J. Phycol. **58,** 146–160 (2022)
© 2021 Phycological Society of America
DOI: 10.1111/jpy.13221

AUREOCOCCUS ANOPHAGEFFERENS (PELAGOPHYCEAE) GENOMES IMPROVE EVALUATION OF NUTRIENT ACQUISITION STRATEGIES INVOLVED IN BROWN TIDE DYNAMICS[1]

Eric R. Gann, Alexander R. Truchon, Spiridon E. Papoulis

Department of Microbiology, University of Tennessee, Knoxville, Tennessee 37996, USA

Sonya T. Dyhrman

Biology and Paleo Environment Division, Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York 10964, USA

Department of Earth and Environmental Sciences, Columbia University, Palisades, New York 10964, USA

Christopher J. Gobler

School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York 11790, USA

and Steven W. Wilhelm[2]

Department of Microbiology, University of Tennessee, Knoxville, Tennessee 37996, USA

The pelagophyte *Aureococcus anophagefferens* causes harmful brown tide blooms in marine embayments on three continents. *Aureococcus anophagefferens* was the first harmful algal bloom species to have its genome sequenced, an advance that evidenced genes important for adaptation to environmental conditions that prevail during brown tides. To expand the genomic tools available for this species, genomes for four strains were assembled, including three newly sequenced strains and one assembled from publicly available data. These genomes ranged from 57.11 to 73.62 Mb, encoding 13,191–17,404 potential proteins. All strains shared ~90% of their encoded proteins as determined by homology searches and shared most functional orthologs as determined by KEGG, although each strain also possessed coding sequences with unique functions. Like the original reference genome, the genomes assembled in this study possessed genes hypothesized to be important in bloom proliferation, including genes involved in organic compound metabolism and growth at low light. Cross-strain informatics and culture experiments suggest that the utilization of purines is a potentially important source of organic nitrogen for brown tides. Analyses of metatranscriptomes from a brown tide event demonstrated that use of a single genome yielded a lower read mapping percentage (~30% of library reads) as compared to a database generated from all available genomes (~43%), suggesting novel information about bloom ecology can be gained from expanding genomic space. This work demonstrates the continued need to sequence ecologically relevant algae to understand the genomic potential and their ecology in the environment.

Key index words: **HABs; metatranscriptomes; organic nitrogen utilization; pan-genomes; xanthine**

Abbreviations**: AaV, Aureococcus anophagefferens Virus; CCMP, culture collection of marine phytoplankton; KEGG, Kyoto encyclopedia of genes and genomes**

Aureococcus anophagefferens causes harmful brown tide blooms, costing millions of dollars in losses due to high cell densities causing shading and potentially being toxic to bivalves (Gobler and Sunda 2012). These blooms were first detected in the northeast United States in the mid 1980s (Sieburth et al. 1988), but have since spread to distinct locations globally including Africa and China (Probyn et al. 2001, Zhang et al. 2012). Studies have shown that *A. anophagefferens* is physiologically well-adapted to the environmental conditions that are dominant during brown tides, specifically low light levels and limited availabilities of inorganic nutrients. These adaptions include an ability to assimilate both organic carbon (Dzurica et al. 1989, Lomas et al. 2001) and organic nitrogen (Lomas et al. 2001, Berg et al. 2002), to grow at low irradiance levels (Milligan and Cosper 1997), persist in complete darkness for prolonged periods of time (Popels et al. 2007) and form resting cysts (Ma et al. 2020). Many of these physiological studies have been conducted on different strains of *A. anophagefferens*, and it is known that differences between strains exist. As an example, some strains are susceptible to

[1] Received 25 March 2021. Accepted 2 November 2021.

[2] Author for correspondence: e-mail wilhelm@utk.edu.

Editorial Responsibility: M. Cock (Associate Editor)

infection by the isolated Aureococcus anophagefferens Virus (AaV), while others are not (Gobler et al. 2007, Brown and Bidle 2014). Some strains of *A. anophagefferens* are harmful to bivalves, while others are not (Bricelj et al. 2004, Harke et al. 2011).

Despite multiple strains of *Aureococcus anophagefferens* existing in culture for decades, and known differences existing physiologically, only a single strain, *A. anophagefferens* CCMP1984, has a publicly available genome to date (Gobler et al. 2011), although another strain, *A. anophagefferens* CCMP1794, has had its genome sequenced (Huff et al. 2016). The reference *A. anophagefferens* CCMP1984 genome encodes the genetic potential for utilization of various organic substrates, growth at low light and other potentially beneficial traits for competition during the blooms (Gobler et al. 2011), providing genomic support for the many physiological studies. Sequencing of *A. anophagefferens* has also provided relevant information into methylation patterns and transposon distributions within the genomes of the harmful bloom former (Huff and Zilberman, 2014, Huff et al. 2016). Besides insights into the genomic architecture, the reference genome provided a way of using sequencing data to understand the ecology of *A. anophagefferens*. Multiple studies have used the reference genome to map metatranscriptomic reads to help understand differences in gene expression over the course of a brown tide bloom (Wurch et al. 2019, Gann et al. 2021). Despite the reference genome improving our understanding of this organism, the succession of different strains of the same organism over the course of other harmful blooms has been shown to occur through several differing methods (Tarutani et al. 2000, Martinez et al. 2012, Park et al. 2014). Strain specificity may lead to different expression patterns inside and outside of a bloom (Liang et al. 2020). As is the case for many phytoplankton, the lack of annotated genomes outside of one or two references leads to an oversimplification of the very complex system that is an algal bloom (Ogura et al. 2018, Chen et al. 2019, Jackrel et al. 2019). Understanding the genomic diversity of algal strains holds the promise to reveal the genetic underpinnings of interclonal variation and ecological succession of strains in an ecosystem setting, as well as create a strong informatic database from which to study algal blooms.

The purpose of this study was to improve our understanding of the genetic potential of *Aureococcus anophagefferens* through the generation of new genomic assemblies from multiple strains. We sequenced and assembled genomes of three strains (*A. anophagefferens* CCMP1984, CCMP1707, and CCMP1850), assembled a genome of one strain (*A. anophagefferens* CCMP1794) from publicly available sequencing data (Huff et al. 2016), and re-annotated the original reference *A. anophagefferens* CCMP1984 genome to better compare differences between the strains. The genomes of the strains sequenced in this study using both long read and short read technologies had higher quality assemblies than the strain where public short read data were used. Even though ~90% of the proteins in each strain had a top BLASTp hit to the reference *A. anophagefferens* CCMP1984 strain, unique functions did exist in individual genomes. Finally, we used meta transcriptomic data from a 2016 brown tide bloom event to assess the informatic utility of the new pan-genomic data for providing insights into the ecology of *A. anophagefferens*.

MATERIALS AND METHODS

Culturing, DNA extractions, sequencing. Non-axenic *Aureococcus anophagefferens* strains CCMP1707, CCMP1850, and CCMP1984, were cultured in modified $ASP_{12}A$ (Gann, 2016), at 19°C with a 14:10 h light:dark cycle that included an irradiance level of 90 μmol photons $\cdot$ m^{-2} $\cdot$ s^{-1}. Cultures (1 L) were pelleted by centrifugation (5000*g*, 5 min) in a Sorvall Lynx 4000 Centrifuge (Thermo Fisher Scientific, Waltham, MA, USA) with a Fiberlite F14-14 × 50cy rotor (Thermo Fisher Scientific, Waltham, MA, USA). DNA extractions were performed using standard phenol-chloroform methods with ethanol precipitation (Sambrook 2001). Long reads were generated using Nanopore sequencing (Jain et al. 2016). Libraries generated using the ligation sequencing kit (Oxford Nanopore Technologies, Oxford, UK), were sequenced on a MinION Mk1B (Oxford Nanopore Technologies, Oxford, UK) with a R9.4.1 flow cell (Oxford Nanopore Technologies, Oxford, UK). Library preparation and short-read sequencing was conducted by the Microbial Genome Sequencing Center (Pittsburgh, PA, USA). Paired-end reads (2 × 150 bp) were generated using the NextSeq 500 system (Illumina, San Diego, CA, USA).

Assembly and gene prediction. For *Aureococcus anophagefferens* strains CCMP1707, CCMP1850, and CCMP1984, bases were called from Nanopore sequencing reads with Guppy version 4.0.15 + 56940742 using the configuration file dna_r9.4.1_450bps_fast.cfg (Wick et al. 2019). Nanopore reads were trimmed for adaptors using Porechop version 0.2.4 (Wick et al. 2017), and trimmed for quality (9) and length (500 bp) using NanoFilt version 2.7.1 (De Coster et al. 2018). Nanopore sequencing statistics were generated and visualized using NanoPlot version 1.33.1 (De Coster et al. 2018). Illumina reads were trimmed using default settings in CLC Genomics Workbench version 12.0 (Qiagen, Hilden, Germany). Genomes were assembled using Canu version 2.0 (Koren et al. 2017). Nanopore and Illumina reads were mapped to the contigs using Bowtie2 version 2.2.3 (Langmead and Salzberg 2012). Contigs were polished with the read mappings using Pilon version 1.23 (Walker et al. 2014). As short-read sequencing of *A. anophagefferens* CCMP1794 was performed previously (Huff et al. 2016), this information was accessed from the NCBI Sequence Read Archive (Accession: SRX2068919). Reads were trimmed using default settings in CLC Genomics Workbench version 12.0 (Qiagen, Hilden, Germany). All four strains were also assembled using SPAdes version 3.11.1 (Bankevich et al. 2012) using the Illumina read data. For the three strains where Nanopore data were present, the assemblies generated by Canu produced longer contigs and therefore were used for this analysis. The SPAdes assemblies did produce complete, circular mitochondria and chloroplast chromosomes, while the Canu assemblies did not, and therefore for the organelle chromosomes, these contigs were used. As *A. anophagefferens* is believed to be diploid (Huff et al. 2016), redundant or heterozygous contigs assembled due to heterogeneity in diploid genomes were removed using Redundans version 0.14a (Pryszcz and Gabaldón 2016) using default settings, with the trimmed Nanopore (if present) and Illumina reads. This pipeline clusters heterozygous contigs, keeping the longest of those clustered (Pryszcz and Gabaldón 2016). To assess bacterial contamination within the assemblies, contigs were queried against the *A. anophagefferens* CCMP1984 reference genome (accession: NZ_ACJI00000000.1; Gobler et al. 2011) using BLASTn (BLAST version 2.8.1+) (Camacho et al. 2009). Also, contigs were split into 500 bp segments and submitted to the Kaiju web server (Menzel et al. 2016) to predict taxonomic origin. Following previously established protocols (Hackl et al. 2020), contigs were considered bacterial in origin if > 50% of the segments within the contig were called bacterial by Kaiju. Mitochondria and chloroplast contigs (see below) were also removed from the assessment of the nuclear genome. Completeness of the nonbacterial contigs were assessed using BUSCO version 4.1.3 (Seppey et al. 2019), using the Stramenopile markers dataset. Coding sequences were called using the online web server for MAKER (Cantarel et al. 2008). To train the pipeline, proteins from the reference *A. anophagefferens* CCMP1984 genome were used (Gobler et al. 2011), as were assembled transcripts from a control time point from the infection cycle transcriptome performed previously (accession: SRR6627647; Moniruzzaman et al. 2018). Transcripts from this transcriptome were assembled in CLC Genomics Workbench version 12.0 (Qiagen, Hilden, Germany). Any contigs that did not include coding sequences were also not included in the final assemblies. Few contigs (< 5) in the three hybrid assemblies did not possess any coding sequences but were greater than 10 kb in length. It would be expected that with segments of DNA this length coding sequences would be present, which could indicate that these were other contaminating contigs that the MAKER pipeline could not call coding sequences on. Therefore, to be stringent, these contigs were removed. tRNAs were predicted using tRNA-scan-SE version 2.0.6 (Chan and Lowe 2019). To predict functions of the coding sequences, the translated amino acid sequences were uploaded to the eggNOG-mapper web server (Huerta-Cepas et al. 2017). All protein sequences from the reference *A. anophagefferens* CCMP1984 genome were also reannotated with the eggNOG-mapper web server. KEGG K numbers, COG categories, GO numbers, and names of proteins used in this study were those generated from eggNOG.

Chloroplast and mitochondria genomes were generated from the SPAdes assemblies. These were determined based on size and protein complement, as the reference *Aureococcus anophagefferens* CCMP1984 chloroplast (accession: NC_012898.1) and mitochondria (accession: MK922345) genomes have previously been sequenced and annotated (Ong et al. 2010, Liu et al. 2019). Translated SPAdes contigs were queried against mitochondria and chloroplast proteins using BLASTx (BLAST version 2.8.1+; Camacho et al. 2009). For each strain, complete, circular, contigs of the appropriate size (~42 kb for the mitochondria and ~89 kb for the chloroplast) with all expected proteins were present. Mitochondria and chloroplast chromosomes were annotated using the PROKKA annotation pipeline in Kbase (Seemann 2014, Arkin et al. 2018).

Comparing assemblies phylogenetically and their protein complements. To compare phylogenetic relationships of the four assemblies from this study and the reference *Aureococcus anophagefferens* CCMP1984 genome, the concatenated alignment of 12 shared single-copy orthologous genes were used. Orthologs were determined using BUSCO version 4.1.3 (Seppey et al. 2019), with the Stramenopile lineage dataset. Only orthologs shared between the five *A. anophagefferens* assemblies and the outgroup *Hondaea fermentalgiana* (accession: GCA_014084085.1), with a BUSCO score greater than 150 were used (Table S1 in the Supporting Information). Concatenated amino acid sequences were aligned using MAFFT version 7 (Katoh and Standley 2013), and trimmed for no gaps using trimAl version 1.3 (Capella-Gutierrez et al. 2009) in Phylemon 2.0 (Sanchez et al. 2011). A maximum likelihood phylogenetic tree was generated using PhyML version 3.0 (Guindon et al. 2010). To compare genomes at the protein level, all predicted proteins were queried against one another in an all versus all BLASTp (BLAST version 2.8.1+; Camacho et al. 2009). Only top BLASTp hits that had a query coverage >30% and an e-value $<1 \times 10^{-10}$ were considered for each genome. Clustering of genomes based on the presence/absence of distinct KEGG K numbers was performed using Bray–Curtis similarity in Primer version 7 (Clarke and Gorley 2015). Fisher's exact test was performed to determine enrichment/depletion of KEGG pathways found within only a subset of the genomes compared to the overall coding potential within all the genomes using R (R Core Team 2018). Assembled mitochondria and chloroplast chromosomes were aligned using mVISTA (Frazer et al. 2004).

Read mappings to 2016 brown tide metatranscriptomes. To assess how the new assemblies could improve the understanding of *Aureococcus anophagefferens* in the environment, metatranscriptomes from a 10-week sampling during the initiation, peak, and collapse of a 2016 brown tide bloom event in Quantuck Bay, New York, USA (Latitude = 40.81° N; Longitude = 72.62° W) were used (BioProject Number: PRJNA689205; Gann et al. 2021). Reads were trimmed for quality using default parameters in CLC Genomic Workbench version 12 (Qiagen, Hilden, Germany). All coding sequences from the five assemblies were clustered at a sequence identity threshold of 0.9 using CD-HIT-EST (Li and Godzik 2006; Appendix S1 in the Supporting Information). Reads were mapped to the clustered coding sequences, and coding sequences from individual genomes, using Bowtie2 (Langmead and Salzberg 2012). Using the clustered coding sequence mappings, specific pathways and genes of interest were searched for (Appendix S2 in the Supporting Information). If multiple coding sequences were present for the same function, the library and coding sequence normalized reads for each coding sequence were summed to provide normalized read mappings for that function. Pearson correlations between the library-normalized read mappings of each genome pair were performed in GraphPad Prism version 8 (GraphPad, San Diego, CA, USA).

Comparing expression of nitrogen transporters from a culture dataset. A previous study performed RNA sequencing on a cultured *Aureococcus anophagefferens* strain isolated in China grown in different conditions to assess nitrogen utilization (Dong et al. 2014). The reads generated were then mapped back to the reference CCMP1984 genome. To determine if trends gleaned from the 2016 brown tides' metatranscriptomes dataset about various nitrogen transporters were supported by culture experiments, expression data for the seven transcriptomes was downloaded from the NCBI Gene Expression Omnibus (GEO) database (Experiment Accession Number: GSE60576; Barrett et al. 2013). RPKM values for various nitrogen transporters from the reference CCMP1984 strain were pulled from those datasets and if multiple coding sequences were present for the same function, the RPKM values for each coding sequence were summed to provide normalized read mappings for that transporter type.

Assessing the ability of Aureococcus anophagefferens to grow on xanthine. Finally, given the importance of organic nitrogen for brown tide ecology and the predicted genetic capacity for *Aureococcus anophagefferens* to grow using purines (Wurch et al. 2014), culture experiments were performed to explore the ability of a non-axenic strain *A. anophagefferens* CCMP1984 to grow on xanthine. To remove carryover of nitrate from original cultures, cultures maintained in modified $ASP_{12}A$ (Gann 2016) were pelleted by centrifugation (5,000 *g*, 5 min) in a Sorvall Lynx 4000 Centrifuge (Thermo Fisher Scientific, Waltham, MA, USA) with a Fiberlite F14-14 x 50cy rotor (Thermo Fisher Scientific, Waltham, MA, USA). Cells were resuspended in modified $ASP_{12}A$ without a nitrogen source. Cells were then added to modified $ASP_{12}A$ + 10 nM NiCl hexahydrate with either xanthine, urea, or nitrate as the sole nitrogen source at concentrations of 0.0735 mM, 0.0735 mM, and 0.147 mM, respectively. All strains were transferred multiple times (>3) during mid-exponential growth to the respective nitrogen source to ensure any residual nitrate was removed during mid exponential phase. To begin the growth curve, mid exponential phase cultures in each nitrogen source were transferred to fresh growth medium in the respective nitrogen source with four biological replicates. Growth was measured for the entire progression of the growth curve. *Aureococcus anophagefferens* cell concentrations were determined via flow cytometry using a FACSCalibur flow cytometer (Becton, Dickinson and Company, Franklin Lakes, NJ, USA). Cells were gated on red fluorescence and forward scatter as described previously (Moniruzzaman et al. 2018). Doubling times were calculated using the following equation, where days eleven and two were $time_N$ and $time_{No}$, respectively:

$$Doubling\ time \frac{time_N - time_{No}}{(\log(cell\ concentration_N) - \log(cell\ concentration_{No}))/\log(2)}$$

A one-way ANOVA followed by Tukey's HSD post hoc testing was used to assess differences in growth rates performed in GraphPad Prism version 8 (GraphPad, San Diego, CA, USA).

RESULTS

Strains and assembly statistics. The four *Aureococcus anophagefferens* strains (CCMP1707, CCMP1794, CCMP1850, CCMP1984) used in this study were all isolated from the northeast United States (Fig. 1A, Table S2 in the Supporting Information), but in different years (Table S2). *Aureococcus anophagefferens* strains CCMP1707, CCMP1850, and CCMP1984 were sequenced by both Nanopore and Illumina sequencing technologies (Table 1). We took advantage of previous Illumina sequencing of *A. anophagefferens* strain CCMP1794 that had not been assembled into larger contigs and the reference *A. anophagefferens* CCMP1984 genome to compare the gene complement of multiple strains (Table 1; Gobler et al. 2011, Huff et al. 2016). The ˜56 Mb reference genome of *A. anophagefferens* CCMP1984 was sequenced using 454 pyrosequencing, and was assembled into >2,000 scaffolds, and >5,000 contigs. The hybrid (Nanopore and Illumina sequencing technologies) assemblies of all three cultured strains in this study provided better assemblies than the original reference *A. anophagefferens* CCMP1984 genome, producing genomes assembled into fewer contigs, with higher N50s, lower L50s, and producing larger contigs (Table 1). The assembly sizes for *A. anophagefferens* CCMP1707, CCMP1850, and CCMP1984 were 64.43 Mb, 57.11 Mb, and 73.62 Mb, respectively. All had a high GC content (˜70%) like the reference *A. anophagefferens* CCMP1984 genome (69.44%; Table 1). The three hybrid assemblies had 13,191, 15,302, and 17,404 coding sequences for *A. anophagefferens* CCMP1707, CCMP1850, and CCMP1984, respectively (Table 1). The number of tRNAs ranged from 67 to 86 (Table 1). To assess completeness of the genomes, single-copy maker orthologs for Stramenopiles were searched for within each of the genomes using BUSCO. The reference genome and the three strains assembled from both Illumina and Nanopore reads had a similar number of complete and fragmented single-copy orthologs (Between 64 and 72 out of 100 Stramenopile single-copy marker orthologs), while the *A. anophagefferens* CCMP1794 had fewer (40; Table 1). The number of these single copy orthologs can provide a relative comparison of completeness, but not a definitive number as more sequenced Pelagophyceae genomes would be required to define what the total coding potential is for the class. Complete chloroplast and mitochondria chromosomes were assembled for all four strains (Table S3 in the Supporting Information). All chloroplast chromosomes were >99% identical with one another (Fig. S1 in the Supporting Information), as were the mitochondria (Fig. S2 in the Supporting Information). Genome similarity is addressed below.

Comparison of the assemblies. As the five genomes from this study were sequenced and assembled in different ways and have differing amounts of completeness (Table 1), we decided for this initial analysis to only focus on the similarities and differences of the encoded proteins. Phylogenetic analysis of shared concatenated single-copy orthologs revealed the reference *Aureococcus anophagefferens* CCMP1984 genome and our re-sequenced *A. anophagefferens* CCMP1984 assembly clustered with one another, as expected, while *A. anophagefferens* CCMP1850 and CCMP1707 were most closely related with one another (Fig. 1B). To compare the protein complement within each strain, all amino acid sequences were queried against the NCBI nonredundant database as well as all of the other proteins from the *A. anophagefferens* genomes. For all genomes assembled in the study, ˜90% of the proteins had a top BLASTp hit to the reference *A. anophagefferens* CCMP1984 genome, ~5% had a top BLASTp hit to another eukaryote, ~3–5% had no hits in the nonredundant database (cutoff e-value $< 1 \times 10^{-10}$), and < 1% of the proteins had a top BLASTp hit to bacteria, archaea, or viruses (Table S4 and Appendix S3 in the Supporting Information). Comparing the proteins encoded in the genomes to all other assemblies showed the strains had many similar proteins (Table S5 and Fig. S3 in the Supporting Information). Excluding *A. anophagefferens* CCMP1794 due to its low completeness (Table 1), ~50% of the

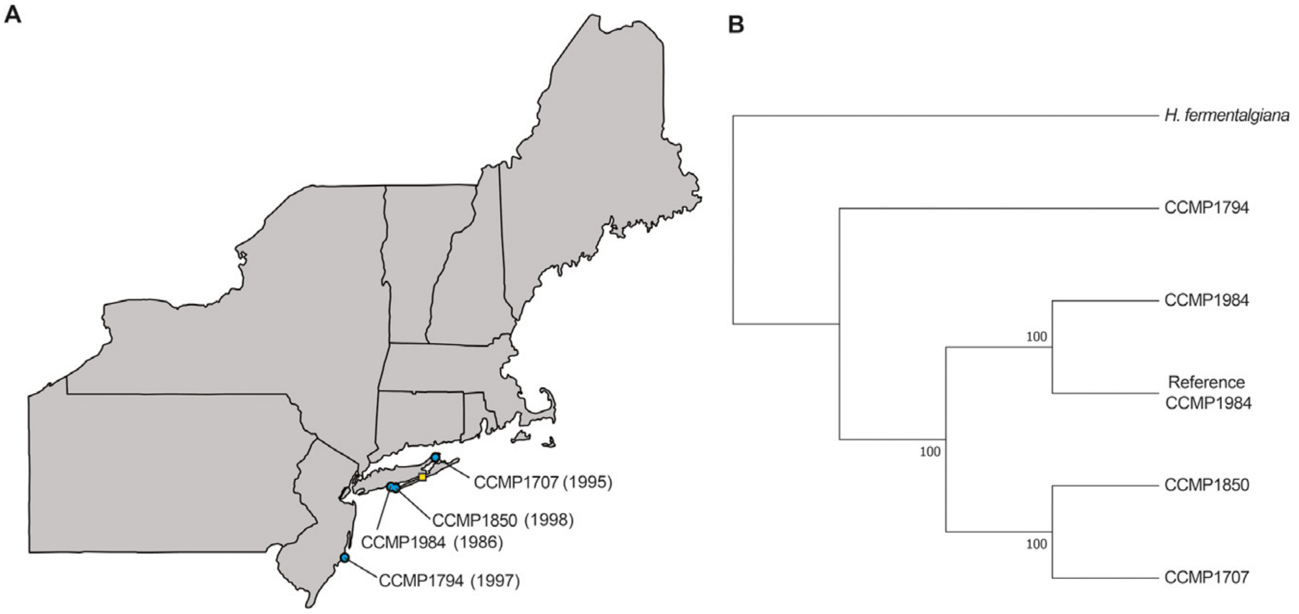


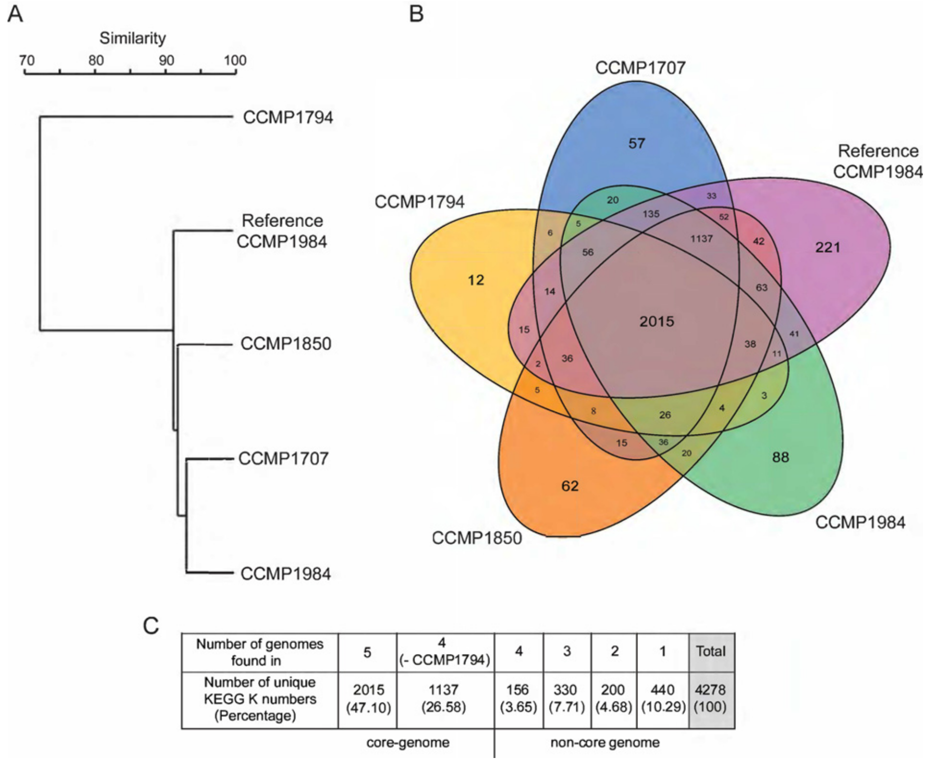
FIG. 1. Description of strains used in this study. (A) Locations where the four *Aureococcus anophagefferens* were isolated along with year of isolation. Strain isolation locations are denoted as blue circles. Quantuck Bay is denoted as an orange square. (B) Maximum likelihood tree of shared concatenated single-copy orthologs. Node support (aLRT-SH statistic) > 50% is shown.

TABLE 1. Nuclear Genome Assembly Statistics of all *Aureococcus* strains.

Strain	Reference CCMP1984	CCMP1794	CCMP1984	CCMP1850	CCMP1707
Sequencing Technology	454 pyrosequencing	Illumina	Nanopore + Illumina	Nanopore + Illumina	Nanopore + Illumina
Assembler	JAZZ	SPAdes	canu	canu	canu
Assembly Size	56.67 Mb	17.32 Mb	73.62 Mb	57.11 Mb	64.43 Mb
Contigs	5239	1815	215	212	149
N50 (Contigs)	33.74 Kb	9.86 Kb	522.78 Kb	483.16 Kb	844.79 Mb
L50 (Contigs)	2078	547	25	33	21
Largest Contig	277.37 Kb	66.74 Kb	8.12 Mb	3.50 Mb	4.50 Mb
%GC	69.44	71.79	70.39	69.80	70.18
Number of single-copy Stramenopile orthologs defined by BUSCO	Total: 72 Complete: 66 Fragmented: 6	Total: 40 Complete: 35 Fragmented: 5	Total: 67 Complete: 52 Fragmented: 15	Total: 64 Complete: 52 Fragmented: 12	Total: 67 Complete: 52 Fragmented: 15
Coding Sequences	11520	4993	17404	13191	15302
tRNAs	27	14	86	67	76
Reference	Gobler et al. (2011)	Huff et al. (2016)	This Study	This Study	This Study

proteins encoded in the assemblies were similar (e-value $< 1 \times 10^{-100}$). Seventy-five percent of the proteins in *A. anophagefferens* CCMP1794 were similar to the other strains (e-value $< 1 \times 10^{-100}$; Table S5, Fig. S3). Each strain had proteins that did not have a BLASTp hit to another strain. For *A. anophagefferens* CCMP1850 and CCMP1794, the number was very low: 85 (0.64% of encoded proteins) and 47 (0.94% of encoded proteins), respectively. For *A. anophagefferens* CCMP1707, CCMP1984, and the reference *A. anophagefferens* CCMP1984 genomes, the number of unique proteins was greater: 1,055 (6.89% of encoded proteins), 1,524 (8.76% of encoded proteins), and 1,693 (14.70% of encoded proteins), respectively (Appendix S4 in the Supporting Information).

Annotation of coding sequences and analysis of core versus non-core-genome functions. The encoded proteins from the genomes assembled in this study were annotated using eggNOG (Huerta-Cepas et al. 2017; Appendix S5, Table S6 in the Supporting Information). To directly compare the assemblies generated in this study to the reference *A. anophagefferens* CCMP1984 genome, the encoded proteins within the reference genome were also reannotated in the same way (Appendix S5, Table S6). Comparing the genomes based on COG categories (Table S7, Fig. S4 in the Supporting Information), or by KEGG categories (Table S8, Fig. S5, Appendix S6 in the Supporting Information) showed the proportions of categories/pathways for each genome were similar. To further compare similarities and differences, we focused on KEGG K numbers, which represent functional orthologs (Kanehisa et al. 2017), as ~50% of the coding sequences annotated could be assigned one (Table S6). We recognize that this biases our comparison to only known proteins but allows for a more comprehensive understanding of shared functionality. Specifically, we examined distinct KEGG K numbers found within each genome, generating 4278 KEGG K numbers as part of the pan-genome for this species (Fig. 2). Clustering the assemblies based on the presence/absence of the 4278 distinct KEGG K numbers (Fig. 2C), revealed all but *A. anophagefferens* CCMP1794 to be > 90% similar (Fig. 2 A). It is worth noting the reference *A. anophagefferens* CCMP1984 genome and the assembled *A. anophagefferens* CCMP1984 genome from this study did not cluster most closely with one another, but those sequenced and annotated from this study did (Fig. 2A). One potential reason for this, is the reference CCMP1984 genome had more distinct KEGG K numbers unique to its genome (221) than the other genomes (CCMP1984 – 88, CCMP1850 – 62, CCMP1794 – 12, CCMP1707 – 57; Appendix S6). Roughly half (47.10%) of the distinct KEGG K numbers were found within all genomes, while another 26.58% were found in all genomes excluding *A. anophagefferens* CCMP1794 (Fig. 2, B and C). As the *A. anophagefferens* CCMP1794 is not as complete as the other assemblies (Table 1), we considered KEGG K numbers found in all five genomes or the four more complete genomes to be the core-genome for this study (Fig. 2C). The remaining 26.33% of the distinct KEGG K numbers were considered not a part of the core-genome, with 10.29% (440 KEGG K numbers) only found in one of the five genomes (Fig. 2C). 80.59% to 89.32% of the distinct KEGG K numbers in each genome were those found in the core-genome (Table S9 in the Supporting Information). Between 0.53 and 5.65% of distinct KEGG K numbers within a genome were unique to that genome, specifically (Table S9). Although there were distinct K numbers and therefore functions found within each genome. The processes these were found in were similar including: K numbers pertaining to metabolism of various amino acid and nucleotide sugars and those pertaining to polyketide



Number of genomes found in	5	4 (- CCMP1794)	4	3	2	1	Total
Number of unique KEGG K numbers (Percentage)	2015 (47.10)	1137 (26.58)	156 (3.65)	330 (7.71)	200 (4.68)	440 (10.29)	4278 (100)

Fig. 2. Comparison of distinct KEGG K numbers found within all genomes. A) Hierarchical clustering of genomes based on the presence/absence of distinct KEGG K numbers using Bray–Curtis similarity. B) Venn-diagram of shared distinct KEGG K numbers in all five genomes. C) Description of distinct KEGG K numbers found in the assemblies.

and macrolide biosynthesis found in all genomes except *A. anophagefferens* CCMP1794. Also, unique K numbers pertaining to glycotransferases, and lectins were found in the genomes of *A. anophagefferens* CCMP1794 and the reference CCMP1984. To determine whether specific pathways/functions were enriched or depleted in a subset of genomes (non-core) of the species compared to the overall coding potential of the genomes, KEGG K numbers were clustered in categories/pathways (Table S8). Eight categories were significantly (Fisher's exact test, *P* value < 0.05) enriched in a subset of genomes including carbohydrate metabolism, nucleotide metabolism, metabolism of cofactors and vitamins, and metabolism of terpenoids and polyketides (Table S10 in the Supporting Information). Five categories were significantly (Fisher's exact test, *P* value < 0.05) depleted, including unclassified metabolism and amino acid metabolism (Table S10).

Comparison of the encoded gene complement with the reference genome. In the initial sequencing of *Aureococcus anophagefferens* CCMP1984, it was hypothesized that many of the proteins encoded in the genome allowed *A. anophagefferens* to outcompete other phytoplankton in the water column during the blooms (Gobler et al. 2011). This included encoding many proteins involved in light harvesting, uptake, and utilization of organic nitrogen and carbon, and numerous transporters. It appears this complement of genes is conserved in these other *A. anophagefferens* assemblies. *A. anophagefferens* CCMP1707, CCMP1850, CCMP1984, and the reference *A. anophagefferens* CCMP1984 possessed 77, 60, 88, and 63 light harvesting complex proteins, respectively (Appendix S7 in the Supporting Information). *Aureococcus anophagefferens* CCMP1794 had fewer light harvesting complex proteins (24; Appendix S7), which is unsurprising with its less complete genome (Table 1). *Aureococcus anophagefferens* is hypothesized to not be limited for nitrogen during blooms, unlike the rest of the community (Gobler et al. 2004), due to its ability to utilize organic nitrogen sources (Berg et al. 2002). The reference *A. anophagefferens* CCMP1984 genome was shown to encode proteins allowing the utilization of a wide range of organic nitrogen compounds (Gobler et al. 2011). This genetic potential for organic nitrogen utilization was found in all strains assembled in this study including enzymes (Table S11, Appendix S8 in the Supporting Information) and transporters (Table S12, Appendix S9 in the Supporting Information) required for the utilization of organic sources including urea, nucleotides, asparagine, and nitriles. Organic carbon utilization is also believed to be a competitive advantage for *A. anophagefferens* during the peak of blooms due to low light caused by high cell densities (Gobler and Sunda 2012). The strains assembled in this study contained a large number of polysaccharide-degrading enzymes (Table S13, Appendix S10 in the Supporting Information), and transporters for the uptake of various polysaccharide (Table S12, Appendix S9), as was reported for reference *A. anophagefferens* CCMP1984 genome (Gobler et al. 2011). These included enzymes for the utilization of simple sugars (i.e., xylose, glucose) and those that can break down more complex polysaccharides (i.e., pectin, heparan, cellulose, xylan; Table S13, Appendix S10).

Comparing genomes for assessment of environmental samples. To assess how the ecological understanding of brown tide blooms might be altered with these new genomes, a metatranscriptomic dataset from a 2016 brown tide bloom event at Quantuck Bay, NY was used (Fig. 1A). This dataset is composed of 18 metatranscriptomes from 10 weekly samples that followed the entire progression of the bloom (initiation, peak, decline; Table S14 in the Supporting Information). Reads were mapped to coding sequences of each genome assembled in this study, the reference genome, and all coding sequences from all strains clustered at a percent identity of 0.9 at the nucleotide level (Table S14). This clustered coding sequence dataset was used as a proxy for the species pan-genome. The reads mapped to each of the assemblies with similar completeness (*A. anophagefferens* CCMP1984, CCMP1707, CCMP1850, and the reference *A. anophagefferens* CCMP1984 genome; Table 1) all increased from ˜1.2% of the library reads during bloom initiation to ˜30% at the peak of the bloom, and then declined again (Fig. 3A). The less complete *A. anophagefferens* CCMP1794 assembly followed the same pattern but accounted for a smaller percentage of the library's reads mapped (ranged from 0.48 to 13.19%; Fig. 3A). Reads mapped to the pan-genome database began at a similar percentage of total library reads (1.78%) but increased to ˜43% of total library reads mapped at peak bloom (Fig. 3A). At peak bloom (6/27/2016; Fig. 3A) there were ˜12 million more reads mapped to the pan-genome database than to any of the other near complete genomes (Table S14). All the different assemblies and clustered coding sequences strongly correlated with one another ($r > 0.99$; Table S15 in the Supporting Information).

As there was an increased *Aureococcus anophagefferens* signal in the metatranscriptomes using the pan-genome database, we used those read mappings to assess the importance of purine/xanthine utilization by *A. anophagefferens* during this brown tide bloom. Purine metabolism has been suggested to be important during brown tide blooms based on the expression of purine transporters during growth on many nitrogen sources (Berg et al. 2008), and the observed overexpression of a xanthine permease during periods of nitrogen limitation (Wurch et al. 2014). As a proxy for nitrogen utilization, read mappings to xanthine transporters as well as other nitrogen sources were used (Appendix S2, Fig. 3B). Reads mapped to xanthine, ammonia and nitrate/

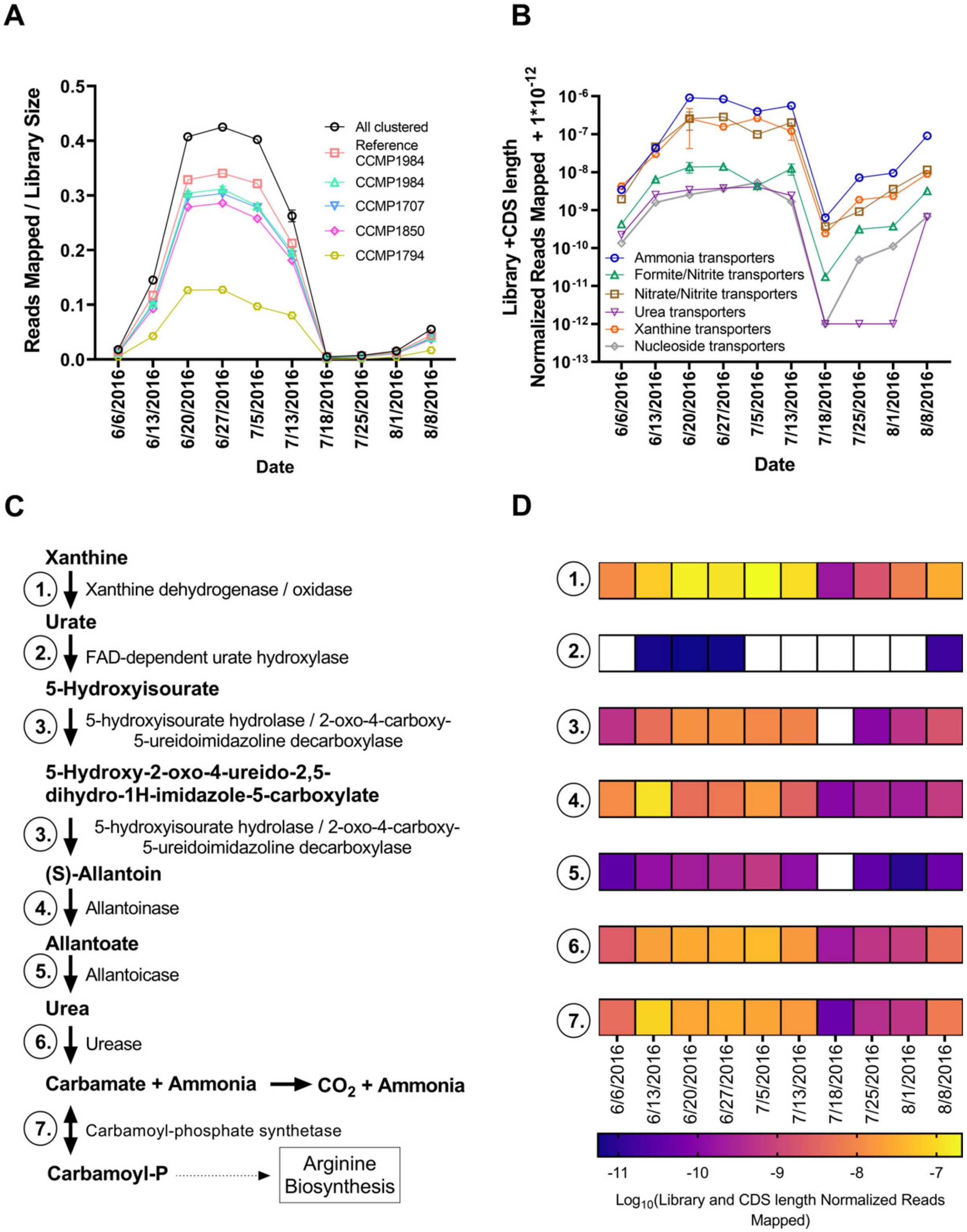


FIG. 3. Read mappings to *Aureococcus anophagefferens* genomes from a 2016 brown tide bloom event in Quantuck Bay, NY. (A) Library normalized reads mapped to coding sequences from each *A. anophagefferens* genome and all coding sequences clustered at a percent identity of 0.9. (B) Library and coding sequence length-normalized reads mapped to nitrogen transporters from all coding sequences clustered at a percent identity of 0.9. Mean values are shown for all points. Error bars are shown where multiple samples were taken. Error bars that did not extend past the data point were omitted. (C) Metabolic pathway by which xanthine is converted to ammonia based on KEGG pathway: map00230. (D) Heatmap of library and coding sequence normalized reads mapped to coding sequences involved in the metabolism of xanthine. Where multiple samples were taken on the same day, the mean of those samples plotted. White squares designate samples where no reads mapped to coding sequences.

nitrite transporters all increased approximately two orders of magnitude from the bloom initiation to the peak bloom and all had around the same number of normalized reads mapped, while reads mapped to formate/nitrite, nucleoside, and urea transporters only increased one order of magnitude as the bloom progressed and had over an order of magnitude fewer reads mapped to them (Fig. 3B). To provide more support for information gained from this analysis, we utilized gene expression data from a transcriptomics dataset of a Chinese strain of *A. anophagefferens* grown in various nitrogen conditions (Dong et al. 2014). As seen in the 2016 brown tide blooms metatranscriptomics dataset, formate/nitrite, nucleoside, and urea transporters had ˜1–2 order of magnitudes less relative expression compared to the other transporters (Table S16 in the Supporting Information). Xanthine transporter expression was similar to both the ammonia transporters and the nitrate/nitrite transporters with the exception of nitrate/nitrite transporters in cultures grown in replete nitrate (Table S16). Finally, it should be noted that xanthine transporters had the highest expression in nitrogen limiting conditions.

Xanthine is converted to ammonia through multiple enzymatic reactions including the final step of converting urea to ammonia (Fig. 3, B and C). To assess whether *Aureococcus anophagefferens* has the genetic potential to convert xanthine to ammonia, KEGG K numbers for each of the enzymatic reactions were searched within the genomes. Each enzyme in the pathway was identified in at least one of the genomes, except for allantoicase (Table S17 in the Supporting Information). Although there was not the KEGG K number for this enzyme, eggNOG predicted multiple coding sequences within the allantoicase family to be present in the genomes (Table S17), providing evidence that *A. anophagefferens* has the genetic potential to convert xanthine to ammonia. Transcripts were detected for all genes encoding enzymes in this pathway during the bloom, with transcripts for genes encoding the enzymes for the first step of converting xanthine to urate (xanthine dehydrogenase/oxidase), and the last step of converting urea to ammonia (urease) being the most abundant (Fig. 3C).

Experimentally it has been shown *Aureococcus anophagefferens* can incorporate the carbon from urea (Lomas et al. 2001): this can occur through either the fixation of respired carbon (CO_2) or potentially through the possible transformation of carbamate generated as an intermediate in urea degradation (Krausfeldt et al. 2019). To examine the latter, we also looked for expression patterns of carbamoyl phosphate synthetase (the enzyme that converts carbamate into carbamoyl phosphate). Carbamoyl phosphate can then be utilized in the biosynthesis of arginine (Fig. 3C). A similar number of normalized reads mapped to the carbamoyl phosphate synthetases during the peak bloom and followed the same pattern as ureases and xanthine dehydrogenases/oxidases (Fig. 3D). To provide evidence *A. anophagefferens* can grow on xanthine as a sole nitrogen source, non-axenic cultures were acclimated to urea, xanthine, and nitrate as sole nitrogen sources by multiple (>3) transfers. We used *A. anophagefferens* CCMP1984 to perform growth curves and calculate doubling times (Fig. S6 in the Supporting Information). There were no significant differences (P value > 0.05) in doubling times for cultures grown on xanthine (average: 1.025 days, SD = 0.033), urea (average = 1.065 days, SD = 0.016), or nitrate (average = 1.059 days, SD = 0.023; Fig. S6B).

DISCUSSION

Brown tides caused by *Aureococcus anophagefferens* cause millions of dollars in annual losses in distinct coastal locations across the globe (Gobler and Sunda 2012). To date, studies of the alga's physiology (e.g., Berg et al. 2002), the reference *A. anophagefferens* CCMP1984 genome (Gobler et al. 2011), and metatranscriptomes from natural blooms (Wurch et al. 2019) and cultures (Dong et al. 2014, Frischkorn et al. 2014), have helped identify the ecological niche of the causative agent of brown tide blooms. Although different strains have been used for physiological studies of this alga, *A. anophagefferens* CCMP1984 has been the only source of publicly available assembled genomic potential to date, despite existing Illumina sequencing of *A. anophagefferens* CCMP1794. Here, we assembled genomes of three strains of *A. anophagefferens* that have never been assembled publicly and resequenced the type strain, *A. anophagefferens* CCMP1984, and compared their coding potential to determine whether the genomes of other isolates might improve our understanding of *A. anophagefferens* physiology and the brown tides it generates. Assemblies generated from a combination of long and short reads improved on the sequencing of the reference genome performed a decade ago as these new assemblies were of similar sizes and completeness (as determined by BUSCO) but are composed of fewer, larger contigs. Improvement on the assembly of genomes using a combination of long and short reads for eukaryotic algae has been shown previously (Cecchin et al. 2019). We believe that having the long reads generated by Nanopore sequencing produced larger contigs as these could help resolve repeat regions and similar regions found within the genome. The reference *A. anophagefferens* genome has a high GC content (Gobler et al. 2011) and contains many repeat regions (Moniruzzaman et al. 2014) and many transposable elements surrounded by inverted repeats (Huff et al. 2016), all of which could make assembling the genome with just short reads challenging. Completeness of eukaryotic algal genomes, as determined by BUSCO, range from < 25% to few being >90%, with over one third being >75% (Hanschen and Starkenburg 2020). This range is also seen in sequenced Stramenopiles, ranging from 7 to >90% (Hackl et al. 2020, Labarre et al. 2021, Tan et al. 2021). It has been hypothesized that poor representation of specific organisms can account for lower than expected shared single-copy orthologs which has been seen in multiple cultured Stramenopiles (Hackl et al. 2020). Therefore, the completeness percentage of ~70% is in line with other eukaryotic algal species in general, as well as within Stramenopiles. More sequencing will be required to determine the genetic complement of this species.

All four *Aureococcus anophagefferens* strains used in this study were isolated in different years, each came from the Northeast United States, and encoded many similar proteins. The highly similar nature of the nuclear genomes was also seen in the organelles, as the chloroplast and mitochondria genomes were nearly identical to one another. This high sequence similarity has been reported for the mitochondria for multiple isolates of this species previously (Sibbald et al. 2021). Most of the coding sequences from assemblies generated in this study best BLASTp hits were to the reference *A. anophagefferens* CCMP1984 genome when compared to the nonredundant database, and for the four strains generated in this study, < 10% of the coding sequences were not found in another assembly. Lack of high BLAST hits to closely related Eukarya is due to lack of reference genomes of other *Pelagophyceae*. Continued sequencing of other algal organisms will improve databases for better definition (Tajima et al. 2016, Hamada et al. 2020). Using distinct KEGG K numbers as a proxy for functional orthologs found within the genomes, the majority of these were shared (>80%), while < 6% were unique to one genome. There was an enrichment in pathways and metabolisms that have been hypothesized to promote bloom proliferation including nucleotide metabolism, carbohydrate metabolism, and metabolism of cofactors and vitamins for K numbers found only in a subset of the genomes (Gobler et al. 2011). Also, K numbers unique to a single genome were found in similar pathways and metabolisms including nucleotide and amino acid metabolism as well as the biosynthesis of polyketides and macrolides. These pathways and functions found only in a subset of genomes may be an example of niche partitioning for certain resources which has been shown to occur in algae on both a phylum (Cheung et al. 2021) and strain-specific level (Majda et al. 2019). Unique strategies of nutrient uptake have also been seen in *Emiliania huxleyi.* The sequencing of multiple strains of *E. huxleyi*, which were isolated from distinct locations globally, demonstrated genes for specific types of nutrient uptake and metabolism were variable in number in the genomes (Read et al. 2013).

Despite both the reference and re-sequenced *Aureococcus anophagefferens* CCMP1984 assemblies being phylogenetically closest as determined from the concatenation of 12 single-copy orthologs, differences did exist. It is difficult to address questions about synteny, or genomic rearrangements, between the two reference strains for multiple reasons. Most importantly, these genomes were sequenced and assembled using completely different methods, as were the calling of the coding sequences. We believe this to be the driving reason for why, when clustering based on KEGG K numbers, CCMP1984 clusters away from the three genomes assembled in the same way. Some of the differences may also be due to biology and evolution of the strain in culture, but there unfortunately is no way of resolving this. The specific reference *A. anophagefferens* CCMP1984 strain from the study a decade ago was not cryopreserved, and therefore we cannot directly compare the current strain to the cryopreserved strain to disentangle what differences are caused by transferring of cultures and what are caused by improvements in informatic methods. Therefore, the re-sequencing should be regarded more as novel genomic information instead of a traditional resequencing effort. This is unfortunately a common problem for work with eukaryotic algae, where the strains may be in culture for decades, and for many there are no methods for cryopreservation. Although the domestication of strains is a problem, we have shown that at peak bloom nearly half of the metatranscriptome reads from 2016 mapped to these strains, suggesting they still are environmentally relevant. It should be noted that all the strains sequenced in this study have been in culture collections for over 20 years and were isolated from a similar geographical location, and therefore the similarities that are being seen might not be reflective of the diversity of this organism in nature. Given that brown tides of *A. anophagefferens* occur annually and in distinct parts of the globe, a continued effort to isolate and sequence new strains is required. This has occurred already in China and initial sequencing work has already been performed (Dong et al. 2014). Assessing differences between strains from the United States and other countries would likely allow for a more comprehensive understanding of this species.

Having more sequencing information allows us to understand not only the genetic potential of the strains sequenced but also provides more information to understand the ecology of *Aureococcus anophagefferens* using environmental sequencing datasets. Using metatranscriptomes capturing the entirety of a three-month brown tide bloom event in Quantuck Bay in 2016 revealed that clustering all coding sequences to generate a pan-genome database increased the number of reads mapped (>10 million more reads) at peak bloom. Although the genomes of these strains are very similar, there are coding sequences only found in one or two genomes, so using a single genome would not capture all of the coding potential that would exist in the species pan-genome database. Having more sequencing information will allow for a more comprehensive view of the *A. anophagefferens* dynamics and may provide new insights into bloom dynamics. Increasing available genomic information can also prove relevant in understanding ecosystem-wide nutrient cycling by way of both nutrient acquisition and incorporation (Nelson et al. 2019) and nutrient biosynthesis (McRose et al. 2014).

The ability of *Aureococcus anophagefferens* to use purines or other organic nitrogen compounds as a sole nitrogen source has been hypothesized to confer an advantage to *A. anophagefferens* during the blooms. This trait has been observed in other blooming algal species under stress (Shi et al. 2021). Physiological studies have shown that purine transporters are expressed during growth on many nitrogen sources in the laboratory, and that expression of the xanthine permease is a physiological diagnostic of nitrogen limitation of this species (Berg et al. 2008, Dong et al. 2014, Wurch et al. 2014). With the new genomic information presented here, we have identified the genes encoding for all steps in the conversion of xanthine to ammonia. In the laboratory, *A. anophagefferens* CCMP1984 can grow on xanthine as a sole nitrogen source, suggesting purines can be utilized readily by this species, although there does appear to be an increased lag time. Despite there not being a significant difference in doubling times, there was a larger distribution of doubling times for cultures grown on xanthine compared to the other two nitrogen sources, and therefore we are currently unable to conclusively state whether *A. anophagefferens* grows equally well with xanthine as the sole nitrogen source compared to urea and nitrate. The increased length of lag time between cultures suggest that xanthine may not be as accessible to the species as the other two nitrogen sources. Future work will be needed to conclusively determine how well this species can grow on this nitrogen source, but from this early data we can state *A. anophagefferens* can grow on this nitrogen source supporting the genomic information.

We used the new pan-genome database to gain insight into xanthine/purine utilization during a brown tide. As a proxy for nitrogen utilization in the blooms, the relative number of mapped reads to various inorganic and organic transporters were examined during the 2016 brown tide bloom event. The specific usage of metatranscriptomic reads for this purpose is common in HAB studies, as often genomics alone do not define bloom dynamics (Ji et al. 2020, Martin et al. 2020, Metegnier et al. 2020). Interestingly, reads mapped to transporters for xanthine, were as numerous as those for inorganic nitrogen sources (ammonia and nitrate/nitrite), and genes encoding the enzymes required for the multiple steps in this conversion had many reads mapped to them during the bloom. These trends were also seen in a transcriptomics dataset of a Chinese strain grown in different nitrogen conditions. These data highlight the potential importance of purines in addition to other organic nitrogen sources, like urea, during blooms, as xanthine is converted to ammonia in multiple steps, including the conversion of urea to ammonia. *Aureococcus anophagefferens* encodes multiple ureases, and it has been shown *A. anophagefferens* can grow on urea. These ureases are also constitutively expressed when grown on multiple nitrogen sources (Lomas et al. 2001, Fan et al. 2003). Although we cannot speculate on the relative importance of xanthine versus other nitrogen transporters based on the abundance of mapped reads during the bloom, the data are consistent with the importance of purines as a nitrogen source. Further study could determine the abundance of nucleotides during blooms and their relative importance as a source of nitrogen for *A. anophagefferens. Aureococcus anophagefferens* cultures are able to incorporate carbon from organic nitrogen sources (Lomas et al. 2001, Mulholland et al. 2002), which may supplement carbon requirements during peak bloom conditions when irradiance levels are low due to the high cell densities (Gobler and Sunda 2012). The enzyme to convert carbamate to carbamoyl phosphate, carbamoyl-phosphate synthetase, was expressed during the bloom, like other enzymes in the conversion of xanthine to ammonia. This carbamoyl phosphate can then be incorporated into arginine, potentially allowing *A. anophagefferens* to utilize the carbon from this pathway. This provides further evidence that the metabolism of carbon from organic nitrogen sources is occurring in natural blooms.

In conclusion, this study generated new genomes for the harmful brown tide bloom forming pelagophyte *Aureococcus anophagefferens* and provided novel insights into the diversity of coding potential in several strains, as well as the utilization of purines in brown tide blooms. Although sequenced strains were very similar, differences did exist, expanding our knowledge of the genetic potential of this species and the utilization of nitrogen during brown tides. The new pan-genome presented here will provide additional insight into the ecology of brown tides in the future.

The authors thank Gary LeCleir, Mohammad Moniruzzaman, Helena Pound, Caleb Schuler, and Robbie Martin for discussions about this work. This work was supported by a National Science Foundation grant (IOS1922958) and an award from the Simons Foundation (735077) to SWW, and by NOAA grants NA09NOS4780206 and NA15NOS4780199 (to STD) and NA15NOS4780183 (to CJG) through the ECOHAB Program (contribution number #1004). The authors have no conflict of interest to declare.

AUTHOR CONTRIBUTIONS

E. Gann: Conceptualization (equal); Data curation (lead); Formal analysis (equal); Investigation (lead); Methodology (lead); Writing-original draft (lead); Writing-review & editing (equal). **A. Truchon:** Data curation (supporting); Formal analysis (supporting); Methodology (supporting); Writing-review & editing (supporting). **S. Papoulis:** Formal analysis (equal); Methodology (equal); Software (equal); Writing-review & editing (equal). **S. Dyhrman:** Conceptualization (supporting); Data curation (equal); Funding acquisition (equal); Writing-review & editing (equal). **C.J. Gobler:** Conceptualization (supporting); Funding acquisition (equal); Project administration (supporting); Writing-review & editing (equal). **S. Wilhelm:** Conceptualization (equal); Formal analysis (supporting); Funding acquisition (lead); Investigation (supporting); Methodology (supporting); Project administration (lead); Writing-original draft (equal); Writing-review & editing (equal).

DATA AVAILABILITY

Raw sequencing data for the genomes were deposited in the Sequence Read Archive under the BioProject number PRJNA692237. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAFCAG000000000, JAFCAH000000000, and JAFCAI000000000 for *A. anophagefferens* CCMP1984, CCMP1850, and CCMP1707, respectively. The version described in this paper is version JAFCAG010000000, JAFCAH010000000, and JAFCAI010000000 for *Aureococcus anophagefferens* CCMP1984, CCMP1850, and CCMP1707, respectively. Scripts used for this study were deposited to GitHub (https://github.com/Wilhelmlab/Gann_2021_Aureococcus_genomes).

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., Dehal, P. et al. 2018. KBase: the United States Department of Energy systems biology knowledgebase. *Nat. Biotechnol* 36:566–9.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M. et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19:455–77.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A. et al. 2013. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41:D991–5.

Berg, G. M., Repeta, D. J. & Laroche, J. 2002. Dissolved organic nitrogen hydrolysis rates in axenic cultures of *Aureococcus anophagefferens* (Pelagophyceae): comparison with heterotrophic bacteria. *Appl. Environ. Microbiol.* 68:401–4.

Berg, G. M., Shrager, J., Glockner, G., Arrigo, K. R. & Grossman, A. R. 2008. Understanding nitrogen limitation in *Aureococcus anophagefferens* (Pelagophyceae) through cDNA and qRT-PCR analysis. *J. Phycol.* 44:1235–49.

Bricelj, V. M., MacQuarrie, S. P. & Smolowitz, R. 2004. Concentration-dependent effects of toxic and non-toxic isolates of the brown tide alga *Aureococcus anophagefferens* on growth of juvenile bivalves. *Mar. Ecol. Prog. Ser.* 282:101–14.

Brown, C. M. & Bidle, K. D. 2014. Attenuation of virus production at high multiplicities of infection in *Aureococcus anophagefferens*. *Virology* 466–467:71–81.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A. & Yandell, M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18:188–96.

Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–3.

Cecchin, M., Marcolungo, L., Rossato, M., Girolomoni, L., Cosentino, E., Cuine, S., Li-Beisson, Y., Delledonne, M. & Ballottari, M. 2019. *Chlorella vulgaris* genome assembly and annotation reveals the molecular basis for metabolic acclimation to high light conditions. *Plant J.* 100:1289–305.

Chan, P. P. & Lowe, T. M. 2019. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* 1962:1–14.

Chen, T., Xiao, J., Liu, Y., Song, S. & Li, C. 2019. Distribution and genetic diversity of the parasitic dinoflagellate *Amoebophrya* in coastal waters of China. *Harmful Algae* 89:101633.

Cheung, Y. Y., Cheung, S., Mak, J., Liu, K., Xia, X., Zhang, X., Yung, Y. & Liu, H. 2021. Distinct interaction effects of warming and anthropogenic input on diatoms and dinoflagellates in an urbanized estuarine ecosystem. *Glob. Chang. Biol.* 27:3463–73.

Clarke, K. R. & Gorley, R. N. 2015. *PRIMER v7: User Manual/Tutorial*. PRIMER-E, Plymouth.

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 34:2666–9.

Dong, H. P., Huang, K. X., Wang, H. L., Lu, S. H., Cen, J. Y. & Dong, Y. L. 2014. Understanding strategy of nitrate and urea assimilation in a Chinese strain of *Aureococcus anophagefferens* through RNA-seq analysis. *PLoS ONE* 9:e111069.

Dzurica, S. L. C., Cosper, E. M. & Carpenter, E. J. 1989. Role of environmental variables, specifically organic compounds and micronutrients, in the growth of the chrysophyte *Aureococcus anophagefferens*. *In* Cosper, E. M., Bricelj, V. M. & Carpenter, E. J. (Eds.) *Novel Phytoplankton Blooms*. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp. 229–52.

Fan, C., Glibert, P. M., Alexander, J. & Lomas, M. W. 2003. Characterization of urease activity in three marine phytoplankton species, *Aureococcus anophagefferens*, *Prorocentrum minimum*, and *Thalassiosira weissflogii*. *Mar. Biol.* 142:949–58.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32:W273–9.

Frischkorn, K. R., Harke, M. J., Gobler, C. J. & Dyhrman, S. T. 2014. *De novo* assembly of *Aureococcus anophagefferens* transcriptomes reveals diverse responses to the low nutrient and low light conditions present during blooms. *Front. Microbiol.* 5:375.

Gann, E. R. 2016. *ASP12A recipe for culturing Aureococcus anophagefferens*. https://www.protocols.io/view/asp12a-recipe-for-culturing-aureococcus-anophageff-f3ybqpw/forks Last accessed Nov 1, 2021.

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

Gobler, C. J., Anderson, O. R., Gastrich, M. D. & Wilhelm, S. W. 2007. Ecological aspects of viral infection and lysis in the harmful brown tide alga *Aureococcus anophagefferens*. *Aquat. Microb. Ecol.* 47:25–36.

Gobler, C. J., Berry, D. L., Dyhrman, S. T., Wilhelm, S. W., Salamov, A., Lobanov, A. V., Zhang, Y. et al. 2011. Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics. *Proc. Natl. Acad. Sci. USA* 108:4352–7.

Gobler, C. J., Boneillo, G. E., Debenham, C. J. & Caron, D. A. 2004. Nutrient limitation, organic matter cycling, and plankton dynamics during an *Aureococcus anophagefferens* bloom. *Aquat. Microb. Ecol.* 35:31–43.

Gobler, C. J. & Sunda, W. G. 2012. Ecosystem disruptive algal blooms of the brown tide species, *Aureococcus anophagefferens* and *Aureoumbra lagunensis*. *Harmful Algae* 14:36–45.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59:307–21.

Hackl, T., Martin, R., Barenhoff, K., Duponchel, S., Heider, D. & Fischer, M. G. 2020. Four high-quality draft genome assemblies of the marine heterotrophic nanoflagellate *Cafeteria roenbergensis*. *Sci. Data* 7:29.

Hamada, M., Satoh, N. & Khalturin, K. 2020. A reference genome from the symbiotic hydrozoan, *Hydra viridissma*. *G3* 10:3883–95.

Hanschen, E. R. & Starkenburg, S. R. 2020. The state of algal genome quality and diversity. *Algal Res.* 50:101968.

Harke, M. J., Gobler, C. J. & Shumway, S. E. 2011. Suspension feeding by the Atlantic slipper limpet (*Crepidula fornicata*) and the northern quahog (*Mercenaria mercenaria*) in the presence of cultured and wild populations of the harmful brown tide alga, *Aureococcus anophagefferens*. *Harmful Algae* 10:503–11.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C. & Bork, P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* 34:2115–22.

Huff, J. T. & Zilberman, D. 2014. Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryote. *Cell* 156:1286–97.

Huff, J. T., Zilberman, D. & Roy, S. W. 2016. Mechanism for DNA transposons to generate introns on genomic scales. *Nature* 538:533–6.

Jackrel, S. L., White, J. D., Evans, J. T., Buffin, K., Hayden, K., Sarnelle, O. & Denef, V. J. 2019. Genome evolution and host-microbiome shifts correspond with intraspecific niche divergence within harmful algal bloom-forming *Microcystis aeruginosa*. *Mol. Ecol.* 28:3994–4011.

Jain, M., Olsen, H. E., Paten, B. & Akeson, M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17:239.

Ji, N., Zhang, Z., Huang, J., Zhou, L., Deng, S., Shen, X. & Lin, S. 2020. Utilization of various forms of nitrogen and expression regulation of transporters in the harmful alga *Heterosigma akashiwo* (Raphidophyceae). *Harmful Algae* 92: 101770.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45:D353–D61.

Katoh, K. & Standley, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772–80.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. & Phillippy, A. M. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27:722–36.

Krausfeldt, L. E., Farmer, A. T., Castro Gonzalez, H. F., Zepernick, B. N., Campagna, S. R. & Wilhelm, S. W. 2019. Urea is both a carbon and nitrogen source for *Microcystis aeruginosa*: tracking ^{13}C incorporation at bloom pH conditions. *Front. Microbiol.* 10:1064.

Labarre, A., López-Escardó, D., Latorre, F., Leonard, G., Bucchini, F., Obiol, A., Cruaud, A. et al. 2021. Comparative genomics reveals new functional insights in uncultured MAST species. *ISME J.* 15:1767–81.

Langmead, B. & Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357–9.

Li, W. & Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–9.

Liang, D., Wang, X., Huo, Y., Wang, Y. & Li, S. 2020. Differences in the formation mechanism of giant colonies in two *Phaeocystis globosa* strains. *Int. J. Mol. Sci.* 21:5393.

Liu, F., Liu, S., Huang, T. & Chen, N. 2019. Construction and comparative analysis of mitochondrial genome in the brown tide forming alga *Aureococcus anophagefferens* (Pelagophyceae, Ochrophyta). *J. Appl. Phycol.* 32:441–50.

Lomas, M. W., Glibert, P. M., Clougherty, D. A., Huber, D. R., Jones, J., Alexander, J. & Haramoto, E. 2001. Elevated organic nutrient ratios associated with brown tide algal blooms of *Aureococcus anophagefferens* (Pelagophyceae). *J. Plankton Res.* 23:1339–44.

Ma, Z., Hu, Z., Deng, Y., Shang, L., Gobler, C. J. & Tang, Y. Z. 2020. Laboratory culture-based characterization of the resting stage cells of the brown tide causing Pelagophyte, *Aureococcus anophagefferens*. *J. Mar. Sci. Eng.* 8:1027.

Majda, S., Boenigk, J. & Beisser, D. 2019. Intraspecific variation in protists: clues for microevolution from *Poteriospumella lacustris* (Chrysophyceae). *Genome Biol. Evol.* 11:2492–504.

Martin, S. F., Doherty, M. K., Salvo-Chirnside, E., Tammireddy, S. R., Liu, J., Le Bihan, T. & Whitfield, P. D. 2020. Surviving starvation: proteomic and lipidomic profiling of nutrient deprivation in the smallest known free-living eukaryote. *Metabolites* 10:273.

Martinez, J. M., Schroeder, D. C. & Wilson, W. H. 2012. Dynamics and genotypic composition of *Emiliania huxleyi* and their co-occurring viruses during a coccolithophore bloom in the North Sea. *FEMS Microbiol. Ecol.* 81:315–23.

McRose, D., Guo, J., Monier, A., Sudek, S., Wilken, S., Yan, S., Mock, T., Archibald, J. M., Begley, T. P., Reyes-Prieto, A. & Worden, A. Z. 2014. Alternatives to vitamin B1 uptake revealed with discovery of riboswitches in multiple marine eukaryotic lineages. *ISME J.* 8:2517–29.

Menzel, P., Ng, K. L. & Krogh, A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* 7:11257.

Metegnier, G., Paulino, S., Ramond, P., Siano, R., Sourisseau, M., Destombe, C. & Le Gac, M. 2020. Species specific gene expression dynamics during harmful algal blooms. *Sci. Rep.* 10:6182.

Milligan, A. J. & Cosper, E. M. 1997. Growth and photosynthesis of the 'brown tide' microalga *Aureococcus anophagefferens* in subsaturating constant and fluctuating irradiance. *Mar. Ecol. Prog. Ser.* 153:67–75.

Moniruzzaman, M., Gann, E. R. & Wilhelm, S. W. 2018. Infection by a giant virus (AaV) induces widespread physiological reprogramming in *Aureococcus anophagefferens* CCMP1984 - a harmful bloom algae. *Front. Microbiol.* 9:752.

Moniruzzaman, M., LeCleir, G. R., Brown, C. M., Gobler, C. J., Bidle, K. D., Wilson, W. H. & Wilhelm, S. W. 2014. Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host–virus coevolution. *Virology* 466–467:60–70.

Mulholland, M. R., Gobler, C. J. & Lee, C. 2002. Peptide hydrolysis, amino acid oxidation, and nitrogen uptake in communities seasonally dominated by *Aureococcus anophagefferens*. *Limnol. Oceanogr.* 47:1094–108.

Nelson, D. R., Chaiboonchoe, A., Fu, W., Hazzouri, K. M., Huang, Z., Jaiswal, A., Daakour, S., Mystikou, A., Arnoux, M., Sultana, M. & Salehi-Ashtiani, K. 2019. Potential for heightened sulfur-metabolic capacity in coastal subtropical microalgae. *iScience* 11:450–65.

Ogura, A., Akizuki, Y., Imoda, H., Mineta, K., Gojobori, T. & Nagai, S. 2018. Comparative genome and transcriptome analysis of diatom, *Skeletonema costatum*, reveals evolution of genes for harmful algal bloom. *BMC Genom.* 19:765.

Ong, H. C., Wilhelm, S. W., Gobler, C. J., Bullerjahn, G., Jacobs, M. A., McKay, J., Sims, E. H. et al. 2010. Analysis of the complete chloroplast genome sequences of two members of the Pelagophyceae: *Aureococcus anophagefferens* CCMP1984 and *Aureoumbra lagunensis* CCMP15071. *J. Phycol.* 46:602–15.

Park, B. S., Wang, P., Kim, J. H., Kim, J. H., Gobler, C. J. & Han, M. S. 2014. Resolving the intra-specific succession within

Cochlodinium polykrikoides populations in southern Korean coastal waters via use of quantitative PCR assays. *Harmful Algae* 37:133–41.

Popels, L. C., MacIntyre, H. L., Warner, M. E., Zhang, Y. H. & Hutchins, D. A. 2007. Physiological responses during dark survival and recovery in *Aureococcus anophagefferens* (Pelagophyceae). *J. Phycol.* 43:32–42.

Probyn, T., Pitcher, G., Pienaar, R. & Nuzzi, R. 2001. Brown tides and mariculture in Saldanha Bay, South Africa. *Mar. Pollut. Bull.* 42:405–8.

Pryszcz, L. P. & Gabaldón, T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* 44: e113–e13.

R Core Team 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria Available online at https://www.R-project.org/

Read, B. A., Kegel, J., Klute, M. J., Kuo, A., Lefebvre, S. C., Maumus, F., Mayer, C. et al. 2013. Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature* 499:209–13.

Sambrook, J. 2001. *Molecular Cloning: A Laboratory Manual*, 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Sanchez, R., Serra, F., Tarraga, J., Medina, I., Carbonell, J., Pulido, L., de Maria, A. et al. 2011. Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic Acids Res.* 39:W470–4.

Seemann, T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–9.

Seppey, M., Manni, M. & Zdobnov, E. M. 2019. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* 1962:227–45.

Shi, X., Xiao, Y., Liu, L., Xie, Y., Ma, R. & Chen, J. 2021. Transcriptome responses of the dinoflagellate *Karenia mikimotoi* driven by nitrogen deficiency. *Harmful Algae* 103:101977.

Sibbald, S. J., Lawton, M. & Archibald, J. M. 2021. Mitochondrial genome evolution in pelagophyte algae. *Genome Biol. Evol.* 13:evab018.

Sieburth, J. M., Johnson, P. W. & Hargraves, P. E. 1988. Ultrastructure and ecology of *Aureococcus anophagefferens* gen. et sp. nov. (Chrysophyceae) - the dominant picoplankter during a bloom in Narragansett Bay, Rhode Island, summer 1985. *J. Phycol.* 24:416–25.

Tajima, N., Saitoh, K., Sato, S., Maruyama, F., Ichinomiya, M., Yoshikawa, S., Kurokawa, K., Ohta, H., Tabata, S., Kuwata, A. & Sato, N. 2016. Sequencing and analysis of the complete organellar genomes of Parmales, a closely related group to Bacillariophyta (diatoms). *Curr. Genet.* 62:887–96.

Tan, M. H., Loke, S., Croft, L. J., Gleason, F. H., Lange, L., Pilgaard, B. & Trevathan-Tackett, S. M. 2021. First genome of *Labyrinthula* sp., an opportunistic seagrass pathogen, reveals novel insight into marine protist phylogeny, ecology and CAZyme cell-wall degradation. *Microb. Ecol.* 82:498–511.

Tarutani, K., Nagasaki, K. & Yamaguchi, M. 2000. Viral impacts on total abundance and clonal composition of the harmful bloom-forming phytoplankton *Heterosigma akashiwo*. *Appl. Environ. Microb.* 66:4916–20.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K. & Earl, A. M. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 9:e112963.

Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genom.* 3:e000132.

Wick, R. R., Judd, L. M. & Holt, K. E. 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* 20:129.

Wurch, L. L., Alexander, H., Frischkorn, K. R., Haley, S. T., Gobler, C. J. & Dyhrman, S. T. 2019. Transcriptional shifts the role of nutrients in harmful brown tide dynamics. *Front. Microbiol.* 10:136.

Wurch, L. L., Gobler, C. J. & Dyhrman, S. T. 2014. Expression of a xanthine permease and phosphate transporter in cultures and field populations of the harmful alga *Aureococcus anophagefferens*: tracking nutritional deficiency during brown tides. *Environ. Microbiol.* 16:2444–57.

Zhang, Q. C., Qiu, L. M., Yu, R. C., Kong, F. Z., Wang, Y. F., Yan, T., Gobler, C. J. & Zhou, M. J. 2012. Emergence of brown tides caused by *Aureococcus anophagefferens* Hargraves *et* Sieburth in China. *Harmful Algae* 19:117–24.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Figure S1. Alignment of the reference *Aureococcus anophagefferens* CCMP1984 chloroplast with all assembled chloroplasts from this study.

Figure S2. Alignment of whole reference *Aureococcus anophagefferens* CCMP1984 mitochondria with all assembled mitochondria from this study.

Figure S3. Comparison of protein complement within strains from all v. all BLASTp results.

Figure S4. Percentage of each COG (clusters of orthologous groups) category as determined by eggNOG per strain.

Figure S5. Percentage of each KEGG pathway as determined by eggNOG per strain excluding BRITE proteins.

Figure S6. Growth of *Aureococcus anophagefferens* CCMP1984 on different nitrogen sources.

Table S1. BUSCO orthologs concatenated for phylogenetic analysis.

Table S2. Information of strains used in this study.

Table S3. Chloroplast and mitochondria assembly statistics.

Table S4. Domain of the top BLASTp hit for each protein when queried against the nr database.

Table S5. All v. all BLASTp e-values separated by query strain.

Table S6. Overview of eggNOG annotations for each strain.

Table S7. Number of COGs (clusters of orthologous groups) by categories by strain.

Table S8. Number of K numbers by categories by strain.

Table S9. Unique K numbers found in each genome separated by the number of genomes each K number is found in.

Table S10. Analysis of enrichment of grouped KEGG pathways found in only a subset of the genomes (non-core).

Table S11. Nitrogen metabolism genes found within the various strains.

Table S12. Transporters found within the various strains.

Table S13. Carbohydrate degrading enzymes found within the various.

Table S14. Number of reads mapped to coding sequences from each genome and all coding sequences clustered at a 0.9 percent identity from the 2016 Quantuck Bay dataset.

Table S15. Pearson's r values for comparing reads mapped to individual strains in the 2016 Quantuck Bay brown tide bloom metatranscriptome.

Table S16. Summed RPKM values from Dong et al. 2014 transcriptomic dataset of nitrogen transporters within the reference CCMP1984 genome.

Table S17. Xanthine metabolism coding sequences detected in the *Aureococcus* genomes.

Appendix S1. Coding sequences found within all coding sequences clustered at a percent identity of 0.9.

Appendix S2. List of coding sequences of interest that were searched for in the transcriptomic analyses.

Appendix S3. Top BLASTp hits for each encoded protein when queried against the nr database.

Appendix S4. List of coding sequences that are unique to that genome when compared to the other genomes in the study. BLASTp e value cutoff $< 1 \times 10^{-10}$.

Appendix S5. Description of all coding sequences annotated with eggNOG.

Appendix S6. Presence absence of all unique KEGG KO numbers within each genome.

Appendix S7. Subset of all coding sequences annotated with eggNOG that are light harvesting complex proteins.

Appendix S8. Subset of all coding sequences annotated with eggNOG that are transporters.

Appendix S9. Subset of all coding sequences annotated with eggNOG that are nitrogen metabolism genes.

Appendix S10. Subset of all coding sequences annotated with eggNOG that are carbon metabolism genes.