Metagenomic characterization of a harmful algal bloom using nanopore sequencing

Peter W. Schafran[*], Victor Cai[†], Hsiao-Pei Yang[*], Fay-Wei Li[*,‡]

[*]Boyce Thompson Institute, Ithaca, NY 14853, USA

[†]Department of Biology, Duke University, Durham, NC 27708, USA

[‡]Plant Biology Section, Cornell University, Ithaca, NY 14853, USA

Running Title: Nanopore HAB metagenome

Keywords: Nanopore, long-read sequencing, metagenomics, cyanobacteria, algal bloom,

Corresponding author:
Fay-Wei Li
Boyce Thompson Institute
533 Tower Rd.
Ithaca, NY 14853 USA
607-254-1244
fl329@cornell.edu

ABSTRACT

Water bodies around the world are increasingly threatened by harmful algal blooms (HABs) under current trends of rising water temperature and nutrient load. Metagenomic characterization of HABs can be combined with water quality and environmental data to better understand and predict the occurrence of toxic events. However, standard short-read sequencing typically yields highly fragmented metagenomes, preventing direct connection of genes to a single genome. Using Oxford Nanopore long-read sequencing, we were able to obtain high quality metagenome-assembled genomes, and show that dominant organisms in a HAB are readily identified, though different analyses disagreed on the identity of rare taxa. Genes from diverse functional categories were found not only in the most dominant genera, but also in several less common ones. Using simulated datasets, we show that the Flongle flowcell may provide an option for HAB monitoring with less data, at the expense of failing to detect rarer organisms and increasing fragmentation of the metagenome. Based on these results, we believe that Nanopore sequencing provides a fast, portable, and affordable method for studying HABs.

INTRODUCTION

Harmful algal blooms (HABs) pose serious threats to not only the aquatic biomes but also human health. HABs occur when toxic or ecosystem-disrupting species of algae and cyanobacteria rapidly increase in abundance in water bodies, leading to contamination of drinking water supplies and closure of recreational waterways and fisheries. The presence of HABs is associated with both acute toxicity, as well as chronic health issues such as cancer (Etheridge 2010, Gorham et al. 2020). The frequency of HABs has increased with warming water temperatures over the last few decades and may increase with additional change in temperature, acidification, nutrient load, and oxygenation of fresh and marine aquatic environments (Glibert et al. 2005). However, strong species and strain specific interactions between the environment and aquatic microbial communities make prediction of HABs difficult (Griffith and Gobler 2020), necessitating ongoing monitoring of species composition and presence of toxins and taste-and-odor compounds, such as microcystins, geosmin, and 2-methylisoborneol (Dietrich 2006). A survey of reservoir managers found that the control of algal blooms was their highest concern (Schafran 2005). Considerable costs are incurred to control HABs and their downstream effects, for example approximately $70 million spent over 10 years in Waco, TX (Dunlap et al. 2015) and $80 million spent annually by the US Army Corps of Engineers (Schafran 2005). Other estimates place costs of HABs at approximately $100 million dollars annually in the US (Hoagland and Scatasta 2006).

Metabarcoding and metagenomic approaches to studying HABs are increasingly used (Anderson 1995, Otten et al. 2016, Hennon and Dyhrman 2020, Hatfield et al. 2020), but as generally applied have some limitations. Molecular probes and PCR primers are designed with specific targets, and can introduce quantitative bias or may fail to amplify unknown members of a microbial community (Krehenwinkel et al. 2017). The large (i.e. 23S/28S) and small (i.e. 16S/18S) ribosomal subunits and ribosomal internal transcribed spacers (ITS) have been most commonly employed as a metabarcode (Lepère et al. 2000, Sherwood and Presting 2007, Stern et al. 2012, Gamez 2018, Zhang et al. 2020), though primers targeting toxin and taste-and-odor compound-producing genes have also been developed (Tsao et al. 2014, Suurnäkki et al. 2015, Legrand et al. 2016). Metabarcoding and PCR-targeted approaches can only provide taxonomic identity and presence/absence data for subjects of interest. Total metagenomic assembly allows discovery of additional taxa and genes, though the typical 150-500 bp Illumina reads limit the ability of assembly algorithms to link broad genomic regions. Library preparation and sequencing typically rely on instrumentation, such as thermocyclers, qPCR machines, and DNA sequencers, that are too large or unstable to function under field conditions. The time between sample collection and data analysis is often longer than the 24-36 hours that can be necessary to take action to prevent toxins and taste-and-odor compounds from infiltrating drinking water systems.

The Oxford Nanopore MinION device offers a unique solution to overcome the above limitations on HAB microbiome research. MinION is a third generation sequencing platform that can produce an average of 15 Gbp of data consisting of long (> 10 kbp) reads, is portable, and is powered by USB (Oxford Nanopore 2020). Compared to Illumina short-read sequencing platforms with estimated error rates from 0.1-0.5% (Pfeiffer et al. 2018), the Nanopore error rate is much higher at 5-15% (Rang et al. 2018), though this can be reduced to about 1% by polishing the assembled contigs with read sequences (Vaser et al. 2017). The MinION's portability has made it a popular choice for DNA sequencing where controlled laboratory conditions are unavailable and/or real-time data are necessary, such as on Arctic/Antarctic expeditions (Johnson et al. 2017, Gowers et al. 2019), on the International Space Station (Castro-Wallace et al. 2017, Burton et al. 2020), in tropical rainforests (Menegon et al. 2017, Pomerantz et al. 2018), and in clinical settings (Wongsurawat et al. 2019). Oxford Nanopore recently released the inexpensive Flongle flow cell which allows for sequencing of up to 2 Gbp at a cost of <20% that of the MinION flow cell (https://nanoporetech.com/products/ accessed 10 August 2020). This may be valuable in applications where increasing the number of sampling events (e.g. long-term monitoring) is more valuable than deep sequencing depth.

In this study, we applied MinION to generate the first HAB metagenome on a water sample from Cayuga Lake, New York. To assess consistency of taxonomic annotations among the current tools on long-read DNA sequences, we evaluated out-of-the-box performance of Centrifuge (Kim et al. 2016), Kaiju (Menzel et al. 2016), Kraken2 (Wood et al. 2019), and BLAST+ (Camacho et al. 2009). We also created *in silico* datasets to mimic the Flongle flow cell output and evaluate its consistency with MinION data.

MATERIALS AND METHODS

Sampling

One surface water sample was collected from an ongoing HAB in Cayuga Lake at Taughannock Falls State Park in Tompkins Co., NY (approximately 42.5470 N, 76.5984 W) on July 15 2019. The water sample was split into 6 tubes (6ml each), and centrifuged at 21,000*g* for 10min to pellet bacteria. For each tube, DNA was extracted using E.Z.N.A. Plant DNA kit (Omega Bio-tek) following the manufacturer's protocol. The DNA was then pooled and concentrated by AMPure XP beads to a total of 50µl for library prep.

Sequencing and assembly

The Nanopore sequencing library was prepared using the Ligation Sequencing kit (SQK-LSK109) and sequenced on MinION R9 flowcell (FLO-MIN106D) for 60 hours. Signal files were basecalled with Guppy v3.1.5 (Oxford Nanopore) with the high accuracy flip-flop mode.

Reads were adapter trimmed with Porechop (https://github.com/rrwick/Porechop) and assembled using Flye v2.7 (Kolmorogov et al. 2019) in its metagenomic mode (--meta flag) with an estimated genome size of 50 Mb (based on a previous assembly of the data). Assembled contigs were polished with five iterations of Racon v1.4.3 (Vaser et al. 2017) followed by one round of polishing with Medaka (https://nanoporetech.github.io/medaka). Assembly graph structure was visualized with Bandage (Wick et al. 2015).

Taxonomic assignment and gene annotation

Polished contigs were taxonomically annotated using four tools in an out-of-the-box fashion (i.e. without creation of custom search databases or extensive search parameter optimization) in order to most closely replicate usage by end users. BLAST+ was used with the NCBI *nt* database downloaded on 10 June 2020. Centrifuge was used with the NCBI nucleotide non-redundant database last updated 3 March 2018 available from its website (https://ccb.jhu.edu/software/centrifuge/). Kaiju was run with its *nr_euk* database, which is the most inclusive and includes a subset of the NCBI *nr* database with sequences belonging to archaea, bacteria, viruses, fungi, and microbial eukaryotes (https://github.com/bioinformatics-centre/kaiju). Kraken2 was used with its option to create the NCBI *nt* database (kraken2-build --download-library nt --db nt) which was constructed on 26 February 2020. Results were interpreted based on either the best annotation for each contig selected by each program (Centrifuge, Kaiju, Kraken2) or the result with the highest bitscore (BLAST+). NCBI taxonomy database files (nodes.dmp and names.dmp) were used to parse taxonomic relationships of annotations. Gene prediction was performed using MetaGeneMark (Besemer and Borodovsky 1999, Zhu et al. 2010), while ribosomal RNA genes were identified with barrnap 0.9 (https://github.com/tseemann/barrnap) and compared to SILVA 138 LSU/SSU Ref NR 99 databases (Quast et al. 2013, https://www.arb-silva.de/documentation/release-138/). rRNA sequences were aligned with MAFFT v7.464 (Katoh and Standley 2013) and phylogenies inferred with IQ-TREE v1.6.12 (Nguyen et al. 2015). Genes were annotated into clustered orthologous groups (COGs) with eggNOG v5.0 (Huerta-Cepas et al. 2019). BUSCO 4.3.1 was used to estimate completeness of gene annotation (Seppey et al. 2019).

Flongle data simulation

To mimic output by the Nanopore Flongle flowcell, data were subset to approximately 2 Gbp by pseudo-randomly selecting reads using shuf (GNU coreutils 8.21). Ten subsets were generated and checked to ensure a similar length distribution to the complete dataset. Each subset was assembled, polished, and annotated with BLASTN as described above.

Data Availability

Supplemental files available through Figshare (https://figshare.com/projects/Nanopore_HAB_metagenome/92567) include the polished metagenome (File S1, Predicted genes and their locations within the metagenome (File S2), amino acid sequences (File S3), and taxonomic annotations for each contig at each taxonomic level (File S4).

RESULTS

Following adapter trimming, 3.74 million reads (9.35 Gbp) were retained with an N50 length of 3651 bp and a mean length of 2498 bp. Maximum read length was 110 kbp. Assembly and contig polishing produced a metagenome size of 48.9 Mbp consisting of 2183 contigs with an N50 of 90 kbp (Table 1; File S1).

Taxonomic Annotation

Taxonomic annotation varied considerably among the classification tools. The number of contigs that could not be classified or could only be classified at the root of the taxonomy ranged from 255 (Centrifuge) to 500 (Kaiju). Because the unclassified contigs tended to be short, these represent only 1.5% (Centrifuge, Kraken) to 15% (Kaiju) of the total length of the assembly (Figure S1). At the superkingdom level, the majority of classified contigs were identified as Bacteria (range 81-90%), while Eukaryota were represented by 10-18% of contigs. Archaea and viruses made up a very small proportion of the results, representing less than 0.5% of classified contigs. At the genus level, the number of taxa identified ranged from 263 (Kaiju) to 309 (Centrifuge). Only 53 genera were shared among all classification tools, representing between 43% to 77% of the total assembly, and of these only 37 genera individually represent >0.5% of the contigs (Figure 1). In each case, over 50% of contigs were identified as *Anabaena* or *Dolichospermum* (here combined due to recent taxonomic revisions between these genera; Wacklin et al. 2009). Taxa shared similar proportions among tools at other taxonomic levels (Figure S2, File S4). For individual contigs, 27% were identified to the same genus by all tools, while for an additional 17% of contigs three of the tools agreed on one genus. For 20% of contigs, there was no agreement between any classification tools (including failure to classify at genus level). These ambiguously annotated contigs on average tended to be longer than the average length of other contigs, with an N50 length of 152 kbp vs 90 kbp, respectively. The number of contigs vs. total assembly size per genus showed a positive linear relationship ($p < 2e^{-16}$, Kendall's $\tau = 0.54$).

Of the largest, high coverage cluster within the assembly graph—presumably representing *Anabaena/Dolichospermum* genome(s)—526 contigs form a relatively 'knotty' structure with many edges generally converging on two nodes (Figure 2). About 72% of contigs in this cluster were annotated as *Anabaena* or *Dolichospermum*. Coloration of edges by coverage suggests two levels, supported by Hartigan's dip test of unimodality ($p < 2.2e^{-16}$). This cluster may represent two genomes from closely related species, but attempts to isolate and reassemble them into more contiguous genomes were unsuccessful. There are 47 long contigs (>100 kbp) that are disconnected to the *Anabaena/Dolichospermum* cluster, which tend to have a lower coverage, except one 648 kbp contig annotated as *Anabaena/Dolichospermum* above 50X coverage. Three circular contigs with coverage from 390-628 X and lengths from 86-95 kbp were also annotated as *Anabaena/Dolichospermum*, and might be their plasmids.

Gene Annotation

Gene annotation identified 66,490 genes with an N50 length of 756 bp (Files S2, S3). BUSCO analysis with the Bacteria v10 dataset found 99% complete genes in our metagenome (including 88% in the *Anabaena/Dolichospermum* cluster), suggesting this annotation is fairly robust. Genes were found representing every COG category (Figure 3). After genes with unknown function, the largest COG represented genes involved with amino acid transport and metabolism. Several COGs had very few representative genes, including: 1) RNA processing and modification; 2) chromatin structure and dynamics; 3) extracellular structures; 4) nuclear structure; 5) cytoskeleton. While the most abundant genera in the assembly also had the highest number of genes, several genera that were relatively rare in the assembly showed a high number of genes distributed across the predominant COG categories (Figure 3).

Barrnap identified thirty-seven 16S and thirty-two 23S sequences with average lengths of 1185 bp and 2469 bp, respectively, though several fragmented sequences needed manual correction. Twenty-three 16S and fifteen 23S sequences matched entries in SILVA databases with greater than 95% identity and 95% coverage of the query sequence length. Phylogenies showed the most common rRNA sequences are nested in *Anabaena* (Figures S3, S4). Pairs of 16S and 23S from the same contig supported the presence of *Pseudanabaena*, *Candidatus* Nanopelagicus, *Candidatus* Fonsibacter, *Cercopagis pengoi,* and an unknown taxon sister to *Curvibacter, Rhodoferax,* and *Polaromonas.* In other cases 16S/23S pairs conflicted in their position to reference taxa. Several 16S/23S pairs nested within Eukaryota but with low similarity to all SILVA references may represent organellar rRNA sequences. However, these sequences also failed to match any sequence in NCBI *nt* with identity and alignment length above 90%.

Flongle Data Simulations

The 10 data subsets created to mimic output of the Flongle flowcell resulted in assemblies with an average size of 17 Mbp (standard deviation = 0.4 Mbp) composed of an average of 965 contigs (standard deviation = 46). While each subset represented 21% of the total data, the subset assemblies averaged 35% of the size of the total assembly. Similarly, the number of genes and total size of the genes both averaged 35% as large as the total assembly.

Taxonomic annotation by BLAST was largely consistent at the superkingdom level, though a small percentage of the subset assemblies was identified as Archaea, which was not present in the total assembly (Figure 4). For other superkingdoms, the total assembly falls within the kernel density estimate of the subset results. At the genus level, larger discrepancies appeared (Figure 4), with similar trends apparent at other taxonomic levels (Figure S5). *Anabaena/Dolichospermum* was more highly represented in the subset assemblies than the total assembly, while most other genera showed the opposite trend. The range of subset results was large for some genera, such as *Sphingomonas* with a range of 15 percentage points, *Sphingopyxis* with 11 percentage points, and unannotated contigs with 11 percentage points.

DISCUSSION

Our exploratory analysis using Nanopore DNA sequencing to characterize a harmful algal bloom highlights some potential uses for this technology, as well as some areas where caution should be used in interpreting the data. The main drawback of long-read DNA sequencing technology is a relatively high error rate compared to short-read sequencers. The error rate of raw sequences for the Nanopore flowcell and basecaller we used is estimated at roughly 10% (Wick et al. 2019), too high to be useful for k-mer based analysis or taxonomic annotation (data not shown). The development of Nanopore-focused assemblers that take into account the difference in coverage typical of metagenomic data, as well as polishing tools that remove errors through consensus calling and neural networks trained on specific models of Nanopore errors, provides methods to create highly accurate metagenomes without short-read sequences. These methods offer an advantage over short-read metagenomes by greatly increasing the contiguity of assembly with a small trade-off in accuracy. Our assembly returned many large contigs (N50 = 90 kbp), including one 2.3 Mbp contig identified as Hyphomonadacae bacterium that is about two-thirds the length of its most closely related bacterial genome in NCBI's Genome database. We expected a more contiguous genome of the predominant cyanobacteria in our sample (*Anabaena* and/or *Dolichospermum*), but those contigs formed a few highly connected and structurally unresolved clusters. This can occur when a genome contains long repeat-rich regions or when multiple organisms share portions of their genomes that are similar. The structure of the assembly graph combined with significant bimodal distribution in coverage suggests that at least two highly similar organisms may be present in the *Anabaena/Dolichospermum* cluster, resulting in several shared edges connecting otherwise distinct clusters of contigs. However, our attempts to isolate and reassemble the raw data contributing to the *Anabaena/Dolichospermum* cluster did not improve the assembly.

BLAST, Centrifuge, Kaiju, and Kraken2 all were able to provide taxonomic annotations for a majority of contigs. Kaiju failed to annotate longer contigs than the others, possibly because it uses a smaller protein search database derived from NCBI *nr*, as opposed to the nucleotide-based *nt* used by the other tools. The low level of shared genus richness—where only 27% of contigs were placed in the same genus by every tool and roughly 20% of the genera identified by each tool are shared among all four—shows that caution needs to be exercised in determining community composition. While the more abundant organisms in this community may be identified with some certainty, contigs receiving mixed taxonomic annotations may need to be manually inspected. We processed BLAST results using the match between each contig and the reference database with the highest score to assign an identity to the contig. Individual matches tended to be short, with an average of 537 bp, so it may be more informative to examine all matches across different parts of the contig and use taxonomic distance (least common ancestor) between best hits to refine the overall identity, similar to how Centrifuge, Kaiju, and Kraken2 operate. Kraken2 provides an option to filter results by a user-defined confidence score, requiring a given fraction of a sequence to match a taxonomic level to receive an annotation. However, even at a low confidence level of 0.1, Kraken2 failed to provide many meaningful results from our data with 68% of contigs unclassified and of those classified, many were placed at the order level and above, suggesting that the application of this approach for metagenomics is limited.

Any classification is only as good as the reference database being searched. While NCBI *nt/nr* are the traditional databases for classification software, inclusion of the NCBI Whole Genome Shotgun (WGS) database can greatly improve results (Martí and Garay 2019). The short matches between our metagenome and *nt/nr* databases highlight the need to incorporate more genomic data for comparison against the increasing number of long-read metagenomic surveys. The software we tested all provide ways to create custom search databases, but this can require technical knowledge and resources many end users do not possess (e.g. our server with 768 GB RAM was memory-limited when trying to create a Centrifuge database from the current *nt* database). As a result, software that are not provided with regularly updated search databases, such as the most recent Centrifuge *nt* release from April 2018, quickly become obsolete as new genomic data are regularly added to the source databases. Additionally, updates in nomenclature, such as the transfer of some *Anabaena* into *Dolichospermum,* and persistence of polyphyly in bacterial phylogenies will confound attempts to use a least common ancestor approach to reconcile multiple matches (Wacklin et al. 2009, Cirés and Ballot 2016). This may partially explain why the proportion of contigs placed into *Anabaena* vs. *Dolichospermum* varies between BLAST/Kraken2 vs. Centrifuge/Kaiju, but in combination the genera are roughly equal. An independent sample from the same HAB event was microscopically identified to *Dolichospermum* (Bloom code 19-3435-B2, Community Science Institute, Ithaca, NY), supporting our decision to combine *Anabaena* and *Dolichospermum* results.

Our gene annotation and classification showed a broad diversity of genes obtained not only from the most abundant organisms, but also from many with lower abundance in our sample. This demonstrates one of the advantages of long-read sequencing in that much longer contigs allow connecting more genes to a single taxon, in addition to being able to taxonomically place previously unknown genes or genes without a known function. Efforts to incorporate genomic and proteomic data into HAB prediction models will be aided by this technologic development (Hennon and Dyhrman 2020). While we did not identify any in our sample, genes related to production of toxin and taste-and-odor compounds, such as geosmin synthase, 2-methylisoborneol (MIB) synthase, the anatoxin-a synthetase gene cluster, and the microcystin synthetase gene cluster, could be easily gleaned from this type of metagenome. This is consistent with failure to detect microcystin in this HAB at the drinking water threshold (<3 μg/L) by the Community Science Institute (Ithaca, NY). The ability of the Nanopore platform to also sequence RNA suggests that monitoring the expression of these genes could provide a real-time, field-based method to track the production of these dangerous compounds.

We created ten datasets *in silico* to estimate the potential of Oxford Nanopore's Flongle flow cell to assemble and characterize a metagenome as well as the MinION flow cell that produced 5 times more data. The average Flongle assembly size was disproportionately large compared to the reduced size of the datasets, suggesting that some of the MinION data could be considered excess. However, the distribution of contig length shows the Flongle assemblies varied tremendously, with the length of the longest contig ranging from 2.5 to 0.6 Mbp and N50 ranging from 60 to 78 Kbp. Since the length distribution of the reads in each dataset was similar, this suggests the difference in assemblies is due to random assortment of reads derived from different organisms. While the proportion of the Flongle assemblies annotated to each taxonomic group was similar to the MinION assembly at the superkingdom level, increasing levels of taxonomic resolution showed the most prevalent organisms were overrepresented in the Flongle results, while the least prevalent organisms tended to be underrepresented. This is consistent with the methodology of the assembly and polishing software, where contigs with very low coverage are removed from the final assembly. Based on our results the Flongle may be useful for capturing large genomic fractions of the predominant members of an HAB, though it is more likely to miss minor taxa in the community.

Conclusions

The profound effects of harmful algal blooms (HABs) on human health, economic activity, and ecological function compounded with the expectation that changes to the environment will alter the frequency and intensity of HABs necessitates a better understanding of the patterns and processes driving these events. Genomics provide complementary evidence to traditional data focused on taxonomic identification, quantification, and water quality measurements. Comparing four tools used for classifying DNA sequences, we found that assigning taxonomy with high confidence in our Nanopore-derived HAB metagenome was often difficult, though as more complete genomes are added to reference databases, we expect this to be resolved. We recovered a diversity of genes from many functional groups, even from organisms present at relatively low abundance in the metagenome. Our *in silico* datasets mimicking the Flongle flow cell for the Nanopore MinION and GridION sequencers suggests this may be a valuable tool for water quality managers who need to repeatedly monitor for HABs, with the tradeoffs being decreased detection of low abundance organisms and greater fragmentation of the metagenome. Given the portability, affordability, and short time from sample-to-sequence of the Nanopore platform, we believe this form of long-read sequencing shows potential for the study and monitoring of algal blooms.

ACKNOWLEDGEMENTS

We thank Taughannock Falls State Park for access to the collection site and the Community Science Institute for sharing their data from this bloom event. V.C. thanks Duncan Hauser and Alaina Petlewski for laboratory training.

FUNDING

This project was supported by National Science Foundation (NSF) grant no. DEB-1831428 to F.-W.L, and NSF Research Experience for Undergraduates award no. DBI-1850796 to V.C.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

Table 1. Summary statistics of assemblies and predicted genes. Flongle results represent mean (standard deviation) of all 10 subsets.

Assembly	**Total Length (Mbp)**	**Number**	**Mean Length (bp)**	**Longest (Mbp)**	**Shortest (bp)**	**N50**	**N70**	**N90**
MinION	48.9	2183	22408	2.30	80	90015	42028	12372
Flongle	17.1 (0.441)	965 (46.4)	17773 (754)	1.31 (0.625)	67 (28)	64864 (4341)	37296 (1432)	10570 (1369)
Predicted Genes	**Total Length (Mbp)**	**Number**	**Mean Length (bp)**	**Longest (bp)**	**Shortest (bp)**	**N50**	**N70**	**N90**
MinION	12.3	66490	184.8	4721	18	252	166	90
Flongle	4.28 (0.103)	23268 (720)	184.2 (1.92)	3379 (572)	18 (0)	255.3 (2.75)	164.1 (2.02)	89.1 (1.10)

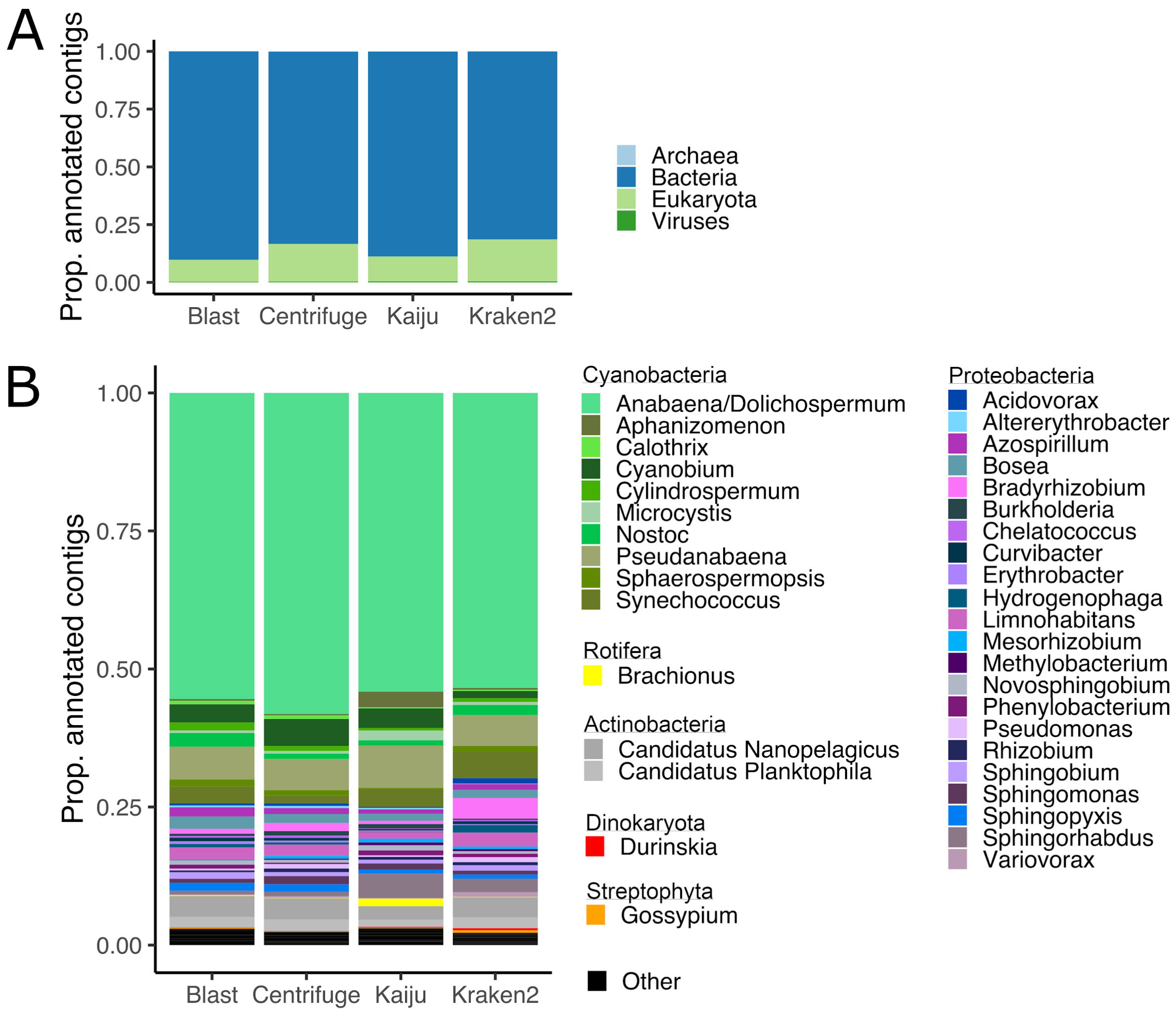
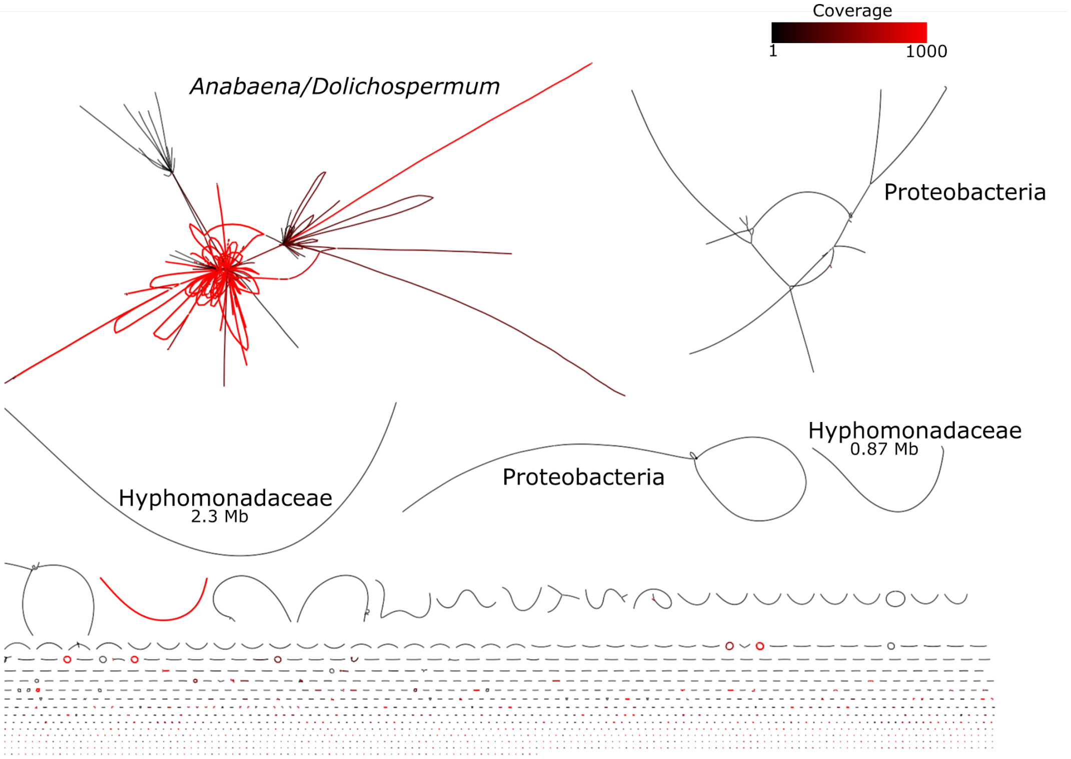


Figure 1. Proportion of contigs classified to superkingdoms (A) and genera (B) by each classification tool. Only genera identified by all classification tools that represent more than 0.5% of assembly shown. Individual genera present at <0.5% were combined into 'Other'. Genera grouped by phylum.



405

406 **Figure 2**. Assembly graph colored by coverage. Five largest clusters annotated with taxon at

407 lowest level that provides majority agreement among all contigs. Contig lengths to scale.

408

409

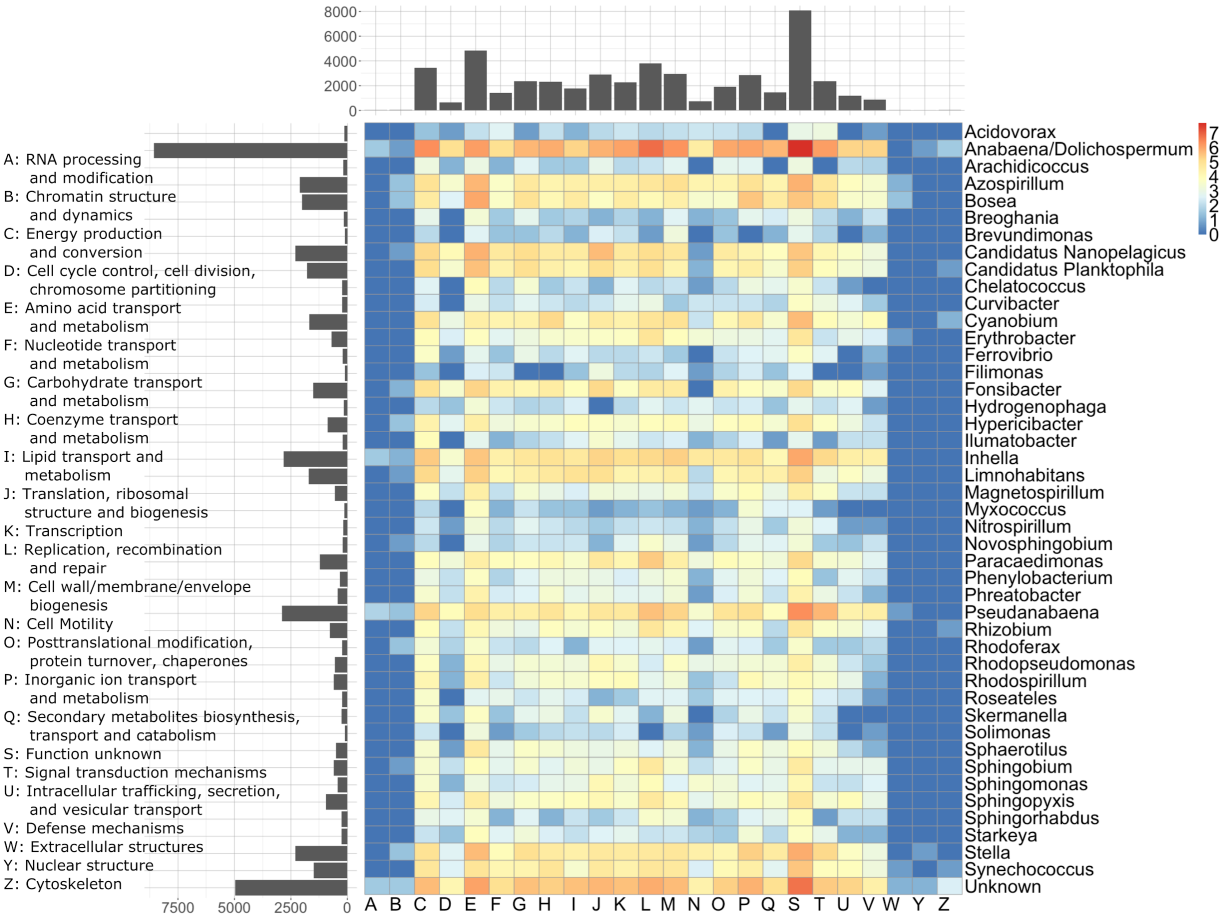


Figure 3. 2D histogram of log-transformed COG annotations for genera with more than 100 annotated genes. Color legend represents log10-transformed number of genes. Histograms represent number of genes in respective row/column. X-axis categories described in left-hand legend.

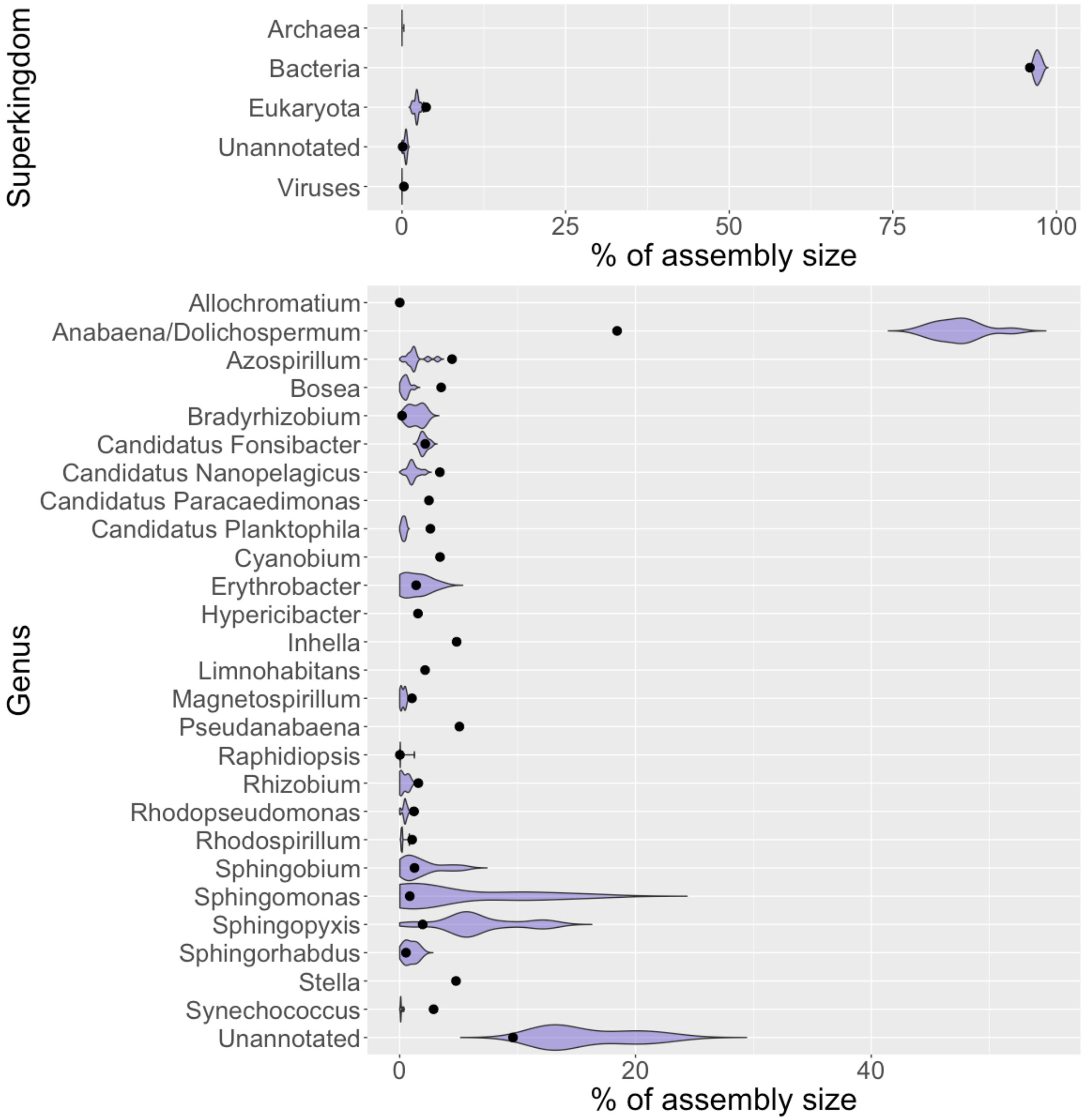
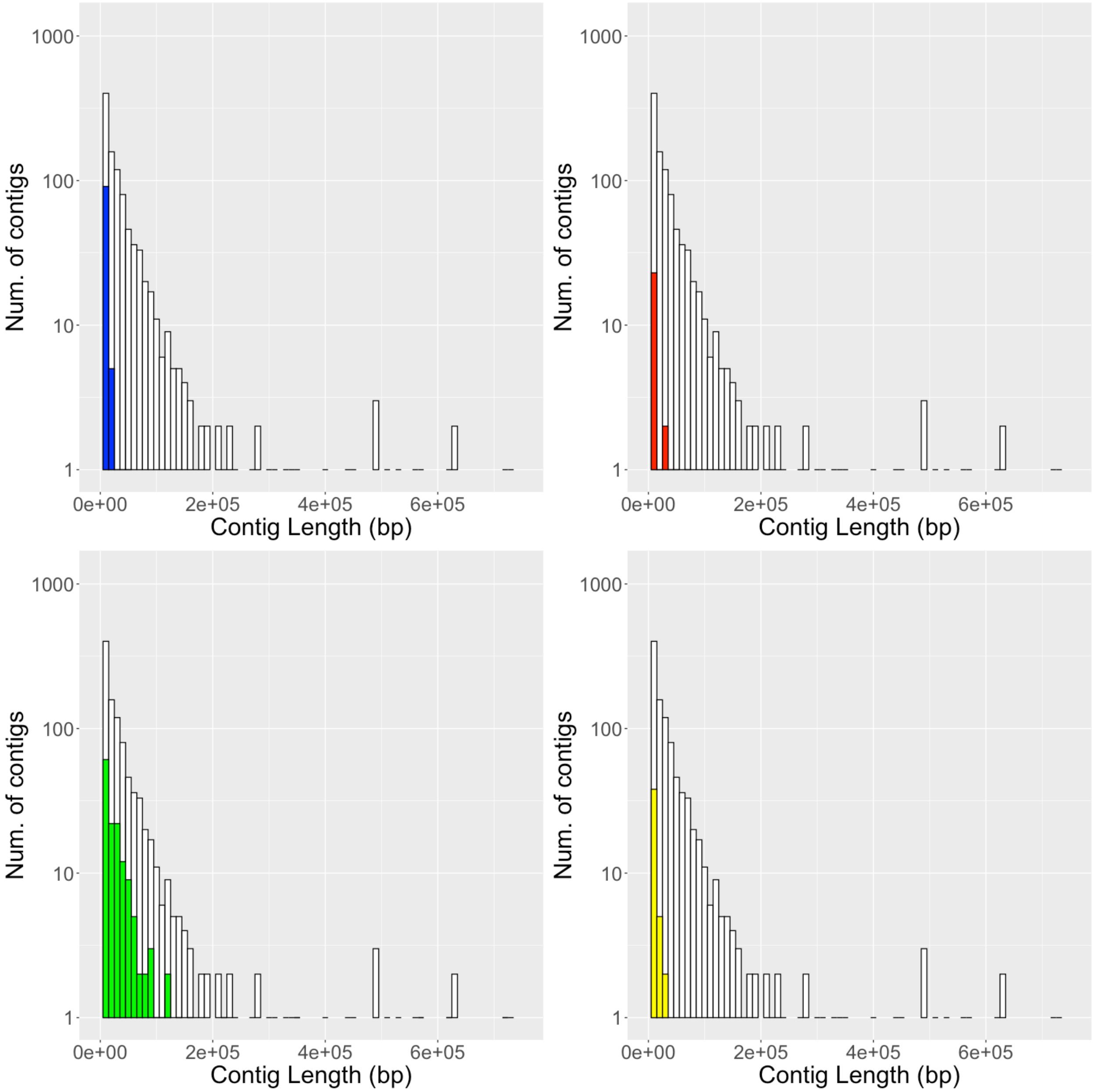


Figure 4. Comparison of assembly sizes for superkingdoms and genera between total assembly (dots) and Flongle-mimic subsets (violin plots) as annotated by BLAST. Only genera representing >1% of assembly shown.



423

424 **Figure S1**. Histograms of unclassified contigs by length. Blue = Blast, Red = Centrifuge, Green
425 = Kaiju, Yellow = Kraken2, White = all contigs.

426

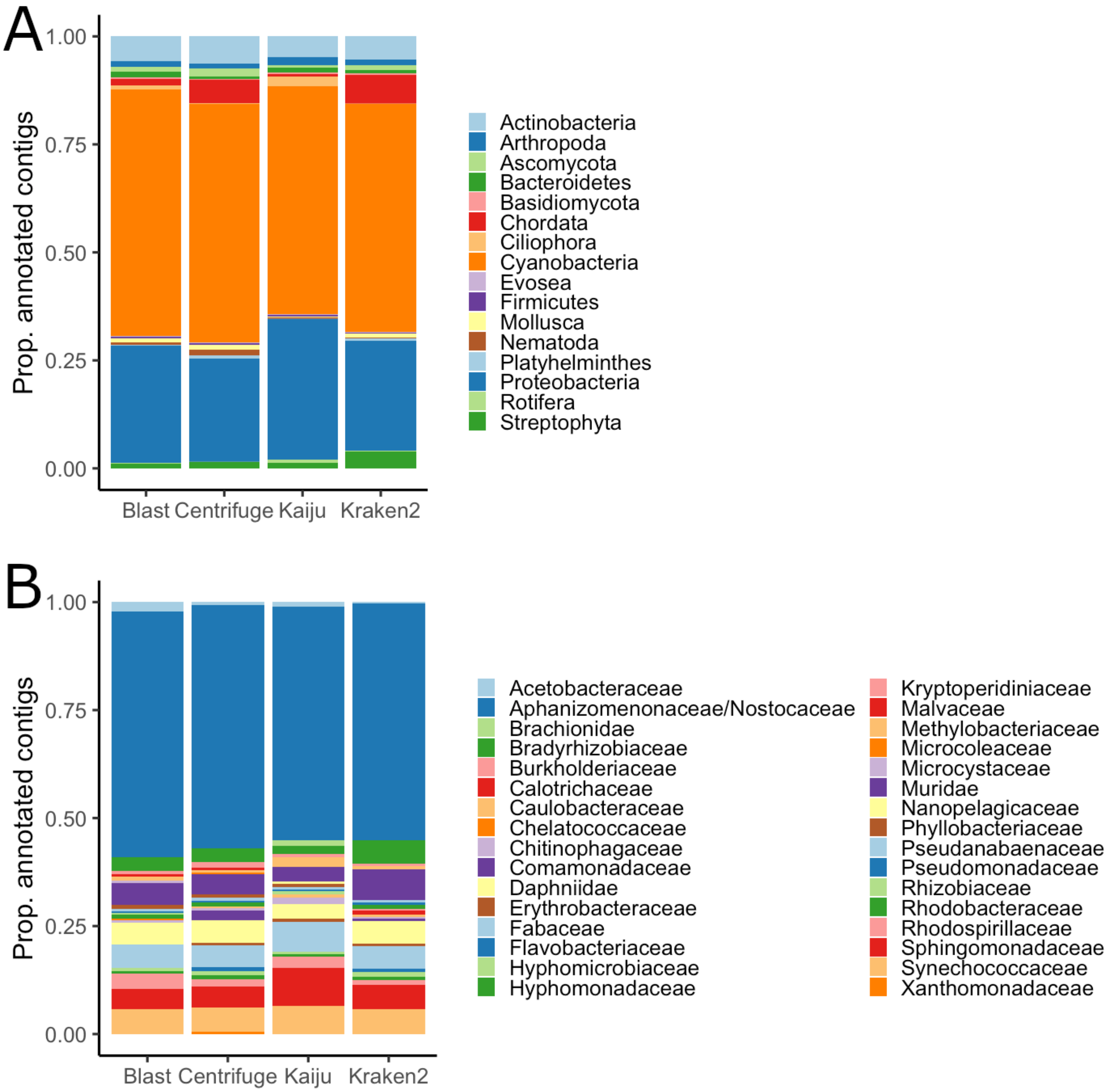
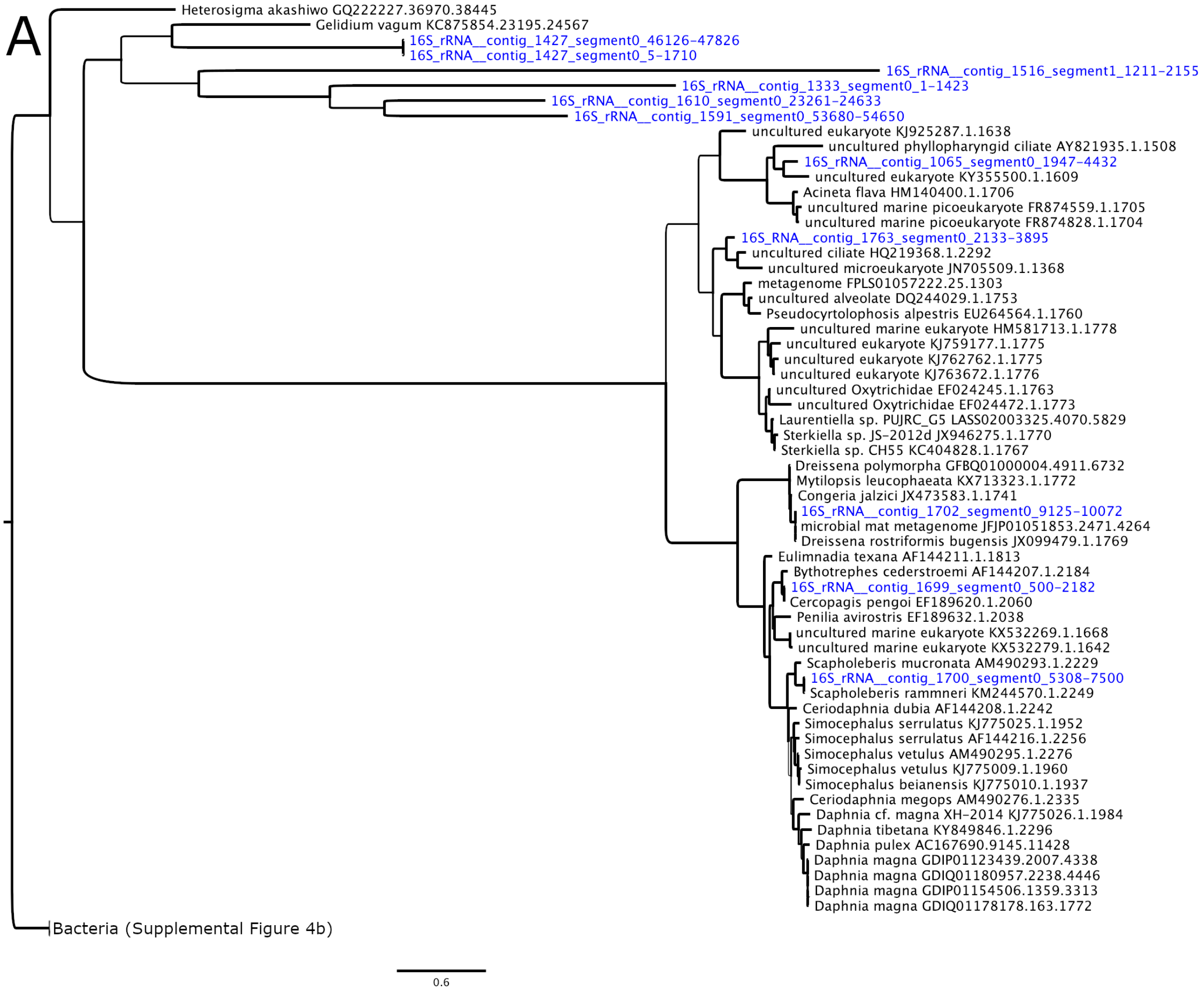


Figure S2. Proportion of contigs classified to phylum (A) and family (B) by each classification tool. Only taxonomic groups identified by all classification tools shown.


A
Heterosigma akashiwo GQ222227.36970.38445
Gelidium vagum KC875854.23195.24567
16S_rRNA__contig_1427_segment0_46126-47826
16S_rRNA__contig_1427_segment0_5-1710
16S_rRNA__contig_1516_segment1_1211-2155
16S_rRNA__contig_1333_segment0_1-1423
16S_rRNA__contig_1610_segment0_23261-24633
16S_rRNA__contig_1591_segment0_53680-54650
uncultured eukaryote KJ925287.1.1638
uncultured phyllopharyngid ciliate AY821935.1.1508
16S_rRNA__contig_1065_segment0_1947-4432
uncultured eukaryote KY355500.1.1609
Acineta flava HM140400.1.1706
uncultured marine picoeukaryote FR874559.1.1705
uncultured marine picoeukaryote FR874828.1.1704
16S_RNA__contig_1763_segment0_2133-3895
uncultured ciliate HQ219368.1.2292
uncultured microeukaryote JN705509.1.1368
metagenome FPLS01057222.25.1303
uncultured alveolate DQ244029.1.1753
Pseudocyrtolophosis alpestris EU264564.1.1760
uncultured marine eukaryote HM581713.1.1778
uncultured eukaryote KJ759177.1.1775
uncultured eukaryote KJ762762.1.1775
uncultured eukaryote KJ763672.1.1776
uncultured Oxytrichidae EF024245.1.1763
uncultured Oxytrichidae EF024472.1.1773
Laurentiella sp. PUJRC_G5 LASS02003325.4070.5829
Sterkiella sp. JS-2012d JX946275.1.1770
Sterkiella sp. CH55 KC404828.1.1767
Dreissena polymorpha GFBQ01000004.4911.6732
Mytilopsis leucophaeata KX713323.1.1772
Congeria jalzici JX473583.1.1741
16S_rRNA__contig_1702_segment0_9125-10072
microbial mat metagenome JFJP01051853.2471.4264
Dreissena rostriformis bugensis JX099479.1.1769
Eulimnadia texana AF144211.1.1813
Bythotrephes cederstroemi AF144207.1.2184
16S_rRNA__contig_1699_segment0_500-2182
Cercopagis pengoi EF189620.1.2060
Penilia avirostris EF189632.1.2038
uncultured marine eukaryote KX532269.1.1668
uncultured marine eukaryote KX532279.1.1642
Scapholeberis mucronata AM490293.1.2229
16S_rRNA__contig_1700_segment0_5308-7500
Scapholeberis rammneri KM244570.1.2249
Ceriodaphnia dubia AF144208.1.2242
Simocephalus serrulatus KJ775025.1.1952
Simocephalus serrulatus AF144216.1.2256
Simocephalus vetulus AM490295.1.2276
Simocephalus vetulus KJ775009.1.1960
Simocephalus beianensis KJ775010.1.1937
Ceriodaphnia megops AM490276.1.2335
Daphnia cf. magna XH-2014 KJ775026.1.1984
Daphnia tibetana KY849846.1.2296
Daphnia pulex AC167690.9145.11428
Daphnia magna GDIP01123439.2007.4338
Daphnia magna GDIQ01180957.2238.4446
Daphnia magna GDIP01154506.1359.3313
Daphnia magna GDIQ01178178.163.1772
Bacteria (Supplemental Figure 4b)
0.6

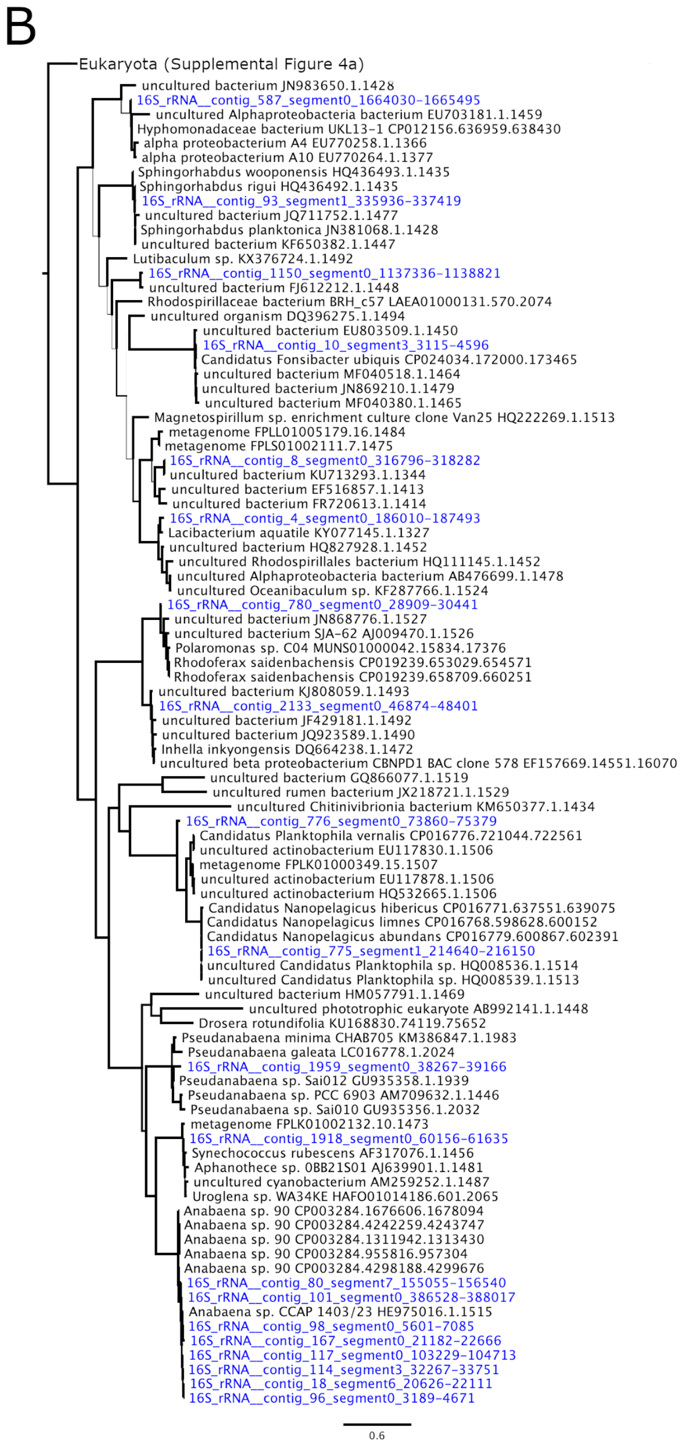
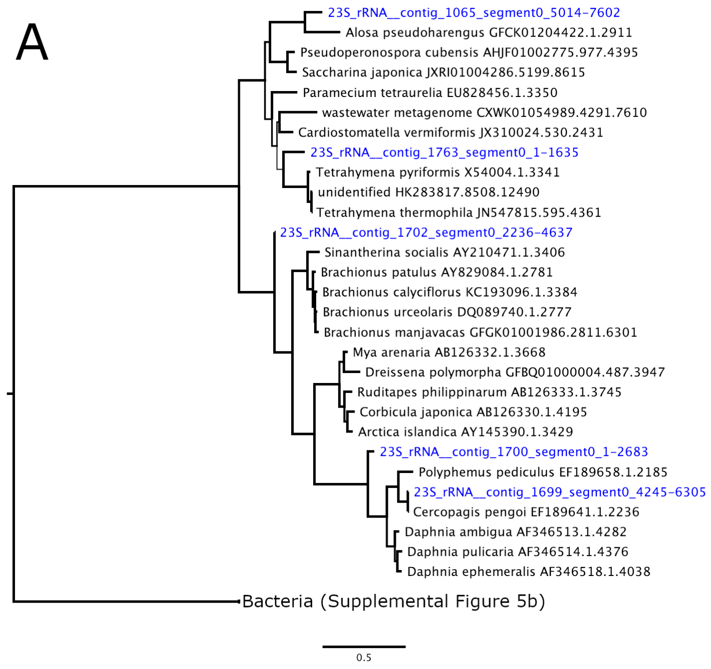

B
Eukaryota (Supplemental Figure 4a)
uncultured bacterium JN983650.1.1428
16S_rRNA__contig_587_segment0_1664030-1665495
uncultured Alphaproteobacteria bacterium EU703181.1.1459
Hyphomonadaceae bacterium UKL13-1 CP012156.636959.638430
alpha proteobacterium A4 EU770258.1.1366
alpha proteobacterium A10 EU770264.1.1377
Sphingorhabdus wooponensis HQ436493.1.1435
Sphingorhabdus rigui HQ436492.1.1435
16S_rRNA__contig_93_segment1_335936-337419
uncultured bacterium JQ711752.1.1477
Sphingorhabdus planktonica JN381068.1.1428
uncultured bacterium KF650382.1.1447
Lutibaculum sp. KX376724.1.1492
16S_rRNA__contig_1150_segment0_1137336-1138821
uncultured bacterium FJ612212.1.1448
Rhodospirillaceae bacterium BRH_c57 LAEA01000131.570.2074
uncultured organism DQ396275.1.1494
uncultured bacterium EU803509.1.1450
16S_rRNA__contig_10_segment3_3115-4596
Candidatus Fonsibacter ubiquis CP024034.172000.173465
uncultured bacterium MF040518.1.1464
uncultured bacterium JN869210.1.1479
uncultured bacterium MF040380.1.1465
Magnetospirillum sp. enrichment culture clone Van25 HQ222269.1.1513
metagenome FPLL01005179.16.1484
metagenome FPLS01002111.7.1475
16S_rRNA__contig_8_segment0_316796-318282
uncultured bacterium KU713293.1.1344
uncultured bacterium EF516857.1.1413
uncultured bacterium FR720613.1.1414
16S_rRNA__contig_4_segment0_186010-187493
Lacibacterium aquatile KY077145.1.1327
uncultured bacterium HQ827928.1.1452
uncultured Rhodospirillales bacterium HQ111145.1.1452
uncultured Alphaproteobacteria bacterium AB476699.1.1478
uncultured Oceanibaculum sp. KF287766.1.1524
16S_rRNA__contig_780_segment0_28909-30441
uncultured bacterium JN868776.1.1527
uncultured bacterium SJA-62 AJ009470.1.1526
Polaromonas sp. C04 MUNS01000042.15834.17376
Rhodoferax saidenbachensis CP019239.653029.654571
Rhodoferax saidenbachensis CP019239.658709.660251
uncultured bacterium KJ808059.1.1493
16S_rRNA__contig_2133_segment0_46874-48401
uncultured bacterium JF429181.1.1492
uncultured bacterium JQ923589.1.1490
Inhella inkyongensis DQ664238.1.1472
uncultured beta proteobacterium CBNPD1 BAC clone 578 EF157669.14551.16070
uncultured bacterium GQ866077.1.1519
uncultured rumen bacterium JX218721.1.1529
uncultured Chitinivibrionia bacterium KM650377.1.1434
16S_rRNA__contig_776_segment0_73860-75379
Candidatus Planktophila vernalis CP016776.721044.722561
uncultured actinobacterium EU117830.1.1506
metagenome FPLK01000349.15.1507
uncultured actinobacterium EU117878.1.1506
uncultured actinobacterium HQ532665.1.1506
Candidatus Nanopelagicus hibericus CP016771.637551.639075
Candidatus Nanopelagicus limnes CP016768.598628.600152
Candidatus Nanopelagicus abundans CP016779.600867.602391
16S_rRNA__contig_775_segment1_214640-216150
uncultured Candidatus Planktophila sp. HQ008536.1.1514
uncultured Candidatus Planktophila sp. HQ008539.1.1513
uncultured bacterium HM057791.1.1469
uncultured phototrophic eukaryote AB992141.1.1448
Drosera rotundifolia KU168830.74119.75652
Pseudanabaena minima CHAB705 KM386847.1.1983
Pseudanabaena galeata LC016778.1.2024
16S_rRNA__contig_1959_segment0_38267-39166
Pseudanabaena sp. Sai012 GU935358.1.1939
Pseudanabaena sp. PCC 6903 AM709632.1.1446
Pseudanabaena sp. Sai010 GU935356.1.2032
metagenome FPLK01002132.10.1473
16S_rRNA__contig_1918_segment0_60156-61635
Synechococcus rubescens AF317076.1.1456
Aphanothece sp. 0BB21S01 AJ639901.1.1481
uncultured cyanobacterium AM259252.1.1487
Uroglena sp. WA34KE HAFO01014186.601.2065
Anabaena sp. 90 CP003284.1676606.1678094
Anabaena sp. 90 CP003284.4242259.4243747
Anabaena sp. 90 CP003284.1311942.1313430
Anabaena sp. 90 CP003284.955816.957304
Anabaena sp. 90 CP003284.4298188.4299676
16S_rRNA__contig_80_segment7_155055-156540
16S_rRNA__contig_101_segment0_386528-388017
Anabaena sp. CCAP 1403/23 HE975016.1.1515
16S_rRNA__contig_98_segment0_5601-7085
16S_rRNA__contig_167_segment0_21182-22666
16S_rRNA__contig_117_segment0_103229-104713
16S_rRNA__contig_114_segment3_32267-33751
16S_rRNA__contig_18_segment6_20626-22111
16S_rRNA__contig_96_segment0_3189-4671
0.6

Figure S3. 16S rRNA phylogeny of metagenome-recovered sequences (blue) and SILVA references (black). Tree rooted along the branch separating Eukaryota (A) and Bacteria (B). Branch thickness is proportional to bootstrap support with scale bar width = 100.


A
23S_rRNA__contig_1065_segment0_5014-7602
Alosa pseudoharengus GFCK01204422.1.2911
Pseudoperonospora cubensis AHJF01002775.977.4395
Saccharina japonica JXRI01004286.5199.8615
Paramecium tetraurelia EU828456.1.3350
wastewater metagenome CXWK01054989.4291.7610
Cardiostomatella vermiformis JX310024.530.2431
23S_rRNA__contig_1763_segment0_1-1635
Tetrahymena pyriformis X54004.1.3341
unidentified HK283817.8508.12490
Tetrahymena thermophila JN547815.595.4361
23S_rRNA__contig_1702_segment0_2236-4637
Sinantherina socialis AY210471.1.3406
Brachionus patulus AY829084.1.2781
Brachionus calyciflorus KC193096.1.3384
Brachionus urceolaris DQ089740.1.2777
Brachionus manjavacas GFGK01001986.2811.6301
Mya arenaria AB126332.1.3668
Dreissena polymorpha GFBQ01000004.487.3947
Ruditapes philippinarum AB126333.1.3745
Corbicula japonica AB126330.1.4195
Arctica islandica AY145390.1.3429
23S_rRNA__contig_1700_segment0_1-2683
Polyphemus pediculus EF189658.1.2185
23S_rRNA__contig_1699_segment0_4245-6305
Cercopagis pengoi EF189641.1.2236
Daphnia ambigua AF346513.1.4282
Daphnia pulicaria AF346514.1.4376
Daphnia ephemeralis AF346518.1.4038
Bacteria (Supplemental Figure 5b)
0.5

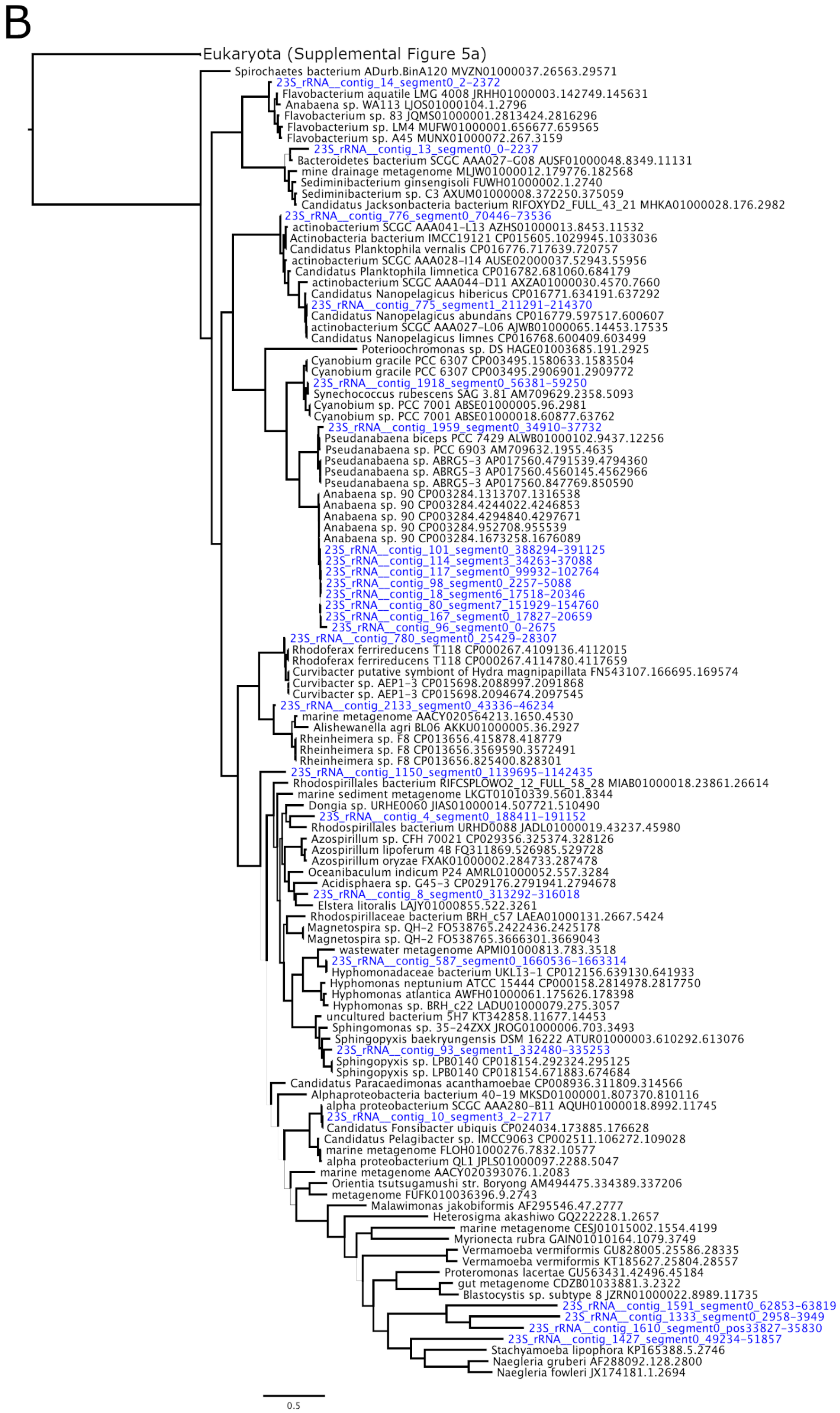

B
Eukaryota (Supplemental Figure 5a)
Spirochaetes bacterium ADurb.BinA120 MVZN01000037.26563.29571
23S_rRNA__contig_14_segment0_2-2372
Flavobacterium aquatile LMG 4008 JRHH01000003.142749.145631
Anabaena sp. WA113 LJOS01000104.1.2796
Flavobacterium sp. 83 JQMS01000001.2813424.2816296
Flavobacterium sp. LM4 MUFW01000001.656677.659565
Flavobacterium sp. A45 MUNX01000072.267.3159
23S_rRNA__contig_13_segment0_0-2237
Bacteroidetes bacterium SCGC AAA027-G08 AUSF01000048.8349.11131
mine drainage metagenome MLJW01000012.179776.182568
Sediminibacterium ginsengisoli FUWH01000002.1.2740
Sediminibacterium sp. C3 AXUM01000008.372250.375059
Candidatus Jacksonbacteria bacterium RIFOXYD2_FULL_43_21 MHKA01000028.176.2982
23S_rRNA__contig_776_segment0_70446-73536
actinobacterium SCGC AAA041-L13 AZHS01000013.8453.11532
Actinobacteria bacterium IMCC19121 CP015605.1029945.1033036
Candidatus Planktophila vernalis CP016776.717639.720757
actinobacterium SCGC AAA028-I14 AUSE02000037.52943.55956
Candidatus Planktophila limnetica CP016782.681060.684179
actinobacterium SCGC AAA044-D11 AXZA01000030.4570.7660
Candidatus Nanopelagicus hibericus CP016771.634191.637292
23S_rRNA__contig_775_segment1_211291-214370
Candidatus Nanopelagicus abundans CP016779.597517.600607
actinobacterium SCGC AAA027-L06 AJWB01000065.14453.17535
Candidatus Nanopelagicus limnes CP016768.600409.603499
Poterioochromonas sp. DS HAGE01003685.191.2925
Cyanobium gracile PCC 6307 CP003495.1580633.1583504
Cyanobium gracile PCC 6307 CP003495.2906901.2909772
23S_rRNA__contig_1918_segment0_56381-59250
Synechococcus rubescens SAG 3.81 AM709629.2358.5093
Cyanobium sp. PCC 7001 ABSE01000005.96.2981
Cyanobium sp. PCC 7001 ABSE01000018.60877.63762
23S_rRNA__contig_1959_segment0_34910-37732
Pseudanabaena biceps PCC 7429 ALWB01000102.9437.12256
Pseudanabaena sp. PCC 6903 AM709632.1955.4635
Pseudanabaena sp. ABRG5-3 AP017560.4791539.4794360
Pseudanabaena sp. ABRG5-3 AP017560.4560145.4562966
Pseudanabaena sp. ABRG5-3 AP017560.847769.850590
Anabaena sp. 90 CP003284.1313707.1316538
Anabaena sp. 90 CP003284.4244022.4246853
Anabaena sp. 90 CP003284.4294840.4297671
Anabaena sp. 90 CP003284.952708.955539
Anabaena sp. 90 CP003284.1673258.1676089
23S_rRNA__contig_101_segment0_388294-391125
23S_rRNA__contig_114_segment3_34263-37088
23S_rRNA__contig_117_segment0_99932-102764
23S_rRNA__contig_98_segment0_2257-5088
23S_rRNA__contig_18_segment6_17518-20346
23S_rRNA__contig_80_segment7_151929-154760
23S_rRNA__contig_167_segment0_17827-20659
23S_rRNA__contig_96_segment0_0-2675
23S_rRNA__contig_780_segment0_25429-28307
Rhodoferax ferrireducens T118 CP000267.4109136.4112015
Rhodoferax ferrireducens T118 CP000267.4114780.4117659
Curvibacter putative symbiont of Hydra magnipapillata FN543107.166695.169574
Curvibacter sp. AEP1-3 CP015698.2088997.2091868
Curvibacter sp. AEP1-3 CP015698.2094674.2097545
23S_rRNA__contig_2133_segment0_43336-46234
marine metagenome AACY020564213.1650.4530
Alishewanella agri BL06 AKKU01000005.36.2927
Rheinheimera sp. F8 CP013656.415878.418779
Rheinheimera sp. F8 CP013656.3569590.3572491
Rheinheimera sp. F8 CP013656.825400.828301
23S_rRNA__contig_1150_segment0_1139695-1142435
Rhodospirillales bacterium RIFCSPLOWO2_12_FULL_58_28 MIAB01000018.23861.26614
marine sediment metagenome LKGT01010339.5601.8344
Dongia sp. URHE0060 JIAS01000014.507721.510490
23S_rRNA__contig_4_segment0_188411-191152
Rhodospirillales bacterium URHD0088 JADL01000019.43237.45980
Azospirillum sp. CFH 70021 CP029356.325374.328126
Azospirillum lipoferum 4B FQ311869.526985.529728
Azospirillum oryzae FXAK01000002.284733.287478
Oceanibaculum indicum P24 AMRL01000052.557.3284
Acidisphaera sp. G45-3 CP029176.2791941.2794678
23S_rRNA__contig_8_segment0_313292-316018
Elstera litoralis LAJY01000855.522.3261
Rhodospirillaceae bacterium BRH_c57 LAEA01000131.2667.5424
Magnetospira sp. QH-2 FO538765.2422436.2425178
Magnetospira sp. QH-2 FO538765.3666301.3669043
wastewater metagenome APMI01000813.783.3518
23S_rRNA__contig_587_segment0_1660536-1663314
Hyphomonadaceae bacterium UKL13-1 CP012156.639130.641933
Hyphomonas neptunium ATCC 15444 CP000158.2814978.2817750
Hyphomonas atlantica AWFH01000061.175626.178398
Hyphomonas sp. BRH_c22 LADU01000079.275.3057
uncultured bacterium 5H7 KT342858.11677.14453
Sphingomonas sp. 35-24ZXX JROG01000006.703.3493
Sphingopyxis baekryungensis DSM 16222 ATUR01000003.610292.613076
23S_rRNA__contig_93_segment1_332480-335253
Sphingopyxis sp. LPB0140 CP018154.292324.295125
Sphingopyxis sp. LPB0140 CP018154.671883.674684
Candidatus Paracaedimonas acanthamoebae CP008936.311809.314566
Alphaproteobacteria bacterium 40-19 MKSD01000001.807370.810116
alpha proteobacterium SCGC AAA280-B11 AQUH01000018.8992.11745
23S_rRNA__contig_10_segment3_2-2717
Candidatus Fonsibacter ubiquis CP024034.173885.176628
Candidatus Pelagibacter sp. IMCC9063 CP002511.106272.109028
marine metagenome FLOH01000276.7832.10577
alpha proteobacterium QL1 JPLS01000097.2288.5047
marine metagenome AACY020393076.1.2083
Orientia tsutsugamushi str. Boryong AM494475.334389.337206
metagenome FUFK010036396.9.2743
Malawimonas jakobiformis AF295546.47.2777
Heterosigma akashiwo GQ222228.1.2657
marine metagenome CESJ01015002.1554.4199
Myrionecta rubra GAIN01010164.1079.3749
Vermamoeba vermiformis GU828005.25586.28335
Vermamoeba vermiformis KT185627.25804.28557
Proteromonas lacertae GU563431.42496.45184
gut metagenome CDZB01033881.3.2322
Blastocystis sp. subtype 8 JZRN01000022.8989.11735
23S_rRNA__contig_1591_segment0_62853-63819
23S_rRNA__contig_1333_segment0_2958-3949
23S_rRNA__contig_1610_segment0_pos33827-35830
23S_rRNA__contig_1427_segment0_49234-51857
Stachyamoeba lipophora KP165388.5.2746
Naegleria gruberi AF288092.128.2800
Naegleria fowleri JX174181.1.2694
0.5

Figure S4. 23S rRNA phylogeny of metagenome-recovered sequences (blue) and SILVA references (black). Tree rooted along the branch separating Eukaryota (A) and Bacteria (B). Branch thickness is proportional to bootstrap support with scale bar width = 100.

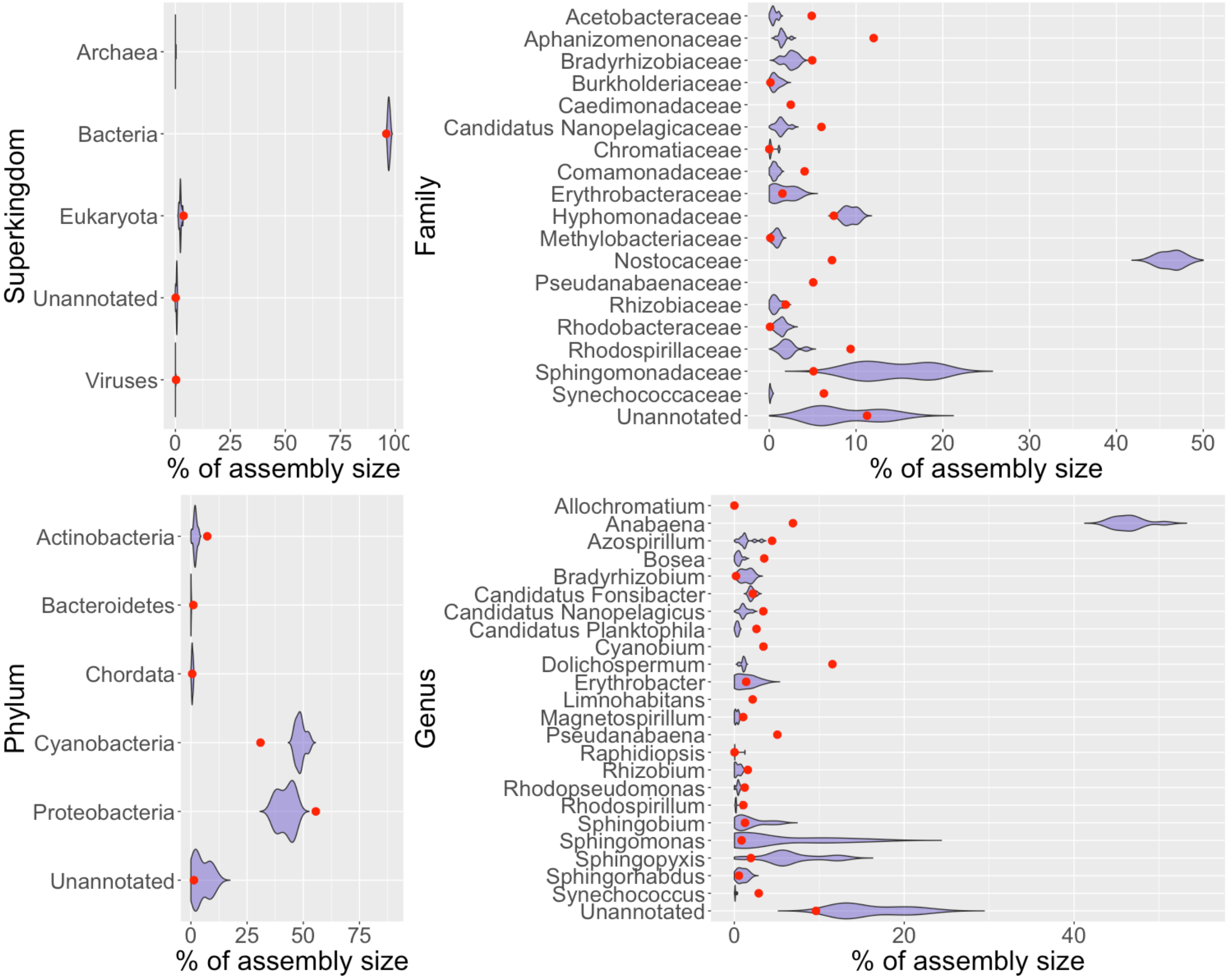


Figure S5. Comparison of assembly sizes for superkingdoms, phyla, families, and genera between total assembly (dots) and Flongle-mimic subsets (violin plots) as annotated by BLAST. Only taxonomic groups representing > 1% of assembly shown.

REFERENCES

Besemer, J., and M. Borodovsky, 1999 Heuristic approach to deriving models for gene finding. Nucleic Acids Research 27(19): 3911-3920. https://doi.org/10.1093/nar/27.19.3911

Burton, A.S., S.E. Stahl, K.K John, M. Jain, S. Juul, et al., 2020 Off Earth identification of bacterial populations using 16S rDNA nanopore sequencing. Genes 11: 76.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, et al., 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Castro-Wallace, S.L., C.Y. Chiu, K.K. John, S.E. Stahl, K.H. Rubins *et al.,* 2017 Nanopore DNA sequencing and genome assembly on the International Space Station. Scientific Reports 7: 18022. http://doi.org/10.1038/s41598-017-18364-0

Dietrich, A., 2006 Aesthetic issues for drinking water. Journal of Water and Health 4(S1): 11-16.

Dunlap, C.R., K.S. Sklenar, and L.J. Blake, 2015 A costly endeavor: addressing algae problems in a water supply. Journal of the American Water Works Association 107(5): E255-E262.

Etheridge, S.M., 2010 Paralytic shellfish poisoning: Seafood safety and human health perspectives. Toxicon 56(2): 108-122.

Gamez, T., 2018 The use of 23S ribotyping to detect harmful and nuisance phytoplankton in a large, subtropical reservoir during an extended drought period (Unpublished thesis). Texas State University, San Marcos, Texas.

Gilbert, P.M., D.M. Anderson, P. Gentien, E. Granéli, and K.G. Sellner, 2005 The global, complex phenomena of harmful algal blooms. Oceanography 18(2): 136-147.

Gorham, T., E.D. Root, Y. Jia, C.K. Shum, and J. Lee, 2020. Relationship between cyanobacterial bloom impacted drinking water sources and hepatocellular carcinoma incidence rates. Harmful Algae 95: 101801.

Gowers, G.-O. F., O. Vince, J.-H. Charles, I. Klarenberg, T. Ellis, et al., 2019 Entirely off-grid and solar-powered DNA sequencing of microbial communities during an ice cap traverse expedition. Genes 10: 902. https://doi.org/10.3390/genes10110902

Griffith, A.W., and C.J. Gobler, 2020 Harmful algal blooms: a climate change co-stressor in marine and freshwater ecosystems. Harmful Algae 91: 101590.

Hatfield, R.G., F.M. Batista, T.P. Bean, V.G. Fonseca, A. Santos, et al., 2020 The application of nanopore sequencing technology to the study of dinoflagellates: a proof of concept study for rapid sequence-based discrimination of potentially harmful algae. Frontiers in Microbiology 11: 844.

Hennon, G.M.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.

Huerta-Cepas, J., D. Szklarczyk, D. Heller, A. Hernández-Plaza, S.K. Forslund, et al., 2019 eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research 47(D1): D309-D314. https://doi.org/10.1093/nar/gky1085

Hoagland, P., and S. Scatasta, 2006 The economic effects of harmful algal blooms, in *Ecology of Harmful Algae. Ecological Studies (Analysis and Synthesis) vol 189*, edited by E. Granéli and J.T. Turner. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-32210-8_30

Johnson, S.S., E. Zaikova, D.S. Goerlitz, Y. Bai, and S.W. Tighe, 2017 Real-time DNA sequencing in the Antarctic dry valleys using the Oxford Nanopore sequencer. Journal of Biomolecular Techniques 28(1): 2-7. https://doi.org/10.7171/jbt.17-2801-009.

Katoh, K., and D.M. Standley, 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Phylogenetics and Evolution 30(4): 772-780.

Kim, D., L. Song, F.P. Breitweiser, and S.L. Salzberg, 2016 Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Research 26(12): 1721-1729.

Krehenwinkel, H., M. Wolf, J.Y. Lim, A.J. Rominger, W.B. Simison, et al., 2017 Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific Reports 7: 17668. https://doi.org/10.1038/s41598-017-17333-x

Kolmogorov, M., J. Yuan, Y. Lin, and P.A. Pevzner, 2019 Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology 37**:** 540-546. https://doi.org/10.1038/s41587-019-0072-8

Legrand, B., J. Lesobre, J. Colombet, D. Latour, and M. Sabart, 2016 Molecular tools to detect anatoxin-a genes in aquatic ecosystems: Toward a new nested PCR-based method. Harmful Algae 58: 16-22.

Lepère, C., A. Wilmotte, and B. Meyer, 2000 Molecular diversity of *Microcystis* strains (Cyanophyceae, Chroococcales) based on 16S rDNA sequences. Systematics and Geography of Plants 70(2): 275-283

Li, H, 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18): 3094-3100. https://doi.org/10.1093/bioinformatics/bty191

Martí, J.M., and C.P. Garay, 2019 Not just BLAST nt: WGS database joins the party. *bioRxiv* 653592. https://doi.org/10.1101/653592 (Preprint posted June 4, 2019)

Menegon, M., C. Cantaloni, A. Rodriguez-Prieto, C. Centomo, A. Abdelfattah, et al., 2017 On site DNA barcoding by nanopore sequencing. PLOS ONE *12*(10): e0184741. https://doi.org/10.1371/journal.pone.0184741

Nguyen, L.-T., H.A. Schmidt, A. von Haeseler, and B.Q. Minh, 2015 IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Molecular Biology and Evolution 32: 268-274. https://doi.org/10.1093/molbev/msu300

Pfeiffer, F., C. Gröber, M. Blank, K. Händler, M. Beyer, *et al.,* 2018 Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific Reports 8: 10950. https://doi.org/10.1038/s41598-018-29325-6

Pomerantz, A., N. Peñafiel, A. Arteaga, L. Bustamante, F. Pichardo, et al., 2018 Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. Gigascience 7: 1-14.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, et al., 2013 The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41: D590-D596.

Rang, F.J., W.P. Kloosterman, and J. de Ridder, 2018 From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biology 19: 90. https://doi.org/10.1186/s13059-018-1462-9

Schafran, G.C, 2005 Reservoir Management Techniques to Improve Raw Water Quality. Algal Metabolytes Workshop, December 9, 2005, Sarasota, FL.

Seppey M., M. Manni, and E.M. Zdobnov, 2019 BUSCO: Assessing Genome Assembly and Annotation Completeness in *Gene Prediction. Methods in Molecular Biology vol 1962,* edited by M. Kollmar. Humana, New York, NY. doi.org/10.1007/978-1-4939-9173-0_14 .

Sherwood, A.R., and G.G. Presting, 2007 Universal primers amplify a 23S rDNA plastid marker in eukaryotic algae and cyanobacteria. Journal of Phycology 43(3): 605-608. https://doi.org/10.1111/j.1529-8817.2007.00341.x

Stern, R.F., R.A. Andersen, I. Jameson, F.C. Küpper, M.A. Coffroth, et al., 2012 Evaluating the ribosomal internal transcribed spacer (ITS) as a candidate dinoflagellate barcode marker. PLOS ONE *7*(8): e42780. https://doi.org/10.1371/journal.pone.0042780

Suurnäkki, S., G.V. Gomez-Saez, A. Rantala-Ylinen, J. Jokela, D.P. Fewer, et al., 2015 Identification of geosmin and 2-methylisoborneol in cyanobacteria and molecular detection methods for the producers of these compounds. Water Research 68: 56-66.

Tsao, H.-W., A. Michinaka, H.-K. Yen, S. Giglio, P. Hobson, et al., 2014 Monitoring of geosmin producing *Anabaena circinalis* using quantitative PCR. Water Research 49:416-425. http://dx.doi.org/10.1016/j.watres.2013.10.028.

Vaser, R., I. Sović, N. Nagarajan, and M. Šikić, 2017 Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research *27*: 737-746. https://doi.org/10.1101/gr.214270.116

Wacklin P., L. Hoffmann, and J. Komárek, 2009 Nomenclatural validation of the genetically revised cyanobacterial genus *Dolichospermum* (Ralfs ex Bornet et Flahault) comb. nova. Fottea 9(1): 59-64.

Wick, R.R., L.M. Judd, and K.E. Holt, 2019 Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biology 20: 129. https://doi.org/10.1186/s13059-019-1727-y

Wick R.R., M.B. Schultz, J. Zobel, and K.E. Holt, 2015. Bandage: interactive visualisation of *de novo* genome assemblies. Bioinformatics 31(20): 3350-3352.

Wongsurawat T., M. Nakagawa, O. Atiq, H.N. Coleman, P. Jenjaroenpun, et al., 2019 An assessment of Oxford Nanopore sequencing for human gut metagenome profiling: A pilot study of head and neck cancer patients. Journal of Microbiological Methods 166: 105739. https://doi.org/10.1016/j.mimet.2019.105739

Wood, D.E., J. Lu, and B. Langmead, 2019 Improved metagenomic analysis with Kraken2. Genome Biology 20: 257.

Zhang, Y., X. Lin, T. Li, H. Li, L. Lin, et al., 2020 High throughput sequencing of 18S rRNA and its gene to characterize a *Prorocentrum shikokuense* (Dinophyceae) bloom. Harmful Algae 94: 101809.

Zhu, W., A. Lomsadze, and M. Borodovsky, 2010 *Ab initio* gene identification in metagenomic sequences. Nucleic Acids Research 38(12): e132. https://doi.org/10.1093/nar/gkq275