LIMNOLOGY
and
OCEANOGRAPHY: METHODS

Limnol. Oceanogr.: Methods 19, 2021, 846–854
© 2021 The Authors. *Limnology and Oceanography: Methods* published by Wiley Periodicals LLC on behalf of Association for the Sciences of Limnology and Oceanography.
doi: 10.1002/lom3.10465

A comparative study of metatranscriptomic assessment methods to characterize *Microcystis* blooms

Helena L. Pound, Eric R. Gann, Steven W. Wilhelm*
Department of Microbiology, University of Tennessee, Knoxville, Tennessee

Abstract

Harmful algal blooms are increasing in duration and severity globally, resulting in increased research interest. The use of genetic sequencing technologies has provided a wealth of opportunity to advance knowledge, but also poses a risk to that knowledge if handled incorrectly. The vast numbers of sequence processing tools and protocols provide a method to test nearly every hypothesis, but each method has inherent strengths and weaknesses. Here, we tested six methods to classify and quantify metatranscriptomic activity from a harmful algal bloom dominated by *Microcystis* spp. Three online tools were evaluated (Kaiju, MG-RAST, and GhostKOALA) in addition to three local tools that included a command line BLASTx approach, recruitment of reads to individual *Microcystis* genomes, and recruitment to a combined *Microcystis* composite genome generated from sequenced isolates with complete, closed genomes. Based on the analysis of each tool presented in this study, two recommendations are made that are dependent on the hypothesis to be tested. For researchers only interested in the function and physiology of *Microcystis* spp., read recruitments to the composite genome, referred to as "Frankenstein's *Microcystis*," provided high total estimates of transcript expression. However, for researchers interested in the entire bloom microbiome, the online GhostKOALA annotation tool, followed by subsequent read recruitments, provided functional and taxonomic characterization, in addition to transcript expression estimates. This study highlights the critical need for careful evaluation of methods before data analysis.

The ecological and economic ramifications of harmful algal blooms draw significant attention from researchers and the public. Toxic algae cause major challenges for recreational and commercial fisheries as well as municipal water supplies (Bullerjahn et al. 2016). In fact, blooms of the cyanobacteria genus *Microcystis* have caused problems for water treatment facilities around the world, leading to water shortages that included significant recent crises in both the United States and China (Qin et al. 2010; Steffen et al. 2017). As blooms continue to increase in duration and severity, there is an associated increase in research to understand how blooms function (Harke et al. 2016; Tang et al. 2018). For many years, researchers have investigated the environmental conditions that stimulate blooms and attempted to mimic them in the laboratory (Orr and Jones 1998; Kaebernick et al. 2000; Sandrini et al. 2014; Steffen et al. 2014). Studies have relied on water quality metrics such as pH, toxin estimates, nutrient concentrations, chlorophyll *a* estimates, and basic microscopy until recent advances in DNA/RNA sequencing (Krüger and Eloff 1978; Reynolds et al. 1981; Seitzinger 1991; Paerl et al. 2016). The democratization of new sequencing technologies has opened novel opportunities in harmful algal bloom research. The ability to explore the genetic potential or activity of a single organism or entire community ([meta]genomics or [meta]transcriptomics) within the natural background allows scientists to investigate both novel and historical ecology and physiology (Steffen et al. 2014; Hennon and Dyhrman 2020).

As sequencing technologies continue to advance, so does the volume of data they generate, and all at decreasing costs (Shakya et al. 2019). As an example, recent metatranscriptomic studies have readily reached 20 million reads per sample, as compared to the 1 million reads per sample only a few years before (Steffen et al. 2015; Stough et al. 2017). While sample collection, processing, and sequencing must be carefully planned and executed, a parallel challenge in sequencing lies in the analysis of the data generated. The opportunity for error in analysis can range from inaccurate estimations of gene

*Correspondence: wilhelm@utk.edu

Author Contribution Statement: H.L.P. and E.R.G. contributed equally.

Additional Supporting Information may be found in the online version of this article.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

expression to misidentifying taxonomy. Every researcher is faced with the challenge of deciding which informatics method(s) to use to test their hypotheses, while also balancing needs for efficiency and accuracy. For example, online analysis tools may be more approachable to computationally limited labs, as these web-based tools can handle large datasets using remote servers and rarely possess a steep learning curve to implement (Kanehisa et al. 2016; Keegan et al. 2016; Menzel et al. 2016). Meanwhile, in-house coding methods can incorporate personally curated databases and allow for multiple tools to be selected and combined into a highly specialized and project specific workflow (Moniruzzaman et al. 2017; Stough et al. 2017; Pound et al. 2020). Tools also vary in the type of input data required, as some require just the libraries of trimmed reads, while others require the libraries to be assembled into contigs. Completely analyzing sequencing data may involve some combination of these or other techniques. All have their strengths and weaknesses that must be carefully considered, as these idiosyncrasies can easily influence the data and conclusions made.

The use of molecular tools to study *Microcystis* bloom events has exploded in recent years as early work demonstrated how polymerase chain reaction (PCR)/quantitative PCR/sequencing can be used to distinguish toxic from nontoxic populations and do so in a quantitative manner (Kurmayer and Kutzenberger 2003; Rinta-Kanto et al. 2005). Water treatment facilities, government agencies tasked with monitoring, and private commercial groups have all engaged molecular biological approaches to study *Microcystis* in recent years (Chorus and Welker 2021). Currently, most (meta)transcriptomic-based *Microcystis* research characterizes activity in the environment by recruiting reads back to a single genome. Many studies use the reference genome *Microcystis aeruginosa* NIES-843 (Harke and Gobler 2013; Steffen et al. 2017; Tang et al. 2018) since it was the first to be sequenced, but other strains have also been used (Davenport et al. 2019; Morimoto et al. 2019). However, it has been demonstrated that there is a great diversity of genetic potential encoded by different *Microcystis* sp. strains, and in the rest of the microbiome, within a single bloom (Meyer et al. 2017; Cook et al. 2020; Pound and Wilhelm 2020*b*).

This study sought to compare multiple techniques and tools in the analyses of transcriptomic sequencing data generated from natural harmful algal bloom communities in fresh waters. Metatranscriptomic libraries were generated from a 2019 *Microcystis* spp.-dominated bloom in Lake Erie, USA. A variety of techniques were used to characterize the functional activity and taxonomic composition, using both trimmed reads and assembled contigs. Techniques were evaluated for three primary metrics. The first was the number of reads recruited, either to all *Microcystis* genes or to specific *Microcystis* marker genes important to central metabolism (resolved by phylogeny). It is assumed that methods capable of detecting and estimating the natural diversity in an environmental sample will recruit more reads than methods that are unable to resolve the fine scale genetic variation present in the environment. The second was the ability for a tool to classify sequence data by function and taxonomy, independent of any recruitments. Some tools classify the reads themselves, while others classify assembled contigs, which then require reads to be recruited back to them to generate relative quantitative results. The third was to determine how well each method correlated with one another, which provides support that the variation between samples is reflective of the ecology of this bloom and not the methods themselves. Based on our analyses, we make two recommendations that are dependent on the hypotheses being tested. The following study details how we came to those conclusions and discusses the various strengths and weaknesses of each approach examined.

Methods

Sampling and sequencing

Samples were collected from the surface of a 2019 *Microcystis* spp.-dominated bloom in Lake Erie, USA, on 21 July 2019, and incubated in 1.0 L acid-washed polycarbonate bottles for 48 h. Whole water was then passed through a 0.2 μm Sterivex filter (Millipore) and flash frozen. RNA was extracted using an acid-phenol-based extraction protocol with an added DNase treatment to remove any residual DNA (Pound and Wilhelm 2020*a*). RNA quality was checked using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific) and quantified using a Qubit RNA HS assay (Invitrogen). Extracted RNA was processed using a Illumina® Stranded Total RNA Prep, Ligation with Ribo-Zero Plus and then 50-million 100-bp paired-end reads were generated on the Illumina NovaSeq platform at Hudson Alpha Discovery Life Sciences (Huntsville, Alabama). Sequence processing and assembly was described in detail in online protocol (Pound and Wilhelm 2021). Briefly, sequences were quality controlled and trimmed in the CLC Genomics workbench version 20.0.4 (Qiagen). Residual ribosomal rRNA reads were removed in silico using SortMeRNA version 4 (Kopylova et al. 2012). The nonribosomal reads classified by SortMeRNA from all samples were jointly assembled in MegaHit version 1.2.9 (Li et al. 2015). This was done to reduce redundancy among identical sequences in various samples as done previously (Pound et al. 2020). Trimmed nonribosomal reads have been uploaded to and are available on MG-RAST (Keegan et al. 2016) under project name "LE2019MT" and raw reads are available on the NCBI SRA database under BioProject number PRJNA737197.

Background on approaches

We explored three widely used online platforms for sequencing analysis: Kaiju (Menzel et al. 2016), MG-RAST (Keegan et al. 2016), and the Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology and Links Annotation (GhostKOALA) (Kanehisa et al. 2016). GhostKOALA and MG-RAST both provide functional and taxonomic classifications, while Kaiju provides only a taxonomic classification (Table 1). Kaiju and MG-RAST accept trimmed sequencing reads as input, while GhostKOALA requires coding sequences to be identified from assembled contigs. We also explored in-house approaches, including variations of a customizable BLASTx database approach (Pound et al. 2020; Pound and Wilhelm 2020*b*), read recruitments to coding sequences of individual *Microcystis* genomes (Harke and Gobler 2013; Davenport et al. 2019; Krausfeldt et al. 2019*b*), and recruitments to a *Microcystis* composite genome created for this analysis. The latter is referred to as "Frankenstein's *Microcystis*" in reference to Mary Shelley's fictional monster comprised of parts from many individuals (Shelley 1818). The BLASTx approach taxonomically classifies a gene of interest within assembled contigs, while the recruitments to the individual and composite genomes use trimmed reads to functionally classify reads against the coding sequences of a single organism (Table 1).

BLASTx approach

Identification and transcript quantification for genes of interest was performed using a BLASTx method previously described (Pound and Wilhelm 2019, 2020*b*). In short, protein databases were established for the following genes of interest using the KEGG orthology K number: DNA-directed RNA polymerase, beta subunit (*rpo*B, K03043), ribulose bisphosphate carboxylase large chain (*rbc*L, K01601), ribose-phosphate pyrophosphokinase 1 (*prp*S, K00948), bifunctional purine biosynthesis protein (*pur*H, K00602), and orotidine 5′-phosphate decarboxylase (*pyr*F, K01591). All reference sequences were downloaded from UniProt KB database version 2020_03 (*see* Data Availability). The assembled contigs from all samples was then queried using the command-line BLASTx against each database specific to a gene of interest, retaining sequences with a minimum contig length of 300 bp and an e-value $< 1 \times 10^{-30}$ (Camacho et al. 2009). Those with BLASTx hits were considered "candidates" and only the aligned portion of the gene sequence was used for recruitments (Pound and Wilhelm 2019). Trimmed reads from each sample were recruited to each trimmed candidate contig in CLC Genomics workbench version 20.0.4 (Qiagen) using a 90% length fraction and 90% identity (Pound and Wilhelm 2020*b*). Any reads that could be recruited to more than one contig with equal identity was randomly assigned, to mimic MG-RAST and Kaiju data handling approaches and to ensure all reads of interest were quantified. The reference trees were established with maximum likelihood phylogenies based on reference proteins from isolated organisms using PhyML version 3.0 (Guindon et al. 2010). To determine taxonomy, candidate sequences were placed on reference phylogenetic trees using the pplacer algorithm (Matsen et al. 2010).

Kaiju

Trimmed reads were uploaded to the online Kaiju platform and default parameters were used to taxonomically classify reads (Menzel et al. 2016). Reads were classified as *Microcystis* spp. by summing all the reads that were classified as the *Microcystis* genus.

MG-RAST

Trimmed reads were uploaded to the online MG-RAST platform (Keegan et al. 2016). All parameters were default except for no dereplication and the percent identity threshold (increased to > 90%). Reads were classified as *Microcystis* spp. by summing all the reads classified as the *Microcystis* genus based on RefSeq annotations. The number of reads classified as five genes important in central metabolism (*rpo*B, *rbc*L, *prp*S, *pur*H, and *pyr*F) annotated by the KEGG orthology database within MG-RAST were also collected.

GhostKOALA

The nucleotide and translated amino acid sequences of coding sequences were predicted within the assembled contigs

Table 1. Comparison of methods analyzed in this study. Y indicates a requirement or function provided, while N indicates a lack of requirement or output provided.

	BLASTx	Kaiju	MG-RAST	Ghost KOALA	Individual genomes	Frankenstein genome
Community taxonomy	Y	Y	Y	Y	N	N
Gene function	Y	N	Y	Y	Y	Y
User-defined database	Y	N	N	N	Y	Y
Input	Contigs	Reads	Reads	Contigs	Reads	Reads
Additional read recruitment	Y	N	N	Y	N	N
Online platform	N	Y	Y	Y	N	N
Example references	Moniruzzaman et al. (2017), Pound et al. (2020)	Chen et al. (2019)	Zhang et al. (2016), Krausfeldt et al. (2019*a*)	Xie et al. (2016), Li et al. (2018), Cook et al. (2020)	Harke and Gobler (2015), Steffen et al. (2017)	This study

using MetaGeneMark version 3.25 (Besemer and Borodovsky 1999; Zhu et al. 2010). Amino acid sequences were functionally and taxonomically classified via KEGG orthology using the prokaryote + eukaryote + virus database using GhostKOALA (Kanehisa et al. 2016; Pound et al. 2021*b*). Predicted coding sequences were assigned KEGG orthology numbers (K numbers) and trimmed reads were recruited to the coding sequences using a 90% similarity fraction over a 90% length fraction in CLC Genomic Workbench. Reads that could be recruited to more than one contig with equal identity were randomly assigned. The number of reads recruited to the five genes important in central metabolism mentioned above was also pulled from the read recruitments.

Recruitment to individual genomes

Coding sequences from complete, closed *Microcystis* spp. genomic assemblies were used in this study. The coding sequences from the following genomes were used: *M. aeruginosa* FD4 (accession: NZ_CP046973.1), *Microcystis* sp. MC19 (accession: NZ_CP020664.1), *Microcystis viridis* NIES-102 (accession: NZ_AP019314.1), *M. aeruginosa* NIES-298 (accession: NZ_CP046058.1), *M. aeruginosa* NIES-843 (accession: NC_010296.1), *M. aeruginosa* NIES-2481 (accession: NZ_CP012375.1), *M. aeruginosa* NIES-2549 (accession: NZ_CP011304.1), *Microcystis panniformis* FACHB-1757 (accession: CP011339.1), and *M. aeruginosa* PCC7806SL (accession: NZ_CP020771.1). Reads were recruited using a 90% similarity fraction over a 90% length fraction to the coding sequences. Any reads that could be recruited to more than one coding sequence with equal identity were randomly assigned. The number of reads recruited to the five markers of central metabolism mentioned above was also pulled from the read recruitments to the individual genomes.

Frankenstein's *Microcystis*: Construction and recruitment

Coding sequences from all complete, closed *Microcystis* genomes (see above) were combined into a single FASTA file and clustered at different nucleotide identities using CD-HIT-EST (Fu et al. 2012; Pound et al. 2021*a*). Coding sequences from the individual genomes were clustered at the nucleotide identities of 0.95, 0.90, 0.85, and 0.80 with the word size set to 10, 8, 6, and 5, respectively, to establish composite genomes of various stringencies. Trimmed reads were recruited using a 90% similarity fraction over a 90% length fraction to each composite genome in CLC Genomics Workbench. Any reads that could be recruited to more than one coding sequence with equal identity were randomly assigned. After comparing the various stringencies (Supplemental Fig. S1), the 0.95 identity cluster showed the greatest number of coding sequences, while reducing redundancy, and was used for all subsequent analyses. This synthetic library is referred to as "Frankenstein's *Microcystis*" and is available on GitHub (*see* Data Availability). The number of reads recruited to five genes important in central metabolism mentioned above was also pulled from the read recruitments to the individual genomes.

Methods correlation

The total number of reads recruited to, or classified as, *Microcystis* spp. per sample, and those recruited to, or classified as, specific *Microcystis* spp. genes important in central metabolism per sample were normalized by the total number of trimmed reads in each sample library. Pearson correlation coefficients were established with Benjamini-Hochberg corrections for multiple comparisons (Benjamini and Hochberg 1995). All statistical analyses were carried out in R studio. The Pearson's *r* coefficients and corrected *p*-values are reported in Supplemental File S1.

Assessment

We used a metatranscriptomic dataset generated from a *Microcystis* spp.-dominated harmful algal bloom in Lake Erie in August 2019 to compare the methods/tools described above. RNA was extracted from samples collected from bottle incubations (where nutrients were being manipulated) as well as in situ samples and then sequenced. Environmental variables were intentionally disregarded, in order to evaluate method performance independent of abiotic variables. A total of ~ 923 million processed reads were generated across twenty sample metatranscriptomes (~ 46 million reads per library, *see* Supplemental File S1). These reads assembled into 2,335,243 contigs that varied in length from 157 to 66,630 nucleotides. Total *Microcystis* reads are reported per sample, while reads recruited to or classified as individual genes reads are reported as the average across all samples.

Total reads recruited to *Microcystis* spp.

One of the primary metrics used to rate the performance of the methods tested was the number of total reads recruited to, or classified, as *Microcystis* spp. All but one of our methods (the BLASTx approach) provided an estimate of total *Microcystis* spp. reads, but not all methods performed equally (Fig. 1). The Kaiju method recruited the fewest reads, as total reads classified as *Microcystis* spp. ranged from 4.08×10^6 to 1.18×10^7, which was between 11.5% and 22.6% of the total reads per sample library (Supplemental File S1). Total *Microcystis* spp. read estimates from MG-RAST (13.8% and 28.9% of library), individual *Microcystis* spp. genome recruitment (17.0% and 37.7% of library), and Frankenstein's *Microcystis* genome recruitment (21.4–43.8% of library) surpassed Kaiju's estimates but did not provide the largest estimates. GhostKOALA recruited the most reads, as total reads recruited to *Microcystis* spp. annotated genes ranged from 8.38×10^6 to 2.36×10^7, which was between 24.2% and 45.2% of the total available reads per sample library.

Reads recruited to *Microcystis*-specific marker genes

It was also important that the methods tested be evaluated for the number of reads recruited to, or classified, as individual genes, as this is important for future studies concerned with

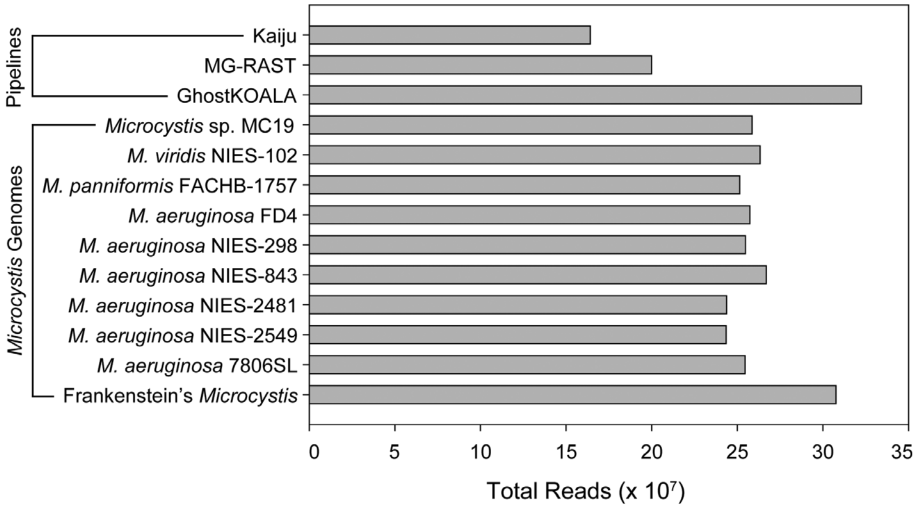

Fig. 1. The sum of all reads in all samples that were recruited to or classified as *Microcystis* in each method.

the dynamics of individual genes, as opposed to total shifts in the genome. All but one of our methods (the Kaiju approach) provided read estimates of individual *Microcystis*-specific genes, and most methods performed relatively equally (Supplemental Fig. S2). MG-RAST had the fewest reads classified to any of the genes we tested. On average, 1.77 $\times$ 10^4 reads per sample were classified as *rpo*B, 2.39 $\times$ 10^4 reads per sample were classified as *rbc*L, 6.57 $\times$ 10^3 reads per sample were classified as *prp*S, 4.30 $\times$ 10^3 reads per sample were classified as *pur*H, and 3.72 $\times$ 10^2 reads per sample were classified as *pyr*F (Supplemental File S1). The other methods evaluated were remarkably similar to each other, including the Ghost KOALA annotated gene recruitment, individual *Microcystis* spp. genome recruitment, and Frankenstein's *Microcystis* genome recruitment, although the BLASTx method outperformed the others in four of the five genes. On average, the BLASTx method recruited 2.86 $\times$ 10^4 reads per sample to *rpo*B, 3.64 $\times$ 10^4 reads per sample to *rbc*L, 8.95 $\times$ 10^3 reads per sample to *prp*S, 4.20 $\times$ 10^3 reads per sample to *pur*H, and 6.22 $\times$ 10^2 reads per sample to *pyr*F (Supplemental File S1).

Methods evaluation

The BLASTx method involves the use of in-house curated databases that each contains a single marker gene from many species; it thus does not allow for a summary of all pathways as the other methods do. However, this approach provides the user full control over the database used and allows for greater confidence in taxonomic predictions based on the subsequent placement of predicted coding sequences on phylogenetic trees (Matsen et al. 2010). These curated databases can be updated by the user at any point and can be accessed locally. This approach characterizes the assembled contigs, which allows for a more accurate estimate of true community diversity, as opposed to recruiting genes to a reference genome. The number of reads recruited for each gene was > 99% correlated to the number of reads recruited by any of the other methods (Supplemental File S1). The primary weakness of this method is the lack of efficiency. This method is restricted to individual gene investigations, which may work well for specific questions, or small genomes (e.g., RNA viruses; Moniruzzaman et al. 2017) but is impractical for the complete characterization of the approximately 3600 genes in a *Microcystis* genome.

While the Kaiju method was the easiest to execute, by our measures it was one of the poorest performers. The program did not provide an efficient way to summarize functional data, as each read was individually characterized to a specific organism, few of which were annotated in the same manner. Therefore, this tool was primarily used to classify taxonomic data, by summing all reads to the *Microcystis* genus. Kaiju does excel in characterizing the entire community, as opposed to recruiting reads to a single genome. However, it classified fewer reads to *Microcystis* than all other methods. It is unclear if the "missing" reads were not annotated at all or mistakenly annotated as a different organism. It is worth noting that this tool and its databases have not been updated since 2017. This method was the only method tested that was not as strongly correlated to the other methods, with correlation coefficients ranging from 0.788 to 0.831, depending on the method compared, but none of the differences were considered significantly different (Supplemental File S1). This suggests that Kaiju may not be consistent in the way it analyzes samples, reducing how well the read estimates correlate to other methods, regardless of the magnitude of reads recruited.

MG-RAST was like Kaiju in that it provided an easy online interface that accepted trimmed reads and functionally characterized the entire community, not just *Microcystis* spp. We note that MG-RAST outperformed Kaiju, but none of the other methods tested in this study. This method did provide the opportunity to characterize individual genes, including our five marker genes. The number of reads recruited to each gene and total *Microcystis* spp. was > 98% correlated to the number of reads recruited by any of the other methods (Supplemental File S1). The decrease in total, and gene specific, reads recruited is suspected to arise from many probable *Microcystis* spp. reads being incorrectly annotated as *Cyanothece* spp., a species that was not regularly detected in our other methods that were capable of characterizing the entire community, such as the Kaiju or BLASTx methods.

The GhostKOALA method provided both community taxonomy and gene functional characterization, all from an online database. Therefore, this method can easily characterize all active species present in any environmental system, not just a *Microcystis* spp. in a *Microcystis*-dominated bloom (Pound et al. 2021*c*). The annotation is based on assembled data and therefore required the recruitment of trimmed reads to genes of interest to make the results quantitative. In our hands this method identified more total reads as *Microcystis* spp. than any other method. The number of reads recruited for each gene and total *Microcystis* spp. was > 98% correlated to the number of reads recruited by any of the other methods (Supplemental File S1). The only notable disadvantage to GhostKOALA methodology is that annotations are limited to KEGG orthologies, so some transcripts within samples may not be annotated.

To assess whether there was a difference in the number of reads that recruited to an individual *Microcystis* genome, we recruited reads to coding sequences from all nine closed, complete genomes in NCBI including six *M. aeruginosa* strains, one *M. panniformis* strain, one *M. viridis* strain, and one *Microcystis* sp. strain (Supplemental File S1). This method does not rely on assembled contigs, and all genomes were easily downloaded from the NCBI database, which is updated regularly to include new genomes and annotations. Each of the five specific genes analyzed in this study showed the highest recruitments in a different genome strain, although the variation between each strain was minimal (1–2%) (Supplemental Fig. S2). The number of reads recruited for each gene and all coding sequences in *Microcystis* spp. was > 98% correlated to the number of reads recruited by any of the other methods (Supplemental File S1). The primary disadvantage of this method is the lack of taxonomic or functional data on the rest of the microbial community within a sample.

While the Frankenstein's genome method is primarily a read recruitment method, it has an additional step as a composite genome must first be generated. However, the approach allows the user to customize and update "Frankenstein's *Microcystis*" as other genomes are completed and published. There was a total of 44,950 coding sequences between the nine strains (average = 4994). Clustering all the coding sequences at decreasing nucleotide identity reduced the total number of clusters from 13,600 to 8920, when clustered at a nucleotide identity of 0.95 or 0.80, respectively. When all coding sequences were clustered, regardless of nucleotide identity, 20.7–43.4% of the total library reads recruited (Supplemental Fig. S1). The 95% identity composite genome performed well in both total reads recruited and reads recruited to the individual genes. The number of reads recruited for each gene and total *Microcystis* was > 98% correlated to the number of reads recruited by any of the other methods (Supplemental File S1). As with the individual genomes though, this method only allows for the characterization of the *Microcystis* community, not the rest of the microbiome.

Discussion

While the preparation of (meta)transcriptomic samples before sequencing can shape the overall outcome of an analysis for microbial communities (Gann et al. 2021), our analyses have indicated that the way sequences are evaluated can have equally large impacts on the conclusions reached. As the volume of sequencing data generated grows and more researchers turn to molecular tools to address environmental questions, researchers may be tempted to use publicly available, automated pipelines or easily downloaded single-strain genomes. Each of these methods has strengths and weaknesses, and it falls upon the researcher to establish best practices. This study however provides guidelines to the growing community of *Microcystis* spp. researchers but can also serve as advisory to those interested in other organisms or communities.

One of the primary concerns for any methodology for metatranscriptomic sequencing analysis is the ability to characterize the true diversity present efficiently and accurately in a sample, regardless of whether it is from a lab culture or an environmental sample. For many years, the common practice has been to quantify activity by recruited reads back to a single reference genome (Harke and Gobler 2013; Steffen et al. 2017; Davenport et al. 2019). Here, we have tested a composite of available complete, closed genomes, Frankenstein's *Microcystis*, where coding sequences from different isolates of the same genus were clustered together to reduce redundancy: this provides a database containing both the common core genes associated within the *Microcystis* genus as well as the unique coding potentials associated with different isolates. We note that this approach can be extended to any microorganism of interest if multiple, well-curated genomes are available for clustering. However, it is important to note that even a composite genome is limited to the diversity of sequenced, well-annotated isolates of microorganisms, and still may not represent a natural community's true diversity. Moving forward, researchers will need to update Frankenstein's *Microcystis* with newly sequenced isolates as they become available. While the BLASTx method we adopted from Pound et al. (2020) and Moniruzzaman et al. (2017) was also capable of more fully characterizing community diversity, the restriction to individual gene databases makes it cumbersome and less ideal for broader hypotheses.

A key detail to consider *a priori* with sequencing data is the resolution to which a researcher may want to characterize the community. As mentioned above, recruitments to Frankenstein's *Microcystis* provide an efficient and comprehensive way to analyze the *Microcystis* spp. community. However, it is well known that the rest of the microbiome is likely important to how harmful algal blooms function (Cook et al. 2020). While the BLASTx method mentioned above can characterize the entire microbial community, marker-gene database inefficiencies are still present. Many tools such as Kaiju, or even 16S rRNA gene sequencing, can provide information on what organisms are present, but few tools also provide information on the functional genes present.

Many other factors should also be considered when choosing a method, including the frequency the tool is updated. Even though GhostKOALA performed well in our analysis, the KEGG ontology database will require regular updates to stay relevant. The only way to have full control over a database is to curate it manually, although we recognize that this can be inefficient and can lead to biases from individual laboratory groups based on annotations. Computational power and coding expertise should also be considered, as some of methods discussed are extremely user friendly while other require some knowledge of command line coding. Many of the online tools use remote servers, while many of the command line tools would likely require local computational power.

Recommendations

For researchers sequencing *Microcystis* spp.-dominated harmful algal blooms, we make two suggestions based on our analyses and the resolution of a proposed study. For those wishing to focus solely on *Microcystis* spp. function and activity, we would recommend using a regularly updated Frankenstein's *Microcystis* composite genome as it does not require the transcriptomes being assembled into contigs and can provide detailed data on every potential gene currently sequenced in *Microcystis* genus. It is also important to note that the combined genome approach used to establish Frankenstein's *Microcystis* could easily be applied to other organisms and study systems. However, for researchers wishing to focus on microbiome ecology and interactions between species, we recommend using the GhostKOALA approach, which provides both functional and taxonomic characterization of the entire community, although it does require additional read recruitments to estimate transcriptional activity. Regardless of the study system, we highly stress the critical importance of taking great care in selecting an appropriate method when processing sequence data.

DATA AVAILABILITY STATEMENT

Trimmed nonribosomal reads have been uploaded to and are available on MG-RAST (Keegan et al. 2016) under project name "LE2019MT," and raw reads are on NCBI SRA database under BioProject number PRJNA737197. Reference protein sequences used to characterize individual genes and the Frankenstein's Microcystis genome are publicly available as FASTA files at https://github.com/Wilhelmlab/PoundGannWilhelm2021.

References

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. B. Methodol. **57**: 289–300.

Besemer, J., and M. Borodovsky. 1999. Heuristic approach to deriving models for gene finding. Nucleic Acids Res. **27**: 3911–3920.

Bullerjahn, G. S., and others. 2016. Global solutions to regional problems: Collecting global expertise to address the problem of harmful cyanobacterial blooms. A Lake Erie case study. Harmful Algae **54**: 223–238. doi:10.1016/j.hal.2016.01.003

Camacho, C., and others. 2009. BLAST+: Architecture and applications. BMC Bioinformatics **10**: 1–9. doi:10.1186/1471-2105-10-421

Chen, H., L. Jing, Z. Yao, F. Meng, and Y. Teng. 2019. Prevalence, source and risk of antibiotic resistance genes in the sediments of Lake Tai (China) deciphered by metagenomic assembly: A comparison with other global lakes. Environ. Int. **127**: 267–275. doi:10.1016/j.envint.2019.03.048

Chorus, I., and M. Welker. 2021. Toxic cyanobacteria in water: A guide to their public health consequences, monitoring and management. Taylor & Francis.

Cook, K. V., and others. 2020. The global *Microcystis* interactome. Limnol. Oceanogr. **65**: S194–S207. doi:10.1002/lno.11361

Davenport, E. J., and others. 2019. Metatranscriptomic analyses of diel metabolic functions during a *Microcystis* bloom in Western Lake Erie (United States). Front. Microbiol. **10**: 2081. doi:10.3389/fmicb.2019.02081

Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics **28**: 3150–3152. doi:10.1093/bioinformatics/bts565

Gann, E. R., Y. Kang, S. Dyhrman, C. J. Gobler, and S. W. Wilhelm. 2021. Metatranscriptome library preparation influences analyses of viral community activity during a brown tide bloom. Front. Microbiol. **12**: 1126. doi:10.3389/fmicb.2021.664189

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst. Biol. **59**: 307–321. doi:10.1093/sysbio/syq010

Harke, M. J., and C. J. Gobler. 2013. Global transcriptional responses of the toxic cyanobacterium, *Microcystis aeruginosa*, to nitrogen stress, phosphorus stress, and growth on organic matter. PLoS One **8**: e69834. doi:10.1371/journal.pone.0069834

Harke, M. J., and C. J. Gobler. 2015. Daily transcriptome changes reveal the role of nitrogen in controlling microcystin synthesis and nutrient transport in the toxic cyanobacterium, *Microcystis aeruginosa*. BMC Genomics **16**: 1068. doi:10.1186/s12864-015-2275-9

Harke, M. J., and others. 2016. A review of the global ecology, genomics, and biogeography of the toxic cyanobacterium, *Microcystis* spp. Harmful Algae **54**: 4–20. doi:10.1016/j.hal.2015.12.007

Hennon, G. M., and S. T. Dyhrman. 2020. Progress and promise of omics for predicting the impacts of climate change on harmful algal blooms. Harmful Algae **91**: 101587. doi:10.1016/j.hal.2019.03.005

Kaebernick, M., B. A. Neilan, T. Börner, and E. Dittmann. 2000. Light and the transcriptional response of the microcystin biosynthesis gene cluster. Appl. Environ. Microbiol. **66**: 3387–3392. doi:10.1128/aem.66.8.3387-3392.2000

Kanehisa, M., Y. Sato, and K. Morishima. 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J. Mol. Biol. **428**: 726–731. doi:10.1016/j.jmb.2015.11.006

Keegan, K. P., E. M. Glass, and F. Meyer. 2016. MG-RAST, a metagenomics service for analysis of microbial community structure and function, p. 207–233. *In* Microbial Environmental Genomics (MEG). Springer.

Kopylova, E., L. Noé, and H. Touzet. 2012. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics **28**: 3211–3217. doi:10.1093/bioinformatics/bts611

Krausfeldt, L. E., A. T. Farmer, H. Castro Gonzalez, B. N. Zepernick, S. R. Campagna, and S. W. Wilhelm. 2019a. Urea is both a carbon and nitrogen source for *Microcystis aeruginosa*: Tracking 13C incorporation at bloom pH conditions. Front. Microbiol. **10**: 1064. doi:10.3389/fmicb.2019.01064

Krausfeldt, L. E., M. M. Steffen, R. M. McKay, G. S. Bullerjahn, G. L. Boyer, and S. W. Wilhelm. 2019b. Insight into the molecular mechanisms for microcystin biodegradation in Lake Erie and Lake Taihu. Front. Microbiol. **10**: 2741. doi:10.3389/fmicb.2019.02741

Krüger, G., and J. Eloff. 1978. The effect of temperature on specific growth rate and activation energy of *Microcystis* and *Synechococcus* isolates relevant to the onset of natural blooms. J. Limnol. Soc. S. Afr. **4**: 9–20.

Kurmayer, R., and T. Kutzenberger. 2003. Application of real-time PCR for quantification of microcystin genotypes in a population of the toxic cyanobacterium *Microcystis* sp. Appl. Environ. Microbiol. **69**: 6723–6730. doi:10.3389/fmicb.2019.02741

Li, D., C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics **31**: 1674–1676.

Li, Q., and others. 2018. A large-scale comparative metagenomic study reveals the functional interactions in six bloom-forming *Microcystis*-epibiont communities. Front. Microbiol. **9**: 746. doi:10.3389/fmicb.2018.00746

Matsen, F. A., R. B. Kodner, and E. V. Armbrust. 2010. pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics **11**: 538. doi:10.1186/1471-2105-11-538

Menzel, P., K. L. Ng, and A. Krogh. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun. **7**: 1–9.

Meyer, K. A., T. W. Davis, S. B. Watson, V. J. Denef, M. A. Berry, and G. J. Dick. 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. doi:10.1371/journal.pone.0183859

Moniruzzaman, M., L. L. Wurch, H. Alexander, S. T. Dyhrman, C. J. Gobler, and S. W. Wilhelm. 2017. Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. Nat. Commun. **8**: 16054. doi:10.1038/ncomms16054

Morimoto, D., and others. 2019. Cooccurrence of broad-and narrow-host-range viruses infecting the bloom-forming toxic cyanobacterium *Microcystis aeruginosa*. Appl. Environ. Microbiol. **85**: e01170-01119. doi:10.1128/AEM.01170-19

Orr, P. T., and G. J. Jones. 1998. Relationship between microcystin production and cell division rates in nitrogen-limited *Microcystis aeruginosa* cultures. Limnol. Oceanogr. **43**: 1604–1614. doi:10.4319/lo.1998.43.7.1604

Paerl, H. W., and others. 2016. It takes two to tango: When and where dual nutrient (N&P) reductions are needed to protect lakes and downstream ecosystems. Environ. Sci. Technol. **50**: 10805–10813. doi:10.1021/acs.est.6b02575

Pound, H., E. R. Gann, and S. W. Wilhelm. 2021a. Creating a Frankenstein's genome. protocols.io. doi:10.17504/protocols.io.bv2zn8f6

Pound, H., E. R. Gann, and S. W. Wilhelm. 2021b. Functional and taxonomic characterization of sequence data using GhostKOALA. protocols.io. doi:10.17504/protocols.io.buvbnw2n

Pound, H., and S. W. Wilhelm. 2020a. RNA extraction from Sterivex using phenol:chloroform. protocols.io. doi:10.17504/protocols.io.bhu6j6ze

Pound, H., and S. W. Wilhelm. 2021. Sequence processing and assembly workflow using CLC Workbench, SortMeRNA, and MegaHit. protocols.io. doi:10.17504/protocols.io.buvdnw26

Pound, H. L., and S. W. Wilhelm. 2019. Metatranscriptomic screening for genes of interest. protocols.io. doi:10.17504/protocols.io.7vyhn7w

Pound, H. L., and S. W. Wilhelm. 2020. Tracing the active genetic diversity of Microcystis and Microcystis phage through a temporal survey of Taihu. PLOS ONE. **15**: e0244482. doi:10.1371/journal.pone.0244482

Pound, H. L., and others. 2020. The "Neglected Viruses" of Taihu: Abundant transcripts for viruses infecting eukaryotes and their potential role in phytoplankton succession. Front. Microbiol. **11**: 338. doi:10.3389/fmicb.2020.00338

Pound, H. L., and others. 2021c. Environmental studies of cyanobacterial harmful algal blooms should include interactions with the dynamic microbiome. Environ. Sci. Technol. **5**: 12776–12779. doi:10.1021/acs.est.1c04207

Qin, B., and others. 2010. A drinking water crisis in Lake Taihu, China: Linkage to climatic variability and lake management. Environ. Manag. **45**: 105–112. doi:10.1007/s00267-009-9393-6

Reynolds, C. S., G. Jaworski, H. Cmiech, and G. Leedale. 1981. On the annual cycle of the blue-green alga *Microcystis aeruginosa* Kütz. emend. Elenkin. Philos. Trans. R. Soc. Lond. B Biol. Sci. **293**: 419–477.

Rinta-Kanto, J. M., A. J. A. Ouellette, G. L. Boyer, M. R. Twiss, T. B. Bridgeman, and S. W. Wilhelm. 2005. Quantification of toxic *Microcystis* spp. during the 2003 and 2004 blooms in western Lake Erie using quantitative real-time PCR. Environ. Sci. Technol. **39**: 4198–4205. doi:10.1021/es048249u

Sandrini, G., H. C. Matthijs, J. M. Verspagen, G. Muyzer, and J. Huisman. 2014. Genetic diversity of inorganic carbon uptake systems causes variation in CO_2 response of the cyanobacterium *Microcystis*. ISME J. **8**: 589–600. doi:10.1038/ismej.2013.179

Seitzinger S. P. 1991. The effect of pH on the release of phosphorus from Potomac estuary sediments: Implications for blue-green algal blooms. Estuarine, Coastal and Shelf Science. **33**: 409–418. doi:10.1016/0272-7714(91)90065-j

Shakya, M., C.-C. Lo, and P. S. Chain. 2019. Advances and challenges in metatranscriptomic analysis. Front. Genet. **10**: 904. doi:10.3389/fgene.2019.00904

Shelley, M. W. 1818. Frankenstein, or, the modern Prometheus. Lackington, Hughes, Harding, Mavor, and Jones.

Steffen, M. M., B. S. Belisle, S. B. Watson, G. L. Boyer, R. A. Bourbonniere, and S. W. Wilhelm. 2015. Metatranscriptomic evidence for co-occurring top-down and bottom-up controls on toxic cyanobacterial communities. Appl. Environ. Microbiol. **81**: 3268–3276. doi:10.1128/AEM.04101-14

Steffen, M. M., and others. 2014. Nutrients drive transcriptional changes that maintain metabolic homeostasis but alter genome architecture in *Microcystis*. ISME J. **8**: 2080–2092. doi:10.1038/ismej.2014.78

Steffen, M. M., and others. 2017. Ecophysiological examination of the Lake Erie *Microcystis* bloom in 2014: Linkages between biology and the water supply shutdown of Toledo, OH. Environ. Sci. Technol. **51**: 6745–6755. doi:10.1021/acs.est.7b00856

Stough, J. M., and others. 2017. Molecular prediction of lytic *vs* lysogenic states for *Microcystis* phage: Metatranscriptomic evidence of lysogeny during large bloom events. PLoS One **12**: e0184146. doi:10.1371/journal.pone.0184146

Tang, X., and others. 2018. Seasonal gene expression and the ecophysiological implications of toxic *Microcystis aeruginosa* blooms in Lake Taihu. Environ. Sci. Technol. **52**: 11049–11059. doi:10.1021/acs.est.8b01066

Xie, M., and others. 2016. Metagenomic analysis reveals symbiotic relationship among bacteria in *Microcystis*-dominated community. Front. Microbiol. **7**: 56. doi:10.3389/fmicb.2016.00056

Zhang, Z., X.-X. Zhang, B. Wu, J. Yin, Y. Yu, and L. Yang. 2016. Comprehensive insights into microcystin-LR effects on hepatic lipid metabolism using cross-omics technologies. J. Hazard. Mater. **315**: 126–134. doi:10.1016/j.jhazmat.2016.05.011

Zhu, W., A. Lomsadze, and M. Borodovsky. 2010. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. **38**: e132. doi:10.1093/nar/gkq275

Acknowledgments

We would like to thank Justin Chaffin and the Ohio State SeaGrant Stone Lab for their assistance in sample collection. Robbie Martin and Brittany Zepernick were instrumental in the sampling efforts. Funding was provided by a grant to the *Great Lakes Center for Fresh Waters and Human Health* (NSF Grant OCE-1840715 and NIH Grant 1P01ES028939-01).

Submitted 19 July 2021
Revised 28 September 2021
Accepted 19 October 2021

Associate editor: Gordon T. Taylor