W., & Carmichael, W. W. (2010). A
916 drinking water crisis in Lake Taihu, China: Linkage to climatic variability and lake management.
917 *Environmental Management, 45,* 105-112.
918 Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2012). The SILVA
919 ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids*
920 *research*, *41*, D590-D596.
921 Raikow, D. F., Sarnelle, O., Wilson, A. E., & Hamilton, S.K. (2004). Dominance of the noxious
922 cyanobacterium *Microcystis aeruginosa* in low-nutrient lakes is associated with exotic zebra mussels.
923 *Limnology and Oceanography, 49,* 482-487.
924 Rambaut, A. (2012). FigTree v1. 4. Available via http://tree.bio.ed.ac.uk/software/ figtree/.
925 Accessed Dec 12, 2018.
926 Richter, M., & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic
927 species definition. *Proceedings of the National Academy of Sciences, 106,* 19126-19131.
928 Reynolds, C. S. (2006). The ecology of phytoplankton. Cambridge University Press.
929 Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A. et al. (2003).
930 Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature*, *424*,
931 1042-1047.
932 Rohrlack, T., Henning, M., & Kohl, J. G. (1999). Mechanisms of the inhibitory effect of the
933 cyanobacterium *Microcystis aeruginosa* on *Daphnia galeata*'s ingestion rate. *Journal of Plankton*
934 *Research*, *21*, 1489-1500.
935 Sahm, A., Berns, M., Platzer, M., & Szafranski, K. (2017). PosiGene: Automated and easy-to-use
936 pipeline for genome-wide detection of positively selected genes. *Nucleic Acids Research, 45,* 1-11.
937 Sarnelle, O., Morrison, J., Kaul, R., Horst, G., Wandell, H. & Bednarz, R. (2010). Citizen
938 monitoring: Testing hypotheses about the interactive influences of eutrophication and mussel invasion on
939 a cyanobacterial toxin in lakes. *Water Research*, *44,* 141-150.
940 Sarnelle, O., Wilson, A. E., Hamilton, S. K., Knoll, L. B., & Raikow, D.F. (2005). Complex
941 interactions between the zebra mussel, *Dreissena polymorpha*, and the harmful phytoplankter, *Microcystis*
942 *aeruginosa*. *Limnology and Oceanography, 50,* 896-904.
943 Sarnelle, O. (1992). Contrasting effects of *Daphnia* ratios of nitrogen to phosphorus in a
944 eutrophic, hard-water lake. Limnology and Oceanography 37:1527-1542.
945 Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B. et al.
946 (2009). Introducing mothur: open-source, platform-independent, community-supported software for

947 describing and comparing microbial communities. *Applied and Environmental Microbiology, 75*, 7537-
948 7541.
949 Sebastian, M., & Ammerman, J. W. (2009). The alkaline phosphatase PhoX is more widely
950 distributed in marine bacteria than the classical PhoA. *ISME Journal, 3,* 563.
951 Seip, K. L., & Reynolds, C. S. (1995). Phytoplankton functional attributes along trophic gradient
952 and season. *Limnology and Oceanography, 40,* 589-597.
953 Seymour, J. R., Amin, S. A., Raina, J., & Stocker, R. (2017). Zooming in on the phycosphere: the
954 ecological interface for phytoplankton–bacteria relationships. *Nature Microbiology, 2*, 17065.
955 Shapiro, B. J., & Polz, M. F. (2014). Ordering microbial diversity into ecologically and
956 genetically cohesive units. *Trends in Microbiology, 22,* 235 – 247.
957 Smith, R. J. (2017). Solutions for loss of information in high-beta-diversity community data.
958 *Methods in Ecology and Evolution*, *8*, 68-74.
959 Soranno, P. A., Bacon, L. C., Beauchene, M. Bednar, K. E. Bissell, E. G., Boudreau, C. K., Boyer
960 M. G. et al. (2017). LAGOS-NE: A multi-scaled geospatial and temporal database of lake ecological
961 context and water quality for thousands of US lakes. *GigaScience*, *6*, 12.
962 Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with
963 thousands of taxa and mixed models. *Bioinformatics, 22,* 2688-2690.
964 Steffen, M. M., et al. (2017). Ecophysiological examination of the Lake Erie *Microcystis* bloom in
965 2014: Linkages between biology and the water supply shutdown of Toledo, OH. *Environmental Science*
966 *and Technology, 51,* 6745-6755.
967 Sterner, R. W. (2008). On the phosphorus limitation paradigm of lakes. *International Review of*
968 *Hydrology, 93, 433-445.*
969 Strom, S., Bright, K., Fredrickson, K., & Brahamsha, B. (2017). The *Synechococcus* cell surface
970 protein SwmA increases vulnerability to predation by flagellates and ciliates. *Limnology and*
971 *Oceanography, 62,* 784-794.
972 Tanabe, Y., Okazaki, Y., Yoshida, M., Matsuura, H., Kai, A., Shiratori, T., Ishida, K., Nakano, S.,
973 & Watanabe, M. M. (2015). A novel alphaproteobacterial ectosymbiont promotes the growth of the
974 hydrocarbon-rich green alga *Botryococcus braunii*. *Scientific Reports, 5,* 10467.
975 Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing:
976 computational challenges and solutions. *Nature Reviews Genetics, 13,* 36-46.
977 Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M.
978 L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., &
979 Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, *457*, 480-484.

Vimr, E. R. (1994). Microbial sialidases: Does bigger always mean better? *Trends in Microbiology, 2,* 271-277.

Vox, M., Hesselman, M. C., te Beek, T. A., van Passel, M. W. J., & Eyre-Walker, A. (2015). Rates of lateral genes transfer in prokaryotes: High but why? *Trends in Microbiology, 23,* 598-605.

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology,* 5261-5267.

Wehr, J. D., & Sheath, R. G. (2003). Freshwater algae of North America: Ecology and classification. Academic Press, San Deigo.

Wetzel, R. G. (2001). Limnology: Lake and River Ecosystems. Ed. 3. Elsevier Academic Press, San Diego, CA.

White, J. D., Kaul, R. B., Knoll, L. B., Wilson, A. E., & Sarnelle, O. (2011). Large variation in vulnerability to grazing within a population of the colonial phytoplankter, *Microcystis aeruginosa*. *Limnology and Oceanography, 56,* 1714-1724.

Whitham, T. G., Bailey, J. K., Schweitzer, J. A., Shuster, S. M., Bangert, R. K., LeRoy, C. J., Lonsdorf, E. V., Allan, G. J., DiFazio, S. P., Potts, B. M., Fisher, D. G., Gehring, C. A., Lindroth, R. L., Marks, J. C. , Hart, S. C., Wimp, G. M., & Wooley, S. C. (2006). A framework for community and ecosystem genetics: From genes to ecosystems. *Nature Reviews Genetics, 7,* 510-523.

Wilson, A. E., Kaul, R. B., & Sarnelle, O. (2010). Growth rate consequences of coloniality in a harmful phytoplankter. *PLoS ONE*, 5, e8679.

Wilson, A. E., Sarnelle, O., Neilan, B. A., Salmon, T. P., Gehringer, M. M., & Hay M. E. (2005). Genetic variation of the bloom-forming cyanobacterium *Microcystis aeruginosa* within and among lakes: Implications for harmful algal blooms. *Applied and Environmental Microbiology, 71,* 6126-6133.

Wilson, A. E., W. A. Wilson, & Hay, M. E. (2006). Intraspecific variation in growth and morphology of the bloom-forming cyanobacterium *Microcystis aeruginosa*. *Applied & Environmental Microbiology, 72,* 7386-7389.

Xiao, M., Li, M., & Reynolds, C. S. (2018). Colony formation in the cyanobacterium *Microcystis*. *Biological Reviews, 93,* 1399-1420.

Yanofsky, C., Cox E. C., & Horn, V. (1966). The unusual mutagenic specificity of an *E. coli* mutator gene. *Proceedings of the National Academy of Sciences, 55,* 274-281.

Zippel, B., & Neu, T. R. (2011). Characterization of glycoconjugates of extracellular polymeric substances in tufa-associated biofilms by using fluorescence lectin-binding analysis. *Applied and Environmental Microbiology, 77,* 505-516.

Data Accessibility:

- Metagenomes sequences: NCBI SRA study # SRP092358, accession # SRR9658105 - SRR9658151.
- 16S Sequences: NCBI SRA project # PRJNA351875, accession # SRX6419375 - SRX6419329.
- Metadata and analysis scripts: https://github.com/sjackrel/Microcystis_Inland_Lake_Genome_Evolution.

Author Contributions:

S.L.J. analyzed all sequencing data; J.D.W. collected all *M. aeruginosa* samples and field metadata, maintained lab cultures, performed growth rate experiments, and extracted DNA; J.T.E. created a custom genome annotation pipeline; K.B. and K.H. extracted *M. aeruginosa* bins from metagenome datasets and evaluated experimental tools; J.D.W., O.S. and V.J.D. designed study; S.L.J. and V.J.D wrote the manuscript; J.D.W. and O.S. contributed to editing the manuscript.

FIGURE 1 Population genomic analysis of *Microcystis aeruginosa* sequence bins. (A) Polymorphic sites, including single nucleotide variants, insertions, and deletions, are more common in genomes of the Low Phosphorus Lake/Low Phosphorus Genotype (LL/LG) phylogenetic groups, suggesting non-clonal heterogeneity. (B) Within the LL/LG group, frequency of polymorphic sites within the genome was inversely correlated with genome completeness estimates by checkM. Note we used robust linear regression to ensure that statistical significance was not overly influenced by the single value in the bottom right-hand corner, however excluding this point entirely renders the association insignificant. (C) Extensive polymorphisms due to isolate heterogeneity lead to many low-coverage contigs representing alternate assembly paths in LL/LG genomes and was visually evident as a bimodal distribution of contig coverage. X-axis is a standardized Z-score with $\mu = 0$, s.d. = 1. (D) This frequent division of a single contig into two separate contigs when extensive polymorphism occurred caused an increase in the genome-wide percent occurrence of the reference versus alternate allele variant in the main sequence bin, as sequence reads carrying the alternate nucleotide were now no longer aligned to the main sequence bin contigs but to the corresponding low coverage contigs. The Y-axis sums forward (F) and reverse (R) strands using the equation: (F+R of Reference)/[(F+R of Reference)+(F+R of Alternate)].

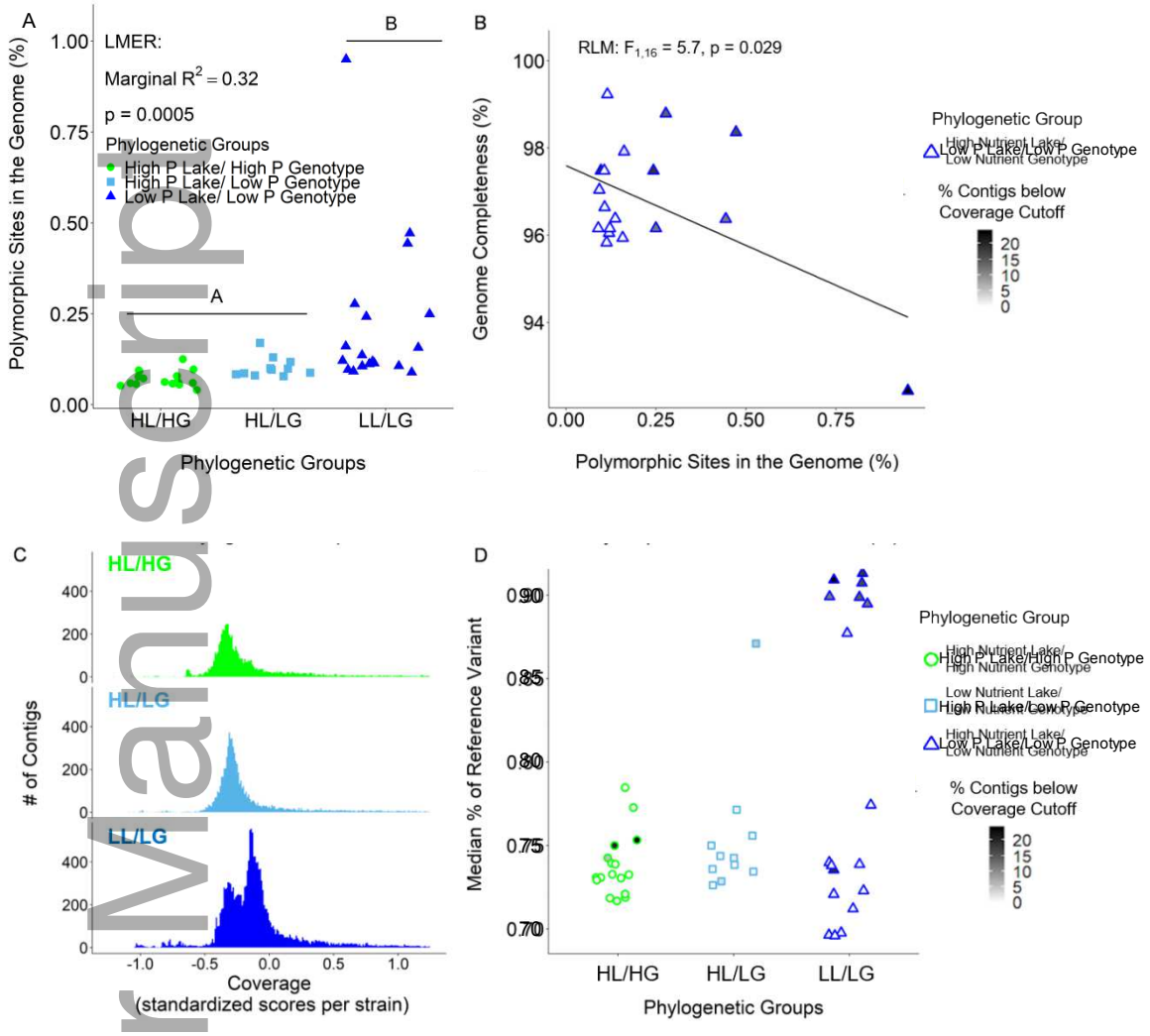

FIGURE 2 Divergent genome structure across a phylogeny of 46 isolates of *Microcystis aeruginosa* collected from 14 inland lakes in Michigan, USA. Multi-locus sequencing typing was used to infer evolutionary history with RAxML based on five concatenated housekeeping genes (FtsZ, glnA, gltX, gyrB and pgi). Dark blue: isolates from oligotrophic lakes ('Low Phosphorus Lake, Low Phosphorus Genotype LL/LG'); light blue: isolates from phosphorus-rich lakes, but related to oligotrophic isolates ('High Phosphorus Lake/Low Phosphorus Genotype, HL/LG''); green: isolates from phosphorus-rich lakes ('High Phosphorus Lake/High Phosphorus Genotype, HL/HG'). All significant trends, as determined using linear mixed effects models that control for collection date and lake of origin, are noted with one asterisk at the $p < 0.10$ level and two asterisks at the $p < 0.05$ level. Values for each metric are shown for each phylogenetic group ($\mu \pm 1$ S.E.). See Fig. S12 for values for each strain. Except for genome size, which is shown in megabases, all metrics are percentage data. Note that genome size, completeness, and GC content consider all contigs, regardless of length, while coding DNA, paralogs, and sigma factors as a

percentage of total genes considers only contigs 2 kb in length and longer. Significance of post-hoc pairwise comparisons are noted with lettering, where groups sharing the same letter do not significantly differ from each other. Nineteen of the 20 publicly available sequences collected worldwide were most closely related to the HL/HG group (Fig. S1).

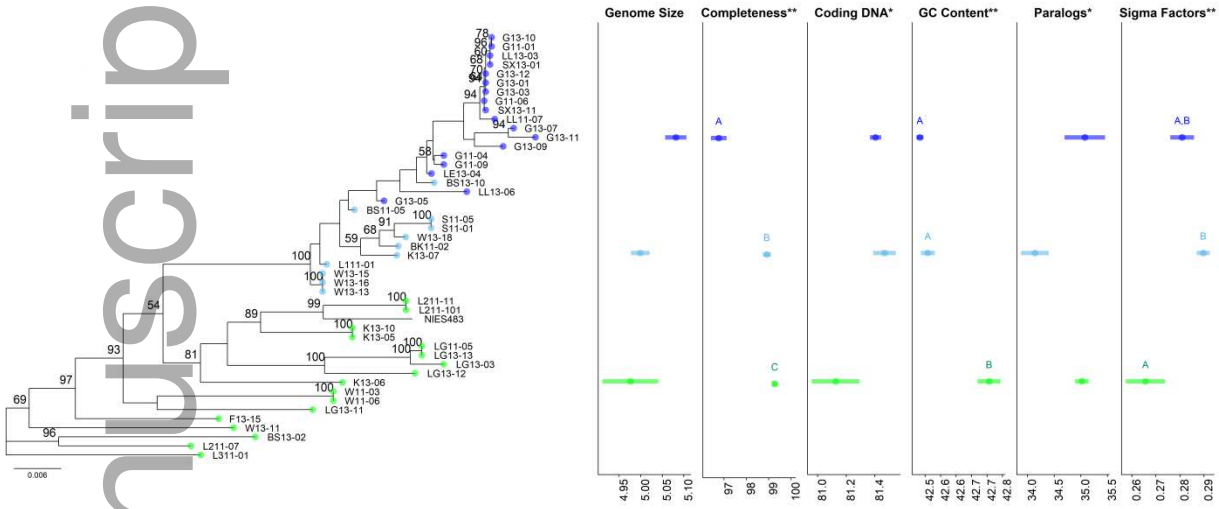

FIGURE 3 Functional analysis of *M. aeruginosa* genomes. (A) Isolates cluster based on genome-wide protein function according to lake of origin and corresponding lake trophic status. Note that all LL/LG isolates originated from oligotrophic lakes that we defined as TP < 10 µg/L, while HL/HG and HL/LG isolates, which frequently co-occurred, originated from mesotrophic and eutrophic lakes that we defined as TP ≥ 10 µg/L. (B) Isolates also cluster in genome-wide protein functions based on the phylogenetic groups shown in Fig. 2. All genes within each genome were assigned to protein families, with points shown closer in principal coordinate space sharing more similarity in protein family composition. Significance of separation was determined using analysis of variance on distance matrices, i.e. adonis.

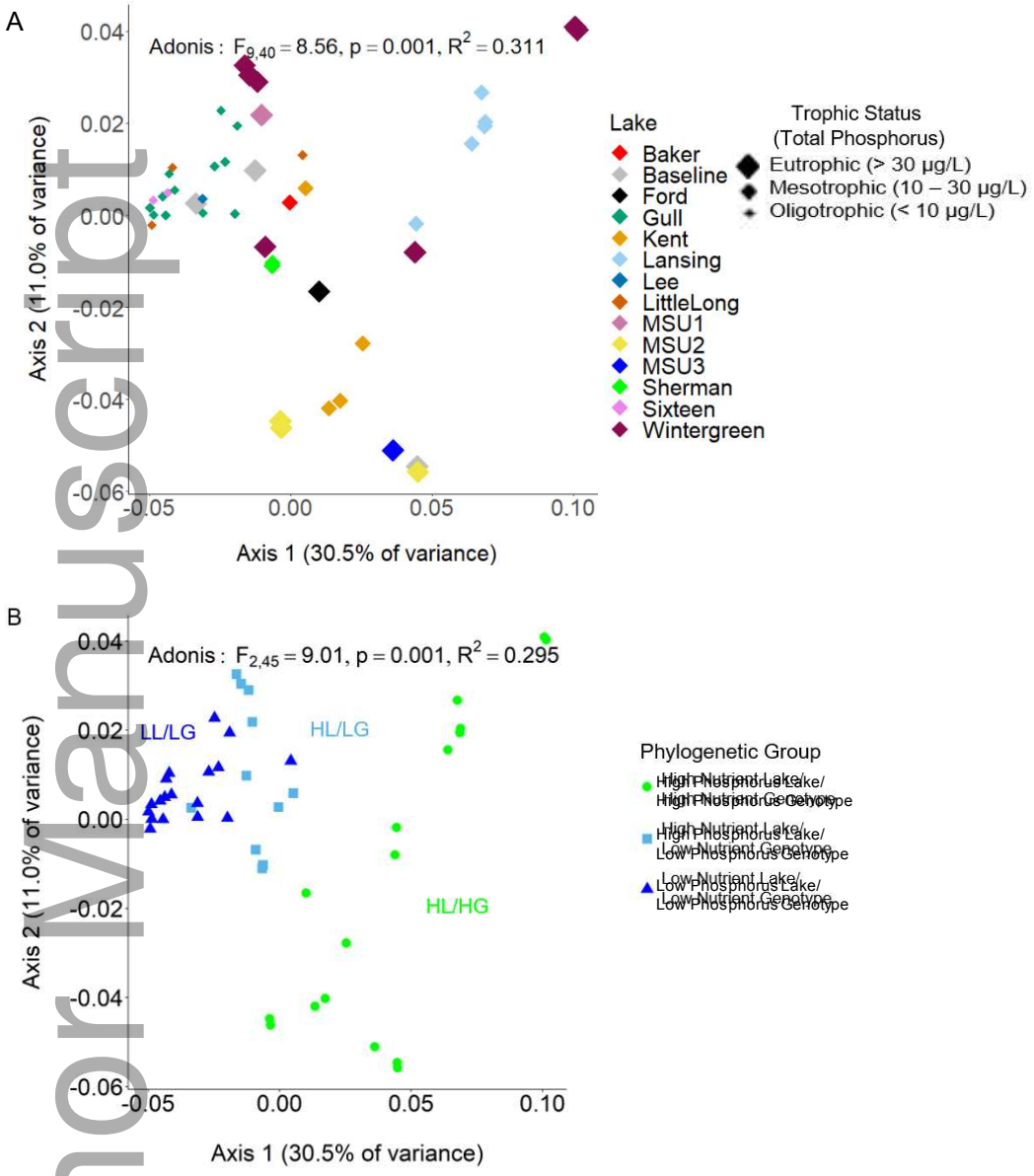

FIGURE 4 Taxonomic and functional analysis of *Microcystis aeruginosa*-associated microbiomes. (A) Isolates of *M. aeruginosa* belonging to different phylogenetic groups harbored taxonomically different communities of phycosphere bacteria. (B) Despite taxonomic differences, phycosphere communities associated with each of the three phylogenetic groups of *M. aeruginosa* tended to have similar protein functions. (C) When controlling for the local pool of phycosphere bacteria, host genomes that were more functionally similar to each other tended to harbor more functionally similar phycosphere communities. The x- and y- axes list functional distance in terms of Bray-Curtis distances from principal coordinate analyses that used the number of genes annotated to different protein families. This analysis used a subset

of samples that were collected on the same day and from the same lake (5 isolates from eutrophic Wintergreen Lake, 8 isolates from oligotrophic Gull Lake, respectively). Note that we show both the best fitting linear model and polynomial model, as results were similar. Also note that taxonomic data using a 16S rRNA gene survey of bacterial taxa is shown in (A), while functional data shown in (B) and (C) uses metagenome data in which all genes have been annotated to protein families to confer gene function.

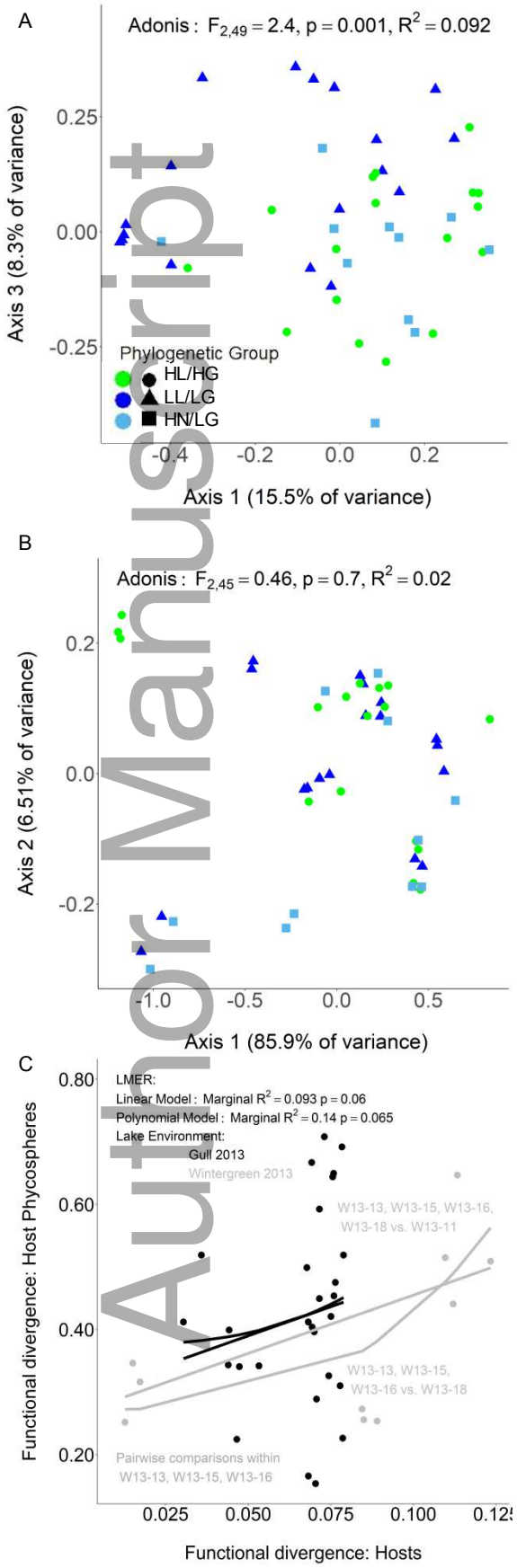

FIGURE 5 Growth rate measurement of *M. aeruginosa* isolates. Isolates of *Microcystis aeruginosa* belonging to the Low Phosphorus Lake/Low Phosphorus Genotype (LL/LG) phylogenetic group grew slower than isolates belonging to the High Phosphorus Lake/High Phosphorus Genotype (HL/HG) phylogenetic group. Isolates from the High Phosphorus Lake/Low Phosphorus Genotype (HL/LG) group, which originated from phosphorus-rich lakes but had a genomic architecture more closely resembling isolates in oligotrophic lakes, grew at intermediate rates. A linear mixed-effects model controls for Lake as a random effect to account for multiple isolates originating from a single lake. Group means illustrated with a dashed line.