



OPEN

DATA DESCRIPTOR

Annotation of protein-coding genes in 49 diatom genomes from the Bacillariophyta clade

Natalia Nenasheva^{1,4}, Clara Pitzschel^{1,4}, Cynthia N. Webster², Alexander J. Hart², Jill L. Wegrzyn², Mia M. Bengtsson³ & Katharina J. Hoff¹✉

Diatoms, a major group of microalgae, play a critical role in global carbon cycling and primary production. Despite their ecological significance, comprehensive genomic resources for diatoms are limited. To address this, we have annotated previously unannotated genome assemblies of 49 diatom species. Genome assemblies were obtained from NCBI Datasets and processed for repeat elements using RepeatModeler2 and RepeatMasker. For gene prediction, BRAKER2 was employed in the absence of transcriptomic data, while BRAKER3 was utilised when transcriptome short read data were available from the Sequence Read Archive. The quality of genome assemblies and predicted protein sets was evaluated using BUSCO, ensuring high-quality genomic resources. Functional annotation was performed using EnTAP, providing insights into the biological roles of the predicted proteins. Our study enhances the genomic toolkit available for diatoms, facilitating future research in diatom biology, ecology, and evolution.

Background & Summary

Diatoms are a diverse group of algae that significantly contribute to global carbon fixation and marine and freshwater ecosystem function¹. In addition to their ecological role, their ability to tolerate and quickly acclimate to rapidly changing environmental conditions is remarkable². These photosynthetic microalgae may capture and transmit CO₂ into diverse compounds, including lipids, omega-3 fatty acids, pigments, antioxidants, and polysaccharides³. They produce a variety of phytosterols, which offer possible health benefits such as cholesterol-lowering properties⁴. Diatoms can be cultivated indoors and outdoors, and their biomass productivity can be doubled in high-technology photobioreactors. A few selected species are used as model organisms in genetics and biochemistry research, while several taxa could be a bioprocess platform for biofuels³.

Diatoms play a critical role in the global carbon cycle^{3,5,6}. Through photosynthesis, diatoms convert carbon dioxide into organic carbon, forming the basis of marine food webs and assisting in the sequestration of carbon in ocean sediments⁶. Diatoms fix atmospheric carbon dioxide, accounting for around 20% of the world's primary production⁷. Their silica-based cell walls contribute to long-term carbon storage as they cause diatom cells to sink and settle on the ocean floor or the bottom of lakes and rivers. This process may be especially important during diatom blooms, which characterise temperate ocean margin zones and freshwater bodies in the spring. Various environmental factors in interactions with marine ecosystems affect the onset and progression of blooms, such as temperature, light intensity, and fluctuations of nutrients^{8,9}.

Interaction and coexistence with bacterial communities are an integral part of the life of diatom algae. They also form consortia and heterogeneous cohorts building networks of numerous cell-to-cell interactions for e.g. nutrient exchange. In this mutually beneficial deal, bacteria contribute by assimilating nutrients from the water and sequester minerals released by diatoms efficiently. Further, bacteria supply nutrients that diatoms are not able to produce themselves, for example, vitamins and fixed nitrogen¹⁰. Additionally, diatom blooms influence bacterial communities, showcasing their interconnectedness in marine ecosystems (e.g.^{11–13}). At the same time,

¹University of Greifswald, Institute of Mathematics and Computer Science and Center for Functional Genomics of Microbes, Walther-Rathenau-Str. 47, 17489, Greifswald, Germany. ²University of Connecticut, Department of Ecology and Evolutionary Biology, Plant Computational Genomics Lab, 75 N. Eagleville Road, Unit 3043, Storrs, CT, 06269-3043, USA. ³University of Greifswald, Institute of Microbiology, Felix-Hausdorff-Straße 8, 17489, Greifswald, Germany. ⁴These authors contributed equally: Natalia Nenasheva, Clara Pitzschel. ✉e-mail: katharina.hoff@uni-greifswald.de

Species name	Literature Reference(s)	Data Reference(s)	Species name	Literature Reference(s)	Data Reference(s)
<i>Asterionella formosa</i>	unknown	75,76	<i>Nitzschia inconspicua</i>	77	78
<i>Asterionellopsis glacialis</i>	79	80	<i>Nitzschia palea</i>	unknown	81–87
<i>Bacterosira constricta</i>	88	89,90	<i>Nitzschia putrida</i>	91	92
<i>Chaetoceros muellerii</i>	93	94–100	<i>Phaeodactylum tricornutum</i>	101	102
<i>Chaetoceros tenuissimus</i>	103	104	<i>Porosira glacialis</i>	88	105,106
<i>Conticribrbra guillardii</i>	88	107,108	<i>Psammoneis japonica</i>	unknown	109
<i>Conticribrbra weissflogii</i>	88	67,110,111	<i>Pseudo-nitzschia multistrata</i>	112,113	114
<i>Craspedostauros australis</i>	115	116	<i>Pseudo-nitzschia multiseriis</i>	unknown	117–124
<i>Cyclostephanos invisitatus</i>	88	125,126	<i>Pseudo-nitzschia pungens</i>	unknown	127
<i>Cyclostephanos tholiformis</i>	88	128	<i>Seminavis robusta</i>	129	130
<i>Cyclotella atomus</i>	88	131–134	<i>Skeletonema costatum</i>	135	136–142
<i>Cyclotella baltica</i>	88	143,144	<i>Skeletonema marinoi</i>	145	27,146–151
<i>Cyclotella choctawhatcheeana</i>	88	152,153	<i>Skeletonema menzelii</i>	88	154,155
<i>Cyclotella cryptica</i>	88	156–161	<i>Skeletonema potamos</i>	88	162,163
<i>Cylindrotheca closterium</i>	164	165	<i>Skeletonema tropicum</i>	88	166,167
<i>Cylindrotheca fusiformis</i>	93	168	<i>Stephanocyclus meneghinianus</i>	88	169,170
<i>Detonula confervacea</i>	88	171–173	<i>Stephanodiscus minutulus</i>	88	174,175
<i>Discostella pseudostelligera</i>	88	176–178	<i>Stephanodiscus triporus</i>	88	179,180
<i>Discostella stelligera</i>	88	181,182	<i>Thalassiosira allenii</i>	88	183,184
<i>Discostella stelligeroides</i>	88	185,186	<i>Thalassiosira delicatula</i>	88	187–189
<i>Epithemia pelagica</i>	unknown	190	<i>Thalassiosira exigua</i>	88	191,192
<i>Fistulifera pelliculosa</i>	193	194	<i>Thalassiosira gravida</i>	88	195,196
<i>Fistulifera solaris</i>	193	197,198	<i>Thalassiosira livingstoniorum</i>	88	199
<i>Fragilaria crotonensis</i>	200,201	202	<i>Thalassiosira mediterranea</i>	88	203,204
<i>Fragilaria radians</i>	205	206–211	<i>Thalassiosira oceanica</i>	93	212–218
<i>Fragilariopsis cylindrus</i>	219	18,220–226	<i>Thalassiosira ordinaria</i>	88	227,228
<i>Licmophora abbreviata</i>	unknown	229	<i>Thalassiosira pacifica</i>	88	230–232
<i>Mayamaea pseudoterrestris</i>	233	234	<i>Thalassiosira pseudonana</i>	235	236
<i>Mediolabrus comicus</i>	88	237–239	<i>Thalassiosira profunda</i>	88	240,241

Table 1. References for sequence data used in this study, either for genome annotation or for comparison to annotations of previously existing genome annotations. For some species, we were unable to retrieve an article reference (indicated with “unknown”). Data references are provided both for genome assemblies and transcriptome data.

bacteria impact the dynamics of diatom growth¹⁴. The ecological roles of diatoms and their interaction with other organisms are now better-understood thanks to molecular techniques, which have provided new insights into cell death, silicon metabolism, environmental sensing, and community-level interactions¹⁵.

However, despite the frequency and importance of diatoms in the ecosystem, complete genetic resources for diatoms are scarce. When starting this study, we found 89 Bacillariophyta genome assemblies at National Center for Biotechnology Information (NCBI) Datasets (<https://www.ncbi.nlm.nih.gov/datasets/>, April 1st, 2024, see Supplementary Table S1). Of these, 66 were flagged as “representative genomes”. In total, 13 of these genome assemblies had an annotation of protein coding genes, but only seven of the genome assemblies flagged as “representative genomes” had such an annotation. This means for four of the annotated assemblies, a younger and better but yet unannotated genome assembly existed (but the assembly of *Thalassiosira pseudonana* was not flagged as representative, had been annotated, and no alternative representative genome assembly was available). For three species available at the NCBI, we found an annotation of protein coding genes in PhycoCosm¹⁶ but not at the NCBI. Knowledge about the protein coding genes is essential to fully exploit genome sequences¹⁷, and thus we made it our mission to annotate previously unannotated genome assemblies of the Bacillariophyta.

Initially, we set out to annotate the genome assemblies of all Bacillariophyta that did not have an annotation of protein coding genes, or where a younger and better representative genome has been made available without annotation. Looking at redundancy (sometimes more than one genome assembly for the same species is available), we selected one assembly from each species. However, we decided later to exclude 10 genome assemblies (see Supplementary Table S2), either due to technical problems during download or annotation, or due to data quality. We ended up successfully annotating 49 Bacillariophyta genome assemblies¹⁸ (references to the original sequence data publications are listed in Table 1, genome assembly details are given in Supplementary Table S3, a taxonomic tree is shown in Fig. 1).

With this study, we present the annotation data of protein coding genes for 49 Bacillariophyta genome assemblies that were previously stored as unannotated at NCBI Datasets. Combined with the previously existing annotations, this now makes a total of 58 Bacillariophyta genome annotations accessible for further studies (Fig. 2 visualises how these 58 species cover the taxonomic clades of Bacillariophyta). Together, these data can be applied to various scientific problems and help researchers better understand many of the processes in diatom algae.

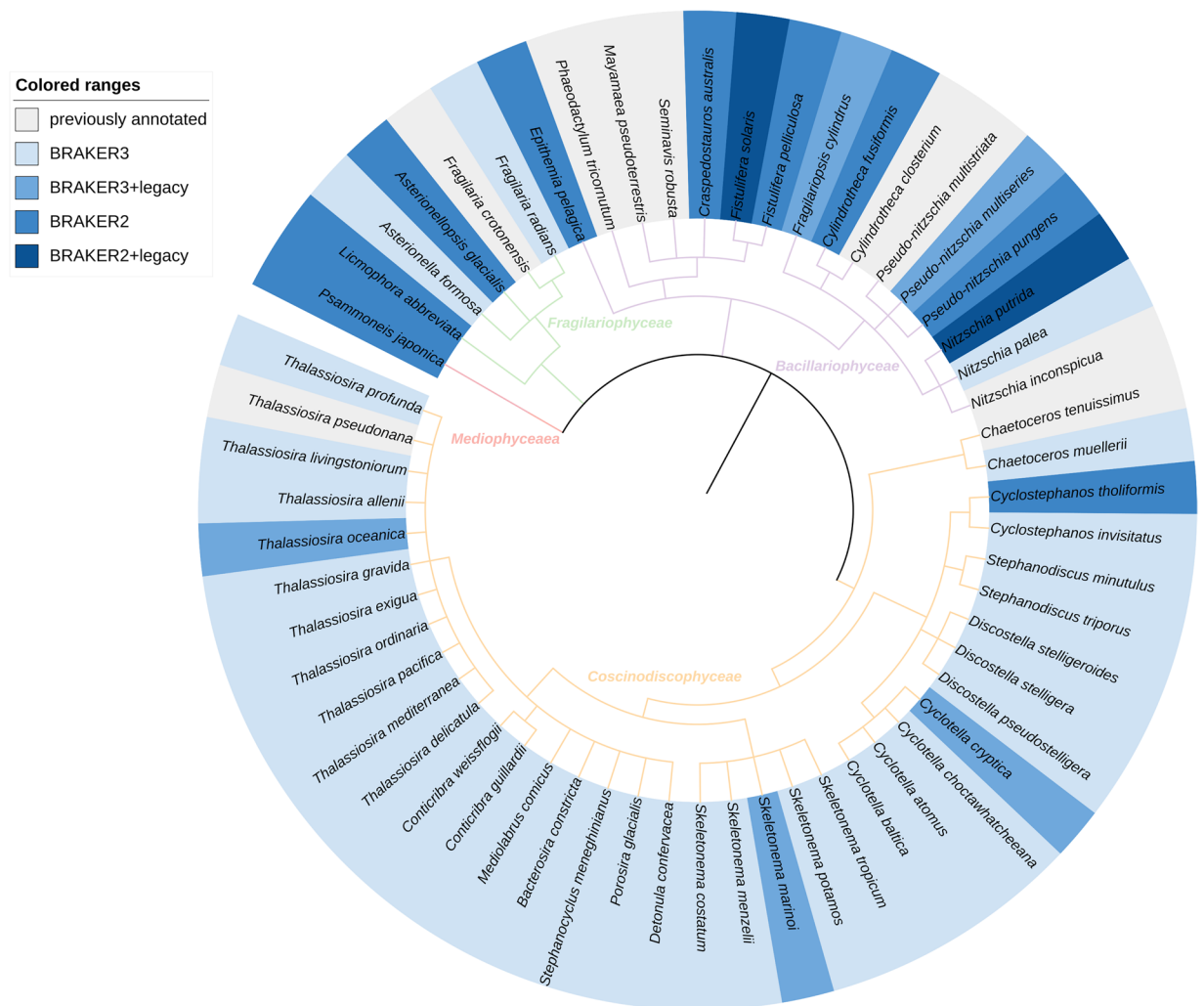


Fig. 1 Taxonomy tree of selected Bacillariophyta genomes. This tree displays species of selected Bacillariophyta genome assemblies available from NCBI datasets between June 14th and 26th 2024. The tree was generated by PhyloT (<https://phylot.biobyte.de/>, August 21st 2024), visualised with iTol⁵⁷. Species with representative genome assemblies with a previously existing annotation at NCBI are labelled in grey. Genomes that we annotated are colored in different shades of blue. From lightest to darkest blue: with BRAKER3; with BRAKER3 including proteins from the same species that were already available for an older assembly at NCBI or from PhycoCosm; with BRAKER2; with BRAKER2 including proteins from the same species that were already available for an older assembly at NCBI or from PhycoCosm. We excluded “uncultured” entries and those matching only two letters followed by a dot, e.g. “sp.”.

Methods

The genome annotations presented here were generated using publicly available genome, transcriptome, and protein data. Data analysis was performed in three steps: (1) data preparation, (2) structural genome annotation, and (3) functional genome annotation. After annotation, we performed assembly contamination analysis (4) and identified horizontal gene transfer candidates (5). Steps 1 and 2 were executed using a semi-automated and reproducible Snakemake workflow¹⁹ that is publicly available at <https://github.com/KatharinaHoff/braker-snake> (August 30th, 2024). Singularity²⁰ was employed to manage software dependencies. Steps 3–5 were performed manually. In addition to genome annotation, we also estimated ploidy in a large number of genome assemblies. All software version numbers are listed in Supplementary Table S4.

Data preparation. In short, we used the NCBI Datasets tool to retrieve Bacillariophyta genome assembly information from the NCBI database (in this case available at <https://www.ncbi.nlm.nih.gov/datasets/> via web browser). Assembly information was filtered to exclude ‘uncultured’ samples and species names ending in ‘sp.’ If multiple assemblies were available for the same species, we prioritized the ‘representative’ assembly, or, if unavailable, the assembly with the largest N50. Genomes with fewer than or equal to 1,000 annotated proteins were selected as candidates for further annotation. This threshold was set to include genome assemblies for annotation that have only a protein coding gene annotation for organelle genomes. For each candidate genome, we checked

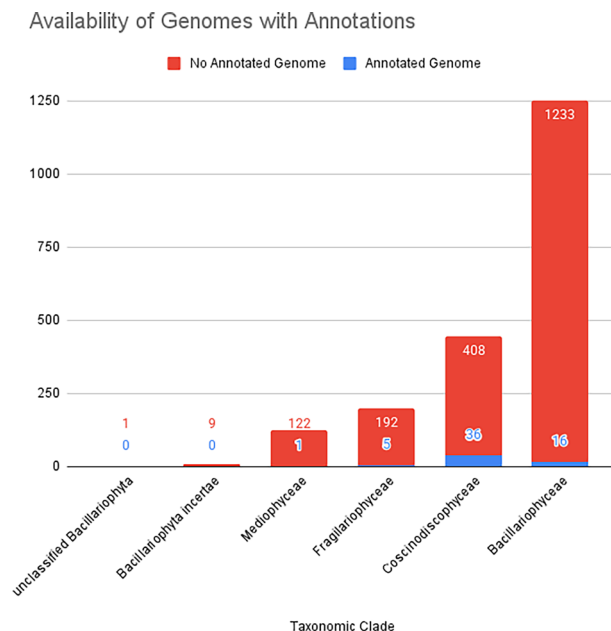


Fig. 2 Stacked bar plot showing the distribution of species with structurally annotated genomes (9 previously annotated, 49 newly annotated in this study) across taxonomic subclades of Bacillariophyta. The lower portion of each bar represents species with annotated genomes, while the full bar height represents the total number of known species according to NCBI Taxonomy.

if an older assembly had existing protein-coding gene annotations (referred to as ‘legacy proteins’) and stored this information. All genome assemblies and any associated legacy proteins were downloaded using the datasets tool.

The workflow automatically retrieves the appropriate OrthoDB v11 partition²¹ for the specified taxon from https://bioinf.uni-greifswald.de/bioinf/partitioned_odb11/ (in the case of diatoms, that is a combination of the following two files: https://bioinf.uni-greifswald.de/bioinf/partitioned_odb11/Stramenopiles.fa.gz and https://bioinf.uni-greifswald.de/bioinf/partitioned_odb11/Viridiplantae.fa.gz). For Bacillariophyta, this corresponds to the Stramenopiles partition, which we combined with the Viridiplantae partition to ensure a larger sequence set.

For species lacking genome annotations, RNA-seq data availability was verified using the Biopython/Entrez API to query the Sequence Read Archive (<https://www.ncbi.nlm.nih.gov/sra>)²². Up to six Illumina paired-end libraries were selected (the top six entries from the Entrez results), and downloaded using fasterq-dump (<https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>, accessed August 21st, 2024). RNA-seq data were aligned to the genome using HISAT2²³. Co-culture libraries were not excluded, as they often provide critical data for diatoms, but libraries with an alignment rate below 20% were discarded. The resulting SAM files were converted to BAM, merged if multiple files existed, sorted, and indexed using SAMtools²⁴.

Before proceeding with automated annotation, we manually queried the PhycoCosm portal (Joint Genome Institute) for existing protein-coding gene annotations for species in our dataset. For *Cyclotella cryptica*^{25,26}, *Nitzschia putrida*²⁷ and *Pseudo-nitzschia multiseriata*, we downloaded available protein sequences and included them as ‘legacy proteins’ in the BRAKER annotation process.

The final output of this data preparation phase was a CSV file that specifies the input files required for the subsequent annotation workflow for each species.

Structural genome annotation. Each selected genome assembly was processed individually using a consistent pipeline. First, RepeatModeler2²⁸ was used to construct a species-specific repeat library, followed by RepeatMasker (<http://www.repeatmasker.org>, accessed August 21st, 2024) to soft mask the repeats in the genome. Depending on the availability of extrinsic data, either BRAKER2 or BRAKER3^{29,30} was employed to predict protein-coding gene structures from the soft-masked genome.

Protein evidence was always used during annotation. For many genomes, the combined Stramenopiles/Viridiplantae protein partition was used as input. Additionally, legacy proteins were incorporated when available. In cases where RNA-seq data were absent, BRAKER2 was run with an option to enrich the predicted gene set using BUSCOs from the Stramenopiles_odb10 dataset³¹, enhanced with compleasm³². BRAKER2 first uses GeneMark-EP +³³, which self-trains GeneMark-ES^{34,35} to identify seed gene sequences. These sequences are then compared to the protein database using DIAMOND³⁶, followed by accurate spliced alignment with Spaln2³⁷. GeneMark-EP + generates an intermediate gene set based on protein evidence, which is refined using AUGUSTUS^{38,39}. TSEBRA⁴⁰ then combines and filters the predictions using protein evidence and BUSCOs as guides⁴¹.

When RNA-seq alignments were available, BRAKER3 was used. This workflow employed GeneMark-ETP⁴², which processes RNA-seq alignments using StringTie2⁴³ to assemble transcripts. GeneMarkS-T⁴⁴ then screens the assembled transcripts for potential genes. DIAMOND and GeneMark-EP +’s protein evidence pipeline were

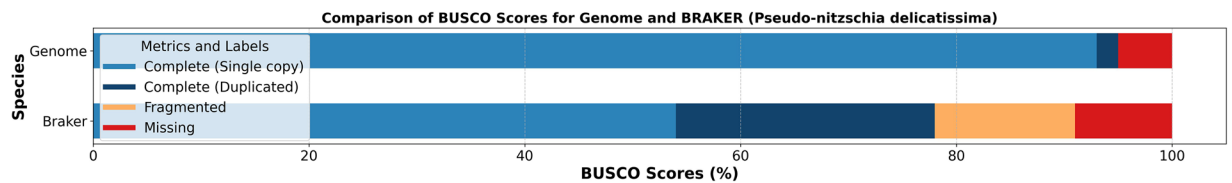


Fig. 3 BUSCO scores of *Pseudo-nitzschia delicatissima*. We decided to exclude this species from further analysis because of the discrepancy of BUSCO scores between genome and protein level.

used to filter the genes, and GeneMark-ETP also performed initial gene predictions based on self-training. AUGUSTUS was again trained on a reliable subset of predicted genes, and the final gene set was merged using TSEBRA.

Not all BRAKER jobs completed successfully; assemblies affected by these failures were excluded from further analysis (see Supplementary Table S2).

For quality control, we ran BUSCO with the stramenopiles_odb10 dataset on both the genome assemblies and the predicted protein sequences. Genomes were excluded if there was a significant discrepancy between BUSCO completeness scores at the genome level and the predicted protein level. For example, despite a 95% BUSCO completeness score at the genome level, *Pseudo-nitzschia delicatissima* achieved only 72% completeness at the annotation level and was excluded (see Fig. 3). Additionally, *Thalassiosira sundarbana* was excluded due to low genome BUSCO completeness (15%) and contamination in the database. *Epithemia catenata* was also excluded due to low genome BUSCO completeness (56%).

Functional gene annotation. The EnTAP functional annotation software was employed to provide functional descriptors and identify potential contaminants for the predicted proteins⁴⁵. EnTAP was configured with two curated databases, NCBI's RefSeq Protein⁴⁶ and UniProtKB/Swiss-Prot⁴⁷, for similarity searches, utilising a 50% target and query coverage minimum, and a DIAMOND E-value threshold of 0.00001. An optimal alignment was selected for each protein query based on phylogenetic relevance, informativeness, and standard alignment quality metrics. Additionally, EnTAP performed independent searches against the EggNOG database⁴⁸ using the EggNOG-mapper toolbox⁴⁹. The resulting gene family assignments, along with high-quality similarity search alignments, facilitated the subsequent connections to Gene Ontology terms^{50,51}, protein domains from Pfam⁵², and pathway associations via KEGG⁵³.

Contamination and HGT analysis. We screened each assembly for potential contamination by leveraging the EnTAP classification of individual transcripts as either contaminated or uncontaminated. In EnTAP, contaminant transcripts aligned with high confidence to the NCBI RefSeq microbial database or exclusively to the microbial gene families housed in EggNOG. Annotated transcripts in each assembly were mapped back to their corresponding contigs, and the proportion of “contaminated” versus “uncontaminated” transcripts was computed per contig. Any contig with more than 75% of its transcripts flagged as contamination was classified as potentially contaminated and a Note was added to each CDS feature in the gff3 file for this assembly. We detected between 1 and 318 contaminated contigs in 39 of the assemblies (see Supplementary Table S6).

Furthermore, we evaluated each predicted proteome, using the longest isoform per gene, for potential horizontally transferred genes (HGT). In specific, we identified HGT candidates that occurred in one or more Bacillariophyta, but were not conserved in other members of the Ochrophyta. For this, we performed additional DIAMOND searches against donor databases (NCBI RefSeq microbial and plant) and a recipient database (NCBI RefSeq Ochrophyta with all Bacillariophyta removed), using coverage thresholds of at least 50% for both query and subject and an e-value cutoff of 1e-5. Candidate HGTs were initially identified as those aligning to the microbial donor database while failing to align, either to the plant donor, or to the recipient database. We then filtered the HGT candidates by removing those lacking two flanking neighboring genes belonging to the target species, or with (either) flanking genes identified as a contaminant, or at the end of a scaffold (lacking two flanking genes for evaluation). The methodology for HGT identification and downstream filtering is available in EnTAP (v.2.3.0). The remaining genes were retained as HGT candidates (see Supplementary Table S7) and each corresponding CDS feature in the gff3 file was tagged accordingly. We identified between 1 to 129 HGT candidates, per species, in 42 of the annotations.

Orthogroup analysis. We used OrthoFinder to identify orthologous gene groups across species by performing an all-versus-all comparison of protein sequences after removing proteins that are located on genomic sequences that were suspected to be contaminants, and excluding horizontal gene transfer candidates (using the longest isoform of each gene). Based on sequence similarities, genes were grouped into orthogroups, which represent sets of genes descended from a common ancestor. To ensure the reliability of the species tree, we included species from nine publicly available annotations (see Table 3) and also the *Oomycota* clade for *Phytophthora cinnamomi*, *Phytophthora infestans*, *Phytophthora ramorum*, *Phytophthora sojae*, and *Bremia lactucae* (see Table 4).

A species tree (Fig. 4) was generated using OrthoFinder with the `-M msa` option, which builds gene trees based on multiple sequence alignments (using MAFFT⁵⁴) and infers their topology with FastTree⁵⁵. FastTree uses an approximate maximum-likelihood approach and provides SH-like (Shimodaira–Hasegawa-like) support values for each branch, which offer a fast estimate of how reliable each split is—though they are not traditional bootstrap values.

Species name	#Genes	#Tx	Mono:Mult	Median Mult	Max Mult	Species name	#Genes	#Tx	Mono:Mult	Median Mult	Max Mult
Statistics of raw BRAKER output											
<i>Asterionella formosa</i>	17643	19690	2.44	2	15	<i>Nitzschia palea</i>	18945	19154	1.75	2	16
<i>Asterionellopsis glacialis</i>	19134	20492	1.59	2	16	<i>Nitzschia putrida</i>	20530	22856	1.3	2	13
<i>Bacterosira constricta</i>	22749	25062	1.98	2	30	<i>Porosira glacialis</i>	48999	49187	6.65	2	31
<i>Chaetoceros muellerii</i>	13586	15972	1.85	2	26	<i>Psammoneis japonica</i>	22827	24371	1.17	2	19
<i>Conticribra guillardii</i>	17169	19531	1.28	3	23	<i>Pseudo-nitzschia multiseriis</i>	41854	44440	2.93	2	24
<i>Conticribra weissflogii</i>	15985	17945	1.29	3	20	<i>Pseudo-nitzschia pungens</i>	17930	19302	1.36	2	25
<i>Craspedostauros australis</i>	19297	21075	1.58	2	14	<i>Skeletonema costatum</i>	21327	22430	1.56	2	16
<i>Cyclostephanos invisitatus</i>	13354	16070	1.18	3	23	<i>Skeletonema marinoi</i>	26297	26524	2.46	2	17
<i>Cyclostephanos tholiformis</i>	19600	20902	0.8	3	34	<i>Skeletonema menzelii</i>	14214	17490	1.92	2	12
<i>Cyclotella atomus</i>	17063	19759	1.02	3	29	<i>Skeletonema potamos</i>	14321	17682	1.98	2	14
<i>Cyclotella baltica</i>	20650	23038	1.28	3	54	<i>Skeletonema tropicum</i>	23816	27647	2.77	2	14
<i>Cyclotella choctawhatcheeana</i>	16876	19331	1.01	3	50	<i>Stephanocyclus meneghinianus</i>	17489	20102	1.29	3	27
<i>Cyclotella cryptica</i>	27018	30077	1.83	3	55	<i>Stephanodiscus minutulus</i>	13530	15516	1.31	2	27
<i>Cylindrotheca fusiformis</i>	18449	19938	1.28	2	14	<i>Stephanodiscus triporus</i>	14465	16751	1.38	2	39
<i>Detonula confervacea</i>	25217	27369	2.29	2	23	<i>Thalassiosira allenii</i>	25993	27929	2.08	2	34
<i>Discostella pseudostelligera</i>	10867	14109	0.93	3	28	<i>Thalassiosira delicatula</i>	34282	36115	3.56	2	38
<i>Discostella stelligera</i>	14843	17755	1.36	3	40	<i>Thalassiosira exigua</i>	25701	28025	1.99	3	40
<i>Discostella stelligeroides</i>	16328	18945	1.36	2	40	<i>Thalassiosira gravida</i>	26431	27989	1.89	2	39
<i>Epithemia pelagica</i>	18983	21013	1.71	2	18	<i>Thalassiosira livingstoniorum</i>	44697	47357	3.79	3	38
<i>Fistulifera pelliculosa</i>	17328	21235	0.32	3	37	<i>Thalassiosira mediterranea</i>	19055	21492	1.38	3	49
<i>Fistulifera solaris</i>	26808	31306	0.8	2	18	<i>Thalassiosira oceanica</i>	26307	29507	1.14	3	42
<i>Fragilaria radians</i>	21031	23879	2.45	2	13	<i>Thalassiosira ordinaria</i>	28461	31353	1.8	3	32
<i>Fragilariopsis cylindrus</i>	21946	25304	1.74	2	24	<i>Thalassiosira pacifica</i>	29416	31996	2.54	2	29
<i>Licmophora abbreviata</i>	15275	16375	1.25	2	15	<i>Thalassiosira profunda</i>	25451	29337	1.66	2	19
<i>Mediolabrus comicus</i>	15450	18098	2.26	2	14						
Statistics of filtered BRAKER output											
<i>Asterionella formosa</i>	14868	16684	1.93	2	15	<i>Nitzschia palea</i>	14103	16798	1.40	2	16
<i>Asterionellopsis glacialis</i>	16229	17400	1.32	2	16	<i>Nitzschia putrida</i>	17981	20020	1.07	2	13
<i>Bacterosira constricta</i>	17271	19412	1.36	2	30	<i>Porosira glacialis</i>	20698	21786	2.51	2	31
<i>Chaetoceros muellerii</i>	11855	14061	1.50	2	26	<i>Psammoneis japonica</i>	19726	21063	0.92	2	19
<i>Conticribra guillardii</i>	14400	16598	0.95	3	23	<i>Pseudo-nitzschia multiseriis</i>	25913	30745	1.69	2	24
<i>Conticribra weissflogii</i>	13188	15000	0.93	3	20	<i>Pseudo-nitzschia pungens</i>	15463	16661	1.09	2	25
<i>Craspedostauros australis</i>	16473	17876	1.28	2	14	<i>Skeletonema costatum</i>	18452	20086	1.36	2	16
<i>Cyclostephanos invisitatus</i>	11843	14437	0.98	2	23	<i>Skeletonema marinoi</i>	22888	26524	2.43	2	17
<i>Cyclostephanos tholiformis</i>	17114	18321	0.58	3	34	<i>Skeletonema menzelii</i>	12968	16052	1.66	2	12
<i>Cyclotella atomus</i>	14665	17168	0.78	3	29	<i>Skeletonema potamos</i>	13005	16171	1.71	2	14
<i>Cyclotella baltica</i>	16981	19122	0.91	3	54	<i>Skeletonema tropicum</i>	19435	22729	2.10	2	14
<i>Cyclotella choctawhatcheeana</i>	14743	17008	0.78	3	50	<i>Stephanocyclus meneghinianus</i>	14567	16694	0.92	3	27
<i>Cyclotella cryptica</i>	20891	23573	1.23	3	55	<i>Stephanodiscus minutulus</i>	11622	13500	1.04	2	27
<i>Cylindrotheca fusiformis</i>	16067	17426	1.11	2	14	<i>Stephanodiscus triporus</i>	12523	14698	1.12	2	39
<i>Detonula confervacea</i>	17910	19888	1.42	2	23	<i>Thalassiosira allenii</i>	19936	21725	1.45	2	34
<i>Discostella pseudostelligera</i>	10212	13359	0.84	2	28	<i>Thalassiosira delicatula</i>	23338	25042	2.33	2	38
<i>Discostella stelligera</i>	12677	15474	1.12	2	40	<i>Thalassiosira exigua</i>	18569	20692	1.25	3	40
<i>Discostella stelligeroides</i>	14509	16984	1.17	2	40	<i>Thalassiosira gravida</i>	20588	22036	1.22	2	39
<i>Epithemia pelagica</i>	16107	17669	1.39	2	18	<i>Thalassiosira livingstoniorum</i>	27460	29797	2.18	3	38
<i>Fistulifera pelliculosa</i>	16965	20824	0.27	3	37	<i>Thalassiosira mediterranea</i>	16320	18575	1.07	3	49
<i>Fistulifera solaris</i>	26502	30972	0.73	2	18	<i>Thalassiosira oceanica</i>	25415	28595	1.11	3	42
<i>Fragilaria radians</i>	18463	21050	2.06	2	13	<i>Thalassiosira ordinaria</i>	22181	24814	1.27	3	32
<i>Fragilariopsis cylindrus</i>	21464	24791	1.68	2	24	<i>Thalassiosira pacifica</i>	20560	22904	1.60	2	29
<i>Licmophora abbreviata</i>	14203	15232	1.15	2	15	<i>Thalassiosira profunda</i>	24879	28765	1.66	2	19
<i>Mediolabrus comicus</i>	13615	16061	1.97	2	14						

Table 2. Top: Statistics on the raw and intermediate BRAKER gene sets, bottom: statistics of the filtered and final BRAKER gene sets. #Genes, number of genes; #Tx, number of transcripts; Mono:Mult, mono-exon to multi-exon ratio; Median Mult, median number of exons in multi-exon genes; Max Mult, largest number of exons in multi-exon genes.

Species name	Genome assembly accession number	Annotation source	#Genes	#Tx	Mono:Mult	Median Mult	Max Mult
<i>Chaetoceros tenuissimus</i>	GCA_021927905.1	DDBJ	18397	18670	1.38	2	96
<i>Cylindrotheca closterium</i>	GCA_933822405.4	EMBL	24371	24633	1.53	2	18
<i>Fragilaria crotonensis</i>	GCA_022925895.1	Genbank	26015	26015	2.14	2	15
<i>Mayamaea pseudoterrestris</i>	GCA_027923505.1	DDBJ	11017	11017	1.04	2	20
<i>Nitzschia inconspicua</i>	GCA_019154785.2	Genbank	38391	38393	1.37	2	14
<i>Phaeodactylum tricornutum</i>	GCA_000150955.2	Genbank	10321	10339	1.19	2	13
<i>Pseudo-nitzschia multistriata</i>	GCA_900660405.1	EMBL	11909	11952	1.46	2	18
<i>Seminavis robusta</i>	GCA_903772945.1	EMBL	35858	35865	1.44	2	44
<i>Thalassiosira pseudonana</i>	GCA_000149405.2	Genbank	11681	11686	0.66	3	77

Table 3. Descriptive statistics of the previously existing annotations of protein coding genes in representative genome assemblies at NCBI. #Genes, number of genes; #Tx, number of transcripts; Mono:Mult, mono-exon to multi-exon ratio; Median Mult, median number of exons in multi-exon genes; Max Mult, largest number of exons in multi-exon genes.

Species name	Genome assembly accession number	Annotation source	#Genes	#Tx	Size (Mb)	Scaffolds
<i>Phytophthora infestans</i>	GCF_000142945.1	RefSeq	17,797	19,337	228.5	4,921
<i>Phytophthora sojae</i>	GCF_000149755.1	RefSeq	26,489	28,142	82.6	82
<i>Bremia lactucae</i>	GCF_004359215.1	RefSeq	9,766	9,766	115.9	220
<i>Phytophthora cinnamomi</i>	GCF_018691715.1	RefSeq	19,973	19,973	109.7	133
<i>Phytophthora ramorum</i>	GCF_020800215.1	RefSeq	15,267	15,268	57.45	28

Table 4. Descriptive statistics of genome assemblies and protein-coding gene annotations in Oomycota. #Genes, number of genes; #Tx, number of transcripts.

These gene trees were then combined using the STAG⁵⁶ (Species Tree from All Genes) algorithm, which reconstructs the species tree by integrating information from genome-wide orthogroup data, including multi-copy gene families. The support values shown on internal nodes of the species tree reflect how often each grouping is supported across all gene trees. Finally, the tree was rooted using STRIDE (Species Tree Root Inference from Duplication Events)⁵⁷, which uses gene duplication patterns to determine the most likely root. Altogether, this approach combines gene family structure and duplication history to produce a comprehensive view of species relationships.

The OrthoFinder results files, including orthogroups, are available at⁵⁸.

Filtering of false positive single exon genes. Descriptive statistics of the raw BRAKER output (see Table 2) and the EnTAP annotation rate (see Supplementary Table S5) suggested that BRAKER overpredicted single-exon genes in some cases. This issue has previously been reported in land plant annotations⁵⁹.

To address this and filter out potential false positive single-exon gene predictions—while retaining gene models that may be of scientific interest—we applied the following filtering approach: We discarded single-exon gene models that lacked a functional annotation by EnTAP, did not have a significant hit in a DIAMOND search against the NCBI RefSeq non-redundant proteins (NR) database (February 2nd, 2024), and were not part of an orthologous group spanning more than one species in the OrthoFinder results.

File processing. In order to prepare NCBI-compliant GFF3 files, the filtered BRAKER output files were decorated with product names and notes according to EnTAP results (command lines at https://github.com/Gaius-Augustus/Diatom_annotation_scripts).

Ploidy Estimation with Smudgeplot. GenBank accessions were used to retrieve additional metadata from NCBI, including read type, DNA SRA accessions, genome size, and assembly level. Ploidy was not estimated if the SRA accession was unavailable or corresponded to long-read data (i.e., PacBio, ONT). A Nextflow pipeline (available at https://github.com/Gaius-Augustus/Diatom_annotation_scripts) was developed to estimate the ploidy for all remaining individuals in parallel. Paired-end SRA accessions were first fetched using sra-tools and filtered for fungal, bacterial, archaeal, and viral contaminants using Kraken’s⁶⁰ default parameters. Coverage was calculated before and after contaminants were removed, ranging between 12–499x. Next, FastK built a database for each contaminant-free library using a k-mer size of 21. With the FastK database, Smudgeplot ‘hetmers’ found all k-mer pairs⁶¹. The lower k-mer threshold (–L) was estimated with Smudgeplot ‘cutoff’. The final ploidy estimate and proportion of heterozygosity carried by paralogs was extracted from the verbose summary text file resulting from the ‘plot’ module (results in Supplementary Table S8).

Data Records

The data set is available at Zenodo (ref. ⁶² version v6). It consists of an archive file called Bacillariophyta_anno-tations.tar.gz. After extraction, the resulting folder Bacillariophyta_anno-tations contains gff3 format files with gene models (summarized in Table 2) that each correspond to a FASTA format genome file. The accession num-bers of the genome assemblies are for user convenience listed in the additionally included file README.md.

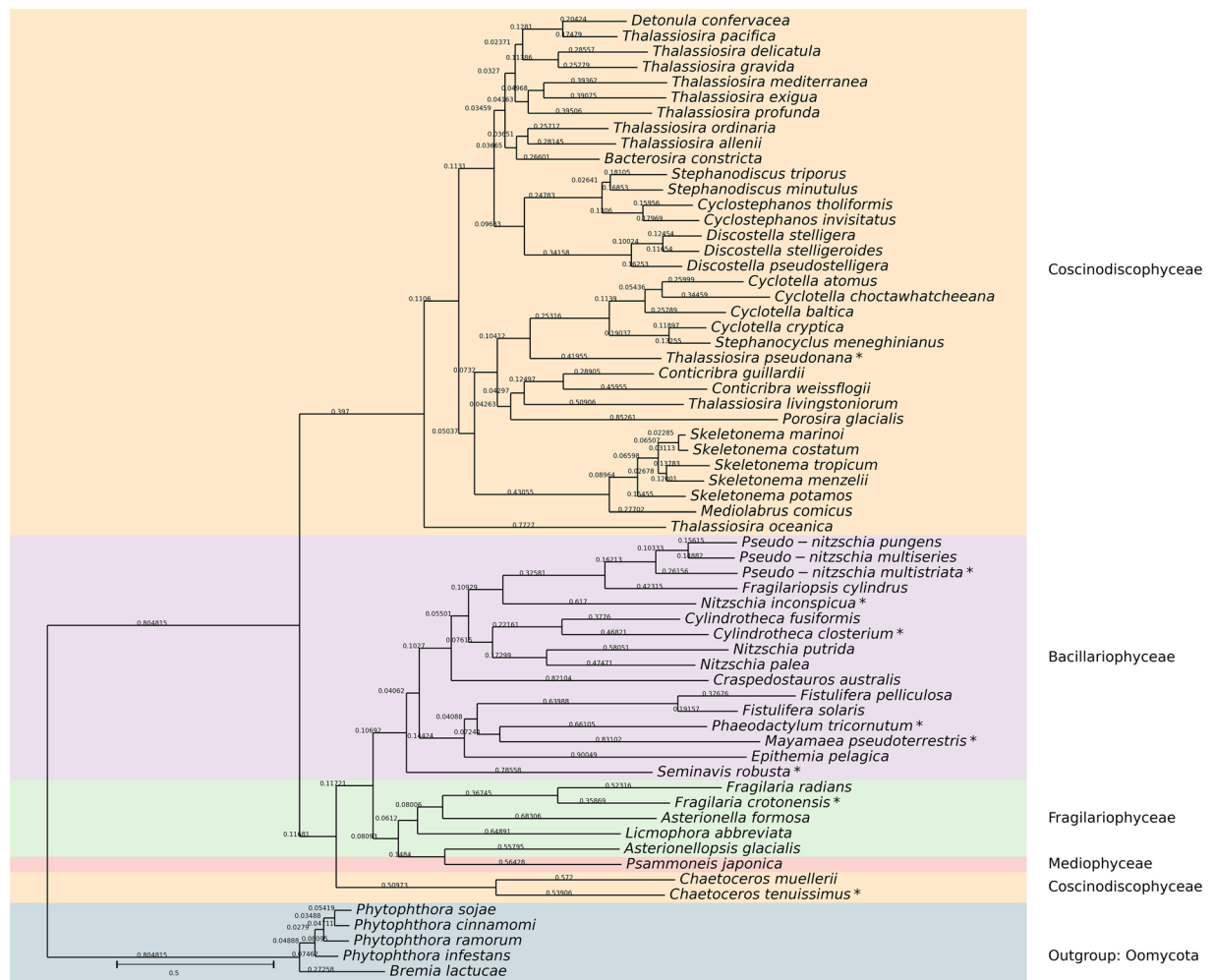


Fig. 4 Rooted species tree of Bacillariophyta with an annotation of protein coding genes. Major diatom lineages are labelled on the right. The previously annotated species (*C. tenuissimus*, *C. closterium*, *F. crotonensis*, *M. pseudoterrestris*, *N. inconspicua*, *P. tricornutum*, *Pseudo-nitzschia multistriata*, *S. robusta*, and *T. pseudonana*) are labeled with a star. The numbers displayed on branches correspond to support values according to the Shimodaira-Hasegawa-like method²⁴².

Technical Validation

We performed a genome annotation study focusing on 49 diatom species, aiming to create a robust genomic dataset that supports future research into diatom biology and evolution. To emphasize the need for our work, we plotted the distribution of existing Bacillariophyta genome annotations in the context of all known species within this taxon (Fig. 2). This analysis highlights the limited representation of annotated diatom species in current genomic resources. Our work significantly expands the number of annotated assemblies from 9 (or 15, including legacy assemblies) to 58, providing a valuable resource for diatom research.

Descriptive statistics for the gene structures of the newly annotated genomes are provided in Table 2. Previously annotated diatom genomes at NCBI contain between 10,321 and 38,391 protein-coding gene models (see Table 3). The gene numbers in the newly generated gene sets fall within this range. Vuruputoor *et al.* (2023)⁵⁹ recommend using the ratio of mono-exon to multi-exon genes as a quality measure for genome annotations, with a suggested ratio of 0.2 for land plants. In contrast, diatom genomes exhibit a higher proportion of single-exon genes, ranging from 0.66 to 2.14 (based on existing annotations; see Table 3). The BRAKER2 and BRAKER3 pipelines tend to overpredict single-exon genes, and we hypothesise that this phenomenon extends to diatom genomes as well. After applying our filtering approach, only five species - *Porosira glacialis* (3.34), *Skeletonema marinoi* (2.46), *Skeletonema tropicum* (2.6), *Thalassiosira delicatula* (2.87), and *Thalassiosira mediterranea* (2.71) - exceeded this range. These deviations are modest and may partly be attributed to selfish DNA elements, such as unmasked transposons and inserted retroviruses. The exon structure of the novel annotations aligns with previously annotated genomes in terms of the median number of exons per gene (2–3) and the largest number of exons per transcript (13–96) (compare Tables 2 and 3).

Evaluating the quality of novel genome annotations is challenging. We used BUSCO to assess genome completeness at both the genome and protein levels (only the longest isoform per gene), following Earth BioGenome

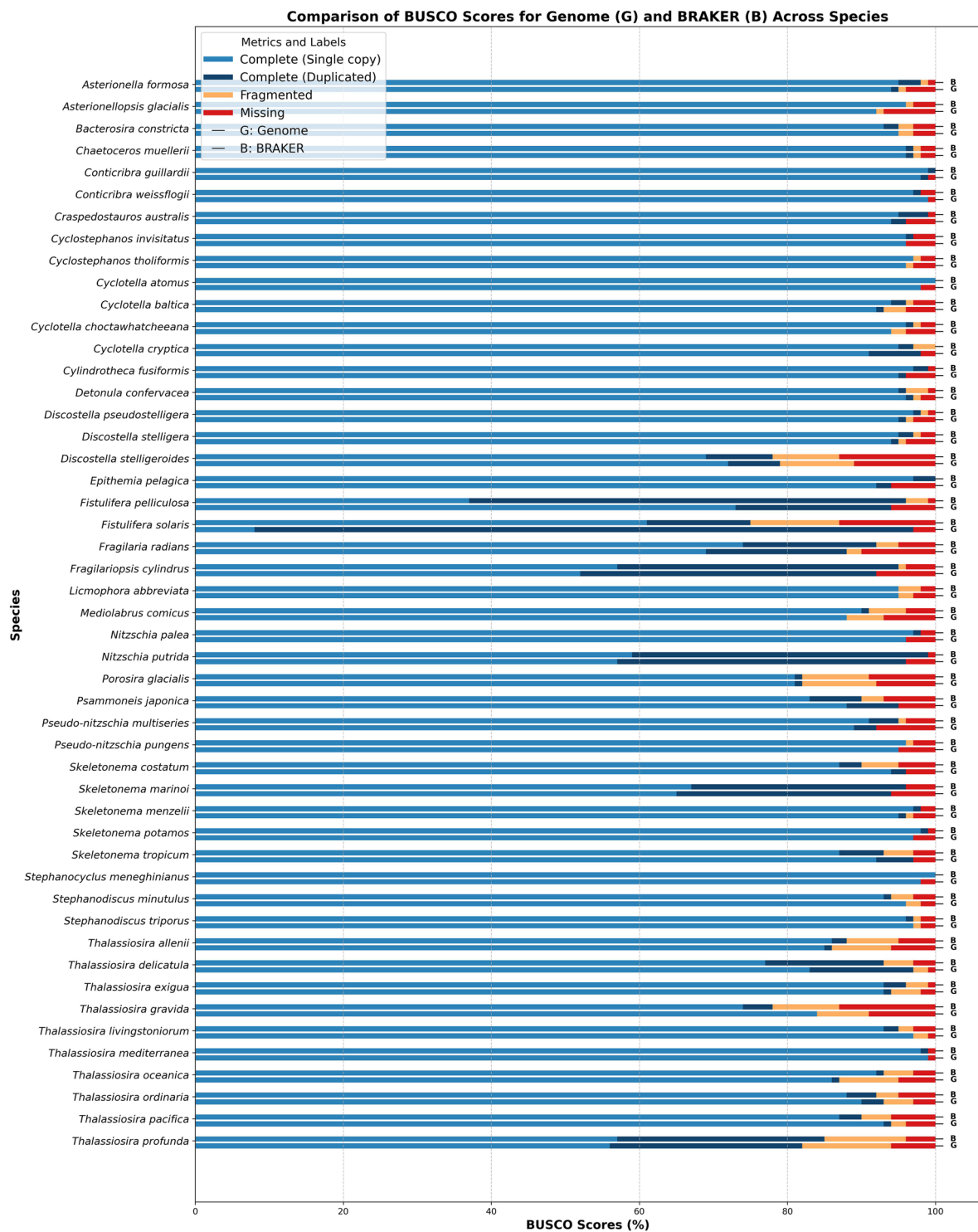


Fig. 5 BUSCO results of genomes and protein sets (only the longest isoform per gene was used in this analysis). This plot demonstrates the quality of genome assemblies (G = Genome) and predicted protein sets (B = BRAKER) across all here annotated species; species ordered alphabetically. The categories Complete (Single copy or duplicated), Fragmented, or Missing BUSCOs are shown.

Project guidelines¹⁷. BUSCO estimates the proportion of genes typically present as single copies within a clade. However, the stramenopiles_odb10 dataset applicable to diatoms is relatively small (100 marker genes). While BUSCO scores measure sensitivity within this limited dataset (see Fig. 5), a close agreement between genome- and protein-level scores suggests that the new annotations do not lack a significant portion of BUSCO genes detectable at the genome level. This is expected, as the stramenopiles_odb10 dataset was used as input for BRAKER.

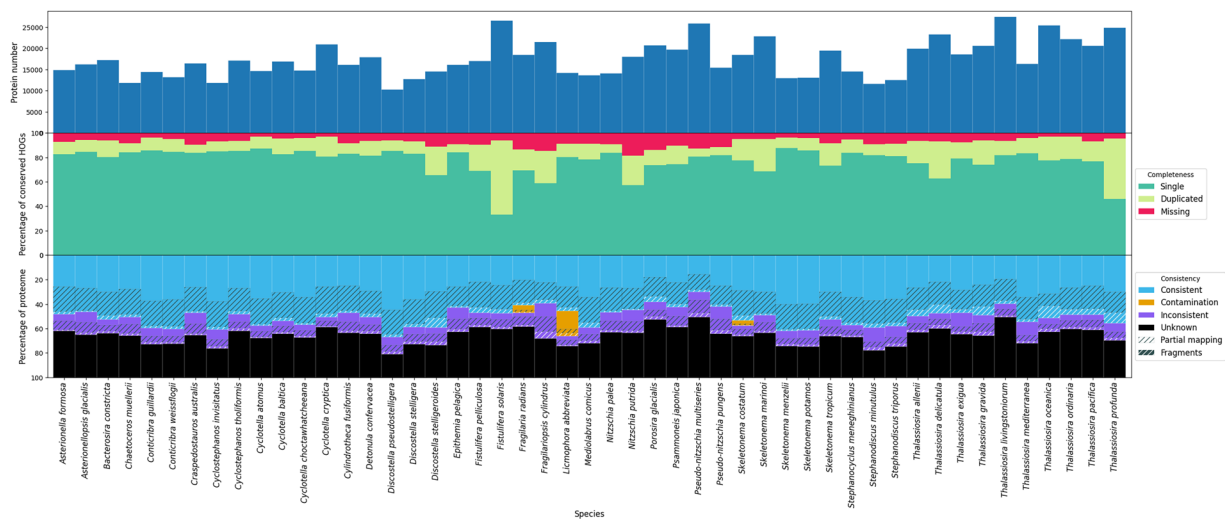


Fig. 6 OMARk results of newly annotated Bacillariophyta genomes. The top bar graph displays the number of canonical proteins per proteome, the middle graph presents completeness metrics based on single-copy, duplicated, or missing conserved genes, and the bottom graph illustrates the consistency assessment. Proteins are categorized as consistent, contamination, inconsistent, unknown, partial mapping, or fragments. Consistent proteins align with taxonomically expected gene families, while contamination refers to proteins matching gene families from other species. Inconsistent proteins belong to gene families outside the expected lineage but are not contaminants. Unknown proteins cannot be assigned to known gene families and may represent novel or misannotated sequences. Partial mapping indicates proteins aligning with gene families over less than 80% of their sequence, and fragments are proteins shorter than half the median length of their gene family.

We also applied OMARk⁶³ to further assess the quality of protein-coding gene annotations. OMARk uses conserved homologous genes (HOGs) from the OMA database⁶⁴ and the OMAR software for fast protein placement⁶⁵. For Bacillariophyta, the relatively small Ochrophyta subset of 942 HOGs is applicable. While this is a limited number of marker genes, OMARk provides additional metrics, assessing contamination, consistency, and fragmentation. Figure 6 shows OMARk results for our newly annotated genomes, while Fig. 7 displays results for previously available reference genomes. Unlike BUSCO, OMARk correctly handles alternative transcript isoforms, suggesting that the observed duplicates are likely real. Notably, we observed a high level of HOG completeness across most assemblies. However, *Thalassiosira profunda* and *Fistulifera solaris* showed a surprisingly high number of duplicate HOGs. For *T. profunda*, this is consistent with BUSCO scores at the genome level, indicating agreement between different metrics. In contrast, the source of duplicates in *F. solaris* remains unclear. We explored the genome assembly statistics (Table 5) but found no obvious explanation. Additionally, OMARk identified a significant level of contamination in the genome of *Licmophora abbreviata*, which had not been flagged as contaminated in public databases (Fig. 8).

To better explain variation in genome-level BUSCO duplication across the diatoms, ploidy was estimated. *Fistulifera* has already been recognized as an allopolyploid^{66,67}, yielding BUSCO duplication rates between 21% and 89%. While elevated BUSCO duplication can indicate polyploidy in some cases, it may also be a result of incomplete purging, mixed samples, or elevated heterozygosity. *Skeletonema marinoi* and *Thalassiosira profunda*, for example, have BUSCO duplication rates of 29% and 26%, respectively, but are still estimated to be diploid. Roberts *et al.* (2024)⁶⁸ report the same ploidy levels. Interestingly, the only exception is *Stephanodiscus minutulus*, which was estimated to be triploid in this study (see Supplementary Table S8).

In the current study, we mainly used the orthogroups constructed by OrthoFinder to filter likely false positive predicted single exon genes. However, the OrthoFinder results themselves are also an interesting result of this study. Across species, the percentage of genes assigned to orthogroups ranged from 85.4% to 99.1%, indicating a generally high rate of orthogroup recovery. For most species, over 90% of genes were successfully assigned, with especially high assignment rates observed in *Skeletonema marinoi* (99.1%), *Discostella pseudostelligera* (98.9%) and *Skeletonema menzeli* (98.9%) (Fig. 8). A few species, such as *Thalassiosira delicatula* (85.4%) and *Bremia lactucae* (species from the outgroup used for the OrthoFinder analysis) (89.8%), showed slightly lower assignment rates, potentially reflecting lineage-specific gene content. It should be noted that OrthoFinder also constructs intra-species orthogroups, which consist of genes from a single species.

In total, 1,115,003 genes (96.8% of the dataset) were assigned to inter-species orthogroups, emphasizing the significant degree of genetic overlap among the species included in this study. Orthogroup inference resulted in a total of 7,092 species-specific orthogroups, comprising 29,717 genes, which represents 2.6% of all input genes. It points to potential species-specific adaptations, with these gene families possibly linked to unique ecological roles or environmental responses. The mean orthogroup size was 32.7 genes, while the median size was 6.0, reflecting a skewed distribution with some large, highly conserved orthogroups. The G50 (i.e., the orthogroup size above which 50% of all assigned genes are found) was 87 for assigned genes and 119 when considering all input genes. The corresponding O50 values—representing the number of the largest orthogroups containing

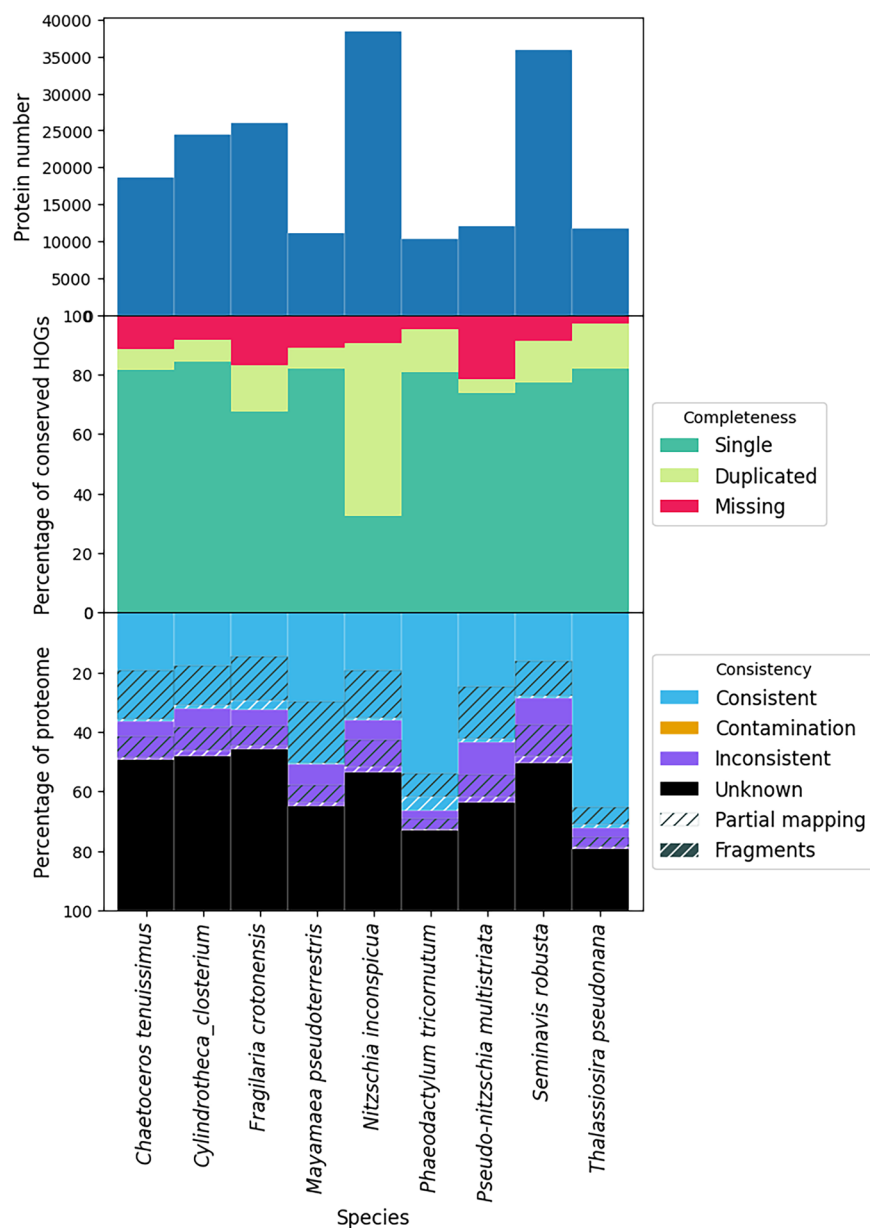


Fig. 7 OMArk results of previously annotated Bacillariophyta reference genome assemblies. Since it was not straight-forward to extract alternative isoform nesting from the GFF3 files, we extracted the longest isoform for each locus with TSEBRA instead of generating an isoform information file for OMArk. The top bar graph displays the number of canonical proteins per proteome, the middle graph presents completeness metrics based on single-copy, duplicated, or missing conserved genes, and the bottom graph illustrates the consistency assessment. Proteins are categorized as consistent, contamination, inconsistent, unknown, partial mapping, or fragments. Consistent proteins align with taxonomically expected gene families, while contamination refers to proteins matching gene families from other species. Inconsistent proteins belong to gene families outside the expected lineage but are not contaminants. Unknown proteins cannot be assigned to known gene families and may represent novel or miss-annotated sequences. Partial mapping indicates proteins aligning with gene families over less than 80% of their sequence, and fragments are proteins shorter than half the median length of their gene family. These metrics provide a comprehensive evaluation of annotation quality beyond completeness alone.

half of the genes—were 2,577 and 4,381, respectively. Notably, only 178 orthogroups included genes from all species. The relatively low number of orthogroups containing genes from all species (262) suggests a high level of gene family diversification, likely reflecting extensive evolutionary divergence and possible lineage-specific expansions or losses across the dataset. The total number of genes per species varied widely, from less than 10,000 in *Discostella pseudostelligera* to almost 36,000 in *Seminais robusta* (see statistics per species in the Supplementary Table S9), highlighting the diversity in genome size across the dataset.

Species name	#Seq	#Nuc (Mbp)	N50	Longest Seq	Species name	#Seqs	#Nuc (Mbp)	N50	Longest Seq
<i>Asterionella formosa</i>	15436	68	15906	173875	<i>Nitzschia palea</i>	3279	41	40280	272293
<i>Asterionellopsis glacialis</i>	6922	67	21950	209170	<i>Nitzschia putrida</i>	234	47	545464	3790763
<i>Bacterosira constricta</i>	43159	105	4147	59115	<i>Porosira glacialis</i>	223117	307	1674	33843
<i>Chaetoceros muellerii</i>	8639	38	21121	213903	<i>Psammoneis japonica</i>	597	91	377693	1221928
<i>Conticribra guillardii</i>	19149	72	7408	58278	<i>Pseudo-nitzschia multiseries</i>	121	252	28641012	39424003
<i>Conticribra weissflogii</i>	35218	130	5543	82792	<i>Pseudo-nitzschia pungens</i>	62	67	6147877	9664134
<i>Craspedostauros australis</i>	90	75	1724159	3871151	<i>Skeletonema costatum</i>	1282	51	97960	756974
<i>Cyclostephanos invisitatus</i>	13656	58	9661	110704	<i>Skeletonema marinoi</i>	52	65	3004636	5945085
<i>Cyclostephanos tholiformis</i>	19322	73	6992	68474	<i>Skeletonema menzelii</i>	2364	33	32299	166604
<i>Cyclotella atomus</i>	9734	54	12871	88620	<i>Skeletonema potamos</i>	2853	35	30345	184032
<i>Cyclotella baltica</i>	26003	99	5685	111838	<i>Skeletonema tropicum</i>	202	79	3169793	5737774
<i>Cyclotella choctawhatcheeana</i>	10213	55	12136	86732	<i>Stephanocyclus meneghinianus</i>	102014	189	2787	114624
<i>Cyclotella cryptica</i>	662	171	494169	2497727	<i>Stephanodiscus minutulus</i>	17827	56	5555	46828
<i>Cylindrotheca fusiformis</i>	13423	48	17542	346874	<i>Stephanodiscus triporus</i>	15611	58	7327	85084
<i>Detonula confervacea</i>	78700	160	3046	113179	<i>Thalassiosira allenii</i>	53254	106	3047	68091
<i>Discostella pseudostelligera</i>	3358	30	20928	129794	<i>Thalassiosira delicatula</i>	142633	248	1502	107858
<i>Discostella stelligera</i>	22180	59	6077	109993	<i>Thalassiosira exigua</i>	60034	166	4614	96147
<i>Discostella stelligeroides</i>	14971	49	5805	63052	<i>Thalassiosira gravida</i>	114748	167	1841	53391
<i>Epithemia pelagica</i>	23	61	3856736	6983076	<i>Thalassiosira livingstoniorum</i>	140151	373	4508	54271
<i>Fistulifera pelliculosa</i>	573	36	176105	1938242	<i>Thalassiosira mediterranea</i>	20845	77	7468	58704
<i>Fistulifera solaris</i>	62	52	1355535	2726884	<i>Thalassiosira oceanica</i>	48312	84	3960	49405
<i>Fragilaria radians</i>	3892	98	100875	1348726	<i>Thalassiosira ordinaria</i>	55674	125	3527	37037
<i>Fragilariopsis cylindrus</i>	1028	69	190760	1265703	<i>Thalassiosira pacifica</i>	116337	221	2749	67851
<i>Licmophora abbreviata</i>	7734	29	6984	97466	<i>Thalassiosira profunda</i>	13246	58	7689	137826
<i>Mediolabrus comicus</i>	9873	36	6507	98331					

Table 5. Assembly statistics according to seqstats (<https://github.com/clwgg/seqstats>) of Bacillariophyta genome assemblies annotated with BRAKER. #Seq = number of sequences, #Nuc (Mbp) = total number of nucleotides in megabase pairs, N50, Longest Seq = longest sequence.

OrthoFinder’s analysis is based on the construction of gene trees, allowing for the classification of orthologous and paralogous relationships. The gene trees can be summarized in species trees, which are particularly useful for identifying variable rates of sequence evolution (through branch lengths) and the order in which sequences diverged (tree topology). The resulting species tree for Bacillariophyta gene sets, including both novel and previously annotated genomes from the International Nucleotide Sequence Database Collaboration (INSDC), is shown in Fig. 4.

The species were grouped into major sub-lineages: Coscinodiscophyceae, Mediophyceae, Fragilariophyceae, and Bacillariophyceae. Consistent with findings from earlier phylogenetic research^{69,70}, diatom sub-lineages are not recovered as a monophyletic group: radial centrics (Coscinodiscophyceae) form a paraphyletic clade, while Mediophyceae and pennate diatoms (Fragilariophyceae and Bacillariophyceae) form separate, well-supported clades. *Chaetoceros muelleri* and *Chaetoceros tenuissimus* are often placed outside the main Coscinodiscophyceae (radial centric) clade and instead fall within the Mediophyceae, a group of polar centric diatoms^{69–71}. Mediophyceae regularly emerge as the sister group to pennate diatoms (Fragilariophyceae and Bacillariophyceae), rather than to radial centrics. *Chaetoceros* species cluster with other Mediophyceae such as *Thalassiosira*, *Biddulphia*, and *Rhizosolenia*, forming a distinct group separate from radial centrics and generally closer to pennate diatoms. This placement supports earlier morphological and phylogenetic studies^{72,73} showing that chain-forming centrics like *Chaetoceros* are more closely related to pennates than to traditional radial centrics.

In some cases, the effect of excluding contaminant and HGT candidate sequences may have been slightly too stringent, potentially leading to overfiltering. To illustrate, we provide BUSCO scores for both the original and genome assemblies and gene sets without contaminant and HGT-labeled sequences (see Supplementary Table S10).

While the PhycoCosm database includes additional annotated Bacillariophyta genomes, our workflow was specifically designed to rely on automatic querying of NCBI datasets for genome downloads. Therefore, we did not include PhycoCosm genomes in this study.

The novel annotations presented here will be valuable for studying interactions between diatoms and bacteria, particularly in the context of algal blooms that play a significant role in global carbon cycling. Given that methods for recovering full eukaryotic genomes from metagenomes are still developing, reference-based binning approaches, such as BlobTools⁷⁴ using DIAMOND, may provide a viable strategy, especially as databases like NCBI NR expand for this clade.

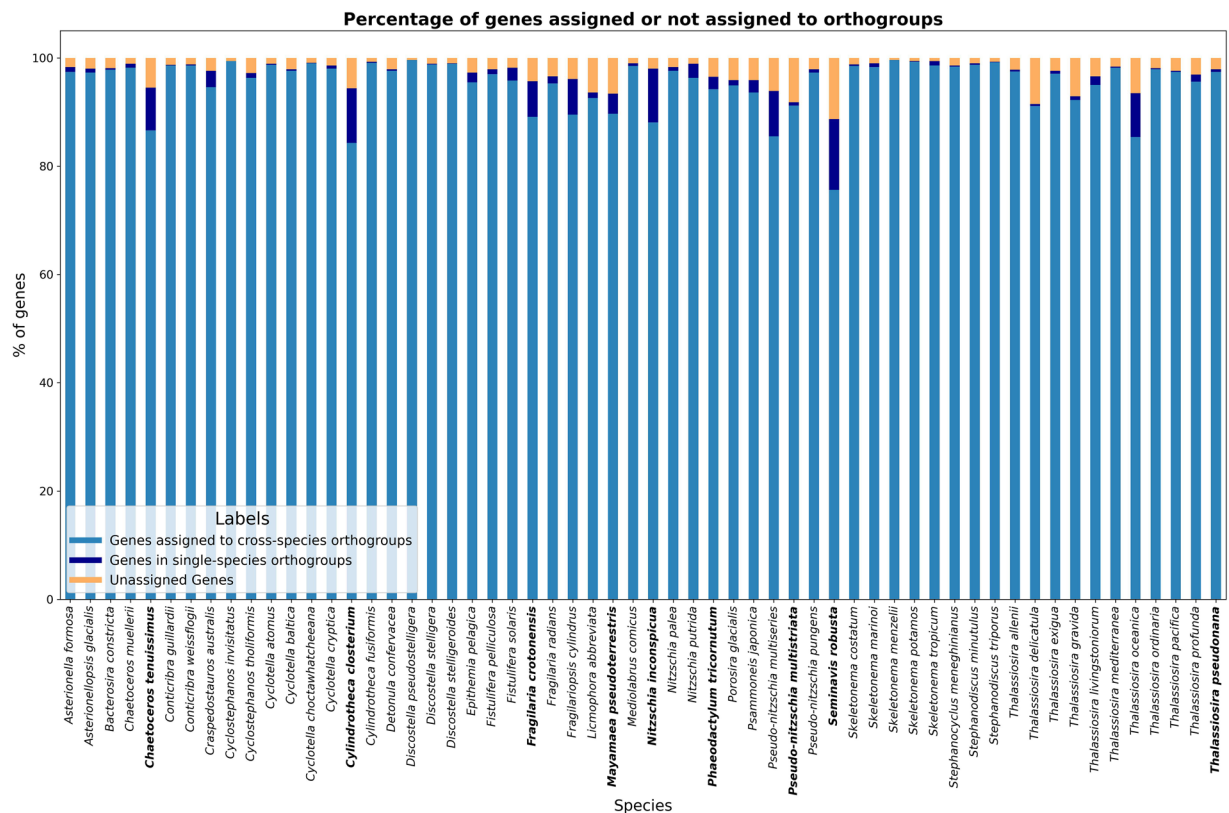


Fig. 8 Assignment rate of proteins to orthogroups. The figure shows the percentage of genes assigned to cross-species orthogroups, to single-species orthogroups (paralog only groups), or not assigned to orthogroups across different species as a stacked barplot. The light blue bars represent the genes assigned to orthogroups, while the orange bars represent the unassigned genes. The dark blue bars represent the genes assigned to single-species orthogroups. Previously annotated species are marked in bold face.

Code availability

The snakemake workflow used to generate this data set is freely available at <https://github.com/KatharinaHoff/braker-snake>. The postprocessing steps including custom scripts are freely available at https://github.com/Gaius-Augustus/Diatom_annotation_scripts. Container and software versions are listed in Supplementary Table S4.

Received: 14 October 2024; Accepted: 29 May 2025;

Published online: 11 June 2025

References

- Falciatore, A., Jaubert, M., Bouly, J.-P., Bailleul, B. & Mock, T. Diatom molecular research comes of age: model species for studying phytoplankton biology and diversity. *Plant Cell* **32**, 547–572 (2020).
- Schaum, C.-E., Buckling, A., Smirnov, N., Studholme, D. J. & Yvon-Durocher, G. Environmental fluctuations accelerate molecular evolution of thermal tolerance in a marine diatom. *Nat. Commun.* **9**, 1719 (2018).
- Sethi, D., Butler, T. O., Shuhaili, F. & Vaidyanathan, S. Diatoms for Carbon Sequestration and Bio-Based Manufacturing. *Biology (Basel)* **9** (2020).
- Jaramillo-Madrid, A. C., Ashworth, J., Fabris, M. & Ralph, P. J. Phytosterol biosynthesis and production by diatoms (Bacillariophyceae). *Phytochemistry* **163**, 46–57 (2019).
- Matsuda, Y. & Kroth, P. G. Carbon fixation in diatoms. In *The structural basis of biological energy generation* (ed. Hohmann-Marriott, M. F.) vol. 39 335–362 (Springer Netherlands, 2014).
- Li, X., Roevros, N., Dehairs, F. & Chou, L. Biological responses of the marine diatom *Chaetoceros socialis* to changing environmental conditions: A laboratory experiment. *PLoS ONE* **12**, e0188615 (2017).
- Ellegaard, M. *et al.* The fascinating diatom frustule—can it play a role for attenuation of UV radiation? *J. Appl. Phycol.* **28**, 3295–3306 (2016).
- Bayramova, E. M., Bedoshvili, Y. D. & Likhoshvay, Y. V. Molecular and cellular mechanisms of diatom response to environmental changes. *Limnology and Freshwater Biology* 20–30 (2023).
- Moreno, C. M. *et al.* Molecular physiology of Antarctic diatom natural assemblages and bloom event reveal insights into strategies contributing to their ecological success. *mSystems* **9**, e0130623 (2024).
- Di Costanzo, F., Di Dato, V. & Romano, G. Diatom-Bacteria Interactions in the Marine Environment: Complexity, Heterogeneity, and Potential for Biotechnological Applications. *Microorganisms* **11** (2023).
- Amin, S. A., Parker, M. S. & Armbrust, E. V. Interactions between diatoms and bacteria. *Microbiol. Mol. Biol. Rev.* **76**, 667–684 (2012).
- Teeling, H. *et al.* Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* **336**, 608–611 (2012).

13. Siebers, R. *et al.* Marine particle microbiomes during a spring diatom bloom contain active sulfate-reducing bacteria. *FEMS Microbiol. Ecol.* **100** (2024).
14. Branscombe, L. *et al.* Cryptic bacterial pathogens of diatoms peak during senescence of a winter diatom bloom. *New Phytol.* **241**, 1292–1307 (2024).
15. Coyne, K. J., Wang, Y. & Johnson, G. Algicidal bacteria: A review of current knowledge and applications to control harmful algal blooms. *Front. Microbiol.* **13**, 871177 (2022).
16. Grigoriev, I. V. *et al.* PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res.* **49**, D1004–D1011 (2021).
17. Lawniczak, M. K. N. *et al.* Standards recommendations for the Earth BioGenome Project. *Proc Natl Acad Sci USA* **119** (2022).
18. Plant and Algal Genomics Lab, Institut de Biologie, École Normale Supérieure. SRX18295403; GSM6736688: Return_to_Light_24h.Biorep2; *Fragilariopsis cylindrus*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR22322166> (2023).
19. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).
20. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLoS ONE* **12**, e0177459 (2017).
21. Zdobnov, E. M. *et al.* OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **49**, D389–D393 (2021).
22. Katz, K. *et al.* The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* **50**, D387–D390 (2022).
23. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
24. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10** (2021).
25. Traller, J. C. *et al.* Genome and methylome of the oleaginous diatom *Cyclotella cryptica* reveal genetic flexibility toward a high lipid phenotype. *Biotechnol. Biofuels* **9**, 258 (2016).
26. Roberts, W. R., Downey, K. M., Ruck, E. C., Traller, J. C. & Alverson, A. J. Improved Reference Genome for *Cyclotella cryptica* CCMP332, a Model for Cell Wall Morphogenesis. *Salinity Adaptation, and Lipid Production in Diatoms (Bacillariophyta)*. *G3 (Bethesda)* **10**, 2965–2974 (2020).
27. University of Gothenburg. SRX20979717: RNA-Seq of *Skeletonema marinoi*. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR25233272> (2023).
28. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457 (2020).
29. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.* **3**, lqaa108 (2021).
30. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
31. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).
32. Huang, N. & Li, H. compleasm: a faster and more accurate reimplement of BUSCO. *Bioinformatics* **39** (2023).
33. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom. Bioinform.* **2**, lqaa026 (2020).
34. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).
35. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).
36. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
37. Iwata, H. & Gotoh, O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res.* **40**, e161 (2012).
38. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
39. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr. Protoc. Bioinformatics* **65**, e57 (2019).
40. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).
41. Brůna, T., Gabriel, L. & Hoff, K. J. Navigating Eukaryotic Genome Annotation Pipelines: A Route Map to BRAKER, Galba, and TSEBRA. *arXiv* <https://doi.org/10.48550/arxiv.2403.19416> (2024).
42. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Res.* **34**, 757–768 (2024).
43. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
44. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78 (2015).
45. Hart, A. J. *et al.* EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Mol. Ecol. Resour.* **20**, 591–604 (2020).
46. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).
47. UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
48. Hernández-Plaza, A. *et al.* eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res.* **51**, D389–D394 (2023).
49. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
50. Ashburner, M. *et al.* Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
51. Gene Ontology Consortium. *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).
52. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
53. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
54. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
55. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 — approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
56. Emms, D. & Kelly, S. STAG: Species Tree Inference from All Genes. *BioRxiv* <https://doi.org/10.1101/267914> (2018).
57. Emms, D. M. & Kelly, S. STRIDE: Species Tree Root Inference from Gene Duplication Events. *Mol. Biol. Evol.* **34**, 3267–3278 (2017).
58. Nenashina, N. *et al.* OrthoFinder Results of Gene Annotations of 49 Bacillariophyta Genome Assemblies. *Zenodo* <https://doi.org/10.5281/zenodo.15380858> (2025).
59. Vuruputoor, V. S. *et al.* Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes. *Appl. Plant Sci.* **11**, e11533 (2023).
60. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
61. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).

62. Nenasheva, N. *et al.* Gene Annotations of 49 Bacillariophyta Genome Assemblies. *Zenodo* <https://doi.org/10.5281/zenodo.13745090> (2025).
63. Nevers, Y. *et al.* Quality assessment of gene repertoire annotations with OMArk. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-024-02147-w> (2024).
64. Altenhoff, A. M. *et al.* OMA orthology in 2024: improved prokaryote coverage, ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA Ecosystem. *Nucleic Acids Res.* **52**, D513–D521 (2024).
65. Rossier, V., Warwick Vesztrocy, A., Robinson-Rechavi, M. & Dessimoz, C. OMamer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches. *Bioinformatics* **37**, 2866–2873 (2021).
66. Tanaka, T. *et al.* Oil accumulation by the oleaginous diatom *Fistulifera solaris* as revealed by the genome and transcriptome. *Plant Cell* **27**, 162–176 (2015).
67. University of Sydney. SRX17002520: Diverse RNA viruses associated with diatom, eustigmatophyte, dinoflagellate and rhodophyte microalgae cultures. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR20985051> (2022).
68. Roberts, W. R., Siepielski, A. M. & Alverson, A. J. Diatom abundance in the polar oceans is predicted by genome size. *PLoS Biol.* **22**, e3002733 (2024).
69. Parks, M. B., Wickett, N. J. & Alverson, A. J. Signal, uncertainty, and conflict in phylogenomic data for a diverse lineage of microbial eukaryotes (diatoms, bacillariophyta). *Mol. Biol. Evol.* **35**, 80–93 (2018).
70. Theriot, E. C. A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. *Plant Ecol. Evol.* **143**, 278–296 (2010).
71. Xu, Q., Cui, Z. & Chen, N. Comparative analysis of chloroplast genomes of seven chaetoceros species revealed variation hotspots and speciation time. *Front. Microbiol.* **12**, 742554 (2021).
72. Medlin, L. K. & Kaczmarek, I. Evolution of the diatoms: V. Morphological and cytological support for the major clades and a taxonomic revision. *Phycologia* **43**, 245–270 (2004).
73. Nakov, T., Beaulieu, J. M. & Alverson, A. J. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). *New Phytol.* **219**, 462–473 (2018).
74. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. *F1000Res.* **6**, 1287 (2017).
75. CNRS-AMU. Genome assembly ASM225602v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_002256025.1/ (2017).
76. CNRS-AMU. SRX2949863: RNA-seq of *Asterionella formosa* BG1 in mixed culture. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR5749612> (2017).
77. Oliver, A. *et al.* Diploid genomic architecture of *Nitzschia inconspicua*, an elite biomass production diatom. *Sci. Rep.* **11**, 15592 (2021).
78. University of California, San Diego. Genome assembly GAI293_CANU_175m_combined - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_019154785.2/ (2021).
79. Shibl, A. A. *et al.* Diatom modulation of select bacteria through use of two unique secondary metabolites. *Proc Natl Acad Sci USA* **117**, 27445–27455 (2020).
80. NYUAD. Genome of *Asterionellopsis glacialis* - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_014885115.2/ (2020).
81. Picardie Jules Verne University. Genome assembly ASM1959358v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_019593585.1/ (2021).
82. OKAYAMA. DRX185892: Illumina HiSeq. 2500 paired end sequencing of SAMD00191360. <https://www.ncbi.nlm.nih.gov/sra/?term=DRR195451> (2021).
83. EcoLab. SRX4900118: RNA-seq of *Nitzschia palea* exposed to high FLG concentration. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR8072139> (2019).
84. EcoLab. SRX4900117: RNA-seq of *Nitzschia palea* exposed to high CNT concentration. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR8072140> (2019).
85. EcoLab. SRX4900116: RNA-seq of *Nitzschia palea* control condition. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR8072141> (2019).
86. EcoLab. SRX4900114: RNA-seq of *Nitzschia palea* exposed to FLG shading. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR8072143> (2019).
87. EcoLab. SRX4900113: RNA-seq of *Nitzschia palea* exposed to low CNT concentration. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR8072144> (2019).
88. Roberts, W. R., Ruck, E. C., Downey, K. M., Pinseel, E. & Alverson, A. J. Resolving Marine–Freshwater Transitions by Diatoms Through a Fog of Gene Tree Discordance. *Syst. Biol.* **72**, 984–997 (2023).
89. University of Arkansas. Genome assembly ASM3735623v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_037356235.1/ (2024).
90. University of Arkansas. SRX14833690: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733610> (2023).
91. Kamikawa, R. *et al.* Genome evolution of a nonparasitic secondary heterotroph, the diatom *Nitzschia putrida*. *Sci. Adv.* **8**, eabi5075 (2022).
92. Graduate School of Agriculture, Kyoto University. Genome assembly Nputr_1.0 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_016586335.1/ (2020).
93. Davison, H. R., Hurst, G. D. D. & Siozios, S. “Candidatus Megaira” are diverse symbionts of algae and ciliates with the potential for defensive symbiosis. *Microb. Genom.* **9** (2023).
94. New York University Abu Dhabi, UAE. Genome assembly ASM1969354v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_019693545.1/ (2021).
95. University of Technology Sydney. SRX5445774: mRNA-seq of *Chaetoceros muelleri* exposed to Flu at t = 72hrs, replicate 2. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR8647946> (2021).
96. University of Technology Sydney. SRX5445772: mRNA-seq of *Chaetoceros muelleri* exposed to control at t = 48hrs, replicate ctl. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR8647948> (2021).
97. University of Technology Sydney. SRX5445771: mRNA-seq of *Chaetoceros muelleri* exposed to Fen at t = 48hrs, replicate 3. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR8647949> (2021).
98. University of Technology Sydney. SRX5445770: mRNA-seq of *Chaetoceros muelleri* exposed to Fen at t = 48hrs, replicate 2. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR8647950> (2021).
99. University of Technology Sydney. SRX5445741: mRNA-seq of *Chaetoceros muelleri* exposed to Ro at t = 24hrs, replicate 3. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR8647979> (2020).
100. University of Technology Sydney. SRX5445737: mRNA-seq of *Chaetoceros muelleri* exposed to Flu at t = 72hrs, replicate 3. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR8647983> (2021).
101. Bowler, C. *et al.* The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239–244 (2008).
102. Diatom Consortium. Genome assembly ASM15095v2 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000150955.2/ (2008).
103. Hongo, Y. *et al.* The genome of the diatom *Chaetoceros tenuissimus* carries an ancient integrated fragment of an extant virus. *Sci. Rep.* **11**, 22877 (2021).

104. Environmental Genomics Group, Research Center for Bioinformatics and Biosciences, National Research Institute of Fisheries Science, Japan Fisheries Research and Education Agency. Genome assembly Cten210_1.0 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_021927905.1/ (2021).
105. University of Arkansas. Genome assembly ASM3693939v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939395.1/ (2024).
106. University of Arkansas. SRX14833699: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733601> (2023).
107. University of Arkansas. Genome assembly ASM3693933v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939335.1/ (2024).
108. University of Arkansas. SRX14833779: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733581> (2023).
109. Chicago Botanic Garden. Genome assembly ASM863298v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_008632985.1/ (2019).
110. University of Arkansas. Genome assembly ASM3694002v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036940025.1/ (2024).
111. CNRS. SRX1078771: Tw reference transcriptome. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR2085009> (2016).
112. Turk Dermastia, T., Dall'Ara, S., Dolenc, J. & Mozetič, P. Toxicity of the Diatom Genus *Pseudo-nitzschia* (Bacillariophyceae): Insights from Toxicity Tests and Genetic Screening in the Northern Adriatic Sea. *Toxins (Basel)* **14** (2022).
113. Basu, S. *et al.* Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. *New Phytol.* **215**, 140–156 (2017).
114. Stazione Zoologica Anton Dohrn. Genome assembly ASM90066040v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_900660405.1/ (2019).
115. Zackova Suchanova, J. *et al.* Diatom adhesive trail proteins acquired by horizontal gene transfer from bacteria serve as primers for marine biofilm formation. *New Phytol.* **240**, 770–783 (2023).
116. B CUBE. Genome assembly DD_uo2022 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_026770025.1/ (2022).
117. Institute of Oceanology, Chinese Academy of Sciences. Genome assembly ASM3735574v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_037355745.1/ (2024).
118. Assembly: GCA_900660405.1 - ENA - EBI. https://www.ebi.ac.uk/ena/browser/view/GCA_900660405.1 (2023).
119. J. Crag Venter Institute. SRX3899616: P. multiseriis grown at 20uM phosphate, 400ppm pCO₂, rep C (totalRNA). <https://www.ncbi.nlm.nih.gov/sra/?term=SRR6956622> (2018).
120. J. Crag Venter Institute. SRX3899615: P. multiseriis grown at 20uM phosphate, 400ppm pCO₂, rep B (totalRNA). <https://www.ncbi.nlm.nih.gov/sra/?term=SRR6956623> (2018).
121. J. Crag Venter Institute. SRX3899610: P. multiseriis grown at 0.5uM phosphate, 220ppm pCO₂, rep A (totalRNA). <https://www.ncbi.nlm.nih.gov/sra/?term=SRR6956628> (2018).
122. J. Crag Venter Institute. SRX3899605: P. multiseriis grown at 0.5uM phosphate, 730ppm pCO₂, rep C (polyA). <https://www.ncbi.nlm.nih.gov/sra/?term=SRR6956633> (2018).
123. J. Crag Venter Institute. SRX3899601: P. multiseriis grown at 20uM phosphate, 400ppm pCO₂, rep A (polyA). <https://www.ncbi.nlm.nih.gov/sra/?term=SRR6956637> (2018).
124. J. Crag Venter Institute. SRX3899600: P. multiseriis grown at 20uM phosphate, 400ppm pCO₂, rep B (polyA). <https://www.ncbi.nlm.nih.gov/sra/?term=SRR6956638> (2018).
125. University of Arkansas. Genome assembly ASM3693967v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939675.1/ (2024).
126. University of Arkansas. SRX14833793: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733507> (2023).
127. Institute of Oceanology, Chinese Academy of Sciences. Genome assembly ASM3735585v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_037355855.1/ (2024).
128. University of Arkansas. Genome assembly ASM3693997v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939975.1/ (2024).
129. Osuna-Cruz, C. M. *et al.* The *Seminavis robusta* genome provides insights into the evolutionary adaptations of benthic diatoms. *Nat. Commun.* **11**, 3320 (2020).
130. Center for Plant Systems Biology, VIB-Ugent. Genome assembly Semro_V1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_903772945.1/ (2022).
131. University of Arkansas. Genome assembly ASM3693993v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939935.1/ (2024).
132. University of Arkansas. SRX14833788: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733512> (2023).
133. University of Arkansas. SRX14833751: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733549> (2023).
134. University of Arkansas. SRX14833688: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733612> (2023).
135. Sorokina, M. *et al.* Draft genome assembly and sequencing dataset of the marine diatom *Skeletonema cf. costatum* RCC75. *Data Brief* **41**, 107931 (2022).
136. Balance of the Microverse. Genome assembly FSU_Scostatum_1.0 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_018806925.1/ (2021).
137. NAGAHAMA_BIO. DRX141855: Illumina HiSeq. 2000 paired end sequencing of SAMD00137718. <https://www.ncbi.nlm.nih.gov/sra/?term=DRR151104> (2018).
138. NAGAHAMA_BIO. DRX141856: Illumina HiSeq. 2000 paired end sequencing of SAMD00137719. <https://www.ncbi.nlm.nih.gov/sra/?term=DRR151105> (2018).
139. NAGAHAMA_BIO. DRX141857: Illumina HiSeq. 2000 paired end sequencing of SAMD00137720. <https://www.ncbi.nlm.nih.gov/sra/?term=DRR151106> (2018).
140. NAGAHAMA_BIO. DRX141858: Illumina HiSeq. 2000 paired end sequencing of SAMD00137721. <https://www.ncbi.nlm.nih.gov/sra/?term=DRR151107> (2018).
141. NAGAHAMA_BIO. DRX141859: Illumina HiSeq. 2000 paired end sequencing of SAMD00137722. <https://www.ncbi.nlm.nih.gov/sra/?term=DRR151108> (2018).
142. NAGAHAMA_BIO. DRX141860: Illumina HiSeq. 2000 paired end sequencing of SAMD00137723. <https://www.ncbi.nlm.nih.gov/sra/?term=DRR151109> (2018).
143. University of Arkansas. Genome assembly ASM3693963v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939635.1/ (2024).
144. University of Arkansas. SRX14833803: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733497> (2023).

145. Liu, S., Xu, Q. & Chen, N. Expansion of photoreception-related gene families may drive ecological adaptation of the dominant diatom species *Skeletonema marinoi*. *Sci. Total Environ.* **897**, 165384 (2023).
146. Institute of Oceanology, Chinese Academy of Sciences. Genome assembly ASM3054422v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_030544225.1/ (2023).
147. University of Gothenburg. Genome assembly GU_Smar_1.1.8 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_030871285.1/ (2023).
148. University of Gothenburg. SRX20979719: RNA-Seq of *Skeletonema marinoi*. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR25233270> (2023).
149. University of Gothenburg. SRX20979712: RNA-Seq of *Skeletonema marinoi*. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR25233277> (2023).
150. University of Gothenburg. SRX20979710: RNA-Seq of *Skeletonema marinoi*. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR25233279> (2023).
151. University of Gothenburg. SRX20979704: RNA-Seq of *Skeletonema marinoi*. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR25233285> (2023).
152. University of Arkansas. Genome assembly ASM3693985v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939855.1/ (2024).
153. University of Arkansas. SRX14833746: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733553> (2023).
154. University of Arkansas. Genome assembly ASM3694000v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036940005.1/ (2024).
155. University of Arkansas. SRX14833727: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733572> (2023).
156. University of Arkansas. Genome assembly ASM1318728v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_013187285.1/ (2020).
157. Lewis Lab, Biological Sciences, University of Arkansas. SRX15830763: GSM6261389: CCMP332_1hr_0ppt_rep3; *Cyclotella cryptica*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR19991157> (2022).
158. Lewis Lab, Biological Sciences, University of Arkansas. SRX15830764: GSM6261390: CCMP332_2hr_0ppt_rep1; *Cyclotella cryptica*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR19991158> (2022).
159. Lewis Lab, Biological Sciences, University of Arkansas. SRX15830768: GSM6261394: CCMP332_30m_0ppt_rep2; *Cyclotella cryptica*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR19991160> (2022).
160. Lewis Lab, Biological Sciences, University of Arkansas. SRX15830773: GSM6261399: CCMP332_8hr_0ppt_rep1; *Cyclotella cryptica*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR19991165> (2022).
161. Lewis Lab, Biological Sciences, University of Arkansas. SRX15830778: GSM6261404: CCMP332_Control_rep3; *Cyclotella cryptica*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR19991170> (2022).
162. University of Arkansas. Genome assembly ASM3694010v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036940105.1/ (2024).
163. University of Arkansas. SRX14833771: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733529> (2023).
164. Audoor, S. *et al.* Transcriptional chronology reveals conserved genes involved in pennate diatom sexual reproduction. *Mol. Ecol.* **33**, e17320 (2024).
165. Universiteit Gent. Genome assembly Cyc-CA1.15-REFERENCE-ANNOTATED - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_933822405.4/ (2023).
166. Institute of Oceanology, Chinese Academy of Sciences. Genome assembly ASM3717862v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_037178625.1/ (2024).
167. University of Arkansas. SRX14833772: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733528> (2023).
168. New York University Abu Dhabi, UAE. Genome assembly ASM1969352v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_019693525.1/ (2021).
169. University of Arkansas. Genome assembly ASM3694004v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036940045.1/ (2024).
170. University of Arkansas. SRX14833722: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733577> (2023).
171. University of Arkansas. Genome assembly ASM3693941v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939415.1/ (2024).
172. University of Arkansas. SRX14833776: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733524> (2023).
173. University of Arkansas. SRX14833691: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733609> (2023).
174. University of Arkansas. Genome assembly ASM3693943v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939435.1/ (2024).
175. University of Arkansas. SRX14833769: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733531> (2023).
176. University of Arkansas. Genome assembly ASM3694008v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036940085.1/ (2024).
177. University of Arkansas. SRX14833794: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733506> (2023).
178. University of Arkansas. SRX14833734: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733565> (2023).
179. University of Arkansas. Genome assembly ASM3693975v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939755.1/ (2024).
180. University of Arkansas. SRX14833796: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733504> (2023).
181. University of Arkansas. Genome assembly ASM3693973v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939735.1/ (2024).
182. University of Arkansas. SRX14833791: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733509> (2023).
183. University of Arkansas. Genome assembly ASM3693965v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939655.1/ (2024).
184. University of Arkansas. SRX14833806: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733494> (2023).

185. University of Arkansas. Genome assembly ASM3693955v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_03693955.1/ (2024).
186. University of Arkansas. SRX14833759: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733541> (2023).
187. University of Arkansas. Genome assembly ASM3693983v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_03693983.1/ (2024).
188. University of Arkansas. SRX14833781: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733519> (2023).
189. University of Arkansas. SRX14833705: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733595> (2023).
190. WELLCOME SANGER INSTITUTE. Genome assembly uoEpiScrs1.2. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_946965045.2/ (2023).
191. University of Arkansas. Genome assembly ASM3693989v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_03693989.1/ (2024).
192. University of Arkansas. SRX14833737: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733562> (2023).
193. Maeda, Y. *et al.* Chromosome-Scale Genome Assembly of the Marine Oleaginous Diatom *Fistulifera solaris*. *Mar. Biotechnol.* **24**, 788–800 (2022).
194. Institute of Engineering, Division of Biotechnology and Life Science, Tokyo University of Agriculture and Technology. Genome assembly Fpelliculosa_ONT_v02 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_026008555.1/ (2022).
195. University of Arkansas. Genome assembly ASM3735621v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_03735621.1/ (2024).
196. University of Arkansas. SRX14833780: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733520> (2023).
197. Institute of Engineering, Division of Biotechnology and Life Science, Tokyo University of Agriculture and Technology. Genome assembly Fsolaris_ONT_v03reference - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_030295235.1/ (2022).
198. Tokyo University of Agriculture and Technology. Genome assembly Fsol_1.0 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_002217885.1/ (2017).
199. University of Arkansas. Genome assembly ASM3693959v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_03693959.1/ (2024).
200. Zepernick, B. N., Truchon, A. R., Gann, E. R. & Wilhelm, S. W. Draft Genome Sequence of the Freshwater Diatom *Fragilaria crotonensis* SAG 28.96. *Microbiol. Resour. Announc.* **11**, e0028922 (2022).
201. Zepernick, B. N. *et al.* Morphological, physiological, and transcriptional responses of the freshwater diatom *Fragilaria crotonensis* to elevated pH conditions. *Front. Microbiol.* **13**, 1044464 (2022).
202. University of Tennessee Knoxville. Genome assembly ASM2292589v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_02292589.1/ (2022).
203. University of Arkansas. SRX14833756: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733544> (2023).
204. University of Arkansas. SRX14833742: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733557> (2023).
205. Galachyants, Y. P. *et al.* Sequencing of the complete genome of an araphid pennate diatom *Synedra acus* subsp. *radians* from Lake Baikal. *Dokl. Biochem. Biophys.* **461**, 84–88 (2015).
206. LIMNOLOGICAL INSTITUTE SB RAS. Genome assembly sac1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_900642245.1/ (2019).
207. Limnological Institute SB RAS. SRX4509102: *Fragilaria radians* RNA-seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR7646146> (2019).
208. Limnological Institute SB RAS. SRX4509101: *Fragilaria radians* RNA-seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR7646147> (2019).
209. Limnological Institute SB RAS. SRX4509100: *Fragilaria radians* RNA-seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR7646148> (2019).
210. Limnological Institute SB RAS. SRX4509099: *Fragilaria radians* RNA-seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR7646149> (2019).
211. Limnological Institute SB RAS. SRX4509098: *Fragilaria radians* RNA-seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR7646150> (2019).
212. New York University Abu Dhabi, UAE. Genome assembly ASM1969357v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_01969357.1/ (2021).
213. DFG Cluster of Excellence “Future Ocean.” Genome assembly ThaOc_1.0 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_000296195.2/ (2012).
214. NCBI (GEO). SRX13155147: GSM5692981: LCuA rep2; *Thalassiosira oceanica* CCMP1005; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR16963171> (2021).
215. NCBI (GEO). SRX13155149: GSM5692982: LCuA rep3; *Thalassiosira oceanica* CCMP1005; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR16963172> (2021).
216. Armbrust, Oceanography, University of Washington. SRX18196720: GSM6720087: To Hydrogren peroxide N2; *Thalassiosira oceanica*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR22218830> (2022).
217. Armbrust, Oceanography, University of Washington. SRX18196719: GSM6720086: To Hydrogren peroxide N1; *Thalassiosira oceanica*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR22218832> (2022).
218. Armbrust, Oceanography, University of Washington. SRX18196718: GSM6720085: To Control N3; *Thalassiosira oceanica*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR22218838> (2022).
219. Paajanen, P. *et al.* Building a locally diploid genome and transcriptome of the diatom *Fragilariopsis cylindrus*. *Sci. Data* **4**, 170149 (2017).
220. TGAC. Genome assembly Fc_falcon_quiver_polished - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_900095095.1/ (2016).
221. JGI-PGF. Genome assembly Fracy1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_001750085.1/ (2016).
222. Plant and Algal Genomics Lab, Institut de Biologie, École Normale Supérieure. SRX18295413: GSM6736698: Return_to_Light_5days.Biorep3; *Fragilariopsis cylindrus*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR22322156> (2023).
223. Plant and Algal Genomics Lab, Institut de Biologie, École Normale Supérieure. SRX18295408: GSM6736693: Return_to_Light_3days.Biorep1; *Fragilariopsis cylindrus*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR22322161> (2023).
224. Plant and Algal Genomics Lab, Institut de Biologie, École Normale Supérieure. SRX18295407: GSM6736692: Return_to_Light_48h.Biorep3; *Fragilariopsis cylindrus*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR22322162> (2023).
225. Plant and Algal Genomics Lab, Institut de Biologie, École Normale Supérieure. SRX18295399: GSM6736684: Return_to_Light_12h.Biorep1; *Fragilariopsis cylindrus*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR22322170> (2023).
226. Plant and Algal Genomics Lab, Institut de Biologie, École Normale Supérieure. SRX18295398: GSM6736683: Return_to_Light_6h.Biorep3; *Fragilariopsis cylindrus*; RNA-Seq. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR22322171> (2023).
227. University of Arkansas. Genome assembly ASM3693969v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_03693969.1/ (2024).
228. University of Arkansas. SRX14833798: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733502> (2023).

229. IBENS. Genome assembly CCMP470 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_900291995.1/ (2018).
230. University of Arkansas. Genome assembly ASM3693987v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939875.1/ (2024).
231. University of Arkansas. SRX14833748: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733551> (2023).
232. University of Arkansas. SRX14833721: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733578> (2023).
233. Suzuki, S. *et al.* Rapid transcriptomic and physiological changes in the freshwater pennate diatom *Mayamaea pseudoterrestris* in response to copper exposure. *DNA Res.* **29** (2022).
234. National Institute for Environmental Studies. Genome assembly Mpse_1.0 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_027923505.1/ (2022).
235. Armbrust, E. V. *et al.* The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79–86 (2004).
236. Diatom Consortium Genome assembly ASM14940v2 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000149405.2/ (2009).
237. University of Arkansas. Genome assembly ASM3694012v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036940125.1/ (2014).
238. Universidad Mayor. SRX10228342: RNA-Seq of RCC4660. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR13846812> (2021).
239. University of Arkansas. SRX14833711: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733589> (2023).
240. University of Arkansas. Genome assembly ASM3693935v1 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036939355.1/ (2024).
241. University of Arkansas. SRX14833695: Phylogenomics of the Thalassiosirales. <https://www.ncbi.nlm.nih.gov/sra/?term=SRR18733605> (2023).
242. Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* **55**, 539–552 (2006).

Acknowledgements

We thank Stepan Saenko for discussing the snakemake workflow for structural genome annotation. NN was funded by the German Research Foundation (DFG) in the framework of the research unit FOR2406 “Proteogenomics of Marine Polysaccharide Utilization (POMPU)” by a grant to KJH (277249973). CW and AH, and their software, were supported by a grant to JLW (NSF DBI 1943371).

Author contributions

K.J.H. and M.M.B. conceptualised the study. C.P. and K.J.H. designed and implemented the software. C.P. executed the snakemake workflow for structural genome annotation under supervision of K.J.H. C.W. and A.H. performed functional annotation under supervision of J.L.W. C.W. performed ploidy analysis under supervision of J.L.W. A.H. performed horizontal gene transfer analysis under supervision of J.L.W. N.N. performed orthology analysis under supervision of K.J.H. N.N. wrote the first draft of the manuscript. All authors contributed to the manuscript in writing. All authors read and approved of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-025-05306-z>.

Correspondence and requests for materials should be addressed to K.J.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025