Molecular fingerprints of cell size sensing and mating type differentiation in pennate diatoms

Darja Belišová*[1], Gust Bilcke*[1,2,3], Sien Audoor[1,a], Sofie D'hondt[1], Lieven De Veylder[2,3], Klaas Vandepoele[2,3,4], Wim Vyverman[1†]

[1] Protistology and Aquatic Ecology, Department of Biology, Ghent University, 9000 Ghent, Belgium

[2] Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium

[3] VIB Center for Plant Systems Biology, 9052 Ghent, Belgium

[4]VIB Center for AI & Computational Biology, VIB, Ghent, Belgium

[a]Present address: Institut de Biologie Physico-Chimique, Laboratory of Chloroplast Biology and Light Sensing in Microalgae, UMR7141 Centre National de la Recherche Scientifique (CNRS), Sorbonne Université, 75005 Paris, France

[†]Corresponding author - wim.vyverman@ugent.be

*These authors contributed equally to this work

The authors declare that they have no conflict of interest.

Abstract

A unique cell size sensing mechanism is at the heart of the life cycle of diatoms. During population growth, cell size decreases until a Sexual Size Threshold (SST) is reached, below which cells become sexually competent. In most pennate diatoms, two genetically determined mating types undergo biochemical and behavioral differentiation below the SST, although the molecular pathways underlying their size-dependent maturation remain unknown. Here, we developed a method to shorten the generation time of *Cylindrotheca closterium* through experimental cell size manipulation, allowing the comparison of genetically identical large and undifferentiated cells with small, sexually competent cells. Transcriptome profiling of six different genotypes revealed a set of 112 size- and mating type-responsive genes. We identified 21 genes upregulated in small cells regardless of the mating type, revealing how cells undergo significant transcriptional reprogramming when passing the SST. Among genes with a mating type specific response, about a third was only expressed below the SST, including a cluster with three sex inducing pheromone-responsive genes. In addition, comparative transcriptomics with the pennate diatom *Seminavis robusta* revealed a shared mating type specificity of Mating-type Related Minus 2 (MRM2) homologs. A combination of phylogenetics, protein structure prediction and transcriptomic profiling during reproduction suggests that these leucine-rich repeat receptor-like proteins are part of a conserved partner-recognition mechanism in pennate diatoms. In conclusion, this study illustrates how different mating types acquire sexual competence in a strictly size-dependent manner, revealing part of the complex machinery underlying size-dependent maturation, mating behavior and the origin of heterothally in pennate diatoms.

Keywords

Diatoms; Microsurgery; Cell size; Life cycle; Mating type differentiation; Comparative transcriptomics; *Cylindrotheca closterium*; *Seminavis robusta*

Introduction

Diatoms are an extremely successful group of unicellular primary producers in freshwater and marine environments [1]. Despite their extraordinary species diversity underlying the colonization of widely different habitats spanning terrestrial to aquatic realms, all diatoms share a unique silica cell wall consisting of two halves (thecae) slightly differing in size. During cell division, the two daughter cells each inherit one of the rigid parental thecae that are dissimilar in size and fit as a "lid on a box" (**Figure 1**). As a result, an asexually growing population of cells will gradually decrease in mean cell size, a phenomenon known as the MacDonald-Pfitzer rule [2, 3]. Different mechanisms of cell size restoration mechanisms have been described [4–6]; but most commonly this happens via the sexual phase. The ability to engage in sexual reproduction is strictly cell-size controlled and triggered once cells drop below a species-specific Sexual Size Threshold (SST). The cellular mechanisms and molecular changes associated with this unique ability of diatom cells to sexually mature by measuring their own cell size are still completely unknown.

In the generally homothallic centric diatoms, cells produce oogonia and sperm cells, while in raphid pennate diatoms vegetative cells become responsive to sexual cues after passing the SST and transform into morphologically identical but functionally differentiated gametangial cells [7–10]. Recent advances revealed the complex nature of diatom sexual behavior, including the identification of multiple sex pheromones involved in the mating process [7, 10–12].

Most raphid pennate diatoms have a heterothallic mating system with two mating types (MT+ and MT-) (reviewed in [5, 13–15]). Although the elucidation of sex determination systems is

still in its infancy, recent studies in *Seminavis robusta* and *Pseudo-nitzschia multistriata* have demonstrated the existence of a single sex-determining locus that drives the expression of downstream mating type related genes [16, 17]. Specifically, the sex-determining A allele of the *P. multistriata* Mating-type Related Plus 3 (*MRP3*) gene segregates with MT+ strains [17]. A simple control mechanism, where differential methylation of the sex locus leads to mono-allelic expression of *MRP3* only in MT+, has been described by Ruggiero et al. (2023). MRP3 acts as a master regulator, activating downstream MT+ genes *MRP1* and *MRP2*, while inactivating the MT- specific genes *MRM1* and *MRM2*. Importantly, despite the widespread occurrence of heterothally in pennate diatoms, no homologs of this sex determining system have yet been identified in other pennate species outside the *Pseudo-nitzschia* - *Fragilariopsis* clade [17].

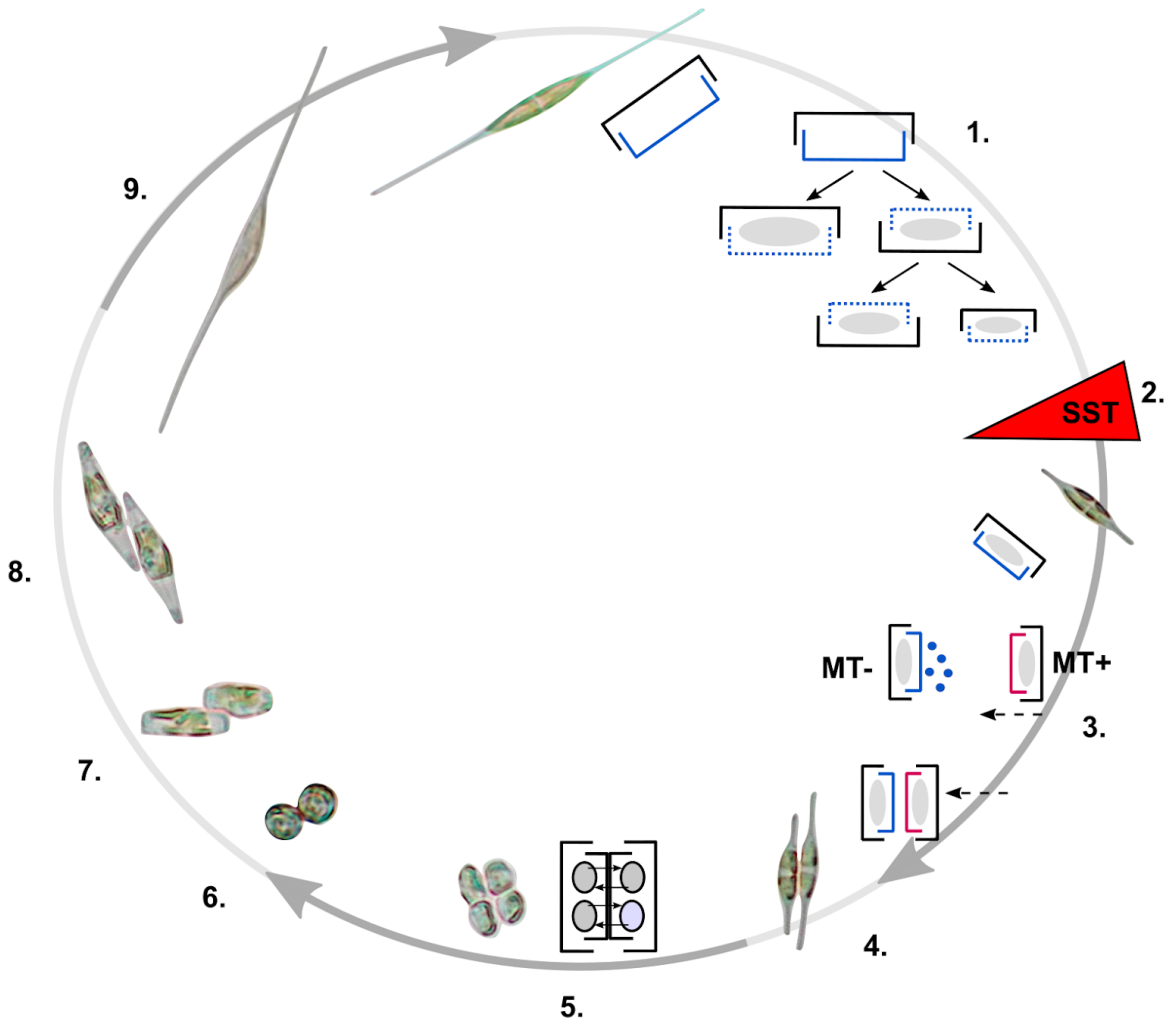

Figure 1: **A typical life cycle of a raphid pennate diatom**, illustrated with the species *C. closterium*. (**1**) Cells divide vegetatively and the mean size in population drops below the sexual size threshold (SST) (**2**). Small cells become sexually mature and seek a compatible mate (**3**). In *C. closterium*, mating type minus (MT-) produces pheromone (SIP-, blue dots) while mating type plus (MT+) evinces oriented motility towards the partner (**3**,**4**). (**5**) Each gametangial cell produces two gametes, but one of the mating types is suggested to be heterogametic and produces gametes determining the sex of the progeny. (**6**) Gametes fuse and form zygotes (**7**). Zygotes expand and become auxospores (**8**) and finally an initial cell starts a new generation.

Here, we studied size sensing and mating type differentiation in the raphid pennate diatom *Cylindrotheca closterium*, a common and globally widespread member of marine, brackish water phytoplankton and microphytobenthos [18]. We developed a protocol to significantly shorten generation time by mechanically reducing cell length, allowing us to compare gene expression of cells above and below the SST for the same genotypes. We carried out differential gene expression analysis using robust transcriptome profiling of six different genotypes, encompassing both mating types in both size variants. Phylogenetic analysis and cross-species transcriptomic analysis were used to reveal conserved size- and mating type-specific genes.

Materials & Methods

Culture conditions, genotypes and size manipulation of *C. closterium* and *S. robusta* cells

Cultures of *C. closterium* and *S. robusta* were grown in artificial seawater (ASW) composed of Tropic Marin BIO-ACTIF sea salt (34.5 g/l; Tropic Marin, Wartenberg, Germany) and sodium bicarbonate (0.08 g/l) supplemented with Guillard's F/2 Marine Water enrichment solution (Sigma-Aldrich, Belgium) and an antibiotic mixture (500 mg/l penicillin, 500 mg/l ampicillin, 100 mg/l streptomycin and 50 mg/l gentamicin). Cells were sustained under constant illumination (25 µmol/m2/s) at 18°C, if not stated otherwise. An initial set of strains was provided by the Belgian Coordinated Collection of Microorganisms (BCCM/DCG) (https://bccm.belspo.be/about-us/bccm-dcg) (**Supplementary Table 1**).

Four parental *C. closterium* strains of two different mating types were crossed to create progeny strains. Monoclonal progeny strains were isolated and their subpopulations were reduced in size using a novel microsurgery approach. Particularly, long apices of single cells were cut off with a sharp and extremely thin tip of a glass Pasteur pipette, with an open diameter ranging between

50 to 200 µm. Such manipulated cells were immediately isolated in a 96 well-plate and maintained in low light conditions (10 – 15 µmol photons/m2/s) to form new monoclonal populations. Altogether, six *C. closterium* strains, three of each mating type, were selected for RNA-sequencing (RNA-seq) of large and small size classes (**Supplementary Table 1**). For *S. robusta*, six strains below the SST, three of each mating type, were selected for additional RNA-seq (**Supplementary Table 1**) to complement existing datasets [19]. More details about strains and their manipulation can be found in **Supplementary Methods**.

RNA extraction and sequencing

Exponentially growing cultures of *C. closterium* and *S. robusta* were harvested on membrane filters (3µm, 25mm, Versapor®, Pall Laboratory, VWRTM) and RNA was extracted using a Qiagen RNeasy extraction kit with optimizations. Detailed protocols can be found in **Supplementary Methods**.

Samples with the highest RNA integrity number and sufficient quantity were subjected to 2x150bp paired-end sequencing using the Illumina NovaSeq 6000 instrument at VIB Nucleomics Core (Leuven, Belgium www.nucleomics.be). On average, 33.8 +- 4.4 million and 32.0 +- 3.7 million reads per sample were sequenced for *C. closterium* and *S. robusta* respectively.

Differential expression analysis and functional annotation

Quality trimmed paired-end reads were mapped to transcripts belonging to the gene annotation model v1.2 of the *C. closterium* reference genome [18] using Salmon v1.3 [20], as explained in **Supplementary Methods**. Read counts were imported into R with tximport v1.22.0 [21], followed by filtering, normalization and generalized linear model fitting as detailed in the **Supplementary Methods**. For *C. closterium*, we defined four contrasts to execute a differential

expression (DE) analysis with EdgeR v3.36.0 [22] (**Figure 2**): (1) cell sizes above and below the SST irrespective of the MT, (2) cells of MT+ versus MT- irrespective of the size, (3) cells of MT+ below the SST versus all other samples, and (4) cells of MT- below the SST versus all other samples. Next, likelihood ratio tests were performed using EdgeR, and p-values were controlled at the 5 % false discovery rate (FDR) level using the Benjamini-Hochberg correction. The analysis of existing [19] and newly generated RNA-seq data of *S. robusta* used for comparative purposes is explained in **Supplementary Methods**. In addition, differential expression calls during the different stages of sexual reproduction of *C. closterium* and *S. robusta* were retrieved from [18].

Protein domain and gene family annotation of *C. closterium* genes was performed using TRAPID 2.0 [23], with default parameters (E-value threshold 10^{-5}) and PLAZA Diatoms as a database for similarity searches. Meanwhile, InterPro and gene family functional annotation for *S. robusta* was retrieved from PLAZA Diatoms 1.0 [19]. Additionally, computational predictions of subcellular localization, phylostratigraphic gene age as well as enrichment analyses were performed, as detailed in the **Supplementary Methods**.

Quantitative reverse transcriptase PCR (qRT-PCR) analysis of response to sex-inducing pheromone

A qRT-PCR experiment was performed on three MT+ specific genes (Ccl_14859, Ccl_14861 and Ccl_14874) to assess their response to sex inducing pheromone of the opposing mating type. Prior to the experiment, culture of MT+ strain ZW2.20 (MT+, DCG 0922) growing in the mid-exponential phase was placed in the dark for 24 h and SIP containing medium from MT- strain CA1.15 (MT-, DCG 0923) was harvested as described previously [10]. Two hours before re-illumination, the supernatant medium of ZW2.20 was replaced with either filtered medium from its mating partner CA1.15 (treatment) or fresh culture medium (control). ZW2.20 cultures

were harvested at six time points (0, 15 min, 1h, 3h, 6h, 9h) after re-illumination and RNA was extracted according to the protocol described in [18]. Details about cDNA preparation, cycling conditions and statistical analysis of the results can be found in **Supplementary Methods**.

Phylogenetic analysis of MRM2 homologs

A hidden Markov model (HMM) was constructed to define the MRM2 gene family (HOM02SEM009752) and identify more distant homologs. A protein sequence alignment created with MAFFT v7.453 [24] was trimmed to retain only the conserved LRR domain, after which a profile HMM was created using the hmmbuild command of HMMER (hmmer.org). The profile HMM is included in **Supplementary Data** as a reference for the MRM2 family. Next, a targeted search for MRM2 homologs in eukaryotic genomes was performed with the hmmsearch command and the top-50 proteins with the highest domain-specific E-value score were selected for phylogenetic analysis. A detailed pipeline can be found in **Supplementary Methods**.

Data availability

New RNA-seq libraries generated for *C. closterium* and *S. robusta* are available at ENA with project number PRJEB52586. Differential expression calls for both *C. closterium* and *S. robusta* as well as gene family predictions for *C. closterium* , curated gene models of MRM2 family proteins in selected pennate diatoms, and the MRM2 family HMM model are available as **Supplementary Data**.

Results & Discussion

Transcriptome profiling reveals size- and mating type-dependent gene expression

We profiled gene expression of large- and small-sized cells for three independent *C. closterium* cultures of each mating type. To identify only genes that show a robust response across strains, three different genotypes of MT+ and MT- each were used as biological replicates. Small cells were generated by microsurgery of the apices of large cells, shortening the time required for cells to become smaller than the SST (**Figure 2A**). Using this technique, we obtained paired cultures for each genotype, where the cells before size manipulation measured on average 76+-7µm (>SST) and afterwards 36+-8µm (<SST). All cultures were tested for their ability to engage in sexual reproduction prior to the experiment. In none of the large-celled cultures, sex was induced after mixing compatible mating types, while this was the case in all cultures whose cell size was experimentally decreased.

To desynchronize the cell cycle and thus prevent the enrichment of a particular cycle phase cultures were maintained in constant light regime. All samples were mapped to the reference genome of *C. closterium* strain CA1.15 (MT-) [18]. The average mapping rate was 62.5 %, with no significant differences in mapping rate between large (63.1+-8.0%) and small cells (61.9+-4.9%) or between MT+ (62.2+-7.0%) and MT- (62.9+-6.4%) ($p>0.05$, two-sample t-test). Multidimensional scaling showed clustering of samples by genotype and not by condition (size or mating type), in agreement with recent observations of significant genotype-specific differences in expression patterns of the centric diatom *Skeletonema marinoi* [25] (**Supplementary Figure 1**).

Differential expression (DE) analysis was performed over four contrasts to detect genes responsive to cell size, mating type, or the interaction of cell size and mating type (**Figure 2B**). Overall, a set of just 112 out of 24.633 annotated genes was identified as differentially expressed in one or more of these contrasts. Given that these DE genes show a conserved expression

pattern throughout three independent genotypes for each mating type, we expect a significant role for these genes in the size- and mating type-related response in *C. closterium*. The small number of DE genes suggests that transcriptomic differentiation between cell size classes and mating types in diatoms is limited to specific traits related to size sensing and behavioral sexual dimorphism, while the overall metabolism of cells remained unchanged.

Whereas 21 genes were upregulated in small cells regardless of their MT, no genes were upregulated in large cells. The cellular changes that take place in both mating types when cells pass the SST thus appears to be mainly driven by the activation of genes below rather than by suppression of genes expressed above the SST.

A sizable number of genes was differentially expressed between mating types regardless of their size: 8 and 54 genes were induced in MT- and MT+, respectively (**Figure 2C**). Hence, the two mating types can be distinguished by their unique gene expression profiles throughout their entire life cycle, i.e. across large and small stages. Once cells decrease below the SST, an additional activation of mating type specific genes occurs. Indeed, 29 genes were upregulated in cells of a specific mating type but only below the SST, with a marked predominance of MT+ specific genes (26 versus only 3 MT- specific) (**Figure 2C**). Hence, size-dependent gene expression changes also reflect the physiological and behavioral differentiation between mating types, even in the absence of a compatible mating partner.

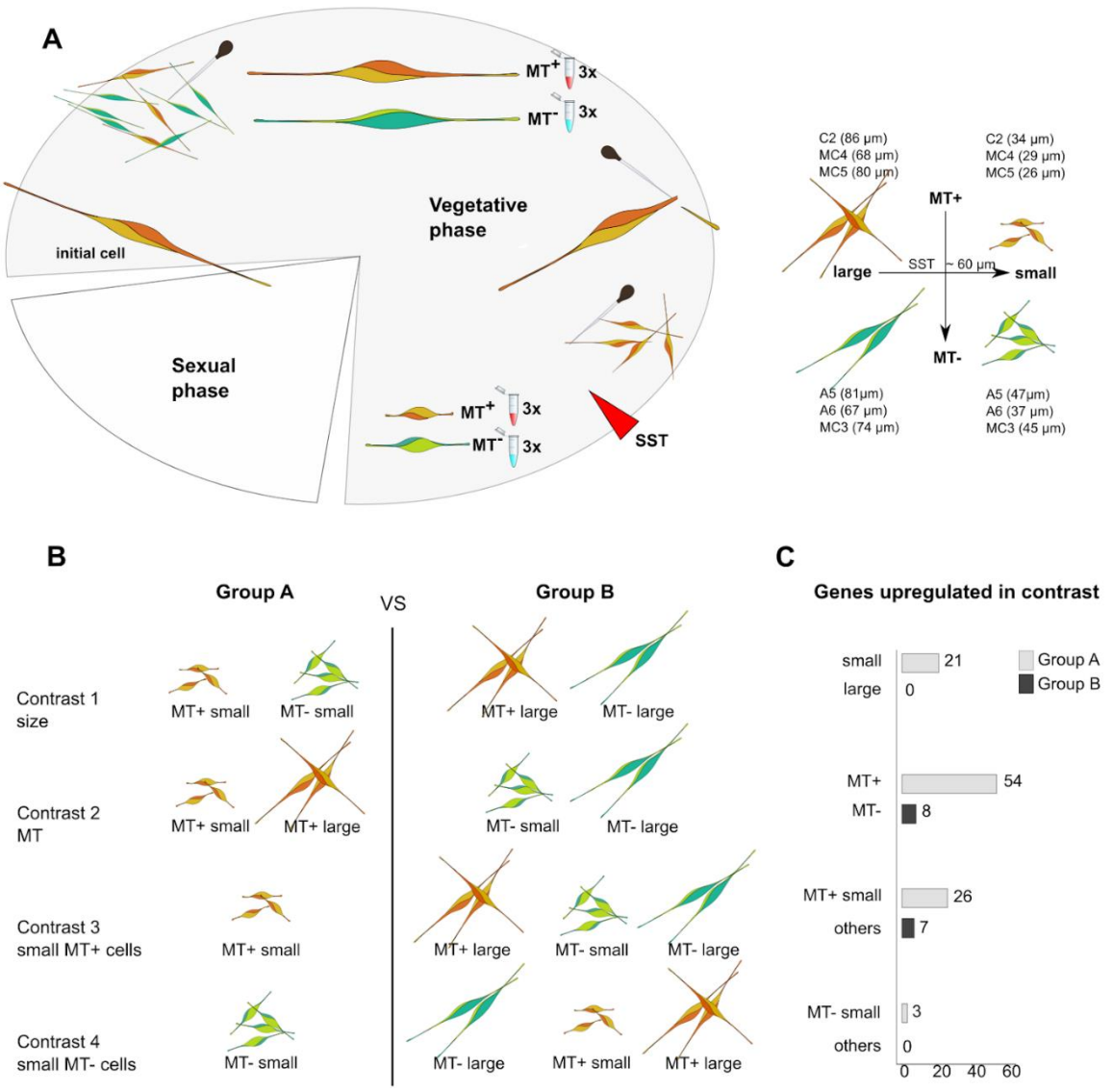

Figure 2: Illustration of the microsurgery-based set-up for size manipulation of *C. closterium*, experimental design of RNA sequencing and contrasts tested in the differential expression analyses. (**A**) Large single cells were isolated to form monoclonal vegetative cultures. These were manipulated in size to drop below the sexual size threshold (SST) and significantly reduce life cycle duration. In such a manner cells above and below SST of the same genotype were cultured simultaneously and harvested for RNA-seq. Cells are coloured to represent the different mating types and size classes, with the line between two shades representing the raphe. (**B**) Transcriptomic differences between mating types, size classes and their interaction were studied by differential expression analysis over four different contrasts.

(**C**) Bar plots showing the number of differentially expressed *C. closterium* genes for each contrast, each time subdivided in genes upregulated in group A vs group B, and genes upregulated in group B vs group A.

Regulatory genes and retrotransposon activity mark small and MT+ cells respectively

For 72 of the differentially expressed genes, InterPro protein domains and/or PLAZA Diatoms gene families could be assigned, whereas 40 genes in our dataset remained without functional annotation (**Supplementary Data**). Enrichment tests were performed to identify InterPro domains and gene families that were significantly more common among the DE genes compared to the background of all expressed genes. We detected eight gene families and five InterPro domains that were significantly enriched among the DE genes for specific contrasts (**Figure 3**). Moreover, one additional gene family was enriched in the set of all DE genes.

Genes upregulated in small cells of both mating types appear to be dominated by regulatory functions. These include a family encoding DNA-binding helicase-like proteins (HOM02SEM000018) (**Figure 3**). Several other genes that interact with DNA or RNA were upregulated in small cells: a nuclease (Ccl_23878), a 5'-Nucleotidase (Ccl_12722) and a ribonuclease H (Ccl_21492). Furthermore, five genes upregulated in cells below the SST contain a potential nuclear localization signal, among which two genes from the family of unknown function HOM02SEM018909. The identification of putative regulatory genes that are transcribed in both mating types below the SST is especially relevant in light of recent reports that the sex-determining gene *MRP3* of *P. multistriata* and a methyltransferase situated nearby are differentially methylated between mating types [26]. A general size-responsive transcription factor is thought to preferentially activate the non-methylated sex-determining allele, causing

mating type differentiation below the SST [26]. Hence, the nuclear and DNA-binding genes identified here are prime candidates as size-dependent molecular switches in diatoms.

Two families of proteins containing reverse transcriptase and DNA integrase domains, indicating a retrotransposon origin, were significantly enriched among MT+ biased genes: HOM02SEM000149 and HOM02SEM000048 (**Figure 3**). Indeed, reverse transcriptase and integrase InterPro domains were significantly more abundant among MT+ biased genes than in the rest of the expressed transcriptome (**Figure 3**). The role of retrotransposons in the evolution of mating type identity may be similar to mechanisms observed in other eukaryotic groups where the mating type locus is protected through suppression of recombination caused by retrotransposon accumulation [27, 28]. Alternatively, an overall higher activity of (former) transposable elements in MT+ may be the result of differences in chromatin accessibility between mating types similar to the above mentioned differential methylation in *P. multistriata* and as is the case in sex chromosomes of higher animals and plants [29, 30].

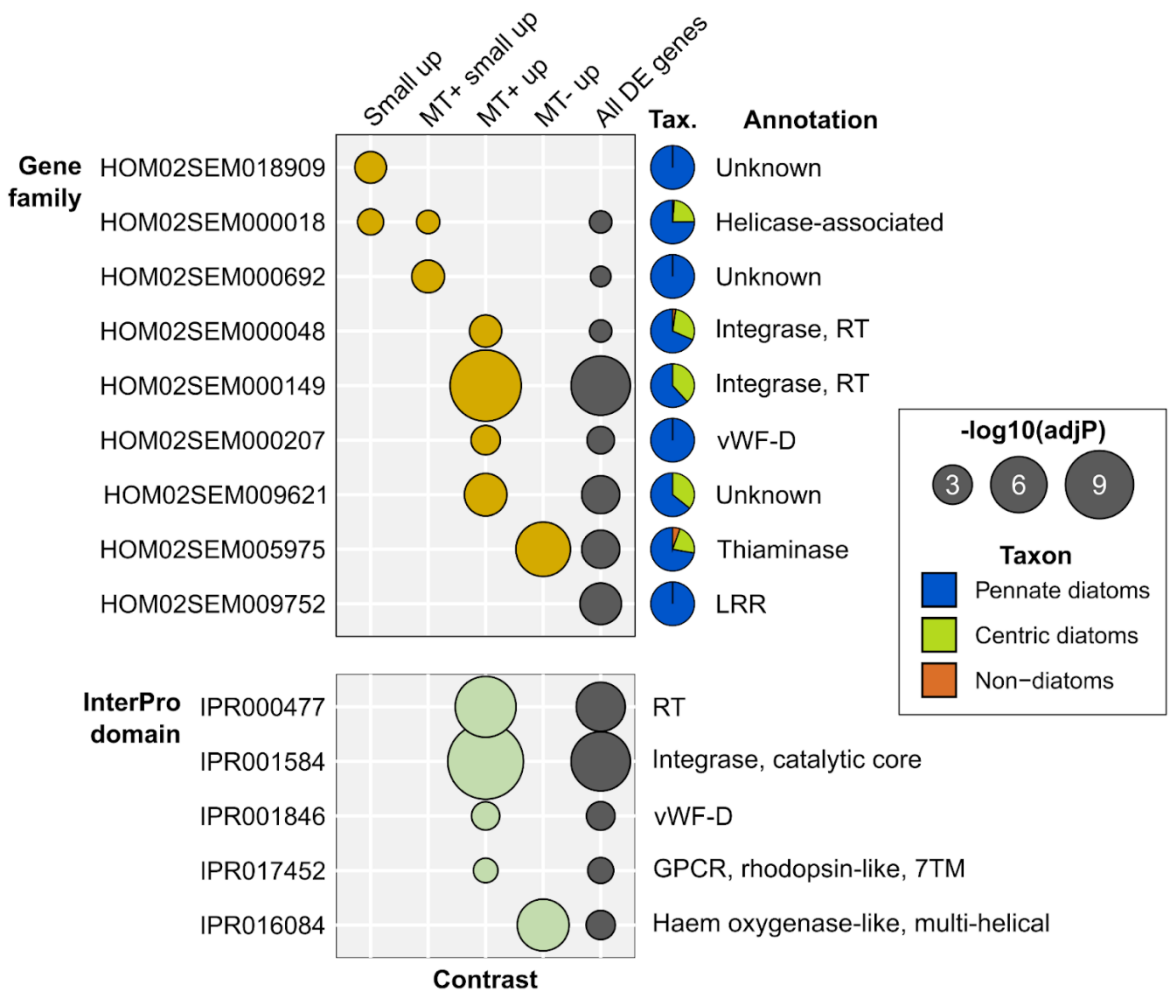

Figure 3: Enrichment of gene families and protein domains among differentially expressed (DE) genes from different contrasts (x-axis). Features that are significantly enriched in a contrast are shown by circles, whose size scales with the significance level (-log10). Pie charts show the taxonomic composition of gene families. MT: mating type RT: reverse transcriptase, PDE: phosphodiesterase, GPCR: G-protein coupled receptor, LRR: Leucine-rich repeat, vWF-D: Von Willebrand factor type D, TM: transmembrane domain.

A gene cluster of MT+ and small size specific genes plays a role during sex pheromone signaling

Among the differentially expressed genes for size and mating type, we identified three *C. closterium* genes that are located nearby on contig81: Ccl_14859, Ccl_14861 and Ccl_14874 (**Figure 4A**). Gene expression of each of the three genes was extremely specific for small cells and restricted to the MT+ (**Figure 4A**). Interestingly, we found that the three aforementioned genes were hyperactive during sexual reproduction in available RNA-seq data (**Figure 4A**) [18]. Inspecting the expression pattern over time, we observed that the induction of Ccl_14859 and Ccl_14861 was the strongest at the first time point T1, which is the stage when sex pheromone signaling, mate finding and mating pair formation is most intense (**Figure 4A**). The expression of Ccl_14874 peaked somewhat later: between stage T1 and T2, when gametogenesis is at its maximum. Since the rapid induction of these three genes during the early stages of mating may indicate a role in the response to sex inducing pheromone (SIP) of the opposite mating type [10], we set up a time-series RT-qPCR experiment (**Figure 4B**). The expression of each of the three genes was significantly induced by a treatment with SIP-containing filtrate. Interestingly, while Ccl_14859 and Ccl_14861 showed their strongest upregulation already 3h after treatment with SIP, the expression pattern of Ccl_14874 peaks after 9h, at the end of the time series.

In terms of functional annotation, the amino acid sequence of Ccl_14859 shares similarity with a *P. tricornutum* phytotransferrin (Phatr3_J54465) [31, 32] and encodes an N-terminal signal peptide. The other two genes (Ccl_14861, Ccl_14874) which differ only by two amino acids at positions 303 and 306, had no functional annotation except for an N-terminal signal peptide, indicating a secretory pathway. Furthermore, no homologs could be detected in any other diatom genome (BLASTp, E < 1e-3), making these two genes highly taxon-specific markers for MT+ identity.

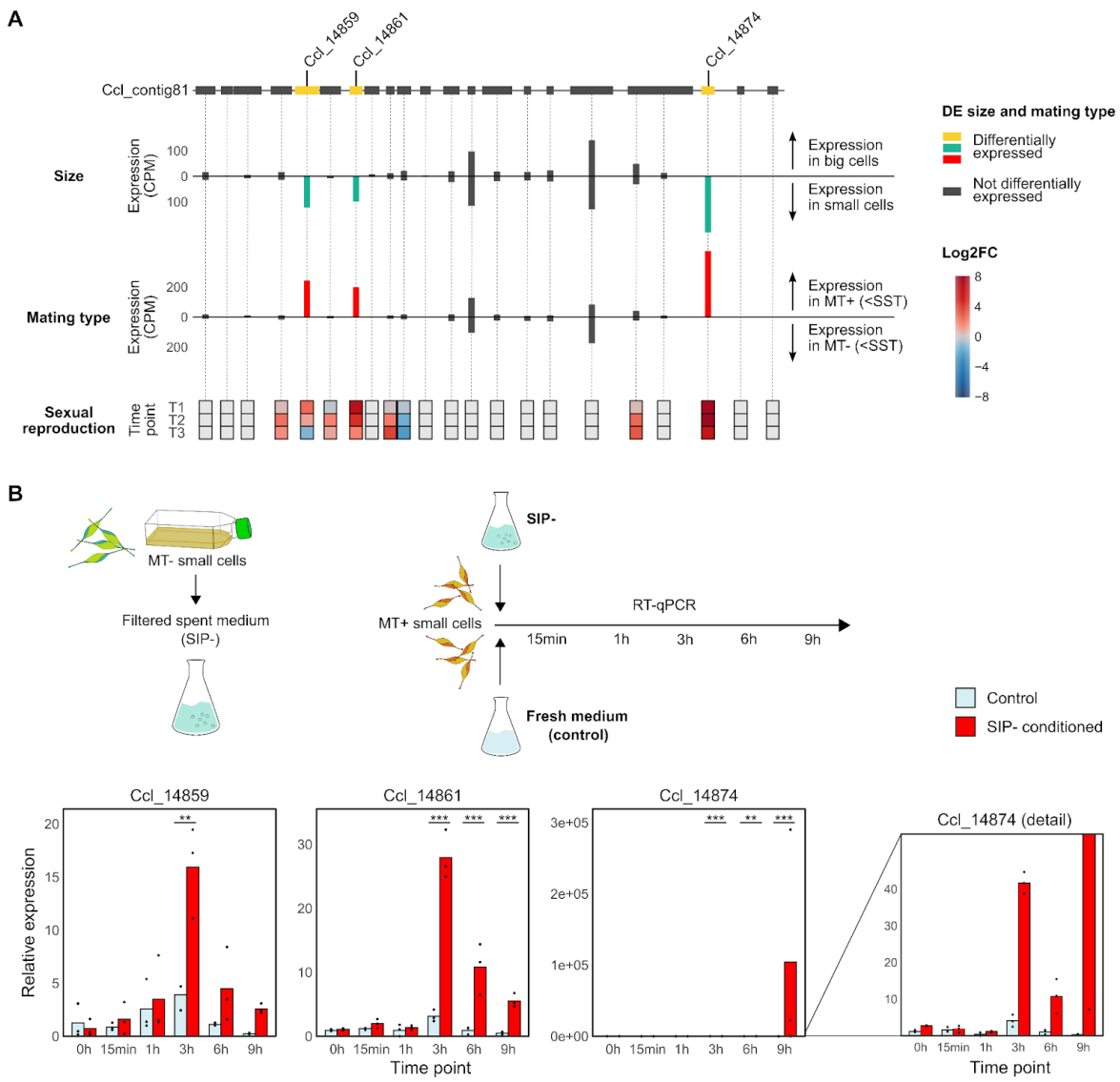

Figure 4: Gene expression of adjacent genes on *C. closterium* contig81 in different size classes, mating types and during three stages of sexual reproduction. (**A**) Bar plots show the average expression over replicates. CPM: counts per million, MT: mating type, SST: sexual size threshold. (**B**) Bar plots showing relative expression throughout a time series of 9h following a treatment of MT+ cells with Sex Inducing Pheromone (SIP) containing spent medium. RT-qPCR: reverse transcriptase quantitative PCR. Stars indicate significant differences in expression between SIP treated and control conditions (pairwise Wald tests on generalized linear model with Poisson distribution, * p<0.05, ** p<0.01, *** p<0.001).

The presence of three closely located genes with a shared MT+ specific and SIP-responsive properties suggests a shared function during early mate finding. Functional gene clusters are commonly found in other eukaryotic taxa such as plants, animals and fungi [33], playing a role in the same cellular process, often a metabolic pathway. Co-expression of clustered genes is typically a result of co-regulation by the same transcription factors, through shared binding sites or common chromatin-level regulation [33]. Further studies are warranted to elucidate the functional significance and regulatory elements associated with this gene cluster.

Mating type differentiation and size sensing genes are associated with pennate- and *Cylindrotheca*-aged genes

Phylostratigraphic analysis, a method used to date the origin of genes or gene families by looking at their homologs across different species, revealed that size- and mating type-responsive DE genes were on average evolutionarily young compared to the entire expressed transcriptome. Indeed, genes upregulated in small cells, in one or both mating types, were significantly enriched in genes that lack homologs outside *C. closterium* (**Figure 5A**). These genes are most likely involved in the species-specific pheromone signaling and mate finding behavior, which become only apparent below the SST [10].

Size-sensing and size-dependent sexual maturation have been observed for most diatom species from both centric and pennate clades [5, 15]. Nevertheless, none of the 21 purely size-responsive genes we found in *C. closterium* were of diatom age, i.e., whose gene family was restricted to pennate and centric diatoms (**Figure 5A**). This suggests that size-sensing in diatoms does not share a common transcriptionally controlled master regulator or that its sequence is strongly diverged between clades. In other eukaryotic organisms, physicochemical signals such as the scaling of the nucleus and membrane-free structures, i.e., nucleolus, spindle, and microtubules, have been used to measure their size in comparison to the cell size, the cell

volume or the surface area [34, 35]. Hence, mating type differentiation and sexual maturation may proceed directly downstream of such physical signals through intracellular scaling without a transcriptionally regulated master regulator.

Genes that were differentially expressed between mating types but only in small cells were significantly more likely to be of pennate age compared to all other *C. closterium* genes (**Figure 5A**). Indeed, many (30%) of the genes involved in mating type differentiation below the SST were rooted within the pennate diatoms. Given that heterothally is a common reproductive strategy in pennate diatoms while centric diatoms typically display homothallic reproduction without mating types [15], these results suggest that the molecular mechanisms underlying mating type determination have a shared origin in the pennate diatoms around 137 million years ago [36].

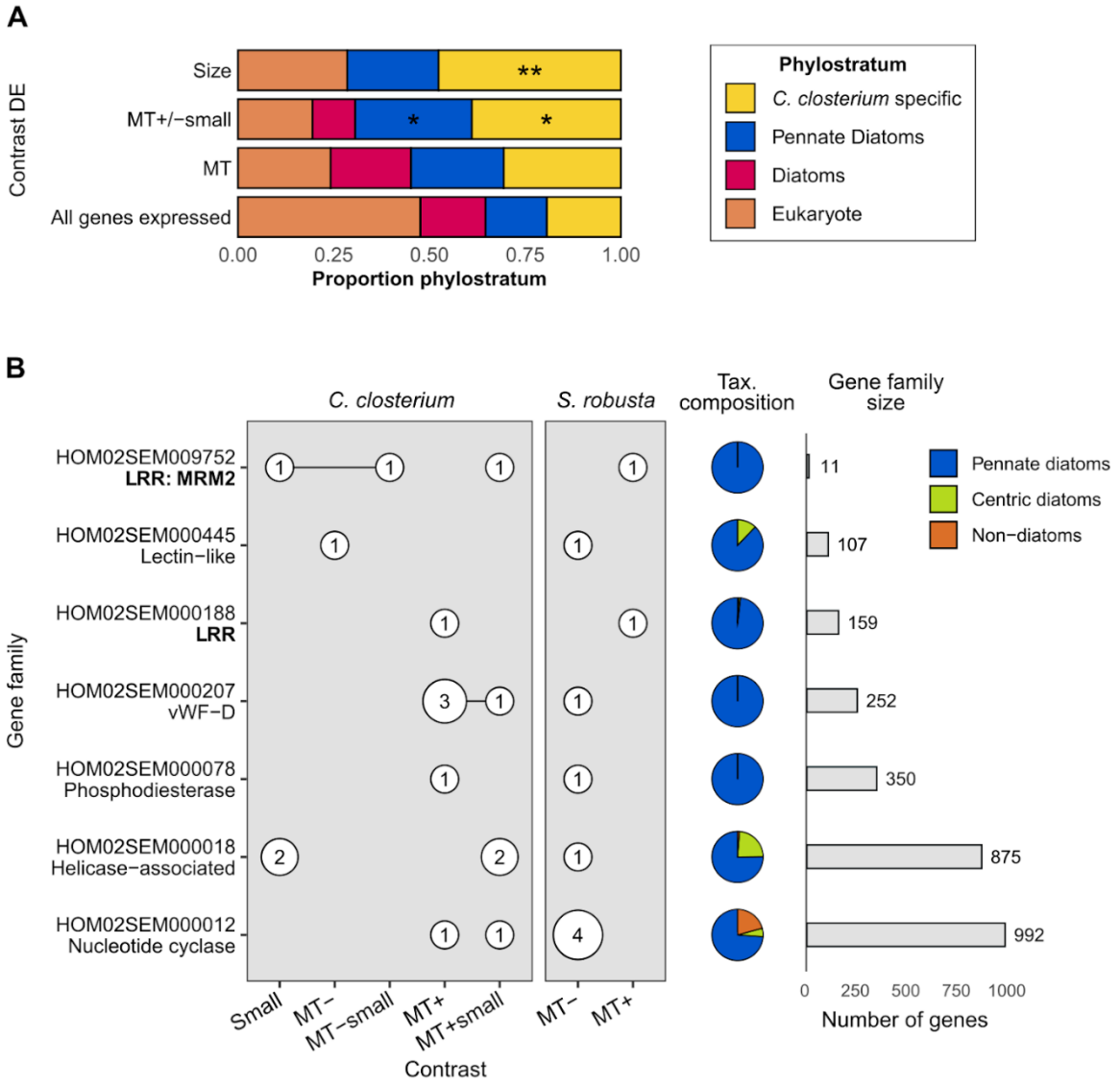

Figure 5: Phylostratigraphic analysis of size and mating type-responsive genes and comparative analysis of gene families containing genes with a mating type- or size-biased expression in two pennate diatoms: *Cylindrotheca closterium* and *Seminavis robusta*. (**A**) bar plots showing the proportion of different phylostrata (phylogenetic gene ages) among sets of differentially expressed (DE) genes as well as the background (all expressed genes). Stars indicate the enrichment significance level of a given phylostratum relative to the background (*: $p < 0.05$, **: $p < 0.01$, hypergeometric test). MT: mating type (**B**) For each gene family, the number of differentially expressed genes in each contrast is shown by circles, with genes that were differentially expressed in more than one contrast connected by a horizontal line. Pie

charts show the taxonomic composition of gene families, while bar charts show the size (number of genes) of each family. LRR: leucine-rich repeat. vWF-D: Von Willebrand factor type D domain.

Leucine-rich repeat proteins are involved in mating type differentiation in pennate diatoms

We next tried to pinpoint which common genes drive mating type differentiation in pennate diatoms by performing a comparative analysis using existing transcriptomic libraries from another heterothallic pennate diatom, *Seminavis robusta* [37]. We took advantage of the presence of 36 and 12 untreated, vegetative RNA-seq samples from small MT+ and MT- cells respectively, to perform a differential expression analysis. This resulted in a set of 183 genes that showed a mating type-specific expression response. We intersected the gene families of these *S. robusta* mating type-specific genes with families to which the 112 *C. closterium* size- and mating type-biased genes belong. Seven gene families contained DE genes in both species (**Figure 5B**).

Although the seven shared gene families range from small (11 genes) to large (992 genes), all of them are predominantly composed of genes belonging to pennate diatoms, further corroborating how mating type-dependent genes have evolved simultaneously with heterothally at the origin of pennate diatoms. Genes from two families containing Leucine-rich repeat (LRR) InterPro domain proteins showed a mating type biased response in both *C. closterium* and *S. robusta* (**Figure 5B**). Since the LRR domain is typically involved in protein-protein interaction or signalization, their conserved mating type dependent response suggests a role in communication between the opposite mating types in pennate diatoms. Although the first LRR family HOM02SEM000188 includes one *S. robusta* and one *C. closterium* gene with a MT+ specific expression, phylogenetic analysis showed that these two genes are not orthologous

(**Supplementary Figure 2**). Hence, the functional and evolutionary significance of their shared expression pattern is unclear.

On the other hand, HOM02SEM009752, a pennate-specific family with only 11 genes in 8 species, comprises two *C. closterium* genes, both of which are specific for small cells, but with a contrasting expression pattern between mating types: gene Ccl_4437 was upregulated in MT+ cells, whereas Ccl_5484 was upregulated in MT- cells (**Figure 5B**). The *S. robusta* homolog (Sro702_g189950) showed a strong MT+ specific expression. Importantly, this family encodes the Mating-type Related Minus 2 (*MRM2*) gene that was previously identified in the raphid pennate diatom *P. multistriata*. *MRM2* is a component of the mating type determination system of *P. multistriata*, where it is downstream of the sex-determining gene *MRP3* [17].

Mating-type Related Minus 2 (MRM2) is widely conserved in pennate diatoms

Multiple sequence alignment showed that proteins from the MRM2 family are conserved through their LRR domain. Therefore, we set out to identify additional homologs of MRM2 by constructing a profile hidden Markov model (HMM) of the LRR domain in the 11 proteins affiliated with HOM02SEM009752. Subsequently, the profile HMM was used to search the PLAZA Diatoms reference dataset for additional proteins that are homologous to MRM2 through their LRR domain [19]. Phylogenetic analysis based on the full-length protein sequence of the top-50 hits revealed a well-resolved clade containing the original MRM2 gene family as well as two genes belonging to other gene families (**Figure 6A, Supplementary Figure 3**). The proteins in this MRM2 clade share structural similarities, consisting of a predicted transmembrane domain on the C terminus and the LRR domain displayed extracellularly (**Figure 6A**). Motif discovery in the LRR domains of MRM2 homologs revealed the NxLt/sGxIPx motif typical to the plant specific (PS) family of diatom LRR proteins (**Figure 6A**) [38]. The short (~50AA) N-terminal intracellular domain does not contain a kinase domain

or other known signaling functionality, similar to the plant LRR-RLPs (leucine-rich repeat receptor-like proteins), which associate with co-receptors to transduce immune signals to the plant cell [39]. Together, MRM2 homologs were found in the genome of seven raphid pennate diatoms, as well as the araphid *Synedra acus*, suggesting a common origin at the root of the pennate diatoms.

Duplicated homologs of *MRM2* show an inverted mating type identity

The *MRM2* homologs of *C. closterium* and *S. robusta* have seemingly undergone gene duplication, with Ccl_4437 and Sro702_g189950 being closest related to the *P. multistriata MRM2* gene, whereas Ccl_5484 and Sro56_g032980 are more distant (**Figure 6A**). These distant out-paralogs of *MRM2* are lost or missing in most other pennate diatoms that encode only one copy of *MRM2*, including the original gene identified in *P. multistriata* [17].

Our data suggest that both Ccl_4437 and Ccl_5484 are conditionally expressed only in cells below the SST (**Figure 6B**). Ccl_4437 was specifically expressed in MT+ strains, while Ccl_5484 showed the opposite expression pattern, being biased towards MT-. We propose calling these homologs *MRM2h+* (*MRM2* homolog expressed in MT+) and *MRM2h-* respectively. Expression profiles of the small cells from the *S. robusta* expression atlas are congruent with what we observed in *C. closterium*: one homolog (Sro702_g189950, *MRM2h+*) is MT+ specific, while the other (Sro56_g032980, *MRM2h-*) is MT- specific (**Figure 6C**). To explore the variability of this response in other genotypes beyond the ones included in the *S. robusta* expression atlas (MT-: 84A, 85B, MT+: 85A), we performed RNA sequencing of six additional *S. robusta* genotypes below the SST, three for each mating type (MT+: 85A-9, P5, P7 and MT-: 85A-1, 85A-6, P4). While this experiment confirmed the MT- specific expression of *MRM2h-* in two out of three strains, *MRM2h+* was expressed at a very low level in all six strains (average expression of 0.76 counts per million). Hence, the mating type specific

expression pattern of *MRM2h+* could not be confirmed in these additional strains (**Supplementary Figure 4**).

Combined, the mating type specific expression of *MRM2* homologs in three species (*P. multistriata, C. closterium* and *S. robusta*) and its distribution throughout the pennates point to a role for *MRM2* as a key gene for mating type identity. As transcription of *MRM2* homologs was activated only when the cell size decreased below the SST, the function of MRM2 appears to be specific for sexually competent cells. Similar studies on the transcriptomic response to cells passing the SST should be undertaken in other life cycle model organisms such as *S. robusta* and *P. multistriata* to determine if MRM2's size-dependency is conserved throughout the pennate diatoms. The duplication and inversion of mating type preference in *S. robusta* and *C. closterium* suggests that MRM2 homologs have evolved to play similar roles, yet in opposite mating types. Finally, our data encourages discussion about how mating type identity is assigned. Mating types are currently distinguished based on cellular behavior during mating: MT- represents passive cells, while either the MT+ gametangia or gametes are motile [40]. Our data suggests that distinguishing mating types by their behavior may not reflect molecular evolution. To illustrate, *MRM2h+* genes Ccl_4437 and Sro702_g189950 exhibited MT+ expression but are phylogenetically closest to the original *P. multistriata MRM2*, which was expressed in MT- (**Figure 6**).

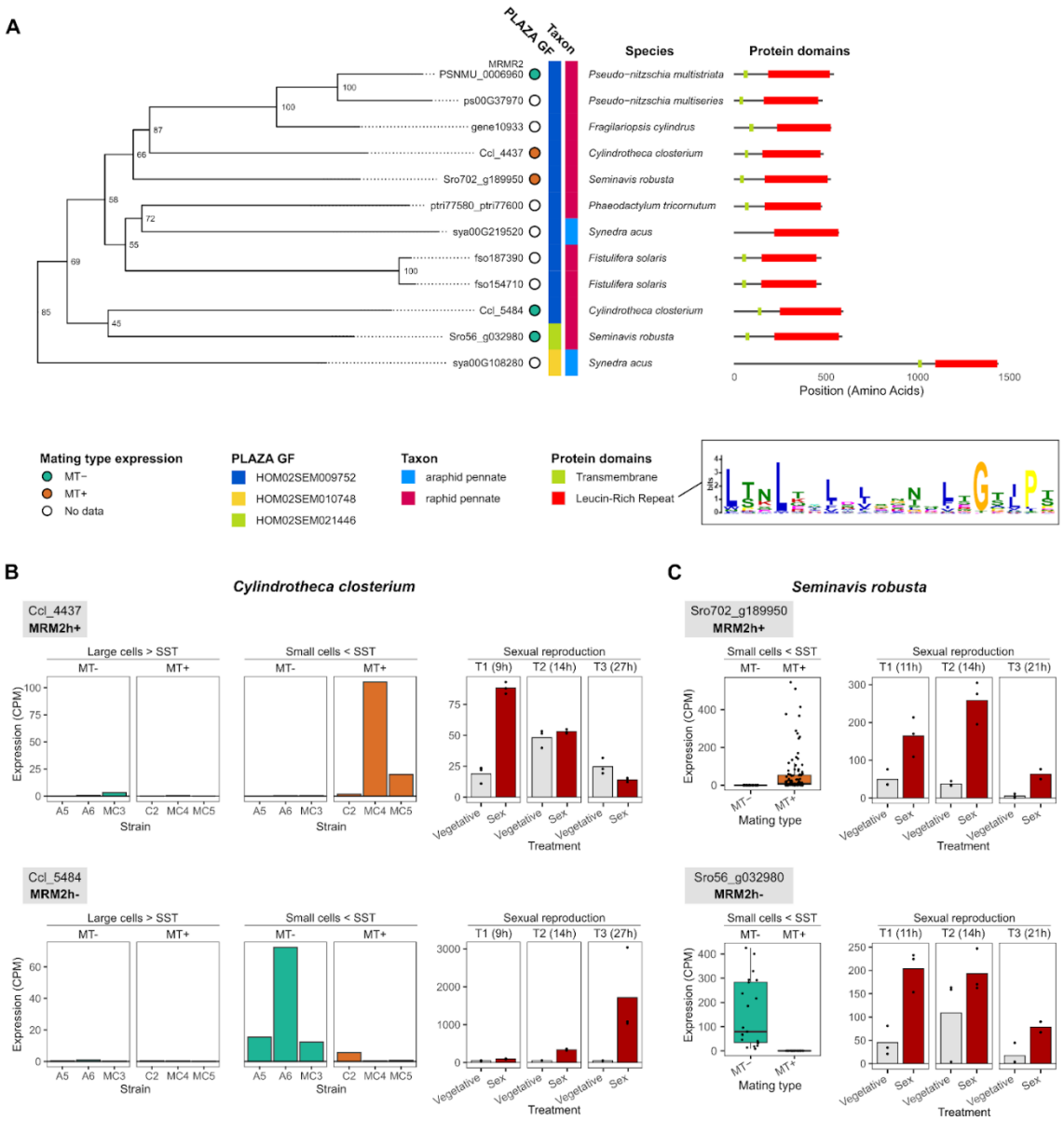

Figure 6: Phylogeny and expression of MRM2 homologs in pennate diatoms. **(A)** Bootstrap-consensus phylogenetic tree of MRM2 proteins in diatoms. Bootstrap support is shown as node labels. The full phylogenetic tree of the top-50 best MRM2 hits is available as Supplementary Figure 3. Next to the gene labels, their mating type-dependent expression pattern, gene family (GF), taxonomy and protein structures (InterPro domains and transmembrane domains predicted by Phobius) are shown. The inset shows the MEME sequence logo of the leucine-rich repeat domains of these proteins **(B)** bar plots showing expression in counts per million (CPM) of MRM2 homologs in large and small cells of different *Cylindrotheca closterium* strains (left) and during three stages of sexual reproduction (right).

For the sexual reproduction, bar plots show the average expression while dots show the individual RNA-seq samples. SST: sexual size threshold, MRM2h+: MT+ biased homolog of MRM2, MRM2h-: MT- biased homolog of MRM2 **(C)** expression of MRM2 homologs in the *S. robusta* expression atlas [19]. Left: boxplots of expression in MT+ (n = 108) and MT- (n = 21) samples in a range of non-sexual conditions, spanning biotic interactions, environmental conditions, sexual reproduction, and diel cycles. Right: expression during sexual reproduction covering three time points after crossing. Bar plots show average expression while dots show individual RNA-seq samples.

Upregulation of *MRM2* homologs during sex suggests a function in recognition between compatible mating types

The differentiation between mating types in raphid pennate diatoms is most apparent during sexual reproduction, where attraction and cell-cell recognition ensure the correct pairing of compatible gametangia and fusion of a MT+ with a MT- gamete (reviewed in Bilcke et al., 2022). Therefore, we inspected the expression of *MRM2* homologs during different stages of sexual reproduction [18, 19] (**Figure 6B,C**). Both *C. closterium MRM2* homologs as well as the *S. robusta MRM2h+* were significantly transcriptionally upregulated in at least one stage of sexual reproduction. Strikingly, the *C. closterium* gene *MRM2h+* is only upregulated during the early stage whereas *MRM2h-* was most strongly upregulated in the last stage. Hence, not only are *MRM2* homologs in *C. closterium* specific for the opposing mating types but they also appear to play a role in different stages of sexual reproduction: gametangia pairing (*MRM2h+*) versus the differentiation of gametes or auxospore development (*MRM2h-*).

Conclusions and Outlook

In this study, we leveraged experimental cell size manipulation to decipher the transcriptional reprogramming associated with size-based sexual maturation in diatoms. By simultaneously harvesting the same six genotypes in large and small cells, we avoided potential confounding due to the use of different genotypes in the two size classes. Furthermore, the use of full biological replication (i.e. independent genotypes) further ensures that the small set of DE genes we identified are robust and valid for a range of *C. closterium* strains.

A sizable set of 29 mating type-specific genes were transcriptionally activated in cells that have passed the SST. We argue that such genes must underlie the chemical and behavioral differentiation between mating types, which only becomes apparent in small cells. Three such co-expressed genes were located nearby on the same contig. Their confined expression in small MT+ cells and coordinated induction in response to sex pheromones suggest an important role during the pairing process, in which MT+ cells are actively gliding towards an attracting MT- cell.

Furthermore, we demonstrated that the *P. multistriata* gene *MRM2* not only has a wider distribution in pennate diatoms than was previously reported [17], but also displays mating type-specificity in at least three raphid pennate diatoms from diverse taxonomic clades (*C. closterium, P. multistriata, S. robusta*). Judging from the fact that these plant-like LRR receptor proteins are most strongly expressed during cell pairing and gamete fusion, we argue that MRM2 is part of an ancient partner recognition system that dates back to the origin of the pennate diatoms. Consequently, the conservation of MRM2 is the first molecular evidence for heterothally being an ancestral feature in pennate diatoms. While most pennate diatoms indeed display a heterothallic mating system, some pennate species display homothallic reproduction or facultative heterothally [15]. Heterothally would consequently have been secondarily lost in these species. Additionally, in accordance with Russo and colleagues (2018) we did not find

any homologs of the *P. multistriata* mating locus *MRP3* outside the *Pseudo-nitzschia/Fragilariopsis* clade [17], suggesting that different clades use different mating type determination systems. In future studies, assessing the conservation of mating type dependent genes such as *MRM2* in a diverse set of heterothallic and homothallic pennate diatoms will be instrumental to understand their evolutionary history. In that respect, the 100 Diatom Genomes project (jgi.doe.gov/csp-2021-100-diatom-genomes) will soon make it possible to investigate life cycle genes at an unprecedented scale.

While the earliest reports of the unique size reduction-restitution life cycle of diatoms date as far back as 1869 [2, 3], the mechanism that cells use to accurately measure their own size is still completely unknown. The fact that the size is constantly monitored by the cell can be deduced from the almost instantaneous sexual maturation after the experimental cell size manipulation we performed using a glass Pasteur pipette. Such microsurgery-based size manipulation, which was previously executed using a razor blade by van Stosch in 1965 [41] and later repeated by Roshchin [42] and Chepurnov [2], represents a time and cost-efficient way to significantly reduce the generation time of experimental models such as *C. closterium*. Using this approach, we identified a set of 21 genes that are upregulated in small cells (<SST) relative to large cells (>SST) in both mating types. These size markers include several nucleus-localized and DNA-interacting genes, suggesting a function as size-dependent switches, activating downstream sexual maturation genes only below the SST, but need to be further validated by functional experiments. However, none of these genes could be clearly linked to a physical function in cell size sensing, and neither did any of them occur specifically in the genome of both pennate and centric diatoms. Hence, it remains unexplained how cells measure their own size and whether a diatom-wide transcriptional master regulator exists. The identification of the first molecular markers for cell size dependent sexual maturation in diatoms - including conserved key genes

such as *MRM2* - paves the road for future research into the physicochemical and molecular changes taking place when cells pass the SST.

Acknowledgements

G.B. is a postdoctoral fellow supported by Fonds Wetenschappelijk Onderzoek (FWO, 1228423N). W.V. acknowledges partial financial support from FWO project G001521N, BOF/GOA projects No. 01G01715 and 01G01323, as well as infrastructure funded by EMBRC Belgium-FWO project GOH3817N. The BCCM/DCG culture collection is supported by the Belgian Science Policy (Belspo).

We sincerely thank Dr. Koen Van den Berge for helpful discussion about experimental design. We are grateful towards Olga Chepurnova for help with maintaining and manipulation of diatom cultures in cryopreservation tanks. We would further like to thank Dr. Katerina Pargana for her help with bioinformatic analysis and Dr. Petra Bulánková for assisting with molecular biology protocols.

Competing Interests

The authors declare that they have no conflict of interest.

References

1. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science (80-)* 1998; **281**: 237–240.
2. Macdonald JD. I.— On the structure of the Diatomaceous frustule, and its genetic cycle . *Ann Mag Nat Hist* 1869; **3**: 1–8.

3. Pfitzer E. Über den Bau und die Zellteilung der Diatomeen. Botanische Zeitung 1869; **1**:1–189.

4. Nagai S, Hori Y, Manabe T, Imai I. Restoration of cell size by vegetative cell enlargement in Coscinodiscus wailesii (Bacillariophyceae). *Phycologia* 1995; **34**: 533–535.

5. Chepurnov VA, Mann DG, Sabbe K, Vyverman W. Experimental studies on sexual reproduction in diatoms. [Review] [248 refs]. *Int Rev Cytol* 2004; **237**: 91–154.

6. Kaczmarska I, Ehrman JM, Mills KE, Sutcliffe SG, Samanta B. Vegetative cell enlargement in selected centric diatom species–an alternative way to propagate an individual genotype. *Eur J Phycol* 2022.

7. Moeys S, Frenkel J, Lembke C, Gillard JTF, Devos V, Van Den Berge K, et al. A sex-inducing pheromone triggers cell cycle arrest and mate attraction in the diatom Seminavis robusta. *Sci Rep* 2016; **6**: 1–13.

8. Basu S, Patil S, Mapleson D, Russo MT, Vitale L, Fevola C, et al. Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. *New Phytol* 2017; **215**: 140–156.

9. Bilcke G, Van Den Berge K, De Decker S, Bonneure E, Poulsen N, Bulankova P, et al. Mating type specific transcriptomic response to sex inducing pheromone in the pennate diatom Seminavis robusta. *ISME J* 2021; **15**: 562–576.

10. Klapper F, Audoor S, Vyverman W, Pohnert G. Pheromone Mediated Sexual Reproduction of Pennate Diatom Cylindrotheca closterium. *J Chem Ecol* 2021; **47**: 504–512.

11. Gillard J, Frenkel J, Devos V, Sabbe K, Paul C, Rempt M, et al. Metabolomics Enables the Structure Elucidation of a Diatom Sex Pheromone. *Angew Chemie Int Ed* 2013; **52**: 854–857.

12. Klapper FA, Kiel C, Bellstedt P, Vyverman W, Pohnert G. Structure Elucidation of the First Sex-Inducing Pheromone of a Diatom. *Angew Chemie - Int Ed* 2023.

13. Davidovich NA, Davidovich OI, Podunai YA, Shorenko KI, Kulikovskii MS. Reproductive properties of diatoms significant for their cultivation and biotechnology. *Russ J Plant Physiol* . 2015. Maik Nauka Publishing / Springer SBM. , **62**: 153–160

14. Bilcke G, Ferrante MI, Montresor M, De Decker SD, Veylder L De, Vyverman W. Life Cycle Regulation. *Mol Life Diatoms* 2022; 205–228.

15. Poulíčková A, Mann DG, Mann DG. Diatom Sexual Reproduction and Life Cycles. *Diatoms: Fundamentals and Applications*. 2019. Wiley, pp 245–272.

16. Vanstechelman I, Sabbe K, Vyverman W, Vanormelingen P, Vuylsteke M. Linkage Mapping Identifies the Sex Determining Region as a Single Locus in the Pennate Diatom Seminavis robusta. *PLoS One* 2013; **8**: e60132.

17. Russo MT, Vitale L, Entrambasaguas L, Anestis K, Fattorini N, Romano F, et al. MRP3 is a sex determining gene in the diatom Pseudo-nitzschia multistriata. *Nat Commun* 2018; **9**.

18. Audoor S, Bilcke G, Pargana A, Belišová D, Thierens S, Van Bel M, et al. Transcriptional chronology reveals conserved genes involved in pennate diatom sexual reproduction. *Press* .

19. Osuna-Cruz CM, Bilcke G, Vancaester E, De Decker S, Bones AM, Winge P, et al. The Seminavis robusta genome provides insights into the evolutionary adaptations of benthic diatoms. *Nat Commun* 2020; **11**.

20. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017; **14**: 417–419.

21. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research* 2016; **4**.

22. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2009; **26**: 139–140.

23. Bucchini F, Del Cortona A, Kreft Ł, Botzki A, Van Bel M, Vandepoele K. TRAPID 2.0: A web application for taxonomic and functional analysis of de novo transcriptomes. *Nucleic Acids Res* 2021; **49**: e101–e101.

24. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 2013; **30**: 772–780.

25. Pinseel E, Nakov T, Van den Berge K, Downey KM, Judy KJ, Kourtchenko O, et al. Strain-specific transcriptional responses overshadow salinity effects in a marine diatom sampled along the Baltic Sea salinity cline. *ISME J* 2022; **16**: 1776–1787.

26. Ruggiero A, Mager S, Manfellotto F, Tuccio V Di, Nishimura T, Pan L, et al. DNA methylation is linked to the monoallelic expression of MRP3, a diatom mating type determining gene.

27. Bachtrog D. A dynamic view of sex chromosome evolution Maintenance and gain of genes on old Y-chromosomes. 2006.

28. Kent T V., Uzunović J, Wright SI. Coevolution between transposable elements and recombination. *Philos Trans R Soc B Biol Sci* . 2017. Royal Society Publishing. , **372**

29. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* . 2007. Nature Publishing Group. , **8**: 272–285

30. Hobza R, Cegan R, Jesionek W, Kejnovsky E, Vyskot B, Kubat Z. Impact of repetitive elements on the Y chromosome formation in plants. *Genes (Basel)* . 2017. MDPI AG. , **8**: 302

31. McQuaid JB, Kustka AB, Oborník M, Horák A, McCrow JP, Karas BJ, et al. Carbonate-sensitive phytotransferrin controls high-affinity iron uptake in diatoms. *Nature* . 2018. ,

555: 534–537

32. Turnšek J, Brunson JK, Martinez Viedma MDP, Deerinck TJ, Horák A, Oborník M, et al. Proximity proteomics in a marine diatom reveals a putative cell surface-to-chloroplast iron trafficking pathway. *Elife* 2021; **10**: 1–113.

33. Nützmann HW, Scazzocchio C, Osbourn A. Metabolic gene clusters in eukaryotes. *Annu Rev Genet* . 2018. Annual Reviews Inc. , **52**: 159–183

34. Reber S, Goehring NW. Intracellular scaling mechanisms. *Cold Spring Harb Perspect Biol* 2015; **7**.

35. Marshall WF. How Cells Measure Length on Subcellular Scales. *Trends Cell Biol* 2015; **25**: 760–768.

36. Nakov T, Beaulieu JM, Alverson AJ. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). *New Phytol* 2018; **219**: 462–473.

37. Osuna-Cruz CM, Bilcke G, Vancaester E, De Decker S, Bones AM, Winge P, et al. The Seminavis robusta genome provides insights into the evolutionary adaptations of benthic diatoms. *Nat Commun* 2020; **11**.

38. Schulze B, Buhmann MT, Río Bártulos C, Kroth PG. Comprehensive computational analysis of leucine-rich repeat (LRR) proteins encoded in the genome of the diatom Phaeodactylum tricornutum. *Mar Genomics* 2015; **21**: 43–51.

39. Snoeck S, Garcia AG, Steinbrenner AD. Plant Receptor-like proteins (RLPs): Structural features enabling versatile immune recognition. *Physiol Mol Plant Pathol* . 2023. Academic Press. , **125**: 102004

40. Kaczmarska I, Poulícková A, Sato S, Edlund MB, Idei M, Watanabe T, et al. Proposals for a terminology for diatom sexual reproduction, auxospores and resting stages. *Diatom Res* 2013; **28**: 263–294.

41. von Stosch HA. Manipulierung der Zellgrosse von Diatomeen im Experiment. *Phycologia* 1965; **5**: 21–44.

42. Roshchin, AM. 1994. Zhiznennye tsiklydiatomovykh vodoroslej. Naukova Dumka, Kiev. pp. 170