Department of Biology
Department of Plant Biotechnology and Bioinformatics
VIB Department of Plant Systems Biology

Sex-signalling and mating type determination in the benthic pennate diatom *Seminavis robusta*

Sara Moeys

Thesis submitted in fulfillment of the requirements for the degree of
Doctor (PhD) in Sciences (Biology)

29th of April 2015

Promoter: Prof. dr. Wim Vyverman
Co-promoter: Prof. dr. Lieven De Veylder

Academic year
2014-2015

Members of the Examination Committee

Prof. Dr. Koen Sabbe (Chairman)

Department of Biology, Ghent University

Prof. Dr. Wim Vyverman (promoter)

Department of Biology, Ghent University

Prof. Dr. Lieven De Veylder (co-promoter)

Department of Plant Biotechnology and Bioinformatics, Ghent University

VIB Department of Plant Systems Biology

Prof. Dr. Atle M. Bones *

Department of Biology, Norwegian University of Science and Technology (NTNU)

Prof. Dr. Tom Beeckman *

Department of Plant Biotechnology and Bioinformatics, Ghent University

VIB Department of Plant Systems Biology

Prof. Dr. Olivier De Clerck *

Department of Biology, Ghent University

Prof. Dr. ir. Sven Mangelinckx

Department of Sustainable Organic Chemistry and Technology, Ghent University

Dr. ir. Marnik Vuylsteke

Department of Animal production, Ghent University

* Members of the Reading committee

Acknowledgments

The work presented in this thesis is not the achievement of one person. Many people helped and supported me during the past four years. Without them, this thesis would never have been written as it is today. Therefore, I would like to thank some people and to apologize beforehand for all people I might have forgotten.

First of all, I would like to thank my promotor Wim Vyverman. When I applied for a job at the lab of Protistology & Aquatic Ecology (PAE), I barely knew what diatoms where and why they could be interesting to study. It was Wim who introduced these beautiful little organisms to me and convinced me that they were for sure worth working on. I am thus grateful that he gave me the chance to pursue a PhD in his lab and for all his support and guidance.

The first months of my PhD, I spent in the PAE lab at the Sterre, learning how to work with diatoms. Later on, I moved to the Plant System Biology department to focus more on the molecular biology of diatoms. I started there in the lab of Marnik Vuysteke. I would like to thank Marnik for all his support and good advice, during the two years I spent in his lab as well as afterwards. After two years, I had to move to another group and luckily Lieven De Veylder accepted me in his Cell Cycle group. Therefore, I would like to thank him, as well as for all fruitful discussions we had during the last two years of my PhD and for being my co-promotor in the project.

I also want to especially thank Marie Huysman, who was involved in designing my project and in numerous discussions about results and future plans. I could always go to her for help with experimental (or other) issues. Furthermore, I also want to thank Koen Sabbe for his input, especially on the mating type locus.

During the last months of my PhD, Barbara Bouillon was of a great help to me, taking over some of my experiments to give me time to write my thesis. I am thankful to her and wish her all the best with her own PhD.

Four years ago, my knowledge of bioinformatics was non-existing. Luckily, I had my colleagues Valerie, Ives and Stephanie to learn me some tips-and-tricks. When doing my RNA-seq analysis, I could always ask Frederik Coppens of the ABB-group for advice. I want to thank him and all others that helped me with my bioinformatics problems, including the people from IT. I bothered them frequently with malfunctioning soft- and hardware and they always succeeding in helping me promptly. Furthermore, I also want to thank Bram Verhelst and Klaas Vandepoele that are now helping us to finally finalize the *Seminavis* genome.

Together with bioinformatics comes biostatistics. I want to thank Lieven Clement and Koen Van den Berghe for their help with this side of my research. They really did a great job analyzing the RNA-seq data.

I would like to thank all my colleagues of the PAE and PSB labs. I really enjoyed working and having fun with them. I especially want to thank the support staff of both labs (Olga, Ilse, Sofie, Tine, Hilde, Ilse, Wilson, Jackie, Kristof and all others) for making life easier in the lab. Although I wasn't there that often, I always felt welcome in the PAE group. I enjoyed working in the small Quantitative Genomics group at PSB and also want to mention the Seed Development group with whom we shared a lab and who make life there even more fun. I also want to thank the Cell Cycle group for making me immediately feel welcome there and for the great atmosphere in the lab.

Apart from my great colleagues in Ghent, I also want to express my gratitude to some collaborators abroad. First, I want to thank Georg Pohnert, Johannes Frenkel and Christine Lembke from the university of Jena for the nice collaboration we had concerning the diatom pheromones. Next, I want to thank Marina Montresor, Mariella Ferrante, Remo Sanges and all others from the Stazione Zoologica Anton Dohrn in Naples. It was nice working with them on the JGI project and I enjoyed my visits to Naples and felt really welcome in there group.

Tenslotte wil ik nog mijn familie, schoonfamilie en vrienden bedanken. Zonder de onvoorwaardelijke steun van mijn ouders had ik hier nooit gestaan. Zij hebben mij altijd alle kansen gegeven en gesteund in al mijn beslissingen. Ik wil ook mijn

broer Jasper en de rest van de familie bedanken voor hun steun en hun interesse in mijn werk. Uiteraard wil ik ook mijn vrienden bedanken, bij wie ik altijd terecht kon voor wat afleiding en plezier, in het bijzonder Radost, Lies, Lot, Lise, Vig en Céel voor de vele leuke momenten. Ik wil ook Lise nog even speciaal bedanken voor de prachtige tekening die de cover van deze thesis siert.

En als laatste, en misschien wel belangrijkste, wil ik Maarten bedanken. Hij heeft mij altijd gesteund, ook toen ik besliste om in het verre Gent te gaan werken. Hij zorgde er zelfs voor die verre verplaatsing draaglijker te maken door lekker eten op tafel te toveren als ik 's avonds thuiskwam. En hij luisterde altijd geduldig naar al mijn verhalen en verzuchtingen. Zonder hem had ik dit nooit gekund.

Sara

Abbreviations

AFLP	amplified fragment length polymorphism
BSA	bulked segregant analysis
BWA	Burrows Wheeler Aligner
cAMP	cyclic adenosine monophosphate
CDD	conserved domains database
c-di-GMP	cyclic diguanylate
CDK	cyclin-dependent kinase
CDP	cyclodipeptide
CDPS	cyclodipeptide synthases
CF-M	conditioning factor produced by MT^-
CF-P	conditioning factor produced by MT^+
cGMP	cyclic guanosine monophosphate
cpm	counts per million
DNA	deoxyribonucleic acid
DE	differential expression
DNMT	DNA methyltransferase
dsCYC	diatom-specific cyclin
dc-SAM	Decarboxylated S-adenosylmethionine
dsx	doublesex
ECFP	enhanced cyan fluorescent protein
EPA	eicosapentaenoic acid
EYFP	enhanced yellow fluorescent protein
FDR	false discovery rate
G_1 phase	gap 1 phase
G_2 phase	gap 2 phase
GC	guanylyl cyclase
gDNA	genomic DNA
GMP	guanosine monophosphate
GTP	guanosine triphosphate
HA	haemagglutinin
HMG	high mobility group
HMM	Hidden Markov Model
LCPA	long-chain polyamines
LOH	loss of heterozygosity
M phase	mitotic phase
MP	mapping population
MRN complex	Mre11, Rad50 and Nbs1 complex
MSA	multiple sequence alignment
MTA	methylthioadenosine
MT	mating type
MTase domain	methyltransferase domain
NJ	Neighbor-Joining
NRPS	non-ribosomal peptide synthases
NSW	natural sea water
P5CS	Δ1-pyrroline-5-carboxylate synthetase

P5CR	Δ1-pyrroline-5-carboxylate reductase
PC	primer combination
PDE	phosphodiesterase
ph-1	*Pseudostaurosira trainorii* pheromone 1
ph-2	*Pseudostaurosira trainorii* pheromone 2
ph-3	*Pseudostaurosira trainorii* pheromone 3
PI	phosphatidylinositol
PIP	phosphatidylinositol phosphate
PKG	cGMP-dependent protein kinase
QTL	quantitative trait locus
RNA	ribonucleic acid
RNA-seq	RNA sequencing
RT-PCR	reverse transcription polymerase chain reaction
RT-qPCR	reverse transcription quantitative polymerase chain reaction
S phase	DNA synthesis phase
SAM	S-adenosylmethionine
SAR	Stramenopiles, Alveolates and Rhizaria
SDV	silica deposition vesicle
sdY	sexually dimorphic on the Y chromosome
SE	standard error
SIT	silicic acid transporter
SNP	single nucleotide polymorphism
SPD	spermidine
SPM	spermine
SRY	Sex Determining Region of Y
SST	sexual size threshold
sxl	sex-lethal
tra	transformer
vcf	variant call format
WGS	whole genome sequencing
xol1	XO lethal 1
ω-3 LC PUFA	omega-3-long-chain polyunsaturated fatty acid

Table of Contents

1
Introduction

Diatoms – of great ecological importance, with great biotechnological potential

All life on earth relies on primary production, which is the fixation of carbon dioxide in organic matter by autotrophic organisms. In this process also oxygen is produced. About half of the global net primary production comes from the oceans (Falkowski and Raven, 1997). Here, diatoms are responsible for at least a quarter of carbon fixation, which is comparable to the primary production of the terrestrial rainforests (Granum et al., 2005, Armbrust, 2009). This places diatoms at the base of marine food webs. Diatoms are, with an estimation of about 200,000 species, the most diverse group of microalgae (Mann and Droop, 1996). They are found in a wide variety of environments where sufficient light and nutrients (nitrogen, phosphorus, silicon and trace elements) are available (Armbrust, 2009, Falkowski and Knoll, 2011). The diatom organic matter serves as food for aquatic organisms in all layers of the ocean as it sinks from the surface. A small fraction of this sinking organic matter settles on the sea floor, where it is sequestered in sediments and contributes to petroleum reserves (Armbrust, 2009). Apart from their key role in the carbon cycle, they also dominate the biogeochemical cycling of silica as the main producers of biogenic silica. Silicon is incorporated in the diatom frustule on average 25 times before reaching the sea floor (Treguer and De La Rocha, 2013).

Besides their great ecological importance, diatoms can also be used in numerous industrial applications. Because of their high lipid content and more specifically their high levels of the omega-3-long-chain polyunsaturated fatty acid (ω-3 LC PUFA) eicosapentaenoic acid (EPA), diatoms are used as feed for filter feeders in aquaculture (Miller et al., 2014). Since vertebrates produce only a limited

amount of ω-3 LC PUFAs, they need a dietary source for these essential compounds. Furthermore, diatoms are also rich in protein and possess all essential amino acids (Guil-Guerrero et al., 2004). Together with the presence of nearly all essential vitamins, this makes them very interesting to be used in human and animal nutrition (Spolaore et al., 2006). Although the algae themselves are generally sold as food additives or added to dietary products, it is also possible to extract high-value molecules, like fatty acids, pigments or stable isotope biochemicals (Medina et al., 1998, Spolaore et al., 2006, Gastineau et al., 2012).

Next to the nutrition industry, also the nanotechnological sector shows interest in diatoms, in particular in their silicified cell wall. Frustules are already used as diatomaceous filters for heavy metals absorption, DNA purification or wastewater treatment (Gilmore et al., 1993, Al-Degs et al., 2001, Panacek et al., 2013). The process of biomineralization is a source of inspiration in material sciences, where knowledge about the biosilicification process is used to develop biomimetic systems for silica nanostructure formation (Lopez et al., 2005). The frustules itself could also be used in nanotechnology, for example as photonic crystals or to be functionalized and used as biosensor (Fuhrmann et al., 2004, Viji et al., 2014). It is also shown that diatom frustules could be used as a natural drug delivery system (Aw et al., 2012, Bariana et al., 2013). Furthermore, silicate can be replaced by magnesium oxide, leading to new possibilities in nanometallurgy (Sandhage et al., 2002, Drum and Gordon, 2003).

Despite the great potential of diatoms in industry, only a few species are being cultivated at large scale. In aquaculture, for example, the diatom *Chaetoceros muelleri* is cultivated to feed larvae and *Haslea ostrearia* is used for the greening of oysters (Turpin et al., 2001, Lopez-Elias et al., 2005). One of the major drawbacks in large-scale diatom cultivation is the high production cost (Lebeau and Robert, 2003). To lower these costs, genetic engineering could be applied to increase the growth rate or enhance the production of high-value molecules (Kroth, 2007). Furthermore, a better knowledge of diatom life cycle regulation is needed to

optimize cultivation techniques and is thus essential to exploit the diatoms potential at its fullest.

Diatoms in the eukaryotic tree of life

Over the past 20 years, phylogenetic methods changed from using mainly morphological data to the use of molecular markers. As a consequence, the image of eukaryotic tree of life evolved from an assembly of four large kingdoms (animals, plants, fungi and protists) to a new eukaryotic tree assembly where most lineages can be assigned to six supergroups, namely Amoebozoa, Opisthokonta, Excavata, Archaeplastida, Rhizaria and Chromalveolata (Parfrey et al., 2006). Since 2012, the full genome is sequenced for at least one species of every major lineage, leading to the use of phylogenomics to build evolutionary trees. This approach confirmed the existence of most supergroups, although some important changes were established (Figure 1) (Burki, 2014). One major change was the disappearance of the Chromalveolates. The stramenopiles, to which diatoms belong, and the alveolates, that previously were part of this supergroup, now fall within the new supergroup called SAR (Stramenopiles, Alveolates, Rhizaria) (Burki et al., 2007). The position of haptophytes and cryptomonads, previously also belonging to the Chromalveolates, remeans unresolved in this analysis. It was assumed that these two protist groups share a common ancestor, but a recent study showed haptophytes branching closer to SAR and cryptomonads to archaeplastida (Burki et al., 2012). Besides stramenopiles and alveolates, also rhizaria are now assigned to the SAR supergroup. This was surprising since none of the groups belonging to rhizaria have plastids of red algal origin, instead most groups are heterotrophic or contains plastids of green algal origin. Altogether, these observations contradict the “chromalveolate hypothesis”. According to this hypothesis, all red algal-derived plastids evolved from one secondary endosymbiosis in a chromalveolate ancestor. The numerous nonphotosynthetic lineages within the chromalveolate group were presumed to have lost their plastid (Cavalier-Smith, 1999).

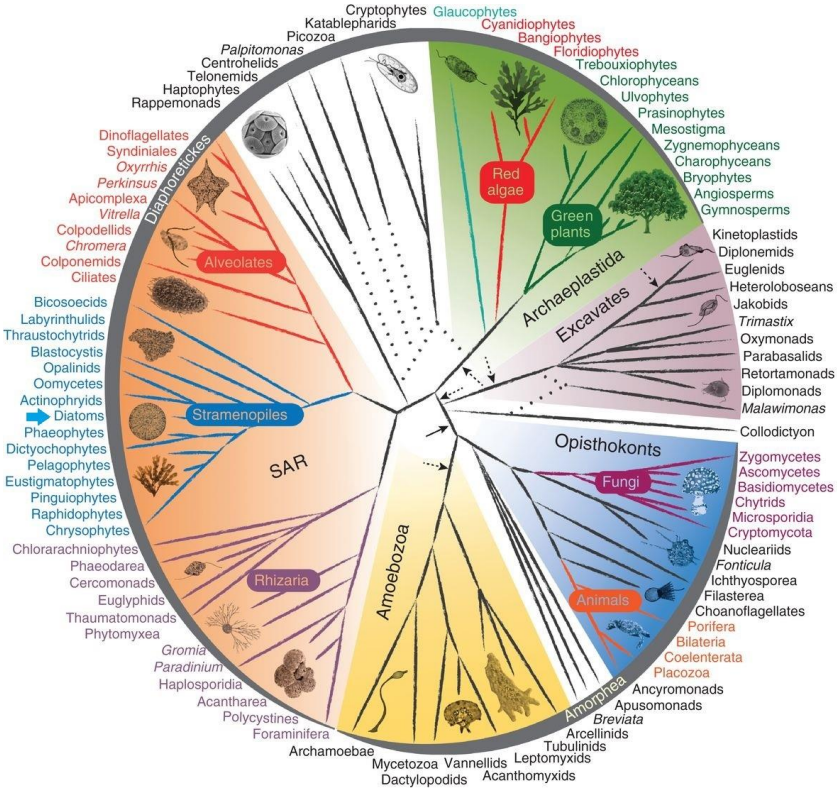


Figure 1: Eukaryotic tree of life based on a consensus of phylogenetic evidence (in particular, phylogenomics), rare genomic signatures, and morphological characteristics (Burki, 2014). Cartoons illustrate the diversity constituting the largest assemblages (colored boxes). The branching pattern does not necessarily represent the inferred relationships between the lineages. Dotted lines denote uncertain relationships, including conflicting positions. The arrows point to possible positions for the eukaryotic root; the solid arrow corresponds to the most popular hypothesis (Amorphea-bikont rooting), the broken arrows represent alternative hypotheses. The blue arrow indicates the position of the diatoms in the tree.

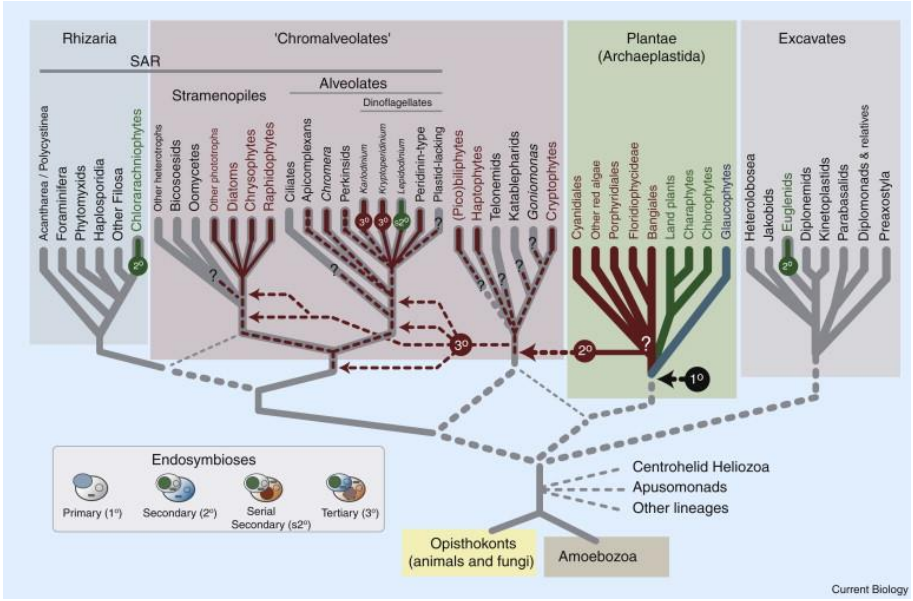


Figure 2: Hypothesis for plastid evolution in eukaryotes (Archibald, 2009). The six supergroups of eukaryotes are shown, with emphasis on lineages containing plastid-bearing groups. Possible secondary, tertiary and serial secondary endosymbioses involving red and green algal endosymbionts are indicated. Primary endosymbiosis occurred in an ancestor of the Archaeplastida. Euglenids and chorarachniophytes obtained green algal-derived plastids in two independent secondary endosymbioses. An ancestor of cryptophytes and haptophytes engulfed a species of an unidentified lineage of red algae in a secondary endosymbiosis event. The stramenopile-alveolate lineage obtained this plastid by tertiary endosymbiosis. This could have been a single event or two separate events.

The above described changes in the assembly of the eukaryotic tree led to a new view on plastid evolution. Land plants, green, red and glaucophyte algae all contain a primary plastid. These primary plastids seem to be derived from the same endosymbiotic event, where a photosynthetic cyanobacterium was engulfed by a eukaryotic host cell (Reyes-Prieto et al., 2007). By intracellular gene transfer, hunderds of endosymbiont-derived genes were transferred to the hosts nucleus. Some lineages, however, obtained their plastid through a secondary endosymbiosis. Both euglenids and chorarachniophytes possess plastids derived from a green alga that was engulfed by a non-photosynthetic ancestor. As these two lineages belong to different supergroups that are dominated by non-

photosynthetic organisms, they probably obtained their plastid via independent events (Rogers et al., 2007). Additionally, six algal lineages have plastids from red algal origin. According to the chromalveolate hypothesis, all red plastids are the results of one secondary endosymbiosis (Cavalier-Smith, 1999). Now, phylogenomic data point towards an alternative theory about plastid evolution (Burki, 2014). Figure 2 depicts the hypothesis of Sanchez-Puerta and Delwiche (2008). According to this theory, a species of an unidentified lineage of red algae was taken up by an ancestor of cryptophytes and haptophytes in a secondary endosymbiosis event. This plastid was then transferred to the stramenopile-alveolate lineage by tertiary endosymbiosis.

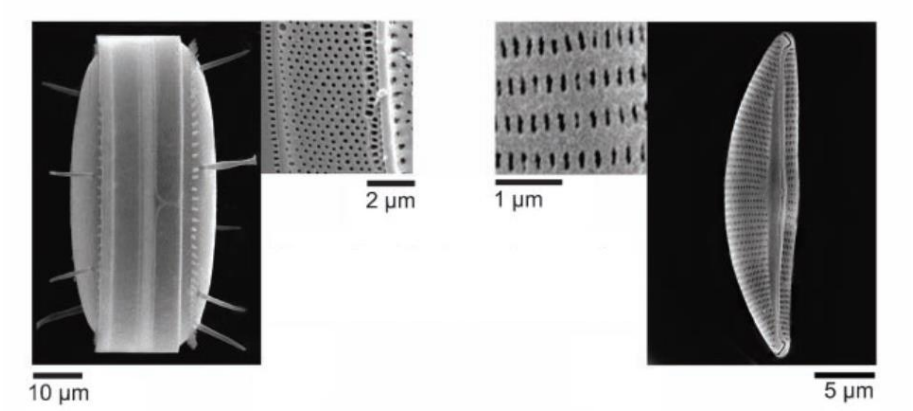


Figure 3: Scanning electron micrographs of the centric diatom *Thalassiosira punctigera* with radially symmetric valves (left) and the pennate diatom *Seminavis robusta* with bilateral symmetry (right) (Chepurnov et al., 2008).

The diatoms, belonging to the straminopiles, evolved somewhere between 250 and 183 Ma ago (Sorhannus, 2007). Based on the shape of their cell wall, three types of diatoms can be distinguished. Diatoms with radially symmetric valves are called centrics, while elongated diatoms with bilateral symmetry are called pennates (Figure 3). In the pennate diatoms, a division is made between diatoms having a fissure through the valve face, called the raphe (raphid pennates) or without this fissure (araphid pennates) (Mann, 1984). Molecular data showed that

the pennate diatoms evolved from the centric diatoms around 125 Ma ago and that raphid pennates descended from araphids 94 ma ago (Sorhannus, 2007).

The diatom life cycle

The defining feature of diatom biology and key to understanding their evolutionary success is their diplontic life cycle, controlled by a unique, endogenous cell size reduction-restitution mechanism (Chepurnov et al., 2004, Kaczmarska et al., 2013). Yet, very little is known about the regulatory mechanisms and signalling pathways underlying diatom sexual reproduction. The vegetative, diploid cells divide mitotically, becoming smaller with each division due to the mechanics of their cell wall formation. The diatom cell wall, also called frustule, is composed of silica, making it rigid. It consists of two overlapping halves, called thecae. The smaller halve is the hypotheca, while the overlapping halve is called epitheca. After mitotic division, each daughter cell inherits one parental theca, which forms the epitheca of its new frustule, and produces a new hypotheca. These new thecae are formed while still enclosed by the cell wall of the parental cell. Consequently, the average size of a proliferating diatom population decreases while the standard deviation increases, a phenomenon known as the McDonald-Pfeitzer rule (Crawford, 1981, Lewis, 1984, Chepurnov et al., 2004). This cell size reduction implies that cells will keep dividing mitotically until they reach a critical minimal cell size and are no longer viable. Although some species developed other strategies to restore cell size, the most common strategy is to undergo sexual reproduction (Chepurnov et al., 2004). When cells reach a certain size threshold, they can enter the sexual phase of the life cycle and, as such, escape cell death since restoration of cell size occurs by the expansion of a specialized zygote, called the auxospore. A new large (initial) cell is formed inside the auxospore's envelope, which then begins a new round of vegetative multiplication. The "cardinal points", including the maximal initial cell size, the maximum size at which cells can sexually reproduce (also known as the sexual size threshold or SST) and the critical minimal size below which cells die, are

quite strict and species-specific (Drebes, 1977, Kaczmarska et al., 2013). Some species also have a lower sexual size threshold, meaning that cells lose the capacity or auxosporulation below a certain size.

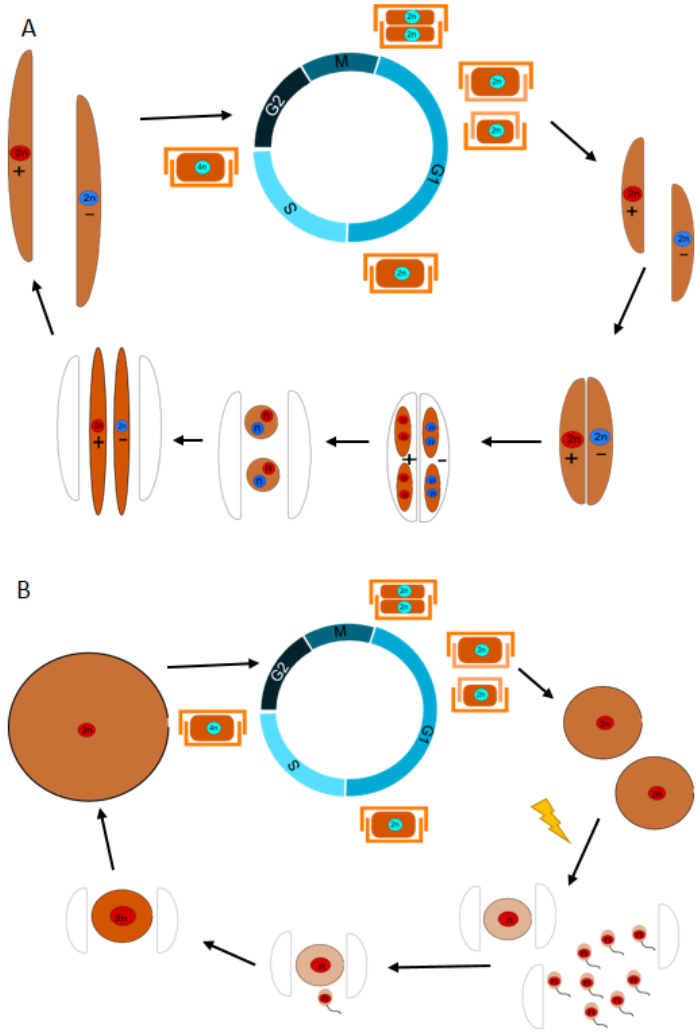


Figure 4: The basic features of the life cycle of pennate (A) and centric (B) diatoms. After every mitotic division, one of the daughter cells will become smaller than the mother cell. Below the sexual size threshold, pennates of different mating type can form mating pairs. In response to external stimuli, centrics produce egg and sperm cells. Gametes fuse to form zygotes that are able to expand (auxospores). Once the auxospores are large enough, initial cells are formed and the vegetative phase starts again.

The secondary cues to switch from vegetative to meiotic division and the specific modes of sexual reproduction itself are fundamentally different between centrics and pennates (Figure 4). Induction of sexual reproduction in centric diatoms is primarily controlled by external factors such as light irradiance, day length, temperature or nutrient availability. The primary determinant for gametogenesis in pennate diatoms is cell-cell interaction between gametangia from different sexually compatible clones (heterothally) (Chepurnov et al., 2002, Chepurnov and Mann, 2004). However, also environmental factors can be important in controlling the induction of sexual reproduction (Chepurnov et al., 2004). Some cases of seasonal control of auxosporulation have been reported, implying an influence of day length, irradiance and/or temperature on sexual reproduction in pennate diatoms (Edlund and Stoermer, 1997, Chepurnov et al., 2004).

In centric diatoms, cells from the same clone produce both male and female gametes (homothally) (Figure 4B). These gametes can be large immotile eggs or motile, flagellated sperm cells (oogamy) (Drebes, 1977, Chepurnov et al., 2004). In contrast, the gametes produced by most pennates are similar in size. The gametes of araphid pennates are morphologically and behaviorally different (anisogamy) (Chepurnov et al., 2004). Female gametes are passive and remain associated with the gametangial thecae, while male gametes are active and migrate towards the female gametes. On the contrary, most raphid pennates form gametes that are morphologically identical (Figure 4A) (Drebes, 1977, Round et al., 1990, Chepurnov et al., 2004). Additionally, they can behave either identical (isogamy) or differently (physiological anisogamy) (Kaczmarska et al., 2013). In the latter case, cis-type (gametes of one gametangium are active, while the gametes of the other gametangium are passive) and trans-type (each gametangium produces one active and one passive gamete) physiological anisogamy exist.

Recent studies have demonstrated that in pennate diatoms the initiation of sexual reproduction between two different mating types (MT) requires the production of multiple sex pheromones (Figure 5) (Sato et al., 2011, Gillard et al.,

2013). Sato and colleagues reported experimental evidence for the involvement of diatom sex pheromones during the sexualization of the araphid pennate diatom *Pseudostaurosira trainorii*. They were the first to demonstrate the induction of gametogenesis by cell-free medium derived from cultures of the opposite sex. The existence of two pheromones was experimentally demonstrated. Ph-1 is secreted by female vegetative clones below the SST and induces sexualization of male clones, leading to the release of two motile gametes. The sexualized male cells and/or the male gametes secrete ph-2, which stimulates sexualization of female cells, which then produce two egg cells. In co-culture, male gametes change shape and become amoeboid when random walk brings them in close proximity of female gametangia or mature eggs. Cell-cell contact is then made through directional movement of the amoeboid male gametes towards the female egg cells. Since, ph-1 does not seem to induce amoeboid movement of the male gametes, it is hypothesized that a third pheromone (ph-3) is present. Ph-3 is secreted by female gametes but cannot be retained in the filtrates or agar gels used in the bio-assays, meaning that it is probably short-lived or very volatile.

Similar bio-assays with the raphid pennate diatom *Seminavis robusta* led to the discovery of a comparable yet different mechanism to induce mating (Gillard et al., 2013). In contrast to *P. trainorii*, gametogenesis in *S. robusta* only occurs after physical contact between cells of different mating type. Nonetheless, they also use pheromones to establish this cell-cell contact. In a first phase, vegetative cells of both mating type below the SST secrete biochemicals, which are called conditioning factors, indicated here as CF-P (secreted by MT^+) and CF-M (secreted by MT^-). When conditioned by the presence of CF-P, MT^- cells start producing a third pheromone that attracts MT^+ clones. At the same time, sensing CF-M makes MT^+ cells susceptible to this attraction pheromone, presumably by inducing the appearance of receptors at the plasma membrane. Combining bio-assays with metabolomics led to the fractionation of this attraction pheromone and its identification as L-diproline. Thus, although the mechanisms can vary between different species, pennate diatoms use multiple chemical signals to induce sexual

reproduction. Presumably, pheromone signalling is used to ensure the presence of a sexually mature mating partner before investing in sexual response.

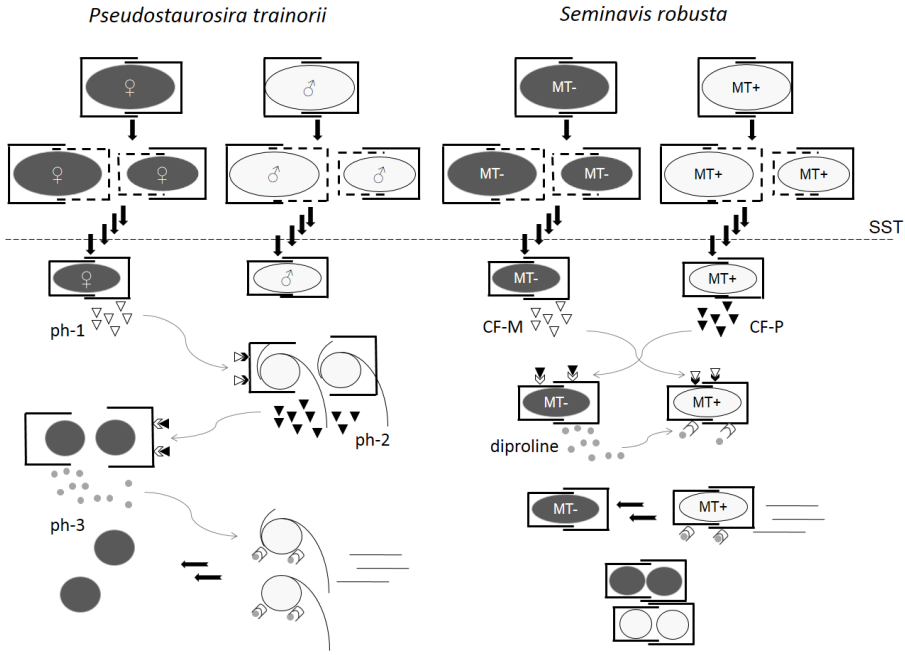


Figure 5: Sex pheromone signalling in *Pseudostaurosira trainorii* (left) and *Seminavis robusta* (right). Both species produce sex pheromones when they pass the SST. In *P. trainorii* first female vegetative cells produce a factor (ph-1) that induces male gametogenesis. Then, the male gametangia and/or gametes produce a second pheromone (ph-2) that induces female gametogenesis. The female gametes probably secrete a third pheromone (ph-3) that directs the male gametes towards them. In *S. robusta*, both mating types secrete a pheromone after passing the SST (CF-P and CF-M). CF-P induces the production of a third pheromone (diproline) by MT^- cells that attracts MT^+ cells. Only after physical contact between MT^+ and MT^- cells is established, gametogenesis is induced.

Sex determination in eukaryotes

Sexual reproduction is conserved in all groups of the eukaryotic tree of life and yet the mechanisms that determine sexual identity are remarkably diverse.

Although the benefits of sexual reproduction are still under debate, two general models have emerged: in one, sexual recombination leads to fitter progeny, in the other sex allows organisms to remove deleterious mutations from their genome (Barton and Charlesworth, 1998, Lenski, 2001).

The most elaborately studied sex determination system is the XY system, as seen in most mammals and in genetic model systems like *Drosophila melanogaster* and *Caenorhabditis elegans*. This created the false impression that sex is usually genotypically determined by one master-switch gene lying on highly heteromorphic sex chromosomes (Bachtrog et al., 2014). However, numerous variations on this theme exist, for example the degree of sex chromosome differentiation can vary greatly and also the master-switch gene can be different, even in closely related species.

Genotypic versus environmental sex determination

In mammals, for example, sexual identity is determined by one or more genes, often lying on sex chromosomes. This genotypic sex determination will be discussed below.

Sometimes sex is determined by environmental instead of genetic factors. In environmental sex determination, external stimuli, like temperature, photoperiod or social factors determine sexual identity (Bachtrog et al., 2014). One well-studied example is temperature-dependent sex determination in reptiles (Bull, 1980), where incubation temperature determines sex by altering the hormone environment of the embryo (Crews, 1996). Also some fish exhibit temperature-dependent sex determination (Ospina-Alvarez and Piferrer, 2008). Recently, it was shown in the European sea bass *Dicentrarchus labrax* that the promoter of gonadal aromatase (cyp19a), that converts androgens into estrogens, is differentially methylated between male and female juveniles and that this methylation is temperature-sensitive (Navarro-Martin et al., 2011). Thus, DNA methylation of the aromatase promoter may be an essential component of the mechanism involved in temperature-dependent sex determination.

In some cases, a combination of genotypic and environmental sex determination is seen. Here, the otherwise genetically determined sex is altered by specific environmental factors. For example, half-smooth tongue sole (*Cynoglossus semilaevis*) has a female heterogametic sex determination system (ZW), but a rise in temperature during the sensitive developmental stage can cause sex reversal of ZW genetic females to phenotypic males (pseudomales) (Chen et al., 2014). It was shown that sexual reversal involves DNA methylation modifications in sex determination pathways (Shao et al., 2014). These methylation modifications in pseudomales are inherited by their ZW offspring, which then naturally develop into pseudomales without temperature incubation.

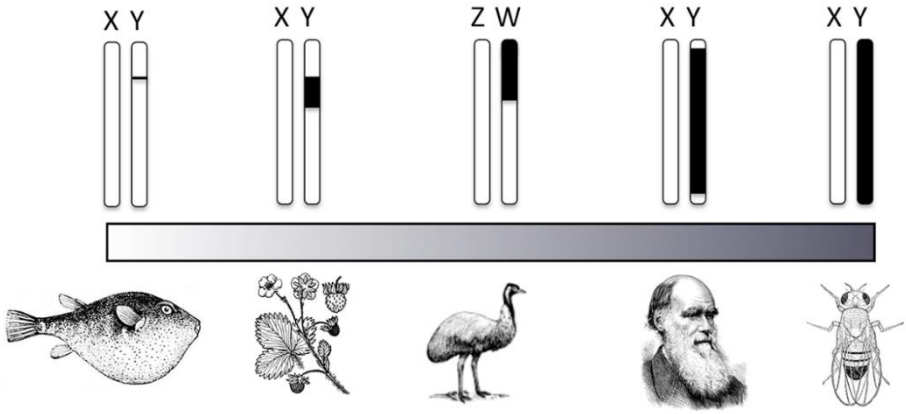


Figure 6: Sex chromosome differentiation ranges from a single sex-determining locus (pufferfish), a small differentiated region (strawberry and emu), over a large differentiated region with short recombining regions (humans), to differentiation along the entire sex chromosome pair (*Drosophila)*. Sex chromosomes are not drawn to scale (Bachtrog et al., 2014).

Sex chromosome differentiation

Sex chromosomes can be either homomorphic or heteromorphic (Figure 6). Heteromorphic sex chromosomes are cytologically distinguishable due to genetic degeneration in one of the chromosomes. These chromosomes cannot recombine with their homolog or they have only a small region that is able to recombine, called the pseudo-autosomal region (PAR) (Bergero and Charlesworth, 2009). In

some species, the PAR is still quite extensive and the non-recombining region can be very small. These so-called homomorphic chromosomes are morphologically the same.

In *Drosophila* species, the X and Y chromosomes share no sequence homology. The Y chromosome contains mainly repetitive DNA and only a few protein-coding genes, that have male-related functions and originated by duplication of autosomal genes (Koerich et al., 2008, Carvalho et al., 2009). The gene content of the Y chromosome differs between different *Drosophila* species. But these Y-linked genes do not determine sex in flies, instead it is determined by the number of X chromosomes present (Erickson and Quintero, 2007). In humans, small recombining regions still exist on the X and Y chromosomes (Graves, 2006). Human sex chromosomes are highly differentiated due to the suppression of recombination in the initially identical chromosomes. The number of genes residing on the X (about 1,000 genes) or Y chromosome (about 50 genes) differ greatly and most of them are not directly involved in sex determination. In many flowering plants, where sex chromosomes evolved only recently, the recombining region is still quite large. In papaya, the male-specific region of the Y chromosome is estimated to be only 8-9 Mbp (Yu et al., 2008). Unlike in plants and animals, in fungi sex (or mating type) is determined by a mating type locus. In *Saccharomyces cerevisiae*, the mating type-specific region is only 642 bp long in a-cells and 747 bp in α-cells, which is only a small fraction of the >300 Mbp chromosome on which the mating type locus resides (Fraser and Heitman, 2004). The two alleles of the mating type locus (MATa and MATα) in yeast encode DNA binding proteins that regulate the expression of mating type determining genes on other chromosomes (Herskowitz et al., 1992). Unlike in *Drosophila*, where the complete sex chromosomes are differentiated, in the tiger pufferfish *Takifugu rubripes* sex is determined by one SNP, which causes an amino acid change (His/Asp384) in the kinase domain of the anti-Müllerian hormone receptor type II (*Amhr2*) gene (Kamiya et al., 2012). While females are homozygous (His/His384), males are

heterozygous (His/Asp384) at this position. The sex determination locus shows no sign of recombination suppression between X and Y chromosomes.

Male versus female heterogamety

The most familiar sex determination system is the XY chromosome system as seen in humans. Here, females have two X chromosomes and males have one X and one Y, meaning that males are the heterogametic sex. But this is not always the case, also females can be heterogametic. This is then called a ZW system and is found in many birds, where males have two Z chromosomes and females have one Z chromosome and one female-specific W chromosome. Although XY and ZW chromosomes evolved from different autosomes with different gene content, they share some similarities (Stiglec et al., 2007, Graves, 2014). Both Y and W are strongly degraded and contain a high number of repetitive sequences and both Y and Z chromosomes have accumulated genes with male-advantage functions. On the other hand, dosage compensation (i.e. the mechanism to equalize expression of X- or Z-linked genes between males and females, as well as between sex and autosomal chromosomes) works differently between XY and ZW systems. As the Y and W chromosomes have lost the majority of their genes, expression differences of X- or Z-borne genes between males and females have to be compensated. Diverse mechanisms are developed to balance the expression of X-linked genes between males and females, like the inactivation of one X chromosome in mammals or the increased expression of X-linked genes in males in *Drosophila* (Nguyen and Disteche, 2006, Gelbart and Kuroda, 2009). This is not the case in birds, where most genes show partial compensation and this compensation ratio can differ between genes, meaning that the expression of only a part of the genes is balanced between males and females (Itoh et al., 2007, Wolf and Bryk, 2011).

Variation in the master-switch gene

In the most-studied model species, sex is determined by one master-switch gene (Bachtrog et al., 2014). This gene is responsible for switching on the male or female developmental program. In mammals, sex is determined by the master-

switch gene *Sry,* that lies on the Y chromosome. Early in embryonic development, *Sry* expression activates testis-specific developmental programs. Without *Sry* expression, the ovary-specific pathways are activated (Lovellbadge, 1992). In Drosophila, *sex-lethal* (*sxl*) is the master-switch gene (Figure 7) (Gilbert, 2000). This gene lies on the X chromosome and is transcribed in the early stages of development, but only when the X-to-autosome ratio is 2X:2A (XX in a diploid background) (Salz et al., 1989). In males, where the X-to-autosome ratio is 1X:2A, another splice-variant of *sxl* is produced, which leads to an inactive protein. In females, sxl controls the processing of *transformer* (*tra*) transcripts, leading to a female-specific mRNA (Boggs et al., 1987). The presence of female-specific tra together with tra2 leads to female-specific processing of the *doublesex* (*dsx*) transcript (Ryner and Baker, 1991). This female-specific dsx protein activates female-specific genes and inhibits male development (Coschigano and Wensink, 1993). In contrast, when tra is not present, male-specific dsx protein is produced, leading to activation of male development and inhibition of female traits (Jursnich and Burtis, 1993). Also in the nematode *C. elegans,* sex is determined by the X-to-autosome ratio; animals with one X are males, while animals with two X chromosomes are hermaphrodite (Zarkower, 2006). This ratio controls the expression of the master-switch gene *XO lethal 1* (*xol1*), that is abundant in males and repressed in hermaphrodites. Xol-1 is then responsible for regulating the downstream sex determination pathway (Rhind et al., 1995).

Some genes are repeatedly found as a master-switch gene in animals. One such gene is *Dsx*- and *mab-3*-related transcription factor 1 (*dmrt1*), which belongs to a family of transcription factors that contain a DM domain (Matson and Zarkower, 2012). Members of this family probably play a role in sexual differentiation downstream of the primary sex determinant in all major animal lineages, for example *dsx* in *Drosophila* (see above) or *mab-3*, *mzb-23* and *dmd-3*, that function in male development of *C. elegans* (Lints and Emmons, 2002, Shukla and Nagaraju, 2010). Members of this gene family are recruited as primary sex determinant in fish, birds and frogs (Matson and Zarkower, 2012). In the teleost

fish medaka (*Oryzias latipes*) a new Y chromosome was formed, which contains a gene that resulted from the duplication of the autosomal *Dmrt1a*, called DM domain on Y (*Dmy*) (Matsuda et al., 2002). *Dmy* is the master switch gene regulating male development analogously to *Sry* in mammals (Matsuda et al., 2007). *Dmrt1* is a Z-linked gene in all birds and knocking it down in chick embryos led to a feminized morphology in ZZ-chicks, suggesting that a higher *Dmrt1* dose in males is responsible for sex determination (Smith et al., 1999, Smith et al., 2009). In the frog *Xenopus laevis*, the W chromosome comprises a duplicated variant of the autosomal *dmrt1*, called *dm-w* (Yoshimoto et al., 2008). This gene probably functions as a dominant-negative inhibitor of male development by binding the autosomal *dmrt1* and as such inducing female development.

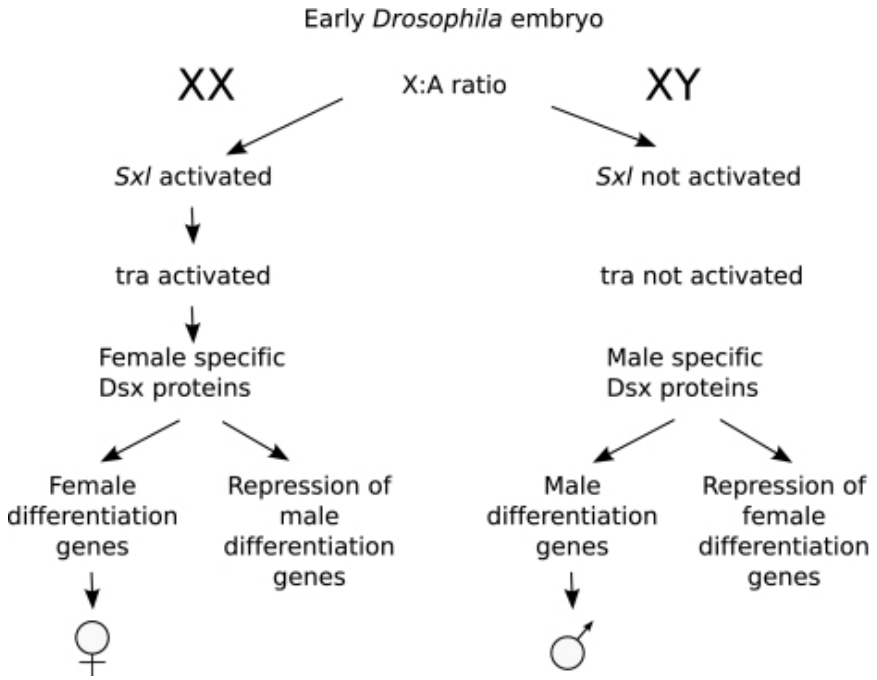


Figure 7: The genetic sex determination pathway of *D. melanogaster* (Ah-King and Nylin, 2010). In XX embryos, the X:A ratio is 2:2. This leads to the activation of *sxl*, which in turn activates *tra.* Tra induces female-specific splicing of *dsx*, of which the product activates female differentiation and represses male developmental pathways. In XY embryos, the X:A ratio is 1:2. Consequently, *sxl* and *tra* are not activated. This leads to male-specific splicing of *dsx*, of which the product activates male differentiation and inhibits female development.

On the other hand, also primary determinants without homologs in closely-related species are known, for example *Sxl* in *Drosophila* or *sdY* in the rainbow trout *Oncorhynchus mykiss*. This last gene lies on the Y chromosome and is necessary for male development in rainbow trout (Yano et al., 2012). It displays sequence similarity to the C-terminal domain of interferon regulatory factor 9, a factor known for its role in immunity, but not yet implicated in sexual differentiation before.

Although one master-switch gene is responsible for sex determination in the organisms described above, sometimes multiple genes may be involved in this process (Bachtrog et al., 2014). For example in zebrafish, several studies report that sex determination is a polygenic trait, with loci on multiple chromosomes being linked to sex (Bradley et al., 2011, Liew et al., 2012). This is also the case for other fish, like the cichlid and European sea bass (Vandeputte et al., 2007, Ser et al., 2010).

Sex determination in the SAR group

Knowledge about sex determination in the SAR lineage is largely limited to the ciliate *Tetrahymena thermophile*, the brown alga *Ectocarpus*, the oomycetes *Phytophthora* and the diatom *S. robusta*.

Mating type determination in the ciliate *Tetrahymena thermophila*

T. thermophila has two nuclei, a germline nucleus, that is inactive during the vegetative stage of its life cycle, and a somatic nucleus, that is responsible for transcription during vegetative growth. This species has seven different mating types and mating is possible with a cell of any mating type but its own (Nanney and Caughey, 1953). After fertilization, a new somatic nucleus is formed by rearranging a copy of the germline nucleus (Figure 8). The germline nucleus contains a tandem array of incomplete gene pairs for each mating type and during these rearrangements the mating type of the offspring is determined by completing one

of these gene pairs and eliminating all others from the new somatic nucleus (Cervantes et al., 2013). These gene pairs code for two transmembrane proteins, MTA and MTB. Since transmembrane proteins can localize to the cell membrane, these proteins might be involved in self/non-self-recognition, which is important for mating in *Tetrahymena* (Mccoy, 1972). MTA and MTB homologs are found in the somatic genome of several other *Tetrahymena* species (Cervantes et al., 2013).

Sex determination in the brown alga *Ectocarpus sp.*

In *Ectocarpus*, sex is determined during the haploid phase of the life cycle. Such a sex determination system is called UV, to distinguish it from the XY and ZW systems, where one of the sexes is heterozygous (Bachtrog et al., 2011). During the haplodiploid life cycle of *Ectocarpus*, the diploid sporophyte produces meiospores that develop into separate male (V) and female (U) gametophytes. Both the U and V chromosome contains a large amount of repetitive DNA and gene density is low (Ahmed et al., 2014). The sex-determining regions constitute only one-fifth of the sex chromosomes (about 1 Mbp), despite their age that is reflected by the high divergence between the male and female sex-determining regions. The male sex determining region is dominant over the female haplotype, implying that femaleness is the default state that is expressed when the male haplotype is absent (Ahmed et al., 2014). One of the nine male-specific genes encodes a high mobility group (HMG) domain (Ahmed et al., 2014). This protein family of transcription factors is involved in mating type or sex determination in fungi and vertebrates (Foster et al., 1992, Idnurm et al., 2008).

Mating type determination in the oomycete *Phytophthora*

Heterothallic species belonging to the oomycete *Phytophtora* genus have two mating types, called A1 and A2 (Brasier, 1992). In *P. infestans*, a single mating type locus is found, which segregates in a non-Mendelian fashion, that could be caused by linkage to balanced lethal loci (Judelson et al., 1995, Judelson, 1996). In *P. parasitica*, on the other hand, Mendelian segregation of the mating type locus was observed (Fabritius and Judelson, 1997). In this species, mating type A1 is

heterozygous, while mating type A2 is homozygous for the mating type locus. Nothing is known about the identity of the mating type locus of *Phytophtora*.

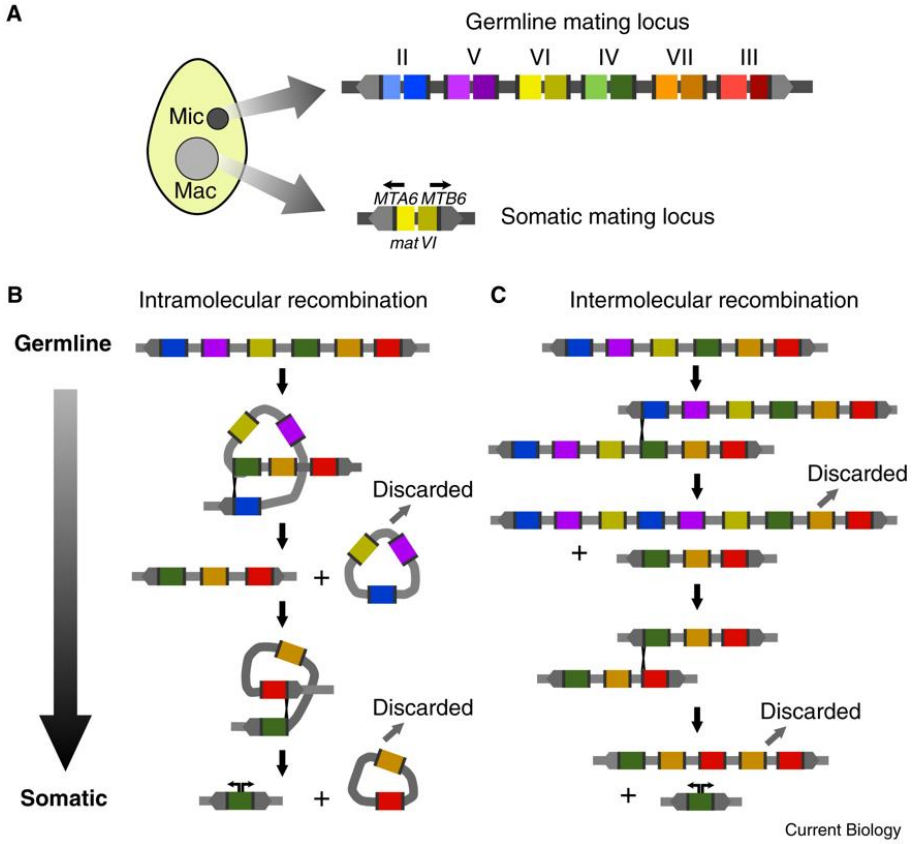


Figure 8: Mating type determination in *Tetrahymena thermophila*. (A) On the left is a schematic of a *T. thermophila* cell with a germline micronucleus (Mic) and a somatic macronucleus (Mac). On the right are diagrams depicting the organization of the germline mating locus (top) and the somatic mating locus (bottom). In this example, the somatic nucleus expresses the mat VI allele, existing of two different genes, MTA6 and MTB6. (B and C) Two models for the elimination of alternative mat-genes during somatic nucleus formation. The different mat-alleles are color coded and regions of high similarity used for recombination are shown as dark boxes. (B) Model adapted from Cervantes et al. (2013) showing elimination of alternative mat-alleles through successive intramolecular recombination events. (C) Alternative model showing elimination of germline mat alleles by intermolecular recombination (Umen, 2013).

Mating type determination in the pennate diatom *S. robusta*

Pennate diatoms are mostly heterothallic, meaning that sexual reproduction requires the presence of two compatible partners (Chepurnov et al., 2004). Two mating types are identified in pennates, called MT^+ and MT^-. Together with the evolution from homothally (centrics) to heterothally (pennates), genetic mating type determination originated in diatoms (Mann et al., 2003, Davidovich et al., 2010). Using AFLP-based sex-specific linkage maps, the mating type locus of *S. robusta* was recently identified as being a single locus (Vanstechelman et al., 2013). This study also showed that MT^+ is the heterogametic mating type. Further analysis and characterization of the genes and sequence polymorphism(s) associated with the mating type locus will be crucial to understand the regulation of the life cycle of *S. robusta* and to evaluate the conservation of the mating type determination system across different diatom species.

Seminavis robusta as a model organism

The two most commonly used model diatoms in molecular biology are the centric diatom *Thalassiosira pseudonana* and the pennate diatom *Phaeodactylum tricornutum*. For both species the genome sequence is available (Armbrust et al., 2004, Bowler et al., 2008) and they can be easily genetically transformed (Apt et al., 1996, Poulsen et al., 2006). During the last decade, these two species yielded much knowledge about diatom biology. The genome projects led to the discovery of the urea cycle in diatoms, countering the assumption that this pathway originated in metazoans (Allen et al., 2011). Also other metabolic pathways were characterized in *P. tricornutum*, like the Entner-Doudoroff glycolytic pathway or sterol biosynthesis (Fabris et al., 2012, Fabris et al., 2014). Also light reception and acclimation is an important research topic in these photosynthetic organisms (Nymark et al., 2009, Lepetit et al., 2013, Wilhelm et al., 2014). In *P. tricornutum*, cell cycle regulation by environmental cues is extensively studied (Huysman et al., 2010, Huysman et al., 2013, Huysman et al., 2014). Furthermore, diatom-specific

properties, like silica deposition, are being elucidated in *T. pseudonana* (Sumper and Brunner, 2008, Poulsen et al., 2013).

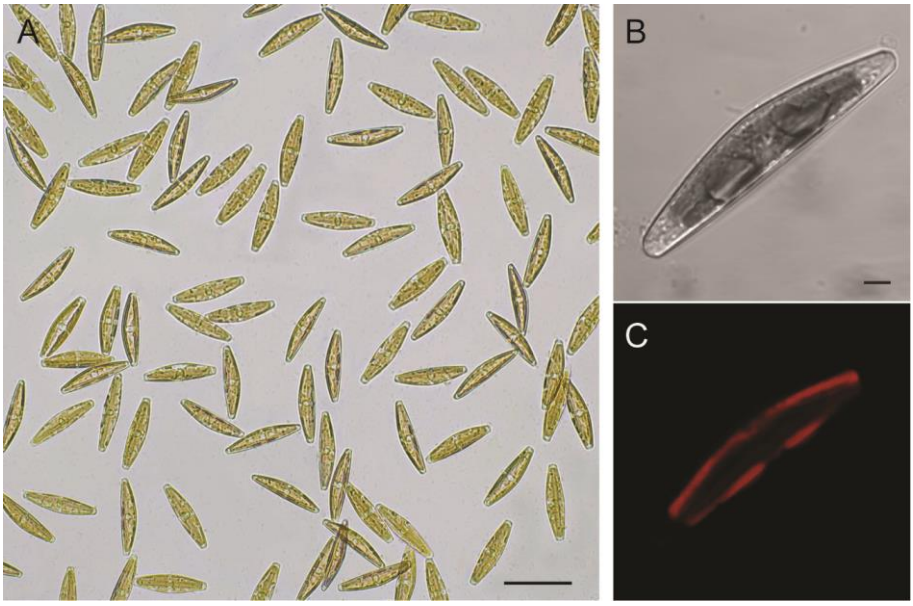


Figure 9: Illustration of our model species *Seminavis robusta*. (A) Light microscopy image of an *S. robusta culture*. Scale bar = 50 µm. (B+C) Confocal laser scanning microscopy of an *S. robusta* cell. B: transmission light. C: autofluorescence of the chloroplasts in red. Scale bar = 5 µm. (Adapted from Huysman et al. (2014))

The major advantage of *P. tricornutum* and *T. pseudonana* is the fact that they don't follow the McDonald-Pfeitzer rule. The observation that the region of valve synthesis is wider than other parts of the cells in *T. pseudonana* can possibly explain the lack of size reduction in this species (Hildebrand et al., 2007). For *P. tricornutum*, the explanation probably lies in the (near) absence of silica in its cell wall, making it less rigid and probably able to expand after vegetative division (Johansen, 1991). Consequently, they can be kept in culture for an infinite amount of time. Despite numerous attempts to induce sex in culture, no auxosporulation has been observed for these two model species. Although being an advantage in many studies, this property is a major disadvantage when studying the diatom life cycle, because these well-established model species don't exhibit this characteristic

diatom life cycle. Moreover, using a sexual model species enables the use of forward genetics to discover genes responsible for a certain phenotype. After mutagenesis and screening of the population, selected mutants have to be crossed to map the mutations to the genome. This is not feasible is asexual species, like *P. tricornutum* of *T. pseudonana* (Saade and Bowler, 2009).

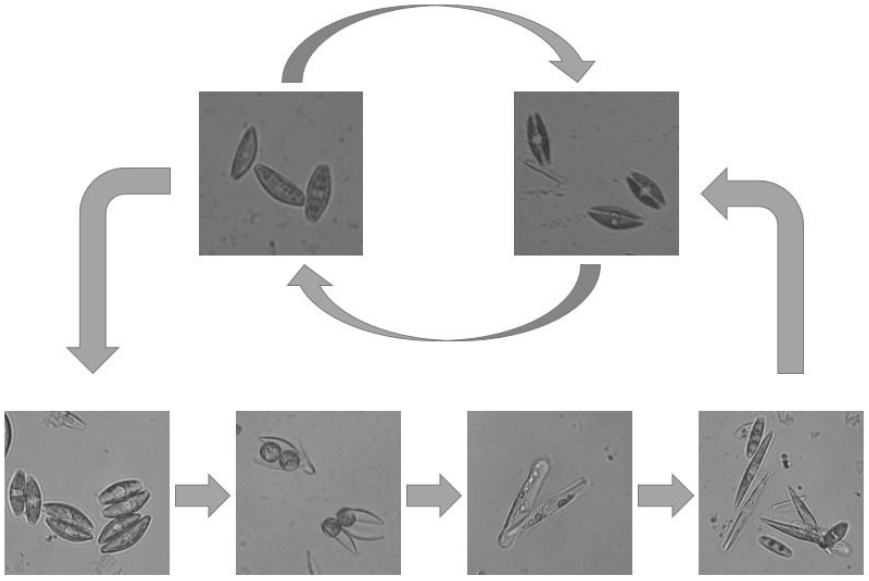


Figure 10: The life cycle of *Seminavis robusta*. Top: vegetative replication. Bottom: sexual reproduction, from left to right: mate pairs, zygotes, auxospores and initial cells.

To resolve this issue, *S. robusta* was introduced as a new model system to study the diatom life cycle (Figure 9) (Chepurnov et al., 2008). This species has a typical pennate life cycle with a size reduction-restitution cycle and size restoration through sexual auxosporulation (Figure 10) (Chepurnov et al., 2002). *S. robusta* is heterothallic, which makes the induction of sex easily controllable, since sexual reproduction can only occur when two compatible clones are mixed. Furthermore, high frequencies of mate pairing and gametogenesis can be obtained when the cell cycle of the culture is synchronized by prolonging the dark period. Moreover, their large cell size (20 - 80 µm), together with their benthic life style (i.e. growing on surfaces, like the bottom of a culture flask) make it easy to monitor cultures

directly under the low magnification of an inverted microscope. It is also easy to isolate single cells under a dissecting microscope.

A breeding program was initiated to generate the first 'long-term' diatom pedigree (Figure 11) (Chepurnov et al., 2008). This demonstrated the tolerance of *S. robusta* to inbreeding, because crossing sibling clones rendered vigorous progeny, even when the parents were themselves progeny of a sibling cross. This would make it possible to create inbred lines without lower fitness. The strains resulting from the breeding program are cryopreserved and deposited in the BCCM/DCG diatoms catalogue (http://bccm.belspo.be), which now contains 172 different *S. robusta* strains.

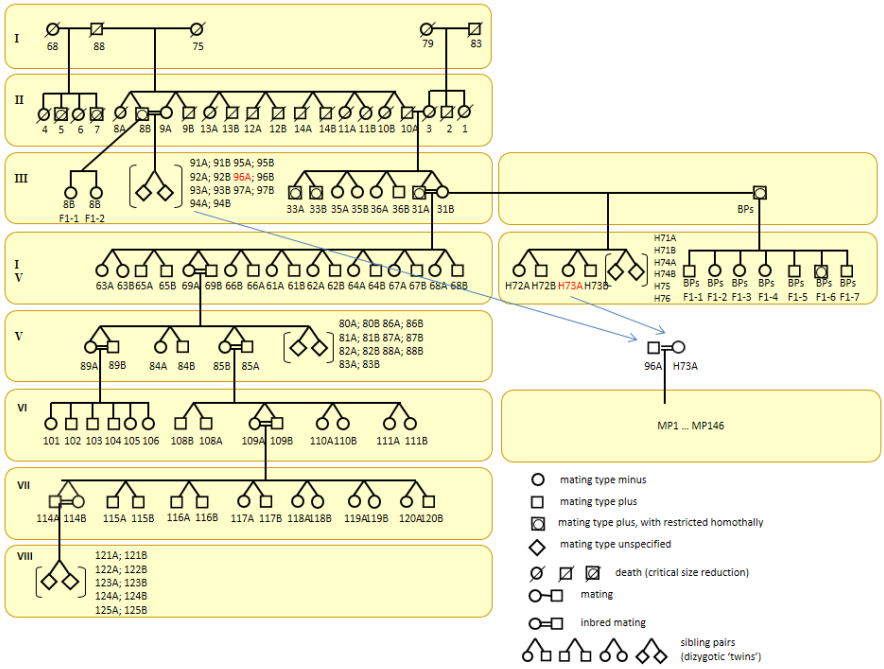


Figure 11: A pedigree of *Seminavis robusta* obtained through interclonal crosses. The founder clones are natural clones isolated from The Netherlands, Zeeland, "Veerse Meer" (51°32'36"N; 3°48'15"E) (Chepurnov et al., 2002). Clone BPs is also a natural clone, isolated from The Netherlands, Westerschelde, Paulinaschor (51°21'N; 3°44'E). Strains 96A and H73A were crossed to generate the F1 mapping population consisting of 146 clones (MP1-MP146) and analysed by Vanstechelman et al. (2013).

In recent years, several high impact papers were published on *S. robusta*. The relationship between cell cycle progression and chloroplast development was explored and a first expression study on *S. robusta* was conducted using cDNA-AFLP (Gillard et al., 2008). *S. robusta* was the first diatom for which the structure of a sex pheromone was elucidated (Gillard et al., 2013) and for which the sex determining region was mapped (Vanstechelman et al., 2013). Last year, the chloroplast genome of *S. robusta* was published, showing that diatom plastids are subject to major changes resulting from horizontal gene transfer (Brembu et al., 2014).

Aims and thesis outline

In this dissertation, we want to expand our knowledge on the life cycle of pennate diatoms, using *S. robusta* as a model organism. To improve *S. robusta* as a model system, a genetic transformation protocol was developed and a large amount of next-generation sequencing data has been generated.

Chapter 2 describes the characterization of the mating type locus of *S. robusta*. Bulked segregant analysis was used to fine-map the mating type locus that was identified in Vanstechelman et al. (2013). By mapping these markers on the draft genome, *DNMT5a* was found to be located at the mating type locus. Furthermore, SNPs cosegregating with the mating type were identified.

Chapter 3 summarizes the results of the "A deep transcriptomic and genomic investigation of diatom life cycle regulation" project, for which RNA-seq was conducted on samples from different stages of the life cycle of *S. robusta*. This transcriptome was used to annotate the cyclin gene family, identify genes involved in meiosis and in silica deposition, and to study their expression profile.

In **chapter 4**, the effect of the conditioning factor produced by MT^+ (CF-P) on gene expression in MT^- cells was more closely examined. It was shown that CF-P induces cell cycle arrest in MT^- and that diproline production increases by upregulating glutamate-to-proline conversion. RNA-seq provided us a first indication that the secondary messenger cGMP is involved in pheromone signalling in *S. robusta*.

Chapter 5 describes the development of a genetic transformation protocol for *S. robusta*. Such a protocol will allow reverse genetics in this model species, which is important to further characterize the molecular regulation of the diatom life cycle.

Finally, in **chapter 6** the main conclusions of this thesis are discussed and also some recommendations for future research are provided.

Literature cited

Ah-King, M. and Nylin, S. (2010) Sex in an Evolutionary Perspective: Just Another Reaction Norm. *Evolutionary Biology* **37**, 234-246.

Ahmed, S., Cock, J.M., Pessia, E., Luthringer, R., Cormier, A., Robuchon, M., Sterck, L., Peters, A.F., Dittami, S.M., Corre, E., Valero, M., Aury, J.M., Roze, D., Van de Peer, Y., Bothwell, J., Marais, G.A. and Coelho, S.M. (2014) A Haploid System of Sex Determination in the Brown Alga Ectocarpus sp. *Current Biology* **24**, 1945–1957.

Al-Degs, Y., Khraisheh, M.A.M. and Tutunji, M.F. (2001) Sorption of lead ions on diatomite and manganese oxides modified diatomite. *Water Research* **35**, 3724-3728.

Allen, A.E., Dupont, C.L., Obornik, M., Horák, A., Nunes-Nesi, A., McCrow, J.P., Zheng, H., Johnson, D.A., Hu, H., Fernie, A.R. and Bowler, C. (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**, 203-207.

Apt, K.E., KrothPancic, P.G. and Grossman, A.R. (1996) Stable nuclear transformation of the diatom Phaeodactylum tricornutum. *Molecular & General Genetics* **252**, 572-579.

Archibald, J.M. (2009) The Puzzle of Plastid Evolution. *Current Biology* **19**, R81-R88.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., Brzezinski, M.A., Chaal, B.K., Chiovitti, A., Davis, A.K., Demarest, M.S., Detter, J.C., Glavina, T., Goodstein, D., Hadi, M.Z., Hellsten, U., Hildebrand, M., Jenkins, B.D., Jurka, J., Kapitonov, V.V., Kröger, N., Lau, W.W.Y., Lane, T.W., Larimer, F.W., Lippmeier, J.C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M.S., Palenik, B., Pazour, G.J., Richardson, P.M., Rynearson, T.A., Saito, M.A., Schwartz, D.C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F.P. and Rokhsar, D.S. (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79-86.

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

Aw, M.S., Simovic, S., Yu, Y., Addai-Mensah, J. and Losic, D. (2012) Porous silica microshells from diatoms as biocarrier for drug delivery applications. *Powder Technology* **223**, 52-58.

Bachtrog, D., Kirkpatrick, M., Mank, J.E., McDaniel, S.F., Pires, J.C., Rice, W.R. and Valenzuela, N. (2011) Are all sex chromosomes created equal? *Trends in Genetics* **27**, 350-357.

Bachtrog, D., Mank, J.E., Peichel, C.L., Kirkpatrick, M., Otto, S.P., Ashman, T.L., Hahn, M.W., Kitano, J., Mayrose, I., Ming, R., Perrin, N., Ross, L., Valenzuela, N., Vamosi, J.C. and Consortium, T.S. (2014) Sex Determination: Why So Many Ways of Doing It? *PLoS Biology* **12**.

Bariana, M., Aw, M.S. and Losic, D. (2013) Tailoring morphological and interfacial properties of diatom silica microparticles for drug delivery applications. *Advanced Powder Technology* **24**, 757-763.

Barton, N.H. and Charlesworth, B. (1998) Why sex and recombination? *Science* **281**, 1986-1990.

Bergero, R. and Charlesworth, D. (2009) The evolution of restricted recombination in sex chromosomes. *Trends in Ecology & Evolution* **24**, 94-102.

Boggs, R.T., Gregor, P., Idriss, S., Belote, J.M. and McKeown, M. (1987) Regulation of sexual differentiation in D. melanogaster via alternative splicing of RNA from the transformer gene. *Cell* **50**, 739-747.

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J.A., Brownlee, C., Cadoret, J.-P., Chiovitti, A., Choi, C.J., Coesel, S., De Martino, A., Detter, J.C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M.J.J., Jenkins, B.D., Jiroutova, K., Jorgensen, R.E., Joubert, Y., Kaplan, A., Kröger, N., Kroth, P.G., La Roche, J., Lindquist, E., Lommer, M., Martin-Jézéquel, V., Lopez, P.J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L.K., Montsant, A., Oudot-Le Secq, M.-P., Napoli, C., Obornik, M., Parker, M.S., Petit, J.-L., Porcel, B.M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T.A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M.R., Taylor, A.R., Vardi, A., von Dassow, P., Vyverman, W., Willis, A., Wyrwicz, L.S., Rokhsar, D.S., Weissenbach, J., Armbrust, E.V., Green, B.R., Van de Peer, Y. and Grigoriev, I.V. (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239-244.

Bradley, K.M., Breyer, J.P., Melville, D.B., Broman, K.W., Knapik, E.W. and Smith, J.R. (2011) An SNP-Based Linkage Map for Zebrafish Reveals Sex Determination Loci. *G3-Genes Genomes Genetics* **1**, 3-9.

Brasier, C.M. (1992) Evolutionary Biology of Phytophthora .1. Genetic System, Sexuality and the Generation of Variation. *Annual Review of Phytopathology* **30**, 153-171.

Brembu, T., Winge, P., Tooming-Klunderud, A., Nederbragt, A.J., Jakobsen, K.S. and Bones, A.M. (2014) The chloroplast genome of the diatom Seminavis robusta: New features introduced through multiple mechanisms of horizontal gene transfer. *Marine Genomics* **16**, 17-27.

Bull, J.J. (1980) Sex Determination in Reptiles. *Quarterly Review of Biology* **55**, 3-21.

Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjaeveland, A., Nikolaev, S.I., Jakobsen, K.S. and Pawlowski, J. (2007) Phylogenomics Reshuffles the Eukaryotic Supergroups. *PLoS ONE* **2**.

Burki, F., Okamoto, N., Pombert, J.F. and Keeling, P.J. (2012) The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. *Proceedings of the Royal Society B-Biological Sciences* **279**, 2246-2254.

Burki, F. (2014) The Eukaryotic Tree of Life from a Global Phylogenomic Perspective. *Cold Spring Harbor Perspectives in Biology* **6**.

Carvalho, A.B., Koerich, L.B. and Clark, A.G. (2009) Origin and evolution of Y chromosomes: Drosophila tales. *Trends in Genetics* **25**, 270-277.

Cavalier-Smith, T. (1999) Principles of protein and lipid targeting in secondary symbiogenesis: Euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *Journal of Eukaryotic Microbiology* **46**, 347-366.

Cervantes, M.D., Hamilton, E.P., Xiong, J., Lawson, M.J., Yuan, D.X., Hadjithomas, M., Miao, W. and Orias, E. (2013) Selecting One of Several Mating Types through Gene Segment Joining and Deletion in Tetrahymena thermophila. *PLoS Biology* **11**.

Chen, S.L., Zhang, G.J., Shao, C.W., Huang, Q.F., Liu, G., Zhang, P., Song, W.T., An, N., Chalopin, D., Volff, J.N., Hong, Y.H., Li, Q.Y., Sha, Z.X., Zhou, H.L., Xie, M.S., Yu, Q.L., Liu, Y., Xiang, H., Wang, N., Wu, K., Yang, C.G., Zhou, Q., Liao, X.L., Yang, L.F., Hu, Q.M., Zhang, J.L., Meng, L., Jin, L.J., Tian, Y.S., Lian, J.M., Yang, J.F., Miao, G.D., Liu, S.S., Liang, Z., Yan, F., Li, Y.Z., Sun, B., Zhang, H., Zhang, J., Zhu, Y., Du, M., Zhao, Y.W., Schartl, M., Tang, Q.S. and Wang, J. (2014) Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nature Genetics* **46**, 253-260.

Chepurnov, V.A., Mann, D.G., Vyverman, W., Sabbe, K. and Danielidis, D.B. (2002) Sexual reproduction, mating system, and protoplast dynamics of *Seminavis* (Bacillariophyceae). *Journal of Phycology* **38**, 1004-1019.

Chepurnov, V.A. and Mann, D.G. (2004) Auxosporulation of *Licmophora communis* (Bacillariophyta) and a review of mating systems and sexual reproduction in araphid pennate diatoms. *Phycological Research* **52**, 1-12.

Chepurnov, V.A., Mann, D.G., Sabbe, K. and Vyverman, W. (2004) Experimental studies on sexual reproduction in diatoms. *International Review of Cytology* **237**, 91-154.

Chepurnov, V.A., Mann, D.G., von Dassow, P., Vanormelingen, P., Gillard, J., Inzé, D., Sabbe, K. and Vyverman, W. (2008) In search of new tractable diatoms for experimental biology. *BioEssays* **30**, 692-702.

Coschigano, K.T. and Wensink, P.C. (1993) Sex-Specific Transcriptional Regulation by the Male and Female Doublesex Proteins of Drosophila. *Genes & Development* **7**, 42-54.

Crawford, R.M. (1981). Some considerations of size reduction in diatom cell walls. in *Proceedings of the 6th International Diatom Symposium, Budapest* 253-265.

Crews, D. (1996) Temperature-dependent sex determination: The interplay of steroid hormones and temperature. *Zoological Science* **13**, 1-13.

Davidovich, N.A., Kaczmarska, I. and Ehrman, J.M. (2010) Heterothallic and homothallic sexual reproduction in *Tabularia fasciculata* (Bacillariophyta). *Fottea* **10**, 251-266.

Drebes, G. (1977) Sexuality. in *The biology of diatoms*, Vol. 13 250-283 (Univ of California Press).

Drum, R.W. and Gordon, R. (2003) Star Trek replicators and diatom nanotechnology. *Trends in Biotechnology* **21**, 325-328.

Edlund, M.B. and Stoermer, E.F. (1997) Evolutionary, and systematic significance of diatom life histories. *Journal of Phycology* **33**, 897-918.

Erickson, J.W. and Quintero, J.J. (2007) Indirect effects of ploidy suggest X chromosome dose, not the X : A ratio, signals sex in Drosophila. *PLoS Biology* **5**, 2821-2830.

Fabris, M., Matthijs, M., Rombauts, S., Vyverman, W., Goossens, A. and Baart, G.J.E. (2012) The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *Plant Journal* **70**, 1004-1014.

Fabris, M., Matthijs, M., Carbonelle, S., Moses, T., Pollier, J., Dasseville, R., Baart, G.J.E., Vyverman, W. and Goossens, A. (2014) Tracking the sterol biosynthesis pathway of the diatom Phaeodactylum tricornutum. *New Phytologist* **204**, 521-535.

Fabritius, A.L. and Judelson, H.S. (1997) Mating type loci segregate aberrantly in Phytophthora infestans but normally in Phytophthora parasitica: implications for models of mating-type determination. *Current Genetics* **32**, 60-65.

Falkowski, P. and Knoll, A.H. (2011) *Evolution of primary producers in the sea*, (Academic Press).

Falkowski, P.G. and Raven, J.A. (1997). Aquatic Photosynthesis. (Blackwell Science: Malden).

Foster, J.W., Brennan, F.E., Hampikian, G.K., Goodfellow, P.N., Sinclair, A.H., Lovellbadge, R., Selwood, L., Renfree, M.B., Cooper, D.W. and Graves, J.A.M. (1992) Evolution of Sex Determination and the Y-Chromosome - Sry-Related Sequences in Marsupials. *Nature* **359**, 531-533.

Fraser, J.A. and Heitman, J. (2004) Evolution of fungal sex chromosomes. *Molecular Microbiology* **51**, 299-306.

Fuhrmann, T., Landwehr, S., El Rharbi-Kucki, M. and Sumper, M. (2004) Diatoms as living photonic crystals. *Applied Physics B-Lasers and Optics* **78**, 257-260.

Gastineau, R., Pouvreau, J.B., Hellio, C., Morancais, M., Fleurence, J., Gaudin, P., Bourgougnon, N. and Mouget, J.L. (2012) Biological Activities of Purified Marennine, the Blue Pigment Responsible for the Greening of Oysters. *Journal of Agricultural and Food Chemistry* **60**, 3599-3605.

Gelbart, M.E. and Kuroda, M.I. (2009) Drosophila dosage compensation: a complex voyage to the X chromosome. *Development* **136**, 1399-1410.

Gilbert, S.F. (2000) Chromosomal sex determination in Drosophila. in *Developmental Biology. 6th edition.* (ed. Gilbert, S.F.), (Sunderland).

Gillard, J., Devos, V., Huysman, M.J.J., De Veylder, L., D'Hondt, S., Martens, C., Vanormelingen, P., Vannerum, K., Sabbe, K., Chepurnov, V.A., Inzé, D., Vuylsteke, M. and Vyverman, W. (2008) Physiological and transcriptomic evidence for a close coupling between chloroplast ontogeny and cell cycle progression in the pennate diatom *Seminavis robusta*. *Plant Physiology* **148**, 1394-1411.

Gillard, J., Frenkel, J., Devos, V., Sabbe, K., Paul, C., Rempt, M., Inze, D., Pohnert, G., Vuylsteke, M. and Vyverman, W. (2013) Metabolomics Enables the Structure Elucidation of a Diatom Sex Pheromone. *Angewandte Chemie-International Edition* **52**, 854-857.

Gilmore, S., Weston, P. and Thomson, J. (1993) A simple, rapid, inexpensive and widely applicable technique for purifying plant DNA. *Australian Systematic Botany* **6**, 139-148.

Granum, E., Raven, J.A. and Leegood, R.C. (2005) How do marine diatoms fix 10 billion tonnes of inorganic carbon per year? *Canadian Journal of Botany-Revue Canadienne De Botanique* **83**, 898-908.

Graves, J.A.M. (2006) Sex chromosome specialization and degeneration in mammals. *Cell* **124**, 901-914.

Graves, J.A.M. (2014) Avian sex, sex chromosomes, and dosage compensation in the age of genomics. *Chromosome Research* **22**, 45-57.

Guil-Guerrero, J.L., Navarro-Juarez, R., Lopez-Martinez, J.C., Campra-Madrid, P. and Rebolloso-Fuentes, M.M. (2004) Functional properties of the biomass of three microalgal species. *Journal of Food Engineering* **65**, 511-517.

Herskowitz, I., Rine, J. and Strathern, J. (1992) Mating-type Determination and Mating-type Interconversion in Saccharomyces cerevisiae. *Cold Spring Harbor Monograph Archive* **21**, 583-656.

Hildebrand, M., Frigeri, L.G. and Davis, A.K. (2007) Synchronized growth of Thalassiosira pseudonana (Bacillariophyceae) provides novel insights into cell-wall synthesis processes in relation to the cell cycle. *Journal of Phycology* **43**, 730-740.

Huysman, M.J.J., Martens, C., Vandepoele, K., Gillard, J., Rayko, E., Heijde, M., Bowler, C., Inzé, D., Van de Peer, Y., De Veylder, L. and Vyverman, W. (2010) Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome Biology* **11**, R17.

Huysman, M.J.J., Fortunato, A.E., Matthijs, M., Costa, B.S., Vanderhaeghen, R., Van den Daele, H., Sachse, M., Inze, D., Bowler, C., Kroth, P.G., Wilhelm, C., Falciatore, A., Vyverman, W. and De Veylder, L. (2013) AUREOCHROME1a-Mediated Induction of the Diatom-Specific Cyclin dsCYC2 Controls the Onset of Cell Division in Diatoms (Phaeodactylum tricornutum). *Plant Cell* **25**, 215-228.

Huysman, M.J.J., Vyverman, W. and De Veylder, L. (2014) Molecular regulation of the diatom cell cycle. *Journal of Experimental Botany* **65**, 2573-2584.

Idnurm, A., Walton, F.J., Floyd, A. and Heitman, J. (2008) Identification of the sex genes in an early diverged fungus. *Nature* **451**, 193-196.

Itoh, Y., Melamed, E., Yang, X., Kampf, K., Wang, S., Yehya, N., Van Nas, A., Replogle, K., Band, M.R. and Clayton, D.F. (2007) Dosage compensation is less effective in birds than in mammals. *Journal of Biology* **6**, 2.

Johansen, J.R. (1991) Morphological Variability and Cell-Wall Composition of Phaeodactylum-Tricornutum (Bacillariophyceae). *Great Basin Naturalist* **51**, 310-315.

Judelson, H.S., Spielman, L.J. and Shattock, R.C. (1995) Genetic-Mapping and Non-Mendelian Segregation of Mating-Type Loci in the Oomycete, Phytophthora-Infestans. *Genetics* **141**, 503-512.

Judelson, H.S. (1996) Genetic and physical variability at the mating type locus of the oomycete, Phytophthora infestans. *Genetics* **144**, 1005-1013.

Jursnich, V.A. and Burtis, K.C. (1993) A Positive Role in Differentiation for the Male Doublesex Protein of Drosophila. *Developmental Biology* **155**, 235-249.

Kaczmarska, I., Poulickova, A., Sato, S., Edlund, M.B., Idei, M., Watanabe, T. and Mann, D.G. (2013) Proposals for a terminology for diatom sexual reproduction, auxospores and resting stages. *Diatom Research* **28**, 263-294.

Kamiya, T., Kai, W., Tasumi, S., Oka, A., Matsunaga, T., Mizuno, N., Fujita, M., Suetake, H., Suzuki, S., Hosoya, S., Tohari, S., Brenner, S., Miyadai, T., Venkatesh, B., Suzuki, Y. and Kikuchi, K. (2012) A Trans-Species Missense SNP in Amhr2 Is Associated with Sex Determination in the Tiger Pufferfish, Takifugu rubripes (Fugu). *PLoS Genetics* **8**.

Koerich, L.B., Wang, X.Y., Clark, A.G. and Carvalho, A.B. (2008) Low conservation of gene content in the Drosophila Y chromosome. *Nature* **456**, 949-951.

Kroth, P. (2007) Molecular biology and the biotechnological potential of diatoms. *Transgenic Microalgae as Green Cell Factories* **616**, 23-33.

Lebeau, T. and Robert, J.M. (2003) Diatom cultivation and biotechnologically relevant products. Part II: Current and putative products. *Applied Microbiology and Biotechnology* **60**, 624-632.

Lenski, R.E. (2001) Genetics and evolution - Come fly, and leave the baggage behind. *Science* **294**, 533-534.

Lepetit, B., Sturm, S., Rogato, A., Gruber, A., Sachse, M., Falciatore, A., Kroth, P.G. and Lavaud, J. (2013) High Light Acclimation in the Secondary Plastids Containing Diatom Phaeodactylum tricornutum is Triggered by the Redox State of the Plastoquinone Pool. *Plant Physiology* **161**, 853-865.

Lewis, W.M. (1984) The Diatom Sex Clock and Its Evolutionary Significance. *American Naturalist* **123**, 73-80.

Liew, W.C., Bartfai, R., Lim, Z.J., Sreenivasan, R., Siegfried, K.R. and Orban, L. (2012) Polygenic Sex Determination System in Zebrafish. *PLoS ONE* **7**.

Lints, R. and Emmons, S.W. (2002) Regulation of sex-specific differentiation and mating behavior in C-elegans by a new member of the DM domain transcription factor family. *Genes & Development* **16**, 2390-2402.

Lopez-Elias, J.A., Voltolina, D., Enriquez-Ocana, F. and Gallegos-Simental, G. (2005) Indoor and outdoor mass production of the diatom Chaetoceros muelleri in a Mexican commercial hatchery. *Aquacultural Engineering* **33**, 181-191.

Lopez, P.J., Descles, J., Allen, A.E. and Bowler, C. (2005) Prospects in diatom research. *Current Opinion in Biotechnology* **16**, 180-186.

Lovellbadge, R. (1992) The Role of Sry in Mammalian Sex Determination. *Ciba Foundation Symposia* **165**, 162-182.

Mann, D.G. (1984). An ontogenetic approach to diatom systematics. in *Proceedings of the 7th International Diatom Symposium* Vol. 113 (O. Koeltz Koenigstein).

Mann, D.G. and Droop, S.J.M. (1996) Biodiversity, biogeography and conservation of diatoms. *Hydrobiologia* **336**, 19-32.

Mann, D.G., Chepurnov, V.A. and Idei, M. (2003) Mating system, sexual reproduction, and auxosporulation in the anomalous raphid diatom Eunotia (Bacillariophyta). *Journal of Phycology* **39**, 1067-1084.

Matson, C.K. and Zarkower, D. (2012) Sex and the singular DM domain: insights into sexual regulation, evolution and plasticity. *Nature Reviews Genetics* **13**, 163-174.

Matsuda, M., Nagahama, Y., Shinomiya, A., Sato, T., Matsuda, C., Kobayashi, T., Morrey, C.E., Shibata, N., Asakawa, S., Shimizu, N., Hori, H., Hamaguchi, S. and Sakaizumi, M. (2002) DMY is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature* **417**, 559-563.

Matsuda, M., Shinomiya, A., Kinoshita, M., Suzuki, A., Kobayashi, T., Paul-Prasanth, B., Lau, E.L., Hamaguchi, S., Sakaizumi, M. and Nagahama, Y. (2007) DMY gene induces male development in genetically female (XX) medaka fish. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 3865-3870.

Mccoy, J.W. (1972) Kinetic Studies on Mating Reaction of Tetrahymena-Pyriformis, Syngen-1. *Journal of Experimental Zoology* **180**, 271-277.

Medina, A.R., Grima, E.M., Gimenez, A.G. and Gonzalez, M.J.I. (1998) Downstream processing of algal polyunsaturated fatty acids. *Biotechnology Advances* **16**, 517-580.

Miller, M.R., Quek, S.Y., Staehler, K., Nalder, T. and Packer, M.A. (2014) Changes in oil content, lipid class and fatty acid composition of the microalga Chaetoceros calcitrans over different phases of batch culture. *Aquaculture Research* **45**, 1634-1647.

Nanney, D.L. and Caughey, P.A. (1953) Mating type determination in Tetrahymena pyriformis. *Proceedings of the National Academy of Sciences of the United States of America* **39**, 1057-1063.

Navarro-Martin, L., Vinas, J., Ribas, L., Diaz, N., Gutierrez, A., Di Croce, L. and Piferrer, F. (2011) DNA Methylation of the Gonadal Aromatase (cyp19a) Promoter Is Involved in Temperature-Dependent Sex Ratio Shifts in the European Sea Bass. *PLoS Genetics* **7**.

Nguyen, D.K. and Disteche, C.M. (2006) Dosage compensation of the active X chromosome in mammals. *Nature Genetics* **38**, 47-53.

Nymark, M., Valle, K.C., Brembu, T., Hancke, K., Winge, P., Andresen, K., Johnsen, G. and Bones, A.M. (2009) An integrated analysis of molecular acclimation to high light in the marine diatom *Phaeodactylum tricornutum*. *PLoS ONE* **4**.

Ospina-Alvarez, N. and Piferrer, F. (2008) Temperature-Dependent Sex Determination in Fish Revisited: Prevalence, a Single Sex Ratio Response Pattern, and Possible Effects of Climate Change. *PLoS ONE* **3**.

Panacek, A., Balzerova, A., Prucek, R., Ranc, V., Vecerova, R., Husickova, V., Pechousek, J., Filip, J., Zboril, R. and Kvitek, L. (2013) Preparation, characterization and antimicrobial efficiency of Ag/PDDA-diatomite nanocomposite. *Colloids and Surfaces B-Biointerfaces* **110**, 191-198.

Parfrey, L.W., Barbero, E., Lasser, E., Dunthorn, M., Bhattacharya, D., Patterson, D.J. and Katz, L.A. (2006) Evaluating support for the current classification of eukaryotic diversity. *PLoS Genetics* **2**, 2062-2073.

Poulsen, N., Chesley, P.M. and Kröger, N. (2006) Molecular genetic manipulation of the diatom *Thalassiosira pseudonana* (Bacillariophyceae). *Journal of Phycology* **42**, 1059-1065.

Poulsen, N., Scheffel, A., Sheppard, V.C., Chesley, P.M. and Kroger, N. (2013) Pentalysine Clusters Mediate Silica Targeting of Silaffins in Thalassiosira pseudonana. *Journal of Biological Chemistry* **288**, 20100-20109.

Reyes-Prieto, A., Weber, A.P.M. and Bhattacharya, D. (2007) The origin and establishment of the plastid in algae and plants. *Annual Review of Genetics* **41**, 147-168.

Rhind, N.R., Miller, L.M., Kopczynski, J.B. and Meyer, B.J. (1995) Xol-1 Acts as an Early Switch in the C-Elegans Male Hermaphrodite Decision. *Cell* **80**, 71-82.

Rogers, M.B., Gilson, P.R., Su, V., McFadden, G.I. and Keeling, P.J. (2007) The complete chloroplast genome of the chlorarachniophyte Bigelowiella natans: Evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Molecular Biology and Evolution* **24**, 54-62.

Round, F.E., Crawford, R.M. and Mann, D.G. (1990) *The diatoms: biology & morphology of the genera*, (Cambridge University Press).

Ryner, L.C. and Baker, B.S. (1991) Regulation of doublesex pre-mRNA processing occurs by 3'-splice site activation. *Genes & Development* **5**, 2071-2085.

Saade, A. and Bowler, C. (2009) Molecular tools for discovering the secrets of diatoms. *Bioscience* **59**, 757-765.

Salz, H.K., Maine, E.M., Keyes, L.N., Samuels, M.E., Cline, T.W. and Schedl, P. (1989) The Drosophila Female-Specific Sex-Determination Gene, Sex-Lethal, Has Stage-Specific, Tissue-Specific, and Sex-Specific Rnas Suggesting Multiple-Modes of Regulation. *Genes & Development* **3**, 708-719.

Sanchez-Puerta, M.V. and Delwiche, C.F. (2008) A hypothesis for plastid evolution in chromalveolates. *Journal of Phycology* **44**, 1097-1107.

Sandhage, K.H., Dickerson, M.B., Huseman, P.M., Caranna, M.A., Clifton, J.D., Bull, T.A., Heibel, T.J., Overton, W.R. and Schoenwaelder, M.E.A. (2002) Novel, bioclastic route to self-assembled, 3D, chemically tailored meso/nanostructures: Shape-preserving reactive conversion of biosilica (diatom) microshells. *Advanced Materials* **14**, 429-433.

Sato, S., Beakes, G., Idei, M., Nagumo, T. and Mann, D.G. (2011) Novel Sex Cells and Evidence for Sex Pheromones in Diatoms. *PLoS ONE* **6**.

Ser, J.R., Roberts, R.B. and Kocher, T.D. (2010) Multiple Interacting Loci Control Sex Determination in Lake Malawi Cichlid Fish. *Evolution* **64**, 486-501.

Shao, C.W., Li, Q.Y., Chen, S.L., Zhang, P., Lian, J.M., Hu, Q.M., Sun, B., Jin, L.J., Liu, S.S., Wang, Z.J., Zhao, H.M., Jin, Z.H., Liang, Z., Li, Y.Z., Zheng, Q.M., Zhang, Y., Wang, J. and Zhang, G.J. (2014) Epigenetic modification and inheritance in sexual reversal of fish. *Genome Research* **24**, 604-615.

Shukla, J.N. and Nagaraju, J. (2010) Doublesex: a conserved downstream gene controlled by diverse upstream regulators. *Journal of Genetics* **89**, 341-356.

Smith, C.A., McClive, P.J., Western, P.S., Reed, K.J. and Sinclair, A.H. (1999) Evolution - Conservation of a sex-determining gene. *Nature* **402**, 601-602.

Smith, C.A., Roeszler, K.N., Ohnesorg, T., Cummins, D.M., Farlie, P.G., Doran, T.J. and Sinclair, A.H. (2009) The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature* **461**, 267-271.

Sorhannus, U. (2007) A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. *Marine Micropaleontology* **65**, 1-12.

Spolaore, P., Joannis-Cassan, C., Duran, E. and Isambert, A. (2006) Commercial applications of microalgae. *Journal of Bioscience and Bioengineering* **101**, 87-96.

Stiglec, R., Ezaz, T. and Graves, J.A.M. (2007) A new look at the evolution of avian sex chromosomes. *Cytogenetic and Genome Research* **117**, 103-109.

Sumper, M. and Brunner, E. (2008) Silica biomineralisation in diatoms: The model organism Thalassiosira pseudonana. *ChemBioChem* **9**, 1187-1194.

Treguer, P.J. and De La Rocha, C.L. (2013) The World Ocean Silica Cycle. *Annual Review of Marine Science, Vol 5* **5**, 477-501.

Turpin, V., Robert, J.M., Goulletquer, P., Masse, G. and Rosa, P. (2001) Oyster greening by outdoor mass culture of the diatom Haslea ostrearia Simonsen in enriched seawater. *Aquaculture Research* **32**, 801-809.

Umen, J.G. (2013) Genetics: Swinging Ciliates' Seven Sexes. *Current Biology* **23**, R475-R477.

Vandeputte, M., Dupont-Nivet, M., Chavanne, H. and Chatain, B. (2007) A polygenic hypothesis for sex determination in the European sea bass - Dicentrarchus labrax. *Genetics* **176**, 1049-1057.

Vanstechelman, I., Sabbe, K., Vyverman, W., Vanormelingen, P. and Vuylsteke, M. (2013) Linkage Mapping Identifies the Sex Determining Region as a Single Locus in the Pennate Diatom Seminavis robusta. *PLoS ONE* **8**.

Viji, S., Anbazhagi, M., Ponpandian, N., Mangalaraj, D., Jeyanthi, S., Santhanam, P., Devi, A.S. and Viswanathan, C. (2014) Diatom-Based Label-Free Optical Biosensor for Biomolecules. *Applied Biochemistry and Biotechnology* **174**, 1166-1173.

Wilhelm, C., Jungandreas, A., Jakob, T. and Goss, R. (2014) Light acclimation in diatoms: From phenomenology to mechanisms. *Marine Genomics* **16**, 5-15.

Wolf, J.B.W. and Bryk, J. (2011) General lack of global dosage compensation in ZZ/ZW systems? Broadening the perspective with RNA-seq. *BMC Genomics* **12**, 91.

Yano, A., Guyomard, R., Nicol, B., Jouanno, E., Quillet, E., Klopp, C., Cabau, C., Bouchez, O., Fostier, A. and Guiguen, Y. (2012) An Immune-Related Gene Evolved into the Master Sex-Determining Gene in Rainbow Trout, Oncorhynchus mykiss. *Current Biology* **22**, 1423-1428.

Yoshimoto, S., Okada, E., Umemoto, H., Tamura, K., Uno, Y., Nishida-Umehara, C., Matsuda, Y., Takamatsu, N., Shiba, T. and Ito, M. (2008) A W-linked DM-domain gene, DM-W, participates in primary ovary development in Xenopus laevis. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 2469-2474.

Yu, Q.Y., Hou, S., Feltus, F.A., Jones, M.R., Murray, J.E., Veatch, O., Lemke, C., Saw, J.H., Moore, R.C., Thimmapuram, J., Liu, L., Moore, P.H., Alam, M., Jiang,

J.M., Paterson, A.H. and Ming, R. (2008) Low X/Y divergence in four pairs of papaya sex-linked genes. *Plant Journal* **53**, 124-132.

Zarkower, D. (2006) Somatic sex determination. in *WorkBook* (ed. Community, T.C.e.R.).

2

A member of the DNA methyltransferase 5 family lies within the mating type locus of the pennate diatom *Seminavis robusta*

Sara Moeys[1,2,3#], Ives Vanstechelman[1,2,3#], Marie J.J. Huysman[2,3], Wim Vyverman[1], Lieven De Veylder[2,3], Tore Brembu[4], Per Winge[4], Atle Bones[4], Koen Sabbe[1*] and Marnik Vuylsteke[2,3,5*]

Sara Moeys and Ives Vanstechelman share first authorship

* Koen Sabbe and Marnik Vuylsteke share senior authorship

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

[2] Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Gent, Belgium

[3] Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, B-9052 Gent, Belgium

[4] NTNU Cell And Molecular Biology Group, Department of Biology, NTNU Realfagbygget, N-7491 Trondheim, Norway

[5] Lab Aquaculture & Artemia Reference Center, Faculty Bioscience Engineering, Gent University, B-9000 Gent, Belgium

Manuscript in preparation

Authors' contributions

SM and IV wrote the manuscript. IV performed the BSA experiments and analyzed the data. SM performed the Sanger sequencing and analyzed the data. MJJH conducted the phylogenic analyses. TB, PW and AB were involved in sequencing the *S. robusta* genome. KS, MV, WV and LDV helped to conceive and design the study, and read and approved the manuscript. MV helped interpreting and processing the data.

Abstract

Linkage analysis recently showed that the mating type of the pennate diatom *Seminavis robusta* is determined by a single locus, with the mating type plus as the heterogametic mating type. In the present study, bulked segregant analysis identified two genetic markers being strongly linked with the mating type locus. Both markers mapped on the same ~25kb scaffold of the *S. robusta* draft genome, containing five protein-coding genes, including a gene belonging to the DNMT5 family of methyltransferases. In addition, SNPs residing in this gene are in full linkage disequilibrium with the mating type segregating in a pedigree of *Seminavis robusta*. This strongly suggests that DNMT5a is responsible for mating type determination in *S. robusta.* DNMT5a is the first identified mating type determinant in diatoms.

Introduction

Diatoms (Bacillariophyceae), which belong to the Stramenopila lineage, are a highly diverse and productive group of algae (Granum et al., 2005). The total number of diatom species is estimated to be ~200,000, which together are responsible for ~20% of global primary production (Mann, 1999). They are increasingly used in biotechnological applications as they produce high-value bioproducts, such as polyunsaturated fatty acids, pigments and lipids for biofuel (Bozarth et al., 2009, Levitan et al., 2014). The life cycle of diatoms is diplontic, meaning that all life cycle stages except the gametes are diploid, and is accompanied by a typical cell-size-reduction-restitution cycle (Chepurnov et al., 2004). During vegetative growth, cell size reduces, ultimately leading to clonal cell death. This can be avoided by switching to sexual reproduction, which is only possible below a species-specific size threshold (SST) (Chepurnov et al., 2002, Chepurnov et al., 2004). During the sexual phase, a specialized zygote is formed that is able to expand to the initial large cell size. Sexual reproduction has never been demonstrated for the most commonly used model diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* (Chepurnov et al., 2008). This prevents the use of forward genetic studies for linking phenotype to genotype, including the use of mutagenesis and QTL mapping and thus also the elucidation of the molecular basis of mating type (MT) determination. The pennate diatom *Seminavis robusta* has been put forward as a model organism to study the pennate life cycle, as it displays a typical diatom life cycle and sexual reproduction can be reliably controlled (Chepurnov et al., 2008).

Sexual development is common in eukaryotic organisms. Sex can be genetically encoded on a specialized sex chromosome or on an autosomal chromosome, or it can be determined by environmental factors, like temperature, local sex ratio or population density. Though little is known about the molecular mechanisms behind environmental sex determination, many genetic sex determining systems have been uncovered (Haag and Doty, 2005). Two homologous sex chromosomes can be either heteromorphic, where one of the homologous chromosomes is smaller (genetic degeneration), or homomorphic (no degeneration in the chromosomes). Classical

sex chromosomes (e.g. in mammals and birds) have small recombining pseudo-autosomal regions (PAR) and large non-recombining regions. It is assumed that this represents an ancient, highly evolved type of sex chromosomes (Bergero and Charlesworth, 2009). On the other hand, in recently evolved sex chromosomes non-recombining regions can be as small as a few megabases and large PAR regions exist (Bergero and Charlesworth, 2009).

Separate sexes or mating types appear to have arisen independently many times during the evolution of major eukaryotic lineages. This is reflected in the variation that exists in both the primary sex determination signal and in the downstream genetic pathways that relay the signal (Haag and Doty, 2005, Bergero and Charlesworth, 2009). However, it appears that across eukaryotic domains, transcription factors are often key determinants regulating sexual development, for example Sry in mammals, Dmrt1 in birds or the HMG-type transcription factor in Zygomycetes (Smith et al., 2009, Koestler and Ebersberger, 2011, Trukhina et al., 2013). On the other hand, in tiger pufferfish *Takifugu rubripes* sex is determined by the hormone receptor Amhr2, and in medaka fish *Oryzias luzonensis* and Patagonian pejerrey *Odontesthes hatcheri* growth factors, Gsdf and Amhy respectively, are the primary sex determinants (Hattori et al., 2012, Kamiya et al., 2012, Myosho et al., 2012).

To date, no MT-determining genes have been identified in the Stramenopila lineage. Only in brown algae, a possible MT-determinant is identified. The brown alga *Ectocarpus* uses the UV system to determine sex, meaning that sex is expressed during the haploid phase of the life cycle and that both the U (female) and V (male) chromosomes contain non-recombining regions (Ahmed et al., 2014). Ahmed et al. (2014) hypothesizes that an HMG gene could be the sex determining gene in *Ectocarpus* sp., since it is only present on the male V chromosome and this gene family is known to be involved in sex or mating type determination in vertebrates and fungi. A recent genome-wide linkage map of *S. robusta* has shown that the MT phenotype segregates as a single locus, flanked by large autosomal-like regions, and has revealed that MT^+ is the heterogametic MT (Vanstechelman et al., 2013). In *S.*

robusta, MTs are defined based on the cell's behavior during mating. MT^- cells are immotile and produce the pheromone diproline, while MT^+ cells are motile and migrate towards MT^- (Gillard et al., 2013).

In the present study, we further investigated the MT-determining region of the pennate diatom *S. robusta*. A bulked segregant analysis (BSA) in combination with AFLP (amplified fragment length polymorphism) and whole genome sequencing (WGS) identified a ~25 kb scaffold of an in-house draft genome, containing five predicted protein-coding genes, including a gene belonging to the DNA methyltransferase 5 (DNMT5) family. By linkage and association mapping, we show that this gene is most likely responsible for MT determination in *S. robusta*.

Materials & methods

Strains and culture conditions

Cell cultures were grown in F/2 medium (Guillard, 1975) made with filtered (GF/C grade microfiber filter; Whatman) autoclaved seawater collected from the North Sea. The cultures were cultivated at 18°C with a 12:12h light:dark period and approximately 85 µmol photons m^{-2} s^{-1} from cool-white fluorescent lights.

All used strains are publicly available in the BCCM/DCG diatom collection (http://bccm.belspo.be) (Chepurnov et al., 2008). For bulked segregant analysis, 25 MT^+ and 25 MT^- progeny from the F_1 mapping population (MP) previously used to build MT-specific linkage maps were used (Vanstechelman et al., 2013). For Sanger sequencing of *DNMT5a*, strains 96A (MT^+), H73A (MT^-), 85A (MT^+), 85B (MT^-), 109A (MT^-), 109B (MT^+), 114A (MT^+) and 114B (MT^-) were used.

Bulked segregant analysis (BSA) using AFLP and Whole Genome Sequencing

For Bulked Segregant Analysis (BSA) using AFLP, genomic DNA was isolated from individual cultures from the F_1-MP sampled in exponential phase as described by Vanstechelman et al. (2013), followed by selective PCR amplification of restriction fragments of a total digest of the genomic DNA obtained with the restriction enzyme

combination *Eco*RI/*Mse*I (Vuylsteke et al., 2007b). Equal quantities of the AFLP pre-amplification products of five MT^+ or MT^- cultures were pooled to form four bulks for each MT. Bulks were screened for presence/absence using 472 *Eco*RI+2/*Mse*I+3 AFLP primer combinations (PCs). Detection of the AFLP fragments was made possible by fluorescent labeling of the *Eco*RI+2 primer in the final selective amplification reaction, and AFLP images were generated using LI-COR automated DNA sequencers. AFLP markers polymorphic between MT^- and MT^+ bulks were further analyzed at the level of all strains of the F_1-MP. Markers completely co-segregating with the MT locus were integrated in the AFLP-based MT-specific linkage maps (Vanstechelman et al., 2013) using the linkage analysis software Joinmap 4.0 (Van Ooijen, 2006), and purified from sequencing gels followed by Sanger sequencing as described by Vuylsteke et al. (2007a).

For BSA using whole genome sequencing (WGS-BSA), equal amounts of cell material of 25 strains of the F_1-MP were pooled for each MT. Genomic DNA was isolated in a similar way as done for the AFLP analysis and single Illumina DNA sequencing libraries were prepared from each bulk. The libraries were prepared from 1 µg DNA using "Illumina TruSeq DNA Sample Preparation Kit" following the standard protocol. DNA fragmentation, end-repair, A-tailing and adapter ligation was followed by PCR amplification. The resulting libraries were quantified by Qubit (Life Technologies) and verified on the Bioanalyzer (Agilent). The libraries were analyzed in a 2x100 bp run on an Illumina HiSeq 2000 instrument.

VarScan (Koboldt et al., 2012) was used to detect SNP loci that reflect heterozygosity in the MT^+ bulk and homozygosity in the MT^- bulk. The Illumina PE reads of the two bulks differing for MT were mapped onto an in-house draft genome using the software Burrows Wheeler Aligner (BWA) (Li and Durbin, 2010) under default conditions. After mapping, vcf (variant call format) files for both MT bulks were constructed by Samtools (Li et al., 2009) with the samtools “mpileup option”. Only the reads mapping uniquely were selected for the construction of the vcf files (option -q=1). Next, the Varscan “somatic option” was ran to screen for allele frequency differences between the two MT bulks (minimal coverage set to 8 for the

2 samples), and their significances were assessed by the Fisher's Exact Test. Scaffolds were ranked according to their accumulation of LOH events (only scaffolds with minimum ten SNPs were considered) using the Fisher's Exact Test.

Annotation of scaffold1897

First, an in-house transcriptome (assembled from JGI projects 402005 and 1006410, http://genome.jgi-psf.org) was mapped to the draft genome using GMAP (Wu and Watanabe, 2005) to identify transcripts mapping to scaffold1897. Based on these transcripts, gene structure predictions were computed with GenomeThreader (Gremme et al., 2005). A Conserved Domains Database (CDD) search (Marchler-Bauer et al., 2011) with the resulting protein predictions against the NCBI CDD v3.11-45746 PSSMs was performed to functionally annotated these proteins.

Phylogenetic analysis of the MT locus

The genomic, transcriptomic and proteomic sequences of the diatoms *Thalassiosira pseudonana*, *Phaeodactylum tricornutum*, *Fragilariopsis cylindricus* and *Pseudo-nitzschia multiseries* are available online through the JGI portal (http://genome.jgi-psf.org/). For *S. robusta* a draft genome and transcriptome assembly was generated in-house.

To retrieve putative DNA methyltransferases (DNMTs), all diatom proteomes were scanned with the Hidden Markov Model (HMM) build based on the seed alignment available in the Pfam database for the C5-cytosine-specific DNA MTase domain (PF00145.12, named DNA_methylase) using HMMsearch. Proteins with a score above the inclusion threshold were retained for further analyses (Suppl. Table S5). Redundant sequences of *S. robusta* were identified by mapping to the draft genome and these sequences were removed from further analyses.

An amino acid sequence alignment was constructed for the diatom DNMT proteins together with reference sequences selected from the DNMT phylogenetic tree of Maumus et al. (2011), representing different DNMT groups. DNMT proteins from *Arabidopsis thaliana*, *Homo sapiens*, *Micromonas pusilla*, *Ostreococcus tauri*

and *Ostreococcus lucimarinus* were selected. Bacterial DNMTs were chosen as outgroup proteins. Based on the multiple sequence alignment generated by MUSCLE (Edgar, 2004) of the MTase domain (DNA_methylase, PF00145.12) of these sequences, a Neighbor-Joining phylogenetic tree was constructed using the MEGA5 software allowing the classification of the diatom DNMTs in the different subfamilies (Saitou and Nei, 1987, Tamura et al., 2011). The Poisson correction method was used for distance calculation and 1000 iterations were applied for bootstrap value calculations (Zuckerkandl and Pauling, 1965, Felsenstein, 1985).

For the DNMT5 subfamily, additional phylogenetic analyses were performed based on the multiple sequence alignments of the MTase (DNA_methylase, PF00145.12) and HEL (SNF2_N, PF00176.18) domain of the diatom DNMT5 sequences and of a selection of fungal (*Aspergillus fumigatus*, *Aspergillus nidulans*, *Coccidioides immitis*, *Coccidioides posadasii*, *Laccaria bicolor*, *Phanerochaete chrysosporium*) and algal (*Aureococcus anophagefferens*, *Bathycoccus prasinos*, *O. tauri*, *O. lucimarinus* and *Micromonas sp.*) DNMT5 sequences retrieved through the OrthoMCL and pico-PLAZA databases (Li et al., 2003, Vandepoele et al., 2013). The multiple sequence alignments were generated with MUSCLE (Edgar, 2004) and non-conserved alignment positions were removed using the Gblocks software (Talavera and Castresana, 2007). Phylogenetic trees were constructed with the Neighbor-Joining (NJ) method as described above.

Sequencing of *DNMT5a*

Twenty eight SNPs in *DNMT5a* for which the MT^+ bulk was heterozygous and the MT^- bulk homozygous were selected from the BSA-WGS data. gDNA was extracted from the two parental strains of the F_1 mapping population (96A and H73A) and six other strains of the *S. robusta* pedigree (MT^+: 85A, 109B, 114A; MT^-: 85B, 109A, 114B) using the Dneasy plant mini kit (Qiagen). About 100 mL of culture (± 3 x 10^6 cells) was scraped from the bottom of the tissue culture flask (Cellstar, 75 cm^2 growth surface, Greiner Bio-One) and filtered on a Versapor filter (3 µm pore size, 25 mm diameter, PALL). Filters were frozen in liquid nitrogen. AP1 buffer (400 mL), 4 µL Rnase A (100

mg/mL) and silicon carbide beads (1.0 mm, BioSpec) were added to the filter. Cells were disrupted by beating on a bead mill (Retsch, 3 x 1 min). All other steps for the DNA extraction were done according to the manufacturer's instructions.

For each strain, five fragments of ~1000 bp, covering the 28 SNP loci selected, were amplified by PCR, followed by sanger sequencing (primers in Suppl. Table S1). Fragments containing insertion/deletions were cloned in pJET1.2 with the CloneJET PCR cloning kit, after which five clones for every fragment were sequenced.

To investigate if the identified amino acid changes are likely to cause structural and/or functional changes, the two alleles were analyzed using Phyre2 (Kelley and Sternberg, 2009) and SusPect (Yates et al., 2014).

Results

Genetic mapping localizes the MT locus to a 24.7 kb genomic scaffold

Initially, AFLP-based BSA was employed to identify MT-linked AFLP markers. Two sets of four MT^+ or MT^- bulks each containing five MT^+ or MT^- strains of the F_1-MP were screened with 472 AFLP PCs. Consistent with a single MT-determining gene model for which MT^+ is heterogametic and MT^- is homogametic (Vanstechelman et al., 2013), we screened for presence of AFLP bands in the four MT^+ bulks and absence in the four MT^- bulks. Eight AFLP fragments were only present in the MT^+ bulks, of which five (E21M442M273.4; E33M132M142.6; E32M231M408.0; E34M221M432.1; E31M321M121.2) strongly co-segregated with the MT locus (LOD ranging from 17.32 to 31.53), which mapped to the previously identified linkage group MT^+_6 (Figure 1) (Vanstechelman et al., 2013). AFLP markers E21M442M273.4 (LOD = 31.53) and E34M221M432.1 (LOD = 26.94), flanking the MT locus on the linkage map, were purified from the gel, sequenced and blasted to the draft genome.

The two marker sequences both mapped on a single scaffold (scaffold1897, length = 24,698 bp) of the draft genome and were separated by only ~8kb, identifying scaffold1897 as most likely carrying the MT locus. Given a total map

distance of ~970 cM (Vanstechelman et al., 2013) and a genome size of 153 Mb (Chepurnov et al., 2008), 1 cM corresponds on average to ~157 kb. As the genetic distance between the two markers is 2.1 cM, the physical to genetic map ratio on scaffold 1897 appears to be strongly distorted (~40-fold). This could be caused by repeats that are not properly assembled or by a higher recombination rate around the MT locus.

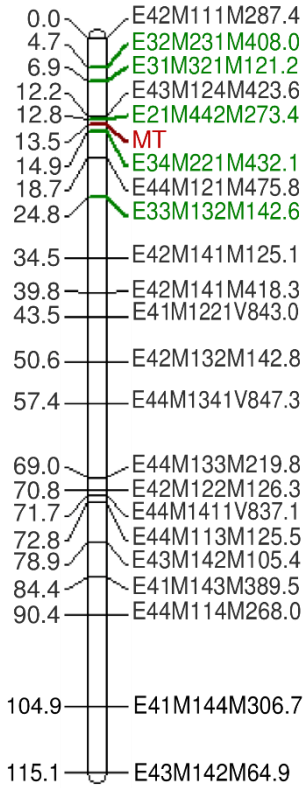


Figure 1: Linkage group MT^+_6 of the *S. robusta* linkage map (AFLP markers in black), identified previously to include the MT locus (red) (Vanstechelman et al., 2013). Five additional AFLP markers (green), identified using AFLP-based BSA, were added.

Scaffold1897 is highly enriched for SNPs that are in full linkage disequilibrium with the MT locus

BSA-WGS generated 85 M and 92 M paired end (PE) reads for the MT^+ and MT^- bulk, respectively. This resulted in an initial coverage of 110X and 120X based on the

genome size of *S. robusta,* which was estimated to be 153 MB using flow cytometry (Chepurnov et al., 2008).

While allele frequencies in the two bulks differing for MT should be approximately equal in genomic regions without loci affecting the MT, they should differ at the single genomic region containing the MT locus. Consistent with a single MT-determinant model in which MT^+ is heterogametic and MT^- is homogametic, we searched for SNP loci for which the MT^+ bulk shows heterozygosity and the MT^- bulk homozygosity using VarScan (Koboldt et al., 2012). The Varscan "somatic option" (Koboldt et al., 2012) was ran, which was initially constructed to detect somatic mutations in tumor samples. It calls loss-of-heterozygosity events (LOH) when the read counts in the two samples are significantly different (P<0.05) and heterozygosity is present in the first sample and is (nearly) absent in the second sample. Scaffolds were then ranked according to the percentages of SNPs showing the target allele frequency difference. Scaffold1897, earlier identified as carrying the two AFLP markers flanking the MT locus, was in the top 0.1% (ranked in the top 35 (p=0.0) out of 72,034 scaffolds) of scaffolds with the highest relative number of SNPs showing the target allele frequency difference (Suppl. Table S2).

Out of the 405 SNPs residing on scaffold1897, 107 (26.4%) were called as an LOH event (Suppl. Table S3), meaning that the allele frequency differs significantly between the two samples and that heterozygosity is present in the first sample (MT^+ bulk) and absent in the second sample (MT^- bulk), thus indicating that these SNPs are in full linkage disequilibrium with the MT locus. Only 3.1% of the SNPs are called as LOH events on a genome-wide scale. Scaffold1897 thus appears to be highly enriched with LOH events (Chi-square, p<0.001), suggesting that this scaffold resides in the MT-determining genomic region.

Annotation of scaffold1897 identified a gene containing a SNF2-type helicase and a C5-specific methyltransferase domain

A gene structure prediction based on transcriptomic data was carried out on scaffold1897. Nineteen different transcripts from the *de novo* assembled

transcriptome mapped on this scaffold. Gene structure prediction identified eleven genes and for five of these genes a protein sequence could be predicted (Figure 2). A CDD search resulted in a functional annotation for these five protein-encoding genes: a serine/threonine protein kinase, a hint-domain containing gene, a gene containing a SNF2-type helicase and a C5-specific methyltransferase domain, a leucine-rich-repeat containing gene and one gene with unknown function. An overview of the gene prediction and the functional annotation of scaffold1897 is given in Figure 2 and supplementary table S4.

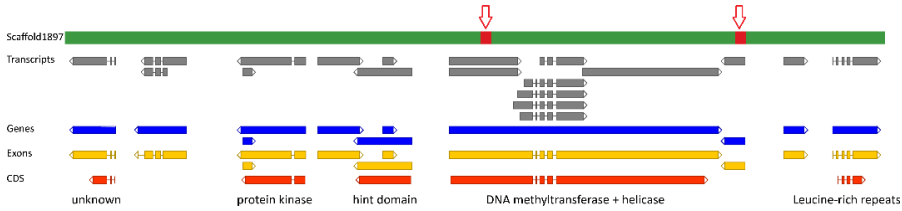


Figure 2: A schematic overview of the gene prediction for scaffold1897. From top to bottom: scaffold1897 (markers flanking the MT locus are indicated with arrows), the mapped transcripts, the predicted genes, exons, coding regions (CDS) and the functions deduced from the CDD search results.

Both MT-flanking AFLP markers mapped on a predicted gene. Marker E34M221M432.1 lies in the gene harboring a SNF2-type helicase and a C5-specific methyltransferase domain and marker E21M442M273.4 lies in a gene for which no protein sequence could be predicted.

As none of the SNPs found in this second gene was called as a LOH event in the BSA-WGS analysis, the gene harboring a SNF2-type helicase and a C5-specific methyltransferase domain is the best candidate for being the MT-determining gene. The combination of these two domains is typically found in the *DNMT5* gene family (Ponger and Li, 2005).

Identification and phylogenetic analysis of diatom DNMTs

DNA methyltransferases (DNMT) sequences were identified in the diatom genomes using a HMMsearch. To this end a HMM profile was build based on the

seed alignment available for the DNA_Methylase domain PF00145.12 in the Pfam database. The results of this HMMsearch are summarized in supplementary table S5. Eukaryotic DNMTs can, on the basis of sequence similarity, be grouped in six functional groups, DNMT1-DNMT6 (Ponger and Li, 2005). The combination of a DNA_methylase and a SNF2-type helicase domain has only been described for the DNMT5 protein family, which has one member in the diatoms *P. tricornutum* and *T. pseudonana* (Ponger and Li, 2005, Maumus et al., 2011).

To determine the sub-classification of all diatom DNMT sequences, a phylogenetic analysis was performed based on a multiple sequence alignment of their MTase domain (Figure 3). Based on this phylogenetic analysis eight diatom sequences could be assigned to the DNMT5 family. Inspection of the domain organization of these sequences confirmed the presence of a N-terminal DNA_methylase and a C-terminal SNF2-type helicase domain in each of these sequences (Figure 4A). To investigate the evolutionary relationship between these sequences, phylogenetic trees were constructed based on multiple sequence alignments of the DNA_methylase (Figure 4B) and SNF2-type helicase (Figure 4C) domain. The *P. tricornutum* DNMT5 sequence and the *F. cylindrus* DNMT5c sequence are the least related to the other diatom DNMT5 proteins as they clearly cluster in another clade together with the DNMT5 proteins of the other algal and fungal representatives. There seems to be no relationship between the phylogeny of the DNMT5 proteins and known evolutionary relationships between the main diatom groups and within the pennate clade (cf. recent phylogenetic studies based on the nuclear-encoded subunit of the rDNA gene (SSU) or on chloroplast data (*rbc*L plus *psb*C) (Chepurnov et al., 2008, Theriot et al., 2010, Medlin, 2011)). No important differences in topology of the trees can be observed between the two analyzed domains.

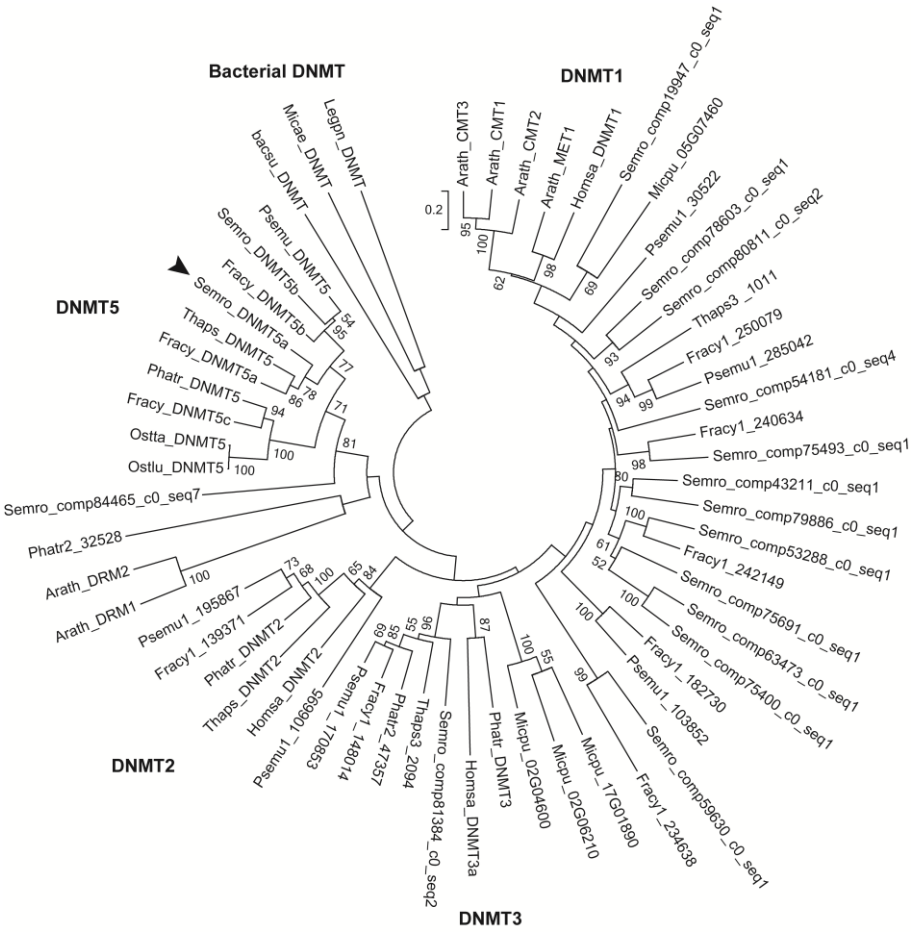


Figure 3: Phylogenetic reconstruction of diatom DNMTs. A phylogenetic tree was generated using the Neighbor-Joining method (Poisson correction, 1000 bootstrap iterations) based on a multiple sequence alignment (generated with MUSCLE) of the DNA_Methylase (PF00145.12) domains of selected DNMT families. Bacterial DNMTs (Legpn, Micae and Bacsu) were selected as outgroup. Legpn, *Legionella pneumophila*; Micae, *Microcystis aeruginosa*; Bacsu, *Bacillus subtilis*; Semro, *Seminavis robusta*; Phatr, *Phaeodactylum tricornutum*; Thaps, *Thalassiosira pseudonana*; Psemu, *Pseudo-nitzschia multiseries*; Fracy, *Fragilariopsis cylindrus*; Homsa, *Homo sapiens*; Arath, *Arabidopsis thaliana*; Micpu, *Micromonas pusilla*; Ostta, *Ostreococcus tauri*; Ostlu, *Ostreococcus lucimarinus*. Bootstrap values >50 are indicated. The gene lying in the MT locus is indicated with an arrow.

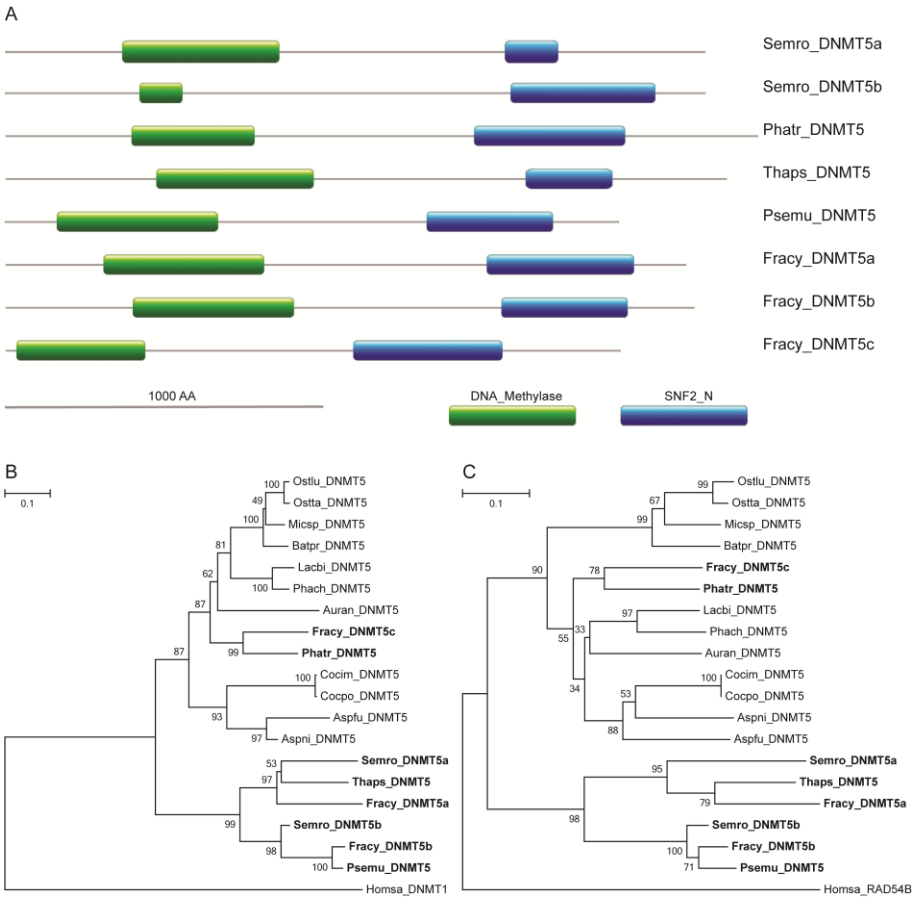


Figure 4: Domain organization and phylogenetic reconstruction of diatom DNMT5 sequences. (A) Representation of domain organization of the diatom DNMT5 sequences generated using the Prosite MyDomains tool. Green: DNA_methylase domain. Blue: SNF2_N domain. (B) NJ tree based on MSA of the MTase domains (Poisson correction, 1000 bootstraps). C. NJ tree based on MSA of the SNF2_N domains (Poisson correction, 1000 bootstraps). Diatom sequences: Semro, *Seminavis robusta*; Phatr, *Phaeodactylum tricornutum*; Thaps, *Thalassiosira pseudonana*; Psemu, *Pseudo-nitzschia multiseries*; Fracy, *Fragilariopsis cylindrus*. Algal sequences: Auran, *Aureococcus anophagefferens*; Batpr, *Bathycoccus prasinos*; Ostta, *Ostreococcus tauri*; Ostlu, *Ostreococcus lucimarinus*; Micsp, *Micromonas sp.*. Fungal sequences: Aspfu, *Aspergillus fumigatus*; Aspni, *Aspergillus nidulans*; Cocim, *Coccidioides immitis*; Cocpa, *Coccidioides posadasii*; Lacbi, *Laccaria bicolor*; Phach, *Phanerochaete chrysosporium*.

From this phylogenetic analysis, we can conclude that the SNF2-type helicase and methyltransferase domain containing gene that co-segregates with the MT indeed belongs to the DNMT5 protein family. As there are two *DNMT5* genes in *S. robusta*, this gene was called *DNMT5a*.

			gene level				protein level				
			MT^+		MT^-		MT^+		MT^-		
bp	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**AA**
6	G	A	22	26	67	0					
261	C	G	6	26	22	0					
318	G	A	17	24	28	0					
687	G	T	17	19	32	0					
771	A	C	10	21	29	0					
831	A	G	25	19	40	0					
974	C	T	29	16	28	0	T	I	T	T	325
1020	T	C	24	8	35	0					
1176	C	A	4	12	22	0					
1302	C	A	27	11	45	0					
1371	T	C	25	17	43	0					
1374	G	A	25	12	44	0					
1384	C	T	27	17	47	0					
1423	A	G	24	12	36	0	I	V	I	I	475
1425	T	G	23	12	34	0					
1689	A	C	10	16	37	0					
1757	A	T	22	13	30	0	K	M	K	K	586
2058	C	G	11	22	36	0					
2064	C	A	12	23	36	0					
2366	A	G	17	9	40	0	I	V	I	I	760
2375	G	+T	14	12	45	0					
2452	T	C	16	15	42	0					
2548	T	C	27	24	46	0					
2566	T	C	21	17	52	0					
2569	A	T	20	17	54	0					
3058	C	T	48	25	53	0	S	L	S	S	905
3098	A	G	37	17	55	0					
6595	T	C	23	23	53	0	V	A	V	V	2084

Table 1: All positions in the *DNMT5a* gene for which MT^+ bulk is heterozygous and MT^- bulk is homozygous are shown. SNPs falling within an intron are indicated in grey. When the nucleotide difference leads to an amino acid change, the amino acids and their position in the protein are shown.

Sequencing *DNMT5a*

To elucidate the sequence polymorphism(s) differentiating both mating types, the BSA-WGS was used to select SNPs in *DNMT5a* for which the MT$^-$ bulk is homozygous and the MT$^+$ bulk is heterozygous. Of the 28 selected SNPs, six are non-synonymous (Table 1). Sanger sequencing of four MT$^+$ (96A, 85A, 109B, 114A) and four MT$^-$ (H73A, 85B, 109A, 114B) strains confirmed that these six non-synonymous SNPs co-segregate with the MT. 96A and H73A are the parental strains used to create the F_1-MP for linkage mapping (Vanstechelman et al., 2013). 85A/B, 109A/B and 114A/B are three sibling pairs belonging the *S. robusta* pedigree depicted in Figure 5.

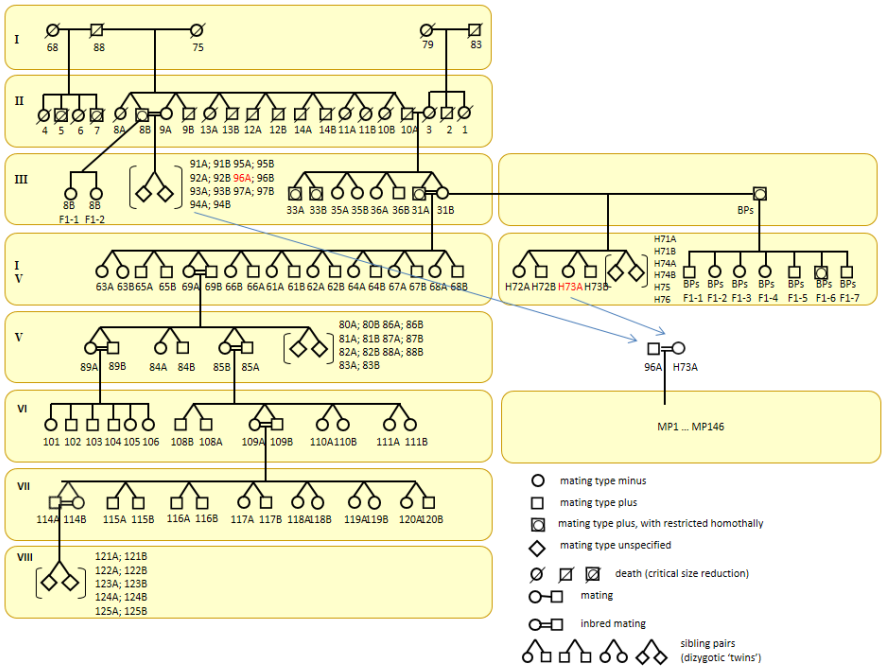


Figure 5: A pedigree of *Seminavis robusta* obtained through interclonal crosses. The founder clones are natural clones isolated from The Netherlands, Zeeland, "Veerse Meer" (Chepurnov et al., 2002). Clone BPs is also a natural clone, but was isolated from The Netherlands, Westerschelde, Paulinaschor. Strains 96A and H73A were crossed to generate the F_1 mapping population consisting of 146 clones (MP1-MP146) (Vanstechelman et al., 2013).

I475V, K586M and I760V are located within the DNA methylase domain, but none of them is located at a conserved position. Using Phyre2 (Kelley and Sternberg, 2009), structural effects of these three polymorphisms were assessed, but no significant differences were found when comparing the predicted structure of the DNA_methylase domain of both alleles. The SuSPect tool (Yates et al., 2014), implemented in Phyre2, can be used to predict whether mutations are likely to have a functional effects, but none of the six non-synonymous SNPs that were identified were likely to have large effects.

Discussion

Linkage analysis previously showed that the MT locus in the heterothallic pennate diatom *S. robusta* segregates as a single locus, disclosing MT^+ as the heterogametic MT (Vanstechelman et al., 2013). Furthermore, the MT locus was restricted to a small segment of linkage group MT^+_6, flanked by large autosome-like regions, suggesting that the MT-determining chromosomes evolved only recently (Bergero and Charlesworth, 2009). In this study, we further investigated the genetic structure of the MT locus in *S. robusta* by a BSA approach using AFLP and WGS technology.

The BSA-AFLP analysis resulted in the identification of two AFLP markers strongly linked to the MT locus mapping to a single 24,698 bp scaffold (scaffold1897) of the draft *S. robusta* genome. The BSA-WGS analysis revealed that this scaffold was highly enriched with SNPs in full linkage disequilibrium with the MT locus displaying heterozygosity in MT^+ and homozygosity in MT^-. A ~40-fold discrepancy was observed in genomic distance estimated from the 2.1 cM intermarker distance of the two AFLP markers in the linkage map (~ 330 kb) and the genomic distance (7905 bp) between the AFLP markers on the scaffold. A ten times higher coverage in the 5' genomic region of this ~8 kb region separating these markers could be caused by the existence of sequence repeats in the 5' region, as such repeat regions are usually concatenated during assembly. These repeats can possibly explain the discrepancy in genomic distance. Another explanation could be a higher recombination rate. This

was already observed in the fungus *Cryptococcus neoformans,* for which there is a discrepancy of the physical/genetic map of around 10- to 50-fold compared to the genome wide average. In this species, recombination is higher in the regions neighboring the MT locus, and the high occurrence of crossovers (including double cross overs) on both sides of the MT locus suggests that the MT locus can be exchanged onto different genetic backgrounds during meiosis (Hsueh et al., 2006).

Scaffold1897 was annotated using the *S. robusta* transcriptome, which made it possible to identify the gene lying in the MT locus as *DNMT5a*, coding for a protein of 2207 amino acids. Members of the DNMT5 protein family contain a SNF2-type DEXDc/HELICc helicase domain next to the C5 methyltransferase domain. This family was discovered in ascomycete and basidomycete fungi (Ponger and Li, 2005). In the meantime, this family was also found in the green algae *Ostreococcus tauri, O. lucimarimus* and *Micromonas pusilla*, in the pelagophyte *Aureococcus anophagefferens* and in the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* (Maumus et al., 2011). Recently, it was shown that DNMT5 is a symmetrical CG methyltransferase and that it probably took over the function of DNMT1 in lineages where this protein was lost (Huff and Zilberman, 2014). DNMT1, and thus probably also DNMT5, acts as a maitenance DNA methyltransferase. After DNA replication, it transfers a methylgroup from s-adenosylmethionine (SAM) to CpG islands in the DNA that are methylated on one of the two strands.

We hypothesize that the *S. robusta* DNMT5a protein plays a role in MT determination by regulating the expression of specific genes through DNA methylation. It has recently been show that DNA methylation plays an important role in sex determination in the half-smooth tongue sole, *Cynoglossus semilaevis* (Chen et al., 2014, Shao et al., 2014). In this species, genetic sex determination (ZW system) co-occurs with temperature-dependent environmental sex determination. During sex reversal, the methylation pattern of the sex determination pathways in ZW females is modified and this modification is inherited by the ZW offspring of these “pseudo-males”. Also in *Populus tomentosa*, a link between sex determination and DNA methylation was found. Two genes involved in DNA methylation, *met1* and

ddm1, are located on sex determination region of Chromosome XIX and are differentially expressed between male and female flowers (Song et al., 2013). *Met1* is a member of the DNMT1 family and is involved in maintenance of CG methylation and *ddm1* is related to the SWI2/SNF2-like proteins. These two functions are also present in *DNMT5a*.

Although no functional effects of the six amino acids changes associated with the MT could be predicted *in silico*, it is still possible that one or more of these SNPs are responsible for a difference in DNMT5a function between the MTs, which can lead to different methylation patterns that determine the MT. A full functional analysis of DNMT5a is necessary to understand its role as a MT-determining factor in *S. robusta* and the importance of the identified amino acid changes. As soon as genetic transformation becomes feasible in *S. robusta*, it will be possible to assess the effects of mutating these six SNPs one-by-one of in combinations.

The SNPs between MTs make it possible to determine the mating type of natural *S. robusta* strains by PCR. This would be a faster method then the one used now, namely make crosses of natural strains with strains from the culture collection of which the mating type is known, since the latter requires cells to be below the SST.

Elucidating the molecular-genetic basis of MT determination in diatoms will not only contribute to a better understanding of the regulation and evolution of their life cycles, but will also establish diatoms as a novel model group to study the evolution of reproductive strategies in eukaryotes. In centric diatoms, only homothallic reproduction is observed, while in pennates, heterothallic reproduction is most common (Mann, 1999, Chepurnov and Mann, 2004). This evolution from homothally (centrics) to heterothally (pennates) represents the evolution to genetic MT determination (Chepurnov et al., 2004, Davidovich et al., 2010).

Literature cited

Ahmed, S., Cock, J.M., Pessia, E., Luthringer, R., Cormier, A., Robuchon, M., Sterck, L., Peters, A.F., Dittami, S.M., Corre, E., Valero, M., Aury, J.M., Roze, D., Van de Peer, Y., Bothwell, J., Marais, G.A. and Coelho, S.M. (2014) A Haploid System of Sex Determination in the Brown Alga Ectocarpus sp. *Current Biology* **24**, 1945–1957.

Bergero, R. and Charlesworth, D. (2009) The evolution of restricted recombination in sex chromosomes. *Trends in Ecology & Evolution* **24**, 94-102.

Bozarth, A., Maier, U.G. and Zauner, S. (2009) Diatoms in biotechnology: modern tools and applications. *Applied Microbiology and Biotechnology* **82**, 195-201.

Chen, S.L., Zhang, G.J., Shao, C.W., Huang, Q.F., Liu, G., Zhang, P., Song, W.T., An, N., Chalopin, D., Volff, J.N., Hong, Y.H., Li, Q.Y., Sha, Z.X., Zhou, H.L., Xie, M.S., Yu, Q.L., Liu, Y., Xiang, H., Wang, N., Wu, K., Yang, C.G., Zhou, Q., Liao, X.L., Yang, L.F., Hu, Q.M., Zhang, J.L., Meng, L., Jin, L.J., Tian, Y.S., Lian, J.M., Yang, J.F., Miao, G.D., Liu, S.S., Liang, Z., Yan, F., Li, Y.Z., Sun, B., Zhang, H., Zhang, J., Zhu, Y., Du, M., Zhao, Y.W., Schartl, M., Tang, Q.S. and Wang, J. (2014) Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nature Genetics* **46**, 253-260.

Chepurnov, V.A., Mann, D.G., Vyverman, W., Sabbe, K. and Danielidis, D.B. (2002) Sexual reproduction, mating system, and protoplast dynamics of *Seminavis* (Bacillariophyceae). *Journal of Phycology* **38**, 1004-1019.

Chepurnov, V.A. and Mann, D.G. (2004) Auxosporulation of *Licmophora communis* (Bacillariophyta) and a review of mating systems and sexual reproduction in araphid pennate diatoms. *Phycological Research* **52**, 1-12.

Chepurnov, V.A., Mann, D.G., Sabbe, K. and Vyverman, W. (2004) Experimental studies on sexual reproduction in diatoms. *International Review of Cytology* **237**, 91-154.

Chepurnov, V.A., Mann, D.G., von Dassow, P., Vanormelingen, P., Gillard, J., Inzé, D., Sabbe, K. and Vyverman, W. (2008) In search of new tractable diatoms for experimental biology. *BioEssays* **30**, 692-702.

Davidovich, N.A., Kaczmarska, I. and Ehrman, J.M. (2010) Heterothallic and homothallic sexual reproduction in *Tabularia fasciculata* (Bacillariophyta). *Fottea* **10**, 251-266.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797.

Felsenstein, J. (1985) Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. *Evolution* **39**, 783-791.

Gillard, J., Frenkel, J., Devos, V., Sabbe, K., Paul, C., Rempt, M., Inze, D., Pohnert, G., Vuylsteke, M. and Vyverman, W. (2013) Metabolomics Enables the Structure Elucidation of a Diatom Sex Pheromone. *Angewandte Chemie-International Edition* **52**, 854-857.

Granum, E., Raven, J.A. and Leegood, R.C. (2005) How do marine diatoms fix 10 billion tonnes of inorganic carbon per year? *Canadian Journal of Botany-Revue Canadienne De Botanique* **83**, 898-908.

Gremme, G., Brendel, V., Sparks, M.E. and Kurtz, S. (2005) Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* **47**, 965-978.

Guillard, R.R.L. (1975) Culture of phytoplankton for feeding marine invertebrates. in *Culture of marine invertebrate animals* 29-60 (Springer).

Haag, E.S. and Doty, A.V. (2005) Sex determination across evolution: Connecting the dots. *PLoS Biology* **3**, 21-24.

Hattori, R.S., Murai, Y., Oura, M., Masuda, S., Majhi, S.K., Sakamoto, T., Fernandino, J.I., Somoza, G.M., Yokota, M. and Strussmann, C.A. (2012) A Y-linked anti-Mullerian hormone duplication takes over a critical role in sex determination. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 2955-2959.

Hsueh, Y.P., Idnurm, A. and Heitman, J. (2006) Recombination hotspots flank the Cryptococcus mating-type locus: Implications for the evolution of a fungal sex chromosome. *PLoS Genetics* **2**, 1702-1714.

Huff, J.T. and Zilberman, D. (2014) Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes. *Cell* **156**, 1286-1297.

Kamiya, T., Kai, W., Tasumi, S., Oka, A., Matsunaga, T., Mizuno, N., Fujita, M., Suetake, H., Suzuki, S., Hosoya, S., Tohari, S., Brenner, S., Miyadai, T., Venkatesh, B., Suzuki, Y. and Kikuchi, K. (2012) A Trans-Species Missense SNP in Amhr2 Is Associated with Sex Determination in the Tiger Pufferfish, Takifugu rubripes (Fugu). *PLoS Genetics* **8**.

Kelley, L.A. and Sternberg, M.J. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols* **4**, 363-371.

Koboldt, D.C., Zhang, Q.Y., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L. and Wilson, R.K. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22**, 568-576.

Koestler, T. and Ebersberger, I. (2011) Zygomycetes, Microsporidia, and the Evolutionary Ancestry of Sex Determination. *Genome Biology and Evolution* **3**, 186-194.

Levitan, O., Dinamarca, J., Hochman, G. and Falkowski, P.G. (2014) Diatoms: a fossil fuel of the future. *Trends in Biotechnology* **32**, 117-124.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, G.P.D.P. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595.

Li, L., Stoeckert, C.J. and Roos, D.S. (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**, 2178-2189.

Mann, D.G. (1999) The species concept in diatoms. *Phycologia* **38**, 437-495.

Marchler-Bauer, A., Lu, S.N., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M.,

Hurwitz, D.I., Jackson, J.D., Ke, Z.X., Lanczycki, C.J., Lu, F., Marchler, G.H., Mullokandov, M., Omelchenko, M.V., Robertson, C.L., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D.C., Zhang, N.G., Zheng, C.J. and Bryant, S.H. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research* **39**, D225-D229.

Maumus, F., Rabinowicz, P., Bowler, C. and Rivarola, M. (2011) Stemming Epigenetics in Marine Stramenopiles. *Current Genomics* **12**, 357-370.

Medlin, L.K. (2011) A Review of the Evolution of the Diatoms from the Origin of the Lineage to Their Populations. in *The Diatom World* 93-118 (Springer).

Myosho, T., Otake, H., Masuyama, H., Matsuda, M., Kuroki, Y., Fujiyama, A., Naruse, K., Hamaguchi, S. and Sakaizumi, M. (2012) Tracing the Emergence of a Novel Sex-Determining Gene in Medaka, Oryzias luzonensis. *Genetics* **191**, 163-170.

Ponger, L. and Li, W.H. (2005) Evolutionary diversification of DNA Methyltransferases in eukaryotic Genomes. *Molecular Biology and Evolution* **22**, 1119-1128.

Saitou, N. and Nei, M. (1987) The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution* **4**, 406-425.

Shao, C.W., Li, Q.Y., Chen, S.L., Zhang, P., Lian, J.M., Hu, Q.M., Sun, B., Jin, L.J., Liu, S.S., Wang, Z.J., Zhao, H.M., Jin, Z.H., Liang, Z., Li, Y.Z., Zheng, Q.M., Zhang, Y., Wang, J. and Zhang, G.J. (2014) Epigenetic modification and inheritance in sexual reversal of fish. *Genome Research* **24**, 604-615.

Smith, C.A., Roeszler, K.N., Ohnesorg, T., Cummins, D.M., Farlie, P.G., Doran, T.J. and Sinclair, A.H. (2009) The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature* **461**, 267-271.

Song, Y.P., Ma, K.F., Ci, D., Chen, Q.Q., Tian, J.X. and Zhang, D.Q. (2013) Sexual dimorphic floral development in dioecious plants revealed by transcriptome, phytohormone, and DNA methylation analysis in Populus tomentosa. *Plant Molecular Biology* **83**, 559-576.

Talavera, G. and Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56**, 564-577.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* **28**, 2731-2739.

Theriot, E.C., Ashworth, M., Ruck, E., Nakov, T. and Jansen, R.K. (2010) A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. *Plant Ecology and Evolution* **143**, 278-296.

Trukhina, A.V., Lukina, N.A., Wackerow-Kouzova, N.D. and Smirnov, A.F. (2013) The Variety of Vertebrate Mechanisms of Sex Determination. *Biomed Research International*.

Van Ooijen, J. (2006) JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations. *Kyazma BV, Wageningen, Netherlands*.

Vandepoele, K., Van Bel, M., Richard, G., Van Landeghem, S., Verhelst, B., Moreau, H., Van de Peer, Y., Grimsley, N. and Piganeau, G. (2013) pico-PLAZA, a

genome database of microbial photosynthetic eukaryotes. *Environmental Microbiology* **15**, 2147-2153.

Vanstechelman, I., Sabbe, K., Vyverman, W., Vanormelingen, P. and Vuylsteke, M. (2013) Linkage Mapping Identifies the Sex Determining Region as a Single Locus in the Pennate Diatom Seminavis robusta. *PLoS ONE* **8**.

Vuylsteke, M., Peleman, J.D. and van Eijk, M.J.T. (2007a) AFLP-based transcript profiling (cDNA-AFLP) for genome-wide expression analysis. *Nature Protocols* **2**, 1399-1413.

Vuylsteke, M., Peleman, J.D. and van Eijk, M.J.T. (2007b) AFLP technology for DNA fingerprinting. *Nature Protocols* **2**, 1387-1398.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875.

Yates, C.M., Filippis, I., Kelley, L.A. and Sternberg, M.J.E. (2014) SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. *Journal of Molecular Biology* **426**, 2692-2701.

Zuckerkandl, E. and Pauling, L. (1965) Evolutionary divergence and convergence in proteins. *Evolving genes and proteins* **97**, 97-166.

Supplementary data

Table S1: Primer sequences used to sequence the 28 selected SNPs in *DNMT5α* in 8 strains of the *S. robusta* pedigree. Start position of the primer sequence on scaffold1987 in bp.

fragment	sequence	start position
1	CAAAAAGCAGAACAAGAGCA	10885
	CTTTTTGGGTTTGGGTTCC	11884
2	TCGACTGGGTGAATCTTTCC	12795
	CCAGTGTGGTGTCCAACAAG	11778
3	TGTGGCGGTAGCTTCTTCTT	13323
	CAGACTGTGCGACAATCCAG	12322
4	GCAAGATTTGTCTCCCGTGT	14259
	AGAATCGTCCCCACAACAAG	13218
5	CGCCCTTACCCACATACAAA	17866
	GGCATCCCCTTGGAGTATTT	16870

Table S2 : Top 0.1% of scaffolds accumulated with LOH events (based on Fisher's Exact Test). The number of SNPs and the number of LOH events are indicated for every scaffold, as well as the p-value of the Fisher's Exact Test. The scaffold containing the MT locus is indicated.

scaffold	SNP	LOH	*p* value
scaffold10997_size2445	45	40	0
scaffold9911_size3546	33	27	0
scaffold7418_size4952	63	51	0
scaffold6699_size6048	102	41	0
scaffold6006_size7647	94	34	0
scaffold5584_size9025	208	62	0
scaffold5159_size10596	61	50	0
scaffold5079_size10925	133	39	0
scaffold5031_size11102	87	41	0
scaffold5005_size11220	77	32	0
scaffold4462_size13617	156	89	0
scaffold4407_size13915	108	66	0
scaffold4271_size14423	145	51	0
scaffold3487_size15880	243	100	0
scaffold3470_size15939	99	45	0
scaffold5467_size16031	192	78	0
scaffold3413_size16110	217	54	0
scaffold3389_size16186	175	69	0
scaffold	**SNP**	**LOH**	***p* value**
scaffold3006_size17883	40	30	0

scaffold2940_size18524	42	25	0
scaffold2602_size20046	128	79	0
scaffold2575_size20200	108	38	0
scaffold2409_size21282	352	123	0
scaffold2311_size21905	32	27	0
scaffold2060_size23468	220	68	0
scaffold2037_size23602	88	58	0
scaffold1933_size24346	232	56	0
scaffold1897_size24698	405	107	0
scaffold1905_size25258	280	72	0
scaffold880_size35938	458	203	0
scaffold866_size38203	436	110	0
scaffold679_size41700	123	83	0
scaffold631_size43421	438	87	0
scaffold556_size50276	636	145	0
scaffold431_size54685	206	50	0
scaffold27913_size684	28	21	3.74E-15
scaffold3335_size20893	125	37	4.3E-15
scaffold2241_size23237	175	43	7.33E-15
scaffold242_size75709	477	74	9.62E-15
scaffold2102_size23526	230	49	1.2E-14
scaffold349_size59093	359	61	8.51E-14
scaffold2619_size19972	138	36	2.91E-13
scaffold6997_size5521	111	31	2.84E-12
scaffold6907_size5675	120	32	3.6E-12
scaffold948_size34817	99	29	5.3E-12
scaffold830_size37219	156	36	6.19E-12
scaffold2630_size20432	184	39	8.01E-12
scaffold8970_size3498	48	21	1.05E-11
scaffold9001_size3480	83	26	1.88E-11
scaffold13689_size1737	28	16	1.46E-10
scaffold1015_size33688	118	29	1.88E-10
scaffold6204_size7162	138	31	3.07E-10
scaffold4319_size14331	173	34	9.32E-10
scaffold4998_size17541	186	35	1.45E-09
scaffold469_size50250	131	29	1.5E-09
scaffold6148_size7280	19	13	1.56E-09
scaffold7108_size5346	85	23	2.89E-09
scaffold1705_size26229	224	38	3.74E-09
scaffold202_size83235	810	10	1.04E-08
scaffold15496_size1438	24	13	1.26E-08
scaffold20485_size1000	50	17	2.05E-08
scaffold8793_size3626	58	18	2.55E-08
scaffold	**SNP**	**LOH**	***p* value**
scaffold1711_size26195	265	40	2.68E-08
scaffold1925_size28972	182	32	2.92E-08

scaffold6876_size5736	90	22	3.01E-08
scaffold16579_size1313	12	10	3.35E-08
scaffold4860_size19523	127	26	4.07E-08
scaffold713_size40284	208	34	5.3E-08
scaffold8904_size3537	71	19	7.72E-08
scaffold8344_size4288	74	19	1.34E-07

Table S3: The position of the SNPs identified as LOH event on the scaffold1897 detected by the software VarScan (Koboldt et al., 2012). The bases for the two different alleles, the number of reads for the first allele of MT^+ (MT^+_1), the number of reads for the second allele of MT^+ (MT^+_2), the number of reads for the first allele of MT^- (MT^-_1), the number of reads for the second allele of MT^- (MT^-_2) and the significance of the allele frequency differences by the Fisher's Exact test. The SNPs falling within the *DNMT5a* gene are indicated.

position	Allele 1	Allele 2	MT^+_1	MT^+_2	MT^-_1	MT^-_2	*p*-value
2881	G	A	15	25	32	0	2.64E-09
3495	T	G	25	19	62	0	3.15E-09
3500	G	A	25	19	63	0	2.59E-09
3558	T	C	23	14	56	0	4.12E-07
3597	G	A	26	12	71	0	8.64E-07
3600	G	T	26	12	72	0	7.70E-07
3669	G	C	15	11	55	0	6.37E-07
3678	G	T	15	12	58	0	1.32E-07
3681	G	C	11	17	53	10	3.15E-05
3749	A	C	26	8	69	0	7.64E-05
3751	G	T	26	9	70	0	2.35E-05
3753	G	C	26	9	71	0	2.15E-05
3769	T	G	23	9	62	0	2.64E-05
3787	A	C	23	9	56	0	4.91E-05
3797	G	T	21	8	45	0	2.85E-04
3799	G	T	20	8	46	0	2.06E-04
3843	G	C	14	5	48	0	0.0012
4048	T	G	7	16	34	8	8.01E-05
4262	T	A	28	7	0	15	7.58E-08
4474	G	C	14	9	41	0	2.97E-05
4921	A	C	22	20	58	0	9.59E-10
5684	G	C	14	11	38	0	7.24E-06
7756	T	C	22	27	52	0	1.90E-11
8931	A	C	20	15	43	0	7.44E-07
position	**Allele 1**	**Allele 2**	**MT^+_1**	**MT^+_2**	**MT^-_1**	**MT^-_2**	***p*-value**
8973	C	A	19	13	40	0	4.90E-06
9078	A	G	24	18	50	0	6.02E-08
9102	T	C	22	11	44	0	2.90E-05

9183	T	C	23	17	46	1	2.23E-06
9203	C	T	21	15	53	0	1.46E-07
9205	G	T	23	15	56	0	1.66E-07
9389	G	A	16	17	48	0	9.08E-09
9633	C	T	27	20	51	0	2.85E-08
9635	A	G	26	15	51	0	9.68E-07
9744	A	G	21	13	14	47	2.09E-04
9816	G	C	17	18	47	0	7.75E-09
9897	A	G	24	19	0	38	3.47E-09
10299	A	G	22	15	24	0	1.33E-04
10924	G	T	25	18	0	28	6.11E-08
10993	A	T	29	23	50	0	8.49E-09
11071	T	A	28	24	77	0	5.75E-12
11077	T	G	26	26	71	0	1.58E-12
11083	A	C	24	25	71	0	1.51E-12
11089	A	C	21	23	70	0	2.77E-12
11107	G	A	22	26	67	0	6.23E-13
11419	A	G	24	17	0	28	6.57E-08
11788	G	T	17	19	32	0	2.57E-07
11865	A	T	32	13	10	40	4.81E-07
11872	C	A	21	10	0	29	5.55E-09
11932	A	G	25	19	40	0	4.26E-07
11938	G	A	15	27	33	7	1.59E-05
12075	C	T	29	16	28	0	1.23E-04
12121	T	C	24	8	35	0	0.0016
12277	A	C	12	4	0	22	6.72E-07
12403	C	A	27	11	45	0	7.45E-05
12472	T	C	25	17	43	0	7.97E-07
12475	G	A	25	12	44	0	2.62E-05
12485	C	T	27	17	47	0	5.97E-07
12524	A	G	24	12	36	0	8.15E-05
12526	T	G	23	12	34	0	9.47E-05
12790	A	C	10	16	37	0	1.45E-08
12858	A	T	22	13	30	0	8.99E-05
12949	G	A	10	26	25	6	1.52E-05
13159	G	C	22	11	0	36	3.29E-10
13165	A	C	23	12	0	36	3.15E-10
13467	A	G	17	9	40	0	8.44E-05
13553	T	C	16	15	42	0	2.07E-07
13649	T	C	27	24	46	0	6.62E-09
13667	T	C	21	17	52	0	3.08E-08
position	**Allele 1**	**Allele 2**	**MT^+_1**	**MT^+_2**	**MT^-_1**	**MT^-_2**	***p*-value**
13670	A	T	20	17	54	0	1.38E-08
14159	C	T	48	25	53	0	1.43E-07
14199	A	G	37	17	55	0	1.44E-06
14466	C	G	45	26	47	5	4.62E-04

14718	T	C	42	24	61	5	5.03E-05
14784	C	G	32	22	43	7	0.0021
16510	A	G	37	11	7	37	2.42E-09
17696	T	C	23	23	53	0	4.30E-10
19257	G	C	23	20	0	38	1.01E-08
19270	A	G	22	16	0	35	8.67E-09
19504	C	T	31	21	0	44	1.31E-11
19506	G	T	31	12	0	24	1.29E-09
20326	A	G	16	7	2	25	4.92E-06
20608	G	T	31	11	53	0	5.48E-05
20751	T	G	29	22	68	0	3.03E-10
20775	G	A	25	21	68	0	1.61E-10
20817	C	T	24	19	49	0	3.50E-08
20880	C	T	23	11	37	0	1.12E-04
20964	T	C	17	26	46	0	2.04E-11
21001	T	C	13	21	41	0	4.41E-10
21059	G	C	17	16	34	0	1.04E-06
21412	C	T	13	38	39	7	2.71E-09
21469	G	T	23	48	53	11	2.11E-09
21518	T	G	41	17	57	0	2.26E-06
21551	G	C	38	19	48	0	1.74E-06
21614	C	T	41	22	48	0	5.55E-07
21624	C	A	41	16	51	0	1.14E-05
21628	A	G	43	16	51	0	1.57E-05
21712	C	T	19	17	55	0	7.48E-09
21815	C	T	22	31	43	0	3.16E-11
21821	G	A	20	29	45	0	1.90E-11
21882	A	C	18	19	41	0	2.63E-08
21962	C	A	22	31	50	0	2.35E-12
22387	T	C	26	17	37	3	5.71E-04
22817	T	C	20	18	34	0	8.10E-07
23672	T	G	29	21	40	0	3.96E-07
24071	C	G	34	18	48	0	1.39E-06

Table S4: Structural and functional annotation of scaffold1897. Gene structure predictions were computed with GenomeThreader (Gremme et al., 2005), based on *the S. robusta* transcriptome. Functional predictions were done using a Conserved Domains Database (CDD)

search. Eleven genes were identified and for five of these genes a protein sequence could be predicted.

	annotation	start	stop	strand	function
gene1	gene	757	1907	-	
	mRNA	757	1907	-	
	exon	757	1679	-	
	CDS	1296	1679	-	
	3'_cis_splice_site	1679	1680	-	
	5'_cis_splice_site	1761	1762	-	unknown
	exon	1762	1791	-	
	CDS	1762	1791	-	
	3'_cis_splice_site	1791	1792	-	
	5'_cis_splice_site	1881	1882	-	
	CDS	1882	1899	-	
	exon	1882	1907	-	
gene2	gene	2526	3850	-	
	mRNA	2526	3850	-	
	exon	2526	2534	-	
	3'_cis_splice_site	2534	2535	-	
	5'_cis_splice_site	2711	2712	-	
	exon	2712	2924	-	
	3'_cis_splice_site	2924	2925	-	
	5'_cis_splice_site	3015	3016	-	
	exon	3016	3143	-	
	3'_cis_splice_site	3143	3144	-	
	5'_cis_splice_site	3213	3214	-	
	exon	3214	3850	-	
gene3	gene	5349	7136	-	
	mRNA	5349	7136	-	
	exon	5349	6731	-	
	CDS	5464	6731	-	serine/threonine
	3'_cis_splice_site	6731	6732	-	protein kinase
	5'_cis_splice_site	6826	6827	-	
	CDS	6827	7112	-	
	exon	6827	7136	-	
gene4	gene	5415	5642	+	
	mRNA	5415	5642	+	
	exon	5415	5642	+	
gene5	gene	7465	8611	+	
	mRNA	7465	8611	+	
	exon	7465	8611	+	

	annotation	start	stop	strand	function
gene6	gene	8559	10050	-	
	mRNA	8559	10050	-	hint-domain
	exon	8559	10050	-	containing protein
	CDS	8616	10001	-	
gene7	gene	9257	9527	+	
	mRNA	9257	9527	+	
	exon	9257	9527	+	
gene8	gene	11067	18455	+	
	mRNA	11067	18455	+	
	exon	11067	13352	+	
	CDS	11102	13352	+	
	5'_cis_splice_site	13353	13354	+	
	3'_cis_splice_site	13439	13440	+	
	exon	13441	13468	+	
	CDS	13441	13468	+	
	5'_cis_splice_site	13469	13470	+	SNF2-type helicase
	3'_cis_splice_site	13551	13552	+	domain C5-specific
	exon	13553	13678	+	methyltransferase
	CDS	13553	13678	+	domain
	5'_cis_splice_site	13679	13680	+	
	3'_cis_splice_site	13761	13762	+	
	exon	13763	13915	+	
	CDS	13763	13915	+	
	5'_cis_splice_site	13916	13917	+	
	3'_cis_splice_site	14002	14003	+	
	CDS	14004	18069	+	
	exon	14004	18455	+	
gene9	gene	18623	19185	-	
	mRNA	18623	19185	-	
	exon	18623	19185	-	
gene10	gene	20259	20807	+	
	mRNA	20259	20807	+	
	exon	20259	20807	+	

	annotation	start	stop	strand	function
gene11	gene	21585	22811	+	
	mRNA	21585	22811	+	
	exon	21585	21615	+	
	5'_cis_splice_site	21616	21617	+	
	3'_cis_splice_site	21688	21689	+	
	exon	21690	21758	+	
	CDS	21737	21758	+	
	5'_cis_splice_site	21759	21760	+	
	3'_cis_splice_site	21842	21843	+	Leucine-rich-repeat containing protein
	exon	21844	21918	+	
	CDS	21844	21918	+	
	5'_cis_splice_site	21919	21920	+	
	3'_cis_splice_site	21988	21989	+	
	exon	21990	22061	+	
	CDS	21990	22061	+	
	5'_cis_splice_site	22062	22063	+	
	3'_cis_splice_site	22152	22153	+	
	CDS	22154	22395	+	
	exon	22154	22811	+	

Table S5: Results of the HMMsearch using PF00145.12 HMM profile.

		E-value	score	bias	E-value	score	bias	
Fragillariopsis cylindrus	jgi|Fracy1|148014|gw1.4.362.1	3.00E-55	187.9	0.1	7.7e-54	183.3	0.0	2.6
	jgi|Fracy1|242149|fgenesh2_pg.9_#_623	1.8e-46	159.1	0.0	1.8e-46	159.1	0.0	5.0
	jgi|Fracy1|250079|fgenesh2_pg.29_#_116	9.2e-41	140.3	8.3	1.2e-38	133.3	0.0	3.3
	jgi|Fracy1|240634|fgenesh2_pg.7_#_972	8.3e-34	117.4	0.3	5.7e-33	114.7	0.0	2.0
	jgi|Fracy1|182730|e_gw1.4.1066.1	1.00E-22	81.0	0.0	9.9e-16	58.0	0.0	2.0
	jgi|Fracy1|205902|estExt_Genewise1Plus.C_20463	8.3e-18	64.8	0.0	2.9e-17	63.0	0.0	1.9
	jgi|Fracy1|139371|gw1.2.229.1	7.5e-17	61.7	0.9	5.5e-16	58.8	0.9	2.4
	jgi|Fracy1|239160|fgenesh2_pg.6_#_562	4.00E-09	36.3	0.0	0.0011	18.5	0.1	3.5
	jgi|Fracy1|212855|estExt_Genewise1Plus.C_290020	5.1e-09	35.9	5.7	0.00091	18.7	0.0	3.5
	jgi|Fracy1|234638|fgenesh2_pg.2_#_746	0.0034	16.8	0.0	0.0034	16.8	0.0	2.2
Phaeodactylum tricornutum	jgi|Phatr2|47357|estExt_fgenesh1_pg.C_chr_130154	1.4e-64	217.8	0.0	5.5e-50	169.8	0.0	2.1
	jgi|Phatr2|16674|e_gw1.28.55.1	5.2e-31	107.4	0.0	6.4e-31	107.1	0.0	1.1
	jgi|Phatr2|45072|estExt_fgenesh1_pg.C_chr_60051	1.8e-16	59.6	0.0	1.3e-15	56.7	0.0	2.0
	jgi|Phatr2|46156|estExt_fgenesh1_pg.C_chr_90163	1.8e-12	46.5	0.0	9.2e-11	40.8	0.0	2.4
	jgi|Phatr2|32528|fgenesh1_pg.C_chr_2000069	0.0011	17.5	0.9	0.0015	17.1	0.9	1.2
Pseudo-nitzschia multiseries	jgi|Psemu1|170853|gw1.1657.17.1	8.3e-53	180.0	0.0	4.2e-51	174.4	0.0	2.1
	jgi|Psemu1|30522|gm1.30522_g	2.3e-51	175.2	0.0	1.8e-25	90.1	0.0	2.5
	jgi|Psemu1|285042|fgenesh1_pg.72_#_6	2.7e-45	155.3	0.2	1.8e-31	109.9	0.0	2.5
	jgi|Psemu1|103852|gw1.770.15.1	3.9e-23	82.5	0.0	5.3e-18	65.6	0.0	2.0
	jgi|Psemu1|195867|e_gw1.179.67.1	1.5e-16	60.8	0.0	7.3e-15	55.3	0.0	3.0
	jgi|Psemu1|106695|gw1.1044.1.1	3.1e-13	49.9	0.0	3.4e-13	49.8	0.0	1.0
	jgi|Psemu1|192957|e_gw1.134.12.1	1.2e-09	38.1	0.2	2.6e-05	23.9	0.0	3.0
Thalassiosira pseudonana	jgi|Thaps3|11011|fgenesh1_pg.C_chr_18000254	7.7e-35	120.2	0.0	4.2e-25	88.2	0.0	3.7
	jgi|Thaps3|2094|fgenesh1_pg.C_chr_2000049	1.00E-27	96.8	0.4	2.00E-24	85.9	0.0	2.3
	jgi|Thaps3|22139|estExt_fgenesh1_pg.C_chr_40165	5.6e-13	48.3	0.2	5.6e-08	31.8	0.0	2.6
	jgi|Thaps3|3158|fgenesh1_pg.C_chr_3000092	1.4e-10	40.4	2.4	7.7e-06	24.8	0.0	3.0

species	Protein ID	full sequence			best domain			exp
		E-value	score	bias	E-value	score	bias	
Seminavis robusta	comp81384_c0_seq2	1.1e-61	211.1	0.0	1.5e-61	210.6	0.0	1.2
	comp86090_c0_seq1	1.00E-60	207.8	0.0	1.3e-60	207.5	0.0	1.1
	comp81384_c0_seq3	6.9e-60	205.1	0.0	1.2e-59	204.3	0.0	1.3
	comp80811_c0_seq2	1.8e-50	174.2	0.0	4.5e-34	120.3	0.0	2.3
	comp78603_c0_seq1	1.9e-48	167.5	0.0	8.2e-38	132.6	0.0	2.2
	comp43211_c0_seq1	9.1e-46	158.7	4.6	6.3e-34	119.8	0.0	2.9
	comp79847_c0_seq1	9.8e-45	155.3	0.0	4.1e-31	110.5	0.0	2.2
	comp54181_c0_seq1	3.7e-44	153.4	0.0	9.4e-31	109.4	0.0	2.3
	comp54181_c0_seq4	8.3e-44	152.3	0.0	1.5e-30	108.7	0.0	2.3
	comp79886_c0_seq1	1.4e-43	151.6	2.4	2.1e-35	124.6	0.0	3.1
	comp70684_c0_seq1	8.7e-43	148.9	0.0	8.00E-29	103.0	0.0	2.1
	comp75691_c0_seq1	5.4e-42	146.3	0.0	5.7e-28	100.2	0.0	2.4
	comp63473_c0_seq3	1.9e-39	137.9	0.0	3.1e-25	91.2	0.0	2.2
	comp63473_c0_seq1	4.6e-39	136.7	0.0	6.5e-25	90.2	0.0	2.5
	comp53288_c0_seq1	1.00E-38	135.5	0.0	9.1e-33	116.0	0.0	2.1
	comp75400_c0_seq1	1.3e-37	131.9	1.0	1.3e-25	92.5	0.0	2.6
	comp75493_c0_seq1	1.4e-32	115.4	0.0	1.4e-32	115.4	0.0	2.8
	comp63473_c0_seq4	2.7e-26	94.7	0.0	9.3e-12	46.9	0.0	2.2
	comp86091_c0_seq1	2.8e-09	38.7	0.0	3.3e-09	38.5	0.0	1.1
	comp19947_c0_seq1	3.1e-09	38.6	0.0	3.2e-09	38.6	0.0	1.0
	comp84465_c0_seq8	3.2e-06	28.7	0.0	3.2e-06	28.7	0.0	2.8
	comp84465_c0_seq1	3.3e-06	28.7	0.0	3.3e-06	28.7	0.0	2.8
	comp83320_c0_seq1	4.3e-05	25.0	0.0	5.8e-05	24.6	0.0	4.4
	comp84465_c0_seq7	0.00015	23.2	0.0	0.00027	22.4	0.0	1.3
	comp59630_c0_seq1	0.0013	20.2	0.1	0.006	17.9	0.0	2.0
	comp63473_c0_seq2	0.0071	17.7	0.2	0.018	16.4	0.0	1.7
	comp27905_c0_seq1	0.0078	17.6	0.0	0.009	17.4	0.0	1.1

3

A transcriptomic study of the life cycle of the pennate diatom *Seminavis robusta*

Sara Moeys [1,2,3], Marie J.J. Huysman [2,3], Valerie Devos [1,2,3], Shrikant Patil [4], Remo Sanges [4], Maria I. Ferrante [4], Marina Montresor [4], Lieven De Veylder [2,3], Wim Vyverman [1]

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

[2] Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Gent, Belgium

[3] Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, B-9052 Gent, Belgium

[4] Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Napoli, Italy

Manuscript in preparation

Authors' contributions

SM analyzed the data and wrote the manuscript. MJJH performed the cyclin annotation. VD performed the sampling. SP, MIF and MM were involved in the identification of the meiosis-related genes. RS performed the transcriptome assembly. SM, MJJH, VD, LDV and WV conceived and designed the study. LDV and WV read and approved the manuscript.

Abstract

Diatoms are evolutionary successful as they comprise around 200,000 species that occupy very diverse habitats. Even though their unique life cycle is of great importance for their evolutionary success, not much is known about its regulation. To learn more on the diatom life cycle, an elaborate RNA-seq experiment was conducted using *Seminavis robusta*, a model species for pennate diatoms. Samples of both mating types, above and below the sexual size threshold, were taken for sequencing during the vegetative and the sexual phase. The resulting transcriptome was parsed for genes involved in several processes that are important in the diatom life cycle. Firstly, a large number of known meiosis-related genes could be found in *S. robusta*, as well as in other available diatom genomes. Secondly, the large cyclin gene family was described in this transcriptome study. A large number of diatom-specific cyclins was found, as well as several A/B-type cyclins, which seem to play a role in meiosis rather than in mitosis. Thirdly, several proteins involved in silica deposition, like silica transporters, frustulins and SPM/SPD synthases, were identified. Several of these genes show expression peaks during G_2/M phase, when cell division takes place, or during the later stages of sexual reproduction, when new valves are formed. Furthermore, large differences in expression between vegetative and sexual stages could be found. In conclusion, this dataset can be used as a starting point for the study of life cycle regulation in pennate diatoms.

Introduction

Despite the importance of understanding the evolutionary success of diatoms, not much is known about the regulation of their life cycle. This lack of knowledge was tackled by conducting an elaborate transcriptomic study of the different life cycle phases of the diatom *Seminavis robusta*. *S. robusta* was chosen here as a model for pennate diatoms, since it displays a typical pennate life cycle that can be easily controlled experimentally (Chepurnov et al., 2008). The diatom life cycle consists of a long vegetative phase and a short sexual phase. During the vegetative phase, the average cell size of the population decreases with each cell division (Figure 1.6) (Drebes, 1977). This is a consequence of their rigid cell wall that consists of two thecae fitting together like the two parts of a petri-dish. Each daughter cell inherits one theca that becomes the upper halve or epitheca. Within this epitheca, the cells form a new hypotheca, leading to a reduced cell size for one of the daughter cells. This cell size reduction will ultimately lead to clonal cell death, unless the cells can undergo sexual reproduction. Sex will lead to the formation of a specialized zygote, called the auxospore, which is able to expand to initial cell size (Figure 1.1-5).

Like the majority of pennate diatoms, *S. robusta* is heterothallic, meaning that cells of two different mating types are required for sexual reproduction to occur. It was recently shown that the interaction between the two mating types is controlled by a multistep pheromone system (Gillard et al., 2013). Once the cell size is below the sexual size threshold (SST) of ± 50 µm, *S. robusta* cells produce a "conditioning" factor that signals the presence of a suitable mating partner to the opposite mating type (MT). Under the influence of the conditioning factor produced by MT^+ (CF-P), MT^--cells secrete the attraction pheromone diproline (Gillard et al., 2013). MT^+-cells are attracted to diproline, but only when they are below the SST and when they have first sensed the conditioning factor secreted by MT^- (CF-M). This attraction can then lead to pair formation or mating between two cells of opposite mating type (Figure 1.1). After this, the differentiation in gametangia begins with meiosis I (Chepurnov et al., 2002). This is followed by cytokinesis, leading to the formation of two protoplasts per gametangium, after which the nuclei undergo meiosis II without a

second cytokinesis (Figure 1.2). The protoplasts differentiate into gametes that are released and almost immediately fuse with the gametes of the other gametangium (Figure 1.3). After plasmogamy, the zygotes are released from the thecae and become spherical. Before auxospores start to develop, two of the four haploid nuclei of the zygote are aborted. Karyogamy occurs at a late stage of auxospore expansion or right after its completion (Figure 1.4). When the auxospore is fully expanded (± 68 µm), two thecae are formed internally and the resulting initial cell will start to divide mitotically again (Figure 1.5).

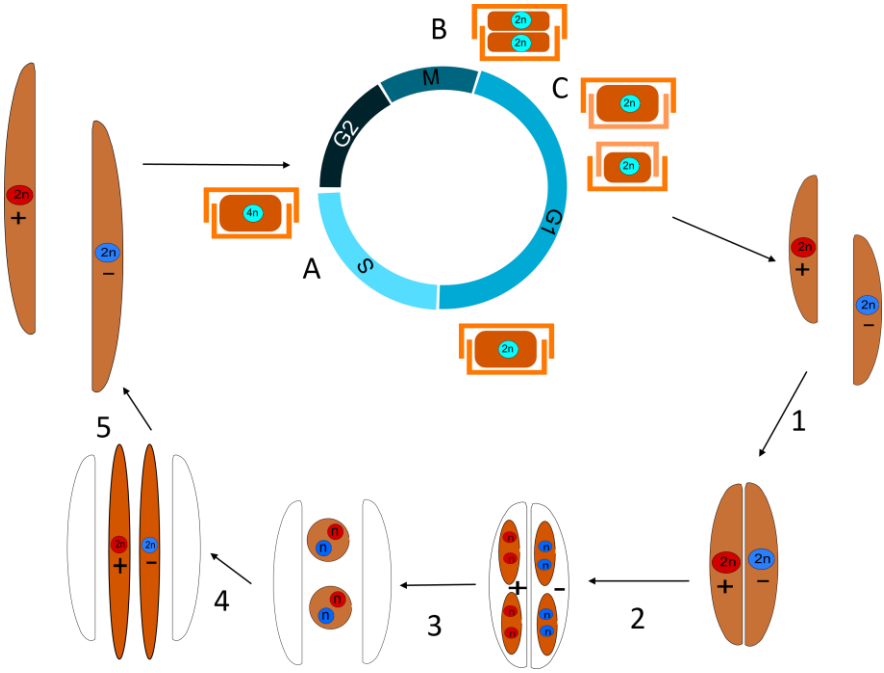


Figure 1: The life cycle of *Seminavis robusta*. When cells of both mating type drop in size below the SST, pairs are formed under the influence of pheromones (1). During gamete formation, both gametangia produce two protoplasts with two haploid nuclei each, of which one will be aborted before auxosporulation (2). When the gametes are released, they immediately fuse to form zygotes (3). These zygotes are called auxospores and are able to expand to initial cell size (4). After that the auxospores are fully expanded, a new frustule is deposited (5) and cells will divide mitotically again (6). During the S phase of the cell cycle, the genome is duplicated (6A) and during M phase mitosis takes place, after which the cell divides (6B). Because both daughter cells inherit one theca and make a new hypotheca, one of the daughter cells will be smaller than the mother cell (6C).

During the vegetative phase of their life cycle the diatom cells pass through the cell cycle to undergo cell division (Figure 1.6). The cell cycle consists of a DNA replication phase (S phase), a phase of mitosis where the two copies of the genome are physically separated (M phase), and finally cell division or cytokinesis itself. S and M phase are separated by two gap phases, G_1 phase precedes S phase and after S phase the cells go through G_2 phase before M phase is initiated. The eukaryotic cell cycle can be arrested at cell cycle checkpoints positioned at the G_1-S or G_2-M transition, or during mid-to-late G_1 phase (Buchanan et al., 2000). By activation of these checkpoints, diatoms respond adequately to environmental changes, like nutrient limitation or variations in light intensity (Olson et al., 1986, Falciatore et al., 2000). Generally, the cell cycle is regulated at these checkpoint by the cyclin-dependent kinases (CDKs) and cyclins (Inze and De Veylder, 2006). These proteins form complexes in which CDKs act as catalytic subunits, while cyclins determine substrate specificity. Diatoms appear to possess a majorly expanded cyclin gene family, with a large number of genes belonging to a new cyclin class, the diatom-specific cyclins (dsCYCs) (Huysman et al., 2010). In *P. tricornutum,* some of these dsCYCs were shown to be induced upon nutrient availability (phosphate: dsCYC5, dsCYC7 and dsCYC10; silica: dsCYC9) or light (dsCYC2), pointing towards a role in environmental signalling to control the cell cycle in a constantly changing environment (Sapriel et al., 2009, Valenzuela et al., 2012, Huysman et al., 2013). It was shown that the onset of cell division after light deprivation in *P. tricormutum* is regulated by dsCYC2 (Huysman et al., 2013). This diatom-specific cyclin controls the G_1-S transition and its transcription is triggered by light onset through the activation of the blue light sensor AUREOCHROME1a. As light deprivation causes cell cycle arrest in diatoms, this strategy can be used to easily synchronize their cell cycle (Huysman et al., 2010). This method has also been applied for *S. robusta*, where a prolonged dark period arrests the majority of cells in G_1 phase (Gillard et al., 2008).(Huysman et al., 2013)

Here, an explorative RNA-seq dataset is described for which samples that arerepresentative of the main life cycle stages of *S. robusta* were taken, including

different phases of vegetative and sexual reproduction, of strains of both mating types, above as well as below the SST. A *de novo* transcriptome was assembled and used to search for genes involved in mitosis, meiosis and silica deposition. GO enrichment analyses were performed to assess the difference between mitotically dividing cells and cells undergoing sexual reproduction.

Materials & methods

Strains and sampling

S. robusta strains 85A and 85B were used, which are publicly available in the diatom culture collection of the Belgian Coordinated Collection of Micro-organisms (BCCM/DCG, http://bccm.belspo.be, accession numbers DCG 0105 and DCG 0107). Cultures were grown in F/2 medium (Guillard, 1975) made with autoclaved filtered natural sea water collected from the North Sea and Guillard's F/2 solution (Sigma-Aldrich).

For the vegetative libraries, *S. robusta* strains 85A (MT^+) and 85B (MT^-) with an average cell size above and below the sexual size threshold (SST) were grown at 18°C in a 12:12h light:dark regime with cool white fluorescent lamps at approximately 80 µmol photons m^{-2} s^{-1}. Before sampling, the dark period was extended with 12 hours to synchronize cells in G_1 phase. After illumination, synchronization was assessed by light microscopy. Pictures were taken using a digital camera connected to a Zeiss Axiovert 40 light microscope and the percentage of dividing cells (distinguished from interphase cells by the newly built cell wall between the two valve-appressed chloroplasts) was counted using cell counter plug-in of the ImageJ software. Cells were harvested hourly until 12 hours post-illumination. The cells were brought in suspension and transferred to a 50 mL Falcon tube and centrifuged for 6 to 10 min at 1,500 to 2,000 × g at 4°C. The supernatant was poured off, and the tube containing the cell pellet was frozen in liquid nitrogen and stored at -80 °C until RNA preparation. Also for the sampling of the sexual stages, cultures were dark syncronized as described above. Then, 85A cell suspensions were added three hours

before illumination to 85B cultures from which the medium was removed. This step was carried out in darkness. Harvesting was done analogous to the vegetative samples from 9 to 24 hours post-illumination. Pictures were taken as discribed above and the different cells types (interphase cells, dividing cells, gametes, zygotes, auxospores and initial cells) were counted.

Total RNA was extracted from each sample using the RNeasy Plant Mini Kit (Qiagen). Cell lysis was achieved by mechanical disruption in 1 mL of RNeasy Lysis buffer (Qiagen) by highest speed agitation with glass/zirconium beads (0.1 mm diameter; Biospec) on a bead mill (Retsch). All other steps for RNA extraction were done according to the manufacturer's instructions. RNA samples were pooled in equal amounts before sequencing.

Transcriptome sequencing and assembly

The RNA was sequenced in collaboration with the JGI institute (http://www.jgi.doe.gov/) within the project "A deep transcriptomic and genomic investigation of diatom life cycle regulation". Poly-A RNA was isolated from 5 µg total RNA using Dynabeads mRNA isolation kit (Invitrogen). The isolation procedure was repeated to ensure that the samples were depleted of rRNA. Purified RNA was subsequently fragmented using RNA Fragmentation Reagents (Ambion) at 70^{o}C for three minutes, targeting fragments of 200-300 bp. Fragmented RNA was purified using Ampure XP beads (Agencourt). Reverse transcription was performed using SuperScript II Reverse Transcription (Invitrogen) with an initial annealing of random hexamer (Fermentas) at 65^{o}C for five minutes, followed by an incubation of 42^{o}C for 50 minutes and an inactivation step at 70^{o}C for ten minutes. cDNA was then purified with Ampure XP beads. This was followed by second strand synthesis using a dNTP mix where dTTP is replaced by dUTP. Reaction was performed at 16^{o}C for one hour. Double stranded cDNA fragments were purified and selected for targeted fragments (200-300 bp) using Ampure XP beads. The ds-cDNA were then blunt-ended, poly-adenylated, and ligated with library adaptors using Kapa Library Amplification Kit (Kapa Biosystems). Adaptor-ligated DNA was purified using Ampure XP beads. Digestion of dUTP was then performed using AmpErase UNG (Applied Biosystems)

to remove second strand cDNA. Digested cDNA was again cleaned up with Ampure XP beads. This is followed by amplification by ten cycles PCR using Kapa Library Amplification Kit (Kapa Biosystems). The final library was cleaned up with Ampure XP beads. Sequencing was done on the Illumina HighSeq2000 platform generating paired end reads of 150 bp each.

Raw reads were filtered and trimmed based on quality and adapter inclusion using Trimmomatic (Bolger et al., 2014). Trimmed and filtered reads were normalized using the normalize_by_kmer_coverage.pl script from the Trinity software release r2013_08_14 (Grabherr et al., 2011). Assembly was performed using Trinity on the trimmed, filtered and normalized . Parameter settings can be found in Suppl. Data.

Results and discussion

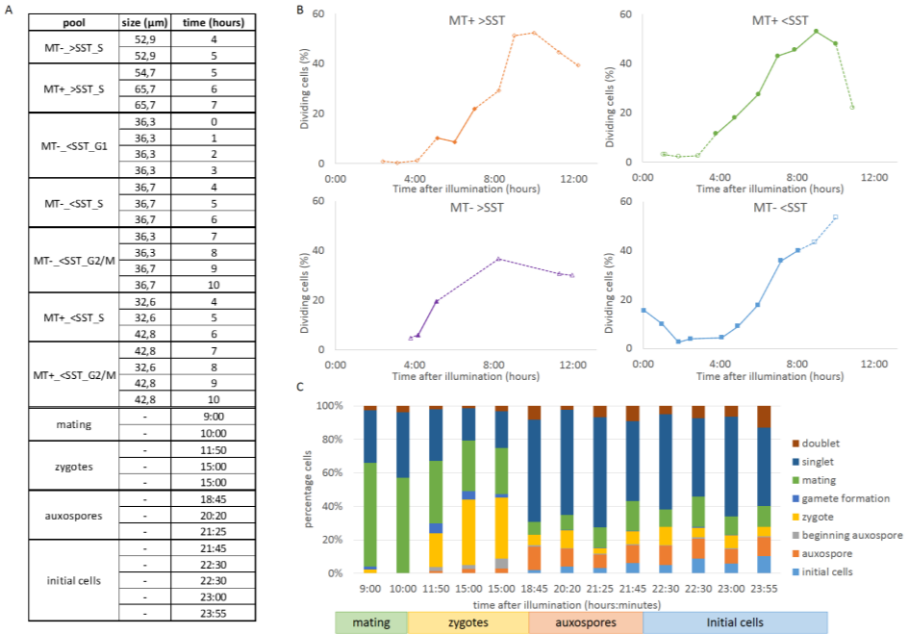
Sequencing and transcriptome assembly

To sample cells during vegetative phase, cultures were synchronized in G_1 phase by prolonging the dark period. After this dark period, lights were switched on and cultures were harvested hourly until twelve hours after illumination. At every time point, the progression through the cell cycle was assessed by counting the number of dividing cells (Figure 2B) (Gillard et al., 2008). By counting dividing cells, it could be estimated that G_1 phase lasted until approximately four hours after illumination, then S phase started and after seven hours G_2/M phase began, which ended around nine hours after illumination. These observations were used to pool RNA from samples belonging to the same cell cycle phase (Figure 2A). This was done for two strains, 85A (MT^+) and 85B (MT^-), both above and below the SST.

For the samples during sexual reproduction, cultures were synchronized in G_1 phase by a prolonged dark period. 85A and 85B cultures were mixed three hours before illumination and cultures were harvested from 9 until 24 hours after illumination. For every time point, dividing cells, cell pairs, gametes, zygotes, auxospores and initial cells were counted to monitor the different phases of sexual

reproduction (Figure 2C). RNA from different samples were pooled according to the frequency of mating cells, zygotes, auxospores and initial cells (Figure 2A).

Eleven pools were selected for sequencing on the Illumina platform of the JGI institute (Figure 2A). This resulted in a total of 1,178,393,498 reads, on average 107,126,682 per library. After de novo assembly, the transcriptome contained 71,790 transcripts covering a total of 101.7 Mb with an N50 of 1,883 bp and transcript length ranging from 201 bp to 13,875 bp.



A

pool	size (µm)	time (hours)
MT-_>SST_S	52,9	4
	52,9	5
MT+_>SST_S	54,7	5
	65,7	6
	65,7	7
MT-_<SST_G1	36,3	0
	36,3	1
	36,3	2
	36,3	3
MT-_<SST_S	36,7	4
	36,7	5
	36,7	6
MT-_<SST_G2/M	36,3	7
	36,3	8
	36,7	9
	36,7	10
MT+_<SST_S	32,6	4
	32,6	5
	42,8	6
MT+_<SST_G2/M	42,8	7
	32,6	8
	42,8	9
	42,8	10
mating	-	9:00
	-	10:00
zygotes	-	11:50
	-	15:00
	-	15:00
auxospores	-	18:45
	-	20:20
	-	21:25
initial cells	-	21:45
	-	22:30
	-	22:30
	-	23:00
	-	23:55

Figure 2: (A) Table with an overview of all RNA pools. (B) The percentage of dividing cells (indication for cell cycle progression) for both MTs above and below SST. Samples marked by full bullets were included in the pools. (C) Percentage of doublets (dividing cells), singlets, mating cells, gametes, zygotes, (beginning) auxospores and initial cells in mixed cultures from 9 to 24 hours after illumination. Below the graph, the four "sexual" pools are indicated.

Meiosis-related genes

Sexual reproduction, and thus meiosis, is widespread in diatoms, although it is sometimes difficult to observe in nature or to induce in culture. As more sequence data is becoming available apart from the two asexual model species *Phaeodactylum*

tricornutum and *Thalassiosira pseudonana*, it is now possible to make an inventory of meiotic genes present in diatoms.

A list of conserved meiotic proteins was taken from Malik et al. (2008) and the list was expanded with additional proteins involved in meiosis according to Hanson et al. (2013). The function of these proteins is described in Supplementary Table S1. *Arabidopsis thaliana* meiotic protein sequences were used as query sequences to search homologs in the available diatom genomes (Suppl. Table S2). Whenever a meiotic protein was not found in the *A. thaliana* genome, corresponding sequences were retrieved from *Saccharomyces cerevisiae*. Homology searches for 61 meiotic proteins were performed in five diatom genomes, those of *T. pseudonana*, *P. tricornutum*, *Fragilariopsis cylindrus*, *Pseudo-nitzschia multiseries*, *Pseudo-nitzschia multistriata* (provided by M. Ferrante), and in the transcriptome of *S. robusta*.

Of the 61 meiosis-related genes selected for this study, 41 were found to be present in diatoms (Suppl. Table S2). Among these 41 genes, 36 are not specific to meiosis and are known to have additional roles other than that in meiosis. These genes mainly play roles in DNA duplication, chromosome maintenance and stability and DNA repair mechanisms.

Of the genes considered in the present study, eleven are known to be specific to meiosis. Of these, five were detected in all diatom genomes surveyed. The most conserved and central of these in meiotic recombination is *SPO11* whose protein product catalyzes programmed double-strand breaks in the genome and initiates meiotic recombination (Keeney et al., 1997). The other four genes include *MND1*, whose protein product forms a heterodimer with Hop2 and facilitates Dmc1-dependent crossover formation (Henry et al., 2006, Chi et al., 2007), *MSH4* and *MSH5*, whose products form a complex and are thought to stabilize crossover intermediates (Snowden et al., 2004, Nishant et al., 2010), and *MER3*, thought to be involved in synaptonemal complex formation (Lynn et al., 2007). Six meiosis-specific genes, namely *ZIP1*, *RED1*, *HOP1*, *HOP2*, *DMC1* and *REC8* could not be identified in any of the diatom genomes. Zip1, Red1 and Hop1 are known to be involved in formation of the synaptonemal complex (Lynn et al., 2007). Rec8 is the meiosis-

specific homolog of Rad21, both proteins are part of the cohesin complex (Watanabe and Nurse, 1999). Rec8 seems to be absent in the majority of the protists (Malik et al., 2007). Intriguingly, Hop2, that forms a heterodimer with Mnd1, could not be found, while Mnd1 was identified in all diatoms investigated.

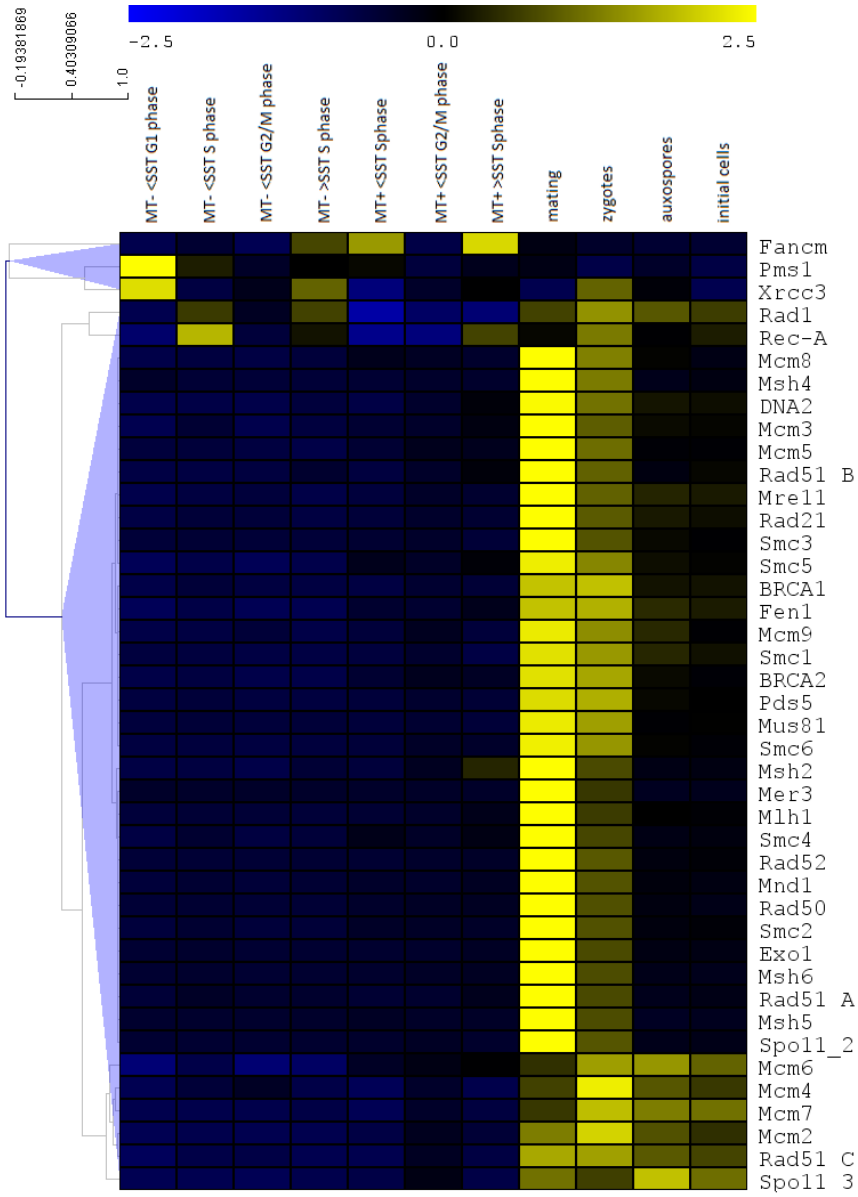


Figure 3: Hierarchical average linkage clustering of normalized cpm values of the *S. robusta* meiosis-related genes.

In *S. robusta*, the expression of the meiosis-related genes was checked over all libraries (Figure 3). All but three genes are highly expressed during cell pairing, the phase at which meiosis takes place, thus confirming their role in meiosis. *FANCM* is mainly expressed during S phase, which corresponds to its role in DNA replication (Luke-Glaser et al., 2010). *PMS1* was higher expressed during G_1 phase, like is the case in *S. cerevisiae* (Spellman et al., 1998). *XRCC3* expression did not seem to be linked to a certain cell cycle phase in this experiment.

Even though no sexual reproduction has been observed in *P. tricornutum* and *T. pseudonana*, all investigated diatom species have a similar set of meiosis-related genes. It is thus possible that these species are still capable of meiosis or that they lost the ability to undergo sexual reproduction only recently and consequently did not yet lose their meiotic genes.

Cyclins

In nature, diatoms encounter rapid environmental fluctuations, for example in nutrient and light availability, which can have large effects on cell physiology (Round et al., 1990). To adapt to these ever-changing circumstances, signalling of environmental cues should be efficiently integrated in cell cycle regulation (Falciatore et al., 2000). The diatoms *T. pseudonana* and *P. tricornutum* have a majorly expanded cyclin gene family. Apart from members of the canonical cyclin gene families, a large number of genes belonging to a new cyclin class, the diatom-specific cyclins (dsCYCs), have been identified (Huysman et al., 2010). In *P. tricornutum,* some of these dsCYCs were shown to be induced upon nutrient availability or light, pointing towards a role in environmental signalling to control the cell cycle in a constantly changing environment (Huysman et al., 2010, Huysman et al., 2013).

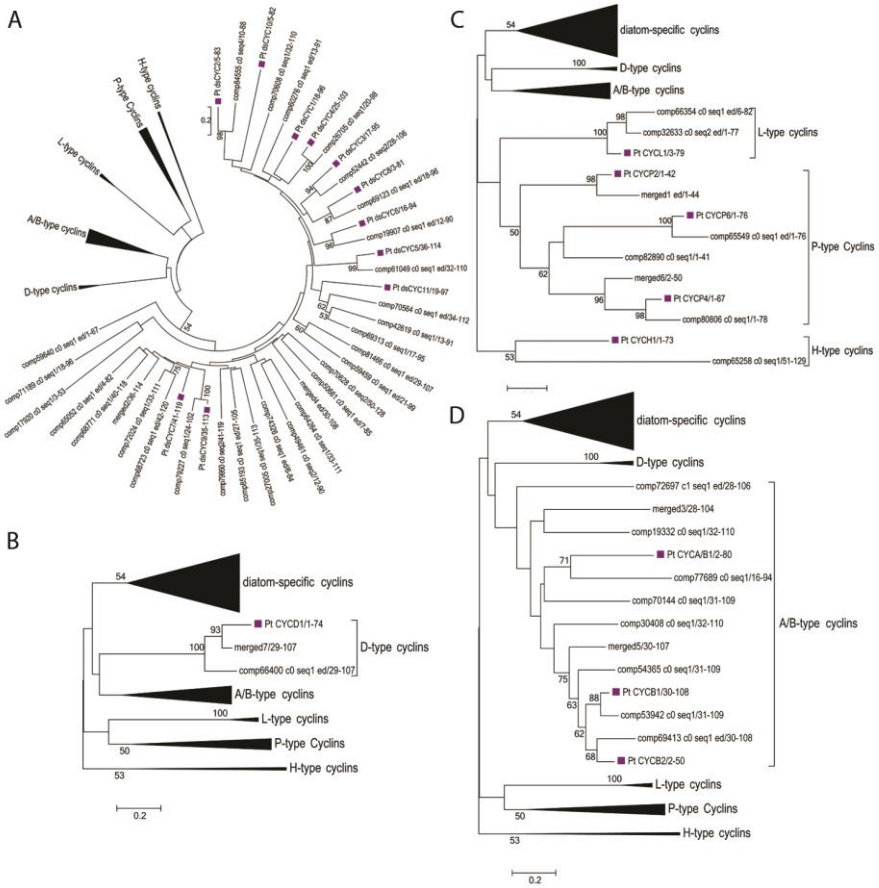


Figure 4: Neighbour-joining tree (Poisson distribution, pairwise deletion, 1000 replicates) of the cyclin family of *P. tricornutum* (purple squares) and *S. robusta*. The tree is depicted as four subtree with (A) the diatom-specific cyclins, (B) D-type cyclins, (C) L-type, P-type and H-type cyclins and (D) the A/B-type cyclins.

HMMER searches were conducted to identify cyclin genes in *S. robusta*. First, protein sequences were predicted based on the transcriptome using Trapid (Van Bel et al., 2013). Then, HMMER 3.0 software (Eddy, 2011) was run on this protein database with a N-terminal cyclin domain HMM profile (PF00134.13) and C-terminal cyclin domain HMM profile (PF02984.8). After reducing redundancy and improving structural annotation by mapping the transcripts to the in-house draft genome of *S. robusta*, 52 cyclins were identified (Suppl. Table S3). A phylogenetic analysis was conducted to assign all *S. robusta* cyclins to the different sub-families. For this, a

neighbor-joining tree based on the MUSCLE alignment of the Cyclin_N domain of *P. tricornutum* and *S. robusta* cyclin sequences was constructed using the MEGA5 software (Figure 4) (Saitou and Nei, 1987, Edgar, 2004). The Poisson correction method was used for distance calculation and 1000 iterations were applied for bootstrap value calculations (Zuckerkandl and Pauling, 1965, Felsenstein, 1985). The expression pattern of the cyclin gene family throughout the life cycle of *S. robusta* can be found in Figure 5 and in Supplementary Figure S1.

A large number of cyclins (52) was identified in *S. robusta*, which may point towards an expansion of this protein family, like in *P. tricornutum* and *T. pseudonana* (Huysman et al., 2010). The largest group of cyclins is the diatom-specific one. Several dsCYCs are expressed early during the cell cycle (G_1 and S phase), as observed for *P. tricornutum* (Suppl. Figure S1) *(Huysman et al., 2010)*. This makes them good candidates to integrate environmental cues into the cell cycle. Also during sexual reproduction, several dsCYCs are highly expressed. These might have acquired novel functions in regulating meiosis to avoid investing in sexual reproduction when favorable conditions are not met.

Two D-type cyclins were found in *S. robusta*. These cyclins are normally associated with G_1-S phase transition, although in plants some D-type cyclins rather seem to be involved in G_2-M transition (Inze and De Veylder, 2006). Similarly, the only D-type cyclin in *P. tricornutum* was highy expressed during late cell cycle and is thus possibly involved in G_2-M transition (Huysman et al., 2010). This is probably also the case for *S. robusta*, as the two D-type cyclins are higher expressed during G_2/M phase (Suppl. Figure S1). On the other hand, their expression is even higher during sexual reproduction than during the vegetative phase (Figure 5).

Furthermore, two L-type cyclins were identified. One of these also shows higher expression during sexual reproduction (Figure 5). For the only H-type cyclin, no clear linkage between its expression and life or cell cycle progression could be found. Additionally, there are five P-type cyclins in *S. robusta* that show different expression patterns (Figure 5).

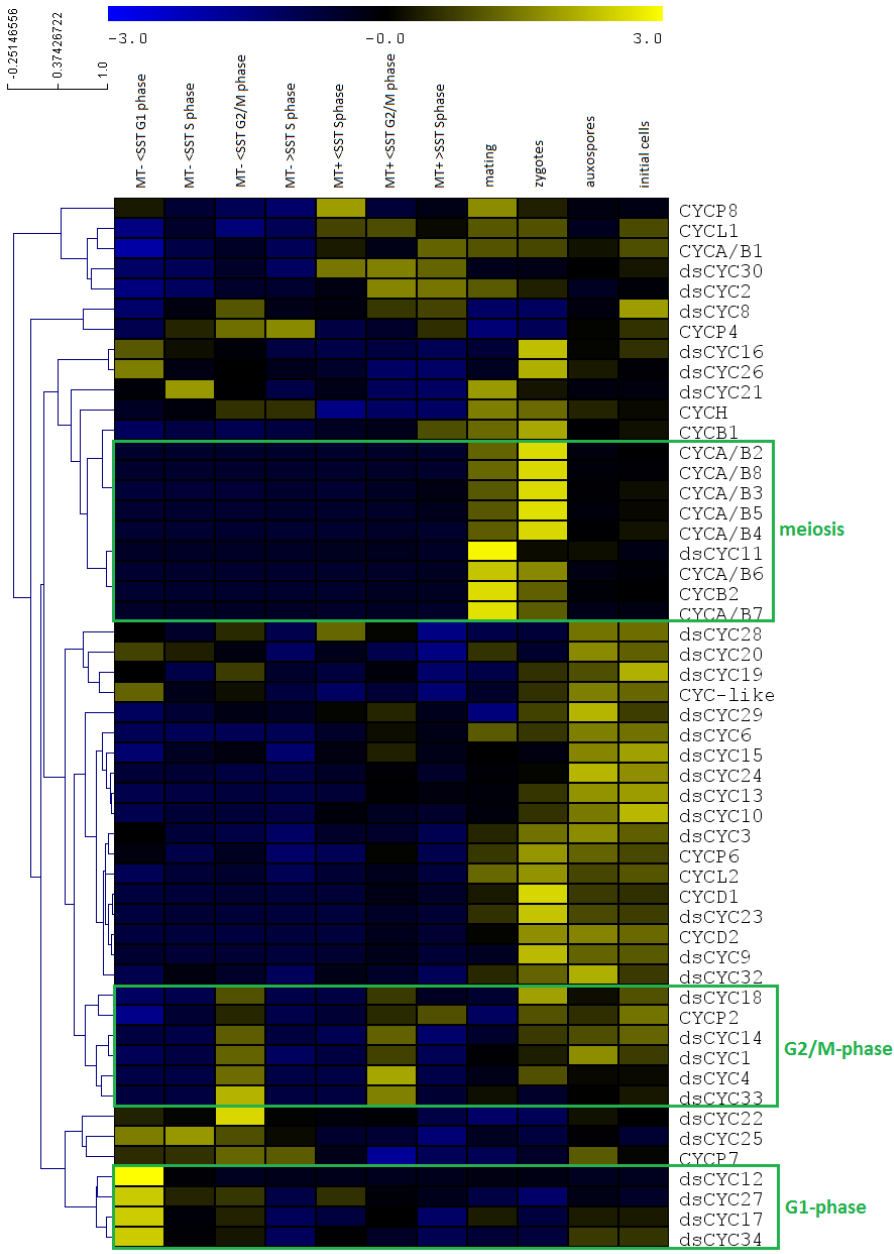


Figure 5: Hierarchical average linkage clustering of normalized cpm values of the *S. robusta* cyclins. Clusters with a clear linkage with certain life/cell cycle stages are indicated in green.

Finally, several A/B-type cyclins were found in *S. robusta*. The *P. tricornutum* CYCA/B1 mRNA accumulated during G_2/M phase, but in the *S. robusta* dataset no

clear link with any cell cycle phase could be seen for its ortholog. On the other hand, a clear linkage between the other A/B-type cyclins and the early sexual stages (mating and zygote formation) could be observed (Figure 5). A/B-type cyclins are known to be involved in the regulation of meiosis in other eukaryotes. For example, B-type cyclins are required for meiotic divisions in budding yeast (Benjamin et al., 2003). CLB1, CLB3, and CLB4 promote progression through the two meiotic divisions and deleting two of these three cyclins results in cells that only execute one meiotic division (Dahmann and Futcher, 1995). Mouse A-type cyclin *CCNA1* and B-type cyclin *CCNB3* are only expressed in the germ-line and *Cnna1* knock-out mice are infertile due to arrest in the first meiotic prophase (Wolgemuth and Roberts, 2010). Two plant A/B-type cyclins, SDS and TAM, are important in male meiosis in flowering plants (Chang et al., 2009). TAM is a A-type cyclin that is necessary for the proper execution of meiotic divisions (Wang et al., 2004) and SDS is involved in the pairing of homologues chromosomes (Azumi et al., 2002).

Silica metabolism and transport

During mitotic cell division and auxosporulation new valves are formed. During vegetative divisions, each daughter cell synthesizes a new hypotheca within the parental cell wall. After auxosporulation, a completely new cell wall is formed. The diatom cell wall is made of amorphous SiO_2 through biomineralization. Silica is taken up by diatoms as orthosilicic acid $Si(OH)_4$ by specific silicic acid transporters (SITs) (Hildebrand et al., 1997). These proteins are found in diverse diatom species, but no homologs are found in other organisms (Thamatrakoln et al., 2006). The shape and the nano- and micropatterns of the diatom cell wall are species-specific, therefore cell wall structure has to be genetically determined. Silica formation takes place in a specialized membrane-bound compartment, called the silica deposition vesicle (SDV) (Drum and Pankratz, 1964). The valves are completely formed within the SDV during cell division and after completion deposited on the cell surface by exocytosis of the SDV (Kroger and Poulsen, 2008). Similarly, girdle bands are formed within another SDV during interphase to allow the increase of cell volume. Microfilaments and microtubules are closely associated with the SDV and were shown to be

important in shaping the silica structure (Schmid, 1979, van de Meene and Pickett-Heaps, 2002).

Additionally, organic compounds that are present inside the SDV are suggested to be involved in silica morphogenesis (Swift and Wheeler, 1992). Several silica-associated proteins were identified that could be components of the SDV (Kroger and Poulsen, 2008). Firstly, frustulins are proteins containing multiple acidic and cysteine-rich domains that seem to be present in all diatoms (Kroger et al., 1996). Frustulins are probably not involved in the silicification process, as they only become associated to the cell wall after silica deposition is complete (van de Poll et al., 1999). Secondly, pleuralins are found in the pleural bands of the epitheca of *Cylindrotheca fusiformis*, but seem to be absent from the *P. tricornutum* and *T. pseudonana* genomes (Kroger et al., 1997, Kroger and Poulsen, 2008). Pleuralins are absent from the SDV and are attached to the newly-deposited pleural bands of the epitheca during interphase or to those of the hypotheca when it is transformed to be the epitheca of one of the daughter cells after cell division (Kroger and Wetherbee, 2000). Thirdly, silaffins are phosphoproteins that appear to be embedded in the silica, because they are hard to extract (Kroger et al., 2002). The first silaffin precursor gene was identified in *C. fusiformis* and consists of a acidic N-terminal part and a highly repetitive basic C-terminal domain (Kroger et al., 1999). During maturation, all repeat units are released as individual peptides, that become heavily post-translationally modified (Kroger et al., 2002). Sil1p-derived peptides and long-chain polyamines (LCPA) are shown to induce the rapid formation of spherical silica particles from silicic acid in vitro and are thus most likely involved in silica deposition in vivo (Kroger et al., 2002, Sumper et al., 2003).

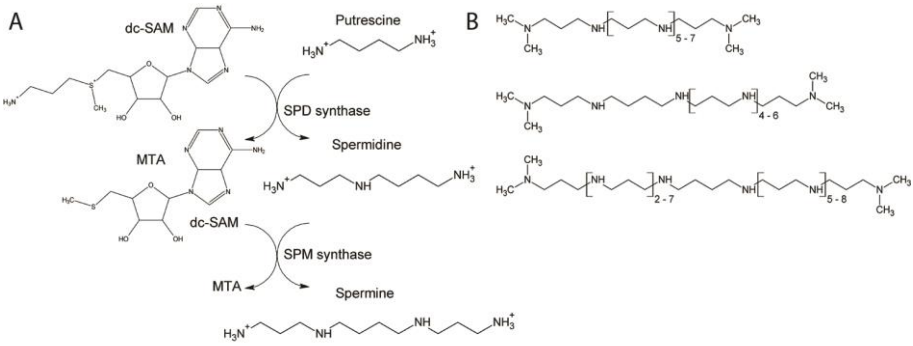


Figure 6: (A) The biosynthesis pathway of spermidine and spermine. SPD synthase transfers a propyleneimine residu from decarboxylated S-adenosylmethionine (dc-SAM) to putrescine to form spermidine and, similary, SPM transfers one from dc-SAM to spermidine to form spermine. MTA = methylthioadenosine. (B) Some examples of diatom LCPAs.

From the chemical structure of the LCPAs, it can be deduced that spermine (SPM) or spermidine (SPD) synthases could be involved in the biosynthesis of LCPAs or the polyamine-type modifications of silaffins (Kroger and Poulsen, 2008). SPD and SPM synthases transfer a propyleneimine residu from decarboxylated S-adenosylmethionine (dc-SAM) to putrescine and spermidine respectively (Wallace et al., 2003). Also the biosynthesis of LCPAs likely involves the transfer of several propyleneimine residues. Several putative SPM/SPD synthases were identified in the genome of *T. pseudonana* (Knott et al., 2007).

The *S. robusta* transcriptome was parsed for SITs, silica-associated proteins and SPM/SPD synthases using reciprocal blast approaches and text-mining in the functional annotation generated by Trapid (Altschul et al., 1990, Van Bel et al., 2013). No pleuralins or silaffins could be identified. Silaffins of *C. fusiformis* and *T. pseudonana* show no sequence conservation. This might reflect the large phylogenetic distance between the two species or can indicate that the arrangement of post-translational modifications, which is less dependent on the exact amino acid sequence, is more important than the tertiary peptide structure (Kroger and Poulsen, 2008). It is thus possible that silaffins are present in *S. robusta*, but that they cannot be identified using simple similarity searches. Since pleuralins are only found in *C. fusiformis* and seem to be lacking in the genome of *P. tricornutum* and *T. pseudonana*

and the transcriptome of *S. robusta*, it is plausible that something similar is going on for pleuralins or that these proteins are specific for a small selection of diatoms.

Four SITs and 35 frustulins could be identified in the *S. robusta* transcriptome using a reciprocal blast approach. The expression profiles for the *S. robusta* SITs and frustulins are shown in Figure 7A and B. Especially SIT3 and a cluster of frustulin genes that are highly expressed during G_2/M phase are of interest, since a new frustule has to be formed after cell division. We can thus hypothesize that SIT3 is strongly expressed late in the cell cycle to take up more silica to be able to construct the new hypothecae of the daughter cells after cell division and that these frustulins are integrated in the cell wall during or shortly after the new valves are formed. Expression of SIT2 is higher during sexual reproduction, mainly during the mating and zygote formation. This transporter could be responsible for silica uptake to make the generation of a new frustule possible during initial cell formation. Also some frustulins are upregulated during the different sexual stages. The other SITs are expressed during S and G_2/M phase and are highest in MT^+-cells above the SST. There is also a cluster of frustulins showing the same expression pattern. Because no biological replicates are available, it is difficult to say whether this is linked to the life cycle stage the cells are in or to the culture conditions in which these large MT^+-samples were grown. Some frustulins are higher expressed during G_1 and S phase. These are potentially incorporated in the girdle bands that are formed during interphase.

Seven putative SPD or SPM synthases could be identified by text-mining the functional annotation of the *S. robusta* transcriptome. Their expression patterns are depicted in Figure 7C. SPM/SPD synthase 1, 3 and 4 are highly expressed during G_2/M phase and during initial cell formation. Since these are the stages where new valves are formed, these synthases could be involved in the biosynthesis of LCPAs.

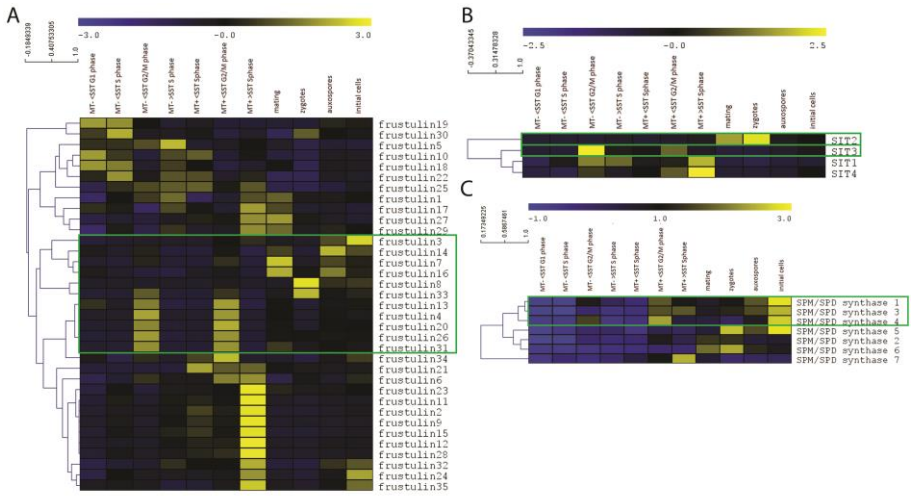


Figure 7: Hierarchical average linkage clustering of normalized cpm values of the *S. robusta* frustulins, SITs and SPM/SPD synthases.

Transcripts involved in sexual reproduction

Because no biological replicates are present in the dataset, differential expression between the vegetative and sexual libraries was examined using the Bioconductor software package edgeR (Robinson et al., 2010), where all vegetative libraries were seen as replicates on the one side and all sexual libraries on the other. This resulted in 2,770 transcripts being overexpressed during vegetative stages and 6,877 during sexual reproduction. On this last category, hierarchical average linkage clustering was conducted to divide the transcripts in four clusters based on their gene expression (Figure 8). For these four clusters, GO enrichment analyses (p-value < 0.05) were conducted using Trapid (Van Bel et al., 2013) to identify the GO terms in which the clusters were enriched in comparison to the complete dataset. GO-module was used to simplify the resulting GO lists (Yang et al., 2011). The resulting key GO terms are represented in Supplementary Table S4.

The mating cluster is enriched in transcripts that are involved in meiosis (DNA metabolism, DNA replication, chromosome organization, nuclear division, ...). Also cell cycle functions are overrepresented, like some A/B-type cyclins and MSH4. Furthermore, some transcripts potentially involved in cell motility (e.g. GO:0000146,

microfilament motor activity) are highly expressed during mating. These proteins are possibly used by MT^+-cells to move towards the diproline-producing MT^--cells. In the zygote cluster, different functions are overrepresented, for instance response to stimuli, intracellular signalling, methylation, chromatin related functions and purine biosynthesis. The transcripts with functions in "response to stimuli" or signalling are possibly involved in signalling between gametes to start gamete fusion and zygote formation. The auxospore cluster is smaller and is only enriched for myo-inositol transport. Inositol can be converted in inositol-phosphates that are known secondary messengers in eukaryotes. It is also a component of phosphatidylinositol (PI) and phosphatidylinositol phosphate (PIP), lipids that are present on the cytosolic side of the cell membrane, where they are involved in signalling and in protein recruitment (Schink et al., 2013). Inositol is also found in the extracellular polymeric substance (EPS) of diatoms (Pierre et al., 2012). The initial cell cluster is the smallest (38 genes) and the only transcript with a predicted function is a peptidase.

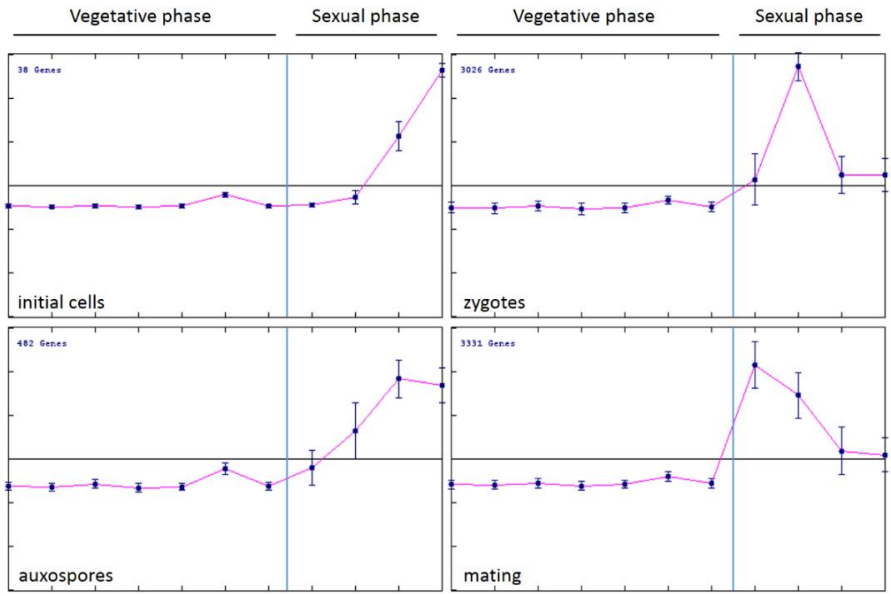


Figure 8: Hierarchical average linkage clustering on the transcripts that are overexpressed during sexual reproduction. The transcripts were divided in four clusters based on their gene expression.

When comparing transcripts that are significantly higher expressed in the vegetative libraries to the complete transcriptome, this transcript set seems enriched in genes with functions in metabolism/biosynthesis (Suppl. Table S4). This probably reflects the fact that during sexual stages no energy is invested in biosynthesis, because a significant amount of energy is needed for meiosis (Ray and Ye, 2013, Dukowic-Schulze et al., 2014).

Conclusions

The RNA-seq dataset reported upon can be used as a resource when investigating the life cycle of pennate diatoms. Since no biological replicates are included in this experiment, additional and targeted experiments would be useful to confirm expression patterns and to identify key regulatory genes. The dataset shows large expression differences between vegetative and sexual stages, where during vegetative phase genes involved in metabolism are overrepresented and during sexual stages, mostly genes involved in DNA replication and meiosis can be found.

We further investigated differential expression for meiotic and mitotic processes and silicification. Parsing the transcriptome for several gene families showed that this dataset is a rich source of sequence information. It can be used to find specific genes of interest and to get a first impression of their expression pattern throughout the life cycle. Firstly, a search for meiosis-related genes in several diatom species could not discriminate between sexual and asexual diatom species. It thus appears that the core meiotic machinery is also conserved in *P. tricornutum* although sexual reproduction has never been observed. Secondly, the cyclin gene family was described. Members of the cyclin family are numerous in *S. robusta,* confirming previous observations in *P. triconutum* and *T. pseudonana* that this protein family is expanded in diatoms. A large group of diatom-specific cyclins was identified, most likely necessary to integrate a wide range of environmental cues with cell cycle regulation and possibly also with the regulation of meiosis. Furthermore, the A/B-type cyclins appear to be involved in meiosis, like they are in other eukaryotes.

Finally, frustulins, several previously unknown SITs and SPM/SPD synthases were identified that are probably involved in silica deposition during vegetative cell division and auxosporulation / initial cell formation.

Literature cited

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403-410.

Azumi, Y., Liu, D.H., Zhao, D.Z., Li, W.X., Wang, G.F., Hu, Y. and Ma, H. (2002) Homolog interaction during meiotic prophase I in Arabidopsis requires the SOLO DANCERS gene encoding a novel cyclin-like protein. *EMBO Journal* **21**, 3081-3095.

Benjamin, K.R., Zhang, C., Shokat, K.M. and Herskowitz, I. (2003) Control of landmark events in meiosis by the CDK Cdc28 and the meiosis-specific kinase Ime2. *Genes & Development* **17**, 1524-1539.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.

Buchanan, B.B., Gruissem, W. and Jones, R.L. (2000) *Biochemistry & molecular biology of plants*, (American Society of Plant Physiologists Rockville).

Chang, L., Ma, H. and Xue, H.W. (2009) Functional conservation of the meiotic genes SDS and RCK in male meiosis in the monocot rice. *Cell Research* **19**, 768-782.

Chepurnov, V.A., Mann, D.G., Vyverman, W., Sabbe, K. and Danielidis, D.B. (2002) Sexual reproduction, mating system, and protoplast dynamics of *Seminavis* (Bacillariophyceae). *Journal of Phycology* **38**, 1004-1019.

Chepurnov, V.A., Mann, D.G., von Dassow, P., Vanormelingen, P., Gillard, J., Inzé, D., Sabbe, K. and Vyverman, W. (2008) In search of new tractable diatoms for experimental biology. *BioEssays* **30**, 692-702.

Chi, P., San Filippo, J., Sehorn, M.G., Petukhova, G.V. and Sung, P. (2007) Bipartite stimulatory action of the Hop2-Mnd1 complex on the Rad51 recombinase. *Genes & Development* **21**, 1747-1757.

Dahmann, C. and Futcher, B. (1995) Specialization of B-Type Cyclins for Mitosis or Meiosis in Saccharomyces-Cerevisiae. *Genetics* **140**, 957-963.

Drebes, G. (1977) Sexuality. in *The biology of diatoms*, Vol. 13 250-283 (Univ of California Press).

Drum, R.W. and Pankratz, H.S. (1964) Post Mitotic Fine Structure of Gomphonema Parvulum. *Journal of Ultrastructure Research* **10**, 217-223.

Dukowic-Schulze, S., Sundararajan, A., Mudge, J., Ramaraj, T., Farmer, A.D., Wang, M.H., Sun, Q., Pillardy, J., Kianian, S., Retzel, E.F., Pawlowski, W.P. and Chen, C.B. (2014) The transcriptome landscape of early maize meiosis. *BMC Plant Biology* **14**.

Eddy, S.R. (2011) Accelerated Profile HMM Searches. *PLoS Computational Biology* **7**.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797.

Falciatore, A., d'Alcalà, M.R., Croot, P. and Bowler, C. (2000) Perception of environmental signals by a marine diatom. *Science* **288**, 2363-2366.

Felsenstein, J. (1985) Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. *Evolution* **39**, 783-791.

Gillard, J., Devos, V., Huysman, M.J.J., De Veylder, L., D'Hondt, S., Martens, C., Vanormelingen, P., Vannerum, K., Sabbe, K., Chepurnov, V.A., Inzé, D., Vuylsteke, M. and Vyverman, W. (2008) Physiological and transcriptomic evidence for a close coupling between chloroplast ontogeny and cell cycle progression in the pennate diatom *Seminavis robusta*. *Plant Physiology* **148**, 1394-1411.

Gillard, J., Frenkel, J., Devos, V., Sabbe, K., Paul, C., Rempt, M., Inze, D., Pohnert, G., Vuylsteke, M. and Vyverman, W. (2013) Metabolomics Enables the Structure Elucidation of a Diatom Sex Pheromone. *Angewandte Chemie-International Edition* **52**, 854-857.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q.D., Chen, Z.H., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644-652.

Guillard, R.R.L. (1975) Culture of phytoplankton for feeding marine invertebrates. in *Culture of marine invertebrate animals* 29-60 (Springer).

Hanson, S.J., Schurko, A.M., Hecox-Lea, B., Welch, D.B.M., Stelzer, C.P. and Logsdon, J.M. (2013) Inventory and Phylogenetic Analysis of Meiotic Genes in Monogonont Rotifers. *Journal of Heredity* **104**, 357-370.

Henry, J.M., Camahort, R., Rice, D.A., Florens, L., Swanson, S.K., Washburn, M.P. and Gerton, J.L. (2006) Mnd1/Hop2 facilitates Dmc1-dependent interhomolog crossover formation in meiosis of budding yeast. *Molecular and Cellular Biology* **26**, 2913-2923.

Hildebrand, M., Volcani, B.E., Gassmann, W. and Schroeder, J.I. (1997) A gene family of silicon transporters. *Nature* **385**, 688-689.

Huysman, M.J.J., Martens, C., Vandepoele, K., Gillard, J., Rayko, E., Heijde, M., Bowler, C., Inzé, D., Van de Peer, Y., De Veylder, L. and Vyverman, W. (2010) Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome Biology* **11**, R17.

Huysman, M.J.J., Fortunato, A.E., Matthijs, M., Costa, B.S., Vanderhaeghen, R., Van den Daele, H., Sachse, M., Inze, D., Bowler, C., Kroth, P.G., Wilhelm, C., Falciatore, A., Vyverman, W. and De Veylder, L. (2013) AUREOCHROME1a-Mediated Induction of the Diatom-Specific Cyclin dsCYC2 Controls the

Onset of Cell Division in Diatoms (Phaeodactylum tricornutum). *Plant Cell* **25**, 215-228.

Inze, D. and De Veylder, L. (2006) Cell cycle regulation in plant development. *Annual Review of Genetics* **40**, 77-105.

Keeney, S., Giroux, C.N. and Kleckner, N. (1997) Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* **88**, 375-384.

Knott, J.M., Romer, P. and Sumper, M. (2007) Putative spermine synthases from Thalassiosira pseudonana and Arabidopsis thaliana synthesize thermospermine rather than spermine. *FEBS Letters* **581**, 3081-3086.

Kroger, N., Bergsdorf, C. and Sumper, M. (1996) Frustulins: Domain conservation in a protein family associated with diatom cell walls. *European Journal of Biochemistry* **239**, 259-264.

Kroger, N., Lehmann, G., Rachel, R. and Sumper, M. (1997) Characterization of a 200-kDa diatom protein that is specifically associated with a silica-based substructure of the cell wall. *European Journal of Biochemistry* **250**, 99-105.

Kroger, N., Deutzmann, R. and Sumper, M. (1999) Polycationic peptides from diatom biosilica that direct silica nanosphere formation. *Science* **286**, 1129-1132.

Kroger, N. and Wetherbee, R. (2000) Pleuralins are involved in theca differentiation in the diatom Cylindrotheca fusiformis. *Protist* **151**, 263-273.

Kroger, N., Lorenz, S., Brunner, E. and Sumper, M. (2002) Self-assembly of highly phosphorylated silaffins and their function in biosilica morphogenesis. *Science* **298**, 584-586.

Kroger, N. and Poulsen, N. (2008) Diatoms-From Cell Wall Biogenesis to Nanotechnology. *Annual Review of Genetics* **42**, 83-107.

Luke-Glaser, S., Luke, B., Grossi, S. and Constantinou, A. (2010) FANCM regulates DNA chain elongation and is stabilized by S-phase checkpoint signalling. *EMBO Journal* **29**, 795-805.

Lynn, A., Soucek, R. and Borner, G.V. (2007) ZMM proteins during meiosis: Crossover artists at work. *Chromosome Research* **15**, 591-605.

Malik, S.B., Ramesh, M.A., Hulstrand, A.M. and Logsdon, J.M. (2007) Protist homologs of the meiotic Spo11 gene and topoisomerase VI reveal an evolutionary history of gene duplication and lineage-specific loss. *Molecular Biology and Evolution* **24**, 2827-2841.

Malik, S.B., Pightling, A.W., Stefaniak, L.M., Schurko, A.M. and Logsdon, J.M. (2008) An Expanded Inventory of Conserved Meiotic Genes Provides Evidence for Sex in Trichomonas vaginalis. *PLoS ONE* **3**.

Nishant, K.T., Chen, C., Shinohara, M., Shinohara, A. and Alani, E. (2010) Genetic Analysis of Baker's Yeast Msh4-Msh5 Reveals a Threshold Crossover Level for Meiotic Viability. *PLoS Genetics* **6**.

Olson, R.J., Vaulot, D. and Chisholm, S.W. (1986) Effects of Environmental Stresses on the Cell-Cycle of 2 Marine-Phytoplankton Species. *Plant Physiology* **80**, 918-925.

Pierre, G., Graber, M., Rafiliposon, B.A., Dupuy, C., Orvain, F., De Crignis, M. and Maugard, T. (2012) Biochemical Composition and Changes of Extracellular Polysaccharides (ECPS) Produced during Microphytobenthic Biofilm Development (Marennes-Ol,ron, France). *Microbial Ecology* **63**, 157-169.

Ray, D. and Ye, P. (2013) Characterization of the Metabolic Requirements in Yeast Meiosis. *PLoS ONE* **8**.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140.

Round, F.E., Crawford, R.M. and Mann, D.G. (1990) *The diatoms: biology & morphology of the genera*, (Cambridge University Press).

Saitou, N. and Nei, M. (1987) The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution* **4**, 406-425.

Sapriel, G., Quinet, M., Heijde, M., Jourdren, L., Tanty, V., Luo, G.Z., Le Crom, S. and Lopez, P.J. (2009) Genome-Wide Transcriptome Analyses of Silicon Metabolism in Phaeodactylum tricornutum Reveal the Multilevel Regulation of Silicic Acid Transporters. *PLoS ONE* **4**.

Schink, K.O., Raiborg, C. and Stenmark, H. (2013) Phosphatidylinositol 3-phosphate, a lipid that regulates membrane dynamics, protein sorting and cell signalling. *BioEssays* **35**, 900-912.

Schmid, A.M.M. (1979) Wall Morphogenesis in Diatoms - Role of Microtubules during Pattern Formation. *European Journal of Cell Biology* **20**, 125-125.

Snowden, T., Acharya, S., Butz, C., Berardini, M. and Fishel, R. (2004) hMSH4-hMSH5 recognizes Holliday junctions and forms a meiosis-specific sliding clamp that embraces homologous chromosomes. *Molecular Cell* **15**, 437-451.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273-3297.

Sumper, M., Lorenz, S. and Brunner, E. (2003) Biomimetic control of size in the polyamine-directed formation of silica nanospheres. *Angewandte Chemie-International Edition* **42**, 5192-5195.

Swift, D.M. and Wheeler, A.P. (1992) Evidence of an Organic Matrix from Diatom Biosilica. *Journal of Phycology* **28**, 202-209.

Thamatrakoln, K., Alverson, A.J. and Hildebrand, M. (2006) Comparative sequence analysis of diatom silicon transporters: Towards a molecular model of silicon transport. *Journal of Phycology* **42**, 10-10.

Valenzuela, J., Mazurie, A., Carlson, R.P., Gerlach, R., Cooksey, K.E., Peyton, B.M. and Fields, M.W. (2012) Potential role of multiple carbon fixation pathways during lipid accumulation in *Phaeodactylum tricornutum*. *Biotechnology for Biofuels* **5**, 40.

Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y. and Vandepoele, K. (2013) TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. *Genome Biology* **14**.

van de Meene, A.M.L. and Pickett-Heaps, J.D. (2002) Valve morphogenesis in the centric diatom Proboscia alata Sundstrom. *Journal of Phycology* **38**, 351-363.

van de Poll, W.H., Vrieling, E.G. and Gieskes, W.W.C. (1999) Location and expression of frustulins in the pennate diatoms Cylindrotheca fusiformis, Navicula pelliculosa, and Navicula salinarum (Bacillariophyceae). *Journal of Phycology* **35**, 1044-1053.

Wallace, H.M., Fraser, A.V. and Hughes, A. (2003) A perspective of polyamine metabolism. *Biochemical Journal* **376**, 1-14.

Wang, Y.X., Magnard, J.L., McCormick, S. and Yang, M. (2004) Progression through meiosis I and meiosis II in Arabidopsis anthers is regulated by an A-type cyclin predominately expressed in prophase I. *Plant Physiology* **136**, 4127-4135.

Watanabe, Y. and Nurse, P. (1999) Cohesin Rec8 is required for reductional chromosome segregation at meiosis. *Nature* **400**, 461-464.

Wolgemuth, D.J. and Roberts, S.S. (2010) Regulating mitosis and meiosis in the male germ line: critical functions for cyclins. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**, 1653-1662.

Yang, X.A., Li, J.R., Lee, Y. and Lussier, Y.A. (2011) GO-Module: functional synthesis and improved interpretation of Gene Ontology patterns. *Bioinformatics* **27**, 1444-1446.

Zuckerkandl, E. and Pauling, L. (1965) Evolutionary divergence and convergence in proteins. *Evolving genes and proteins* **97**, 97-166.

Supplementary data

Materials & Methods

Trimmomatic parameters

Task	Command	Parameters
Nr. of threads	-threads 20	
Phred score	-phred64	
Remove Illumina adapters	ILLUMINACLIP	fastaWithAdaptersEtc = adapters.fa seed mismatches = 2 palindrome clip threshold = 40 simple clip threshold = 15
Remove low quality bases from the beginning	LEADING	Quality = 5
Remove low quality bases from the end	TRAILING	Quality = 5
sliding window trimming	SLIDINGWINDOW	Window size = 5 Required quality = 20
Minimal length	MINLEN	Length = 100

normalize_by_kmer_coverage.pl parameters

Task	Command	Parameter
Input file type is fastq	--seqType	fq
Maximum GB memory for k-mer counting by Jellyfish	--JM	240G
Maximum coverage for reads	--max_cov	30
Strand-specific RNA-Seq read orientation	--SS_lib_type	RF
Number of threads for Jellyfish	--JELLY_CPU	24

Trinity.pl

Task	Command	Parameter
Input file type is fastq	--seqType	fq
Maximum GB memory for k-mer counting by Jellyfish	--JM	220G
Number of CPUs to use by Inchworm	--inchworm_cpu	22
java initial hap space settings for Butterfly	--bflyHeapSpaceInit	22G
java max heap space setting for butterfly	--bflyHeapSpaceMax	220G
Calculate CPUs based on 80% of max_memory divided by maxbflyHeapSpaceMax	--bflyCalculateCPU	
Number of CPUs to use	--CPU	22
Strand-specific RNA-Seq read orientation	--SS_lib_type	RF
min count for K-mers to be assembled by Inchworm	--min_kmer_cov	2

Figure S1: Hierarchical average linkage clustering of normalized cpm values of the *S. robusta* cyclins during the cell cycle.

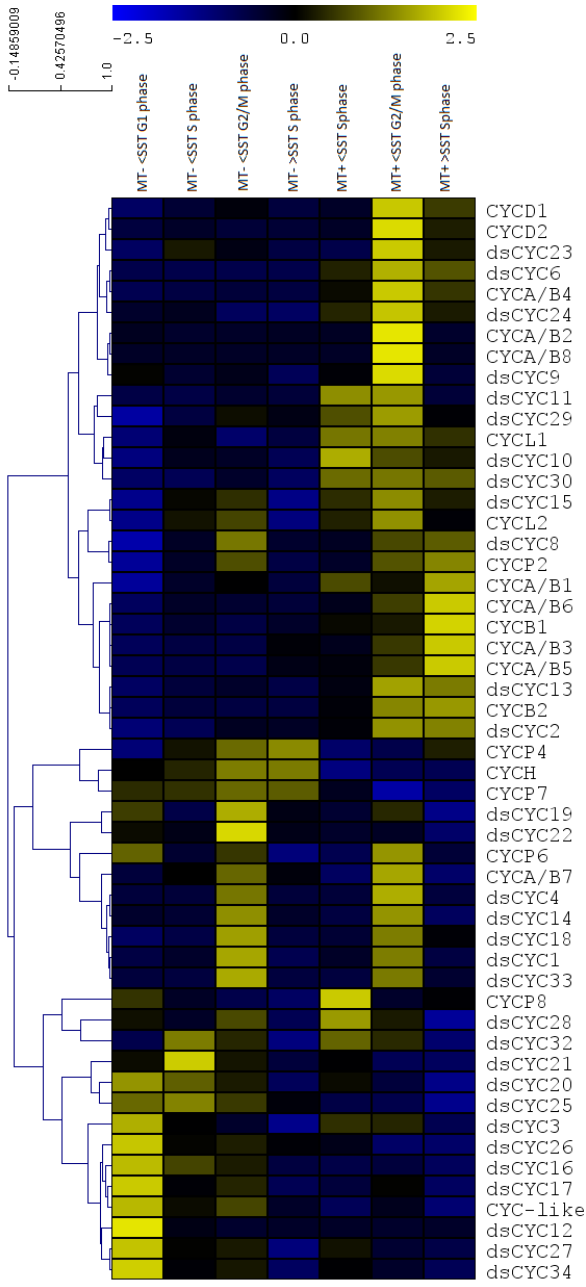


Table S1: Function of the meiotic genes included in this study. Gene functions are taken from Malik et al. (2008). Genes marked with * are meiosis-specific genes.

Protein	Role in meiosis
Dmc1*	Meiotic homolog of Rad51, Binds to ssDNA end of DSBs and involved in inter-homologous recombination
Hop1 *	Binding to DSBs, component of lateral and axial synaptonemal complex
Hop2*	Homology search together with Mnd1, works in Dmc1 dependent homology search pathway downstream of Rad51
Mcm2-7	Mcm2-7 form hexamer and are involved in DNA replication
Mcm8-9	Mcm8 and Mcm9 are involved in meiotic recombination
Mer3	Meiosis-specific DEAD-box helicase that promotes Holliday junction resolution with crossover interference together with ZMM proteins, including Msh4 and Msh5
Mlh1	DNA mismatch repair protein, forms heterodimers with Mlh2, Mlh3 and Pms1, interacts with Msh2/Msh6 and Msh4/Msh5
Mlh2	DNA mismatch repair protein, forms heterodimer with Mlh1
Mlh3	Forms heterodimer with Mlh1, interacts with Msh4/Msh5 to promote meiotic crossovers
Mnd1*	Together with Hop2 works in homology searching and also required in stable DNA heteroduplex
Mre11	3'–5' dsDNA exonuclease and ssDNA endonuclease; trims broken DNA ends after DSBs and hairpins
Msh2	Forms heterodimer with Msh3 or Msh6, works in DNA mismatch repair
Msh4*	Forms heterodimer with Msh5, together with Mlh1/Mhl3 heterodimer directs Holliday junction resolution with crossover interference
Msh5*	Forms heterodimer with Msh4 and interacts with Mlh1/Mhl3 heterodimer to direct Holliday junction resolution
Msh6	Forms heterodimer with Msh2, works in DNA mismatch repair
Pds5	Involved in maintenance of sister chromatid cohesion in late prophase
Pms1	Forms heterodimer with Mlh1, involved in DNA mismatch repair
Rad1	5'-3' endonuclease, required in meiotic crossing over, functions during nucleotide excision repair
Rad21	Holds Smc1 and Smc3 together and thus holding sister chromatids together during meiosis and mitosis
Rad50	DNA binding ATPase, holds broken DNA strands while Mre11 trims DSBs
Protein	**Role in meiosis**

Rad51	Mediates homologous pairing and strand invasion, involved in DNA repair mechanisms in mitosis and meiosis
Rad52	Binds to ssDNA and initiates homologous recombination, stimulates Rad51 mediated strand invasion
Rec8*	Meiotic homolog of Rad21, involved in holding sister chromatids together during meiotic recombination
Scc3	Interacts with cohesin complex Smc1-Smc3 and Rad21/Rec8 and helps in holding cohesin ring together
Smc1	Part of sister chromatids cohesin subunit, forms heterodimer with Smc3
Smc2	Essential for chromosome assembly and segregation, part of core condensing subunit that forms heterodimer with Smc4
Smc3	Forms heterodimer with Smc1 and acts as sister chromatids cohesin subunit
Smc4	Part of core condesin subunit, forms heterodiamer with Smc2, essential for chromosome assembly and segregation
Smc5	Involved in DNA repair and checkpoint response, Functions as a heterodimer with Smc6 (Rad18)
Smc6 (RAD18)	Important in post-replication DNA repair, forms heterodimer with Smc5 and binds to ssDNA
Spo11-2*	Creates double-strand breaks (DSBs) in homologous chromosomes in meiotic recombination
Spo11-3	Creates DSBs in homologous chromosomes

Table S2: Accession numbers for the proteins involved in meiosis analyzed in this study. Arabidopsis thaliana meiotic proteins were used as query sequence Accession numbers in bold are for Saccharomyces cerevisiae proteins used as a query sequence whenever the specific query gene was not found in A. thaliana. Protein IDs are given for the diatom genomes available at the JGI portal, scaffold IDs are given for the Pseudo-nitzschia multistriata genome and transcripts IDs for Seminavis robusta. Genes marked with * are genes that do not have known functions outside of meiosis.

(see next pages)

Protein name	Accession numbers of proteins used as query	*Thaps*	*Phatr*	*Fracy*	*Pnmse*	*Pnmst*	*Semro*
DNA replication and chromosome maintenance							
Mcm2	NP_001185154.1	29936	18622	204899	209470	PsnmuV1_4_scaffold_29	Semro_comp78811_c0_seq1
Mcm3	Q9FL33.1	34975	51597	264318	318351	PsnmuV1_4_scaffold_106	Semro_comp50104_c0_seq1
Mcm4	NP_179236.3	269123	51412	146869	203268	PsnmuV1_4_scaffold_387	Semro_comp82592_c0_seq1
Mcm5	NP_001189521.1	31609	11490	224321	255529	PsnmuV1_4_scaffold_85	Semro_comp83065_c1_seq1
Mcm6	AED95141.1	26545	468	193082	321321	PsnmuV1_4_scaffold_15	Semro_comp61600_c0_seq1
Mcm7	P43299.2	262526	13243	184349	243980	PsnmuV1_4_scaffold_70	Semro_comp70058_c0_seq2
Mcm8	NP_187577.1	261512	52561	189062	213178	PsnmuV1_4_scaffold_32	Semro_comp79168_c0_seq3
Mcm9	NP_179021.3	37362	981	156569	183315	PsnmuV1_4_scaffold_269	Semro_comp59174_c0_seq1
Mcm10	NP_179694.2	NF	NF	NF	NF	NF	NF
Smc1	AEE79265.1	35499	25506	212269	162817	PsnmuV1_4_scaffold_81	Semro_comp83927_c0_seq1
Smc2	NP_201047.1	1393	30352	210755	191984	PsnmuV1_4_scaffold_52	Semro_comp61213_c0_seq1
Smc3	NP_180285.4	259020	52607	208027	251818	PsnmuV1_4_scaffold_396	Semro_comp78328_c0_seq1
Smc4	AED95695.1	42365	44165	212991	144962	PsnmuV1_4_scaffold_172	Semro_comp76089_c0_seq1
Smc5	AED92224.1	9851 23096	54192	193562	286374	PsnmuV1_4_scaffold_6	Semro_comp64598_c0_seq1
Smc6 (rad18)	NP_196383.1	1743	36853	168857	165557	PsnmuV1_4_scaffold_47	Semro_comp65344_c0_seq1
Pds5	NP_177883.5	5929	1590	136077	203285	PsnmuV1_4_scaffold_162	Semro_comp82484_c0_seq6
Scc3	AEC10920.1	NF	NF	NF	NF	NF	NF
Rec8*	NP_196168.1	NF	NF	NF	NF	NF	NF
Rad21	NP_851110.1	8557	44595	245879	324402	PsnmuV1_4_scaffold_343	Semro_comp80503_c0_seq6

Protein name	Accession numbers of proteins used as query	*Thaps*	*Phatr*	*Fracy*	*Pnmse*	*Pnmst*	*Semro*
Double-strand break formation							
Spo11_1*	AEE75304.1	NF	NF	NF	NF	NF	NF
Spo11_2*	AEE34178.1	263510	36531	242364	156625	PsnmuV1_4_scaffold_67	Semro_comp74200_c0_seq1
Spo11_3	NP_195902.1	42646	24838	239125	251788	PsnmuV1_4_scaffold_4	Semro_comp59497_c0_seq2
Crossover regulation							
Dmc1*	AAC49617.1	NF	NF	NF	NF	NF	NF
Hop1*	NP_172691.1	NF	NF	NF	NF	NF	NF
Hop2*	CAF28783.1	NF	NF	NF	NF	NF	NF
Mer3*	AAX14498.1	11979	39994	239915	285411	PsnmuV1.4_scaffold_45	Semro_comp60890_c0_seq1
Mnd1*	NP_194646.2	25513	54296	273989	295346	PsnmuV1_4_scaffold_40	Semro_comp20014_c0_seq1
Msh4*	AAT70180.1	261368	51916	144820	259109	PsnmuV1_4_scaffold_8	Semro_comp57561_c0_seq2
Msh5*	NP_188683.3	16039	52173	149505	183820	PsnmuV1_4_scaffold_154	Semro_comp80580_c0_seq6
Red1/Asy3*	AEC10782.1	NF	NF	NF	NF	NF	NF
Zip1*	AEE30217.1	NF	NF	NF	NF	NF	NF
Asy1	AEE34638.1	NF	NF	NF	NF	NF	NF
Asy2	AEE86019.1	NF	NF	NF	NF	NF	NF
Asy3	AEC10782.1	NF	NF	NF	NF	NF	NF
DNA damage sensing and response							
Rad50	AEC08614.1	9195	51876	243939	320939	PsnmuV1_4_scaffold_1	Semro_comp61512_c0_seq1
Mre11	NP_200237.1	34332	54699	275781	233741	PsnmuV1_4_scaffold_44	Semro_comp82091_c0_seq3
Xrs2/Nbs1	ABA54896.1	NF	NF	NF	NF	NF	NF

Protein name	Accession numbers of proteins used as query	*Thaps*	*Phatr*	*Fracy*	*Pnmse*	*Pnmst*	*Semro*
Double-strand break repair (recombinational repair)							
Rad51 A	BAE99388.1					PsnmuV1_4_scaffold_44	Semro_comp76648_c0_seq1
Rad51 A1	BAE99388.1					PsnmuV1_4_scaffold_268	Semro_comp76648_c0_seq1
Rad51 B	NP_180423.3					PsnmuV1_4_scaffold_634	Semro_comp71219_c0_seq1
Rad51 C	CAC14091.1	54137	257784	201530	29459	PsnmuV1_4_scaffold_61	Semro_comp71710_c0_seq3
Rad51 D	NP_001077479.1	NF	NF	NF	NF	NF	NF
Xrcc2	NP_201257.2	NF	NF	NF	NF	NF	NF
Xrcc3	NP_200554.1					PsnmuV1.4 scaffold 108	Semro_comp70556_c0_seq2
Rec-A	BAE99388.1	267595	51425	186275	166360	PsnmuV1_4_scaffold_30	Semro_comp77000_c0_seq1
Rad52	**CAA86623.1**	25447	49083	238228	50181	PsnmuV1_4_scaffold_46	Semro_comp79910_c0_seq1
Rad1	Q9LKI5.2	22869	30908	208467	230429	PsnmuV1_4_scaffold_358	Semro_comp82187_c0_seq2
Msh2	AEE76112.1	32661	19604	159571	153636	PsnmuV1_4_scaffold_41	Semro_comp80478_c0_seq1
Msh6	NP_001190656.1	261781	53969	212924	190397	PsnmuV1_4_scaffold_42	Semro_comp80580_c0_seq6
Mlh1	NP_567345.2	263509	54331	136590	257081	PsnmuV1_4_scaffold_98	Semro_comp75421_c0_seq1
Mlh2	**NP 013135.1**	NF	NF	NF	NF	NF	NF
Mlh3	NP_195277.5	NF	NF	NF	NF	NF	NF
Pms1	AAM00563.1	264783	29228	248102	242883	PsnmuV1_4_scaffold_81	Semro_comp73186_c0_seq1
Mms4/Eme1	**AAF06816.1**	NF	NF	NF	NF	NF	NF
Mus81	NP_194816.2	NF	36625	241086	63674	PsnmuV1_4_scaffold_59	Semro_comp84506_c0_seq2
Fancm	NP_001185141.1	11922	47619	248113	68428	PsnmuV1_4_scaffold_119	Semro_comp74927_c0_seq1
Fen1	AED93576.1	269347	48638	206746	260195	PsnmuV1_4_scaffold_11	Semro_comp51200_c0_seq1
Exo1	Q8L6Z7.2	4742	48206	261553	110816	PsnmuV1_4_scaffold_311	Semro_comp61722_c0_seq1
DNA2	NP_001184943.1					PsnmuV1_4_scaffold_162	Semro_comp83726_c0_seq1
BRCA1	AAO39850.1						Semro_comp80973_c0_seq2
BRCA2	AEE81814.1						Semro_comp82255_c0_seq4

Table S3: Overview of the 52 *S. robusta* cyclins. 51 of them contain a N-terminal cyclin domain (N-CYC), 25 a C-terminal cyclin domain, 8 contain a D-box and no KEN-boxes are found. 10 belong to the A/B-type cyclins, 5 are P-type cyclins, 1 H-type, 2 L-type, 2-D-type and 31 dsCYCs.

transcript name	N-CYC	C-CYC	D-box	subfamily	name
comp79535_c0_seq1_ed		yes			CYC-like
comp77689_c0_seq1	yes	yes		CYCA-B	CYCA/B1
comp19332_c0_seq1	yes	yes		CYCA-B	CYCA/B2
comp30408_c0_seq1	yes	yes		CYCA-B	CYCA/B3
comp70144_c0_seq1	yes	yes	27-30	CYCA-B	CYCA/B4
comp54365_c0_seq1	yes	yes	141-144	CYCA-B	CYCA/B5
merged5	yes	yes	8-11, 40-43	CYCA-B	CYCA/B6
comp72697_c1_seq1_ed	yes	yes		CYCA-B	CYCA/B7
merged3	yes	yes	18-21	CYCA-B	CYCA/B8
comp53942_c0_seq1	yes	yes		CYCA-B	CYCB1
comp69413_c0_seq1_ed	yes		407-410	CYCA-B	CYCB2
merged7	yes	yes		CYCD	CYCD1
comp66400_c0_seq1_ed	yes	yes		CYCD	CYCD2
comp65258_c0_seq1	yes			CYCH	CYCH
comp32633_c0_seq2_ed	yes			CYCL	CYCL1
comp66354_c0_seq1_ed	yes			CYCL	CYCL2
merged1_ed	yes			CYCP	CYCP2
comp80806_c0_seq1	yes		280-283	CYCP	CYCP4
comp65549_c0_seq1_ed	yes			CYCP	CYCP6
comp82890_c0_seq1	yes			CYCP	CYCP7
merged6	yes			CYCP	CYCP8
comp60276_c0_seq1_ed	yes			dsCYC	dsCYC1
comp84555_c0_seq4	yes			dsCYC	dsCYC2
comp52442_c0_seq2	yes	yes		dsCYC	dsCYC3
comp26705_c0_seq1	yes	yes		dsCYC	dsCYC4
comp19907_c0_seq1_ed	yes			dsCYC	dsCYC6
comp69123_c0_seq1_ed	yes			dsCYC	dsCYC8
comp79227_c0_seq1	yes			dsCYC	dsCYC9
comp70608_c0_seq1	yes	yes		dsCYC	dsCYC10
comp17920_c0_seq1	yes			dsCYC	dsCYC11
comp27005_c0_seq1	yes			dsCYC	dsCYC12
comp42619_c0_seq1	yes	yes	241-244	dsCYC	dsCYC13

transcript name	N-CYC	C-CYC	D-box	subfamily	name
comp49461_c0_seq2	yes			dsCYC	dsCYC14
comp50661_c0_seq1_ed	yes			dsCYC	dsCYC15
comp59459_c0_seq1_ed	yes	yes		dsCYC	dsCYC16
comp59640_c0_seq1_ed	yes			dsCYC	dsCYC17
comp61049_c0_seq1_ed	yes	yes		dsCYC	dsCYC18
comp64264_c0_seq1	yes			dsCYC	dsCYC19
comp65052_c0_seq1_ed	yes			dsCYC	dsCYC20
comp68723_c0_seq1_ed	yes	yes		dsCYC	dsCYC21
comp68771_c0_seq1	yes	yes		dsCYC	dsCYC22
comp69313_c0_seq1	yes	yes		dsCYC	dsCYC23
comp70564_c0_seq1_ed	yes	yes		dsCYC	dsCYC24
comp70628_c0_seq2	yes	yes		dsCYC	dsCYC25
comp71189_c0_seq1	yes			dsCYC	dsCYC26
comp72024_c0_seq1	yes			dsCYC	dsCYC27
comp74326_c0_seq1_ed	yes	yes		dsCYC	dsCYC28
comp79950_c0_seq2	yes			dsCYC	dsCYC29
comp81466_c0_seq1_ed	yes		271-274	dsCYC	dsCYC30
comp85193_c0_seq1_ed	yes			dsCYC	dsCYC32
merged2	yes			dsCYC	dsCYC33
merged4_ed	yes			dsCYC	dsCYC34

Table S4: GO enrichment analyses (p-value < 0.05) were conducted using TRAPID (Van Bel et al., 2013) to identify the GO terms where the clusters are enriched in compared to the complete dataset. GO-module (Yang et al., 2011) was used to simplify the resulting GO lists. The resulting key GO terms are represented for the four clusters (mating, zygote formation, auxosporulation and initial cell) and for the transcripts that are significantly higher expressed in the vegetative phase compared to the sexual stage.

Key GO term	p-value	GO description
Mating		
GO:0044085	3.45E-03	cellular component biogenesis
GO:0005622	2.36E-08	intracellular
GO:0005488	4.85E-05	binding
GO:0043170	4.06E-02	macromolecule metabolic process
GO:0008094	5.11E-35	DNA-dependent ATPase activity
GO:0003677	3.38E-62	DNA binding
GO:0007059	6.97E-03	chromosome segregation
GO:0008076	1.34E-02	voltage-gated potassium channel complex
GO:0065004	6.06E-11	protein-DNA complex assembly
GO:0019948	2.93E-02	SUMO activating enzyme activity
GO:0030983	2.74E-05	mismatched DNA binding
GO:0006913	5.31E-03	nucleocytoplasmic transport
GO:0043234	1.84E-07	protein complex
GO:0006338	4.63E-03	chromatin remodeling
GO:0006259	9.85E-63	DNA metabolic process
GO:0004221	2.88E-11	ubiquitin thiolesterase activity
GO:0006139	6.93E-17	nucleobase-containing compound metabolic process
GO:0004386	1.21E-23	helicase activity
GO:0051276	1.63E-17	chromosome organization
GO:0032446	1.95E-03	protein modification by small protein conjugation
GO:0044428	3.47E-14	nuclear part
GO:0006268	8.03E-04	DNA unwinding involved in DNA replication
GO:0009262	3.31E-03	deoxyribonucleotide metabolic process
GO:0003887	3.35E-07	DNA-directed DNA polymerase activity
GO:0000146	3.93E-03	microfilament motor activity
GO:0009186	4.58E-04	deoxyribonucleoside diphosphate metabolic process
GO:0031497	1.56E-07	chromatin assembly
GO:0051726	2.93E-02	regulation of cell cycle

Key GO term	p-value	GO description
GO:0006511	3.30E-06	ubiquitin-dependent protein catabolic process
GO:0016881	4.49E-03	acid-amino acid ligase activity
GO:0004842	4.72E-03	ubiquitin-protein transferase activity
GO:0015462	5.52E-05	protein-transmembrane transporting ATPase activity
GO:0003896	5.57E-03	DNA primase activity
GO:0044257	7.80E-06	cellular protein catabolic process
GO:0016787	1.93E-22	hydrolase activity
GO:0044427	6.20E-18	chromosomal part
GO:0004798	1.60E-02	thymidylate kinase activity
GO:0005694	1.23E-32	chromosome
GO:0003916	1.07E-03	DNA topoisomerase activity
GO:0032993	5.13E-09	protein-DNA complex
GO:0000279	1.73E-03	M phase
GO:0006807	1.83E-09	nitrogen compound metabolic process
GO:0008270	4.84E-03	zinc ion binding
GO:0005634	6.22E-26	nucleus
GO:0007126	1.86E-03	meiotic nuclear division
GO:0022402	2.35E-04	cell cycle process
GO:0007049	3.29E-04	cell cycle
Zygote formation		
GO:0005626	1.27E-03	insoluble fraction
GO:0010150	4.06E-09	leaf senescence
GO:0015935	3.29E-02	small ribosomal subunit
GO:0042598	7.59E-04	vesicular fraction
GO:0009611	8.39E-05	response to wounding
GO:0005792	7.59E-04	microsome
GO:0009605	2.65E-04	response to external stimulus
GO:0009751	1.61E-05	response to salicylic acid
GO:0080095	9.04E-04	phosphatidylethanolamine-sterol O-acyltransferase activity
GO:0000922	3.45E-04	spindle pole
GO:0006333	3.22E-02	chromatin assembly or disassembly
GO:0080096	9.04E-04	phosphatidate-sterol O-acyltransferase activity
GO:0000226	5.30E-03	microtubule cytoskeleton organization
GO:0004525	4.15E-03	ribonuclease III activity
GO:0000785	1.36E-02	chromatin

Key GO term	p-value	GO description
GO:0006306	8.39E-05	DNA methylation
GO:0003676	1.18E-08	nucleic acid binding
GO:0031407	9.81E-09	oxylipin metabolic process
GO:0003999	4.68E-05	adenine phosphoribosyltransferase activity
GO:0005315	4.15E-03	inorganic phosphate transmembrane transporter activity
GO:0004854	1.07E-02	xanthine dehydrogenase activity
GO:0000267	1.27E-03	cell fraction
GO:0034434	1.15E-03	sterol esterification
GO:0005815	1.01E-03	microtubule organizing center
GO:0015114	4.49E-02	phosphate ion transmembrane transporter activity
GO:0016629	2.82E-09	12-oxophytodienoate reductase activity
GO:0006730	3.46E-02	one-carbon metabolic process
GO:0010149	4.06E-09	senescence
GO:0016127	1.15E-03	sterol catabolic process
GO:0010181	1.40E-02	FMN binding
GO:0032993	3.18E-03	protein-DNA complex
GO:0043232	8.28E-04	intracellular non-membrane-bounded organelle
GO:0004386	1.34E-08	helicase activity
GO:0006144	2.25E-06	purine nucleobase metabolic process
GO:0008270	8.04E-05	zinc ion binding
GO:0005624	1.27E-03	membrane fraction
GO:0007186	4.54E-02	G-protein coupled receptor signaling pathway
GO:0003886	6.59E-07	DNA (cytosine-5-)-methyltransferase activity
Auxosporulation		
GO:0005366	9.12E-03	myo-inositol:proton symporter activity
GO:0015798	1.34E-02	myo-inositol transport
Initial cells		
GO:0004222	4.79E-02	metalloendopeptidase activity
GO:0005576	5.70E-03	extracellular region
Vegetative phase		
GO:0009065	4.04E-03	glutamine family amino acid catabolic process
GO:0009765	8.92E-10	photosynthesis, light harvesting
GO:0006614	3.11E-03	SRP-dependent cotranslational protein targeting to membrane
GO:0070972	3.11E-03	protein localization to endoplasmic reticulum
GO:0071539	3.11E-03	protein localization to centrosome

Key GO term	p-value	GO description
GO:0046148	1.13E-14	pigment biosynthetic process
GO:0006779	1.02E-12	porphyrin-containing compound biosynthetic process
GO:0015979	1.15E-10	photosynthesis
GO:0016114	1.38E-06	terpenoid biosynthetic process
GO:0006006	3.75E-11	glucose metabolic process
GO:0044282	2.11E-09	small molecule catabolic process
GO:0008610	9.36E-13	lipid biosynthetic process
GO:0019748	7.69E-04	secondary metabolic process
GO:0018106	1.39E-08	peptidyl-histidine phosphorylation
GO:0006066	1.40E-08	alcohol metabolic process
GO:0044262	1.19E-08	cellular carbohydrate metabolic process
GO:0000160	4.95E-07	phosphorelay signal transduction system
GO:0055114	1.34E-23	oxidation-reduction process
GO:0034641	1.76E-02	cellular nitrogen compound metabolic process
GO:0008152	2.26E-15	metabolic process
GO:0042651	9.61E-06	thylakoid membrane
GO:0044436	3.45E-06	thylakoid part
GO:0009579	1.37E-05	thylakoid
GO:0044434	2.11E-12	chloroplast part
GO:0009536	7.62E-14	plastid
GO:0005737	1.53E-05	cytoplasm
GO:0005623	3.57E-02	cell
GO:0042286	3.74E-04	glutamate-1-semialdehyde 2,1-aminomutase activity
GO:0004365	5.15E-11	glyceraldehyde-3-phosphate dehydrogenase (NAD+) (phosphorylating) activity
GO:0004021	1.38E-04	L-alanine:2-oxoglutarate aminotransferase activity
GO:0004324	2.12E-03	ferredoxin-NADP+ reductase activity
GO:0003989	9.18E-10	acetyl-CoA carboxylase activity
GO:0004802	6.99E-03	transketolase activity
GO:0017099	6.99E-03	very-long-chain-acyl-CoA dehydrogenase activity
GO:0008943	2.95E-09	glyceraldehyde-3-phosphate dehydrogenase activity
GO:0016851	9.50E-04	magnesium chelatase activity
GO:0004312	5.40E-04	fatty acid synthase activity
GO:0004332	2.96E-02	fructose-bisphosphate aldolase activity
GO:0031406	2.16E-15	carboxylic acid binding
GO:0051287	4.14E-05	NAD binding

Key GO term	p-value	GO description
GO:0019842	9.89E-15	vitamin binding
GO:0005506	1.30E-13	iron ion binding
GO:0000156	2.27E-08	phosphorelay response regulator activity
GO:0020037	4.15E-04	heme binding
GO:0009055	8.80E-05	electron carrier activity
GO:0000155	2.46E-04	phosphorelay sensor kinase activity
GO:0016746	2.62E-02	transferase activity, transferring acyl groups
GO:0016491	5.11E-22	oxidoreductase activity

4

A sex-inducing conditioning factor triggers cell cycle arrest and diproline production in *Seminavis robusta*

Sara Moeys [1,2,3], Jeroen Gillard [1,2,3], Barbara Bouillon [2,3], Valerie Devos [1,2,3], Koen Van den Berge [4], Johannes Frenkel [5], Marie J.J. Huysman [2,3], Georg Pohnert [5], Lieven Clement [4], Koen Sabbe [1], Lieven De Veylder [2,3] and Wim Vyverman [1]

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

[2] Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium

[3] Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

[4] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281-S9, B-9000 Ghent, Belgium

[5] Institute of Inorganic and Analytical Chemistry, Friedrich Schiller University Jena, Lessingstrasse 8, D-07743 Jena, Germany

Manuscript in preparation

Authors' contributions

SM, JG, VD, LDV and WV conceived and designed the study. SM performed the RNA-seq experiment, analyzed the data and wrote the manuscript. JG and VD performed the physiological experiments. BB performed the RT-qPCRs. KvdB and LC conducted the statistical analyses. JF and GP provided the purified fraction. MJJH and KS were involved in experimental design. LDV and WV read and approved the manuscript.

Abstract

In the marine, benthic pennate diatom *Seminavis robusta*, a multistep pheromone system is used to control mate pair formation. First, cells below the sexual size threshold produce sex-inducing conditioning factors that are sensed by the opposite mating type, after which MT^- cells produce the attraction pheromone diproline. Here, the effect of the MT^+-conditioning factor (CF-P) on MT^- is investigated, using bio-assays and RNA sequencing. Sensing of CF-P by MT^- leads to a G_1 cell cycle arrest and to the production of the attraction pheromone diproline. Under the influence of CF-P, expression of mitosis-related genes decreases, while meiosis-related genes are upregulated. Furthermore, an upregulation of the glutamate-to-proline pathway is seen, which probably leads to higher diproline production. We hypothesize that this signal is passed on to its targets by the secondary messenger cGMP, since a bifunctional guanylyl cyclase/phosphodiesterase is rapidly induced in CF-P treated cultures.

Introduction

Despite the great ecological importance of diatoms, not much in know about the regulation of their unique life cycle. The diatom life cycle is characterized by a size-dependent transition from mitotically dividing cells to cells capable of sexual reproduction (Drebes, 1977, Chepurnov et al., 2004). Without sexual reproduction, the population would undergo clonal cell death. This is a consequence of the way the new cell wall is formed after a vegetative division. The cell wall, or frustule, consists of two silicified thecae, called the hypotheca (smaller halve) and the epitheca (larger halve) that fit together as the two halves of a petri-dish. After cell division, both daughter cells inherit one valve from the mother cell and synthesize a new hypotheca within the constraints of the mother valve. Consequently, one of the daughter cells will be smaller than the mother cell, which will lead to a decrease in average cell size in the population. During sexual reproduction, which is only possible below a species-specific sexual size threshold (SST), a specialized zygote, called the auxospore, is formed. This auxospore is free of the constraining frustule and elongates, resulting in offspring of initial cell size.

It was recently shown that the differentiation of cells passing the SST involves the production of pheromones that make it possible to find a mating partner (Sato et al., 2011, Gillard et al., 2013). In the pennate diatom *Seminavis robusta*, the structure of one of these pheromones was elucidated (Gillard et al., 2013). It was shown that MT^+ cells are attracted to diproline, produced by MT^- cells. These cells excrete diproline when they are below the SST, but only after they perceived the presence of a suitable mating partner (MT^+ cells below SST). Likewise, the attraction of MT^+ cells to diproline is dependent on cell size and on the prior perception of metabolites produced by sexually mature MT^- cells. These findings point to a multistep control system for cell pairing. In this way, *S. robusta* probably avoids investing in sexual reproduction when no suitable mating partner is present.

Here, we show that MT^- cells below the SST are arrested in the G_1 phase when exposed to a diffusible factor produced by MT^+ cells. This info-chemical, called

conditioning factor plus (CF-P), is responsible for the "conditioning" of MT^- cells to arrest their cell cycle and to start the production of the attraction pheromone diproline. An active medium fraction containing CF-P was used to set up an RNA-seq experiment to study the effects of CF-P on genome-wide expression. The results indicated that diproline production was increased by upregulating the proline biosynthesis from glutamate. Furthermore, it points towards the secondary messenger cGMP as a key regulator of pheromone signaling.

Materials & Methods

Strains and culture conditions

Seminavis robusta strains 8B and 85A (MT^+) and 9A and 85B (MT^-) were used, which are available in the diatom culture collection of the Belgian Coordinated Collection of Micro-organisms (BCCM/DCG, http://bccm.belspo.be, accession numbers DCG 0097, DCG 0105, DCG 0096 and DCG 0107). Cultures were grown in F/2 medium (Guillard, 1975) made with autoclaved filtered natural sea water collected from the North Sea and Guillard's F/2 solution (Sigma-Aldrich).

RNA-seq on CF-P-treated MT^- cultures

Cultures of strain 85B (MT^-) with an average cell size below the sexual size threshold were grown at 18°C in a 12:12h light:dark regime with cool white fluorescent lamps at approximately 80 µmol photons m^{-2} s^{-1}. Before sampling, the dark period was extended with 12 hours to synchronize cells in G_1 phase. A first control sample was taken in darkness (T0). Then, an active fraction containing CF-P (provided by Prof. dr. Georg Pohnert) was added to halve of the cultures in darkness. After 15 minutes, the light was turned on and again 15 minutes later treated and untreated cultures were harvested by filtration on versapor filters (3 µm pore size, 25 mm diameter, PALL). The filters were rinsed with 1 mL PBS and put in liquid N_2. This was repeated 1 hour and 3 hours after illumination. Sampling was carried out in triplicate.

The protocol used for RNA extraction was based on the method developed by Apt et al. (1995) with some modifications. The frozen filters were put in extraction buffer (100 mM Tris-HCl pH 7.5, 2% CTAB, 1.5 M NaCl, 50 mM EDTA, 10% β-mercaptoethanol) and sharp carbid beads were added. The samples were shaken in a bead mill (Retsch) at room temperature for 30 min. 100 µL of 10X Chelex-100 was added and the sample was incubated for 15 min at 56°C, then one volume of chloroform:isoamylic alcohol (24:1, V/V) was added and the samples were shaken for 25 min at low speed. After centrifugation, the upper phase was transferred to a new tube and mixed with 0.3 V of absolute ethanol to precipitate the polysaccharides, and extracted with 1 V of chloroform. After centrifugation the upper phase was transferred to a fresh tube and RNA was precipitated overnight at -20°C by addition of 0.25 V of 12 M LiCl and β-mercapto-ethanol to 1% final concentration. After centrifugation, the pellet was washed with 70% ethanol, air-dried and suspended in Rnase-free water for DNAse treatment by an RNAse-free DNAse I (Turbo DNAse, Ambion) according to the manufacturer's instructions. An extraction was then carried out by adding Phenol-Chloroform (1:1, V/V). After centrifugation the upper phase was transferred to a fresh tube, and extracted with 1 V of chloroform:isoamylic alcohol (24:1, V/V) and centrifuged again. The upper phase was precipitated with 0.3 M NaOAc pH 5.5 and 75% ice cold ethanol by incubating overnight at 20°C. After centrifugation, the pellet was washed with 70% ethanol and air-dried. The pellet was resuspended in an appropriate volume of RNAse-free water.

Sequencing libraries were prepared using Illumina TruSeq Stranded mRNA kit. All libraries were analysed in a 2x 150bp run on 2 lanes of a flowcell of the Illumina HiSeq 2500 at VIB Nucleomics Core (www.nucleomics.be). Every library was sequenced twice, as all libraries were pooled together in every lane.

After adaptor trimming with CLC Assembly Cell 4.2.0 (CLC Inc, Aarhus, Denmark), all paired reads were mapped to an in-house draft genome of *S. robusta* using GSNAP (GMAP-GSNAP version 2013-06-27) (Wu and Nacu, 2010) with default settings. Based on this mapping, a genome-guided trinity approach was used to assemble the

transcriptome, using Trinity r20131110 (Grabherr et al., 2011) and PASA r20130907 (Haas et al., 2003). All reads were mapped to this transcriptome using GSNAP (Wu and Nacu, 2010) after which all unmapped reads were extracted and de novo assembled using Trinity (Grabherr et al., 2011) using default settings. This de novo assembly was joined together with the genome-guided transcriptome assembly. CD_HIT 4.6.1 (cd-hit-est) (Li and Godzik, 2006, Fu et al., 2012) was used to optimize the assembly with the following settings: -c 0.9 -n 8 -t 1. Finally, all transcripts shorter than 500 bp were discarded. For functional annotation, the final transcriptome was loaded into TRAPID (reference database: Plaza 2.5) (Proost et al., 2009, Van Bel et al., 2013).

For every library, the reads were mapped to the final transcriptome using GSNAP (Wu and Nacu, 2010) with default settings and mapped reads per transcript per library were counted. Transcripts with low overall counts (threshold of at least 1 cpm in at least six samples) have been removed from the analysis because they have little power for detecting differential expression. Hence, filtering these transcripts leads to a negligible information loss and results in a gain of statistical power (Bourgon et al., 2010). Upon filtering, the sequencing libraries have been normalized using TMM normalization (Robinson and Oshlack, 2010). Transcript-wise RNA-seq counts were analysed using a negative binomial model with a factor with 7 levels, one for each treatment (Control, CF-P) x time (T0, T15min, T1h, T3h) combination (C0, C15min, C1h, C3h, CFP15min, CFP1h,CFP3h). The control taken in darkness (C0) was used as the reference level. Technical replicates have been pooled as to avoid underestimation of the dispersion parameter. Both differential expression (DE) between CF-P treated samples and untreated samples within each time point as well as DE changes over time have been assessed on a log scale using appropriate contrasts (DE1 = CFP15min – C15min, DE2 = CFP1h – C1h, DE3 = CFP3h – C3h, DE3-DE1, DE2-DE1, DE3-DE2). A hierarchical two-stage approach has been used for assessing statistical significance, which controls the overall false discovery rate (FDR) at the transcript level (Heller et al., 2009). In the first stage, the overall null hypothesis that no differential expression occurred at none of the timepoints

(DE1=DE2=DE3=0) has been assessed. Next, the six contrasts of interest have been tested for all genes that passed the screening stage. An advanced, within-gene, multiple testing correction that accounts for the relatedness of different hypotheses (contrasts) has been adopted in the second stage (Shaffer, 1986). Hence, the overall FDR has efficiently been controlled at a 5% level, while maintaining significant power for testing the hypotheses of interest. All analyses were conducted with the R/BioConductor package edgeR (version 3.8.5) (Robinson et al., 2010) in R version 3.1.2 and likelihood ratio tests were used for assessing all hypotheses of interest (McCarthy et al., 2012).

RT-qPCR on CF-P-treated MT$^-$ cultures

Samples were taken as described for the RNA-seq on CF-P-treated MT$^-$ cultures just before illumination and 15 min, 1h, 3h, 6h, 9h and 12h after illumination. Progression through the cell cycle was monitored by counting dividing cells under an inverted microscope. About 200 mL of culture (± 5 x 10^6 cells) was scraped from the bottom of a tissue culture flask (Cellstar, 175 cm^2 growth surface, Greiner Bio-One) and filtered on a Versapor filter (3 µm pore size, 25 mm diameter, PALL). Filters were frozen in liquid nitrogen. Total RNA was extracted using the RNeasy Plant Mini Kit (Qiagen). RLT buffer (1mL), β-mercaptoethanol (10µL) and silicon carbide beads (1.0 mm, BioSpec) were added to the filter. Cells were disrupted by beating on a bead mill (Retsch, 3 x 1 min). All other steps were done according to the manufacturer's instructions. cDNA was prepared using the iScript cDNA Synthesis Kit (bio-rad) according to the manufacturer's instructions.

15 ng of cDNA was used as template for each RT-Q-PCR reaction. Samples were amplified in triplicate on the Lightcycler480 platform with the Lightcycler 480 SYBR Green I Master mix (Roche) in the presence of 0.5 µM gene-specific primers (primer sequences in Suppl. Table S1). The cycling conditions were 10 minutes polymerase activation at 95°C and 45 cycles at 95°C for 10 s, 58°C for 15 s, 72°C for 15 s. Afterwards, amplicon dissociation curves were recorded by heating from 65°C to 97°C with a ramp rate of 2.5°C.

RT-qPCR gene expression data were normalised within each biological replicate through the ΔCt relative quantification method from the qbase+ software (Biogazelle). Two stably expressed normalisation genes, previously selected based on cDNA-AFLP data (Valerie Devos, unpublished data), were used for normalisation. After normalisation, expression values were calculated relative to the unconditioned treatment at time point 0. A linear regression model was built for each gene separately, using the logarithm of relative expression values as response variable. A logarithmic transformation was applied to approximate a Gaussian distribution for the residuals. The regression model contains a discrete time variable, the treatment effect (unconditioned/conditioned) and their interaction, allowing for different time profiles for both treatments. An effect for the biological replicate was added to correct for biological heterogeneity between samples and possible batch (plate) effects, as each biological replicate was analysed on a separate plate. To assess expression differences in time, hypothesis tests based on the sum of the treatment and treatment x time interaction effect were performed for specific time points using an F-test. Hypothesis tests were evaluated at time points 1h, 3h, 6h, 9h and 12h. The Bonferroni procedure was used to correct for within-gene multiple testing.

Results

Cells below the SST produce a cell-cycle-arresting signal

When two strains of opposite mating types are mixed, mating pairs are formed. In these pairs, cell cycle progression is interrupted as a result of the induction of gametogenesis. Surprisingly, also cells that were not part of a mating pair, did not go through mitosis, since no dividing cells were observed in the mixed cultures, even though only about halve of the cells had engaged in mating pair formation (Figure 1A). This observation is a first indication of a cell cycle arrest induced by the presence of the compatible mating type.

Experiments where the medium of cultures was replaced by filtered culture medium of the other mating type (thus containing all excreted metabolites but no

cells) showed that no physical cell-cell contact was required to arrest the cell cycle (Figure 1B). Filtered medium of a sexually reproducing culture could induce cell cycle arrest in both MT^+ and MT^- below the SST. Cells below the SST also exhibited the cell cycle arrest when they were brought in filtered culture medium of cells of the opposite mating type below the SST. Cells above the SST didn't arrest their cell cycle and were not able to induce cell cycle arrest in other cultures. In addition, microscopic observations showed that cells had undivided chloroplasts at the girdle (data not shown), indicative for G_1 phase (Gillard et al., 2008). These results indicate that cells below the SST secrete diffusing info-chemicals with the ability to arrest the cell cycle of the compatible mating type in the G_1 phase.

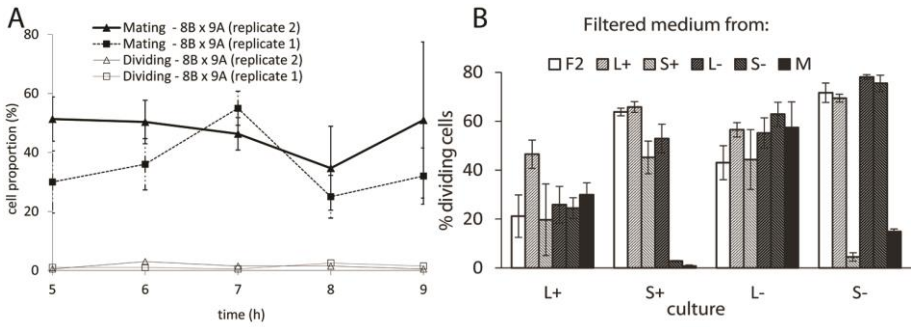


Figure 1: (A) The proportion of dividing (empty symbols) and mating (full symbols) cells in two mixed cultures of MT^+ (8B) x MT^- (9A) strains after dark-synchronization. Error bars are ± SE of 3 replicate estimations. (B) The effect of medium replacements on dark-synchronized cultures of MT^+ (85A) and MT^- (85B) above (L+ and L-) and below (S+ and S-) the SST was assessed by counting dividing cells at 9 hours after illumination. The medium of these cultures was replaced by spent medium from a sexually reproducing culture (M) or from dark-synchronized cultures of MT^+ and MT^- above (L+ and L-) and below (S+ and S-) the SST, or by blank growth medium (F2). Error bars are ± SE of 3 replicates. * indicates statistical significance with p-value < 0.001.

CF-P is responsible for the mitosis-to-meiosis switch in MT^-

It was previously shown that the attraction of *S. robusta* MT^+ towards MT^- cells is controlled by diproline (Gillard et al., 2013) and that diproline production by MT^- cells below the SST greatly increases when these cells were first conditioned by the

spent medium of MT^+ cultures. Likewise, MT^+ cells are only responsive to diproline after conditioning with spent medium of MT^- cells. An active fraction was purified from the medium of MT^+ cells below the SST that was able to induce the diproline production by MT^- cells (Frenkel, 2014). Similar to adding filtered medium of the opposite mating type, administering this active fraction to an MT^- culture below the SST led to a cell cycle arrest. Dividing cells were observed in unconditioned cultures from six hours after illumination onwards; no division was seen in conditioned cultures (Figure 2A). Diproline production and cell cycle arrest are thus most likely induced by the same info-chemical, called conditioning factor plus (CF-P), present in this active fraction. Analogously, the info-chemical secreted by MT^- to induce cell cycle arrest and probably also diproline responsiveness in MT^+ is called conditioning factor minus (CF-M).

The cell cycle arrest induced by CF-P was verified in a RT-qPCR experiment. Cultures of small MT^- cells were grown for three days in 12h:12h light:dark regime and then the cell cycle was synchronized by extending the dark period with twelve hours. After this prolonged dark period, the majority of cells are arrested in the G_1 phase of the cell cycle (Gillard et al., 2008). Before illumination, halve of the cultures was treated with the active fraction containing CF-P. To verify the cell cycle arrest at the molecular level, the expression level of two mitosis markers was measured by RT-qPCR. *CYCB1* and *CYCA/B1* were selected, because of their expression pattern in *Phaeodactylum tricornutum*, where they are both highly expressed during mitosis (Huysman et al., 2010). Accordingly, the mitotic markers *CYCB1* and *CYCA/B1* are indeed expressed during mitosis in untreated *S. robusta* cells. In contrast, they are repressed when cells are treated with CF-P, which corresponds with the observed cell cycle arrest (Figure 2B/C). At nine hours, which is approximately the timing for mitosis, both markers are significantly differentially expressed between conditioned and unconditioned cultures (adjusted p-value for *CYCB1* = 5.71×10^{-8}, adjusted p-value for *CYCA/B1* = 0.004).

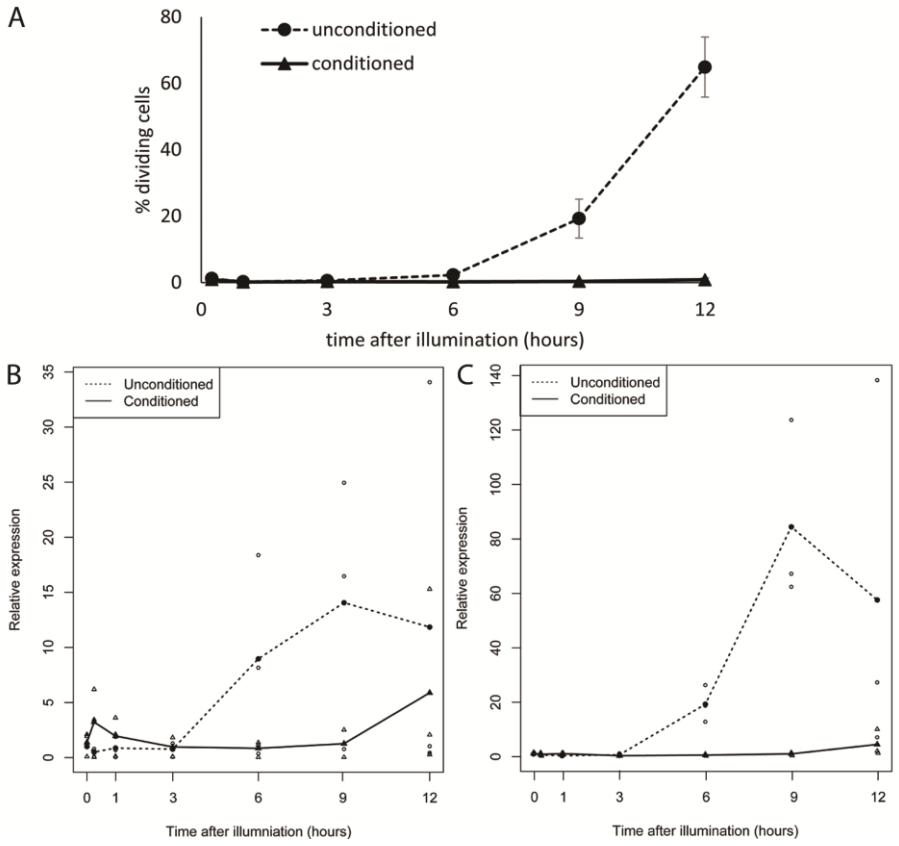


Figure 2: Cell cycle arrest induced by the CF-P containing fraction. (A) Dividing cells were counted in conditioned ( ) and unconditioned ( ) MT⁻ cultures. (B and C) RT-qPCR on *CYCA/B1* (B) and *CYCB1* (C) in conditioned ( ) and unconditioned ( ) MT⁻ cultures. Relative expression values are normalized to two stably expressed genes. The lines represent the average of 3 biological replicates. Relative expression values for all data points are shown.

To obtain a global view on the transcriptional changes induced by CF-P, an RNA-seq experiment was conducted. For this experiment, MT⁻ cultures were treated with the active fraction as described above and samples were taken 15 minutes, 1 hour and 3 hours after illumination. Additionally, a non-treated MT⁻ culture was harvested before illumination. RNA sequencing yielded a total of 152M PE reads or 3.6M/library/lane on average. A combination of genome-guided assembly and de novo assembly of the unmapped reads produced a transcriptome of 50,231 transcripts with an average length of 1,416 bp and ranging from 500 to 17,339 bp.

After functional annotation using TRAPID (Van Bel et al., 2013), based on similarity searches against the PLAZA 2.5 database (Proost et al., 2009), 12,358 transcripts got at least one GO term appointed and 13,675 transcripts one or more interpro domains.

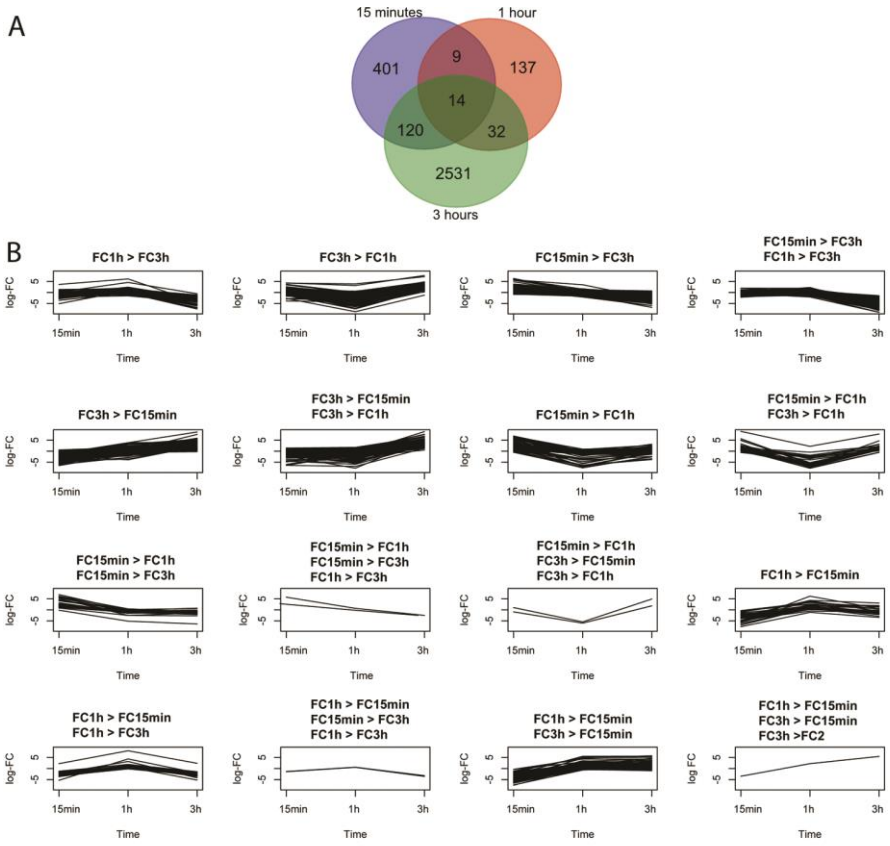


Figure 3: (A) Differential expression between conditioned and unconditioned cultures at 15 minutes, 1 hour and 3 hours. (B) Transcripts for which the fold change (conditioned versus unconditioned cells) significantly changed over time, divided in classes based on the direction of the difference in fold change (FC) between the different time points. The graphs show the logFC in relation to time.

The screening stage, comparing treated and untreated samples within each time point, showed that 3244 unique transcripts were differentially expressed in at least one time point. The majority of DE occurred at 3h (2697 transcripts), compared to

1h and 15m (192 and 544 DE transcripts, respectively; Figure 3A). Furthermore, a significant difference in fold change has been observed for 289, 1308 and 1263 transcripts in the 1h versus 15m, 3h versus 15m and 3h versus 1h comparison, respectively. These results have been used to classify differentially expressed transcripts according to significant fold change patterns (Figure 3B).

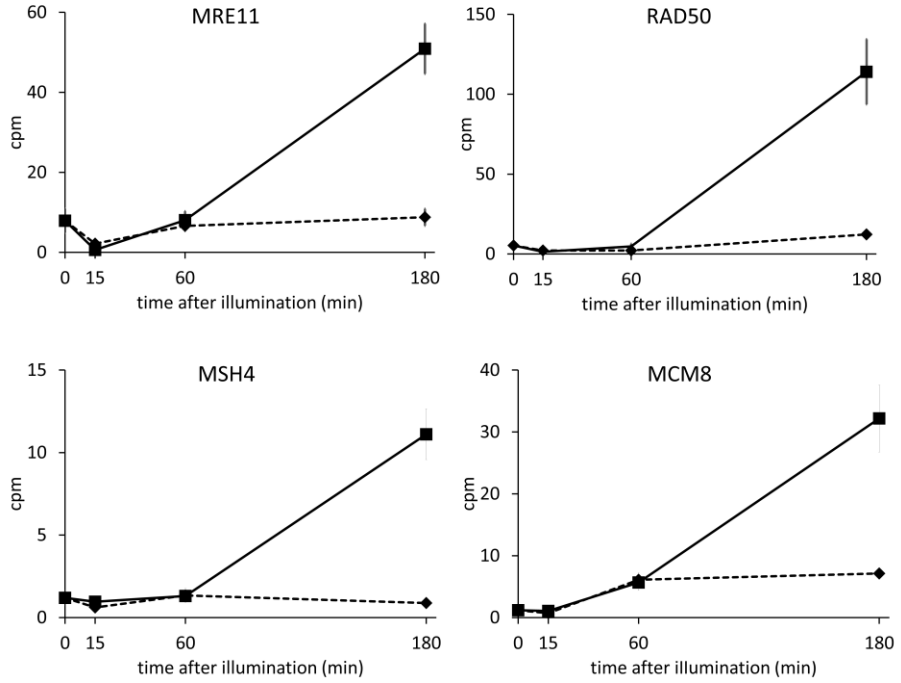


Figure 4: Mean cpm values +- SE error bars (3 replicates) for *MRE11*, *RAD50, MSH4* and *MCM8* in unconditioned cultures (--◆--) and conditioned cultures (—■—).

Amongst the differentially expressed genes after treatment with CF-P, four meiosis-related genes were identified. All four are significantly higher expressed in conditioned compared to unconditioned cultures at three hours after illumination (Figure 4). *MRE11* and *RAD50* belong to the FC3>FC1 class, *MSH4* to the FC3>FC2 class (Figure 3B). These genes were shown to be highly expressed during meiosis in *S. robusta* (Chapter 3). MRE11 and RAD50 are both part of the MRN complex that is involved in repairing DNA double-strand breaks and in homologous recombination during meiosis (Ajimura et al., 1993, Symington, 2002). MSH4 is a mismatch repair

protein that functions specifically in meiosis (Kolas and Cohen, 2004). MCM8 is involved in meiotic recombination (Blanton et al., 2005, Crismani et al., 2013).

Together with the observed cell cycle arrest, these data indicate that the sensing of CF-P by MT^- cells below the SST results in a switch from mitosis-related genes to meiosis-related genes, although meiosis itself will only start when mating pairs have been formed and gametogenesis can start (Chepurnov et al., 2002).

The glutamate-to-proline pathway is upregulated under CF-P treatment

Gillard et al. (2013) detected diproline in the medium of dark-synchronized conditioned MT^- cultures 5h after illumination. Thus, we expect the genes responsible for diproline production to be upregulated in this dataset after administration of CF-P. Cyclodipeptides can be synthesized by non-ribosomal peptide synthases (NRPS) in bacteria and fungi or by tRNA-dependent cyclodipeptide synthases (CDPS), which are not only found in bacteria and lower eukaryotes, but also in animals (Belin et al., 2012).

Several putative NRPSs could be identified based on their functional annotation in TRAPID (Van Bel et al., 2013). However, none of them was upregulated under CF-P treatment. To identify putative members of the CDPS family PSI-BLAST (BLAST 2.2.27+) (Altschul et al., 1997) was used with *Streptomyces noursei* AlbC and *Nematostella vectensis* CDPS2 as queries, but no significant hits were found.

Although no NRPS or CDPS homolog was found to be upregulated, Δ1-pyrroline-5-carboxylate synthetase (*P5CS*) was time-dependently upregulated in conditioned cultures (subset FC3>FC1, FC3>FC2), with a strong increase in expression at three hours (Figure 3B). This enzyme has both γ-glutamyl kinase and glutamic-γ-semialdehyde dehydrogenase activity and mediates the first two steps in the conversion of glutamate to proline (Hu et al., 1992). Consequently, this enzyme could be the first step in diproline biosynthesis (Figure 5B). Also the second enzyme of the pathway, Δ1-pyrroline-5-carboxylate (P5C) reductase, was significantly higher

expressed in conditioned compared to unconditioned cultures three hours after illumination.

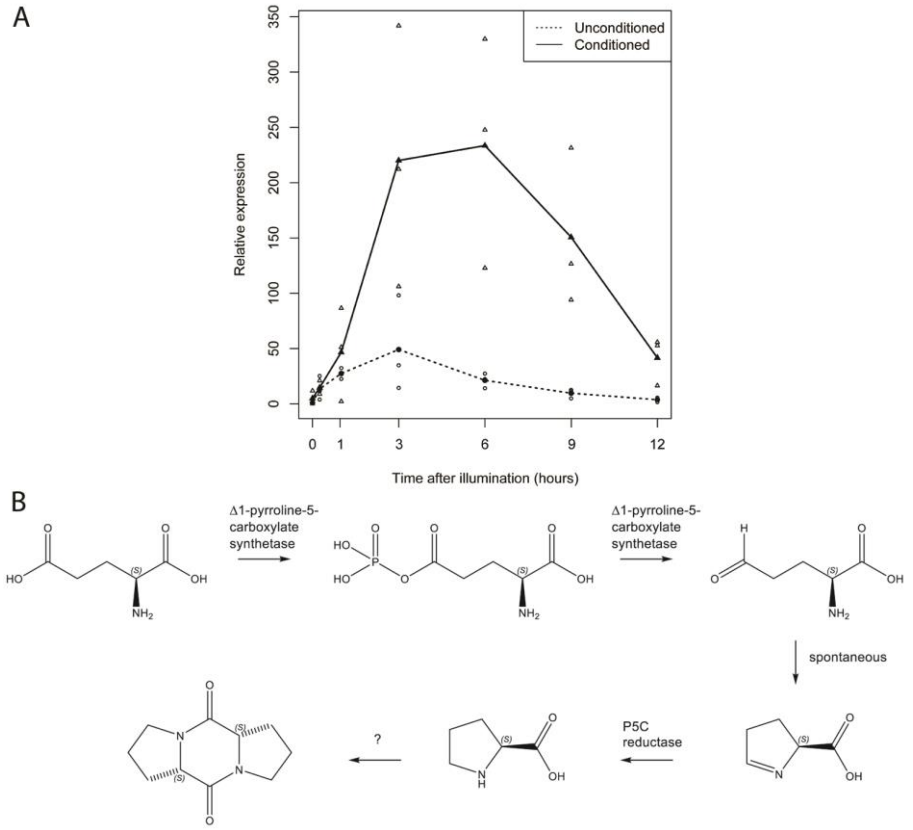


Figure 5: (A) RT-qPCR for *Δ1-pyrroline-5-carboxylate synthetase* in unconditioned (--◆--) and conditioned MT$^-$ cultures (—■—). Relative expression values are normalized to 2 stably expressed genes. The lines represent the average of 3 biological replicates. Relative expression values for all data points are shown. (B) Hypothetical diproline biosynthesis pathway.

The expression of *P5CS* was confirmed in a longer time course by RT-qPCR, showing that *P5CS* is significantly higher expressed after CF-P treatment at 6h (adjusted p-value = 0.003), 9h (adjusted p-value = 6.2 x 10^{-4}) and 12h (adjusted p-value = 0.003) after illumination (Figure 5A). This corresponds with the previous finding that diproline can be detected in conditioned MT$^-$ cultures from five up to ten hours after illumination and with the timing of mate pairing (5-10 hours after illumination) (Gillard et al., 2013).

The expression of a bifunctional guanylyl cyclase/phosphodiesterase increases upon treatment with CF-P

Only one transcript was found for which the fold change in expression (difference between conditioned and unconditioned cultures) at each time point was statistically higher than at the previous time point (class FC3>FC2, FC3>FC1, FC2>FC1) (Figure 3B). This gene encodes a putative bifunctional protein containing an adenylyl/guanylyl cyclase domain and a cyclic nucleotide phosphodiesterase (PDE) domain (Figure 6B). These domains are responsible for cyclic nucleotide (cAMP of cGMP) synthesis and breakdown respectively, pointing to the potential involvement of one of these secondary messengers in pheromone signalling in *S. robusta*.

By aligning the catalytic core of the cyclase domain to that of several adenylyl and guanylyl cyclases that were previously used to model the catalytic mechanism of nucleotide cyclases (Liu et al., 1997), it could be predicted that this enzym represents a guanylyl cyclase. Of the guanine binding residues, all but two are conserved and except for the pyrophosphate-binding site, all catalytic sites are conserved (Suppl. Figure S1A). Richter et al. (2001) identified eight amino acids that are conserved in cAMP-specific PDEs but not in cGMP-specific PDEs. These amino acids are not conserved in the PDE domain of this *S. robusta* protein, indicating that it probably has cGMP-hydrolyzing capacity (Suppl. Figure S1B). From these findings, it can be concluded that this bifunctional cyclase/PDE probably has cGMP specificity.

Guanylyl cyclases can be either transmembrane or soluble proteins. Phobius was used to predict the transmembrane topology of the bifunctional guanylyl cyclase/PDE (Kall et al., 2007). This protein has two predicted transmembrane domains N-terminal from the two catalytic domains (Figure 6B), indicating that this probably is a transmembrane guanylyl cyclase.

The expression of the *GC/PDE* gene was followed during a longer time course by RT-qPCR, showing that its expression is significantly higher in conditioned cultures at 3h (adjusted p-value = 0.013), 9h (adjusted p-value = 0.019) and 12h (adjusted p-

value = 0.012) after illumination, which largely corresponds with the expression of *P5CS* (Figure 6A, Figure 5A).

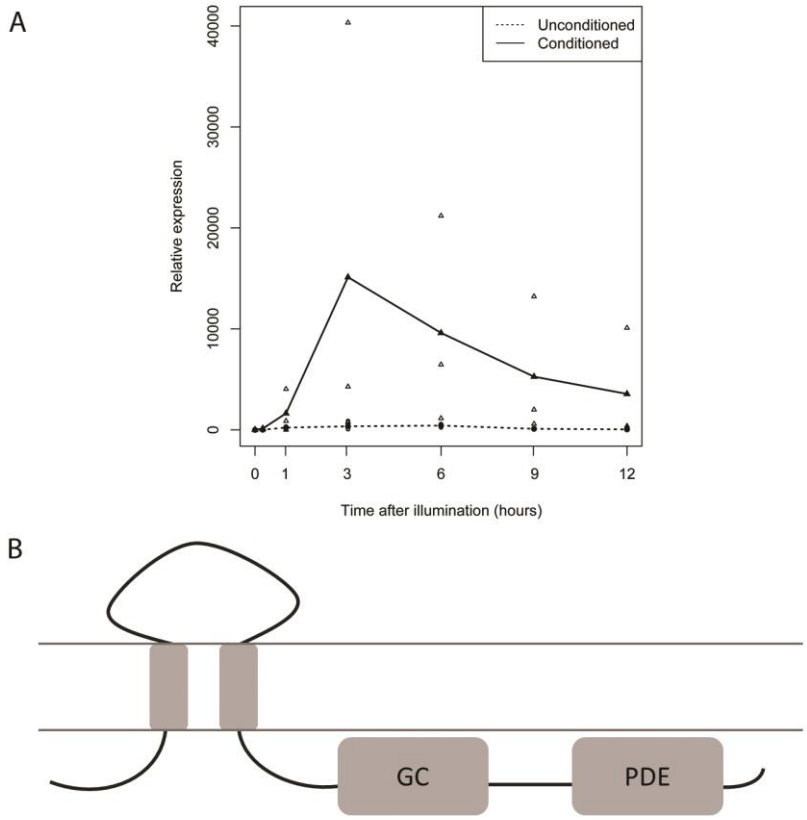


Figure 6: (A) RT-qPCR for the bifunctional GC/PDE in unconditioned (--●--) and conditioned cultures (—▲—). Relative expression values are normalized to 2 constitutively expressed genes. The lines represent the average of 3 biological replicates. Relative expression values for all data points are shown. (B) Schematic of the protein architecture: two transmembrane domains, a guanylyl cyclase domain (GC) and a cyclic nucleotide phosphodiesterase domain (PDE).

Discussion

In this paper, the first steps towards mate pairing in the pennate diatom *Seminavis robusta* were investigated. Under the influence of the info-chemical CF-P, produced by MT^+ below the SST, MT^- cells arrest their cell cycle and start secreting diproline, the pheromone responsible for the attraction of MT^+ cells (Figure 7). Only cells that are below the SST, and thus able to undergo sexual reproduction, respond to CF-P. This molecule is the signal that potential mating partners are in close proximity.

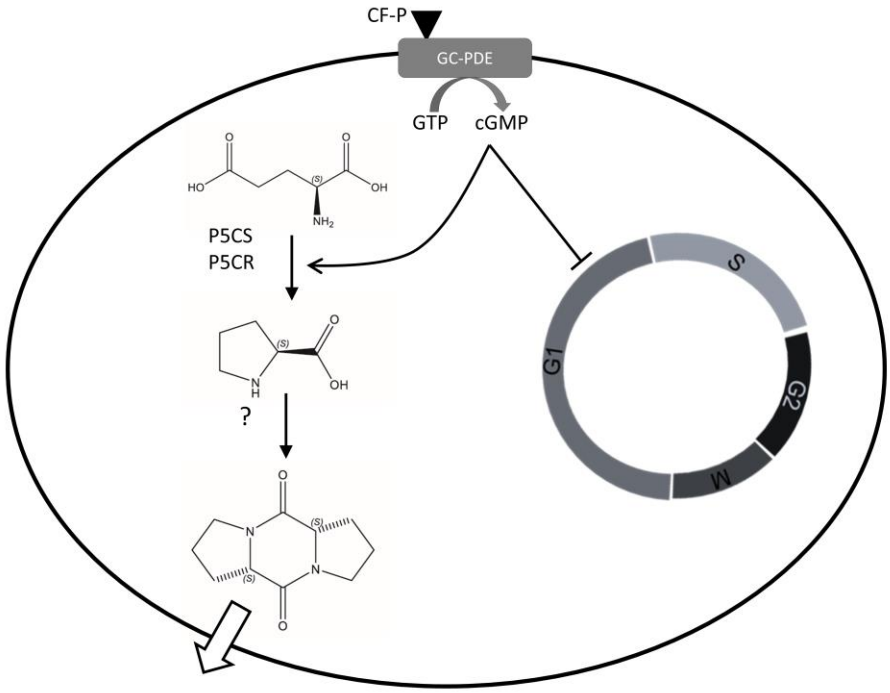


Figure 7: A schematic representation of our hypothesis. Upon binding of CF-P on its receptor (possibly the extracellular part of the GC-PDE protein) cGMP levels rise and the secondary messenger cGMP induces the conversion of glutamate to proline by Δ1-pyrroline-5-carboxylate synthetase (P5CS) and Δ1-pyrroline-5-carboxylate reductase (P5CR). Proline can then be converted to diproline by an unknown mechanism. Diproline is secreted to attract MT^+ cells. Apart from inducing pheromone production, cGMP also triggers a cell cycle arrest in G_1 phase.

It was shown that at the moment of mate pairing, cells stop dividing. It is expected that this is the case for the cells forming pairs, since they will start gametogenesis. Surprisingly, also cells that do not find a mating partner (yet) display no signs of mitosis. Medium replacement experiments confirmed that direct cell-cell contact is not necessary for this cell cycle arrest, but that it is induced by diffusible info-chemicals that are secreted by cells below the SST. Microscopic observations showed that arrested cells had undivided chloroplasts at the girdle, which is indicative for G_1 phase. In vegetatively dividing cells, chloroplasts divide and translocate from the girdle to the valves during S/G_2 phase (Gillard et al., 2008). However, when mating cells prepare for meiosis, chloroplast translocation happens without prior chloroplast division (Chepurnov et al., 2002). This indicates that the transition from the mitotic cell cycle to meiosis happens before chloroplast division and consequently before S/G_2 phase. Taken together, these observations point to a cell cycle arrest in G_1 phase that is induced by info-chemicals secreted by the opposite mating type. Similarly, in yeast, mammals and plants commitment to meiosis is made before the onset of premeiotic S phase (Pawlowski et al., 2007). Also in the centric diatom *Thalassiosira weissflogii*, it was shown that cells could only be induced to undergo spermatogenesis when they are in the early G_1 phase (Armbrust and Chisholm, 1990). We can thus postulate that *S. robusta* arrests its cell cycle in G_1 phase when mating partners are present to prolong the phase wherein the cells are able to make the transition from mitosis to meiosis. In this way, they raise their chances of finding a mating partner within the right time frame. An analogous signalling mechanism is known in yeast, where pheromones induce cell cycle arrest in G_1 phase preceding mating (Bardwell, 2004).

Δ1-pyrroline-5-carboxylate synthetase and Δ1-pyrroline-5-carboxylate reductase are upregulated when cells are treated with CF-P. These two enzymes form the pathway that is responsible for the conversion of glutamate to proline. The upregulation of this pathway indicates that the attraction pheromone diproline is synthesized from proline. However, the enzyme(s) implicated in the last step (proline to diproline) are not yet identified. Cyclodipeptides (CDP), like diproline, are

widespread in nature and are predominantly synthesized by microorganisms. They are thought to be involved in cell-cell communication, for instance in bacterial quorum sensing or in bacteria-plant interactions (Degrassi et al., 2002, Ortiz-Castro et al., 2011). Two distinct protein families are known to be involved in CDP biosynthesis, non-ribosomal peptide synthetases (NRPS) and cyclodipeptide synthases (CDPS) (Belin et al., 2012). NRPS is a large protein family that is responsible for the synthesis of a huge variety of peptides in bacteria and fungi (Caboche et al., 2008). The CDPS family was only discovered in 2002 in *Streptomyces noursei* (Lautru et al., 2002). Members are now identified in a wide range of bacteria, but also in some eukaryotes, namely in annelids, fungi, protozoa and animals (Aravind et al., 2010, Seguin et al., 2011). While CDPSs are small enzymes (about 200 AA), NRPSs are large proteins consisting of several modules, that each add one amino acid to the peptide. Each module consists of three domains: an adenylation domain responsible for substrate activation, a peptidyl carrier domain that binds the activated substrate and a condensation domain where the peptide bond is formed (Marahiel et al., 1997). Apart from these standard domains, accessory domains can be present, which introduce modifications to the peptide. In CDPS pathways, these modifications are done by other enzymes, that are often present in the same operon on the genome (Belin et al., 2012). In contrast to NRPSs, CDPSs lack an adenylation domain for substrate activation, instead they use aminoacyl-tRNAs as substrates. As a consequence, their substrates are restricted to the canonical amino acids, while NRPS can use a wider range of substrates. There are several putative NRPS homologs present in *S. robusta*, but none of them was upregulated upon treatment with CF-P. Nevertheless, it is possible that these genes are constitutively expressed and that their activity is regulated at the posttranscriptional level. And although no CDPS homolog could be identified using PSI-blast, it cannot be excluded that an enzyme with a similar function is present, since sequence similarity between CDPSs is only 19-27% (Belin et al., 2012).

Interestingly, a bifunctional guanylyl cyclase/phosphodiesterase (GC/PDE) was found for which the expression difference between conditioned and unconditioned

cells strongly increases through time. cGMP is a secondary messenger synthesized from GTP by guanylyl cyclases (GCs) and broken down to GMP by phosphodiesterases (PDE). GCs can be soluble or membrane-bound. In animals, soluble GCs are activated by nitric oxide, while membrane-bound GCs are activated by ligands that bind to their extracellular domain (Potter, 2011). The CF-P-responsive GC/PDE has two transmembrane domains and both the GC and PDE catalytic domains are positioned at the inside of the cell. It is thus possible that this enzyme is activated by a ligand (most likely CF-P) that binds to its extracellular domain. Mammalian membrane-bound GCs have one transmembrane domain with the catalytic domain on the inside of the cell and a ligand-binding domain on the outside, but other conformations have been found in other eukaryotes (Schaap, 2005). The main targets of cGMP are cGMP-dependent protein kinase (PKG), cGMP-gated ion channels and proteins containing a GAF-domain. cGMP is known to regulate many processes in mammalians, like for instance smooth muscle relaxation and phototransduction (Potter et al., 2006, Koch and Dell'Orco, 2013). cGMP signalling is often involved in motility, for example in sperm chemotaxis in the sea urchin *Arbacia punctulata* or myosin II regulation in Dictyostelium (Kaupp et al., 2006, Artemenko et al., 2014). Up to now, nothing is known about the function of cGMP signalling in diatoms. The fact that the GC and PDE domains are part of a fusion protein could indicate a tight regulation of cGMP levels. To our knowledge, this combination of domains is only found in diatoms. We hypothesize that CF-P is sensed by either the extracellular domain on GC/PDE or by an activator of this protein. This leads to cGMP synthesis, which can then activate the downstream signalling cascade, probably leading to cell cycle arrest and diproline production.

Literature cited

Ajimura, M., Leem, S.H. and Ogawa, H. (1993) Identification of New Genes Required for Meiotic Recombination in Saccharomyces-Cerevisiae. *Genetics* **133**, 51-66.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402.

Apt, K.E., Clendennen, S.K., Powers, D.A. and Grossman, A.R. (1995) The Gene Family Encoding the Fucoxanthin Chlorophyll Proteins from the Brown Alga Macrocystis-Pyrifera. *Molecular & General Genetics* **246**, 455-464.

Aravind, L., de Souza, R.F. and Iyer, L.M. (2010) Predicted class-I aminoacyl tRNA synthetase-like proteins in non-ribosomal peptide synthesis. *Biology Direct* **5**.

Armbrust, E.V. and Chisholm, S.W. (1990) Role of Light and the Cell-Cycle on the Induction of Spermatogenesis in a Centric Diatom. *Journal of Phycology* **26**, 470-478.

Artemenko, Y., Lampert, T.J. and Devreotes, P.N. (2014) Moving towards a paradigm: common mechanisms of chemotactic signaling in Dictyostelium and mammalian leukocytes. *Cellular and Molecular Life Sciences* **71**, 3711-3747.

Bardwell, L. (2004) A walk-through of the yeast mating pheromone response pathway. *Peptides* **25**, 1465-1476.

Belin, P., Moutiez, M., Lautru, S., Seguin, J., Pernodet, J.L. and Gondry, M. (2012) The nonribosomal synthesis of diketopiperazines in tRNA-dependent cyclodipeptide synthase pathways. *Natural Product Reports* **29**, 961-979.

Blanton, H.L., Radford, S.J., McMahan, S., Kearney, H.M., Ibrahim, J.G. and Sekelsky, J. (2005) REC, Drosophila MCM8, drives formation of meiotic crossovers. *PLoS Genetics* **1**, 343-354.

Bourgon, R., Gentleman, R. and Huber, W. (2010) Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 9546-9551.

Caboche, S., Pupin, M., Leclere, V., Fontaine, A., Jacques, P. and Kucherov, G. (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Research* **36**, D326-D331.

Chepurnov, V.A., Mann, D.G., Vyverman, W., Sabbe, K. and Danielidis, D.B. (2002) Sexual reproduction, mating system, and protoplast dynamics of *Seminavis* (Bacillariophyceae). *Journal of Phycology* **38**, 1004-1019.

Chepurnov, V.A., Mann, D.G., Sabbe, K. and Vyverman, W. (2004) Experimental studies on sexual reproduction in diatoms. *International Review of Cytology* **237**, 91-154.

Crismani, W., Portemer, V., Froger, N., Chelysheva, L., Horlow, C., Vrielynck, N. and Mercier, R. (2013) MCM8 Is Required for a Pathway of Meiotic Double-Strand Break Repair Independent of DMC1 in Arabidopsis thaliana. *PLoS Genetics* **9**.

Degrassi, G., Aguilar, C., Bosco, M., Zahariev, S., Pongor, S. and Venturi, V. (2002) Plant growth-promoting Pseudomonas putida WCS358 produces and secretes four cyclic dipeptides: Cross-talk with quorum sensing bacterial sensors. *Current Microbiology* **45**, 250-254.

Drebes, G. (1977) Sexuality. in *The biology of diatoms*, Vol. 13 250-283 (Univ of California Press).

Frenkel, J. (2014) PhD dissertation, Friedrich-Schiller-University of Jena.

Fu, L.M., Niu, B.F., Zhu, Z.W., Wu, S.T. and Li, W.Z. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152.

Gillard, J., Devos, V., Huysman, M.J.J., De Veylder, L., D'Hondt, S., Martens, C., Vanormelingen, P., Vannerum, K., Sabbe, K., Chepurnov, V.A., Inzé, D., Vuylsteke, M. and Vyverman, W. (2008) Physiological and transcriptomic evidence for a close coupling between chloroplast ontogeny and cell cycle progression in the pennate diatom *Seminavis robusta*. *Plant Physiology* **148**, 1394-1411.

Gillard, J., Frenkel, J., Devos, V., Sabbe, K., Paul, C., Rempt, M., Inze, D., Pohnert, G., Vuylsteke, M. and Vyverman, W. (2013) Metabolomics Enables the Structure Elucidation of a Diatom Sex Pheromone. *Angewandte Chemie-International Edition* **52**, 854-857.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q.D., Chen, Z.H., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644-652.

Guillard, R.R.L. (1975) Culture of phytoplankton for feeding marine invertebrates. in *Culture of marine invertebrate animals* 29-60 (Springer).

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L. and White, O. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654-5666.

Heller, R., Manduchi, E., Grant, G.R. and Ewens, W.J. (2009) A flexible two-stage procedure for identifying gene sets that are differentially expressed. *Bioinformatics* **25**, 1019-1025.

Hu, C.A.A., Delauney, A.J. and Verma, D.P.S. (1992) A Bifunctional Enzyme (Delta-1-Pyrroline-5-Carboxylate Synthetase) Catalyzes the 1st 2 Steps in Proline Biosynthesis in Plants. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 9354-9358.

Huysman, M.J.J., Martens, C., Vandepoele, K., Gillard, J., Rayko, E., Heijde, M., Bowler, C., Inzé, D., Van de Peer, Y., De Veylder, L. and Vyverman, W. (2010) Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome Biology* **11**, R17.

Kall, L., Krogh, A. and Sonnhammer, E.L.L. (2007) Advantages of combined transmembrane topology and signal peptide prediction - the Phobius web server. *Nucleic Acids Research* **35**, W429-W432.

Kaupp, U.B., Hildebrand, E. and Weyand, I. (2006) Sperm chemotaxis in marine invertebrates - Molecules and mechanisms. *Journal of Cellular Physiology* **208**, 487-494.

Koch, K.W. and Dell'Orco, D. (2013) A Calcium-Relay Mechanism in Vertebrate Phototransduction. *Acs Chemical Neuroscience* **4**, 909-917.

Kolas, N.K. and Cohen, P.E. (2004) Novel and diverse functions of the DNA mismatch repair family in mammalian meiosis and recombination. *Cytogenetic and Genome Research* **107**, 216-231.

Lautru, S., Gondry, M., Genet, R. and Pernodet, J.L. (2002) The albonoursin gene cluster of S-noursei: Biosynthesis of diketopiperazine metabolites independent of nonribosomal peptide synthetases. *Chemistry & Biology* **9**, 1355-1364.

Li, W.Z. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659.

Liu, Y., Ruoho, A.E., Rao, V.D. and Hurley, J.H. (1997) Catalytic mechanism of the adenylyl and guanylyl cyclases: Modeling and mutational analysis. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 13414-13419.

Marahiel, M.A., Stachelhaus, T. and Mootz, H.D. (1997) Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chemical Reviews* **97**, 2651-2673.

McCarthy, D.J., Chen, Y.S. and Smyth, G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**, 4288-4297.

Ortiz-Castro, R., Diaz-Perez, C., Martinez-Trujillo, M., del Rio, R.E., Campos-Garcia, J. and Lopez-Bucio, J. (2011) Transkingdom signaling based on bacterial cyclodipeptides with auxin activity in plants. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 7253-7258.

Pawlowski, W.P., Sheehan, M.J. and Ronceret, A. (2007) In the beginning: the initiation of meiosis. *BioEssays* **29**, 511-514.

Potter, L.R., Abbey-Hosch, S. and Dickey, D.M. (2006) Natriuretic peptides, their receptors, and cyclic guanosine monophosphate-dependent signaling functions. *Endocrine Reviews* **27**, 47-72.

Potter, L.R. (2011) Guanylyl cyclase structure, function and regulation. *Cellular Signalling* **23**, 1921-1926.

Proost, S., Van Bel, M., Sterck, L., Billiau, K., Van Parys, T., Van de Peer, Y. and Vandepoele, K. (2009) PLAZA: A Comparative Genomics Resource to Study Gene and Genome Evolution in Plants. *Plant Cell* **21**, 3718-3731.

Richter, W., Unciuleac, L., Hermsdorf, T., Kronbach, T. and Dettmer, D. (2001) Identification of inhibitor binding sites of the cAMP-specific phosphodiesterase 4. *Cellular Signalling* **13**, 287-297.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140.

Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**.

Sato, S., Beakes, G., Idei, M., Nagumo, T. and Mann, D.G. (2011) Novel Sex Cells and Evidence for Sex Pheromones in Diatoms. *PLoS ONE* **6**.

Schaap, P. (2005) Guanylyl cyclases across the tree of life. *Frontiers in Bioscience-Landmark* **10**, 1485-U1485.

Seguin, J., Moutiez, M., Li, Y., Belin, P., Lecoq, A., Fonvielle, M., Charbonnier, J.B., Pernodet, J.L. and Gondry, M. (2011) Nonribosomal Peptide Synthesis in Animals: The Cyclodipeptide Synthase of Nematostella. *Chemistry & Biology* **18**, 1362-1368.

Shaffer, J.P. (1986) Modified Sequentially Rejective Multiple Test Procedures. *Journal of the American Statistical Association* **81**, 826-831.

Symington, L.S. (2002) Role of RAD52 epistasis group genes in homologous recombination and double-strand break repair. *Microbiology and Molecular Biology Reviews* **66**, 630-+.

Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y. and Vandepoele, K. (2013) TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. *Genome Biology* **14**.

Wu, T.D. and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873-881.

Supplementary data

Figure S1: (A) Sequence alignment of the catalytic core of the following adenylyl/guanylyl cyclases: BT_ACI_C1 (NP_776654.1), BT_ACI_C2 (NP_776654.1), BT_ACII_C2 (NP_112269.1),BT_sGCα (NP_786972.1), BT_sGCβ (NP_777066.1), DD_ACG (Q03101.2), HS_rGC1 (AAI48422.1) and SR_NC/PDE (BT = *Bos Taurus*, DD = *Dictyostelium discoideum*, HS = *Homo sapiens*, SR = *Seminavis robusta*). Brown: forskolin binding, light blue: Gsα, dark blue: Gβγ, red: catalytic sites, green: adenine, orange: guanine. Guanine binding residues are indicated in yellow in SR_NC/PDE and catalytic sites are indicated in red, non-conversed residue are underlined (adapted from Liu et al. (1997)). (B) Sequence alignment of part of the following PDE proteins: three cAMP-specific PDEs (PDE4, PDE7 and PDE8), three cGMP-specific PDEs (PDE5, PDE6 and PDE9) and the bifunctional *S. robusta* GC/PDE. The eight amino acids identified to be conserved only in cAMP-hydrolyzing PDEs are indicated in red.

gene	forward primer	reverse primer
V1[1]	CAAGGCAAGAAGGATGGCAAG	GCGACAAGCAAATTAGCAAACC
V4[1]	AGGCTACCGTGGGACTTGG	GATCTGGACTGCGCTGGTTC
CYCA/B1	GCTTGGGTCGATGCAATACT	ATCCGAGAGGGTCAAGAGGT
CYCB1	GGGATGTAACAAACGCCAAT	CCTCTCATCCATGACGGACT
P5CS	AATCGAAAACGGGGTAGCTT	CCCTCGACAATCTCAACCAT
GC/PDE	TCCTTCCCAAGTTTTCATGC	CATGGTTTGGTTGCTTGTTG

Table S1: Primers use in the RT-qPCR on CYCA/B1, CYCB1, P5CS and GC/PDE. [1] normalization genes, based on cDNA-AFLP (Valerie Devos, Unpublished results).

5

A genetic transformation protocol for the pennate diatom *Seminavis robusta*

Sara Moeys [1,2,3], Lieven De Veylder [2,3], Wim Vyverman [1]

and Marie J.J. Huysman [2,3]

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

[2] Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Gent, Belgium

[3] Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, B-9052 Gent, Belgium

Manuscript in preparation

Authors' contributions:

SM, LDV, WV and MJJH conceived and designed the study. SM performed the experiments and wrote the manuscript. MJJH gave advice about the experimental setup. LDV, WV and MJJH revised the manuscript.

Abstract

Diatoms represent a large and widespread group of microalgae. They thank their evolutionary success to some unique properties, like a chimeric genome resulting from secondary endosymbiosis, a silicified cell wall and a peculiar life cycle. Unfortunately, the most commonly used species to study the molecular biology of diatoms, do not display sex. Therefore, *Seminavis robusta* was introduced as a new model diatom species, because it exhibits a typical diatom life cycle and is easy to handle. A genetic transformation protocol for *S. robusta* would enable the application of reverse genetics in this species, allowing functional characterization of genes by creating overexpression or knock-down lines. An additional benefit of *S. robusta* is the ability to create double mutants by crossing two transgene strains, which is not possible in the current model diatoms due to the lack of a sexual cycle. Here, we describe a genetic transformation protocol for *S. robusta* making use of microparticle bombardment.

Introduction

Diatoms belong to one of the most species-rich and abundant classes of microalgae and are responsible for about 40% of the total oceanic carbon fixation (Granum et al., 2005, Armbrust, 2009). They possess a highly chimeric genome due to their evolution through endosymbiotic events (Armbrust et al., 2004, Bowler et al., 2008). This enabled them to develop several unique features that can partially explain their evolutionary success. One of these features is the way they control their life cycle. As a consequence of their rigid, silicified cell wall, the average cell size in a population decreases during the vegetative phase, ultimately leading to clonal death (Lewis, 1984, Chepurnov et al., 2004). This can be avoided by sexual reproduction, which can only be induced when cells are below a certain size threshold. During sexual reproduction, a specialized zygote, called the auxospore, is formed. This auxospore elongates, resulting in offspring of initial cell size. This cell size reduction-restitution mechanism is used as an endogenous clock to switch between cell division and cell differentiation and sexual reproduction (Lewis, 1984). Yet, little is known about the regulatory mechanisms and signalling pathways underlying this clock.

In contrast to the commonly used model diatom species *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*, for which sexual reproduction has never been observed, *Seminavis robusta* is a favourable species in diatom-life-cycle research (Chepurnov et al., 2008). Firstly, the induction of sex can be easily controlled, since cultures of different mating types must be mixed for gametogenesis to occur. Secondly, the cell cycle can be synchronized by prolonging the dark period to obtain high frequencies of cell pairing and gametogenesis. Furthermore, the cells are reasonably large (20 - 70 µm) and grow on the surface of plates or flasks, making it feasible to observe all different life-cycle stages under the low magnification of an inverted microscope. Additionally, transcriptome data is available (Chapter 3 and 4) as well as an in-house genome (unpublished data).

Despite all advantages mentioned above, one major drawback is that up to now reverse genetics is not possible in *S. robusta,* because no genetic transformation protocol is available. As reverse genetics significantly increases the potential of a model species, we developed a genetic transformation protocol for *S. robusta* using biolistic transformation. First, good selection markers were identified and endogenous promoters were cloned in gateway-compatible vectors. Then, the microparticle bombardment protocol was optimized.

Materials & methods

Strains and culture conditions

Seminavis robusta strains 85A (MT^+) and 85B (MT^-) were used, which are publicly available in the diatom culture collection of the Belgian Coordinated Collection of Micro-organisms (http://bccm.belspo.be, accession numbers DCG 0105 and DCG 0107, respectively). Cultures were grown in F/2 medium made with autoclaved filtered natural sea water (NSW) collected from the North Sea and Guillard's F/2 solution (Sigma-Aldrich) (Guillard, 1975). These cultures were grown at 18°C in a 12:12h light:dark regime with cool white fluorescent lamps at approximately 80 µmol photons m^{-2} s^{-1}.

Selection marker

Several antibiotics (nourseothricin, zeocin, chloramphenicol and kanamycin) were tested at different concentrations (50, 100, 500 and 1000 µg/mL). These tests were performed in F/2 medium made with 50% or 100% natural sea water and at different plate densities (25,000, 250,000 and 2,500,000 cells per plate). Cell viability was checked by microscopy three weeks after addition of the antibiotics.

Construction of a vector collection

Vectors are based on the Gateway-compatible destination vectors designed for *P. tricornutum* by Siaut et al. (2007) in which the *FcpB* promoter of *P. tricornutum*

(*Pt-FcpB*) was replaced by the endogenous *S. robusta fucoxanthin chlorophyll a/c protein FCPE* (*pFcpE*) and *histone H4* (*pH4*) promoters (sequences in Suppl. data).

The *FCPE* and *H4* genes were identified in the in-house draft genome of *S. robusta*, using the BLAST algorithm (Altschul et al., 1990). Fragments of 1002 bp upstream of the *FCPE* gene and 996 bp upstream of the *H4* gene were amplified with the following primers containing *Sac*I and *Xba*I restriction sites (underlined): 5'-GATCGAGCTCATCCAAATAGACAGCCTTG-3' and 5'-CATGTCTAGAGGTGGAAATT AGGTCTCTG-3' for *pFcpE*, and 5'-GATCGAGCTCTTCGTGATGTTCTGGGC-3' and 5'-GTCATGTCTAGACTTGTTGGTTGCTAAA GTAT-3' for *pH4*.

The pDEST-C-HA and pDEST-C-YFP vectors of Siaut et al. (2007) contain respectively three repeats of the haemagglutinin (HA) epitope or the gene coding for the enhanced yellow fluorescent protein (EYFP) C-terminal of the gateway cassette controlled by the *Pt-FcpB* promoter. In these vectors, the promoter was replaced by *pFcpE* or *pH4* through restriction with *Sac*I and *Xba*I (Promega) and ligation with T4 DNA ligase (Thermo Scientific), resulting in pDEST-pFcpE-C-HA or pDEST-pFcpE-C-YFP, and pDEST-pH4-C-HA or pDEST-pH4-C-YFP.

To introduce the enhanced cyan fluorescent protein (ECFP) as a C-terminal tag, pFcpB-ECFP (Siaut et al., 2007) was used as a template to amplify *ECFP* with the following primers, containing *Nco*I and *EcoR*I restriction sites (underlined): 5'-GATCCCATGGTGAGCAAGGGCGAGGA-3' and 5'-GATTCGAATTCTTACTTGTACAGCTCG TCCATGC-3'. Then, the YFP-tag of pDEST-N-YFP (Siaut et al., 2007) was replaced by this fragment. First, an LR-reaction (LR clonase® enzyme mix - Life Technologies) with pENTR-YFP and pDEST-N-YFP was conducted and the expression vector was selected for on ampicillin. Second, the *EYFP* gene of this expression vector was replaced by *ECFP* in a restriction-ligation reaction with *Nco*I and *EcoR*I. Third, the *Pt-FcpB* promoter was replaced by *pFcpE* and *pH4* by restriction-ligation with *Sac*I and *Xba*I. Finally, a BP reaction (BP clonase® enzyme mix - Life Technologies) with pDONR221 was performed to switch the *EYFP* gene again for the gateway cassette. To select for the pDEST-pFcpE-C-CFP and pDEST-pH4-C-CFP vectors, DB3.1 *Escherichia coli* cells

were transformed with the BP reaction product and transformants were selected on ampicillin and chloramphenicol.

Expression vectors

The selection marker genes *neomycin phosphotransferase II* (*nptII*) from *E. coli* K12 and *nourseothricin acetyltransferase* (*nat1*) from *Streptomyces noursei* were respectively amplified from the pK7WGF2 vector (Karimi et al., 2002) using the following AttB1- and AttB2-containing primers (AttB sites are underlined) 5'-GGGGACAAGTTTGT ACAAAAAAGCAGGCTTCATGATTGAACAAGATGGATTG-3' and 5'-GGGGACCACTTTGTACAA GAAAGCTGGGTCTCAGAAGAACTCGTCAAGAAG-3', and from pNat (Zaslavskaia et al., 2000) using the following AttB1- and AttB2-containing primers (AttB sites are underlined) 5'-GGGGACAAGTTTGTACAAAAAAGCAGGCTA TGACCACTCTTGACGACAC-3' and 5'-GGGGACCACTTTGTACAAGAAAGCTGGGTTCA GGGGCAGGGCATGCTCA-3'.

The *Histone H2B* gene (sequence in Suppl. Data) was amplified from *S. robusta* cDNA using the following AttB1- and AttB2-containing primers (AttB sites are underlined) 5'-GGGGACAAGTTTGTACAAAAAAGCAGGCTATGGCCAAGACACC ATCCAA-3' and 5'-GGGGACCA CTTTGTACAAGAAAGCTGGGTACGCACTGGAAA ACTTGGTAA-3'. These fragments were then introduced in pDONR221 using BP clonase®, resulting in entry vectors that were used in an LR reaction (LR clonase®) to create expression vectors. *NptII* and *nat1* were introduced in pDEST-pFcpE-C-HA and pDEST-pH4-C-HA, creating pFcpE-nptII, pFcpE-nat1, pH4-nptII and pH4-nat1. Since these genes contain a stop codon, the HA-tag will not be expressed. *H2B* was introduced in pDEST-pFcpE-C-YFP, pDEST-pH4-C-YFP, pDEST-pFcpE-C-CFP and pDEST-pH4-C-CFP, resulting in pFcpE-H2B-YFP, pH4-H2B-YFP, pFcpE-H2B-CFP and pH4-H2B-CFP.

Microparticle bombardment

About 500,000 cells were inoculated in a tissue culture flask (Cellstar, 175 cm^2 growth surface, Greiner Bio-One), containing 200 mL growth medium. After three

days of growth in a 12:12h light:dark regime, approximately 5 x 10^6 cells were spread on agar plates containing 50% NSW + F/2 medium. Microparticle bombardment was performed using a biolistic PDS-1000/He Particle Delivery System (BioRad). About 3 mg tungsten M-17 particles (Ø 1.1 µm, Bio-Rad) was coated with 5 µg plasmid (or 3 µg of each plasmid in co-transformation), as described in the PDS/1000 He instruction manual. The plates containing the *S. robusta* cells were positioned 6 cm below the stopping screen and a burst pressure of 1550 psi (or 1800 psi when testing conditions in first experiment) was applied. After the bombardment, 5 mL of 50% NSW + F/2 was added to the plates and they were kept for 48h in constant light, before being transferred to selective plates. These selective plates were kept for about four weeks in constant light to allow colonies to grow.

Validation of the transformants

After four weeks, colonies were picked up and streaked on a new selective plate as well as transferred to liquid selective medium. The selective medium was replaced two times before cells were transferred to non-selective medium.

Genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen). About 100 mL of culture (± 3 x 10^6 cells) was scraped from the bottom of the tissue culture flask (Cellstar, 75 cm^2 growth surface, Greiner Bio-One) and filtered on a Versapor filter (3 µm pore size, 25 mm diameter, PALL). Filters were frozen in liquid nitrogen. AP1 buffer (400 mL), 4 µL Rnase A (100 mg/mL) and silicon carbide beads (1.0 mm, BioSpec) were added to the filter. Cells were disrupted by beating on a bead mill (Retsch, 3 x 1 min). All other steps for the DNA extraction were done according to the manufacturer's instructions.

Total RNA was extracted using the RNeasy Plant Mini Kit (Qiagen). About 200 mL of culture (± 5 x 10^6 cells) was scraped from the bottom of a tissue culture flask (Cellstar, 175 cm^2 growth surface, Greiner Bio-One) and filtered on a Versapor filter (3 µm pore size, 25 mm diameter, PALL). Filters were frozen in liquid nitrogen. RLT buffer (1mL), β-mercaptoethanol (10µL) and silicon carbide beads (1.0 mm, BioSpec) were added to the filter. Cells were disrupted by beating on a bead mill (Retsch, 3 x

1 min). All other steps for the RNA extraction were done according to the manufacturer's instructions. cDNA was prepared using the iScript cDNA Synthesis Kit (Bio-Rad) according to the manufacturer's instructions.

To verify the presence of the inserted genes, PCR was performed on genomic DNA. As a negative control, gDNA extracted from a wild type strain was used. As a positive control, the vector used for transformation was added. Primers can be found in supplementary data. To check if the transgenes are expressed, a PCR on cDNA (RT-PCR) was performed. cDNA from a wild type strain was used as a negative control. As a positive control, the vector used for transformation was added.

To verify the fluorescent signal in the transformants containing a fluorescent tag (YFP or CFP), cells were examined under an Axiovert 135M microscope using a chroma 41028 (EX = 500 nm, BS = 515 nm, EM = 535 nm) and a CFP (EX = 425 - 440 nm, EM = 460 - 500 nm) filter.

Results

Selection marker

To set up a genetic transformation protocol, a good selection marker is needed. For this, the sensitivity of *S. robusta* to several antibiotics was examined (Table 1). *S. robusta* is very sensitive to chloramphenicol, but this is possibly due to the ethanol in which the chloramphenicol is dissolved, as cell death was also observed on the control plates, for which the same amount of ethanol was added to the medium. Therefore, chloramphenicol cannot be used to select transformants. *S. robusta* seems rather resistant to zeocin, excluding this antibiotic. In contrast, the cells are sensitive to nourseothricin and kanamycin. Based on the test results, kanamycin at a concentration of 300 µg/mL and nourseothricin at a concentration of 30 µg/mL in 50% natural seawater were chosen for the selection of transformants.

		100% seawater			50% seawater			
	µg/mL	2,50 x	2,50 x 10^5	2,50 x 10^6	2,50 x 10^4	2,50 x 10^5	2,50 x 10^6	cells/plate
Chloram-phenicol	0	-	-	+	+	+	+	
	50	-	-	-	+	-	-	
	100	-	-	-	-	-	-	
	500	-	-	-	-	-	-	
	1000	-	-	-	-	-	-	
zeocin	0	ND	ND	ND	++	++	++	
	50	ND	ND	ND	+	+	+	
	100	ND	ND	ND	+	+	+	
	500	ND	ND	ND	+	+	+	
	1000	ND	ND	ND	+	+	+	
kanamycin	0	+	+	+	++	++	++	
	50	+	++	++	+	+	+	
	100	++	++	++	++	++	++	
	200	ND	ND	ND	++	++	-	
	300	ND	ND	ND	+	-	-	
	400	ND	ND	ND	+	-	-	
	500	+	+	+	-	-	-	
	1000	+	+	+	-	-	-	
nourseothricin	0	ND	ND	ND	++	++	++	
	25	ND	ND	ND	+	+	+	
	50	ND	ND	ND	-	-	+	
	100	ND	ND	ND	-	-	+	
	200	ND	ND	ND	-	-	-	
	300	ND	ND	ND	-	-	-	
	400	ND	ND	ND	-	-	-	
	500	ND	ND	ND	-	-	-	

Table 1: The sensitivity of *S. robusta* to chloramphenicol, zeocin, kanamycin and nourseothricin was examined in 100% and 50% natural seawater + F/2 nutrients at different plate densities (2,5 x 10^4, 2,5 x 10^5 and 2,5 x 10^6 cells/plate). ++ indicates living cells, + means that many cells are death and - that all cells are death. ND = not done.

Vector collection

To be able to select for transformants on kanamycin or nourseothricin, a resistance gene (*ntpII* or *nat1* respectively) should be introduced in a vector under the control of a strong promoter. For this, two endogenous promoter regions were selected in the *S. robusta* draft genome, namely the *fucoxanthin chlorophyll a/c protein FcpE* (*pFcpE*) and *histone H4* (*pH4*) promoters. Homologs for the *P. tricornutum FcpB* and *histone H4* genes were identified using blast (Altschul et al., 1990) and those with high expression levels in a transcriptomic dataset covering diverse life cycle stages (Chapter 3) were selected (cpm values in Suppl. data).

Fragments of about 1000 bp upstream of the respective gene were arbitrarily selected and were introduced in gateway-compatible vectors with a C-terminal tag, i.e. a haemagglutinin epitope or one of the two fluorescent tags, enhanced yellow fluorescent protein (eYFP) and enhanced cyan fluorescent protein (eCFP). The gateway® cloning system (Life technologies) was used because it allows easy introduction of any gene to be expressed in the transformed cells.

The pDEST-C-HA and pDEST-C-YFP vectors designed by Siaut et al. (2007) for the genetic transformation of *P. tricornutum* were used as backbone. First, the *P. tricornutum* promoter was replaced by the endogenous *S. robusta* promoters, resulting in the pDEST-pFcpE-C-HA, pDEST-pH4-C-HA, pDEST-pFcpE-C-YFP and pDEST-pH4-C-YFP vectors (Figure 1). Then, a cyan fluorescent tag, eCFP, was introduced in the vector collection. For this, the *EYFP* gene of pDEST-C-YFP (Siaut et al., 2007) was replaced by *ECFP* and the *Pt-FcpB* promoter was replaced by *pH4* or *pFcpE* to create the pDEST-pFcpE-C-CFP and pDEST-pH4-C-CFP vectors (Figure 1).

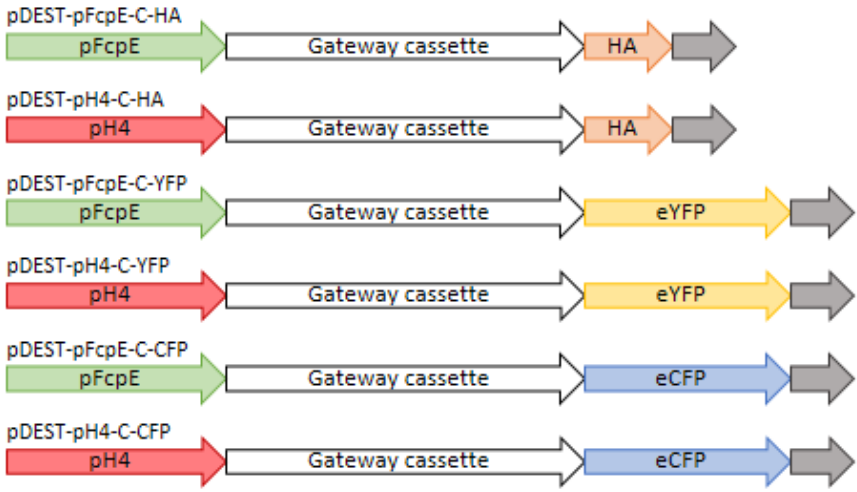


Figure 1: A schematic representation of the destination vectors created for the genetic transformation of *S. robusta*. The promoter (*pFcpE* or *pH4*) was integrated in front of the gateway cassette (containing attB1 and attB2 sites, a chloramphenicol resistance gene and a *ccdB* gene), followed by a tag (HA, eYFP or eCFP) and finally the terminator sequence of the *P. tricornutum FCPA* gene (in grey).

Microparticle bombardment

To date, genetic transformation of diatoms is mostly done using microparticle bombardment (Apt et al., 1996, Poulsen et al., 2006, Buhmann et al., 2014). In this method, also called biolistic transformation, tungsten or golden particles are coated with DNA and fired under high pressure onto the cells in a vacuum chamber.

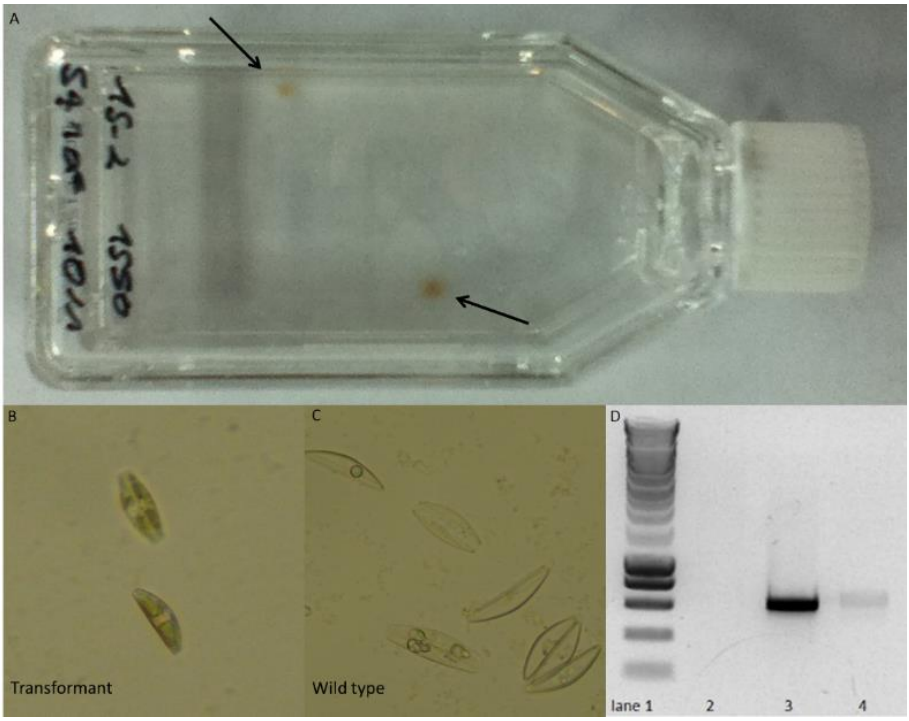


Figure 2: (A) Two colonies growing on selective medium for four weeks after transformation with pFcpE-nptII. A transformant (B) and wild type cells (C) in selective medium under inverted microscope. (D) PCR analysis of *S. robusta* 85A transformants after particle bombardment with pFcpE-nptII. The PCR product is a 620 bp fragment of nptII. lane 1: SmartLadder (Eurogentec), lane 2: wild type *S. robusta* 85A, lane 3: putative transformant, lane 4: positive control (FcpE-nptII vector).

To determine optimal biolistic bombardment conditions, a first experiment was conducted with the pFcpE-nptII vector. Here, the amount of plasmid DNA and the pressure were varied. Transformants were selected in liquid 50% NSW+F/2 medium with 300 µg/mL kanamycin. After four weeks, one colony was found in the culture

transformed with 5 µg DNA at 1800 psi and two colonies were found in the culture transformed with 10 µg DNA at 1550 psi (Figure 2A). All putative transformants showed normal growth when transferred to fresh selective medium (Figure 2B/C). The presence of the *nptII* gene could be confirmed by PCR on genomic DNA, extracted from the three transformants (example for one transformant in Figure 2D). As no large difference was seen between the different conditions, in future experiments 1550 psi and 5 µg plasmid DNA were used. These are the same conditions reported for the transformation of *P. tricornutum* (Apt et al., 1996). In later experiments, no difference in efficiency was observed when selection was done on agar plates instead of liquid medium (results not shown). Therefore, selection was carried out on 50%NSW+F/2+1% agar plates in the next experiments.

Selection marker	Gene-of-interest	Strain	Repl.	# Colonies	Surviving
pH4-nptII	pFcpE-H2B-YFP	85A	1	7	5
pH4-nptII	pFcpE-H2B-YFP	85A	2	0	NA
pFcpE-nptII	pH4-H2B-YFP	85A	1	0	NA
pFcpE-nptII	pH4-H2B-YFP	85A	2	2	2
pH4-nptII	pFcpE-H2B-CFP	85B	1	2	1
pH4-nptII	pFcpE-H2B-CFP	85B	2	4	3
pFcpE-nptII	pH4-H2B-CFP	85B	1	0	NA
pFcpE-nptII	pH4-H2B-CFP	85B	2	3	2
pH4-nat1	pFcpE-H2B-YFP	85B	1	0	NA
pH4-nat1	pFcpE-H2B-YFP	85B	2	0	NA
pFcpE-nat1	pH4-H2B-YFP	85B	1	0	NA
pFcpE-nat1	pH4-H2B-YFP	85B	2	0	NA
pH4-nat1	pFcpE-H2B-CFP	85A	1	4	1
pH4-nat1	pFcpE-H2B-CFP	85A	2	4	1
pFcpE-nat1	pH4-H2B-CFP	85A	1	1	0
pFcpE-nat1	pH4-H2B-CFP	85A	2	0	NA

Table 2: An overview of the co-transformation experiments. A selection marker (*nptII* or *nat1*) under control of the *FcpE* or *H4* promoter was combined with a eYFP- or eCFP-tagged *H2B* behind the *FcpE* or *H4* promoter. Two strains (85A and 85B) were transformed. For every vector combination, there were two replicates. The number of colonies after four weeks is indicated, as well as the number of putative transformants that survived when transferred to fresh selective medium. NA = Not Applicable.

To validate the destination vectors containing a eYFP- or eCFP-tag, *S. robusta* was co-transformed with a selection vector and an expression vector containing the *H2B* gene in front of the eYFP- or eCFP-tag. Based on the number of kanamycin- or nourseothricin-resistant lines, efficiency was estimated to be 4 out of 10 x 10^6 cells for kanamycin and 5 out of 100 x 10^6 cells for nourseothricin, although the number of colonies per plate was quite variable (Table 2). Despite the fact that the presence of the transgenes could not be confirmed by PCR for all transformants, all transfomants that were selected for RT-PCR showed expression of eYFP- or eCFP-tagged *H2B* (results shown in Suppl. Data). Unfortunately, no fluorescent signal could be detected.

Conclusion

We have shown that *S. robusta* can be stably transformed using microparticle bombardment. After testing several antibiotics, the *neomycin phosphotransferase II* (*nptII*) and the *nourseothricin acetyltransferase* (*nat1*) gene, conferring respectively kanamycin (300 µg/mL) or nourseothricin (30 µg/mL) resistance, were chosen as selection markers in *S. robusta*. Co-transformation is possible using one selection vector and one vector with a gene-of-interest. Although the expression of the *H2B-YFP* or *H2B-CFP* fusion gene could be confirmed by RT-PCR, no fluorescent signal was observed. This could be due to low transcription or translation of the tagged gene. Since only tagged H2B was tried, it should be checked if the same problem exists when transforming a different tagged gene. Another explanation could be a different codon usage of *S. robusta* compared to *P. tricornutum*. This could then lead to a bad translation of the tags. Once the *S. robusta* genome is structurally annotated, the codon usage can be determined. Next to the vectors with C-terminal tag, also vectors for N-terminal tagging should be developed, because the position of the tag can sometimes cause misfolding of the protein (Christensen et al., 2009).

The efficiency of transformation is still quite low. To improve this, cell density could be increased to yield more positive colonies. Furthermore, the vectors could be linearized before transformation, because this proved to increase efficiency in *P.*

tricornutum and *Nannochloropsis* transformation (Kilian et al., 2011, Miyahara et al., 2013). Additionally, an electroporation protocol could be tested to transform *S. robusta*. Such a protocol has recently been developed for *P. tricornutum* (Miyahara et al., 2013, Zhang and Hu, 2014).

Furthermore, RNA interference should be tested in *S. robusta*, since this will be important for the functional characterization of interesting genes. In *P. tricornutum*, antisense and inverted repeat sequences are used to make knock-down lines (De Riso et al., 2009). These type of constructs should be put under the control of one of the endogenous promoters identified in this study to be used for the biolistic bombardment of *S. robusta*.

Literature cited

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403-410.

Apt, K.E., KrothPancic, P.G. and Grossman, A.R. (1996) Stable nuclear transformation of the diatom Phaeodactylum tricornutum. *Molecular & General Genetics* **252**, 572-579.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., Brzezinski, M.A., Chaal, B.K., Chiovitti, A., Davis, A.K., Demarest, M.S., Detter, J.C., Glavina, T., Goodstein, D., Hadi, M.Z., Hellsten, U., Hildebrand, M., Jenkins, B.D., Jurka, J., Kapitonov, V.V., Kröger, N., Lau, W.W.Y., Lane, T.W., Larimer, F.W., Lippmeier, J.C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M.S., Palenik, B., Pazour, G.J., Richardson, P.M., Rynearson, T.A., Saito, M.A., Schwartz, D.C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F.P. and Rokhsar, D.S. (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79-86.

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

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J.A., Brownlee, C., Cadoret, J.-P., Chiovitti, A., Choi, C.J., Coesel, S., De Martino, A., Detter, J.C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M.J.J., Jenkins, B.D., Jiroutova, K., Jorgensen, R.E., Joubert, Y., Kaplan, A., Kröger, N., Kroth, P.G., La Roche, J., Lindquist, E., Lommer, M., Martin-Jézéquel, V., Lopez, P.J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L.K., Montsant, A., Oudot-Le Secq, M.-P., Napoli, C., Obornik, M., Parker, M.S., Petit, J.-L., Porcel, B.M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T.A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M.R., Taylor, A.R., Vardi, A., von Dassow, P., Vyverman, W., Willis, A., Wyrwicz, L.S., Rokhsar, D.S., Weissenbach, J., Armbrust, E.V., Green, B.R., Van de Peer, Y. and Grigoriev, I.V. (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239-244.

Buhmann, M.T., Poulsen, N., Klemm, J., Kennedy, M.R., Sherrill, C.D. and Kroger, N. (2014) A Tyrosine-Rich Cell Surface Protein in the Diatom Amphora coffeaeformis Identified through Transcriptome Analysis and Genetic Transformation. *PLoS ONE* **9**.

Chepurnov, V.A., Mann, D.G., Sabbe, K. and Vyverman, W. (2004) Experimental studies on sexual reproduction in diatoms. *International Review of Cytology* **237**, 91-154.

Chepurnov, V.A., Mann, D.G., von Dassow, P., Vanormelingen, P., Gillard, J., Inzé, D., Sabbe, K. and Vyverman, W. (2008) In search of new tractable diatoms for experimental biology. *BioEssays* **30**, 692-702.

Christensen, T., Amiram, M., Dagher, S., Trabbic-Carlson, K., Shamji, M.F., Setton, L.A. and Chilkoti, A. (2009) Fusion order controls expression level and activity of elastin-like polypeptide fusion proteins. *Protein Science* **18**, 1377-1387.

De Riso, V., Raniello, R., Maumus, F., Rogato, A., Bowler, C. and Falciatore, A. (2009) Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic Acids Research* **37**.

Granum, E., Raven, J.A. and Leegood, R.C. (2005) How do marine diatoms fix 10 billion tonnes of inorganic carbon per year? *Canadian Journal of Botany-Revue Canadienne De Botanique* **83**, 898-908.

Guillard, R.R.L. (1975) Culture of phytoplankton for feeding marine invertebrates. in *Culture of marine invertebrate animals* 29-60 (Springer).

Karimi, M., Inze, D. and Depicker, A. (2002) GATEWAY((TM)) vectors for Agrobacterium-mediated plant transformation. *Trends in Plant Science* **7**, 193-195.

Kilian, O., Benemann, C.S.E., Niyogi, K.K. and Vick, B. (2011) High-efficiency homologous recombination in the oil-producing alga Nannochloropsis sp. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 21265-21269.

Lewis, W.M. (1984) The Diatom Sex Clock and Its Evolutionary Significance. *American Naturalist* **123**, 73-80.

Miyahara, M., Aoi, M., Inoue-Kashino, N., Kashino, Y. and Ifuku, K. (2013) Highly Efficient Transformation of the Diatom Phaeodactylum tricornutum by Multi-Pulse Electroporation. *Bioscience Biotechnology and Biochemistry* **77**, 874-876.

Poulsen, N., Chesley, P.M. and Kröger, N. (2006) Molecular genetic manipulation of the diatom *Thalassiosira pseudonana* (Bacillariophyceae). *Journal of Phycology* **42**, 1059-1065.

Siaut, M., Heijde, M., Mangogna, M., Montsant, A., Coesel, S., Allen, A., Manfredonia, A., Falciatore, A. and Bowler, C. (2007) Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. *Gene* **406**, 23-35.

Zaslavskaia, L.A., Lippmeier, J.C., Kroth, P.G., Grossman, A.R. and Apt, K.E. (2000) Transformation of the diatom Phaeodactylum tricornutum (Bacillariophyceae) with a variety of selectable marker and reporter genes. *Journal of Phycology* **36**, 379-386.

Zhang, C.Y. and Hu, H.H. (2014) High-efficiency nuclear transformation of the diatom Phaeodactylum tricornutum by electroporation. *Marine Genomics* **16**, 63-66.

Supplementary data

Seminavis robusta sequences

The *pFcpE* and *pH4* sequences after PCR with primers containing *Sac*I and *Xba*I restriction sites (underlined). The *H2B* sequence after PCR with primers containing AttB sites (underlined).

>pFcpE

GATCGAGCTCATCCAAATAGACAGCCTTGTTGTCGTCTCCCAAGCAACATCAATCATCCAAAATCGTT
CGTTCTCCAAAGTCTTGAGGAGGTCTAAGCGCGTTCACCCCGGATGCGACGAGCCAACTGAATATCT
TTAGGCATGATTGTCACGCGCTTGGCGTGAATAGCACAAAGATTCGTATCCTCGAACAATCCAACCA
AGTAGGCTTCCGATGCCTCCTGAAGAGCAAGGAGTGCCGTGCCTTGGAACCTGCAAGAATGAACAA
GGAGAAGGTGTGATGTGAGTTCGCTGATTGTGCCATGGTGGCAATCATTTCCAATCTCTTCCAACCCA
ACAAGACATGGTACTCACCGGACGTGGGCCTTGAAATCTTGGGAAATCTCACGAACAAGGCGTTGG
AAGGGAAGCTTGCGGAGCAACAAATCAGTACTCTTCTGGTATCGACGAATCTGACGGATAGCAACA
GTACCAGGGCGGTAACGGTGGGGCTTCCGGACACCTCCCTGCCGAGGGGCACGCTCTCGAGCAGCT
ATAGTTCGCAGTTCCATGCGTGGAGCCCTTCCTCCGGTGGACTTACGGGCGGTTTGCTTGGTGCGAG
CCATGGTGAACGAAATAGGTAGAGATAAAGATGAAGATGAAGACGAAGGTAGTTTCAACTTTGCCG
AGATCGGATTATGAGAGAATTGTGAGAGGCGGATTTGAGGGTGGTATTTCGAAATCAGAAAAGTAT
TTGAAATTTTGGCCACTCGAAATGTCGCCTGCCGAGGTGTGGCCGGCTTTGAGCGGACTCTTTTGGC
GGCCCCTTTTGGCGCGGATCTTAGCCCTCTCTCTAAATATGTGACGTCATGATTTAAATCGGGACTGA
ATTCGATTCGTCTTTCAGTTGTTGGTTGAAGGTGGAGAAGCGTATCTTCCATTCTTCGACCTCAACTTG
CCAAGAAGGAACCAAAAGAGGCAAATCGAATGTATGCATCGTTCACAATTTTGCCAGAGACCTAATT
TCCACCTCTAGACATG

>pH4

GATCGAGCTCTTCGTGATGTTCTGGGCGGCACCGAACCAAGGCTGTAGCAATGACAACGACACAACC
AATGCAAAAGATATCCCCCGTAGTCCACTCTGATTGGTATGCCGATACATCAGAGGCTCTACCTCTCA
ACACAATCACAACGGAAGATAGATACTCGGAAGTAGAGCTGGTCTGTTTTATGGGTTTGTTGGCAGC
ACTGAGCTGATCGGTACAGTATCTTTGCCTGGACCACTTTGAACGCACGAAAGATCGCTAAAAGACA
TCGAAAAGGCTTTTCATGTCATGCTGAACAGCGCACCTCGAAATCTAGGTGGTGTAAAACGGCTTCC
ATTGCTAATTGCTCCCCACGGAACCGCGCTGTTGTGAGCCCCCTCTTCCGAGCCAGTGAGAGAGATC
AATCACTGACATCTAGCTAGCTACTAAAGAAAGTTTTTGCCTTGACCCTTTTTATTGATATCTACCTGT
TGGGTACCATGGAGTACGTAATGTTTGTTAGTTGTTGAACCGCTTAACCACCCCGCTTTCTTCACCCCA
AGGAGGTGTGGGGGATCCTTGAAATTTGGACACTCTTGGAAAGGGATACAGCAAACAAATGATAGC
CTGCTAACGTGGTGTGAACTATCATGCTACAGTGTAGTATGGTACCCAAATATGTCAACATCCAATCT
AATCTATCAATACTCTATCAATCAACTACCGGCACCGTACCGGAATGTAAAGTAAGCTTGAAAGAGA
GGCTGACTGAATGATTAATCGACAGGAGTCCAACACATGACGTAATCACAACTTCACTGGCTCTGTTT
TGTTTTTTAAACTGAGCGGTGCTTGCGTTGGCGCTTGTGCTTGGCTGGCGGCACCCGATGCAGCCGC
CATTCTCCTTCCTCTTTTCTTTTGCTGCTGCTGCTCTCATCTCACAATCCCAGCCTCGAGAAACTTGGAA
AGCACTTTCCAAACTCCTCCGTTCTACGCTTACTTTGATACTTTAGCAACCAACAAGTCTAGACATGAC

>H2B

GGGGACAAGTTTGTACAAAAAAGCAGGCTATGGCCAAGACACCATCCAAGCAATCCGCCAAGACCC
CCAAGAAGGCTGCTGGTGGCTCCAAGAAGTCCAAGAAGCGTACCGAGACCTATTCCTCCTACATCTA
CAAGGTGTTGAAGCAAGTCCATCCCGATACTGGTATCTCCAAGAAGGGCATGTCCATCATGAACTCT
TTCATCAATGATATTTTCGAACGCATCGCCACGGAAGCTGGAAAGCTTGCCACCTACAACAAGAAGG
CCACCTTGAGTAGCCGTGAAATCCAGACCGCCGTTCGTTTGATGCTCCCTGGAGAGTTGGCTAAGCA
TGCTGTCAGTGAGGGCACGAAGGCTGTTACCAAGTTTTCCAGTGCGTACCCAGCTTTCTTGTACAAA
GTGGTCCCC

Expression of *FcpE* and *Histone H4* in *S. robusta*

Cpm-values for eleven libraries of RNA-seq: MT^- below the SST in G_1 (85BS_G_1), S (85BS_S) or G_2/M phase (85BS_G_2M), MT^- above the SST in S phase (85BL_S), MT^+ below the SST in S (85AS_S) or G_2/M phase (85AS_G_2M), MT^+ above the SST in S phase (85AL_S) and sexual samples during mating, zygote formation, auxosporulation and initial cell formation.

gene	85BS_G_1	85BS_S	85BS_G_2M	85BL_S	85AS_S	85AS_G_2M	85AL_S
FcpE	258,0136	398,9	260,5673	1368,821	876,2468	565,8118	898,6808
H4	25,79165	51,97544	31,14856	42,97951	158,056	66,95371	136,3394

gene	mating	zygotes	auxospores	Initial cells
FcpE	172,2489	12,28602	56,07871	62,60364
H4	286,7054	136,9052	107,6215	104,2336

Primers for verification PCR

gene	Forward primer	Reverse primer	bp
nptII	GGCTATTCGGCTATGACTGG	AGCCAACGCTATGTCCTGAT	620
nat1	AAAAAAGCAGGCTTCATGACCACTCTTGACGAC A	CAGGGCATGCTCATGTAGAG	560
H2B-YFP/CFP	GAAGCAAGTCCATCCCGATA	GAACTTCAGGGTCAGCTTGC	380
actin	GATTGTCTTGGCAGGTGGAT	GCCACCAACATAGGCAGAAT	200

Validation co-transformation

Overview of the validation results of the co-transformation experiment. For every vector combination, the colony number and the number used in the Figures S1 and S2 are indicated. Furthermore, the presence of a PCR fragment of the right size after PCR on genomic or cDNA is indicated. No = not present, yes = present, ND = not done.

Transformants				gDNA		cDNA
selection	**gene-of-interest**	**colony**	**nr**	**marker**	**gene**	**gene**
pFcpE-nptII	pH4-H2B-YFP	1	1	no	no	yes
		2	2	no	no	yes
pH4-nptII	pFcpE-H2B-YFP	1	3	no	no	ND
		2	4	no	no	yes
		3	5	no	no	ND
		4	6	no	no	ND
		5	7	no	no	ND
		6	8	no	no	ND
pH4-nat1	pFcpE-H2B-CFP	1	9	no	no	yes
		2	10	yes	no	yes
pFcpE-nptII	pH4-H2B-CFP	1	11	yes	no	yes
		2	12	yes	no	ND
pH4-nptII	pFcpE-H2B-CFP	1	13	yes	yes	ND
		2	14	yes	no	yes
		3	15	yes	no	ND
		4	16	yes	no	ND

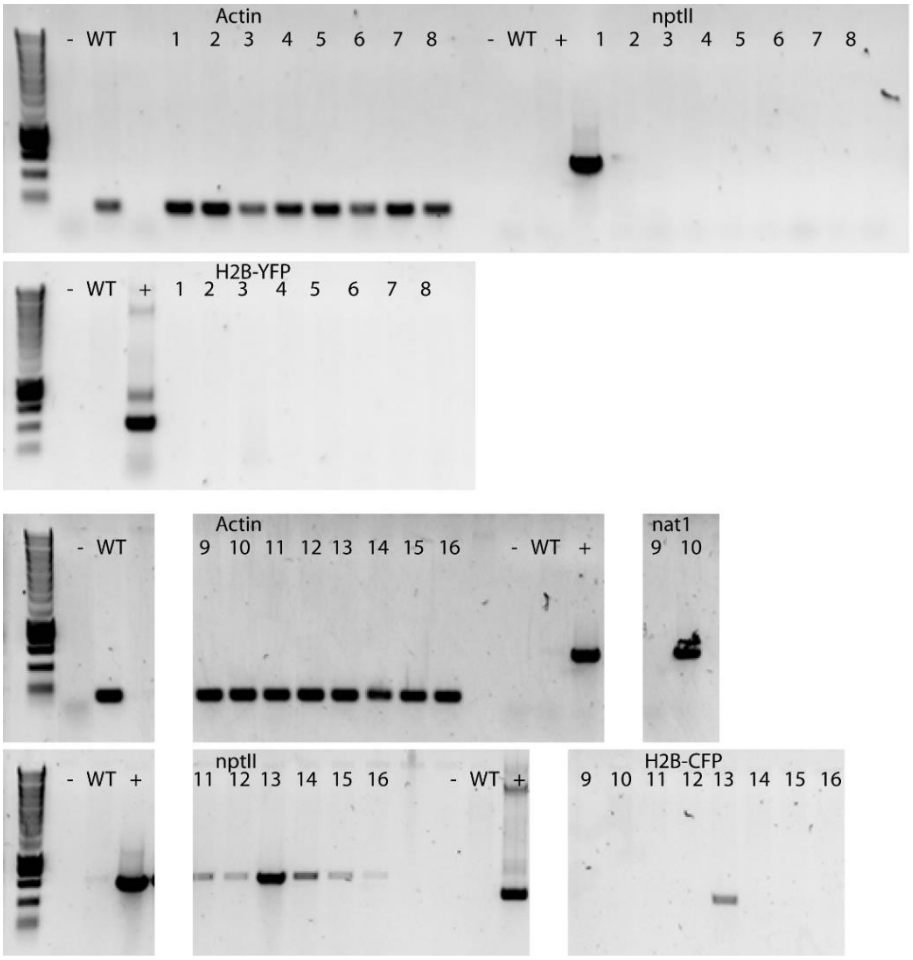


Figure S1: Images of the agarose gels after validation PCRs on genomic DNA with primers for *actin*, *nptII*, *nat1* and *H2B-YFP* or *H2B-CFP*. The first lane shows SmartLadder (Eurogentec). "-" = no DNA added, "WT" = wild type strain, "+" = positive control (vector), for the numbering, see table above.

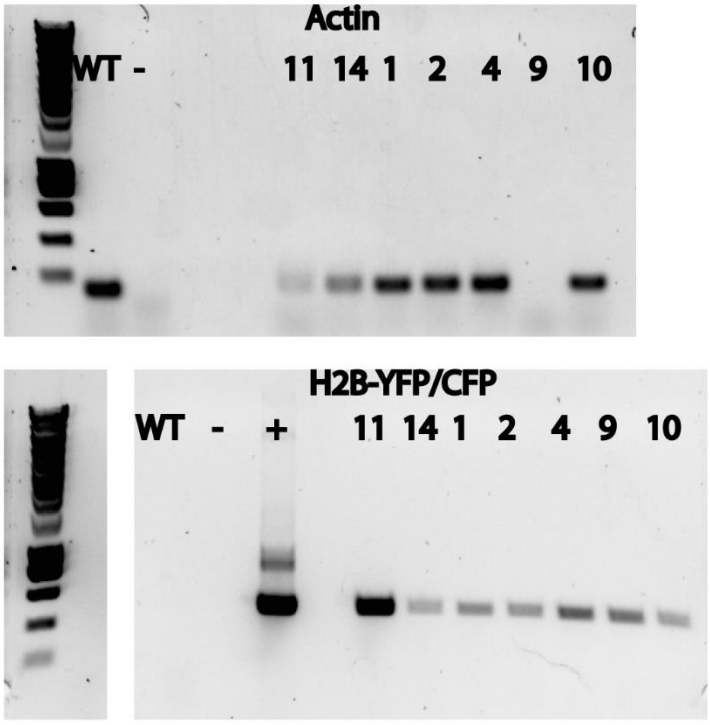


Figure S2: Images of the agarose gels after validation RT-PCRs with primers for *actin* and *H2B-YFP* or *H2B-CFP*. The first lane shows SmartLadder (Eurogentec). "WT" = wild type strain, "-" = no DNA added, "+" = positive control (vector), for the numbering, see table above.

6
General discussion

This thesis aimed to contribute to the knowledge on the life cycle of diatoms and more specifically on the molecular and genetic regulation of sexual reproduction. *Seminavis robusta* was used as a model system for pennate diatoms. By fine-mapping the mating type (MT) locus and combining this with the draft genome of *S. robusta*, it could be determined that only one gene resides on the MT locus and that it belongs to the *DNMT5* family of DNA methyltransferases (Chapter 2). These findings suggest a role for DNA methylation in MT determination in *S. robusta*. A transcriptome study covering the most important stages in the life cycle of *S. robusta* allowed us to identify genes involved in mitosis, meiosis and silica deposition and to get a first impression of their expression patterns (Chapter 3). Gillard et al. (2013) described the multistep pheromone system that *S. robusta* uses to initiate sexual reproduction, with a focus on the attraction pheromone diproline. In chapter 4, the focus lies on the conditioning factor produced by MT^+ (CF-P) that induces cell cycle arrest and diproline biosynthesis in MT^- cells. The genome and transcriptome data generated in the course of this thesis, together with the genetic transformation protocol that has been developed (Chapter 5), contribute to the further establishment of *S. robusta* as a model system in diatom research.

The involvement of DNA methylation in mating type determination in *Seminavis robusta*

S. robusta is, like most pennate diatoms, heterothallic, meaning that (almost) no intraclonal sexual reproduction occurs. This implies the existence of different mating types. By making interclonal crosses, clones can be subdivided in two groups, which correspond to two different mating types (Chepurnov et al., 2002). Although this is generally accepted to be the most common mating system, species with more than

two mating types exist, for example in ciliates (*Tetrahymena* or *Euplotes*) or fungi (*Schizophyllum*) (Whitfield, 2004, Phadke and Zufall, 2009).

In *S. robusta*, the two mating types are physiologically different (Gillard et al., 2013). A migrating (MT^+) and an attracting (MT^-) mating type can be distinguished in mixed cultures. Using AFLP-based linkage mapping, it could be demonstrated that the mating type is determined by one genetic locus in *S. robusta* (Vanstechelman et al., 2013). MT^+ was shown to be the heterogametic mating type, while MT^- is homogametic. Furthermore, the MT locus appears to be surrounded by a large region of recombination.

Here, a bulked segregant analysis combined with AFLP and whole genome sequencing was used to identify markers that are closely linked to the MT locus. By mapping the two most-closely linked markers to a draft genome, it could be determined that a member of the DNA methyltransferase 5 (*DNMT5*) gene family lies within the MT locus (Chapter 2). Besides in diatoms, *DNMT5* is also found in divergent marine algae, for example in the pelagophyte *Aureococcus*, in the coccolithophore *Emiliania huxleyi* and in green algae, which have in common that they lack *DNMT1* (Huff and Zilberman, 2014). DNMT1 is a maintenance DNA methyltransferase that copies CG methylation after DNA replication in many eukaryotes (Law and Jacobsen, 2010). It was shown that DNMT5 exhibits symmetrical CG methylation in species lacking DNMT1 (Huff and Zilberman, 2014). This suggests that DNMT5 is important for the inheritance of DNA methylation patterns.

S. robusta has two paralogs of the *DNMT5* gene, suggesting that one of these paralogs (*DNMT5a*) specialized in MT determination. We hypothesize that the DNMT5a protein is responsible for the correct inheritance of MT-specific DNA methylation patterns after DNA replication and mitosis/meiosis. Probably, target genes of DNMT5a are differentially methylated between MT^+ and MT^- cells. This differential methylation could result in differential gene expression or alternative splicing and thus change the activity (Jones, 2012, Maunakea et al., 2013). These

target genes could be for example involved in the production or reception of the MT-specific conditioning factors (Chapter 4). In order to identify target genes, genome-wide bisulfite sequencing or methylation-sensitive AFLP could be used to find genes that are differentially methylated between MT^+ and MT^- strains.

Also in other eukaryotic lineages, methylation is implicated in sex/mating type determination. For example in maize, it was shown that *Rmr6*, which maintains cytosine methylation patterns, is necessary to maintain monoecy through the repression of female sexual development (Parkinson et al., 2007). Furthermore, in *Melandrium album,* suppression of female sexual development in XY males is dependent on DNA methylation of specific sequences. This suppression can be heritably modified by treatment with the hypomethylating drug, 5-azacytidine (Janousek et al., 1996). In fish, methylation plays a role in temperature-dependent sex determination. In the European sea bass, *Dicentrarchus labrax*, a rise in temperature early in life leads to increased DNA methylation of the *cyp19a* promoter, which leads to decreased expression. Cyp19a is an aromatase that converts androgens into estrogen. Because the androgen-to-estrogen ratio determines whether the gonads of non-mammalian vertebrates develop into testes or ovaries, decreased expression of *cyp19a* leads to masculinization. Also in the half-smooth tongue sole, *Cynoglossus semilaevis*, methylation is important in triggering sex-reversal (Shao et al., 2014). An increase in temperature during the sensitive developmental stage can results in sex reversal of ZW females to phenotypic males (pseudomales) (Chen et al., 2014). In addition, offspring of pseudomales can reverse sex even at lower temperatures, an indication that sex reversal is inheritable. Genes involved in sex determination are differentially methylated in females compared to males and pseudomales (Shao et al., 2014).

Also in *S. robusta*, signs of mating type switching have been observed. Sporadically, sexual reproduction can be seen in monoclonal MT^+ cultures (Victor Chepurnov, Barbara Bouillon, personal communication). Since this is only observed in one of the mating types, mating type switching seems to be linked to the MT locus. Our hypothesis is that under specific circumstances the methylation of DNMT5a-

targets is changed, causing a subpopulation to switch from MT^+ to MT^-. This subpopulation is consequently able to mate with clones that did not switch mating type.

A deep transcriptomic investigation of diatom life cycle regulation

The genetic difference between the two mating types at the MT locus leads to different behavior at several stages of the life cycle. As long as the cells are above the sexual size threshold (SST), no physiological differences have been observed. Once below the SST, a distinction can be made between the two mating types. Both mating types then start the production of a conditioning factor to signal their presence to the other mating type. These cells are only responsive when they are themselves below the SST. As the conditioning factors only have an effect on the other mating type, they must be at least a little divergent (Chapter 4). Both mating types also show a different response to the conditioning factor secreted by the opposite mating type. Although they both arrest their cell cycle, only MT^- produces diproline (Chapter 4 and Gillard et al. (2013)). This diproline is a pheromone that attracts conditioned MT^+ cells. Once mating pairs are formed, no dissimilarities between the two mating types are observed anymore. After mate pairing, both cells produce two gametes that are released from the parental valves. These gametes fuse almost immediately with the gametes release by the other cell of the mating pair, which results in the formation of two zygotes. No morphological or behavioral differences between the gametes coming from opposite mating types are observed, meaning that *S. robusta* is isogamous.

To learn more about these differences between the two mating types and diatom life cycle regulation in general, funding for an exploratory transcriptome study was obtained from the JGI project “A deep transcriptomic and genomic investigation of diatom life cycle regulation”. RNA sequencing was conducted on synchronized cultures of *S. robusta* in the different stages of its life cycle (Chapter 3). About 8,000 transcripts (11% of the total number of transcripts) are differentially expressed when comparing all MT^+ with all MT^- samples. This is possibly due to the

large variety of conditions represented by the samples (different cell cycle stages, cells above and below the SST) and to intrinsic differences between the two strains. Without any biological replicates, it is difficult to distinguish biological relevant expression differences from false positives. Also when comparing the vegetative with the sexual samples, a high number of differentially expressed transcripts was found (over 9,000 transcripts, 13% of the transcripts). However, when doing GO enrichments, some trends can be seen. The vegetative samples are enriched in genes with functions in metabolism/biosynthesis , while the sexual stages are enriched in genes with functions in meiosis, motility and cell signaling. This probably reflects the switch from growing cells with an active metabolism to sexualized cells that undergo gametogenesis.

Apart from *S. robusta*, also some *Pseudo-Nitzschia multistriata* RNA samples were sequenced as part of this JGI project (collaboration with M. Montresor). Comparing both species proved difficult because of several reasons. Firstly, the *P. multistriata* samples were not synchronized and no sexual stages were sampled. Secondly, it was nearly impossible to identify true orthologs between the two species. This is probably due to the redundancy that is present in de novo assembled transcriptomes. As multiple copies of the same gene might be present, the results from the reciprocal blast approach used to identify orthologs are difficult to interpret. This issue can probably be resolved when the genomes of both species become available. Thirdly, the lack of proper replicates makes it hard to distinguish false positives from real expression differences.

Despite some shortcomings, this transcriptome is a rich source of sequence information. As long as the genome of *S. robusta* is not available, this dataset can be used to search for specific gene families or genes related to specific functions. Some examples, like the cyclin gene family or genes involved in silicification, are described in Chapter 3. An advantage of this dataset compared to the genome, is the availability of expression information. Because of the lack of biological replicates, expression patterns should be confirmed using RT-qPCR. However, this dataset is a good starting point for further research. This can clearly been seen in the

identification of meiosis-related genes (Chapter 3). These genes were found based on sequence homology with *Arabidopsis thaliana* or *Saccharomyces cerevisiae*. For most of these genes, the expression pattern (higher expression in mating cells) confirmed their assumed function in meiosis in *S. robusta*.

A multistep pheromone system regulates the initiation of sexual reproduction

Next to the RNA-seq dataset described above, a second RNA-seq experiment was conducted to learn more about pheromone signaling in *S. robusta* (Chapter 4). Here, MT^- cultures were treated with an active biochemical fraction containing conditioning factor plus (CF-P). This factor is produced by MT^+ cells below the SST and was shown to induce cell cycle arrest and diproline production in MT^- cultures below the SST.

Physiological and molecular data are indicative for a cell cycle arrest in G_1/S phase. This is confirmed by the observation that the chloroplasts translocated without prior chloroplast division in mating pairs (Chepurnov et al., 2002). Since chloroplast division normally takes place during S/G_2 phase of the mitotic cell cycle, this supports an arrest in G_1 phase (Gillard et al., 2008). Furthermore, this is in correspondence with the fact that also in other eukaryotes the commitment to meiosis is made before the onset of S phase (Pawlowski et al., 2007). By arresting their cell cycle in this phase, diatoms can prolong the time frame in which they are able to switch to meiosis and thus increase the chance of finding a mating partner in time.

The actual attraction of the two mates is induced by diproline that makes the MT^+ cells more motile and directs them towards the diproline-secreting MT^- cells (Gillard et al., 2013). Only MT^- cells below the SST produce diproline, and this only after they perceived the conditioning factor CF-P (Chapter 4). Likewise, MT^+ cells are only responsive to diproline when they first sensed the conditioning factor CF-M that is produced by MT^- cells below the SST. When the two cells of opposite mating type come into direct contact, they form mating pairs and probably use yet another

signaling system to induce gametogenesis. This is most likely by the use of non-diffusible signaling molecules (paracrine signaling) or via membrane-embedded lipids or proteins (juxtacrine signaling), because gametogenesis is only been observed in *S. robusta* when cells are in direct contact.

This multistep pheromone system seems to be concentration dependent. More sexual events are observed when less culture medium is present, even when the distance between the cells is the same (same number of cells on same surface area) (Gillard et al., 2013). This means that when the pheromones are present in the same amount but in a smaller volume, so in higher concentration, sexual proficiency increases. This implies that they only invest in sexual reproduction when the opposite mating type is present in sufficiently high densities. In other words, they only initiate sexual reproduction when the chances to find a mating partner are high enough. This mechanism is reminiscent of quorum sensing, a mechanism that is normally associated with bacteria, but that has also been described in fungi (Sprague and Winans, 2006). When the population density increases, the concentration of quorum sensing signalling molecules in the extracellular environment also increases. As soon as this concentration reaches a theshold level, a signalling cascade is activated, leading to a cellular response. Because conditioning factor signalling in *S. robusta* appears to act similarly, this could be a first example of quorum sensing in diatoms.

The RNA-seq that was conducted to investigate the effect of CF-P on genome-wide expression gave us a first indication on the downstream signaling pathway used in pheromone signaling. A bifunctional guanylyl cyclase/phosphodiesterase was rapidly induced in MT^- cultures treated with CF-P, implicating the secondary messenger cGMP in sex-signaling in *S. robusta* (Chapter 4). cGMP is found throughout the tree of life, although no guanylyl cyclases are found in fungi or plants (Schaap, 2005). cGMP is known to activate cGMP-dependent protein kinases or cGMP-gated ion channels (Francis and Corbin, 1999, Matulef and Zagotta, 2003). Guanylyl cyclases are found in soluble and in membrane-bound form. Soluble guanylyl cyclases are often regulated by nitric oxide, while transmembrane guanylyl

cyclases act as receptors for peptides and small molecules (Schaap, 2005). To our knowledge, the combination of a cyclase and a phosphodiesterase domain in one protein is not found in eukaryotes other than diatoms. However, this conformation is found in bacteria where an EAL phosphodiesterase domain often occurs in tandem with a diguanylate cyclase domain in a single protein, for example in MorA, a bifunctional c-di-GMP regulator in *Pseudomonas aeruginosa* (Phippen et al., 2014). The combination of two related functions in one protein frequently occurs in nature. Also in diatoms several putative fusion proteins have been identified, for example in carbohydrate metabolism or in the sterol pathway (Kroth et al., 2008, Fabris et al., 2014). These fusion proteins probably evolved to improve co-regulation of two closely-connected functions, indicating that cGMP levels are likely to be strictly regulated in diatoms.

The establishment of *Seminavis robusta* as a genuine model species

S. robusta has been developed as a model species to investigate the diatom life cycle (Chepurnov et al., 2008, Gillard et al., 2008, Gillard et al., 2013, Vanstechelman et al., 2013). It was chosen because it displays the typical diatom life cycle with a size reduction-restitution system where initial size is restored by sexual auxosporulation (Chepurnov et al., 2002). Furthermore, its heterothallic mating system allows reliable control of sexual reproduction, since two clones of opposite mating type have to be mixed to induce sexualization. Moreover, gametogenesis occurs at high frequency and can be synchronized by extending the dark period before mixing compatible clones. Additional benefits are its large cell size, which simplifies microscopic observations, easy cultivation techniques and the possibility to cryopreserve strains (Chepurnov et al., 2008).

To facilitate a more detailed investigation of the regulation of sexual reproduction, molecular tools should become available for *S. robusta*. One important step is the development of a genetic transformation protocol. This requirement is now fulfilled by setting up a protocol for the stable transformation of *S. robusta* using microparticle bombardment (Chapter 5). Although transgene lines could be created,

there is room for improvement. Firstly, no fluorescence could be seen in transformants containing the *H2B-YFP* or *H2B-CFP* fusion genes. This problem should be addressed by, for example, make transformants with other tagged genes or with N-tagged genes. Also the codon usage of *S. robusta* should be checked to make sure the tags can be properly translated. Secondly, transformation efficiency is still quite variable. This could possibly be improved by optimizing cell density and antibiotic concentration or by using a different transformation technique. At the moment, an electroporation protocol, that was developed for *P. tricornutum* (Zhang and Hu, 2014), is tested for *S. robusta* (M. Huysman, personal communication).

Genetic transformation enables the overexpression of genes-of-interest to study their function. Also fluorescent labelling of protein will become possible, which will allow localization of proteins in the cell. Up to now, no attempts to decrease the activity of genes-of-interest are undertaken in *S. robusta*. This could be done by silencing genes through RNA interference or by making knock-outs using for example TALEN or CRISPR-Cas9 technology. RNA interference is routinely applied in *Phaeodactylum tricornutum* (De Riso et al., 2009, Lavaud et al., 2012, Huysman et al., 2013) and also genomic engineering techniques are becoming available for *P. tricornutum* (Daboussi et al., 2014). In the near future, these techniques will be adapted for the use in *S. robusta*.

Another requisite for model organisms nowadays is the availability of sequencing data. A large amount of transcriptome data has become available for this species (Chapters 3 and 4) and the genome is being assembled and annotated at the moment in our department. A proper genome assembly and gene prediction will improve the analyses that now have been conducted on the *de novo* transcriptomes of *S. robusta*, because RNA-seq reads could then be mapped directly to the gene models. This will eliminate the redundant and partial transcripts. The availability of a genome will also allow for forward genetic techniques, like random mutagenesis screens and GWAS, because it will be possible to map mutations to the reference genome.

Future perspectives

Knowing the nature of the MT locus offers the possibility to change the mating type of a clone by manipulating the activity of the MT-determining gene. This would then enable self-fertilization, which would make it easy to make homozygous lines. A way to do this would be to interfere with methylation, for example by administering hypomethylating drugs, like 5-azacytidine (Janousek et al., 1996). Another more direct method would be to change DNMT5a functioning by overexpression of the MT^+-specific allele in MT^- cells or by lowering its activity using hairpin constructs.

The next step in the characterization of MT determination would be to identify the targets of DNMT5a. This can be done by looking for genes that are differentially methylated between the two mating types. Possible techniques to achieve this are whole genome bisulfite sequencing or methylation-sensitive restriction enzymes combined with for example AFLP (Vuylsteke et al., 2000, Zilberman and Henikoff, 2007). The identified targets could be involved in pheromone signaling, for example in the biosynthesis of the conditioning factors or in their perception, or in other, yet unknown, mating type-regulated pathways.

Another approach to learn more about the enzymes involved in the production of the conditioning factors or about the receptors used to detect them is the elucidation of their structure. Following the identification of the attraction pheromone diproline (Gillard et al., 2013), the research group of Georg Pohnert tries to fractionate both conditioning factors with the ultimate goal to determine their exact structure (Personal communication). Once the nature of these factors is known, one could speculate about the enzymes that could synthesize these molecules. The genes coding for these enzymes can be searched for in the *S. robusta* genome or transcriptome and their expression at different life cycle stage could be checked in the RNA-seq data described in chapter 3. It is expected that these genes are highly expressed in cultures of one of the two mating types and only below the SST. Additionally, expression could be related to the cell cycle, because cells are only

able to make the switch to meiosis when they are in G_1 phase. It is thus logical that they only try to find a mating partner when they are themselves in the right cell cycle phase. The same holds true for the receptors of the conditioning factors.

To know whether the findings described in this dissertation can be extrapolated to other diatoms, more research on other species will be necessary. *S. robusta* is the first diatom for which the MT locus is identified. In contrast to most other sequenced diatoms, *S. robusta* has two genes belonging to the *DNMT5* gene family. This led to the hypothesis that the original *DNMT5* was duplicated and that one of the copies could specialize to become the MT determinant. If we look to the other diatoms for which a genome is available, only *Fragilariopsis cylindrus* has multiple (three) copies of *DNMT5*. Although no sexual reproduction is observed in *F. cylindrus*, auxospores of *F. kerguelensis* are described (Assmy et al., 2006), making it plausible that also *F. cylindrus* could be capable of sexual reproduction. However, if mating cannot be reliably induced in laboratory conditions, it is impossible to determine the MT locus in this species. The only species for which a genome is available and sex can be controlled in the lab is *Pseudo-Nitzschia multiseries* (Davidovich and Bates, 1998). On the other hand, only one *DNMT5* gene is found in the genome of *P. multiseries* and this gene seems closer related to *S. robusta DNMT5b* then to *DNMT5a* (Chapter 2, figure 4). This does not completely rule out DNMT5 as a MT determinant in *P. multiseries*, but more research will be needed to elucidate its function.

Furthermore, not much is known about pheromones in diatoms. The araphid pennate diatom *Pseudostaurosira trainorii* uses a pheromone system that is comparable with that of *S. robusta* (Sato et al., 2011). The major difference is that in *P. trainorii* attraction between the two sexes happens at the level of the gametes instead of the gametangia as is the case in *S. robusta*. Additionally, the first two pheromones (ph-1 and ph-2) in the cascade are not secreted simultaneously like in *S. robusta* (CF-P and CF-M) and they induce gametogenesis in contrast to the conditioning factors of *S. robusta* that induce cell cycle arrest, but no initiation of meiosis (Chapter 4). Although the presence of the pheromones in *P. trainorii* could

be detected with bio-assays, nothing is known yet about the chemical structure of these pheromones.

Recently, the first indications were found that also in *P. multistriata* chemical cues are used to initiate sexual reproduction (Scalco et al., 2014). Sexual proficiency is dependent on cell density and higher in physically undisturbed conditions. Additionally, preliminary experiments showed that gametogenesis could occur when a strain was conditioned by the filtered medium of the opposite mating type. In *P. multistriata*, no clear attraction from one mating type to the other was observed. Instead, both mating types seem to be motile. Furthermore, a reduction of the growth rate was seen in sexually active cultures, pointing to a potential cell cycle arrest similar as in *S. robusta*. All these results taken together, it seems that also *P. multistriata* uses pheromones to initiate mating, although the system could be quite different from that of *S. robusta*, which could reflect the large difference in life style between benthic and planktonic species. Next to *P. multistriata*, also *Cylindrotheca closterium* could be a suitable species to study possible pheromone signaling. Sexual reproduction in *C. closterium* can be synchronized by elongating the dark period and high numbers of gametes can be obtained (Vanormelingen et al., 2013), making this a suitable species for the development of standardized bio-assays like was done for *P. trainoii* and *S. robusta*.

Also in centric diatoms, there are indications that pheromone signaling could be involved in synchronizing gametogenesis between potential mating partners. When two *Ditylum brightwellii* clones below the SST were mixed, gametogenesis advanced and higher numbers of auxospores were produced compared to in monoclonal cultures (Koester et al., 2007). Additionally, it can be suspected that some kind of signaling is involved in the interaction between the gametes. In *Actinocyclus* sp., for example, sperm cells are attracted in large numbers to the immotile egg cells (Idei et al., 2012).

In conclusion, this thesis contributed to our knowledge about life cycle regulation in *S. robusta*. This knowledge can be deepened by identifying the target

genes of the MT determinant DNMT5a and the pheromone biosynthesis enzymes and receptors. The ultimate goal is to extend these findings to other diatoms. Therefore, it should be determined if DNMT5 is also the primary MT determinant in other species and if the pheromone system used by *S. robusta* is similar to that of other diatoms.

General conclusion

Diatoms use a size reduction-restitution mechanism to determine when to switch from vegetative division to sexual reproduction. Here, the life cycle of *Seminavis robusta* was studied with a focus on mating type determination and the molecular signaling that precedes mate pair formation. In parallel, the position of *S. robusta* as a model system for pennate diatoms was strengthened by generating a large amount of transcriptome data and by developing a genetic transformation protocol.

We showed that the MT locus encodes the DNA methyltransferase *DNMT5a*, which suggests that the physiological differences between mating types are determined by the differential methylation of certain target genes. Further research will be needed to confirm the link between MT determination and DNA methylation, as well as to identify the target genes.

Amongst the DNMT5a targets could be the genes coding for the enzymes synthesizing the conditioning factors or for their receptors, because both mating types produce a different conditioning factor and respond only to the one produced by the opposite mating type. Physiological experiments and expression data showed that the conditioning factors cause a cell cycle arrest in the other mating type. An RNA-seq experiment confirmed that CF-P was responsible for the induction of diproline production by MT^- and more specifically that this was done by upregulating the glutamate-to-proline conversion. This experiment also indicated a potential role for the secondary messenger cGMP in pheromone signaling. Further experiments should be conducted to determine the exact role of cGMP and to find the enzyme responsible for the last step of the diproline production.

In conclusion, this thesis contributed to our knowledge about life cycle regulation in *S. robusta*. This knowledge can be deepened by identifying the target genes of the MT determinant DNMT5a and the pheromone biosynthesis enzymes and receptors. The ultimate goal is to extend these findings to other diatoms.

Therefore, it should be investigated if DNMT5 is also the primary MT determinant in other species and if the pheromone system used by *S. robusta* is similar to that of other diatoms.

Literature cited

Assmy, P., Henjes, J., Smetacek, V. and Montresor, M. (2006) Auxospore formation by the silica-sinking, oceanic diatom Fragilariopsis kerguelensis (Bacillariophyceae). *Journal of Phycology* **42**, 1002-1006.

Chen, S.L., Zhang, G.J., Shao, C.W., Huang, Q.F., Liu, G., Zhang, P., Song, W.T., An, N., Chalopin, D., Volff, J.N., Hong, Y.H., Li, Q.Y., Sha, Z.X., Zhou, H.L., Xie, M.S., Yu, Q.L., Liu, Y., Xiang, H., Wang, N., Wu, K., Yang, C.G., Zhou, Q., Liao, X.L., Yang, L.F., Hu, Q.M., Zhang, J.L., Meng, L., Jin, L.J., Tian, Y.S., Lian, J.M., Yang, J.F., Miao, G.D., Liu, S.S., Liang, Z., Yan, F., Li, Y.Z., Sun, B., Zhang, H., Zhang, J., Zhu, Y., Du, M., Zhao, Y.W., Schartl, M., Tang, Q.S. and Wang, J. (2014) Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nature Genetics* **46**, 253-260.

Chepurnov, V.A., Mann, D.G., Vyverman, W., Sabbe, K. and Danielidis, D.B. (2002) Sexual reproduction, mating system, and protoplast dynamics of *Seminavis* (Bacillariophyceae). *Journal of Phycology* **38**, 1004-1019.

Chepurnov, V.A., Mann, D.G., von Dassow, P., Vanormelingen, P., Gillard, J., Inzé, D., Sabbe, K. and Vyverman, W. (2008) In search of new tractable diatoms for experimental biology. *BioEssays* **30**, 692-702.

Daboussi, F., Leduc, S., Marechal, A., Dubois, G., Guyot, V., Perez-Michaut, C., Amato, A., Falciatore, A., Juillerat, A., Beurdeley, M., Voytas, D.F., Cavarec, L. and Duchateau, P. (2014) Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology. *Nature Communications* **5**.

Davidovich, N.A. and Bates, S.S. (1998) Sexual reproduction in the pennate diatoms Pseudo-nitzschia multiseries and P-pseudodelicatissima (Bacillariophyceae). *Journal of Phycology* **34**, 126-137.

De Riso, V., Raniello, R., Maumus, F., Rogato, A., Bowler, C. and Falciatore, A. (2009) Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic Acids Research* **37**.

Fabris, M., Matthijs, M., Carbonelle, S., Moses, T., Pollier, J., Dasseville, R., Baart, G.J.E., Vyverman, W. and Goossens, A. (2014) Tracking the sterol biosynthesis pathway of the diatom Phaeodactylum tricornutum. *New Phytologist* **204**, 521-535.

Francis, S.H. and Corbin, J.D. (1999) Cyclic nucleotide-dependent protein kinases: Intracellular receptors for cAMP and cGMP action. *Critical Reviews in Clinical Laboratory Sciences* **36**, 275-328.

Gillard, J., Devos, V., Huysman, M.J.J., De Veylder, L., D'Hondt, S., Martens, C., Vanormelingen, P., Vannerum, K., Sabbe, K., Chepurnov, V.A., Inzé, D., Vuylsteke, M. and Vyverman, W. (2008) Physiological and transcriptomic evidence for a close coupling between chloroplast ontogeny and cell cycle progression in the pennate diatom *Seminavis robusta*. *Plant Physiology* **148**, 1394-1411.

Gillard, J., Frenkel, J., Devos, V., Sabbe, K., Paul, C., Rempt, M., Inze, D., Pohnert, G., Vuylsteke, M. and Vyverman, W. (2013) Metabolomics Enables the

Structure Elucidation of a Diatom Sex Pheromone. *Angewandte Chemie-International Edition* **52**, 854-857.

Huff, J.T. and Zilberman, D. (2014) Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes. *Cell* **156**, 1286-1297.

Huysman, M.J.J., Fortunato, A.E., Matthijs, M., Costa, B.S., Vanderhaeghen, R., Van den Daele, H., Sachse, M., Inze, D., Bowler, C., Kroth, P.G., Wilhelm, C., Falciatore, A., Vyverman, W. and De Veylder, L. (2013) AUREOCHROME1a-Mediated Induction of the Diatom-Specific Cyclin dsCYC2 Controls the Onset of Cell Division in Diatoms (Phaeodactylum tricornutum). *Plant Cell* **25**, 215-228.

Idei, M., Osada, K., Sato, S., Toyoda, K., Nagumo, T. and Mann, D.G. (2012) Gametogenesis and Auxospore Development in Actinocyclus (Bacillariophyta). *PLoS ONE* **7**.

Janousek, B., Siroky, J. and Vyskot, B. (1996) Epigenetic control of sexual phenotype in a dioecious plant, Melandrium album. *Molecular & General Genetics* **250**, 483-490.

Jones, P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13**, 484-492.

Koester, J.A., Brawley, S.H., Karp-Boss, L. and Mann, D.G. (2007) Sexual reproduction in the marine centric diatom Ditylum brightwellii (Bacillariophyta). *European Journal of Phycology* **42**, 351-366.

Kroth, P.G., Chiovitti, A., Gruber, A., Martin-Jezequel, V., Mock, T., Parker, M.S., Stanley, M.S., Kaplan, A., Caron, L., Weber, T., Maheswari, U., Armbrust, E.V. and Bowler, C. (2008) A Model for Carbohydrate Metabolism in the Diatom Phaeodactylum tricornutum Deduced from Comparative Whole Genome Analysis. *PLoS ONE* **3**.

Lavaud, J., Materna, A.C., Sturm, S., Vugrinec, S. and Kroth, P.G. (2012) Silencing of the Violaxanthin De-Epoxidase Gene in the Diatom Phaeodactylum tricornutum Reduces Diatoxanthin Synthesis and Non-Photochemical Quenching. *PLoS ONE* **7**.

Law, J.A. and Jacobsen, S.E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics* **11**, 204-220.

Matulef, K. and Zagotta, W.N. (2003) Cyclic nucleotide-gated ion channels. *Annual Review of Cell and Developmental Biology* **19**, 23-44.

Maunakea, A.K., Chepelev, I., Cui, K.R. and Zhao, K.J. (2013) Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Research* **23**, 1256-1269.

Parkinson, S.E., Gross, S.M. and Hollick, J.B. (2007) Maize sex determination and abaxial leaf fates are canalized by a factor that maintains repressed epigenetic states. *Developmental Biology* **308**, 462-473.

Pawlowski, W.P., Sheehan, M.J. and Ronceret, A. (2007) In the beginning: the initiation of meiosis. *BioEssays* **29**, 511-514.

Phadke, S.S. and Zufall, R.A. (2009) Rapid diversification of mating systems in ciliates. *Biological Journal of the Linnean Society* **98**, 187-197.

Phippen, C.W., Mikolajek, H., Schlaefli, H.G., Keevil, C.W., Webb, J.S. and Tews, I. (2014) Formation and dimerization of the phosphodiesterase active site of the Pseudomonas aeruginosa MorA, a bi-functional c-di-GMP regulator. *FEBS Letters* **588**, 4631-4636.

Sato, S., Beakes, G., Idei, M., Nagumo, T. and Mann, D.G. (2011) Novel Sex Cells and Evidence for Sex Pheromones in Diatoms. *PLoS ONE* **6**.

Scalco, E., Stec, K., Iudicone, D., Ferrante, M.I. and Montresor, M. (2014) The Dynamics of Sexual Phase in the Marine Diatom Pseudo-Nitzschia Multistriata (Bacillariophyceae). *Journal of Phycology* **50**, 817-828.

Schaap, P. (2005) Guanylyl cyclases across the tree of life. *Frontiers in Bioscience-Landmark* **10**, 1485-U1485.

Shao, C.W., Li, Q.Y., Chen, S.L., Zhang, P., Lian, J.M., Hu, Q.M., Sun, B., Jin, L.J., Liu, S.S., Wang, Z.J., Zhao, H.M., Jin, Z.H., Liang, Z., Li, Y.Z., Zheng, Q.M., Zhang, Y., Wang, J. and Zhang, G.J. (2014) Epigenetic modification and inheritance in sexual reversal of fish. *Genome Research* **24**, 604-615.

Sprague, G.F. and Winans, S.C. (2006) Eukaryotes learn how to count: quorum sensing by yeast. *Genes & Development* **20**, 1045-1049.

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

Vanstechelman, I., Sabbe, K., Vyverman, W., Vanormelingen, P. and Vuylsteke, M. (2013) Linkage Mapping Identifies the Sex Determining Region as a Single Locus in the Pennate Diatom Seminavis robusta. *PLoS ONE* **8**.

Vuylsteke, M., Mank, R., Brugmans, B., Stam, P. and Kuiper, M. (2000) Further characterization of AFLP (R) data as a tool in genetic diversity assessments among maize (Zea mays L.) inbred lines. *Molecular Breeding* **6**, 265-276.

Whitfield, J. (2004) Everything you always wanted to know about sexes. *PLoS Biology* **2**, 718-721.

Zhang, C.Y. and Hu, H.H. (2014) High-efficiency nuclear transformation of the diatom Phaeodactylum tricornutum by electroporation. *Marine Genomics* **16**, 63-66.

Zilberman, D. and Henikoff, S. (2007) Genome-wide analysis of DNA methylation patterns. *Development* **134**, 3959-3965.

7
Summary

Diatoms are, with about 200,000 species, the most species-rich group of microalgae. These unicellular photosynthetic organisms are found in all aquatic environments where enough light and nutrients are available. They are of great ecological importance, as they are responsible for about 20% of global primary production. Moreover, they are key players in the biogeochemical cycling of carbon, nitrate and silicon. They probably thank their evolutionary success to their unique life cycle. Due to their rigid, silicified cell wall and their cell division mechanism, the cell size of one of the daughter cells will be smaller than the size of the mother cell after mitotic division. Without a mechanism to restore the original cell size, this would ultimately lead to clonal cell death. Restitution of the initial cell size is achieved through sexual reproduction, whereby a specialized zygote is formed, called the auxospore. Since the auxospore is not restrained by a rigid cell wall, it can grow to initial cell size. In diatoms, sexual reproduction is only possible when cells are smaller than a species-specific sexual size threshold. This cell size reduction-restitution mechanism is used as an internal clock to regulate the switch from vegetative division to sexual reproduction.

To study this mitosis-to-meiosis switch, *Seminavis robusta* was developed as a new model organism. This was necessary because the two widely used model diatom species, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*, do not exhibit this peculiar size reduction-restitution mechanism. *S. robusta* was chosen as a model because of its typical pennate life cycle and its ease in cultivation. Like most pennate diatoms, *S. robusta* is heterothallic, meaning that sex only occurs between clones of a different mating type. This makes the induction of sex easy to control. Moreover, high frequencies of gametogenesis can be obtained when synchronizing the cell cycle of a culture by prolonging the dark period. The genetic transformation protocol and

the large amount of transcriptomic data, described in this thesis, contribute to the establishment of *S. robusta* as a model system in molecular research.

In chapter 2 of this thesis, the gene responsible for mating type (MT) determination is described. Earlier, an AFLP-based linkage map of *S. robusta* was used to identify the mating type locus as a single locus in the genome. This study also showed that MT^+ is the heterogametic mating type. Here, fine-mapping led to the identification of the DNA methyltransferase *DNMT5a* as being present on the MT locus of *S. robusta*. This indicates that MT determination in *S. robusta* is probably regulated by differential methylation of genes between MT^+ and MT^-.

Chapter 3 describes a transcriptome analysis covering the most important stages of the diatom life cycle. This transcriptome was used to identify the orthologs of genes that are known to be involved in meiosis in other eukaryotes. For most of them, the expression pattern confirmed their probable role in meiosis in diatoms. Also the cyclin gene family was annotated, using this transcriptome. A large number of cyclins was found, of which a considerable amount belonged to the diatom-specific cyclins, which are thought to be involved in transmitting environmental signals to the cell cycle. Also some A/B-type cyclins that showed a higher expression during meiosis could be identified. This observation is in correspondence with their role in the regulation of meiotic division in other eukaryotes. Furthermore, some diatom-specific genes involved in silicification of the cell wall were identified in this dataset.

In chapter 4, a second RNA-seq experiment was used to study the effect of conditioning factor plus (CF-P) on the genome-wide expression in MT^- cells. Before gametogenesis can occur, mate pairs between two cells of opposite mating type have to be formed. In *S. robusta*, this mate pair formation is controlled by a multistep pheromone system. First, cells below the sexual size threshold produce conditioning factors. The two mating types produce a different factor and both factors induce cell cycle arrest in G_1 phase in the other mating type. The conditioning factor produced by MT^+ (CF-P) also triggers MT^- to produce the attraction pheromone diproline.

Diproline attracts conditioned MT^+ cells in order to facilitate mate pair formation. The RNA-seq experiment showed that diproline production is increased by upregulating glutamate-to-proline conversion. Additionally, these data gave a first indication that the secondary messenger cGMP could be involved in pheromone signaling.

In conclusion, this thesis expanded our knowledge about the diatom life cycle by identifying *DNMT5a* as the primary mating type determinant in *S. robusta*, by providing an elaborate transcriptomic dataset covering the most important stages of the life cycle and by studying the pheromone system of *S. robusta* in more detail. Furthermore, all these sequence data together with the genetic transformation protocol contributed to the position of *S. robusta* as a model species to study the molecular biology of pennate diatoms.

7
Samenvatting

Diatomeeën zijn, met ongeveer 200 000 soorten, de meest soortrijke groep binnen de microalgen. Deze eencellige fotosynthetiserende organismen komen voor in alle aquatische omgevingen waar voldoende licht en voedingsstoffen beschikbaar zijn. Ze zijn van groot ecologisch belang omwille van hun substantiële bijdrage tot de primaire productie (20% van de globale productie). Verder spelen ze een sleutelrol in de biogeochemische cycli van koolstof, stikstof en silicium. Diatomeeën danken hun evolutionaire succes vermoedelijk aan hun unieke levenscyclus. Door hun onbuigzame, gesilicificeerde celwand en hun manier van celdeling zal een van de twee dochtercellen na mitotische deling iets kleiner zijn dan de moedercel. Zonder een manier om de celgrootte te herstellen, zou dit tot klonale celdood leiden. Ze zijn echter in staat de initiële celgrootte te herstellen via seksuele voortplanting. Daarbij wordt een gespecialiseerde zygote gevormd, die men de auxospore noemt. Deze auxosporen kunnen uitgroeien tot initiële celgrootte, omdat hun celwand minder star is. In diatomeeën is seksuele voortplanting enkele mogelijk wanneer de celgrootte onder een soortspecifieke seksuele groottedrempel ligt. Dit unieke celgroottereductie-herstel-mechanisme wordt gebruikt als een interne klok om de overgang van vegetatieve celdeling naar seksuele voortplanting te reguleren.

Seminavis robusta werd geïntroduceerd als een nieuw modelorganisme om deze overgang van mitose naar meiose te bestuderen. Dit was nodig omdat de twee meest-gebruikte modeldiatomeeën, *Phaeodactylum tricornutum* en *Thalassiosira pseudonana*, geen celgroottereductie en -herstel vertonen en bijgevolg ook geen seksuele voortplanting. *S. robusta* werd gekozen als modelorganisme omwille van zijn typische pennate levenscyclus en omdat ze eenvoudig te cultiveren zijn. Zoals de meeste pennate diatomeeën is *S. robusta* heterothallisch. Dit wil zeggen dat seks enkel voorkomt tussen cellen van verschillend *mating type*, wat ervoor zorgt dat

seksuele voortplanting eenvoudig te controleren is in labcondities. Bovendien worden er grote aantallen gameten gevormd wanneer de celcyclus van een cultuur gesynchronizeerd wordt door de donkerperiode te verlengen. Het genetische transformatieprotocol en de grote hoeveelheid transcriptoomdata, die beschreven worden in deze thesis, dragen bij tot de ontwikkeling van *S. robusta* tot een volwaardig modelorganisme in moleculair onderzoek.

In hoofdstuk 2 van deze thesis wordt het gen dat verantwoordelijk is voor *mating type* (MT) determinatie in *S. robusta* beschreven. Eerder werd een AFLP-gebaseerde koppelingskaart gebruikt om de MT locus te identificeren. Deze studie toonde aan dat het *mating type* van *S. robusta* bepaald wordt door één locus in het genoom en dat MT^+ het heterogamete *mating type* is. In deze thesis tonen we aan dat het DNA methyltransferase gen *DNMT5a* aanwezig is in de MT locus. Dit impliceert dat MT determinatie in *S. robusta* gereguleerd wordt door differentiële methylatie tussen de twee *mating types*.

Hoofdstuk 3 beschrijft een transcriptoomanalyse die de belangrijkste fases van de levenscyclus van *S. robusta* bevat. Dit transcriptoom werd gebruikt om orthologen te identificeren van genen die betrokken zijn bij meiose in andere eukaryoten. Voor de meeste genen werd hun potentiële rol in meiose in diatomeeën bevestigd door hun expressiepatroon. Ook de genfamilie van de cyclines werd geannoteerd op basis van deze dataset. Een groot aantal cyclines werd geïdentificeerd, waarvan een aanzienlijk deel tot de diatomee-specifieke cyclines behoort. Deze dsCYCs zijn waarschijnlijk betrokken bij het doorgeven van omgevingssignalen naar de celcyclus. Verder werden er verschillende A/B-type cyclines gevonden die een hogere expressie vertonen tijdens meiose. Dit is in overeenstemming met hun rol in meiose in andere eukaryoten. Daarnaast werden ook diatomee-specifieke genen geïdentificeerd die betrokken zijn bij de silicificatie van de celwand.

In hoofdstuk 4 wordt een tweede RNA-seq experiment beschreven. Dit werd opgezet om het effect van conditioneringsfactor plus (CF-P) op de genoomwijde

expressie van MT^- te bestuderen. Om te kunnen overgaan tot gametogenese, moeten twee cellen van tegenovergestelde *mating types* eerst een paar vormen. Deze paarvorming wordt in *S. robusta* gecontroleerd door een feromoonsysteem in meerdere stappen. In een eerste stap beginnen cellen een conditioneringsfactor uit te scheiden wanneer ze de seksuele groottedrempel overschreiden. De twee *mating types* produceren een verschillende conditioneringsfactor en beide factoren veroorzaken een celcyclus arrest in de G_1-fase in het andere *mating type*. Bovendien induceert de conditioneringsfactor, die aangemaakt wordt door MT^+, de productie van het aantrekkingsferomoon diproline door MT^-. Diproline trekt geconditioneerde MT^+ cellen aan zodat paarvorming kan optreden. Het RNA-seq experiment toonde aan dat de diproline productie verhoogd wordt door het opreguleren van de glutamaat-naar-proline omzetting. Bovendien geven deze data een eerste indicatie dat het secundaire boodschappermolecule cGMP waarschijnlijk betrokken is bij feromoonsignalering.

Deze thesis draagt bij tot onze kennis over de levenscyclus van diatomeeën door de identificatie van het eerste MT-bepalende gen, *DNMT5a*, door een uitgebreide transcriptoomdataset die de belangrijkste stadia van de levenscyclus beslaat en door een gedetailleerde studie van het feromoonsysteem in *S. robusta*. Bovendien bevorderen het genetische transformatieprotocol en de uitgebreide transcriptoomdatasets de ontwikkeling van *S. robusta* als een volwaardig modelorganisme om de moleculaire biologie van pennate diatomeeën te bestuderen.