Transcriptional chronology reveals conserved genes involved in pennate diatom sexual reproduction

Sien Audoor, Gust Bilcke, Katerina Pargana, Darja Belišová, Sander Thierens, Michiel van Bel, Lieven Sterck, Nadine Rijsdijk, Rossella Annunziata, Maria Immacolata Ferrante, et al.

Transcriptional chronology reveals conserved genes involved in pennate diatom sexual reproduction

Sien Audoor[1,a]*, Gust Bilcke[1,2,3]*, Katerina Pargana[1], Darja Belišová[1,2,3], Sander Thierens[2,3], Michiel Van Bel[2,3], Lieven Sterck[2,3], Nadine Rijsdijk[1,2,3], Rossella Annunziata[4], Maria Immacolata Ferrante[4,5], Klaas Vandepoele[2,3,6]** and Wim Vyverman[1]**§

Author's institution(s)/affiliation(s):

[1] Laboratory of Protistology and Aquatic Ecology, Department of Biology, University Ghent, Krijgslaan 281 S8, 9000 Ghent, Belgium

[2] VIB Center for Plant Systems Biology, Technologiepark 71, 9052 Ghent, Belgium

[3] Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 71, 9052 Ghent, Belgium

[4] Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy

[5] Associate to the National Institute of Oceanography and Applied Geophysics, 34151 Trieste, Italy

[6] VIB Center for AI & Computational Biology, Technologiepark 71, 9052 Ghent, Belgium

*Share first authorship

**Share senior authorship

§Corresponding author: wim.vyverman@ugent.be

[a]Present address: UMR7141 - Biologie du Chloroplaste et Perception de la Lumière chez les Microalgues, Institut de Biologie Physico-Chimique (IBPC), Centre National de la Recherche Scientifique (CNRS), Sorbonne Université, 13 Rue Pierre et Marie Curie, 75005 Paris, France

Abstract

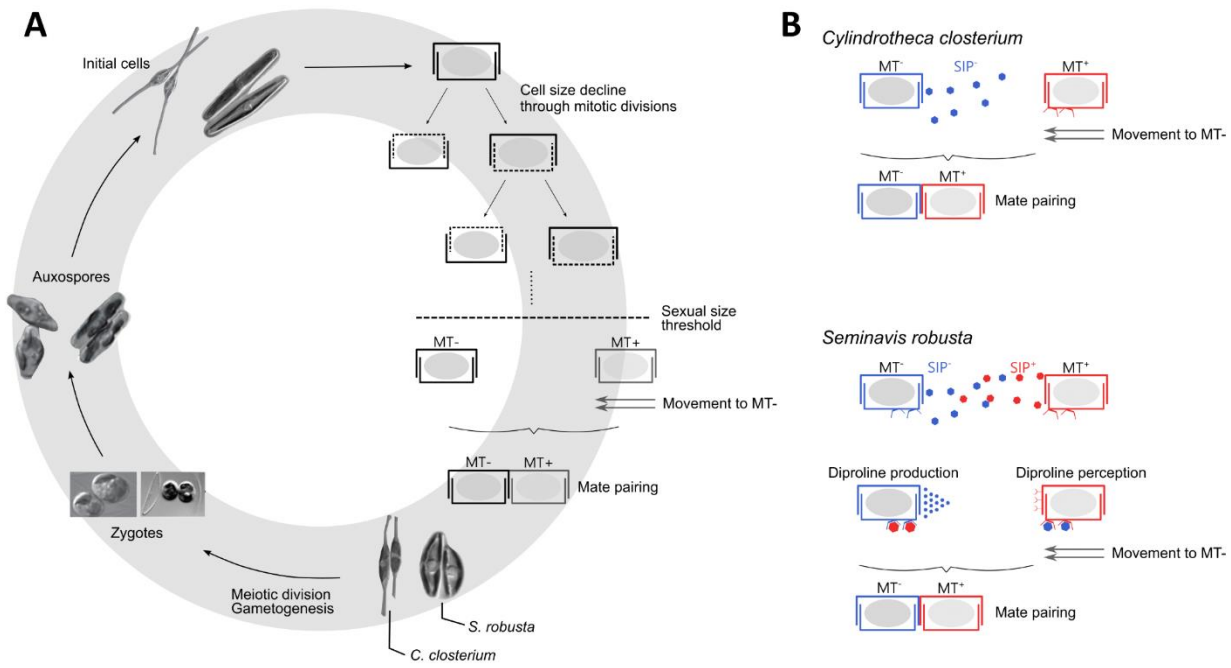
Sexual reproduction is a major driver of adaptation and speciation in eukaryotes. In diatoms, siliceous microalgae with a unique cell size reduction-restitution life cycle and among the world's most prolific primary producers, sex also acts as the main mechanism for cell size restoration through the formation of an expanding auxospore. However, the molecular regulators of the different stages of sexual reproduction and size restoration are poorly explored. Here, we combined RNA sequencing with the assembly of a 55 Mbp reference genome for *Cylindrotheca closterium* to identify patterns of gene expression during different stages of sexual reproduction. These were compared with a corresponding transcriptomic time series of *Seminavis robusta* to assess the degree of expression conservation. Integrative orthology analysis revealed 319 one-to-one orthologs that are differentially expressed during sex in both species, among which 56 genes consistently upregulated during cell pairing and gametogenesis, and 11 genes induced when auxospores are present. Several early, sex-specific transcription factors and B-type cyclins were also upregulated during sex in other pennate and centric diatoms, pointing towards a conserved core regulatory machinery for meiosis and gametogenesis across diatoms. Furthermore, we find molecular evidence that the pheromone-induced cell cycle arrest is short-lived in benthic diatoms, which may be linked to their active mode of mate finding through gliding. Finally, we exploit the temporal resolution of our comparative analysis to report the first marker genes for auxospore identity called AAE1-3 ("Auxospore-Associated Expression"). Altogether, we introduce a multi-species model of the transcriptional dynamics during size restoration in diatoms and highlight conserved gene expression dynamics during different stages of sexual reproduction.

Keywords

Diatoms; Sexual reproduction; Genomics; Transcriptomics; *Cylindrotheca closterium*; *Seminavis robusta*

Introduction

Sexual reproduction is a crucial event in the life cycle of most eukaryotes, with a shared machinery for meiotic recombination dating back to at least their last common ancestor (Goodenough & Heitman, 2014). While the importance of sexual reproduction for adaptation and speciation is widely appreciated (Mayr, 1942; West-Eberhard, 1979), it is also a key mechanism involved in cell size control in diatoms, a species-rich group of unicellular algae characterised by a siliceous cell wall (Mann & Vanormelingen, 2013). This distinctive, rigid frustule forces daughter cells to be formed within the parental cells during mitotic divisions, as new valves are synthesised *de novo* in the existing parental valves. This leads to an average cell size decrease in a natural population, a phenomenon known as the MacDonald-Pfitzer rule (Macdonald, 1869; Pfitzer, 1871), eventually causing cell death as cells become too small to survive (Chepurnov et al., 2004; Lewis, 1984). Although vegetative cell enlargement has been reported for several species (Kaczmarska et al., 2022), most species rely on sexual reproduction to restore their original cell size (Chepurnov et al., 2004; Lewis, 1984). When cells fall below a sexual size threshold (SST), they are able to produce gametes which fuse to form a zygote that expands into an auxospore (**Figure 1A**). This cell type is unique to diatoms and forms an initial cell once maximal cell size is reached (Chepurnov et al., 2004; Kaczmarska et al., 2013; Kaczmarska & Ehrman, 2021).



80

81 **Figure 1: The life cycle of benthic pennate diatoms. (A)** For the majority of their life, diatoms reproduce
82 asexually. However, each round of mitotic divisions leads to a cell size decline as a consequence of the
83 diatom's rigid silica frustule that consists of two overlapping valves of unequal size. During mitosis, each
84 daughter cell inherits one parental valve within which a new valve is synthesised. As a result, one of the
85 daughter cells is always smaller than its parent, leading to a gradual cell size reduction on a population
86 level. To avoid cell death by extreme miniaturisation, most diatoms will reproduce sexually to restore
87 original cell size. *S. robusta* or *C. closterium* cells start producing pheromones once the sexual size
88 threshold (SST) is reached, resulting in the active movement of mating type plus (MT+) cells towards
89 compatible mating type minus (MT-) cells. When partners have successfully paired, gametogenesis takes
90 place ensuring two gametes are formed per parent. Compatible gametes fuse into zygotes, which start
91 expanding and transform into auxospores. Expansion lasts until maximal cell size is reached and initial
92 cells are released, ready to start a new cycle of mitotic divisions. **(B)** Below the SST, *C. closterium* MT-
93 cells produce an attraction pheromone guiding MT+ cells and facilitating movement towards their partner.
94 For *S. robusta*, both MT+ and MT- cells produce a sex inducing pheromone (SIP). Upon SIP+ perception
95 MT- cells will produce a second pheromone, the attraction pheromone diproline to which MT+ cells become
96 sensitive when SIP- is perceived. Consequently, MT+ cells move towards MT- cells along the diproline
97 gradient to form a mating pair.

98

99 While cell-size restoration through auxospore expansion is broadly conserved among diatoms,
100 a great diversity of strategies for mate finding and gametogenesis has evolved, mainly coinciding
101 with the major phylogenetic groups. Most centric diatoms are oogamous and possess a
102 homothallic mating system, in which a single clonal culture gives rise to both egg and sperm cells
103 (Chepurnov et al., 2004; Von Stosch, 1950). In contrast, pennate diatoms are mostly

heterothallic, i.e. two genetically different mating types are required for sexual reproduction to occur. Despite being the evolutionary youngest clade (Kooistra et al., 2007), pennate diatoms are the most species-rich, which has been attributed to the development of a raphe - a longitudinal opening within the cell wall that facilitates active gliding through the excretion of extracellular polymeric substances (EPS) - that further subdivides the pennates in araphids (no raphe) and raphids (Benoiston et al., 2017; Chepurnov et al., 2004; Round et al., 1990). The evolution of gliding motility in raphid diatoms enabled a pivotal change during sexual reproduction: whereas the gametes encounter each other in centric and araphid pennate diatoms, gametangia actively glide towards a partner and pair up in benthic raphid diatoms (Chepurnov & Mann, 2004).

In all species investigated so far, mate finding of raphid diatoms is orchestrated by sex pheromones. Common among species, gametangia produce sex inducing pheromones (SIPs) to induce a G1 phase cell cycle arrest in the other mating type, thus increasing the time window for the formation of a stable mating pair before committing to meiosis (**Figure 1B**) (Annunziata et al., 2022; Basu et al., 2017; Bilcke et al., 2020; Klapper et al., 2021; Moeys et al., 2016). While SIPs are produced by both sexes in *Seminavis robusta*, the cell cycle arrest in the benthic *Cylindrotheca closterium* appears to be asymmetric, with only the mating type minus (MT-) releasing a cytostatic SIP (Klapper et al., 2021). Both *C. closterium* and *S. robusta* produce attraction pheromones to direct the movement of MT+ towards MT- (**Figure 1B**) (Gillard et al., 2008; Klapper et al., 2021). In contrast, the planktonic raphid diatom *Pseudo-nitzschia multistriata* does not appear to produce any attraction pheromone, but rather seems to rely on passive pairing in the water column (Font-Muñoz et al., 2019; Marotta et al., 2022). Finally, in all of the above species, gametogenesis occurs after a stable mating pair has been formed,

suggesting that an unknown local signal is exchanged between paired cells (Klapper et al., 2021; Moeys et al., 2016; Scalco et al., 2016; Vanormelingen et al., 2013).

Although the different stages of the life cycle of diatoms are well-described, knowledge about the genes involved in the progression during different cell stages remains scarce (Bilcke et al., 2022). Indeed, multiple conserved Sex-Inducing Genes (SIGs) have been identified (Armbrust, 1999; Ferrante et al., 2019), yet their function remains unclear. Furthermore, while the biosynthetic pathway for the attraction pheromone diproline in the benthic *S. robusta* and the molecular hallmarks of a cell cycle arrest in *S. robusta* and *P. multistriata* have been established, gene expression patterns associated with mate pairing and gametogenesis have not been resolved (Annunziata et al., 2022; Basu et al., 2017; Bilcke et al., 2020; Marotta et al., 2022; Moeys et al., 2016). Finally, despite auxospores being the hallmark of the diatom life cycle, and being covered by particular, species-specific cell wall elements, the genes responsible for auxospore formation remain unknown (Bilcke et al., 2022).

Here, we aim to identify conserved genes involved in sexual reproduction in pennate diatoms and unravel their shared chronology. To this end, we generated a matched case-control RNA-seq time series of sexual reproduction for *C. closterium* and assembled a reference genome for this species to allow homology-based comparative analyses. We subsequently integrated our transcriptomic dataset with an analogous time series from the benthic *S. robusta*, allowing us to identify successive waves of gene expression during sex. To obtain a diatom-wide view of gene expression during the sexual life cycle, we compared these results with additional sex transcriptomes from the planktonic diatoms *P. multistriata* and *Skeletonema marinoi.*

Material and methods

Selected strains and culture conditions

Strain and culture conditions

Two compatible *Cylindrotheca closterium* strains of opposite mating type, ZW2.20 (DCG 0922, MT+) and CA1.15 (DCG 0923, MT-) were obtained from the Belgian Coordinated Collection of Microorganisms (BCCM/DCG, http://bccm.belspo.be/about-us/bccm-dcg). Cultures were grown in artificial sea water (ASW) composed of Tropic Marin BIO-ACTIF sea salt (34.5 g L^{-1}; Tropic Marin, Wartenberg, Germany) and sodium bicarbonate (0.08 g L^{-1}) and supplemented with Guillard's F/2 Marine Water enrichment solution (Sigma-Aldrich). Bacterial contaminants were removed by treatment with an antibiotics mixture (500 mg L^{-1} penicillin, 500 mg L^{-1} ampicillin, 100 mg L^{-1} streptomycin and 50 mg L^{-1} gentamycin). Cell cultures were grown in large 175 cm^2 cell culture flasks (Greiner Bio One, Vilvoorde, Belgium) at 18°C under constant light in low light conditions of approximately 50 µmol photons m^2 s^{-1}.

Phylogenetic characterisation of strains

For *rbcL* amplicon sequencing, *C. closterium* strains CA1.15 and ZW2.20 were harvested by centrifugation (5 min, 3000 rpm). After DNA extraction using the QIAamp DNA Mini kit (Qiagen) with minor modifications (see Supplementary Methods), the rbcL region was amplified using the DPrbcL1 forward and DPrbcL7 reverse primers (Jones et al., 2005). A subset of previous published *C. closterium* sequences were included in this phylogenetic analysis (Stock et al., 2019), as well as one *Nitzschia sp.* outgroup sequence (GenBank: HF675116.1). All sequences were aligned using MAFFT v7.453 (Katoh & Standley, 2013) and the most appropriate model of

DNA evolution (TIM3+F+R2) was established via modelfinder (Kalyaanamoorthy et al., 2017). Lastly, the phylogenetic tree was reconstructed through IQ-TREE v1.6.12 (Nguyen et al., 2015), as detailed in Supplementary Methods.

Genome assembly of *C. closterium*

Genomic DNA extraction and sequencing

Hundred ml of dense cultures was scraped, resuspended in the medium by gentle shaking and harvested by filtration on Versapor filters (3 µm pore size, 25-mm diameter, PALL). Filters were put in Eppendorf tubes, flash frozen in liquid nitrogen and stored at -80°C until DNA extraction. High-molecular-weight (HMW) DNA for whole-genome sequencing was extracted using an optimised CTAB protocol (see Supplementary Methods), without vortexing or beating to avoid damage to the integrity of the chromosomal DNA. After extraction, all excessive pipetting and freezing/thawing steps were eliminated. DNA concentration and quality was evaluated by spectrophotometry (Nanodrop, Thermo Scientific) and agarose gel electrophoresis.

Library preparation for Illumina sequencing was performed using the Illumina MiSeq Library Prep kit and paired-end 250 bp sequencing was performed on a MiSeq v2 500 system at VIB Nucleomics core (Leuven, Belgium). Reads were obtained in two separate sequencing runs; the first run generating 6,818,277 paired-end reads and the second run 8,321,843 paired-end reads. Additionally, two MinION sequencing runs were performed. Short DNA fragments up to 10 kb were depleted using the Short Read Eliminator (SRE) XS kit (Circulomics, Baltimore, USA) and library preparation was performed using the SQK-LSK109 kit (Oxford Nanopore Technologies, Oxford, UK) following the manufacturer's protocol (1D Sequencing by Ligation). The sequencing libraries were loaded on separate R9.4 FLO-MIN106 flow cells with a MinION Mk1b device and sequenced with live base calling for 72 h. Sequences of the first run were basecalled with Flappie

v1.0.0 yielding 3,347,215 reads. Sequences of the second run were basecalled with Guppy yielding 1,685,675 reads. Quality filtering and trimming were performed for both Illumina and MINion runs, as described in Supplementary Methods.

De novo genome assembly

A two-step procedure was used to assemble the *C. closterium* genome. First, a draft assembly was constructed from long reads of the first MinION run. Next, the initial hybrid assembly was produced combining the Illumina libraries and second MinION run, while using the contigs from the first draft assembly for gap closure and repeat resolution. In detail, the trimmed and filtered reads from the first MinION run were assembled into a draft assembly using Canu v1.8 (Koren et al., 2017) with default parameters. For polishing of this first draft assembly, MinION reads were mapped back to the draft contigs using Minimap2 v2.17 (H. Li, 2018) with parameters: -ax map-ont. The assembly was then polished threefold using Racon v1.4.3 (Vaser et al., 2017) with default settings, followed by another round of mapping and polishing with Nanopolish v0.8.4 (https://github.com/jts/nanopolish) and reduction and gap closing by Redundans v0.13 (Pryszcz & Gabaldón, 2016). In a second stage, the final hybrid assembly was created with SPAdes v3.11.1 (Bankevich et al., 2012) combining all Illumina short reads with the second-run MinION reads, while the Canu draft assembly was introduced as untrusted contigs for gap closure and repeat resolution. Finally, only scaffolds with a length of at least 1000 bp were retained using the get.seqs command from mothur v1.43.0 (Schloss et al., 2009), creating the *C. closterium* v1.0 genome assembly (**Figure 2C**).

Gene prediction

A set of 30 RNA-seq libraries spanning both vegetative and sexual conditions (**Table S1**) was selected for gene prediction in the *C. closterium* v1.0 genome assembly. All reads were trimmed

and mapped to the genome assembly using STAR v2.7.0 (Dobin et al., 2013). After mapping, BAM files were used to infer the coding region (CDS) of protein-coding genes using BRAKER v2.1.2 following the BRAKER1 pipeline (Hoff et al., 2016). The resulting GFF annotation file was cleaned with AGAT v0.8.1 with the "convert_sp_gxf2gxf" command (Dainat et al., 2022). In addition, transfer RNAs (tRNA) were predicted using tRNAscan-SE v2.0.10 in "general" mode using Infernal v1.1 as a sequence search engine (Chan et al., 2021). Ribosomal RNAs (rRNA) were added using RNAmmer v1.2 (Lagesen et al., 2007). Finally, all annotation was merged and CDS and protein sequences were extracted with AGAT v0.8.1 (Dainat et al., 2022). Chloroplast (Ccl_contig161, Ccl_contig104) and mitochondrial scaffolds (Ccl_contig184) were identified through blastn with the respective *Seminavis robusta* contigs as a query (Osuna-Cruz et al., 2020).

Genome quality control

Statistics of the genome assembly were calculated using QUAST v5.0.0 (Mikheenko et al., 2018). Genomescope v1.0.0 (Vurture et al., 2017) was used to verify the haploid genome size, determine heterozygosity and confirm ploidy based on the k-mer distribution of raw Illumina reads. In brief, Jellyfish v2.2.6 (Marçais & Kingsford, 2011) was used to quantify the frequency of all 21-mers in pooled reads of both DNA sequencing runs (see "Illumina sequencing"), which were loaded into the Genomescope web browser (http://qb.cshl.edu/genomescope) using a maximum k-mer coverage of 1000 (**Figure 2D**). To determine the gene space completeness, *C. closterium* coding sequences (CDS) were loaded into TRAPID v2.0 (Bucchini et al., 2021) with default settings. The presence of core diatom (Bacillariophyta) gene families was determined using a conservation threshold of 0.8.

Functional gene annotation and homology detection

To functionally annotate the 24,359 newly predicted protein-coding genes of *C. closterium* in relation to existing gene sets from diatoms and other eukaryotic species, we created a sequence database containing the genomic information of 20 species, based on the PLAZA Diatoms database (Osuna-Cruz et al., 2020). CDS data of nine diatom species and nine eukaryotic outgroup species retrieved from PLAZA Diatoms v1.0 were complemented with the CDS of the reference genome of the sexual species *Skeletonema marinoi* (Pinder et al., in prep.) and the *C. closterium* reference genome introduced here (**Table S2**). We used this sequence database to detect Interpro domains, create homologous gene families and infer gene orthology as explained in supplementary Methods (Van Bel et al., 2022).

RNA sequencing of sexual reproduction in *C. closterium*

Generation of sexual RNA-seq transcriptomes for *C. closterium*

To verify the induction of a cell cycle arrest by darkness or sex pheromones, both strains were dark-incubated or SIP-treated as described in Klapper et al., (2021) (full protocol in Supplementary Methods). The effect of the treatment was tested in triplicate and DNA content was measured on intact fixed cells stained with 1x Sybr Green. The stained cell suspensions were analysed using an S3e TM Cell Sorter flow cytometer and ProSort v1.5.0.15 software (Bio-Rad Laboratories, Inc.). The proportion of G2/M relative to the total cell population was modelled using generalised linear models using a quasibinomial distribution and logit link. Tukey's post-hoc pairwise comparisons were carried out using Wald tests with the multcomp package (Hothorn et al., 2008).

Cultures of both mating types were grown separately in 150 ml ASW medium in large 175 cm^2 cell culture flasks (Greiner Bio One, Vilvoorde, Belgium). Both strains were inoculated at 200 cells/ml and were grown for four days without refreshment of the medium. Subsequently, the dark period was extended to a total of 36 h to synchronise cultures in the G1 phase (Vanormelingen et al., 2013). After synchronisation, still in the dark, cultures were scraped and resuspended in the medium by gentle shaking. Then, mating was initiated by pooling 50 ml of each mating type in a new, large culture flask. Consequently, this rendered three different large culture flasks: one filled with 100 ml MT+, one with 100 ml MT- and one with 100 ml of mating culture (a mixture of 50:50 MT+ and MT-). All flasks were incubated in continuous light at 20 µmol photons m^2s^{-1}.

The progression of mating was verified regularly with a Zeiss Axiovert 135 inverted microscope (Zeiss, Jena, Germany). At three instances, 10 photographs of mating cultures were taken randomly with a Canon Powershot G11 digital camera to later quantify cells in the different stages of sexual reproduction (paired cells, gamete and zygote formation, auxosporulation and initial cell formulation) using the Cell Counter plug-in in ImageJ v1.52 (https://imagej.nih.gov/ij/index.html). At three different time points after the initiation of mating, T1=9 h, T2=14 h and T3=27 h, coinciding with the peak of paired cells, gamete/zygote formation and auxosporulation respectively, cultures were harvested by filtration on Versapor filters (3 µm pore size, 25-mm diameter, PALL). Upon harvesting, the control cultures of both mating types were harvested on the same filter, to obtain the same 50/50 mating type ratio as in the mating cultures. RNA extraction, quality control and 2x150bp Illumina sequencing were performed as shown in Supplementary Methods.

Differential expression analysis and cross-species comparative transcriptomics

Retrieval of public RNA-seq data of sexual reproduction

The *C. closterium* transcriptome data generated in this paper were compared with independent, publicly available sexual reproduction experiments from other diatom species (**Figure 2E**). In particular, we investigated conserved expression patterns with *S. robusta*, a benthic diatom for which the life cycle is relatively well understood (**Figure 1A,B**). For *S. robusta*, we retrieved RNA from three timepoints T1, T2 and T3 from Osuna-Cruz et al., (2020). These time points represent analogous stages of sexual reproduction as in *C. closterium* (**Figure 3A**). To contrast expression patterns with planktonic diatoms, we introduce additional transcriptomic data of sexual reproduction from two planktonic diatoms: T2 (gametes, zygotes) and T3 (auxospores, initial cells) from the pennate *P. multistriata* (Annunziata et al., 2022) as well as T1 (gametangia) and T2 (intermediate stage between gametangia and auxospores) from the centric *S. marinoi* (Ferrante et al., 2019).

Gene expression quantification and differential expression analysis

Quality control of RNA-seq reads for libraries of the four selected species was performed using FastQC v0.11.2 (https://github.com/s-andrews/FastQC). *P. multistriata* and MT- *S. robusta* reads were trimmed with Trimmomatic v0.36 using the command ILLUMINACLIP 2:30:10:2 (Bolger et al., 2014). Next, clean paired-end (*S. robusta, C. closterium*) and single-end (*P. multistriata, S. marinoi*) reads were mapped to each species' respective reference gene set (**Figure 2E**) (Basu et al., 2017; Osuna-Cruz et al., 2020; Pinder et al., in prep.) using Salmon v1.3 (Patro et al., 2017). Salmon read counts were imported into R v3.3.2 (Team, 2017), summarised to gene level

counts using the tximport package v1.16.1 (Soneson et al., 2016) and genes with a counts per million (CPM) >1 in at least three samples were kept. Next, a "Trimmed Mean Of M-values" (TMM) normalisation was performed to normalise for library size and composition effects, as implemented in EdgeR v3.30.3 (Robinson et al., 2010). TMM-normalised CPM were calculated for all libraries using the cpm function in EdgeR, which were used for the visualisation of gene expression throughout the paper.

A differential expression (DE) analysis was performed for each species by fitting a negative binomial model to model gene expression in function of treatment (time point and vegetative/mating) as a single factor using EdgeR. DE analysis was performed with likelihood ratio tests against a log2 fold change threshold of 1, implemented in EdgeR as 'glmTreat'. To control the false discovery rate (FDR) at 5% on the gene level and allow the discovery of a single "overall" set of omnibus DE genes for each species, we adopted a stage-wise testing approach as implemented in the R package stageR v1.10.0 (Van den Berge et al., 2017), as detailed in the Supplementary Methods.

Comparative transcriptomics between *C. closterium* and *S. robusta*

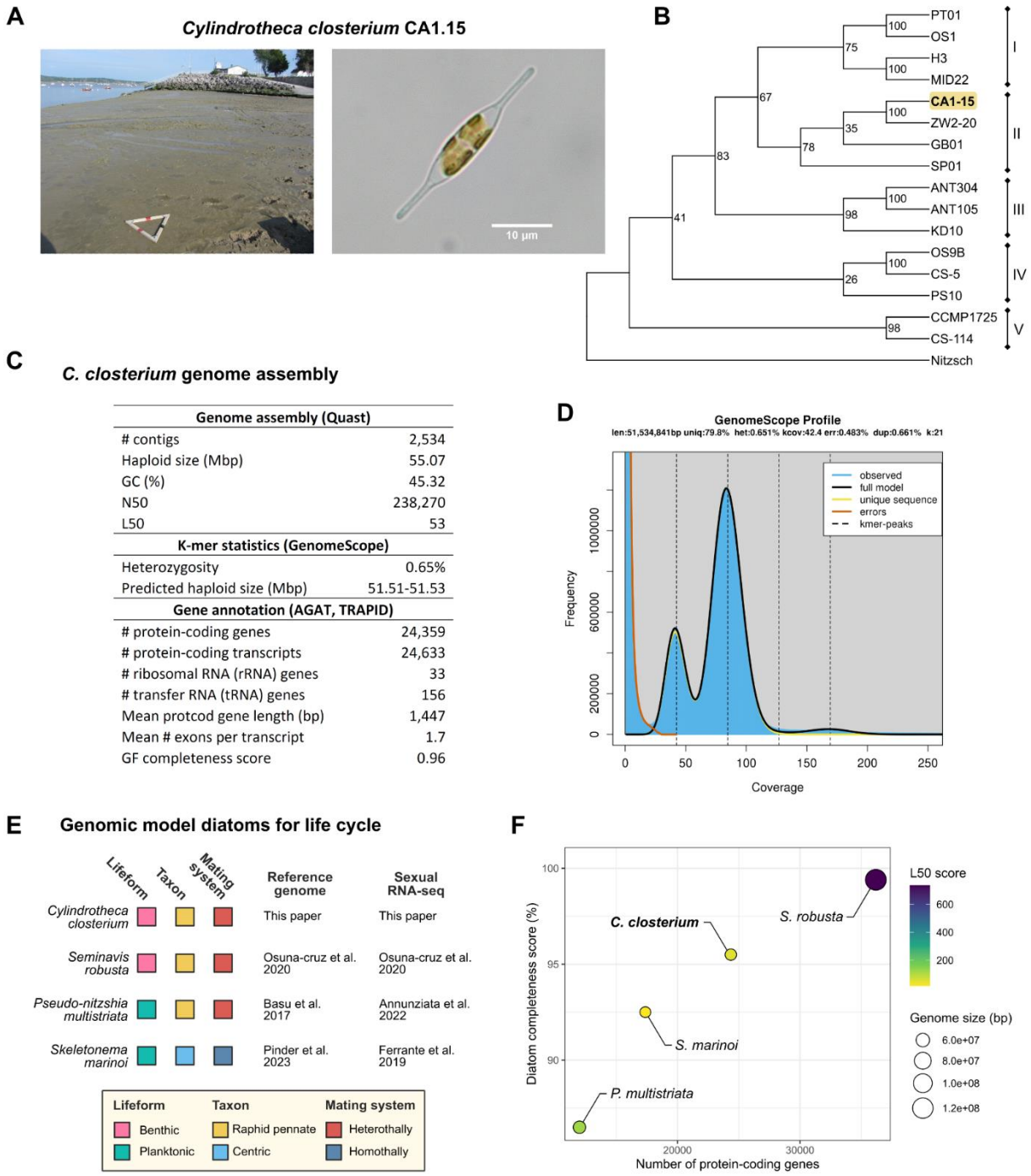
The differential expression results of the sexual time series from benthic diatoms *C. closterium* and *S. robusta* were integrated using two complementary approaches (**Figure 3D**): (1) based on the PLAZA integrative orthology method, we identified 6,659 one-to-one orthologs that fulfil the "BHIF", "TROG", and "ORTHO" requirements. From these, a set of 138 ortholog pairs that are significantly upregulated during sex in both species was selected for further analysis. To prioritise genes with a conserved timing in the sexual process across benthic species, we further subsetted ortholog pairs with maximum fold changes in T1 or T2 ("gametangia, gametes and zygotes") or in T3 ("auxospores") in both species and ortholog pairs were ordered by the sum of the ranks of their screening-stage adjusted p-value. (2) For each species, homologous gene

families that were enriched upregulated genes from each time point (T1, T2 or T3) was determined with a hypergeometric test using the enricher command from ClusterProfiler for R with an FDR-adjusted p-value of 0.05 to infer enrichment (Yu et al., 2012). Gene families enriched in upregulated genes in both species were retained for closer inspection.

Results and Discussion

A reference genome for *C. closterium*

Cylindrotheca closterium (Ehrenberg) Reimann and Lewin is a predominantly benthic species widespread in coastal and estuarine waters (Round et al., 1990). Strain CA1.15 was isolated from a phototrophic biofilm on an intertidal muddy sediment in Baie de la Canche (France), identified as clade II (Stock et al., 2019) and selected for genome sequencing (**Figure 2A,B**). Hybrid assembly with SPAdes of 15 million 250 bp paired-end Illumina reads and 5 million long Oxford Nanopore MinION reads across two runs resulted in a haploid genome size of 55.07 Mbp, distributed across 2,534 scaffolds with an N50 of 238,270 (**Figure 2C**). A k-mer analysis of the Illumina reads corroborates the assembly size, estimating the haploid genome size at 51 Mbp (**Figure 2D**). This falls well within the range of available diatom genome sizes, which range from 20 Mbp in *Plagiostriata* sp. up to 231 Mbp in *C. weissflogii* (**Table S3**) (L. Li et al., 2021; Sato et al., 2020). Modelling the k-mer frequency distribution indicates a heterozygosity level of 0.65% - a similar level as the benthic *S. robusta* (0.79%), but lower than in highly heterozygous genomes such as *Nitzschia inconspicua* (2.7%) and *Chaetoceros tenuissimus* (1.6%) (Hongo et al., 2021; Oliver et al., 2021; Osuna-Cruz et al., 2020). Finally, the k-mer histogram also supports the diploid nature of the *C. closterium genome* (**Figure 2D**).



Genome assembly (Quast)	
# contigs	2,534
Haploid size (Mbp)	55.07
GC (%)	45.32
N50	238,270
L50	53
K-mer statistics (GenomeScope)	
Heterozygosity	0.65%
Predicted haploid size (Mbp)	51.51-51.53
Gene annotation (AGAT, TRAPID)	
# protein-coding genes	24,359
# protein-coding transcripts	24,633
# ribosomal RNA (rRNA) genes	33
# transfer RNA (tRNA) genes	156
Mean protcod gene length (bp)	1,447
Mean # exons per transcript	1.7
GF completeness score	0.96

Figure 2: Genome properties for *C. closterium* and comparison with other genomic models for the diatom life cycle. (A) isolation site of *C. closterium* CA1-15 from a phototrophic biofilm on an intertidal muddy sediment in Baie de la Canche (France). **(B)** Midpoint-rooted maximum likelihood consensus tree based on 1000 rapid bootstrap inferences of an rbcL alignment in *C. closterium* strains. Node labels

indicate bootstrap support. Reference strains for the different *C. closterium* clades were sourced from (Stock et al., 2019). Roman numerals represent the different clades, while *Nitzschia sp.* was used as an outgroup. **(C)** *C. closterium* CA1-15 genome assembly and gene annotation statistics. **(D)** Histogram and modelling of the frequency distribution of 21-mers in all Illumina DNA reads **(E)** Overview and comparison of genomic model diatoms for life cycle. **(F)** Scatter plot visualising the contiguity and diatom completeness score of genome assemblies from selected life cycle model diatoms. The x-axis displays the number of protein-coding genes, while the y-axis depicts the diatom completeness score. Genome assemblies are coloured according to their L50 score and dot size increases with genome size.

Gene model prediction supported by RNA-seq samples covering both vegetative and sexual conditions (**Table S1**) resulted in a total of 24,359 protein-coding genes, placing *C. closterium* among the diatoms with larger gene sets (**Table S3**) (Osuna-Cruz et al., 2020). Combined with its moderate genome size, this suggests a high gene density in *C. closterium*. Indeed, although *C. closterium* and *P. multistriata* share a similar genome size (55 Mbp vs 59 Mbp respectively), the latter encodes approximately half as many genes (Basu et al., 2017). Assessing the completeness of the *C. closterium* gene space (Bucchini et al., 2021), we could successfully retrieve 96% of core diatom gene families (**Figure 2E,F**). Protein-based sequence similarity searches between the proteome of *C. closterium* and 19 other species based on the PLAZA Diatom v1.0 database (Osuna-Cruz et al., 2020) (**Table S2**) resulted in the clustering of the 24,359 protein-coding genes in the *C. closterium* genome into 11,009 distinct homologous gene families (**Figure S1**). Of these genes, 14,821 (60.8%) belonged to a diatom-specific gene family, among which 7446 (30.6%) members of species-specific *C. closterium* gene families, ranking *C. closterium* among the species with the highest proportion of species-specific genes, hence suggesting a high level of genetic specialisation (**Figure S2**).

Differential expression analysis of sex transcriptomes

We generated a time series transcriptome for *C. closterium* undergoing sexual reproduction, with the time points chosen in such a way as to capture the same sexual cell types as an existing RNA-seq dataset in *S. robusta*: T1, dominated by paired cells; T2, consisting mainly of gametes and zygotes; and T3, marking the appearance of auxospores (**Figure 3A**) (Osuna-Cruz et al., 2020). After confirming that sexual reproduction in *C. closterium* can be synchronised through a dark-induced cell cycle arrest in the G1 phase (**Figure 3B, S5,** Vanormelingen et al. (2013)), we made sexual crosses of dark-synchronised strains and harvested at three different time points (**Figure 3C**). For both species, we carried out a differential expression (DE) analysis to compare expression between sexual cultures and their matched control within each time point. A total of 4660 and 7541 genes were detected to be DE between vegetative and sexual conditions in *C. closterium* and *S. robusta* respectively, with a similar number of up- and downregulated genes found in each time point (**Figure 3D**). Notably, while *C. closterium* exhibits a much stronger differential expression in T3 compared to T1 and T2, this was not the case in *S. robusta*. This is likely due to the lower proportion of auxospores in T3 in *S. robusta* compared to *C. closterium* (8.0% and 25.1% respectively, **Figure 3C**).

After confirming that the *C. closterium* genome encodes a complete set of meiosis-specific genes as well as homologs of the Sex Induced Genes SIG5-SIG11 (**Data S1**) (Ferrante et al., 2019; Patil et al., 2015), we verified the correspondence of the time points between species by confirming that a set of five meiosis-specific genes (*SPO11-2*, *Mer3*, *Mnd1*, *Msh4* and *Msh5*) exhibit maximal fold changes in T1 and T2, when meiosis should be ongoing during gametogenesis (**Figure 4A**). Similarly, *SIG7*, which encodes a meiotic homologous-pairing domain (Hop2) and was upregulated in response to the sex-inducing pheromone SIP in *S. robusta* (Bilcke et al., 2020), showed its strongest expression during gametogenesis (T1) in both

species (**Figure S3**). Together, these markers indicate the feasibility of comparing time series across species.

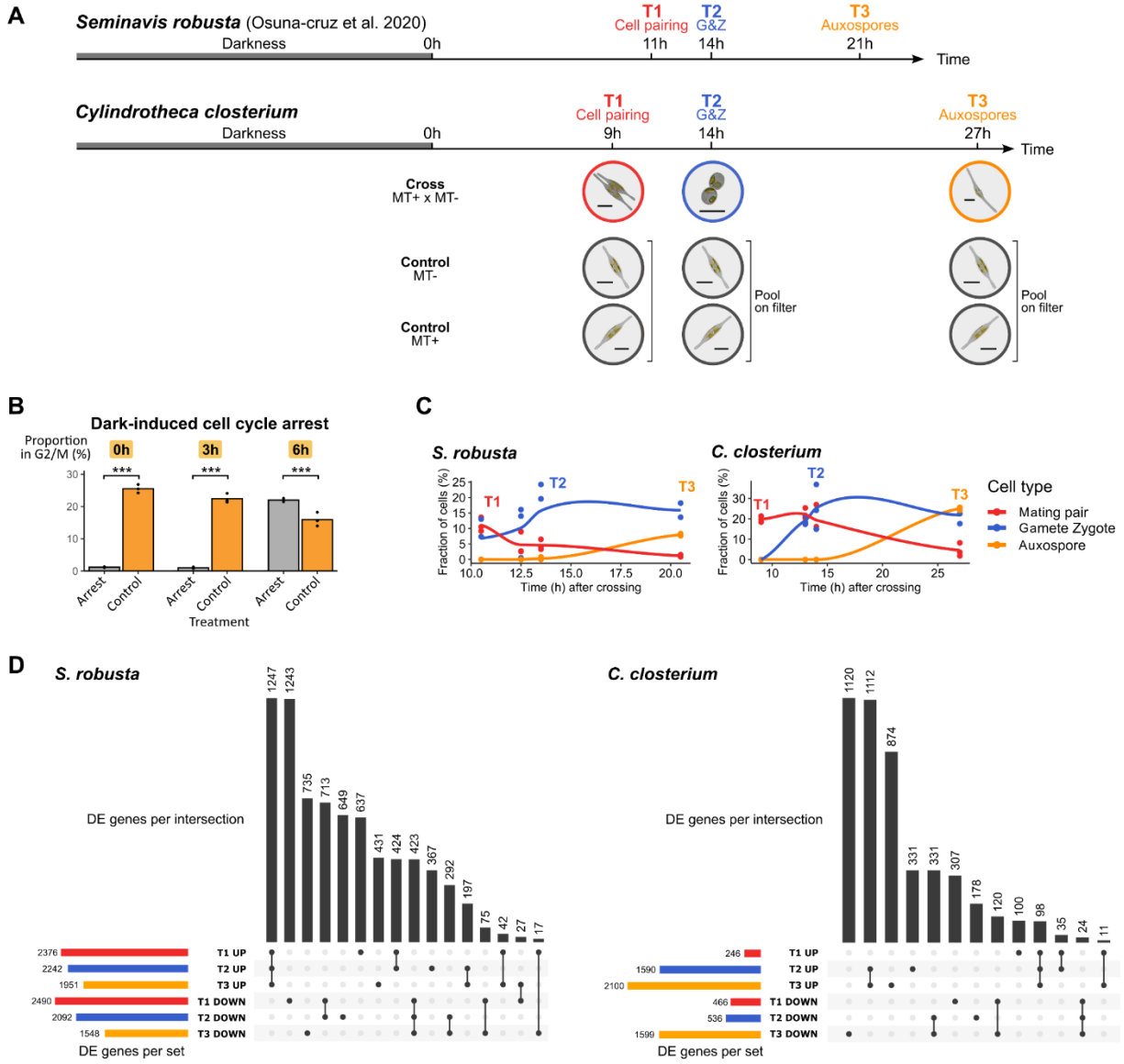


Figure 3: Generation and comparative analysis of RNA-sequencing datasets during sexual reproduction. (A) Experimental set-up of sexual reproduction time series. Mating type plus and minus (MT+/-) strains of *S. robusta* and *C. closterium* were grown as monocultures ("Control") or mixed to induce sexual reproduction ("Cross"). Before mixing and re-illumination, cultures were synchronised via a 36 h dark incubation. Samples were harvested at three stages during sexual reproduction called T1-T3,

corresponding to cell pairing, gametes & zygotes (“G&Z”) and auxospores respectively. Scale bars indicate 10 µm. **(B)** Bar plots showing the proportion of cells in G2 and M phase of the cell cycle in cultures treated with a dark arrest (“Arrest”) and those continuously illuminated (“Control”). Bars indicate the average proportion, while points show individual replicates. **(C)** Change in the proportion of sexual cell types throughout the RNA-seq time series for *S. robusta* and *C. closterium*. Points represent individual replicates, while the smoothed line connects the mean. Selected time points for RNA-seq are indicated in colours (T1, T2, T3). **(D)** Intersection analysis between differentially expressed genes in *C. closterium* and *S. robusta.* For each species, an UpSet plot visualises the number of significantly up-and downregulated (“UP”, “DOWN”) genes that are unique or shared between different time points (T1, T2, T3). Only sets with >10 genes are shown.

Comparative transcriptomics reveals shared responses of *S. robusta* and *C. closterium* to sexual reproduction.

Two complementary approaches were used to identify conserved sex genes between *S. robusta* and *C. closterium* (**Figure 4B**). Firstly, based on the PLAZA integrative orthology method (Proost et al., 2015), we selected a conserved set of one-to-one ortholog pairs that are significantly differentially expressed (DE) during sex in both species. Generally, the overlap between sets of DE genes of *S. robusta* and *C. closterium* was not very large. Out of 1453 DE genes in *C. closterium* and 1027 DE genes in *S. robusta* that belong to one-to-one ortholog pairs, only 319 pairs were differentially expressed in both species. While this is significantly larger than the random expected outcome (Chi-square test, $p < 1e\text{-}15$), this leaves 1842 ortholog pairs that were differentially expressed during sexual reproduction in only one species. Among the conserved differentially expressed pairs, we did observe an enrichment in genes regulated in the same direction (Chi-square test, $p < 1e\text{-}15$). Indeed, orthologs were typically either upregulated in both species ($n = 138$) or downregulated in both ($n = 99$) (**Figure 4B**). Upregulated ortholog pairs were further subsetted by the timing of their maximum fold changes to prioritise genes with a

conserved timing in the sexual process: 56 genes consistently peaked in T1/T2 (gametangia, gametes and zygotes) in both species while 11 did so in T3 (auxospores).

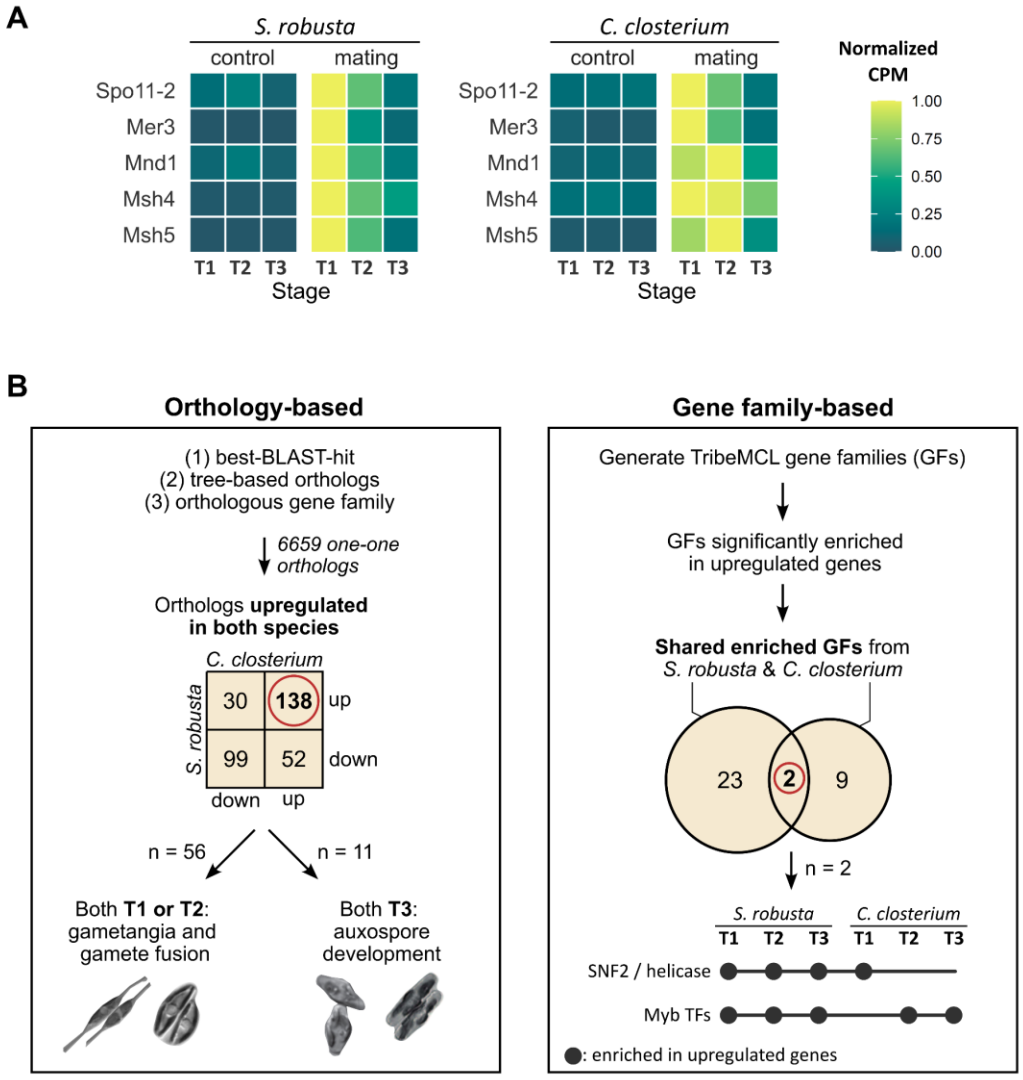


Figure 4: Comparative transcriptomics of sexual reproduction time series in the benthic *C. closterium* and *S. robusta*. (A) Heatmap showing gene expression of meiotic genes in normalised counts per million (CPM), averaged over the replicates for each time point. **(B)** Schematic representation of the two complementary approaches used to detect conserved sexual genes across benthic diatom species. The red circles indicate genes and gene families (GFs) selected for further analysis. For the two shared enriched GFs, a diagram shows in which time points (T1,T2,T3) the two gene families shared between both benthic species exhibited a significant enrichment in upregulated genes during sex.

In a second approach, gene families enriched in upregulated genes were determined at each time point. Out of 25 and 11 gene families enriched in upregulated genes during sex in *S. robusta* and *C. closterium* respectively, only two gene families were enriched in both species, which were retained for further investigation (**Figure 4B**). Together, these two approaches could identify both specific sexual ortholog pairs, as well as broader gene families where different but related genes are induced in either species during sex. In what follows, we will use these gene sets as a starting point to discuss key players in sexual reproduction chronologically, starting with a pheromone-induced cell cycle arrest, cell-cell recognition and gametogenesis, and ending by pinpointing potential auxospore genes.

Pheromone-induced cell cycle arrest precedes gametangial pairing

Raphid pennate diatoms such as *S. robusta* and *C. closterium* exhibit a (partial) cell cycle arrest in response to sex pheromones (**Figure 5A),** which we also confirmed for the *C. closterium* strains used here ($p < 0.001$, **Figure 5B, S6**). This pheromone-induced cell cycle arrest is accompanied by a downregulation of numerous cell cycle genes in *S. robusta* (Bilcke et al., 2020). In contrast, we here observed three cell cycle-related genes among the 56 ortholog pairs induced during the early stages (T1, T2) of sex in *C. closterium* and *S. robusta*: *MCM9*, *Cdc20-1* and a diatom-specific cyclin (*dsCyc2*). To investigate this further, we assessed the expression of cell cycle markers in four diatom species: *C. closterium, S. robusta* and *P. multistriata* (which exhibit a cell cycle arrest), and the centric *S. marinoi* (which does not) (Ferrante et al., 2019). Markers included genes that define the S phase such as PCNA, Origin recognition complex, DNA polymerases and others, as well as the M phase, e.g. condensin complex subunits, separase and cdc20 (for a full list, see **Data S2**). We observed that most cell cycle markers were either non-responsive or upregulated during sexual reproduction (**Figure 5C**). The upregulation of these cell cycle genes likely signals that gametangia are undergoing the pre-meiotic S-phase and meiotic M-phase (meiosis I and II) prior to gametogenesis.

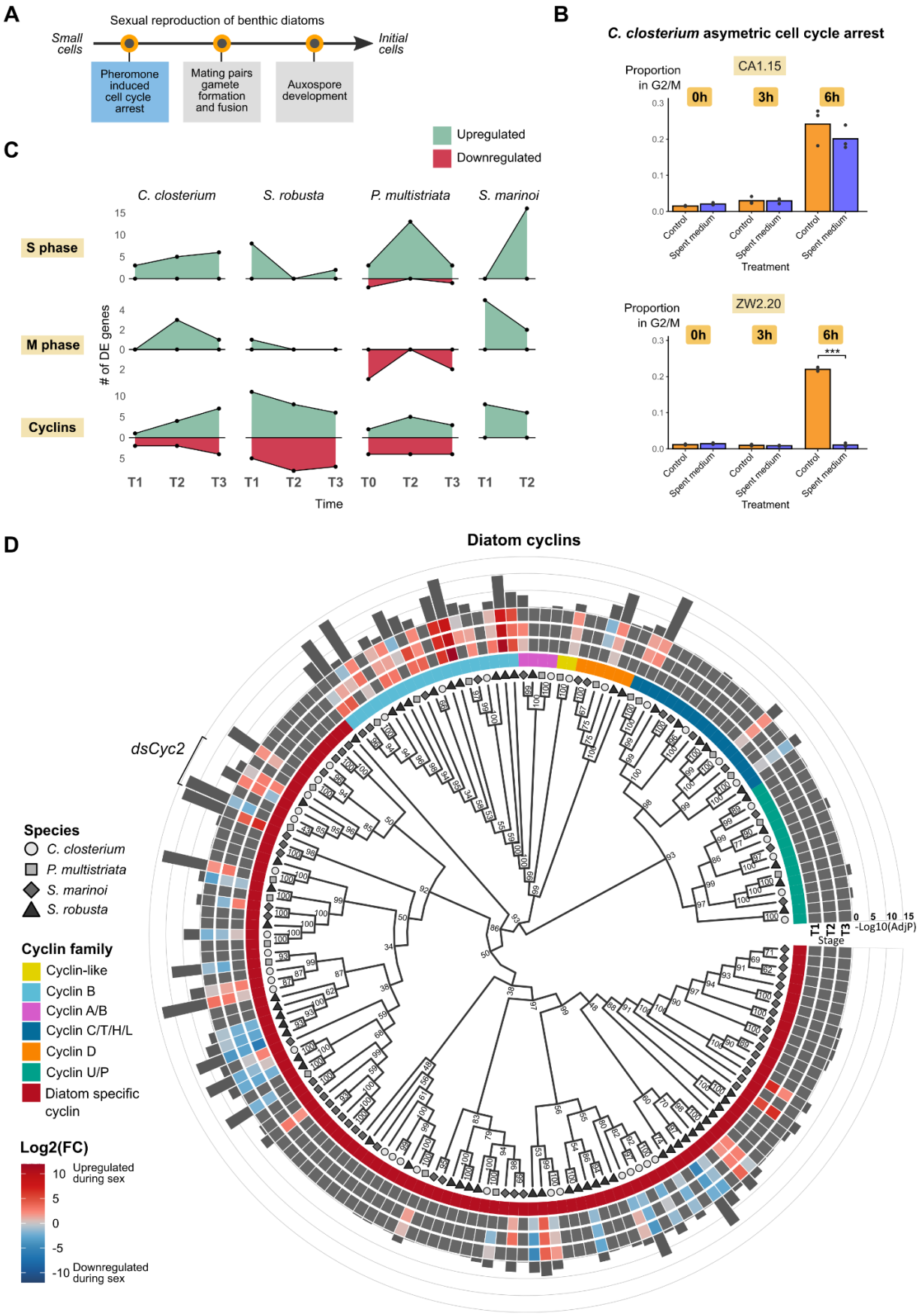

A
Sexual reproduction of benthic diatoms
Small cells
Initial cells
Pheromone induced cell cycle arrest
Mating pairs gamete formation and fusion
Auxospore development
B
C. closterium asymetric cell cycle arrest
CA1.15
Proportion in G2/M
0h
3h
6h
Control
Spent medium
Treatment
ZW2.20
***
C
Upregulated
Downregulated
C. closterium
S. robusta
P. multistriata
S. marinoi
S phase
M phase
Cyclins
# of DE genes
T1
T2
T3
T0
Time
D
Diatom cyclins
dsCyc2
Species
C. closterium
P. multistriata
S. marinoi
S. robusta
Cyclin family
Cyclin-like
Cyclin B
Cyclin A/B
Cyclin C/T/H/L
Cyclin D
Cyclin U/P
Diatom specific cyclin
Log2(FC)
Upregulated during sex
Downregulated during sex
Stage
-Log10(AdjP)

Figure 5: Cell cycle progression during sexual reproduction. (A) Timeline of cell stages during sexual reproduction in benthic diatoms. The stage considered in this figure is coloured in blue. **(B)** Line plot showing the number of differentially expressed cell cycle marker genes over time, for four sexual diatom species. Differentially expressed genes are subdivided in genes upregulated during sex (green, upwards) and those downregulated during sex (red, downwards). T0 is the acute response time point of the *P. multistriata* experiment (Annunziata et al., 2022). **(C)** Evidence from a flow cytometry assay for an asymmetric cell cycle arrest in *C. closterium* strains ZW2.20 (MT+) and CA1.15 (MT-). Bar plots show the proportion of cells in G2 and M phase, just before re-illumination (0 h), as well as 3 h and 6 h after illumination. Colours indicate whether the cultures were untreated (control) or treated with spent medium from cultures of the opposite mating type. Bars indicate the average proportion in G2/M, while points show individual replicates. **(D)** Circularised midpoint-rooted bootstrap consensus phylogenetic tree of cyclins in four species: *S. robusta, S. marinoi, P. multistriata* and *C. closterium*. The rings surrounding the trees indicate the cyclin family classification as well as a heatmap of fold changes during sexual reproduction over three time stages. Only the fold changes of genes with a significant differential expression (FDR < 0.05) are shown. The $-\log_{10}$ adjusted p-value for the screening stage is visualised as bar plots on the outer edge. Node labels indicate bootstrap support. Insets show gene expression in counts per million for two homologs of the light-dependent cell cycle regulator dsCyc2 (left) and a summary of the number of up-and downregulated cyclins in each species (right). Gene IDs and gene names of all cell cycle marker genes and cyclins can be found in **Data S2** and the full phylogenetic tree with gene IDs is included as **Figure S4.**

Meanwhile, several cyclins were downregulated during sex in the pennates (**Figure 5C, D**). The majority of these downregulated cyclins belong to the family of so-called diatom-specific cyclins (**Figure 4D**) (Huysman et al., 2010). Members of this unique cyclin family have been implicated in the interaction of the cell cycle with various environmental cues, such as light, phosphorus concentration, iron depletion and grazing (Amato et al., 2018; Huysman et al., 2010, 2013; Valenzuela et al., 2012). Hence, their downregulation may simply indicate the downregulation of cyclins with a vegetative signalling function. Notably, we observed that two specific cyclin A/B

genes (*Sro331_g119100* and *Sro251_g099220*) and the diatom-specific cyclin *dsCyc2* were upregulated, rather than downregulated, in *S. robusta* and *C. closterium*. The downregulation of these genes is considered a hallmark of an active pheromone-induced cell cycle arrest (Annunziata et al., 2022; Bilcke et al., 2020). Hence, this finding suggests that the cell cycle arrest was broken when paired cells start appearing, in contrast to the planktonic *P. multistriata* (**Figure 3C**). Indeed, both *dsCyc2* and multiple cell cycle marker genes continue to be repressed until the last time point (5 days) in *P. multistriata*, for which a 3-day growth arrest has been reported (**Figure 5C, 5D**) (Annunziata et al., 2022). A shorter cell cycle arrest in benthic compared to planktonic pennates may be attributed to the pheromone-guided cell searching and pairing in benthic diatoms (Gillard et al., 2008; Klapper et al., 2021; Sato et al., 2020). This is also in accordance with the short-lived production of the attraction pheromone in *S. robusta* that lasts only 5 h, suggesting that the mate-finding period is limited in time (Gillard et al., 2008). In contrast, assuming random mate finding in the planktonic *P. multistriata,* a long cell cycle arrest increases the probability of successful pairing of cells in G1 phase. The centric *S. marinoi* was the only species that did not display a downregulation of any cyclin or cell cycle marker (**Figure 5C**), in accordance with the absence of a growth arrest in this species (Ferrante et al., 2019). Contrary to the raphid pennates, centric diatoms release flagellated male gametes that localise and move towards an immobile oogonial cell, and thus do not need a cell cycle arrest before gametogenesis.

Sex-specific B-type cyclins are widespread across diatoms

We observed a strong induction of B1 and B2 type cyclins during sexual reproduction in all four different species (**Figure 5D, Data S1**). This response was conserved across diatom species from different phyla (centric versus pennate) and mating strategies (homothallic oogamy versus heterothallic isogamy) (**Figure 2C**). In animals, plants and ciliates, B-type cyclins have been identified as regulators of meiosis (Bulankova et al., 2013; Carlile & Amon, 2008; Karasu et al.,

2019; Y. Li et al., 2019; Yan et al., 2016). In accordance with a possible function in meiosis, most B-type cyclins showed a maximal expression at T1 or T2, time points where gametangia and gametes were abundant (**Figure 3C, 5D**). While almost all B-type cyclins were associated with meiosis and gametogenesis, certain cyclins from other families showed a different timing of expression. Indeed, the diatom-specific cyclin *Sro299_g111470* was previously found to be upregulated in response to SIP in *S. robusta* and showed a persistent induction throughout the time series presented here. Furthermore, several cyclins were induced predominantly during T3 in *S. robusta* or *C. closterium,* suggesting they are involved in the regulation of auxospore formation or identity (**Figure 5D**).

Conserved meiotic and regulatory genes are expressed during cell pairing and gamete fusion

Among the 138 one-to-one orthologs that are upregulated during sexual reproduction in *C. closterium* and *S. robusta* (**Figure 4B**), we identified 56 orthologous gene pairs that show coordinated peak expression in T1 or T2 in both benthic diatoms (**Figure 6B, Data S3**). These can be considered genes that are involved in the early stages of sexual reproduction: cell pairing, the development of two gametes per gametangium, and fusion into zygotes. In accordance with ongoing meiotic divisions during gamete formation, several of these early ortholog pairs encode meiotic recombination and DNA repair proteins, including msh5, mer3, SIG7/Hop2 and WSS1 (**Figure 6B**).

Additionally, several genes with a regulatory function showed a consistent early expression (**Figure 6B**). After ranking by their joint differential expression level, we found that the top pair corresponds to a diatom-specific zinc-finger transcription factor that was also upregulated during sexual reproduction of *S. marinoi* and *P. multistriata* (**Figure 6B**). Further, a specific F-box protein (a recognition subunit of ubiquitin ligase complexes) was upregulated in both benthics

549 (*S. robusta* and *C. closterium*), whereas the WD repeat protein WDR55 was upregulated during
550 pairing and gametogenesis in all three pennate species (*S. robusta, C. closterium*, *P.*
551 *multistriata*), but not in the centric *S. marinoi* (**Figure S7**). In plants, WDR55 is required for
552 gametogenesis and fusion of polar nuclei in embryo sacs (Bjerkan et al., 2012), raising the
553 possibility that WDR55 has a similar molecular function in pair formation or gamete fusion of
554 pennate diatoms. A pair of casein kinases from a diatom-specific gene family, meanwhile,
555 showed a peak in expression in T1/T2. Casein kinases perform various functions, among which
556 DNA repair and circadian clock regulation (Matsuo & Ishiura, 2010). Finally, we observed three
557 orthologous gene pairs that belong to pennate diatom-specific gene families, all of them without
558 a clear functional annotation (**Figure 6B**). One of these pairs (*Ccl_9979*/*Sro1095_g240650*) has
559 a homolog in the planktonic pennate *P. multistriata* (*Pmu0124150*), which was strongly
560 upregulated during gametogenesis (T2) as well.

561 Besides conserved orthologous gene expression, a second approach to detect homologous
562 expression between *S. robusta* and *C. closterium* entailed intersecting gene families that are
563 enriched in upregulated genes during sex at each time point, resulting in two gene families
564 enriched in both species (**Figure 4B**).
565 The first family, functionally annotated with SNF2/helicase protein domains, is large and diverse,
566 comprising 27 and 46 members in *C. closterium* and *S. robusta,* respectively. Clades with
567 upregulated genes encoded homologous recombination and chromatin regulation machinery:
568 RAD54, several CHR genes (“Chromatin remodelling”), and homologs of the RAD5-related
569 DNA/RNA helicase AT5G05130 (**Figure S8**) (Chen et al., 2008; Mazin et al., 2010). Their shared
570 upregulation is likely associated with chromosomal reorganisation during meiosis. In support of
571 this hypothesis, in *C. closterium*, the gene family was only enriched in upregulated genes in T1,
572 when gametangia undergo meiosis.

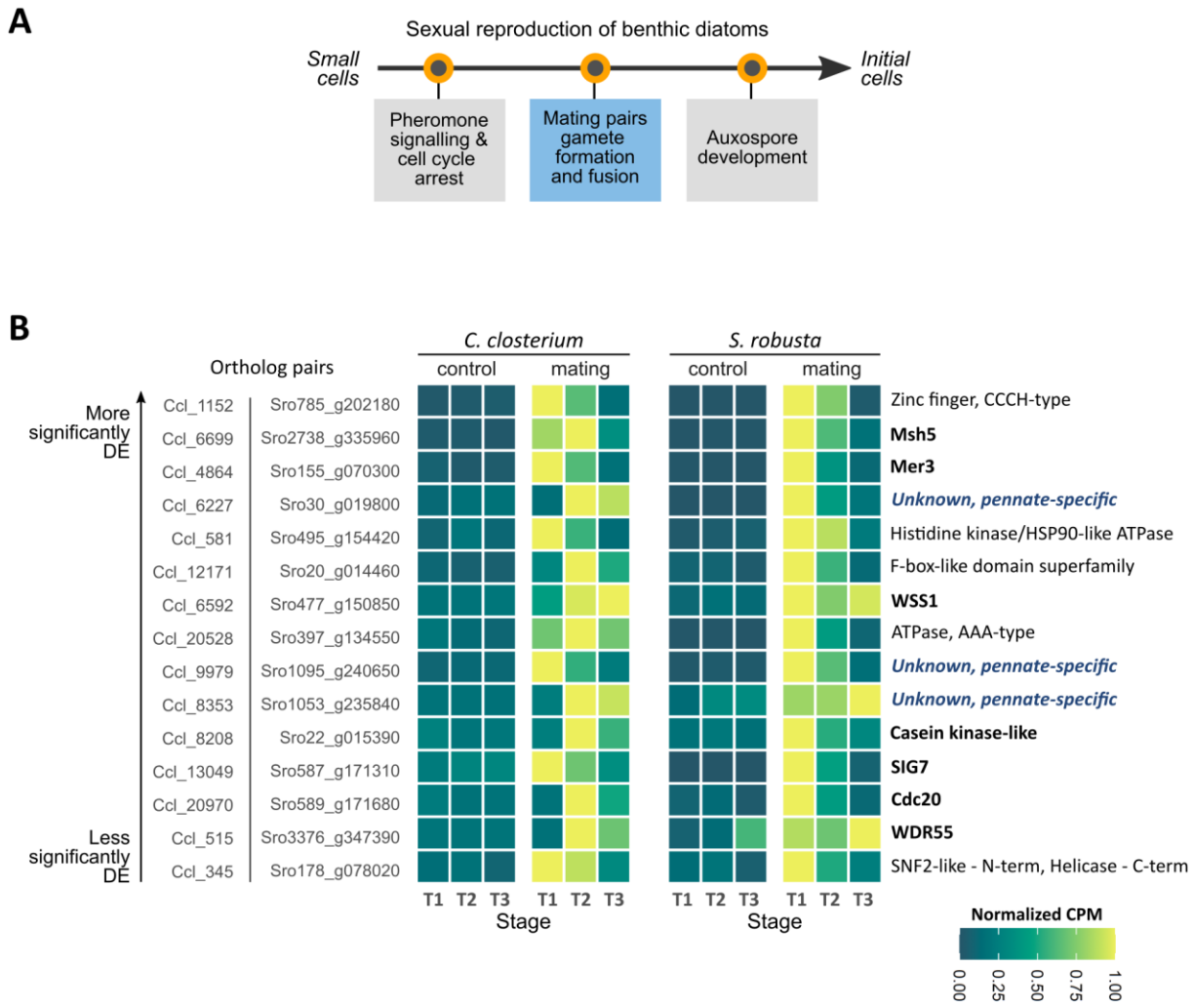


Figure 6: Concerted gene expression during the early stages of sexual reproduction. (A) Timeline of cell stages during sexual reproduction in benthic diatoms. The stage considered in this figure is coloured in blue. **(B)** Heatmap showing gene expression in average normalised counts per million (CPM) over time for the top 15 highest ranked T1/T2-specific ortholog pairs of the benthic *S. robusta* and *C. closterium*. Gene pairs are ordered by differential expression level using the sum of their p-value ranks: the smaller the ranksum, the higher the joint significance of ortholog pairs. On the right, InterPro-based functional annotation is shown. Genes whose function could be inferred based on well-studied homologs from other eukaryotes are indicated in bold.

583 The second enriched gene family encompasses the majority of Myb transcription factors found
584 in diatoms, more specifically the 2R-MYB, 3R-MYB, 4R-MYB and 5R-MYB clades (Wang et al.,
585 2022). This gene family consists of 13 and 11 members in *C. closterium* and *S. robusta*
586 respectively and was enriched in upregulated genes in almost every time point: T1, T2 and T3
587 in *S. robusta* and T2, T3 in *C. closterium* (**Figure 4B**). In plants, Myb transcription factors are
588 involved in both the regulation of female reproductive development as well as participation in the
589 pollen tube reception (Liang et al., 2013; Makkena et al., 2012).

590

591 ## First insights into the genes that define the auxospore

592 Auxospores are a hallmark of the diatom life cycle, and are an essential developmental step
593 responsible for cell size restoration in the majority of species studied so far. While the unique
594 ultrastructure and cellular changes during auxospore development are well-described, the
595 process has not been genetically characterised (Kaczmarska & Ehrman, 2021). Here, we make
596 use of the presence of young auxospores at T3 in both *S. robusta* and *C. closterium* to identify
597 the first putative auxospore markers (**Figure 3C**). Out of the 138 ortholog pairs significantly
598 upregulated during sex in both species, we selected all pairs that had consistent maximum fold
599 changes during the auxospore stage (T3). Ranking the remaining T3-specific genes based on
600 their differential expression significance level, we identified three key orthologous gene pairs
601 whose expression is peaking when auxospores are developing, which we subsequently called
602 “AAE1-3” (“Auxospore-associated expression”) (**Figure 7B,C**). AAE1 encodes high mobility
603 group box (HMGB) proteins, an abundant family of DNA binding proteins with various DNA-
604 related functions in chromatin structuring and transcription (Mallik et al., 2018). This gene seems
605 very specific for auxosporulation, as it was not expressed in vegetative conditions. Even though
606 this gene is found in both pennate and centric diatoms, no homolog could be found in *P.*

multistriata nor in *S. marinoi,* where they were either lost during evolution or have not been annotated in the genome assembly. AAE2 is a homolog of the *P. tricornutum* zinc-finger transcription factor *Pt_CCHH18* (Rayko et al., 2010), whose auxospore-associated expression was confined to pennate diatoms (**Figure 7C)**. Finally, AAE3 belongs to a diatom-specific gene family encompassing both centric and pennate species and encodes dioxygenase/hydroxylase genes with unknown function.

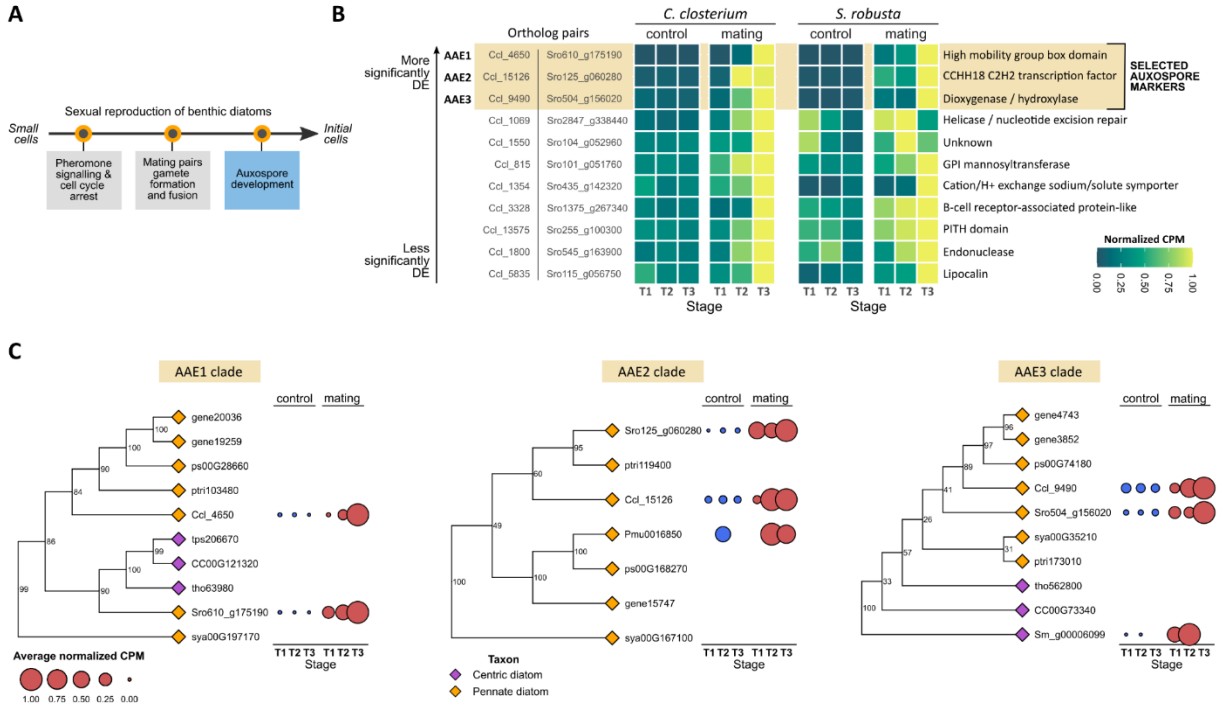


Figure 7: Orthology-based identification of auxospore genes. (A) Timeline of cell stages during sexual reproduction in benthic diatoms. The stage considered in this figure is coloured in blue. **(B)** Heatmap showing gene expression in average normalised counts per million (CPM) over time for T3-specific ortholog pairs in *S. robusta* and *C. closterium*. Gene pairs are ordered by differential expression level using the sum of their p-value ranks. InterPro-based functional annotation is shown on the right. **(C)** Midpoint-rooted consensus phylogenetic tree showing homologs of each of the selected Auxospore-associated expression (AEE) genes. Node labels are bootstrap support values. On the right, the average normalised CPM of genes of four species for which sexual RNA-seq time series data is available (*C. closterium, S. robusta, P. multistriata* and *S. marinoi*) is represented by the radii of circles. The fill colour of circles corresponds with the treatment (blue = control, red = sexual reproduction).

Considering that the expression of these genes consistently peaked during the time point when developing auxospores were observed, we postulate that they play a role during auxospore enlargement and maturation. Since AAE1 and AAE2 appear to be involved in transcription regulation, they may be determining auxospore cell fate, or may be master regulators of unique features of auxospores such as (1) their gradual elongation, (2) auxospore-specific cell wall elements: a layer of incunabula outer elements, and inner siliceous bands called the perizonium or (3) the formation of a silicified initial cell within the auxospore (Kaczmarska et al., 2013). Importantly, AAE2 and AAE3 showed a background expression in vegetative conditions, suggesting their function is not completely auxospore-specific (**Figure 7C**).

Conclusions and future perspectives

Cylindrotheca closterium is a widespread species that inhabits numerous marine to brackish water coastal regions (Round et al., 1990) and has been the subject of many eco-physiological studies, illustrating its ability to adapt to varying temperature (Najdek et al., 2005; Stock et al., 2019) and salinity ranges (Apoya-Horton et al., 2006; Glaser & Karsten, 2020). Furthermore, a well-established crossing procedure is available, as well as detailed information on the cytology of different sexual cell types and assays to study pheromone communication (Klapper et al., 2021; Vanormelingen et al., 2013). Hence, the assembly of a reference genome and sex transcriptomes for *C. closterium* represent new resources to expand our understanding of its ecology, physiology and behaviour.

In this study, we created a chronological overview of the succession of genes during the consecutive stages of the sexual life cycle of benthic diatoms *Seminavis robusta* and *C. closterium*. Most cell cycle marker genes were no longer downregulated during gametogenesis, contrasting their response to SIP in *S. robusta* (Bilcke et al., 2020), most likely reflecting the

meiotic cell cycle during gametogenesis. Furthermore, we observed the upregulation of characteristic markers downregulated during a pheromone-induced cell cycle arrest (Annunziata et al., 2022; Bilcke et al., 2020). We hypothesise this may be linked to the active and more efficient mate finding in benthic diatoms (Klapper et al., 2021; Moeys et al., 2016), thus eliminating the need for a long cell cycle arrest as observed in planktonic diatoms, where a growth arrest can last for three days (Annunziata et al., 2022).

The existence of conserved sexual regulators such as Zn finger transcription factors and B-type cyclins in pennate and centric diatoms alike implies that the regulation of the sexual process has been conserved in the past 190 million years since the origination of diatoms in the late Jurassic (Nakov et al., 2018). Indeed, a potential zinc-finger transcription factor was induced during the early stage of sex in all four species, thus making it a prime candidate as master regulator of sexual cell development and size restoration in diatoms. Moreover, we identified another zinc-finger transcription factor that is among the conserved genes with an auxospore-associated expression (AAEs), although its auxospore-associated expression seems to be limited to the pennate clade. In general, the AAE genes were typically found in the genome of most diatom species (**Figure 7C**), which is in agreement with auxospore expansion being observed in every major clade of diatoms (Kaczmarska & Ehrman, 2021). The identification of the first putative genetic markers for the auxospore marks the first step in the molecular characterization of this unique cell type, which is the missing link that closes the unique size-dependent life cycle of diatoms. Furthermore, the creation of similar multi-species transcriptomes for other clades of diatoms such as the centrics and the araphid diatoms will be instrumental to investigate their unique strategies for mate finding and gametogenesis, such as oogamy in centrics or the peculiar pseudopodium appendage of the male gamete in araphids.

Acknowledgments

SA was supported by the Bijzonder Onderzoeksfonds (BOF-GOA 2015000402), while GB is a postdoctoral fellow supported by Fonds Wetenschappelijk Onderzoek (1228423N). WV acknowledges partial financial support from project FWO G001521N, BOF/GOA No. 01G01715 and infrastructure funded by EMBRC Belgium—FWO project GOH3817N. Collaboration in this project was supported through two ASSEMBLE+ transnational access grants: DIAREP and CRISPI.

The authors would like to express their gratitude towards Sofie D'hondt and Petra Bulankova for helpful advice about molecular techniques, Willem Stock, Sam De Decker and Renaat Dasseville for isolating the *C. closterium* reference genome strain CA1.15 and Peter Chaerle and Olga Chepurnova for cryopreservation of used strains. We thank Matthew Pinder and Mats Töpel for early access to their *S. marinoi* reference gene models.

References

Amato, A., Sabatino, V., Nylund, G. M., Bergkvist, J., Basu, S., Andersson, M. X., Sanges, R., Godhe, A., Kiørboe, T., Selander, E., & Ferrante, M. I. (2018). Grazer-induced transcriptomic and metabolomic response of the chain-forming diatom Skeletonema marinoi. *The ISME Journal*, *12*(6), 1594–1604. https://doi.org/10.1038/s41396-018-0094-0

Annunziata, R., Mele, B. H., Marotta, P., Volpe, M., Entrambasaguas, L., Mager, S., Stec, K., d'Alcalà, M. R., Sanges, R., Finazzi, G., Iudicone, D., Montresor, M., & Ferrante, M. I. (2022). Trade-off between sex and growth in diatoms: Molecular mechanisms and demographic implications. *Science Advances*, *8*(3), eabj9466. https://doi.org/10.1126/sciadv.abj9466

Apoya-Horton, M. D., Yin, L., Underwood, G. J. C., & Gretz, M. R. (2006). Movement Modalities and Responses to Environmental Changes of the Mudflat Diatom Cylindrotheca Closterium (bacillariophyceae)1. *Journal of Phycology*, *42*(2), 379–390. https://doi.org/10.1111/j.1529-8817.2006.00194.x

Armbrust, E. V. (1999). Identification of a New Gene Family Expressed during the Onset of Sexual Reproduction in the Centric Diatom Thalassiosira weissflogii. *Applied and Environmental Microbiology*, *65*(7), 3121–3128.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455. https://doi.org/10.1089/CMB.2012.0021

Basu, S., Patil, S., Mapleson, D., Russo, M. T., Vitale, L., Fevola, C., Maumus, F., Casotti, R., Mock, T., Caccamo, M., Montresor, M., Sanges, R., & Ferrante, M. I. (2017). Finding a partner in the ocean: Molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. *The New Phytologist*, *215*(1), 140–156. https://doi.org/10.1111/nph.14557

Benoiston, A. S., Ibarbalz, F. M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S., & Bowler, C. (2017). The evolution of diatoms and their biogeochemical functions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1728). https://doi.org/10.1098/rstb.2016.0397

Bilcke, G., Ferrante, M., Montresor, M., De Decker, S., Veylder, L., & Vyverman, W. (2022). Life Cycle Regulation. In *The Molecular Life of Diatoms*. Springer Netherlands. https://doi.org/10.1007/978-3-030-92499-7

Bilcke, G., Van den Berge, K., De Decker, S., Bonneure, E., Poulsen, N., Bulankova, P., Osuna, C., Dickenson, J., Sabbe, K., Pohnert, G., Vandepoele, K., Mangelinckx, S., Clement, L., Veylder, L., & Vyverman, W. (2020). Mating type specific transcriptomic response to sex inducing pheromone in the pennate diatom Seminavis robusta. *The ISME Journal*, *15*. https://doi.org/10.1038/s41396-020-00797-7

Bjerkan, K. N., Jung-Roméo, S., Jürgens, G., Genschik, P., & Grini, P. E. (2012). Arabidopsis WD REPEAT DOMAIN55 Interacts with DNA DAMAGED BINDING PROTEIN1 and Is Required for Apical Patterning in the Embryo[C][W]. *The Plant Cell*, *24*(3), 1013–1033. https://doi.org/10.1105/tpc.111.089425

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

731 Bucchini, F., Del Cortona, A., Kreft, Ł., Botzki, A., Van Bel, M., & Vandepoele, K. (2021).
732 TRAPID 2.0: A web application for taxonomic and functional analysis of de novo
733 transcriptomes. *Nucleic Acids Research*, *49*(17), e101.
734 https://doi.org/10.1093/nar/gkab565
735 Bulankova, P., Akimcheva, S., Fellner, N., & Riha, K. (2013). Identification of Arabidopsis
736 meiotic cyclins reveals functional diversification among plant cyclin genes. *PLoS*
737 *Genetics*, *9*(5), e1003508. https://doi.org/10.1371/journal.pgen.1003508
738 Carlile, T. M., & Amon, A. (2008). Meiosis I Is Established through Division-Specific
739 Translational Control of a Cyclin. *Cell*, *133*(2), 280–291.
740 https://doi.org/10.1016/j.cell.2008.02.032
741 Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2021). tRNAscan-SE 2.0: Improved detection
742 and functional classification of transfer RNA genes. *Nucleic Acids Research*, *49*(16),
743 9077–9096. https://doi.org/10.1093/nar/gkab688
744 Chen, I.-P., Mannuss, A., Orel, N., Heitzeberg, F., & Puchta, H. (2008). A Homolog of ScRAD5
745 Is Involved in DNA Repair and Homologous Recombination in Arabidopsis. *Plant*
746 *Physiology*, *146*(4), 1786–1796. https://doi.org/10.1104/pp.108.116806
747 Chepurnov, V. A., & Mann, D. G. (2004). Auxosporulation of Licmophora communis
748 (Bacillariophyta) and a review of mating systems and sexual reproduction in araphid
749 pennate diatoms. *Phycological Research*, *52*(1), 1–12. https://doi.org/10.1111/j.1440-
750 1835.2003.00319.x
751 Chepurnov, V. A., Mann, D. G., Sabbe, K., & Vyverman, W. (2004). Experimental studies on
752 sexual reproduction in diatoms. *International Review of Cytology*, *237*, 91–154.
753 https://doi.org/10.1016/S0074-7696(04)37003-8
754 Dainat, J., Hereñú, D., Davis, E., Crouch, K., LucileSol, Agostinho, N., pascal-git, & tayyrov.
755 (2022). *NBISweden/AGAT: AGAT-v0.9.2*. Zenodo.
756 https://doi.org/10.5281/zenodo.6621429
757 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson,
758 M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner.
759 *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635
760 Ferrante, M. I., Entrambasaguas, L., Johansson, M., Töpel, M., Kremp, A., Montresor, M., &
761 Godhe, A. (2019). Exploring Molecular Signs of Sex in the Marine Diatom Skeletonema
762 marinoi. *Genes*, *10*(7), E494. https://doi.org/10.3390/genes10070494
763 Font-Muñoz, J. S., Jeanneret, R., Arrieta, J., Anglès, S., Jordi, A., Tuval, I., & Basterretxea, G.
764 (2019). Collective sinking promotes selective cell pairing in planktonic pennate diatoms.
765 *Proceedings of the National Academy of Sciences of the United States of America*,
766 *116*(32), 15997–16002. https://doi.org/10.1073/pnas.1904837116
767 Gillard, J., Devos, V., Huysman, M. J. J., De Veylder, L., D'Hondt, S., Martens, C.,
768 Vanormelingen, P., Vannerum, K., Sabbe, K., Chepurnov, V. A., Inzé, D., Vuylsteke, M.,
769 & Vyverman, W. (2008). Physiological and transcriptomic evidence for a close coupling
770 between chloroplast ontogeny and cell cycle progression in the pennate diatom
771 Seminavis robusta. *Plant Physiology*, *148*(3), 1394–1411.
772 https://doi.org/10.1104/pp.108.122176
773 Glaser, K., & Karsten, U. (2020). Salinity tolerance in biogeographically different strains of the
774 marine benthic diatom Cylindrotheca closterium (Bacillariophyceae). *Journal of Applied*
775 *Phycology*, *32*(6), 3809–3816. https://doi.org/10.1007/s10811-020-02238-6
776 Goodenough, U., & Heitman, J. (2014). Origins of Eukaryotic Sexual Reproduction. *Cold*
777 *Spring Harbor Perspectives in Biology*, *6*(3), a016154.
778 https://doi.org/10.1101/cshperspect.a016154
779 Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1:
780 Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and

AUGUSTUS. *Bioinformatics*, *32*(5), 767. https://doi.org/10.1093/BIOINFORMATICS/BTV661

Hongo, Y., Kimura, K., Takaki, Y., Yoshida, Y., Baba, S., Kobayashi, G., Nagasaki, K., Hano, T., & Tomaru, Y. (2021). The genome of the diatom Chaetoceros tenuissimus carries an ancient integrated fragment of an extant virus. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-00565-3

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal. Biometrische Zeitschrift*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

Huysman, M. J., Fortunato, A. E., Matthijs, M., Costa, B. S., Vanderhaeghen, R., Van den Daele, H., Sachse, M., Inzé, D., Bowler, C., Kroth, P. G., Wilhelm, C., Falciatore, A., Vyverman, W., & De Veylder, L. (2013). AUREOCHROME1a-Mediated Induction of the Diatom-Specific Cyclin dsCYC2 Controls the Onset of Cell Division in Diatoms (Phaeodactylum tricornutum). *The Plant Cell*, *25*(1), 215–228. https://doi.org/10.1105/tpc.112.106377

Huysman, M. J., Martens, C., Vandepoele, K., Gillard, J., Rayko, E., Heijde, M., Bowler, C., & Inzé, D. (2010). *Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling*. 19.

Jones, H. M., Simpson, G. E., Stickle, A. J., & Mann, D. G. (2005). Life history and systematics of Petroneis (Bacillariophyta), with special reference to British waters. *European Journal of Phycology*, *40*(1), 61–87. https://doi.org/10.1080/09670260400024675

Kaczmarska, I., & Ehrman, J. M. (2021). Enlarge or die! An auxospore perspective on diatom diversification. *Organisms Diversity & Evolution*, *21*(1), 1–23. https://doi.org/10.1007/s13127-020-00476-7

Kaczmarska, I., Ehrman, J. M., Mills, K. E., Sutcliffe, S. G., & Samanta, B. (2022). Vegetative cell enlargement in selected centric diatom species – an alternative way to propagate an individual genotype. *European Journal of Phycology*, *0*(0), 1–18. https://doi.org/10.1080/09670262.2022.2112760

Kaczmarska, I., Poulíčková, A., Sato, S., Edlund, M., Idei, M., Watanabe, T., & Mann, D. (2013). Proposals for a terminology for diatom sexual reproduction, auxospores and resting stages. *Diatom Research*, *28*, 263–294. https://doi.org/10.1080/0269249X.2013.791344

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), Article 6. https://doi.org/10.1038/nmeth.4285

Karasu, M. E., Bouftas, N., Keeney, S., & Wassmann, K. (2019). Cyclin B3 promotes anaphase I onset in oocyte meiosis. *The Journal of Cell Biology*, *218*(4), 1265–1281. https://doi.org/10.1083/jcb.201808091

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Klapper, F., Audoor, S., Vyverman, W., & Pohnert, G. (2021). Pheromone Mediated Sexual Reproduction of Pennate Diatom Cylindrotheca closterium. *Journal of Chemical Ecology*, *47*, 504–512. https://doi.org/10.1007/s10886-021-01277-8

Kooistra, W., Gersonde, L., Medlin, L. K., & Mann, D. G. (2007). The origin and evolution of diatoms: Their adaptation to a planktonic existance. In *Evolution of primary producers in the Sea* (pp. 207–249). Elsevier.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736.

https://doi.org/10.1101/GR.215087.116/-/DC1
Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*(9), 3100–3108. https://doi.org/10.1093/nar/gkm160
Lewis, W. (1984). The diatom sex clock and its evolutionary significance. *American Naturalist*, *123*, 73–80.
Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191
Li, L., Wang, H., Wang, S., Xu, Y., Liang, H., Liu, H., & Sonnenschein, E. C. (2021). The Draft Genome of the Centric Diatom Conticribra weissflogii (Coscinodiscophyceae, Ochrophyta). *Protist*, *172*(5), 125845. https://doi.org/10.1016/j.protis.2021.125845
Li, Y., Wang, L., Zhang, L., He, Z., Feng, G., Sun, H., Wang, J., Li, Z., Liu, C., Han, J., Mao, J., Li, P., Yuan, X., Jiang, L., Zhang, Y., Zhou, Q., & Li, W. (2019). Cyclin B3 is required for metaphase to anaphase transition in oocyte meiosis I. *The Journal of Cell Biology*, *218*(5), 1553–1563. https://doi.org/10.1083/jcb.201808088
Liang, Y., Tan, Z.-M., Zhu, L., Niu, Q.-K., Zhou, J.-J., Li, M., Chen, L.-Q., Zhang, X.-Q., & Ye, D. (2013). MYB97, MYB101 and MYB120 Function as Male Factors That Control Pollen Tube-Synergid Interaction in Arabidopsis thaliana Fertilization. *PLOS Genetics*, *9*(11), e1003933. https://doi.org/10.1371/journal.pgen.1003933
Macdonald, J. D. (1869). I.—On the structure of the Diatomaceous frustule, and its genetic cycle. *Annals and Magazine of Natural History*, *3*(13), 1–8. https://doi.org/10.1080/00222936908695866
Makkena, S., Lee, E., Sack, F. D., & Lamb, R. S. (2012). The R2R3 MYB Transcription Factors FOUR LIPS and MYB88 Regulate Female Reproductive Development. *Journal of Experimental Botany*, *63*(15), 5545–5558. https://doi.org/10.1093/jxb/ers209
Mallik, R., Kundu, A., & Chaudhuri, S. (2018). High mobility group proteins: The multifaceted regulators of chromatin dynamics. *The Nucleus*, *61*(3), 213–226. https://doi.org/10.1007/s13237-018-0257-4
Mann, D. G., & Vanormelingen, P. (2013). An inordinate fondness? The number, distributions, and origins of diatom species. *Journal of Eukaryotic Microbiology*, *60*(4), 414–420. https://doi.org/10.1111/jeu.12047
Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England)*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011
Marotta, P., Borgonuovo, C., Santin, A., Russo, M. T., Manfellotto, F., Montresor, M., De Luca, P., & Ferrante, M. I. (2022). Mate Perception and Gene Networks Regulating the Early Phase of Sex in Pseudo-nitzschia multistriata. *Journal of Marine Science and Engineering*, *10*(12), Article 12. https://doi.org/10.3390/jmse10121941
Matsuo, T., & Ishiura, M. (2010). Chapter 6—New Insights into the Circadian Clock in Chlamydomonas. In *International Review of Cell and Molecular Biology* (Vol. 280, pp. 281–314). Academic Press. https://doi.org/10.1016/S1937-6448(10)80006-1
Mayr, E. (1942). *Systematics and the Origin of Species, from the Viewpoint of a Zoologist*. Harvard University Press.
Mazin, A. V., Mazina, O. M., Bugreev, D. V., & Rossi, M. J. (2010). Rad54, the motor of homologous recombination. *DNA Repair*, *9*(3), 286–302. https://doi.org/10.1016/j.dnarep.2009.12.006
Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., & Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, *34*(13), i142–i150. https://doi.org/10.1093/bioinformatics/bty266
Moeys, S., Gillard, J., Bouillon, B., Devos, V., Van den Berge, K., Frenkel, J., Huysman, M. J.

J., Pohnert, G., Clement, L., Sabbe, K., De Veylder, L., & Vyverman, W. (2016). A sex-inducing conditioning factor triggers cell cycle arrest and diproline production in Seminavis robusta. *Scientific Reports*, *6*(January), 1–13. https://doi.org/10.1038/srep19252

Najdek, M., Blažina, M., Djakovac, T., & Kraus, R. (2005). The role of the diatom Cylindrotheca closterium in a mucilage event in the northern Adriatic Sea: Coupling with high salinity water intrusions. *Journal of Plankton Research*, *27*(9), 851–862. https://doi.org/10.1093/plankt/fbi057

Nakov, T., Beaulieu, J. M., & Alverson, A. J. (2018). Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). *New Phytologist*, *219*(1), 462–473. https://doi.org/10.1111/nph.15137

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Oliver, A., Podell, S., Pinowska, A., Traller, J. C., Smith, S. R., McClure, R., Beliaev, A., Bohutskyi, P., Hill, E. A., Rabines, A., Zheng, H., Allen, L. Z., Kuo, A., Grigoriev, I. V., Allen, A. E., Hazlebeck, D., & Allen, E. E. (2021). Diploid genomic architecture of Nitzschia inconspicua, an elite biomass production diatom. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-95106-3

Osuna-Cruz, C. M., Bilcke, G., Vancaester, E., De Decker, S., Bones, A. M., Winge, P., Poulsen, N., Bulankova, P., Verhelst, B., Audoor, S., Belisova, D., Pargana, A., Russo, M., Stock, F., Cirri, E., Brembu, T., Pohnert, G., Piganeau, G., Ferrante, M. I., … Vandepoele, K. (2020). The Seminavis robusta genome provides insights into the evolutionary adaptations of benthic diatoms. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-17191-8

Patil, S., Moeys, S., von Dassow, P., Huysman, M. J. J., Mapleson, D., De Veylder, L., Sanges, R., Vyverman, W., Montresor, M., & Ferrante, M. I. (2015). Identification of the meiotic toolkit in diatoms and exploration of meiosis-specific SPO11 and RAD51 homologs in the sexual species Pseudo-nitzschia multistriata and Seminavis robusta. *BMC Genomics*, *16*, 930. https://doi.org/10.1186/s12864-015-1983-5

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Pfitzer, E. (1871). Untersuchungen über Bau und Entwicklung der Bacillariaceen (Diatomaceen. *Botanische Abhandlungen Aus Dem Gebiet Der Morphologie Und Physiologie*, *2*(198), 6.

Pinseel, E., Nakov, T., Van den Berge, K., Downey, K. M., Judy, K. J., Kourtchenko, O., Kremp, A., Ruck, E. C., Sjöqvist, C., Töpel, M., Godhe, A., & Alverson, A. J. (2022). Strain-specific transcriptional responses overshadow salinity effects in a marine diatom sampled along the Baltic Sea salinity cline. *The ISME Journal*, *16*(7), 1776–1787. https://doi.org/10.1038/s41396-022-01230-x

Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inzé, D., Mueller-Roeber, B., & Vandepoele, K. (2015). PLAZA 3.0: An access point for plant comparative genomics. *Nucleic Acids Research*, *43*(Database issue), D974-981. https://doi.org/10.1093/nar/gku986

Pryszcz, L. P., & Gabaldón, T. (2016). Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, *44*(12), e113. https://doi.org/10.1093/nar/gkw294

Rayko, E., Maumus, F., Maheswari, U., Jabbari, K., & Bowler, C. (2010). Transcription factor

families inferred from genome sequences of photosynthetic stramenopiles. *New Phytologist*, *188*(1), 52–66. https://doi.org/10.1111/j.1469-8137.2010.03371.x

Robinson, M. D., Mccarthy, D. J., & Smyth, G. K. (2010). EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Round, F. E., Crawford, R. M., & Mann, D. G. (1990). *The Diatoms: Biology and Morphology of the Genera.* Cambridge University Press. https://diatoms.org/citations/round{_}fe{_}crawford{_}rm{_}and{_}mann{_}dg-1990-the{_}diatoms{_}biology{_}and{_}morphology{_}of

Sato, S., Nanjappa, D., Dorrell, R. G., Vieira, F. R. J., Kazamia, E., Tirichine, L., Veluchamy, A., Heilig, R., Aury, J.-M., Jaillon, O., Wincker, P., Fussy, Z., Obornik, M., Muñoz-Gómez, S. A., Mann, D. G., Bowler, C., & Zingone, A. (2020). Genome-enabled phylogenetic and functional reconstruction of an araphid pennate diatom Plagiostriata sp. CCMP470, previously assigned as a radial centric diatom, and its bacterial commensal. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-65941-x

Scalco, E., Amato, A., Ferrante, M. I., & Montresor, M. (2016). The sexual phase of the diatom Pseudo-nitzschia multistriata: Cytological and time-lapse cinematography characterization. *Protoplasma*, *253*(6), 1421–1431. https://doi.org/10.1007/s00709-015-0891-5

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Soneson, C., Love, M. I., Robinson, M. D., & Floor, S. N. (2016). Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000 Research*, *4*(1521). https://doi.org/10.12688/f1000research.7563.1

Stock, W., Vanelslander, B., Rüdiger, F., Sabbe, K., Vyverman, W., & Karsten, U. (2019). Thermal niche differentiation in the benthic diatom Cylindrotheca closterium(Bacillariophyceae) complex. *Frontiers in Microbiology*, *10*(JUN), 1–12. https://doi.org/10.3389/fmicb.2019.01395

Team, R. C. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.gnu.org/copyleft/gpl.html.

Valenzuela, J., Mazurie, A., Carlson, R. P., Gerlach, R., Cooksey, K. E., Peyton, B. M., & Fields, M. W. (2012). Potential role of multiple carbon fixation pathways during lipid accumulation in Phaeodactylum tricornutum. *Biotechnology for Biofuels*, *5*(1), 40. https://doi.org/10.1186/1754-6834-5-40

Van den Berge, K., Soneson, C., Robinson, M. D., & Clement, L. (2017). stageR: A general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biology*, *18*(1), 151. https://doi.org/10.1186/s13059-017-1277-0

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., & Vandepoele, K. (2022). PLAZA 5.0: Extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Research*, *50*(D1), D1468–D1474. https://doi.org/10.1093/nar/gkab1024

Vanormelingen, P., Vanelslander, B., Sato, S., Gillard, J., Trobajo, R., Sabbe, K., & Vyverman, W. (2013). Heterothallic sexual reproduction in the model diatom Cylindrotheca. *European Journal of Phycology*, *48*(1), 93–105.

https://doi.org/10.1080/09670262.2013.772242
Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/GR.214270.116
Von Stosch, A. H. (1950). Oogamy in a Centric Diatom. *Nature*, *165*(4196), Article 4196. https://doi.org/10.1038/165531a0
Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics (Oxford, England)*, *33*(14), 2202–2204. https://doi.org/10.1093/bioinformatics/btx153
Wang, W., Fang, H., Aslam, M., Du, H., Chen, J., Luo, H., Chen, W., & Liu, X. (2022). MYB gene family in the diatom Phaeodactylum tricornutum revealing their potential functions in the adaption to nitrogen deficiency and diurnal cycle. *Journal of Phycology*, *58*(1), 121–132. https://doi.org/10.1111/jpy.13217
West-Eberhard, M. J. (1979). Sexual Selection, Social Competition, and Evolution. *Proceedings of the American Philosophical Society*, *123*(4), 222–234.
Yan, G.-X., Dang, H., Tian, M., Zhang, J., Shodhan, A., Ning, Y.-Z., Xiong, J., & Miao, W. (2016). Cyc17, a meiosis-specific cyclin, is essential for anaphase initiation and chromosome segregation in Tetrahymena thermophila. *Cell Cycle*, *15*(14), 1855–1864. https://doi.org/10.1080/15384101.2016.1188238
Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Data Accessibility and Benefit Sharing

Data Accessibility

Raw Illumina and MinION reads used for genome assembly are available at the European Nucleotide Archive (ENA) with project number PRJEB49951. Meanwhile, RNA-seq libraries describing control and sexualized *C. closterium* samples are available at ENA with project number PRJEB49955. Additional vegetative RNA-sequencing samples generated for gene prediction (**Table S1**) are available at PRJEB50488 and new amplicon sequences generated in this study were deposited in GenBank OQ948421 (CA1.15) and OQ948422 (ZW2.20).

The final genome assembly, CDS and protein sequences are available through ORCAE at https://bioinformatics.psb.ugent.be/gdb/Cylindrotheca_closterium/Version1.2/. Furthermore,

both the final assembly and genome annotation were submitted to the ENA project PRJEB49951 under accession number GCA_933822405. The *S. marinoi* V1 reference genome is available from NCBI with accession number JATAAI000000000, and the *S. marinoi* gene models v1.1.2 (Pinder et al., in prep.) used for mapping in this study are available as Supplementary data. Differential expression calls for each of the four diatom species as well as functional and homology data obtained from the 20-species sequence database (InterPro domains, gene family prediction, one-to-one orthologs between *C. closterium* and *S. robusta*) are available as Supplementary Data.

Benefit Sharing

Benefits from this research accrue from the sharing of our data and results on public databases as described above.

Author contributions

SA, GB, KV and WV conceived the study. SA and DB were responsible for *C. closterium* amplicon and whole-genome sequencing. KP and GB performed genome assembly, gene prediction and functional annotation. ST and MVB built the custom diatom gene database, including the prediction of gene families and orthologs. LS further cleaned the *C. closterium* genome assembly and provided online access through the ORCAE platform. NR was responsible for microscopy of vegetative and sexual *C. closterium* cells. SA performed the RNA-sequencing experiment. GB and SA carried out all differential expression and comparative transcriptomic analyses. RA and MF contributed to the strategy of comparing sex transcriptomes across diatom species. SA and GB jointly wrote the manuscript with input from WV and KV. The final manuscript was read and reviewed by all co-authors.