Received: 11 November 2024 | Accepted: 16 January 2025

DOI: 10.1111/jpy.13545

GENOME RESOURCE

Three reference genomes for freshwater diatom ecology and evolution

Wade R. Roberts | **Andrew J. Alverson**

Department of Biological Sciences, University of Arkansas, Fayetteville, Arkansas, USA

Correspondence
Wade R. Roberts, Department of Biological Sciences, University of Arkansas, 1 University of Arkansas, Fayetteville, AR 72701, USA.
Email: wader@uark.edu

Funding information
Division of Environmental Biology, Grant/Award Number: 1651087

Editor: N. Poulsen

Abstract

Diatoms are an important component of marine and freshwater ecosystems. Although the majority of described diatom species live in freshwater systems, genome sequencing efforts have focused primarily on marine species. Genomic resources for freshwater species have the potential to improve our understanding of diatom ecology and evolution, particularly in the context of major environmental shifts. We used long- and short-read sequencing platforms to assemble reference genomes for three freshwater diatom species, all in the order Thalalassiosirales, which are abundant in the plankton of oceans, lakes, reservoirs, and rivers worldwide. We targeted three species that cover the breadth of phylogenetic diversity in the cyclostephanoid clade of Thalassiosirales: *Cyclostephanos tholiformis* (JALLPB020000000), *Discostella pseudostelligera* (JALLBG020000000), and *Praestephanos triporus* (JALLAZ020000000). The reference genome for *D. pseudostelligera* was considerably smaller (39 Mb) than those of both *P. triporus* (73 Mb) and *C. tholiformis* (177 Mb). Long-read sequencing allowed for the assembly of scaffold-level genomes, including regions rich in repetitive DNA. Compared to short-read assemblies, long-read assemblies increased the contig N50 length as much as 37-fold and reduced the number of contigs by more than 88%. Transcriptome-guided annotation of the protein-coding genes identified between 10,000 and 12,000 genes. This work provides further demonstration of the value of long-read sequencing and provides novel genomic resources for understanding the ecology and evolution of freshwater diatoms.

KEYWORDS
Cyclostephanos, diatoms, *Discostella*, long-read sequencing, *Praestephanos*, salinity, Thalassiosirales

INTRODUCTION

Diatoms are ubiquitous across marine and freshwater ecosystems where they play key roles in global primary production and form the base of aquatic food webs (Armbrust, 2009; Smol & Stoermer, 2010). Approximately 70% of described diatom species exist in freshwaters (Nakov et al., 2019), yet most genome sequencing projects to date have focused on marine taxa. Available draft genomes of freshwater diatoms are useful for comparative genomics and gene identification but are often incomplete due to poor or missing coverage of repetitive DNA, which are the principal driver of genome size in diatoms (Galachyants et al., 2015; Maberly et al., 2021; Roberts et al., 2024). Long-read sequencing technologies have made it

Abbreviations: AED, annotation edit distance; TSA, transcriptome shotgun assembly; WGS, whole genome shotgun.

possible to produce reference-quality genomes that contain a full representation of an organism's DNA, including regulatory and repetitive regions. However, to date only two reference-quality genomes have been published for freshwater diatoms, and both are in the pennate clade (Suzuki et al., 2022; Zepernick et al., 2022). Reference genomes create new possibilities for comparative genomic studies of individual species (Armbrust et al., 2004), populations (Pinseel et al., 2023), or clades (Roberts et al., 2024).

Thalassiosirales are a lineage of polar centric diatoms that are common and abundant in marine and freshwater phytoplankton communities. They have long been used as experimental models (Guillard & Ryther, 1962; Schultz, 1971; Thamatrakoln & Hildebrand, 2007), and one species, *Cyclotella nana* (formerly *Thalassiosira pseudonana*), was the first diatom with a sequenced genome (Armbrust et al., 2004), establishing it as an important model species (Poulsen & Kröger, 2023). Although genomic resources for Thalassiosirales are steadily increasing, most work has focused on marine and euryhaline species (Figure 1; Roberts et al., 2020, Liu et al., 2023, Liu & Chen, 2024a, 2024b). We previously assembled draft genomes using Illumina short-read sequencing for three freshwater species: *Cyclostephanos tholiformis*, *Discostella pseudostelligera*, and *Praestephanos triporus* (Roberts et al., 2023). Although these assemblies contained most or all the expected genes, differences between the sequenced and estimated genome sizes revealed that genome coverage was incomplete due, most likely, to the misassembly of repetitive regions (Roberts et al., 2024). We have augmented previous assemblies with long-read nanopore sequencing and report scaffold-level, reference-quality genome assemblies for these three freshwater diatoms.

MATERIALS AND METHODS

Culturing, DNA extraction, and sequencing

Clonal cultures of *Cyclostephanos tholiformis* strain UTEX 3231 (AJA228-03), *Discostella pseudostelligera* strain UTEX 3233 (AJA232-27), and *Praestephanos triporus* strain AJA276-08 were established from phytoplankton samples collected from three river systems in the southeastern United States (Table 1; Roberts et al., 2023). The cultures are available through the UTEX Culture Collection of Algae at the University of Texas at Austin. We extracted DNA using a modified CTAB protocol (Doyle & Doyle, 1987; Roberts et al., 2020) and constructed short-read DNA libraries with the KAPA HyperPrep Kit (Roche) for 2×100 bp paired-end sequencing on the Illumina HiSeq 4000 or NovaSeq 6000 platforms. Raw reads were quality-filtered and trimmed of adapters with Ktrim (v1.5.1) (Sun, 2020), followed by

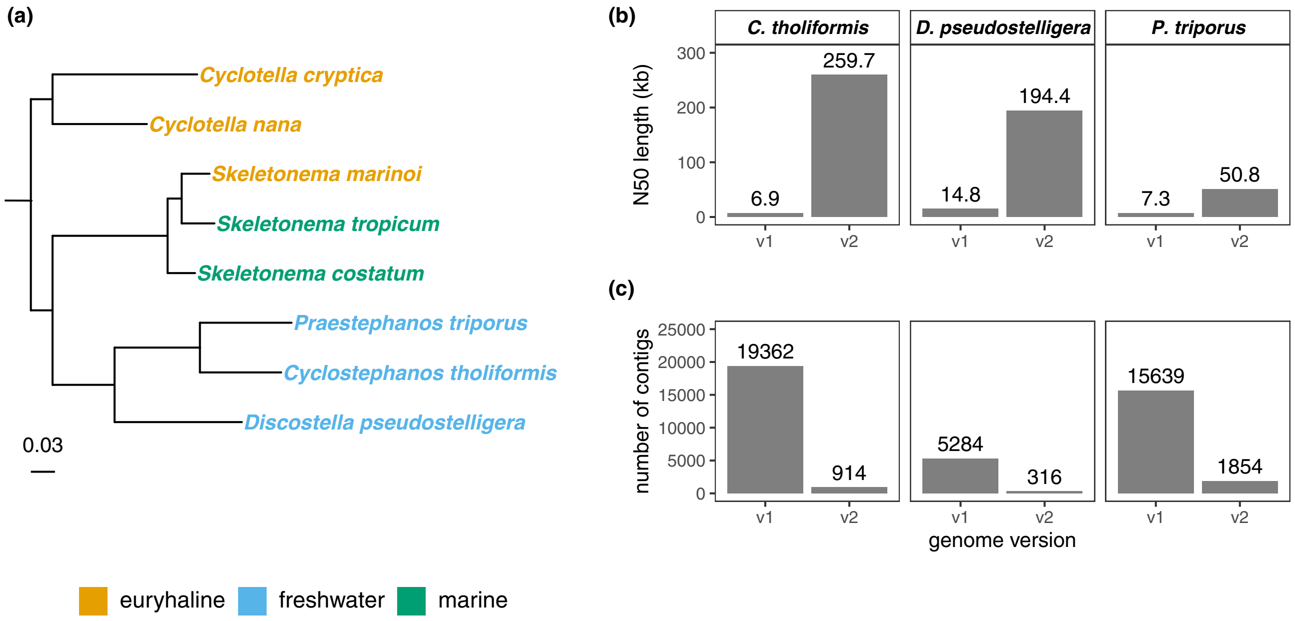

FIGURE 1 Long-read sequencing greatly improved genome assemblies for the freshwater diatoms *Cyclostephanos tholiformis*, *Discostella pseudostelligera*, and *Praestephanos triporus*. (a) Maximum likelihood phylogenetic analysis of Thalassiosirales reference genomes based on 100 concatenated single-copy orthologs. All branches received maximum support based on 10,000 ultrafast bootstrap replicates. The scale bar represents the expected number of substitutions per site. (b) Increased length of the contig N50 length (in kilobases) and (c) reduction in the number of contigs between the long-read (v2) genome assemblies versus the previous short-read (v1) assemblies.

TABLE 1 General features of the genome assemblies.

	Cyclostephanos tholiformis	*Discostella pseudostelligera*	*Praestephanos triporus*
Collection			
Strain	UTEX 3231 (AJA228-03)	UTEX 3233 (AJA232-27)	AJA276-08
Locality	Neosho River, Monkey Island, OK, USA	Mississippi River, St. Louis, MO, USA	Great Pee Dee River, Georgetown, SC, USA
Latitude, longitude	36.558683, −94.8384	38.622652, −90.183879	33.365217, −79.266267
Genome assembly			
Estimated haploid genome size (Mb)	190.4	39.4	85.6
Assembly size (Mb)	177.4	39.1	72.8
Sequencing depth, Illumina	84×	132×	45×
Sequencing depth, ONT	21×	10×	10×
Number of scaffolds	914	316	1854
Contig N50 (kb)	260	194	51
L50	208	58	449
GC content (%)	47.6	47.2	49.9
Repetitive DNA (%)	74.2	21.5	47.1
Stramenopile BUSCO completeness of assembly (%)	94	97	92
Eukaryota BUSCO completeness of assembly (%)	71	75	68
NCBI WGS Accession	JALLPB020000000	JALLBG020000000	JALLAZ020000000
Genome annotation			
Number of protein-coding genes	11,917	10,456	11,380
Average gene length (bp)	2089	2188	1915
Average exon length (bp)	708	861	694
Genes with annotation edit distance ≤0.5 (%)	98	98.4	96.2
Predicted proteins with Pfam or PANTHER domain (%)	70	71	67
BUSCO completeness of annotation (%)	94	95	88
OMArk completeness of annotation (%)	91.7	92.9	89.3

error correction with BayesHammer (part of SPAdes 3.14.1) (Nikolenko et al., 2013).

Additional CTAB DNA extractions that contained sufficient quantities (>3 μg) of high molecular weight DNA were used for long-read sequencing on the MinION platform (Oxford Nanopore Technologies, ONT). The quality and quantity of high molecular weight DNA was assessed using 0.8% agarose gel electrophoresis, Qubit 2.0 (dsDNA BR kit; Thermo Fisher Scientific), and Nanodrop 2000 (Thermo Fisher Scientific). Libraries were constructed with the Ligation Sequencing Kit (ONT) and sequenced on R9.4.1 flowcells (ONT). Raw MinION sequences were base-called with ONT's Guppy software (--trim_adapters -c dna_r.9.4.1_450bps_fast.cfg; v6.5.7); read quality statistics were calculated with NanoStat (v1.2.0; De Coster et al., 2018), and base-called reads were self-corrected with VeChat (v1.1.1; Luo et al., 2022).

Genome assembly

Default settings were used for all subsequent programs unless stated otherwise. Each stage of the assembly process was assessed for quality and completeness with QUAST (v5.0.0; Gurevich et al., 2013), BUSCO (stramenopiles_odb10 dataset; v5.1.3; Simão et al., 2015), and Merqury (v1.3; Rhie et al., 2020). Uncorrected long reads were assembled with metaFlye (--meta --iterations 0; v2.9.1-b1780; Kolmogorov et al., 2019, 2020) and corrected with Racon (-m 8 -x -6 -g -8 -w 500; v1.4.16; Vaser et al., 2017) and Medaka (-m r941_min_fast_g303; v1.2.3; https://github.com/nanoporetech/medaka). For *Cyclostephanos tholiformis* and *Discostella pseudostelligera*, the long-read assemblies were polished five times using POLCA (part of MaSuRCA v.4.1.1; Zimin & Salzberg, 2020) and the short Illumina reads to fix single nucleotide and small indel errors. Next, the contigs were scaffolded and gap-filled into longer contigs with the corrected long-reads using SAMBA (-d ont -m 3000 -f and -d ot -m 2000 -f; part of MaSuRCA v.4.1.1; Zimin & Salzberg, 2022). The scaffolds were then polished with JASPER (Guo et al., 2023). Following these steps, contaminant contigs belonging to bacteria were removed from the assemblies with EukRep (v0.6.6; West et al., 2018), and haplotigs were removed from the *C. tholiformis* assembly with Purge Haplotigs (Roach et al., 2018). For *Praestephanos triporus*, a hybrid assembly was performed using hybridSpades (--nanopore --trusted-contigs -k auto --cov-cutoff off; v3.14.1; Antipov et al., 2016) that combined the short Illumina reads, long nanopore reads, and the corrected long-read contigs. Following hybrid assembly, the short reads were used for scaffolding with SSPACE (v3.0; Boetzer et al., 2011), gap filling with GapCloser (v1.12; Luo et al., 2012), and polishing with POLCA. We then produced scaffolds with the corrected long reads using SAMBA and removed contaminant contigs from the assembly with EukRep. Finally, short contigs and remaining organellar contigs were removed from all three assemblies using BioKIT (remove_short_sequences; v0.0.9; Steenwyk et al., 2022).

Genome annotation and gene prediction

Repetitive sequences were identified with RepeatModeler2 (v.2.0.5; Flynn et al., 2020) and masked using RepeatMasker (-s -a -gff -norna -xsmall; v.4.1.5; https://www.repeatmasker.org/). After masking, gene models were predicted using the MAKER2 pipeline (v2.31.10; Holt & Yandell, 2011), which integrates evidence from transcripts and proteins homologs with trained *ab initio* gene prediction models. For transcript evidence, we used previously generated de novo assembled transcriptomes from the three strains (Roberts et al., 2023). Predicted proteomes from *Cyclotella cryptica* CCMP332 v2 (Roberts et al., 2020), *Cyclotella nana* CCMP1335 v4 (Filloramo et al., 2021), *Fragilariopsis cylindrus* CCMP1102 v1 (Mock et al., 2017), and *Phaeodactylum tricornutum* CCAP1055/1 v3 (Rastogi et al., 2018) were used as protein homolog evidence for gene annotation. We ran MAKER first, using only the transcript and protein evidence to generate initial gene models. Subsequent rounds of MAKER included trained ab initio gene prediction models from Augustus (Stanke et al., 2008), GeneMark-ES (Lomsadze et al., 2005), and SNAP (Korf, 2004). Augustus models were trained using the automated approach implemented in BUSCO with the eukaryota_odb10 dataset (--augustus --long). GeneMark-ES models were trained using a repeat soft-masked version of the genome assembly. SNAP models were trained using the initial gene models predicted by MAKER from the alignments of proteins and transcripts to the genome assembly. MAKER gene models were assessed using annotation edit distance (AED; Holt & Yandell, 2011), recovery of conserved orthologs with BUSCO (stramenopiles_odb10 and eukaryota_odb10 datasets; v.5.1.3; Simão et al., 2015), and proteome completeness with OMArk (LUMA.h5 database; v.0.3.0; Nevers et al., 2024). Annotation edit distance scores were calculated for each predicted gene based on how congruent the gene model is with the input evidence (proteins and transcripts; Holt & Yandell, 2011). Annotations are considered excellent when 90% of gene models have AED ≤0.5 and more than 50% of the proteins have a recognizable domain (Campbell et al., 2014). Single-exon gene overprediction can be a problem for poorly annotated genomes, so we also calculated the ratio of monoexonic:multiexonic genes using the output of gFACs (v.1.1.2; Caballero & Wegrzyn, 2019), as suggested by Vuruputoor et al. (2023).

To identify protein domains and gene ontology terms, we searched the predicted proteins against the Pfam (v.32.0; El-Gebali et al., 2019) and PANTHER (v.14.1; Mi et al., 2019) databases using InterProScan (-iprlookup -dp -goterms; v.5.36-75.0; Jones et al., 2014). We also searched the proteins against the SwissProt database (release 2024_06) using NCBI BLASTP (-evalue 1e-10 -num_alignments 1 -seg yes -soft_masking true -lcase_masking -max_hsps 1; v.2.16.0; Camacho et al., 2009) and the UniProt Reference Proteomes database (release 2020_04) using Diamond BLASTP (--evalue 1e-10 --max-target-seqs 1 --sensitive --max-hsps 1; v.2.1.9; Buchfink et al., 2015).

Genome size estimation

We estimated the haploid genome size of each strain using the kmer-based approach implemented in GenomeScope (v.2.0; Ranallo-Benavidez et al., 2020). Contaminant-free short reads were prepared for each strain by aligning the Illumina reads to the final assembly using minimap2 (commands: -ax sr; v.2.10; Li, 2018), sorting the alignments by coordinate using samtools (v.1.10; Li et al., 2009), and exporting only the aligned and properly paired reads using bam2fastq (--aligned --no-unaligned --no-filtered; https://github.com/jts/bam2fastq). K-mers of length 31 were then counted and summarized into a histogram with Jellyfish (v.2.3.0). Finally, we provided the histogram to GenomeScope (-k 31 -p 2) to estimate the haploid genome sizes.

Phylogenetic analysis

We used the 100 single-copy orthologs from the BUSCO stramenopiles_odb10 dataset to estimate a phylogenetic tree. BUSCO orthologs were extracted from the three genome assemblies and combined with corresponding orthologs from the reference genomes of *Cyclotella cryptica*, *Cyclotella nana*, *Skeletonema costatum* CNS00243 v1 (Liu & Chen, 2024a), *S. marinoi* CNS00100 v1 (Liu et al., 2023), and *S. tropicum* CNS00166 v1 (Liu & Chen, 2024b). Individual ortholog alignments were generated with MAFFT (--localpair --maxiterate 1000; v7.505; Katoh & Standley, 2013), trimmed with ClipKIT (-m smart-gap; v2.1.3; Steenwyk et al., 2020), and concatenated into a supermatrix with phyx (Brown et al., 2017). We estimated a maximum likelihood tree with IQ-TREE (v2.2.0.3; Minh et al., 2020), partitioning the alignment by ortholog, identifying the best-fit amino acid substitution model for each partition (-m MFP -mset JTT, WAG, LG, Q.pfam), and performing 10,000 ultrafast bootstrap replicates (-B 10000; Minh et al., 2013).

RESULTS AND DISCUSSION

We sequenced three libraries on the MinION platform and base-called between roughly 308,000 and 594,000 reads per species that yielded 1–4 Gb of raw data (Appendix S1: Table S1). The N50 read lengths were 9.5–15.8 kb, and average Phred scores of reads across runs had approximately 92% accuracy (Table S1). Previous Illumina sequencing yielded 3–10 Gb of raw reads per species (Table S1).

We assembled long-reads for *Cyclostephanos tholiformis* and *Discostella pseudostelligera* directly into draft genomes, but lower coverage of the *Praestephanos triporus* long-reads necessitated a hybrid assembly that combined the long and short Illumina reads (Appendix S1: Table S2). As expected with nanopore-based long-read sequencing, the initial assemblies for *C. tholiformis* and *D. pseudostelligera* were highly contiguous (contig N50 >150 kb) but had low consensus quality values (QV) of ~17 (~98% accuracy). After multiple rounds of polishing with the higher accuracy Illumina reads, QV for the final assemblies for *C. tholiformis* and *D. pseudostelligera* increased to 32, consistent with ~99.9% accuracy. Since the *P. triporus* genome was assembled in a hybrid manner, there was a tradeoff between slightly smaller contigs (contig N50 of 50.8 kb) but higher accuracy (QV of 44, or 99.99% accuracy). Despite these different approaches, the final genome assemblies for each species were highly complete with Stramenopile BUSCO scores all >92% (Table 1; Table S2), which is standard for high-quality diatom genome assemblies.

The inclusion of long nanopore reads dramatically improved previously published short-read draft assemblies (Roberts et al., 2023). For each species, the contig N50 lengths increased 7-fold (*Praestephanos triporus*), 13-fold (*Discostella pseudostelligera*), and >37-fold (*Cyclostephanos tholiformis*; Figure 1). These contig N50 values are similar to other diatom genomes sequenced on the nanopore sequencing platform (Audoor et al., 2024). Additionally, the inclusion of long-read sequencing reduced the number of contigs for each species genome by >88% (Figure 1). Although chromosome-level genome assemblies are becoming more common for diatoms (Filloramo et al., 2021; Liu & Chen, 2024a), our results demonstrate that long nanopore reads are sufficient to assemble complete or near-complete diatom genomes into a few hundred contigs. Although these are not yet chromosome-level assemblies, we detected telomeric sequence repeats (TTAGGG/CCCTAA; Fulnecková et al., 2013) at the beginning or end of five contigs in *C. tholiformis*, two in *D. pseudostelligera*, and five in *P. triporus*.

At more than 177 Mb in length, *Cyclostephanos tholiformis* had the largest genome assembly, while *Praestephanos triporus* was smaller at 73 Mb, and *Discostella pseudostelligera* was the smallest at 39 Mb (Table 1). These assembly sizes are consistent with the estimated haploid genome size for these species and strains (Table 1) and are also within the size range of other reference genomes published for the Thalassiosirales (*Cyclotella nana*, 33 Mb; *Cyclotella cryptica*, 171 Mb). The larger genomes contained proportionally more repetitive DNA (Table 1), further underscoring repeat content as the principal determinant of genome size in diatoms (Roberts et al., 2024). The large genome of *C. tholiformis* was composed of 74% repetitive DNA (Table 1), which is greater than the similarly sized *Cyclotella cryptica* genome (171 Mb) that had an estimated repetitive DNA content of 60% (Roberts et al., 2020). Unclassified repeats and long terminal retrotransposons were the largest contributors to the total repeat content in each genome (Appendix S1: Table S3). This trend is similar to that of other diatom genomes (Roberts et al., 2020) and confirms the need to understand how repeats and transposable elements contribute to diatom genomic diversity and evolution (Hermann et al., 2014).

Despite substantial differences in assembly lengths among *Cyclostephanos tholiformis*, *Discostella pseudostelligera*, and *Praestephanos triporus*, the three species contained similar numbers of predicted protein-coding genes (Table 1). All three genomes had 10–12k genes, similar to the expected range of 10–20 k genes in other Thalassiosirales genomes (Roberts et al., 2023, 2024). The number of genes was similar between the new assemblies and older draft assemblies (Appendix S1: Table S4), showing the utility of short-read-based genome skimming for phylogenomics (Roberts et al., 2023, 2024). The high percentage (>96%) of gene models with AED ≤0.5 and the large percentage of predicted proteins with an identifiable protein domain (>67%) all suggested high-quality annotations (Table 1). Additionally, the ratios of mono-exonic:multiexonic genes were 0.65, 0.71, and 0.57 for *C. tholiformis*, *D. pseudostelligera*, and *P. triporus*, respectively. These values are lower than other recently published diatom genomes (e.g., *Cyclotella nana*, 0.89; *Skeletonema marinoi*, 1.19), suggesting that our gene models were not likely to suffer from single exon gene overpredictions. Putative functional annotations were identified for roughly 90% of gene models after searching the predicted proteins against the SwissProt, UniProt Reference Proteomes, Pfam, and PANTHER databases (Table S4).

Marine–freshwater transitions have spawned enormous diversity across the tree of life (Jamy et al., 2022), including for diatoms (Nakov et al., 2019). A complete understanding of how diatoms and other microeukaryotes repeatedly have crossed this and other steep environmental gradients will benefit from genetic studies of model species (Poulsen & Kröger, 2023) and cross-species comparative genome studies. The scaffold-level reference genomes and gene models for *Cyclostephanos tholiformis*, *Discostella pseudostelligera*, and *Praestephanos triporus* represent valuable new resources for studying how diatoms scale the "invisible wall" that separates marine and freshwater environments (Bilcke & Kamakura, 2023).

AUTHOR CONTRIBUTIONS

Wade R. Roberts: Conceptualization (equal); data curation (lead); formal analysis (lead); methodology (lead); validation (lead); visualization (lead); writing – original draft (equal); writing – review and editing (equal). **Andrew J. Alverson:** Conceptualization (equal); funding acquisition (lead); resources (lead); writing – original draft (equal); writing – review and editing (equal).

ACKNOWLEDGMENTS

We thank Elizabeth Ruck and Kala Downey for culture care and laboratory assistance. This work was supported by the National Science Foundation (DEB 1651087). This research used resources available through the Arkansas High Performance Computing Center, which is funded through multiple NSF grants and the Arkansas Economic Development Commission. The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The Whole Genome Shotgun (WGS) projects are available from NCBI GenBank under accessions JALLPB020000000 (*Cyclostephanos tholiformis*), JALLBG020000000 (*Discostella pseudostelligera*), and JALLAZ020000000 (*Praestephanos triporus*). The Transcriptome Shotgun Assembly (TSA) projects are available from NCBI GenBank under the accessions GJWN00000000 (*C. tholiformis*), GKZL00000000 (*D. pseudostelligera*), and GKZM00000000 (*P. triporus*). The genome sequences, gene models, functional annotations, and code to reproduce Figure 1 have been deposited into Zenodo (https://doi.org/10.5281/zenodo.14628780).

ORCID

Wade R. Roberts https://orcid.org/0000-0002-5100-7558
Andrew J. Alverson https://orcid.org/0000-0003-1241-2654

REFERENCES

Antipov, D., Korobeynikov, A., McLean, J. S., & Pevzner, P. A. (2016). hybridSPAdes: An algorithm for hybrid assembly of short and long reads. *Bioinformatics*, *32*, 1009–1015.

Armbrust, E. V. (2009). The life of diatoms in the world's oceans. *Nature*, *459*, 185–192.

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., Bechner, M., Brzezinski, M. A., Chaal, B. K., Chiovitti, A., Davis, A. K., Demarest, M. S., Detter, J. C., Glavina, T., Goodstein, D., Hadi, M. Z., … Rokhsar, D. S. (2004). The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science*, *306*, 79–86.

Audoor, S., Bilcke, G., Pargana, K., Belišová, D., Thierens, S., Van Bel, M., Sterck, L., Rijsdijk, N., Annunziata, R., Ferrante, M. I., Vandepoele, K., & Vyverman, W. (2024). Transcriptional chronology reveals conserved genes involved in pennate diatom sexual reproduction. *Molecular Ecology*, *33*, e17320.

Bilcke, G., & Kamakura, S. (2023). Scaling the invisible wall: Molecular acclimation of a salinity-tolerant diatom to freshwater. *Molecular Ecology*, *32*, 2692–2694.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, *27*, 578–579.

Brown, J. W., Walker, J. F., & Smith, S. A. (2017). Phyx: Phylogenetic tools for Unix. *Bioinformatics*, *33*, 1886–1888.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*, 59–60.

Caballero, M., & Wegrzyn, J. (2019). GFACs: Gene filtering, analysis, and conversion to unify genome annotations across alignment and gene prediction frameworks. *Genomics, Proteomics & Bioinformatics*, *17*, 305–310.

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

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. *Current Protocols in Bioinformatics*, *48*, 4.11.1–4.11.39.

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*, 2666–2669.

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, *19*, 11–15.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*, D427–D432.

Filloramo, G. V., Curtis, B. A., Blanche, E., & Archibald, J. M. (2021). Re-examination of two diatom reference genomes using long-read sequencing. *BMC Genomics*, *22*, 379.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 9451–9457.

Fulnecková, J., Sevcíková, T., Fajkus, J., Lukesová, A., Lukes, M., Vlcek, C., Lang, B. F., Kim, E., Eliás, M., & Sykorová, E. (2013).

A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. *Genome Biology and Evolution*, *5*, 468–483.

Galachyants, Y. P., Zakharova, Y. R., Petrova, D. P., Morozov, A. A., Sidorov, I. A., Marchenkov, A. M., Logacheva, M. D., Markelov, M. L., Khabudaev, K. V., Likhoshway, Y. V., & Grachev, M. A. (2015). Sequencing of the complete genome of an araphid pennate diatom *Synedra acus* subsp. *radians* from Lake Baikal. *Doklady Biochemistry and Biophysics*, *461*, 84–88.

Guillard, R. R. L., & Ryther, J. H. (1962). Studies of marine planktonic diatoms: I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (Cleve) Gran. *Canadian Journal of Microbiology*, *8*, 229–240.

Guo, A., Salzberg, S. L., & Zimin, A. V. (2023). JASPER: A fast genome polishing tool that improves accuracy of genome assemblies. *PLoS Computational Biology*, *19*, e1011032.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*, 1072–1075.

Hermann, D., Egue, F., Tastard, E., Nguyen, D.-H., Casse, N., Caruso, A., Hiard, S., Marchand, J., Chénais, B., Morant-Manceau, A., & Rouault, J. D. (2014). An introduction to the vast world of transposable elements – What about the diatoms? *Diatom Research*, *29*, 91–104.

Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*, 491.

Jamy, M., Biwer, C., Vaulot, D., Obiol, A., Jing, H., Peura, S., Massana, R., & Burki, F. (2022). Global patterns and rates of habitat transitions across the eukaryotic tree of life. *Nature Ecology & Evolution*, *6*, 1458–1470.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*, 1236–1240.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772–780.

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T. P. L., & Pevzner, P. A. (2020). metaFlye: Scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, *17*, 1103–1110.

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*, 540–546.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*, 3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

Liu, S., & Chen, N. (2024a). Chromosome-level genome assembly of the cosmopolitan diatom *Skeletonema costatum* provides insights into ecological adaptation. *Algal Research*, *84*, 103761.

Liu, S., & Chen, N. (2024b). Chromosome-level genome assembly of marine diatom *Skeletonema tropicum*. *Scientific Data*, *11*, 403.

Liu, S., Xu, Q., & Chen, N. (2023). Expansion of photoreception-related gene families may drive ecological adaptation of the dominant diatom species *Skeletonema marinoi*. *Science of the Total Environment*, *897*, 165384.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., & Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, *33*, 6494–6506.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., … Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*, 18.

Luo, X., Kang, X., & Schönhuth, A. (2022). VeChat: Correcting errors in long reads using variation graphs. *Nature Communications*, *13*, 1–12.

Maberly, S. C., Gontero, B., Puppo, C., Villain, A., Severi, I., & Giordano, M. (2021). Inorganic carbon uptake in a freshwater diatom, *Asterionella formosa* (Bacillariophyceae): From ecology to genomics. *Phycologia*, *60*, 427–438.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, *47*, D419–D426.

Minh, B. Q., Nguyen, M. A. T., & von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution*, *30*, 1188–1195.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*, 1530–1534.

Mock, T., Otillar, R. P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., Ward, B. J., Allen, A. E., Dupont, C. L., Frickenhaus, S., Maumus, F., Veluchamy, A., Wu, T., Barry, K. W., Falciatore, A., Ferrante, M. I., … Grigoriev, I. V. (2017). Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature*, *541*, 536–540.

Nakov, T., Beaulieu, J. M., & Alverson, A. J. (2019). Diatoms diversify and turn over faster in freshwater than marine environments. *Evolution*, *73*, 2497–2511.

Nevers, Y., Warwick Vesztrocy, A., Rossier, V., Train, C.-M., Altenhoff, A., Dessimoz, C., & Glover, N. M. (2024). Quality assessment of gene repertoire annotations with OMArk. *Nature Biotechnology*, *1–10*, 124–133.

Nikolenko, S. I., Korobeynikov, A. I., & Alekseyev, M. A. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, *14*(Suppl 1), S7.

Pinseel, E., Ruck, E. C., Nakov, T., Jonsson, P. R., Kourtchenko, O., Kremp, A., Pinder, M. I. M., Roberts, W. R., Sjöqvist, C., Töpel, M., Godhe, A., Hahn, M. W., & Alverson, A. J. (2023). *Local adaptation of a marine diatom is governed by genome-wide changes in diverse metabolic processes*. bioRxiv. https://doi.org/10.1101/2023.09.22.559080

Poulsen, N., & Kröger, N. (2023). *Thalassiosira pseudonana* (*Cyclotella nana*) (Hustedt) Hasle et Heimdal (Bacillariophyceae): A genetically tractable model organism for studying diatom biology, including biological silica formation. *Journal of Phycology*, *59*, 809–817.

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*, 1432.

Rastogi, A., Maheswari, U., Dorrell, R. G., Vieira, F. R. J., Maumus, F., Kustka, A., McCarthy, J., Allen, A. E., Kersey, P., Bowler, C., & Tirichine, L. (2018). Integrative analysis of large scale transcriptome data draws a comprehensive landscape of *Phaeodactylum tricornutum* genome and evolutionary origin of diatoms. *Scientific Reports*, *8*, 4834.

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, *21*, 245.

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*, 460.

Roberts, W. R., Downey, K. M., Ruck, E. C., Traller, J. C., & Alverson, A. J. (2020). Improved reference genome for *Cyclotella cryptica*

CCMP332, a model for cell wall morphogenesis, salinity adaptation, and lipid production in diatoms (Bacillariophyta). *G3: Genes, Genomes, Genetics*, *10*, 2965–2974.

Roberts, W. R., Ruck, E. C., Downey, K. M., Pinseel, E., & Alverson, A. J. (2023). Resolving marine-freshwater transitions by diatoms through a fog of gene tree discordance. *Systematic Biology*, *72*, 987–997.

Roberts, W. R., Siepielski, A. M., & Alverson, A. J. (2024). Diatom abundance in the polar oceans is predicted by genome size. *PLoS Biology*, *22*, e3002733.

Schultz, M. E. (1971). Salinity-related polymorphism in the brackish-water diatom *Cyclotella cryptica*. *Canadian Journal of Botany*, *49*, 1285–1289.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*, 3210–3212.

Smol, J. P., & Stoermer, E. F. (2010). *The diatoms: Applications for the environmental and earth sciences*. Cambridge University Press.

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*, 637–644.

Steenwyk, J. L., Buida, T. J., 3rd, Li, Y., Shen, X.-X., & Rokas, A. (2020). ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biology*, *18*, e3001007.

Steenwyk, J. L., Buida, T. J., III, Gonçalves, C., Goltz, D. C., Morales, G., Mead, M. E., LaBella, A. L., Chavez, C. M., Schmitz, J. E., Hadjifrangiskou, M., Li, Y., & Rokas, A. (2022). BioKIT: A versatile toolkit for processing and analyzing diverse types of sequence data. *Genetics*, *221*, iyac079.

Sun, K. (2020). Ktrim: An extra-fast and accurate adapter- and quality-trimmer for sequencing data. *Bioinformatics*, *36*, 3561–3562.

Suzuki, S., Ota, S., Yamagishi, T., Tuji, A., Yamaguchi, H., & Kawachi, M. (2022). Rapid transcriptomic and physiological changes in the freshwater pennate diatom *Mayamaea pseudoterrestris* in response to copper exposure. *DNA Research*, *29*, dsac037.

Thamatrakoln, K., & Hildebrand, M. (2007). Analysis of *Thalassiosira pseudonana* silicon transporters indicates distinct regulatory levels and transport activity through the cell cycle. *Eukaryotic Cell*, *6*, 271–279.

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*, 737–746.

Vuruputoor, V. S., Monyak, D., Fetter, K. C., Webster, C., Bhattarai, A., Shrestha, B., Zaman, S., Bennett, J., McEvoy, S. L., Caballero, M., & Wegrzyn, J. L. (2023). Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes. *Applications in Plant Sciences*, *11*, e11533.

West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C., & Banfield, J. F. (2018). Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Research*, *28*, 569–580.

Zepernick, B. N., Truchon, A. R., Gann, E. R., & Wilhelm, S. W. (2022). Draft genome sequence of the freshwater diatom *Fragilaria crotonensis* SAG 28.96. *Microbiology Resource Announcements*, *11*, e0028922.

Zimin, A. V., & Salzberg, S. L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Computational Biology*, *16*, e1007981.

Zimin, A. V., & Salzberg, S. L. (2022). The SAMBA tool uses long reads to improve the contiguity of genome assemblies. *PLoS Computational Biology*, *18*, e1009860.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Detailed results for sequencing output, genome assembly, repeat classification, and comparison to previous genome versions.

How to cite this article: Roberts, W. R., & Alverson, A. J. (2025). Three reference genomes for freshwater diatom ecology and evolution. *Journal of Phycology*, *00*, 1–8. https://doi.org/10.1111/jpy.13545