UNIVERSITEIT GENT

FACULTY OF SCIENCES

Department of Biology

Department of Plant Biotechnology and Bioinformatics

Department of Plant Systems Biology

Identification, characterization and evolution of the mating type locus in diatoms

Ives Vanstechelman

Promotors: Prof. Dr. Koen Sabbe and Prof. Dr. ir. Marnik Vuylsteke

Thesis submitted in fulfillment of the requirements for the degree of

Doctor (PhD) in Sciences (Biology)

Academic year 2012-2013

VIB

Plant Systems Biology

Protistology

Aquatic Ecology

Exam commission

Promotors:

Prof. Dr. Koen Sabbe

Prof. Dr. ir. Marnik Vuylsteke

Members of the reading commission

Dr. Mariella Ferrante

Prof. Dr. Olivier De Clerck

Prof. Dr. Wout Boerjan

Other members of the exam commission

Prof. Dr. Wim Vyverman

Prof. Dr. Mieke Verbeken

Dr. Marie Huysman

Acknowledgements

The work presented in this thesis could not be successful without the help of several people. Therefore, I want to give a few words of thanks and appreciation for the people who supported me during this 4 years during PhD.

First of all, I would like to thank my two promotors Koen Sabbe and Marnik Vuylsteke. Koen is my promotor with the biological background. He supported me a lot in giving more knowledge about the biology of diatoms and evolution. When I first met him, he could directly convince me to start in this position. The unique features of the diatom life cycle attracted me and I really saw it as a big challenge to get insight into the genetic basis of this. The person who helped me the most with succeeding in this PhD is undoubtedly Marnik Vuylsteke. He is my promotor with the genetic background and as a consequence, I mainly worked in his lab, the Quantitative Genomics lab. Making progression in a PhD project is the same as being constanly confronted with a lot of problems and I am sure, Marnik solved most of them! I could always disturb him in his office, even in the moments that he had a lot of other work to do. Marnik was always early in the lab and sometimes, the first thing I did in the morning was going to his office with a problem. His opinions were necessary to continue another day. I also want to thank him because he gave me the opportunity to work some months on the BlueGene project. Besides this professional part, I liked him very much personally. His funny jokes were welcome to continue another day.

Stephanie De Vos was my neighbor at our desk for four years. As we had a the same project for a different species, we were frequently confronted with the same problems and so, we did learn a lot from each other. We first had to share the AFLP technology and then the next generation sequencing. Working together with her was really a nice experience to me.

Our monthly meetings in ‘de sterre’ were very useful. We talked about our problems and progression with Wim Vyverman, Koen Sabbe, Marnik Vuylsteke, Marie Huysman, Valerie Devos and sara Moeys. These meeting were a source of new ideas and were very useful in my progression.

I would like to thank Pieter Vanormelingen and Olga Chepurnova for learning me the *Seminavis* culturing techniques. Ilse vercauteren learned the AFLP technology to me and as she was our lab manager, we could ask her all lab related things. I also would like to thank

Tine Verstraete who assisted me in the BSA-AFLP procedure. Valerie Devos learned the radio-active AFLP procedure to me. Marie huysman assisted me in a lot of things and in the phylogenetic analyses. Wilson Ardiles assisted in sequencing and a lot of people from the PSB learned little things to me, necessary in this research.

After the linkage mapping, a new era of next generation sequencing analyses started for me. A lot of learning was needed in this new challenging field. Yao-Cheng Lin, Stephane Rombauts, Lieven Sterck and Frederik Delaere from the bio-informatics and IT groups teached me the first steps in Linux and in a lot of programs. The VIB bio-informatics training and service facility (BITS) organized bio-informatics courses and whenever a problem arose, Joachim Jacob from BITS was always willing to help. In the lab, as members of our little corner (Stephanie, myself, Annelies en then Sara), we helped each other a lot and of course, we also had a lot of fun together.

In special I would like to thank the members of the PAE, the Quantitative Genomics and Seed Development group for all the nice moment in and outside the lab, for the nice teambuildings and all the fun we made.

For people at my home front, I will switch to dutch.

Mijn familie, schoonfamilie en vrienden wil ik graag bedanken. Zonder de onvoorwaadelijke steun van mijn ouders, die me alle kansen gegeven hebben, had ik hier nu niet gestaan. Ook mijn broers en schoonzussen wil ik bedanken voor alle steun. Mijn nichtjes, Laura en Julie, zijn erbij gekomen tijdens mijn doctoraat. Jullie komst was enorm mooi.

Tot slot, Shanah, wil ik jou bedanken. Jij bent het mooiste geschenk van deze laatste vier jaar. Gedurende mijn doctoraat heb ik jou leren kennen en zijn we getrouwd. Onze onvoorwaardelijke liefde voor mekaar is het mooiste waar ik ooit van kon dromen.

Ives

Contents

Abbreviations

APS	adenosine 5’ phosphosulfate
AA	Amino acid
AFLP	amplified fragment length polymorphism
bp	Base pairs
BLAST	Basic local alignment search tool
BCCM	Belgian co-ordinated collections of micro-organisms
BS	bisulphite
b	Bubble size
BSA	bulked segregant analysis
BWA	Burrows Wheeler Aligner
CO_2	carbon dioxide
cM	centimorgan
CCD	charge-coupled device
ChIP	chromatin immunoprecipitation
cDNA	Complementary DNA
CIM	composite interval mapping
CDD	conserved domain database
CP	cross pollination
DDX	DEAD box protein
DDX3X	DEAD box protein 3, X-chromosomal
DDX3Y	DEAD box protein 3, Y-chromosomal
DNA	deoxyribonucleic acid
DCG	Diatom Collection Ghent
DNMT	DNA methyltransferase

S-phase	DNA synthesis phase
JGI	DOE Joint Genome Institute
DRM	Domains rearranged methylase
DSX	Doublesex
DMY	*Drosophyla melanogaster* Y
EPA	eicosapentaoenic acid
EDTA	Ethylenediaminetetraacetic acid
E-value	Expect value
EST	Expressed sequence tag
FDR	False discovery rate
FEM	Feminizing gene
FS	Full-sib
FUS1	Fusion 1
GS	Genome sequencer
Hh	Hedgehog
HMG	High mobility group
HD	Homeodomain
IRD	infrared dye
INDEL	Insertion/deletion
LRR	Leucine rich repeat
LOD	Logarithm of the odds
LOH	loss-of-heterozygosity
MSY	Male specific region of the Y chromosome
MP	Mapping population
MP	Mate pair
MAT	Mating type
MT	Mating type

ML	Maximum likelihood
MB	Mega bases
MET1	Methyltransferase1
MY	Million years
MID	Minus dominance
MOG	Myelin oligodendrocyte glycoprotein
NCBI	National Center for Biotechnology Information
NJ	Neighborhood joining
NGS	next generation sequencing
PE	Paired end
PCI	Phenol:chlorophorm:isoamyl alcohol
PCR	Polymerase chain reaction
PC	Primer combination
PK	Protein kinase
PAR	Pseudo-autosomal region
PPi	pyrophosphate
qPCR	quantitative PCR
QTL	Quantitative trait loci
RACE	Rapid amplification of cdna ends
RPKM	Reads Per Kilobase of exon model per Million mapped reads
RT	Real Time
RIL	Recombinant Inbred Line
REC	Recombination frequency
SAR	Rhizaria Stramenopila Alveolata
RNA	Ribonucleic acid
rDNA	ribosomal DNA
SAM	S-adenosyl methyltransferase

SEG	Segregation type
SRY	Sex determining region Y
SXL	Sex-lethal gene
SAD1	Sexual adhesion1
SST	sexual size threshold
HEL-SAM	SF2-family related helicase/S-adenosyl methionine dependent methyltransferase
SiO_2	Silicium dioxide
SIM	Simple interval mapping
SE	Single end
SNP	Single nucleotide polymorphism
TRA	Transformer gene
TPT	Triose phophate transporter
VCF	variant call format
WGS	Whole genome sequencing
w	Word size
Y1H	Yeast One Hybrid

1

Introduction

Diatoms: general description

Ecological importance

Diatoms (Bacillariophyceae) are single-celled or colonial microbial eukaryotes, which are photosynthetic and important members of planktonic and benthic microbial communities. They are one of the most speciose groups of algae and are major players in ocean biogeochemistry, responsible for ~20% of the world's primary production (C fixation; in the same range compared to the contribution of the rainforests) (NELSON *et al.* 1995; FIELD *et al.* 1998). Diatom abundance is generally highest at the beginning of spring and in autumn, when nutrients are not limiting and when light intensity and day length are optimal for photosynthesis (SCALA and BOWLER 2001). Availability (e.g. via upwelling) of the macronutrients nitrate, phosphate and silicic acid and the micronutrient iron has been shown to play a critical role in the development of extensive diatom blooms (BRULAND *et al.* 2005). The diatom requirement for silica means that they are a critical component of global biogeochemical silica cycling (FALKOWSKI *et al.* 1998). Diatom blooms, can sometimes be harmful due to the production of neurotoxins such as domoic acid (one of the causes of amnesic shellfish poisoning). These blooms can have negative impacts on the local ecosystem, as well as on fishing and aquaculture activities (SCALA and BOWLER 2001). Diatom blooms are accompanied by considerable carbon dioxide drawdown in the ocean surface layer. With bloom termination, organic matter sinks out of the surface layer and is remineralized by microbes and zooplankton and serves as a base for marine food webs (SMETACEK *et al.* 2012). A small fraction of this sinking organic matter however escapes consumption and settles on the sea floor, where it is sequestered over geological time scales in sediments and rocks and contributes to petroleum reserves (ARMBRUST 2009). This fraction of carbon is removed out of the atmosphere and thus, this cycling is very important concerning the global C-cycle and hence, global warming.

Semi-enclosed littoral ecosystems (including bays, estuaries, lagoons, and deltas) are very important and highly productive ecosystems worldwide (GUARINI *et al.* 2008). They are often characterized by the presence of extensive intertidal flats that sustain intense benthic microalgal primary production. Furthermore, benthic microalgae inhabiting intertidal sediments (the so-called “microphytobenthos”, which is mainly composed of pennate diatoms in temperate ecosystems) produce exopolysaccharides (STAATS *et al.* 2000) which may increase the stability of cohesive sediment and reduce mud erosion.

Main features and evolution

On the basis of the analyses of molecular and morphological characters, several major eukaryotic lineages (the so-called eukaryotic supergroups) are currently recognized (figure 1). The SAR clade concept (Stramenopila, Alveolata, and Rhizaria) has recently proven to be more stable than the traditional concept of the Chromalveolates (KATZ 2012).

The diatoms form a monophyletic group within the Stramenopila lineage (BHATTACHARYA et al. 1992; MEDLIN et al. 1993; VAN DEN HOEK et al. 1995; KOOISTRA et al. 2003).

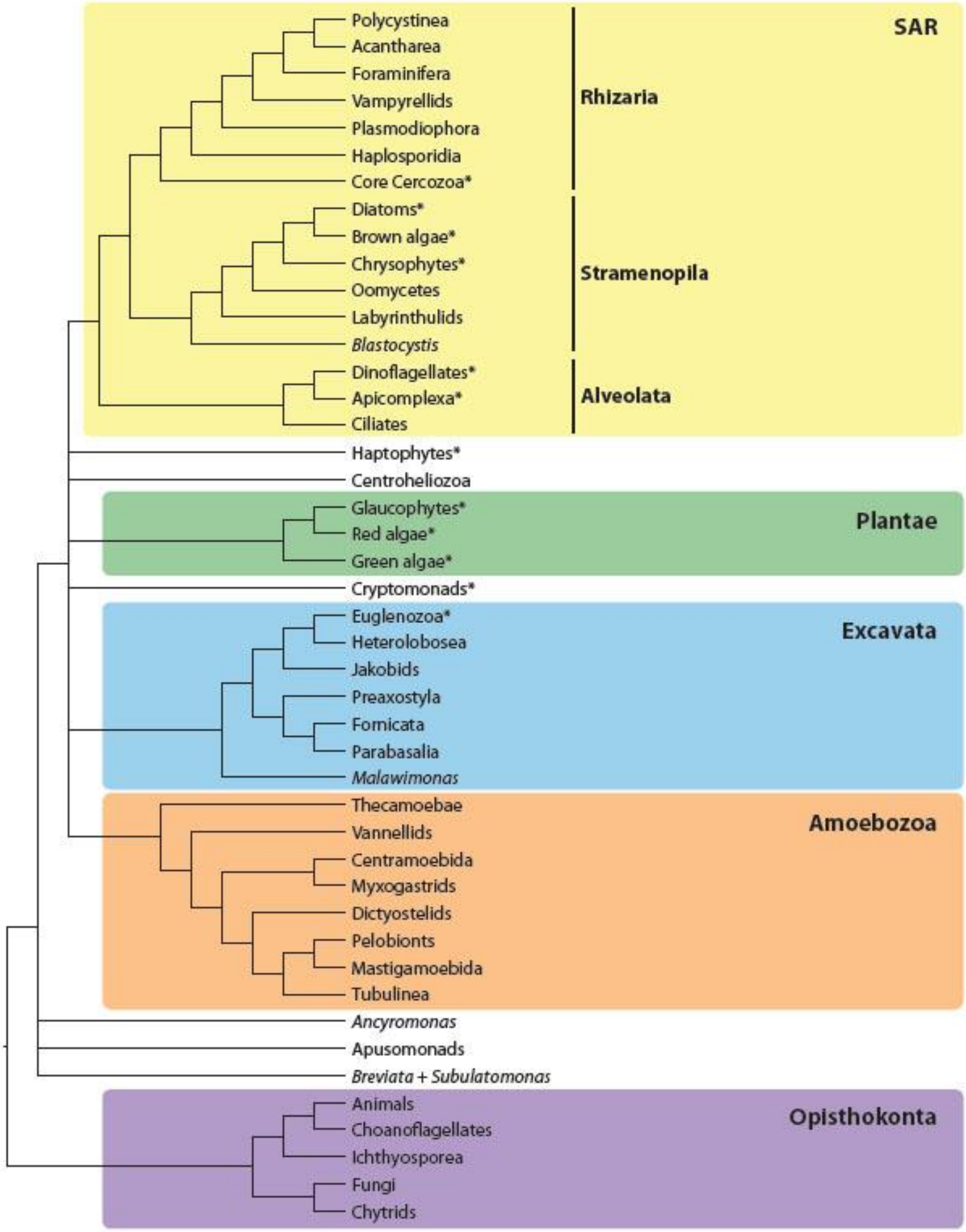


Figure 1. Phylogenetic relationships among representatives of the major eukaryotic lineages. Groups with members that have plastids are marked with an asterisk (KATZ 2012)**.**

Oxygenic photosynthesis found its origin in the prokaryotic cyanobacteria, and spread throughout the eukaryotic branch via a number of endosymbiotic events (Figure 2). Below, the endosymbiotic events leading to the diatom plastid are briefly described. An initial, primary endosymbiosis occurred about 1.5 billion years ago, when a eukaryotic heterotroph engulfed (or was invaded by) a cyanobacterium which later became the photosynthetic plastid of the Plantae, the group that includes land plants and red and green algae (YOON *et al.* 2004). Genes were subsequently transferred from the symbiotic cyanobacterial genome to the host nucleus, with about 5-20 % of the Plantae nuclear genes being derived from the cyanobacterial endosymbiont (REYES-PRIETO *et al.* 2006; DEUSCH *et al.* 2008). A secondary endosymbiosis (the movement from plastids from one eukaryote to another) occurred in which a different eukaryotic heterotroph captured a red alga. Over time, the red algal endosymbiont was transformed into the plastid of the diatoms. Other algal lineages are also known to harbour red-algal-derived plastids, viz. other groups belonging to the Stramenopiles (brown algae, the chrysophytes, etc.), the cryophytes, haptophytes, apicomplexa, dinoflagellates and *Chromera velia* (ARCHIBALD 2009). Gene transfer continued from the red-algal nuclear and plastid genomes to the host nucleus. A potential, transient endosymbiosis of a green algal cell has also been inferred (MOUSTAFA *et al.* 2009). While the Chromalveolate hypothesis supported a single endosymbiosis in the common ancestor of the Chromalveolate clade, recent studies propose that a higher number of separate secondary or even tertiary endosymbiotic events could be responsible for the presence of plastids of red algal descent in representatives of the SAR clade (ARCHIBALD 2009). More than 300 bacterial gene transfers as well have been found in diatoms (BOWLER *et al.* 2008), including the potential invasion of a Chlamydial parasite (ARMBRUST 2009). Most of these bacterial gene transfers have been attributed to horizontal gene transfer between diatoms and bacteria (KEELING and PALMER 2008; MARTENS *et al.* 2008 and references cited in these reviews).

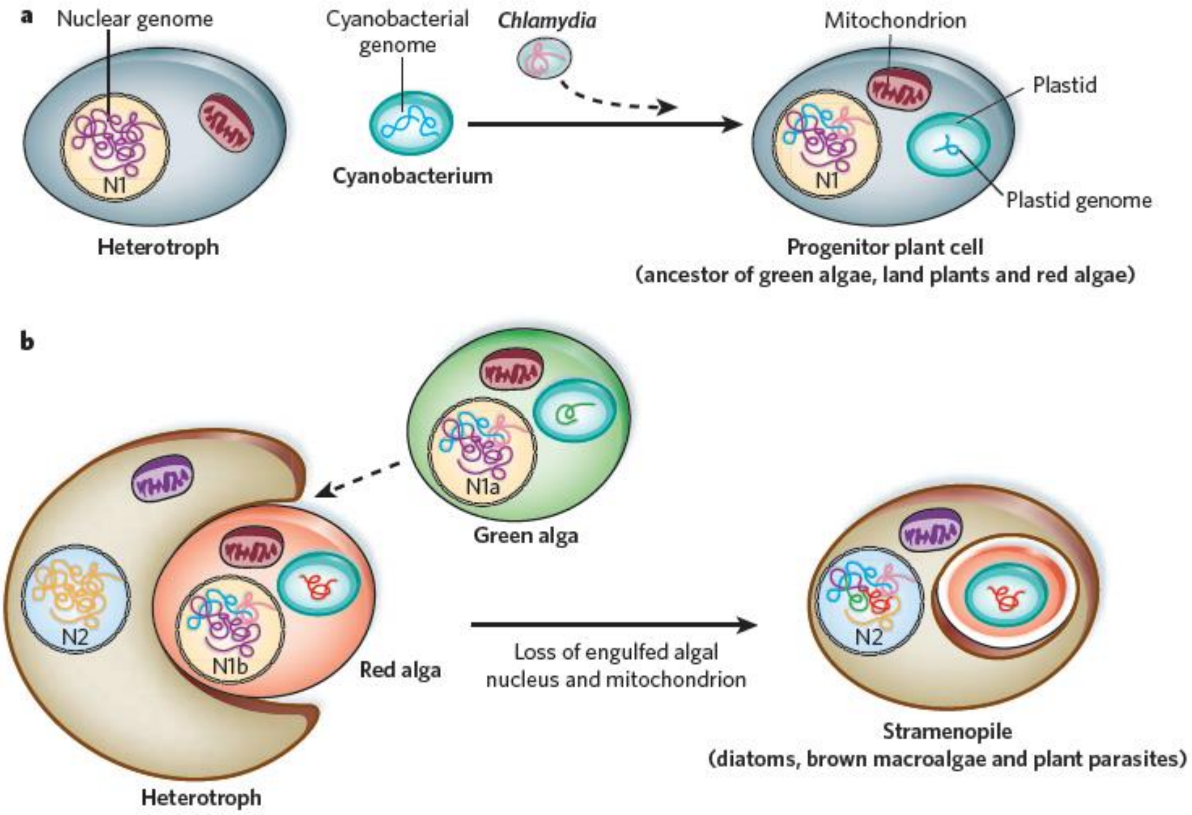


Figure 2. Representation of the origin of diatom plastids through sequential primary (a) and secondary (b) endosymbioses, and their potential effects on genome evolution. (a) During primary endosymbiosis, a large proportion of the engulfed cyanobacterial genome is transferred to the host nucleus (N1), with few of the original genes being retained within the plastid genome. The potential for invasion of the host by a chlamydial parasite is indicated with a dashed arrow, and the ensuing transfer of chlamydial genes to the host nucleus is indicated in pink. The progenitor plant cell subsequently diverged into red and green algae (incl. land plants), readily distinguished by their plastid genomes. (b) During secondary endosymbiosis, a different heterotroph engulfs a eukaryotic red alga. Potential (transient) engulfment of a green algal cell as well is indicated with a dashed arrow. The algal mitochondrion and nucleus are lost, and crucial algal nuclear and plastid genes (indicated in blue, purple and pink) are transferred to the heterotrophic host nucleus, N2. Additional bacterial genes are gained and lost throughout diatom evolution, but for simplicity this is not indicated here (ARMBRUST 2009).

The fossil record and molecular data indicate that diatoms (Bacillariophyceae) are a relatively young group which only appeared in the Mesozoic era (<205 MY) (MEDLIN *et al.* 1997; MEDLIN *et al.* 2000). However, new results based on molecular clock analyses suggest that the diatom lineage evolved sometime near the Devonian-Carboniferous transition (around 350 MY ago), and that the fossils of many diatom groups could be much older than the currently known paleontological record suggests (BROWN and SORHANNUS 2010). Since their origin, diatoms have invaded almost every aquatic and semi-aquatic environment where enough light penetrates to allow photosynthesis and have diversified into hundreds of genera and an estimated 200,000 species (MANN 1999). It is unclear which are the principal factors that have fostered the evolutionary success of diatoms but their unique life cycle (see below) is believed to have played an important role (CHEPURNOV *et al.* 2004a).

The most well-known characteristic of diatoms is their beautiful cell wall made essentially of hydrated glass ($SiO_2.nH_2O$) (DRUM and GORDON 2003). This unique type of cell wall (frustule) takes the form of two silica shells (valves) forming a box with an overlapping lid (VAN DEN HOEK *et al.* 1997). The frustule consists of two thecae (the hypotheca and the epitheca). Each theca consists of a valve and a cingulum (or girdle, composed of hoop-like rings or copulae). The silica cell walls are ornamented with species-specific patterns and structures that have formed the basis of identification and taxonomical classification over the last century (SCALA and BOWLER 2001). The ecological role of the silica wall is not well understood. It has been suggested that it may form a robust first line of defence against grazers (SMETACEK 1999).

Two major diatom groups, the centric and the pennate diatoms, are distinguished on the basis of differences in cell wall symmetry and structure, reproductive types and strategies (VAN DEN HOEK *et al.* 1997) and molecular data (MEDLIN 2011). Pennate diatoms are elongated and bilaterally symmetrical in valve view, whereas most centric diatoms are radially symmetrical (figure 3). Studies on morphology, the fossil record and molecular data leave little doubt that the centric group is ancestral to the pennate group (CHEPURNOV *et al.* 2004).

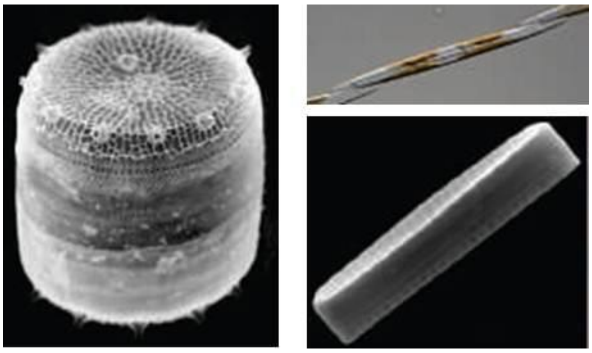


Figure 3. Typical valve symmetry of centric and pennate diatoms. The centric *Thalassiosira pseudonana* (left) and the raphid pennates (right) *Pseudo-nitzschia multiseries* (top) and *Fragilariopsis cylindricus* (bottom) (ARMBRUST 2009).

Recent phylogenetic studies and molecular clock analyses are shown in figure 4a and figure 4b. Authors suggest that the radial centric class (Coscinodiscophyceae) originated as the oldest diatom clade at least over 180 MY ago. The class of the bipolar centrics (Mediophyceae) probably diverged about 183-238 MY ago. The early divergence of the pennates (Bacillariophyceae) into three major clades, basal araphid, core araphid and raphid diatoms took place over a short period of time, with all major clades of araphid diatoms having appeared by the end of the Cretaceous in all analyses (MEDLIN 2011). There is as yet no consensus about the monophyly (figure 4a) of the above-mentioned main diatom groups, as other studies support paraphyly (figure 4b) of these groups (CHEPURNOV *et al.* 2008; THERIOT *et al.* 2009).

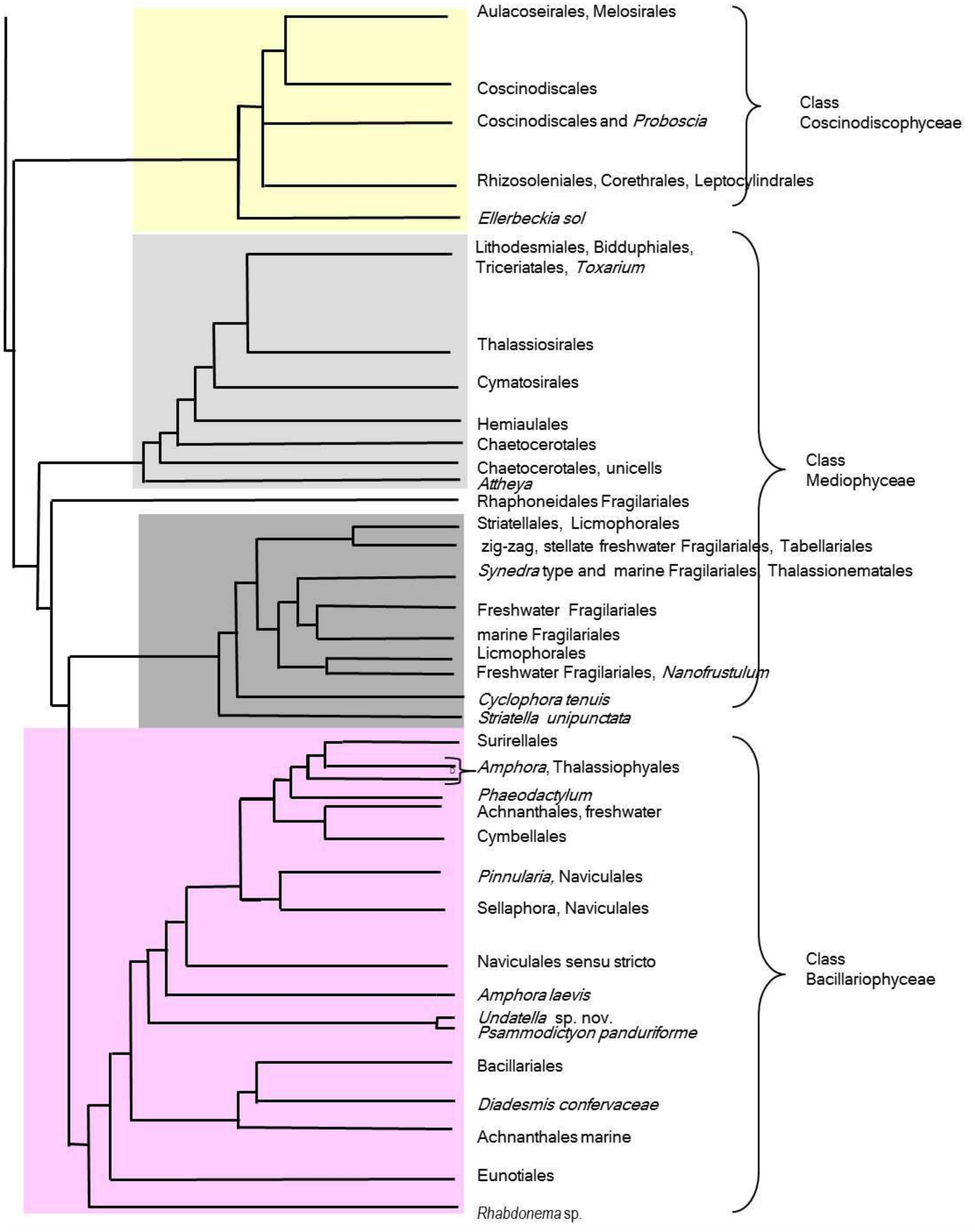

Aulacoseirales, Melosirales
Coscinodiscales
Coscinodiscales and *Proboscia*
Rhizosoleniales, Corethrales, Leptocylindrales
Ellerbeckia sol
Class
Coscinodiscophyceae
Lithodesmiales, Bidduphiales,
Triceriatales, *Toxarium*
Thalassiosirales
Cymatosirales
Hemiaulales
Chaetocerotales
Chaetocerotales, unicells
Attheya
Rhaphoneidales Fragilariales
Striatellales, Licmophorales
zig-zag, stellate freshwater Fragilariales, Tabellariales
Synedra type and marine Fragilariales, Thalassionematales
Freshwater Fragilariales
marine Fragilariales
Licmophorales
Freshwater Fragilariales, *Nanofrustulum*
Cyclophora tenuis
Striatella unipunctata
Class
Mediophyceae
Surirellales
Amphora, Thalassiophyales
Phaeodactylum
Achnanthales, freshwater
Cymbellales
Pinnularia, Naviculales
Sellaphora, Naviculales
Naviculales sensu stricto
Amphora laevis
Undatella sp. nov.
Psammodictyon panduriforme
Bacillariales
Diadesmis confervaceae
Achnanthales marine
Eunotiales
Rhabdonema sp.
Class
Bacillariophyceae

Figure 4a. Phylogenetic tree (based on the 18 S rDNA gene) supporting monophyly of the main diatom clades and their phylogenetic relations. Yellow: radial centrics; light grey: bi- and multipolar centrics; dark grey: araphid pennates; purple: raphid pennates (MEDLIN 2011).

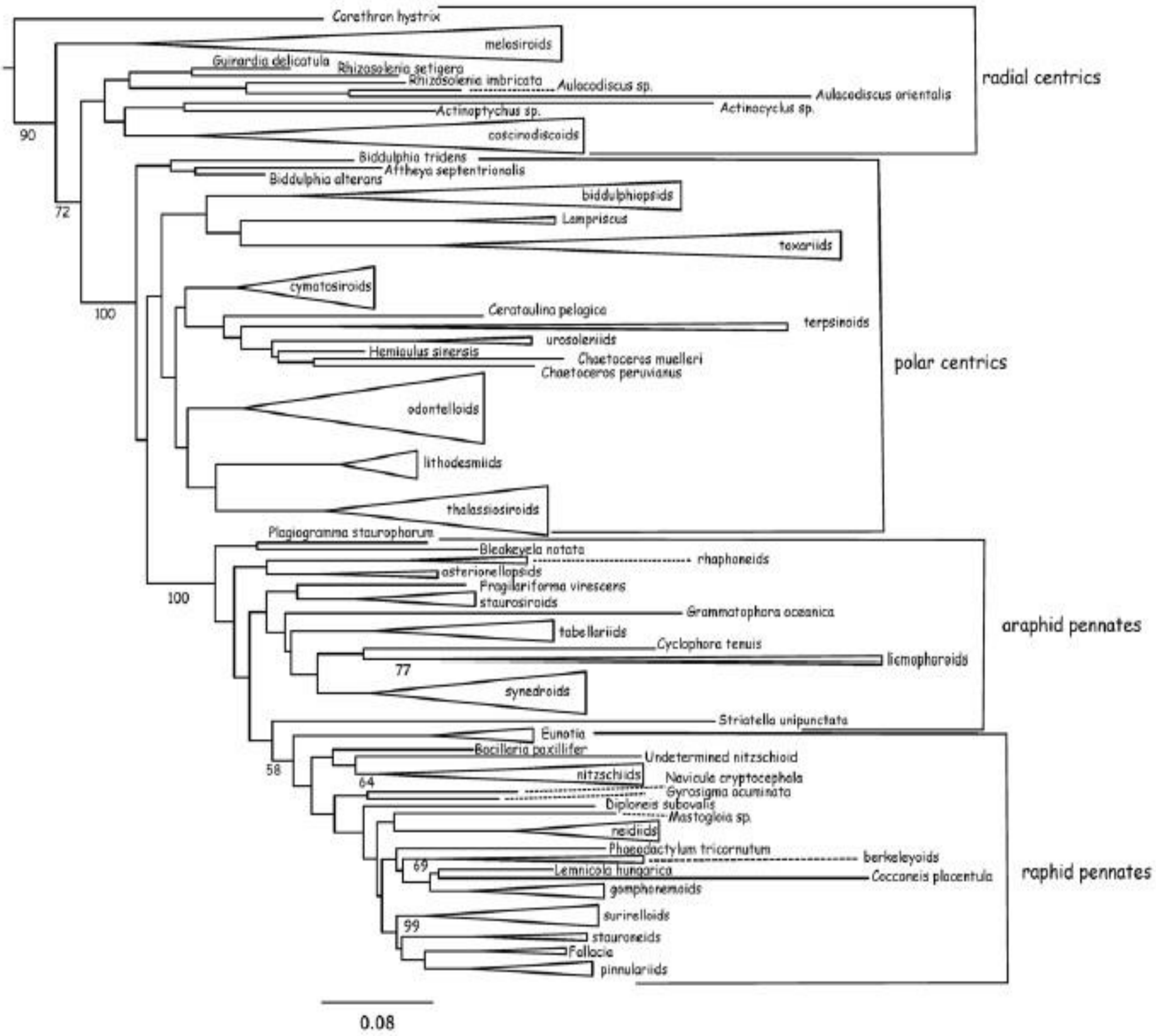


Figure 4b. Maximum likelihood phylogeny inferred from combined SSU, *rbc*L and *psb*C data. Height of triangles reflect relative number of taxa sampled. Length of the triangles reflects longest distance betweena terminal taxon and the most recently shared node for that clade. Bootstrap values are only shown for nodes on the main "trunk" of the tree. This tree is supporting paraphyly of the main diatom clades and their phylogenetic relations (Theriot *et al.* 2009).

In spite of the fact that the pennate lineage has only existed for about 90 million years, their genome structures are dramatically different from those of the centric diatoms, and a substantial fraction of genes (40%) is not shared by these representatives of the two lineages (BOWLER *et al.* 2008). Analysis of molecular divergence compared with yeasts and metazoans reveals rapid rates of gene diversification in diatoms (BOWLER *et al.* 2008).

Current and potential applications of diatoms

The most nutritionally relevant biomolecules produced by diatoms are unsaturated fatty acids like eicosapentaenoic acid (EPA, 3.9–5% of dry weight in Phaeodactylum tricornutum (LEBEAU and ROBERT 2003; MEISER *et al.* 2004), arachidonic acid (LEBEAU and ROBERT 2003), docosahexaenoic acid (KROTH 2007) and other omega-3 fatty acids (SHEEHAN *et al.* 1998). High quality vitamins and food supplements are also synthesised by diatoms (BECKER 2007). These algal products have applications as health food, cosmetics and animal feed (BOZARTH *et al.* 2009).

Silicon originating from frustules can be used as a cleaning agent (pollution remediation) or for nanotechnology applications. Other industrial application include the petroleum reserves and phytoremediation of heavy metals contamination (BOZARTH *et al.* 2009).

Diatoms can also become important for the production of biofuels. Until recently, biodiesel production has been derived from terrestrial plants, leading to competition between biodiesel and food production. Microalgae have the potential to synthesize 30 times more oil per hectare than terrestrial plants without competing for agricultural land (GRAHAM *et al.* 2012).

Protocols for genetic transformation are available for diatoms [see below and Saade and Bowler (SAADE and BOWLER 2009)]. Transgenic diatoms can be cultivated in closed bioreactors, which is an advantage in comparison with the cultivation of transgenic land plants, which are not cultivated in closed systems (BOZARTH *et al.* 2009).

There are still major bottlenecks for diatom applications as the cultivation of diatoms and the extraction of metabolites have not been optimized yet (LEBEAU and ROBERT 2003). The main limitations are the production costs and the palatability (BOZARTH *et al.* 2009). A better understanding and control of the diatom life cycle (including the aging of cultures and the use of genetic transformation) will be essential for diatom culturing. Since life cycle traits have only been studied in a minority of diatom species (CHEPURNOV *et al.* 2008), a better understanding of the vegetative and sexual cycle of diatoms could significantly improve the potential of diatom applications and this thesis will contain a step forward in this understanding.

The diatom cell and life cycle

General features

The diatom life cycle comprises two main phases, namely a long vegetative phase lasting months to years, during which the cells divide mitotically, and a relatively short sexual phase (figure 5) which includes gametogenesis and fertilization and only lasts for several hours. After this, a complex developmental process leading to the formation of a new vegetative initial cell takes place, lasting several hours to a week or more. The diatom life cycle has several unique features (CHEPURNOV *et al.* 2004b):

(1) The life cycle in diatoms is diplontic. This is almost unique for algae (MANN 1993). Vegetative cells are diploid and the only haploid cells are the gametes, which have a very short life span.

(2) Vegetative multiplication is accompanied by gradual cell size reduction. This principal is known as the MacDonald-Pfitzer rule (MACDONALD 1869; PFITZER 1871). The two thecae of a single cell are unequal in size, with one being slightly smaller than the other one. This can be represented as a box-and-lid structure (ROUND *et al.* 1990). After cytokinesis, each daughter cell occupies one of the thecae of the mother cell. As a result, each daughter cell inherits one parental theca which becomes the lid or the epitheca of the new frustule. To complete its frustule, each daughter cell then manufactures a new theca, the hypotheca. These hypothecae are always contained within, and therefore smaller than, the epithecae. As a consequence, one daughter cell is slightly smaller than the parent, whereas the other one has the same size of the parental cell. With repeated cell divisions, the average cell size of the population gradually decreases.

(3) Maximal cell size is restored through development of a specialized cell, called the auxospore. The formation of auxospores results from sexual reproduction. Gametes are produced and form a zygote and then an auxospore, which expands. When expansion is complete, a new initial cell is formed inside the auxospore envelope. This initial cell, which begins dividing after formation, is two to three times larger than the parental cells (CHEPURNOV *et al.* 2004).

(4) Cells that fail to undergo sexual reproduction and auxosporulation continue dividing mitotically until they become critically small and finally die (clonal cell death).

(5) The capacity of cells to become sexualized is cell size dependent. This is called the sexual size threshold (SST) (CHEPURNOV *et al.* 2004; GILLARD *et al.* 2013) which is species specific. Only cells smaller than the SST are able to start sexual reproduction.

Figure 5 summarizes the basic plan of the diatom life cycle (i.c. a pennate diatom, see below) and the place of sexual reproduction in it. There is overwhelming evidence that this operates in all of the main diatom lineages (KOOISTRA *et al.* 2003).

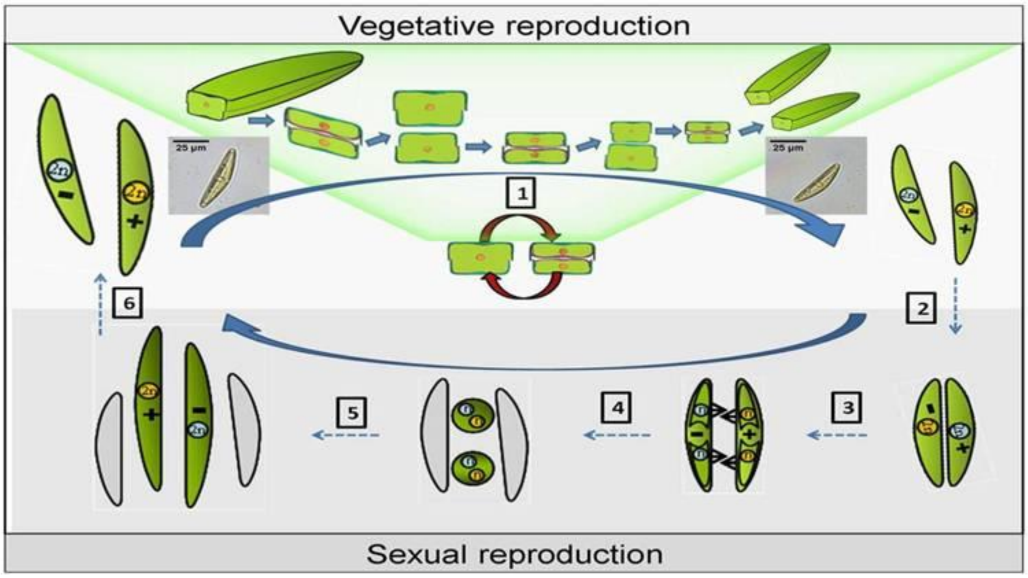


Figure 5. Overview of the life cycle in a heterothallic pennate diatom (*Seminavis robusta*). (1) Vegetative reproduction with cell size reduction. (2) Transition to sexual reproduction where cells of different mating type pair. (3) Formation of gametes. (4) Formation of zygotes. (5) Formation of initial cells via expansion of auxospores. (6) Transition to vegetative reproduction.

Variation in reproductive features

In centric diatoms, oogamy occurs during sexual reproduction. Oogamy involves the fertilization of a large, nonmotile egg by a small, anteriorly uniflagellate sperm (CHEPURNOV *et al.* 2004b). Oogamy is an extremely differentiated form of anisogamy. In pennate diatoms, both anisogamy and isogamy can occur. Anisogamy involves the production of morphologically and/or physiologically (incl. behaviourally) gametes (CHEPURNOV *et al.*

2004; DAVIDOVICH *et al.* 2010). No flagellated gametes have ever been reported in pennates and the mating type plus (MT^{+}) and mating type minus (MT^{-}) gametes do not usually differ in size (CHEPURNOV and MANN 2004). Motile gametes have been observed in araphid pennate diatoms (SATO *et al.* 2011), but here motility is associated with the extrusion and retrieval of microtubule-based ‘threads’. Isogamy involves the production of physiologically and morphologically identical gametes (CHEPURNOV *et al.* 2004b). Diatoms thus show an evolutionary transition from oogamy in centrics to isogamy in pennates (CHEPURNOV *et al.* 2008), which is the reverse of the evolutionary trend present in other eukaryotic lineages (RANDERSON and HURST 2001).

In centric diatoms, production of both types of gametes occurs in monoclonal cultures. These diatoms are called homothallic (CHEPURNOV *et al.* 2008). In pennates, heterothally seems to be the most common form of mating (ROSHCHIN 1994; MANN *et al.* 1999; CHEPURNOV and MANN 2004). Infrequently however, intraclonal (homothallic) reproduction of male clones can be observed in heterothallic pennate diatoms (e.g. *T. tabulata* and *F. delicatissima* (ROSHCHIN 1994).

Different types of sexual reproduction exist in diatoms. Allomixis refers to a fully sexual mode of reproduction, in which the fertilization occurs between gametes derived from different individual cells. Automixis refers to a form of reduced sexuality and genetic recombination in which the fertilization occurs between gametes or haploid nuclei derived from the same cell. Automixis has been observed in both the centric (DREBES 1977) and pennate (MANN 1994) lineages and can be either facultative or obligate. There are two main types of automixis. In both, meiosis occurs in an unpaired mother cell. Then either two normally differentiated gametes fuse together (paedogamy; e.g. *Nitzschia fonticola* (TROBAJO *et al.* 2006)) or meiotic cytokinesis is supressed and two of the four tetrad nuclei fuse in undivided protoplasts (autogamy; e.g. *Pinnularia nodosa* (POULICKOVA and MANN 2008)) (GEITLER 1979). Paedogamy has never been reported to occur in the centric taxa while within the pennate group facultative paedogamy and paedogamy occur (CHEPURNOV *et al.* 2004). Apomixis refers to auxosporulation in which the formation of the auxospore resembles that seen in allo- and automixis, but in which meiosis is replaced by pseudomeiosis (a process involving stages resembling meiotic prophase, but without reduction in ploidy), followed by a mitotic division, or by differentiation of a pseudozygote (a cell resembling the zygote of related allomictic or automictic species) from a diploid vegetative cell. Apomixis is documented in several species (e.g. *Eunotia* and *Amphora* species) (GEITLER 1973; NAGAI *et* *al.* 1995; SABBE *et al.* 2004). As with autogamy, asexual reproduction is a secondary modification of a basically sexual pathway of development. Haploid parthenogenesis involves the development of an unfused gamete into an auxospore and subsequently an initial cell, which may divide to produce a clonal lineage of (at least initially) haploid cells. A number of examples (CHEPURNOV *et al.* 2004b) are known from cultures and laboratory populations but it is unknown whether the phenomenon occurs in natural populations.

Diatom model organisms for molecular biology

Whole genome models

Experimental biology is largely based on a model systems approach. The choice of the model organism is often highly debatable, because uncomfortable compromises must be made between contradictory requirements (for example, large-celled species would be easier to examine microscopically, but small-celled species often grow faster). Nowadays, it is obvious that genomic techniques are essential for diatom research. The most commonly used model diatoms for which whole genome information is available are *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* (ARMBRUST *et al.* 2004; BOWLER *et al.* 2008; CHEPURNOV *et al.* 2008). These two diatoms are only distantly related and have the advantage that they represent the two major diatom groups: *Thalassiosira* is a centric diatom and *Phaeodactylum* is a pennate one. As most used models, these two satisfy some important criteria: they are easy to grow and manipulate experimentally (= laboratory convenience: short generation time, ease of stock maintenance, experimental tractability). They also both have a small genome size (~30 megabases (MB)) (CHEPURNOV *et al.* 2008; SAADE and BOWLER 2009). There is a large amount of cytological, physiological and biochemical data available for both organisms. However, sexual reproduction, cell size reduction and a size restitution cycle, key features of most diatom life cycles (CHEPURNOV *et al.* 2004), to date have never been demonstrated for these model diatoms. Without the ability to control sex, identifying genes and/or polymorphisms (e.g. natural genetic variants, induced mutations) underlying useful traits by forward genetic approaches remains impossible (CHEPURNOV *et al.* 2008).

Recently, additional diatom genomes have been sequenced. These include the pennate diatom *Fragilariopsis cylindricus* (http://genome.jgi-psf.org/Fracy1/Fracy1.home.html) and the toxic pennate diatom *Pseudo-nitzschia multiseries* (http://genome.jgi-psf.org/Psemu1/Psemu1.home.html). The life cycle of *F. cylindricus* is virtually unexplored although observations of natural populations indicate that a size reduction – restitution cycle is present (HASLE 1965). All *Pseudo-nitzschia* species that have been examined experimentally display the typical diatom life cycle. However, they are difficult for microscopical monitoring because cells occur in long chains and offer very few visible markers of cell cycle progression or sexualization. Recently, the genome of the centric oceanic diatom *Thalassiosira oceanica* was published (LOMMER *et al.* 2012). Further sequencing of diatom genomes will undoubtedly result in more insights into diatom molecular biology (CHEPURNOV *et al.* 2008).

Genetic transformation in diatoms

For many years, the major bottleneck for exploring the molecular and cellular biology of diatoms was the absence of a protocol for genetic transformation. The generation of transgenic organisms is essential for dissecting basic molecular biological processes and allows the use of reverse genetic approaches (SCALA and BOWLER 2001). The first successful generation of a diatom took place in 1995 with the production of transgenic lines of *Cyclotella cryptica* and *Navicula saprophila* (DUNAHAY *et al.* 1995). The technology is now most advanced for the pennate diatom *P. tricornutum* (SCALA and BOWLER 2001). Transformation protocols are now also available for the centric model diatom *T. pseudonana* (POULSEN *et al.* 2006) and for the pennate diatom *Cylindrotheca fusiformis* (POULSEN and KROGER 2005). All of these protocols are based on biolistic methods, which lead to random but stable integration of foreign DNA into the chromosomes of the cell nucleus (BOZARTH *et al.* 2009). Restrictions on transformation are that the cell nucleus is presently the exclusive target, that more markers generating antibiotics resistance have to be developed and that silencing techniques have to be further developed (BOZARTH *et al.* 2009).

Seminavis robusta – a new model to study the diatom life cycle

The raphid pennate diatom *Seminavis robusta* (figure 6), which is a benthic diatom colonizing semi-enclosed littoral ecosystems, has recently been advocated as a model organism to study diatom biology, and in particular life cycle regulation because it is a heterothallic species (note however that homothallic reproduction has also very infrequently been observed, P. Vanormelingen, pers. comm.), and sexual crosses can be easily experimentally manipulated and synchronized, allowing full experimental control of the sexual process (CHEPURNOV *et al.* 2008).

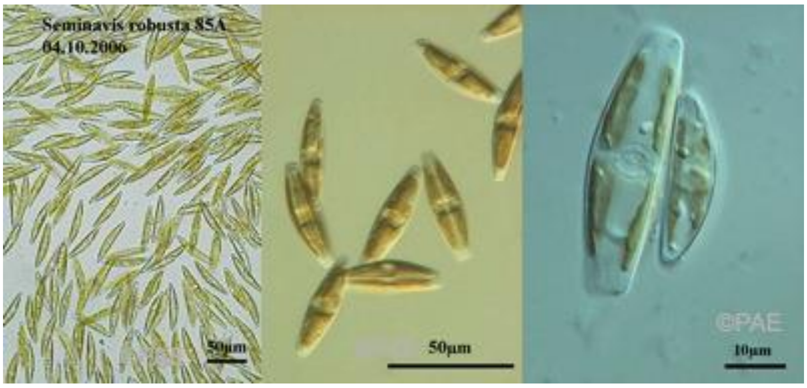


Figure 6. Vegetative cells of the pennate diatom *S. robusta*.

The cytological and reproductive characteristics of *S. robusta* are well documented (CHEPURNOV *et al.* 2002). The life cycle is ‘typical’ for diatoms with a size reduction – restitution life cycle and restoration of size taking place via sexual auxosporulation (figure 7). The potential for sexualization is, as usual for diatom species, size dependent and cells smaller than 50 µm are able to undergo the sexual phase. In *S. robusta*, two paired gametangia form two gametes apiece, which fuse to form two auxospores (figure 7). Sexual reproduction is isogamous. The auxospore develops through bipolar expansion, accompanied by the deposition of “auxospore-specific” siliceous elements (the perizonium), and finally, it is transformed into the enlarged cell of the next generation (figure 7). A key reason to select *S. robusta* as a model is that the mating system is heterothallic (CHEPURNOV *et al.* 2008). This allows the induction of sex to be controlled reliably, because sexual reproduction cannot start until two compatible clones are mixed together. Furthermore, pairing (mediated by pheromones and migration of the MT^+ (GILLARD *et al.* 2013)) and gametogenesis occur with very high frequency and with remarkable efficiency and synchrony. No other examined diatom has approached the same success in mating and F_1 development as *S. robusta* (CHEPURNOV *et al.* 2008). *S. robusta* cells are reasonably large (up to 80 µm long) and grow on surfaces. This allows easy recognition of the various stages of the cell and life cycles and direct monitoring of sexual processes in experimental vessels even under low magnifications. Cells however are only loosely attached to surfaces and can easily be suspended for e.g. molecular or biochemical analyses. The whole *S. robusta* life cycle (i.e. including vegetative and sexual reproduction) is short and can be completed in a few months if cells are grown exponentially in optimal conditions. The cell cycle is correspondingly short (ca. 0.5 days). A controlled breeding program has demonstrated that *S. robusta* is highly tolerant for inbreeding (CHEPURNOV *et al.* 2008). In contrast, severe inbreeding depression occurs in some other diatoms, even over a few generations (ROSHCHIN 1994; CHEPURNOV and MANN 2000; CHEPURNOV *et al.* 2004; CHEPURNOV *et al.* 2006). Furthermore, reliable protocols for synchronization and cryopreservation are available (CHEPURNOV *et al.* 2008; GILLARD *et al.* 2012) and a genetic transformation protocol is being optimized (Moeys S., personal communication April 2013).

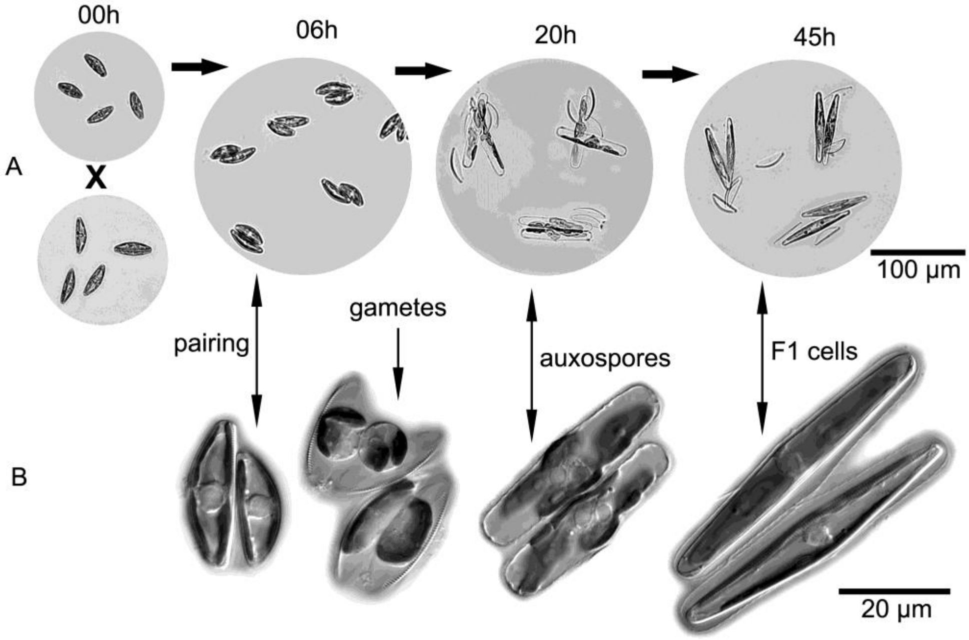


Figure 7. The life cycle of *S. robusta* (CHEPURNOV *et al.* 2008). (A) A cross (X) of two opposite mating type clones: changes in the mixed culture during two days. After 6 h, pairing of opposite MT's can be observed. After 20 h, auxospores develop; the empty parental valves can also be seen. After 45h, initial cells (F1) are formed. (B) Principal stages of sexual reproduction: pairing, gamete and auxospore formation, and the development of F1 offspring.

The phylogenetic position of a model species is also very important (BOLKER and RAFF 1997). *Seminavis* belongs to the species-rich and ecologically important clade of the Naviculaceae, which is very common in both marine and fresh water habitats. *S. robusta* is an motile benthic diatom, whereas all other sequenced diatoms to date are planktonic or belong to the sea ice community (CHEPURNOV *et al.* 2008). Another key feature for a model system is its genome size. The haploid genome size of *S. robusta* has been determined using flow cytometry (CONNOLLY *et al.* 2008) and appeared to be approximately 153 MB (CHEPURNOV *et al.* 2008). This is larger when compared to the two most used model diatoms *P. tricornutum*

(31 MB) and *T. pseudonana* (32 MB) and the other sequenced diatoms *T. oceanica* (92 MB) and *F. cylindricus* (81 MB) but smaller compared to the genome of *P. multiseries* (218 MB).

All these features together make *S. robusta* a suitable model to study the sexual process in diatoms.

Sex determination in eukaryotes

General features

Although many eukaryotic organisms, in particular micro-organisms, can reproduce both sexually and asexually, the vast majority appears to undergo sexual reproduction during their life cycles. Hypotheses advanced to explain why sex is so ubiquitous are myriad, but centre on a few recurrent themes (SMITH 1978; BELL 1982; METIN *et al.* 2010). First, the admixture of genetic material from two genetically distinct individuals that occurs during sexual reproduction may give rise to novel gene combinations that result in offspring better suited to novel or changing environments. Second, sexual reproduction may serve to remove deleterious mutations, such as transposons, that have arisen within the genome. Finally, sex might serve both roles, facilitating combinations of successful alleles and simultaneously purging the genome of deleterious ones. In response to constant environments, asexual reproduction is also likely to be of relative benefit by relieving strains from the metabolic and genetic costs associated with sexual reproduction (XU 2005). Thus, a balance between sexual and asexual reproduction may be struck in response to different environmental conditions.

Sexual development is common in eukaryotic organisms from yeasts to humans. A sexual population generally consists of two genders which is frequently genetically regulated by a pair of sex chromosomes (BERGERO and CHARLESWORTH 2009). The evolution of genes in the sex determining region is influenced by whether the organism is haploid or diploid. In contrast to mammals and other obligate diploid organisms, fungi and most algae are viable as both haploids and diploids. This has influenced the evolution of genes in the sex-determining regions in the two systems as it enables those in obligate diploids to degenerate to non-functional alleles in one of the two sex-determining chromosomes (FRASER *et al.* 2004). Sex determination systems can also be controlled by environmental cues. In many species of reptiles, sexual fate is solely determined by environmental cues present at a crucial stage during development (MANOLAKOU *et al.* 2006).

Different mating types (MTs) are characterized by the genetic compatibility among gametes, while gametes of the same MTs cannot form zygotes. Species with MTs can still have different sex roles and can mate in both the male and female role [e.g. investment in large non-motile gametes (female role) and in fertilization by small motile gametes (male role)]. Mating types cannot be linked to either role (male or female role) (NIEUWENHUIS and AANEN 2012).

Evolution of sex chromosomes

The two homologous sex chromosomes can be either heteromorphic, where one of the homologs is smaller (genetic degeneration), or homomorphic (no degeneration in one of the chromosomes). The non-recombining regions of heteromorphic sex chromosomes, such as the male specific (MSY) region of the mammalian Y chromosome, the entire *Drosophila* Y chromosome and the female specific part of the W chromosome of birds, have lost most or even all genes that were present on the ancestral chromosome by genetic degeneration. These ‘classical’ sex chromosomes have small recombining pseudo-autosomal regions (PAR regions), like in mammals and birds (figure 8a), or do not recombine with their homologues at all (*Drosophila* Y chromosome) (BERGERO and CHARLESWORTH 2009). Other sex chromosomes have extensive recombining PAR regions (in which both homolog chromosomes carry the same gene content). Their non-recombining regions can be as small as a few MB. In such cases, the homologous chromosomes are not morphologically distinguishable (homomorphic, figure 8b). These may represent recently evolved sex chromosome systems and are sometimes called proto-sex or young sex chromosomes. A third kind of sex chromosomes are called neo-sex chromosomes (figure 8c). In these, an autosomal segment fuses with one of the sex chromosomes which can immediately create heteromorphism. In the other case, autosomal segments fuse with both sex chromosomes. This case leads to non-heteromorphic sex chromosomes.

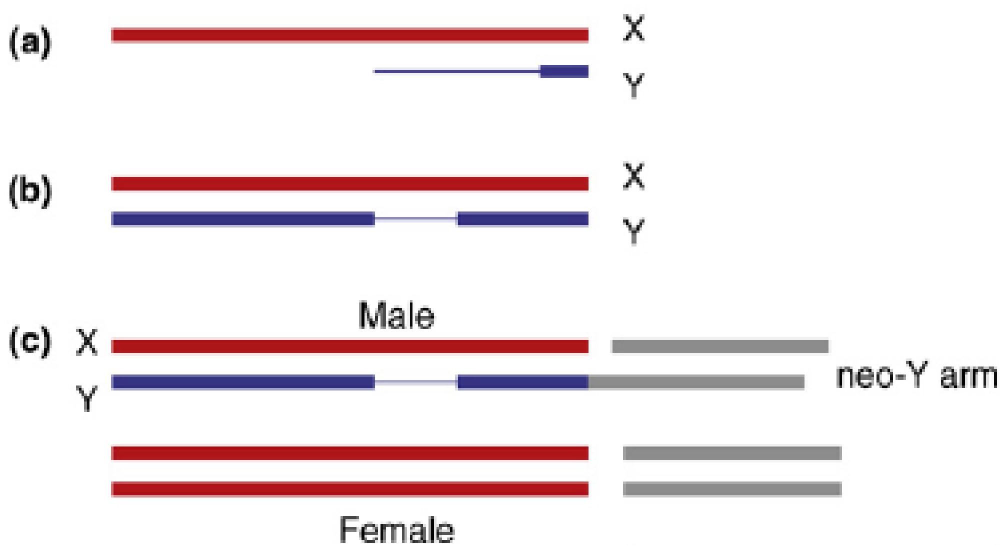


Figure 8. Diagram of different types of sex chromosomes (BERGERO and CHARLESWORTH 2009). Thick lines show regions that can recombine (pseudo-autosomal regions - PAR). The thin lines show non-recombining regions. (a) classical XY chromosomes with a small PAR region. (b) A sex chromosome-like region in an otherwise normal chromosome (large PAR region). (c) Formation of a neo-sex chromosome by fusion between a Y chromosome and an autosome.

Variation in sex differentiation in eukaryotes

Dioecy (separate sexes) has arisen independently many times during the evolution of various major eukaryote lineages (BERGERO and CHARLESWORTH 2009). The mechanisms that specify sex determination are among the least-conserved known. Marked variation exists in both the primary sex determination signal and in the downstream genetic pathways that interpret the signal (HAAG and DOTY 2005). As a consequence, not only differences in genetic structures of sex chromosomes but also differences in genomic content of sex loci in different eukaryotic species exist. Typical for sex determination factors is that they are often transcription factors, regulating downstream cellular processes including proliferation and cell migration (CAPEL 2000). Below, the sex determining systems of some major eukaryotic lineages are described.

Sex determination in the Ophistokonta lineage

In the animal kingdom, there are diverse mechanisms for determining sex, with the predominant ones being chromosomally based leading to oocyte-producing females and sperm-producing males (ZANETTI and PUOTI 2013).

In vertebrates, sex can be determined either by environmental or genetic factors (KIKUCHI *et al.* 2007), or sex can be genetically determined with limited, secondary influences from the environment as has recentlt been reported for the zebrafish (LIEW *et al.* 2012). Even in the case of animals where sex is determined by the genome, the primary sex-determining gene is not conserved in divergent species (CAPEL 2000). DMY (*Drosophila melanogaster* Y) and SRY (Sex determining region Y) have been identified as the primary sex-determining transcription factor genes in the medaka fish (*Oryzias latipes*) and in most mammals, respectively (MATSUDA *et al.* 2002a). These genes are also identified as sex determinants in insects (PAYRE *et al.* 1990). The DMY domain is described as a zinc finger-like DNA binding motif (MATSUDA *et al.* 2002b) and SRY encodes a DNA binding protein of the HMG-box (High Mobility Group box) family that recognizes both chromatin structure and a specific binding sequence (FERRARI *et al.* 1992; RIMINI *et al.* 1995). This regulatory transcription factor initiates cellular processes which in turn influence architectural patterning, fate commitment and differentiation of cells within an organ (CAPEL 2000). Apart from these described genes, the molecular nature of the primary sex-determining gene(s) in most vertebrates remains unknown (SCHARTL 2004).

In the model fly *Drosophila,* sex determination is achieved by a balance of female determinants on the X chromosome and male determinants on the autosomes (GILBERT 2000). The flies have four chromosome sets from which one sex chromosome set, two autosome set and a set of little chromosomes. If there is only one X chromosome in a diploid cell and two haploid autosomes (little chromosome set not included; 1X:2A), the fly is male. If there are two X chromosomes in a diploid cell and two haploid autosomes (little chromosome set not included; 2X:2A), the fly is female (XX). A quantitative chromosomal signal, the X:A ratio decides whether the key gene in sex determination, *SX1* (Sex lethal 1) is active (XX) or inactive (XY). The functional state, ON or OFF, of SXl, regulated via a few subordinate regulatory genes, controls a switch gene (*DSX*) that can express two mutually exclusive functions, M or F. These serve to repress either the female or the male set of differentiation genes, thus directing the cells either into the male or into the female sexual pathway (NOTHIGER and STEINMANNZWICKY 1987). In flies, the Y chromosome is not involved in determining sex. Instead, it contains genes active in forming sperm in adults.

Sex loci have been intensively studied in the fungal kingdom. While the vast majority of sexually reproducing organisms occur as just two sexes or mating types, transitions in sexuality from two to multiple mating types (MTs), and vice versa, have occurred in the fungal kingdom. Sexual reproduction is common in fungi, and mating types occur in two general patterns: bipolar and tetrapolar (KRONSTAD and STABEN 1997; CASSELTON and OLESNICKY 1998; FRASER *et al.* 2007). In the bipolar systems, a single genetic locus occurs in two alternative forms, known as idiomorphs (a or α, a or A, + or -, P or M) and these govern the identity of the cell (METIN *et al.* 2010). Co-incubation of strains of opposite MT under suitable conditions leads to sexual reproduction. Species with bipolar mating systems are found in the ascomycete, basidiomycete and zygomycete phyla, providing evidence that this is an ancestral organization within the fungi. In the basidiomycete phylum, many species however have a more complex sex determining system, known as tetrapolar, in which two unlinked sex determining loci are present (KRONSTAD and STABEN 1997; CASSELTON and OLESNICKY 1998; FRASER *et al.* 2007). One locus encodes homeodomain transcription factors and the other encodes pheromones and pheromone receptors, and both loci must differ for sexual reproduction to occur. In many species, these loci are also multiallelic, resulting in literally thousands of mating types, which promotes outcrossing (KOTHE 1996). Transitions from tetrapolar to bipolar mating type determination have occurred several times independently, and in several examples result from fusions of the two unlinked loci to form one contiguous region (FRASER and HEITMAN 2005), potentially illustrating common evolutionary pressures limiting mating types/sexes to just two.

Ascomycete sex is orchestrated by the *MAT* locus, which encodes key transcription factor genes that govern identity and developmental fate (LEE *et al.* 2010b). The two mating type alleles, MATα and MATa in *Saccharomyces cerevisiae* (HABER 1992), differ by approximately 700 bp of sequences. “a” cells harbor the MATa allele, containing the a1 gene, whose product is a HD2 class homeodomain (HD) transcription factor. In “α” cells, the MAT α locus carries the α1 gene, encoding an alpha box transcription factor, and the α2 gene encodes an HD1 class homeodomain transcription factor. Both α1 and α2 genes regulate the expression of α and a cell-specific genes, encoding pheromone receptors (LEE *et al.* 2010b).

Zygomyceta and their alleged sister taxon, the Microsporidia, exclusively share the presence of a cluster of three genes encoding a triose phosphate transporter (TPT), a high mobility group-type (HMG) transcription factor, and an RNA helicase. In these groups, the HMG-type transcription factor acts as the sole sex determinant. The evolutionary relationship between the gene clusters of the two groups was unraveled by analyzing evolutionary relationships between the TPT and the helicase genes. HMG transcription factors were omitted from this analysis because they lack any phylogenetic information (LEE *et al.* 2010a).

A screen in 15 plant, animal and fungal species for orthologues of the TPT and helicase genes of the Microsporidia and the Zygomyceta resulted in four unrelated ortholog groups which each contained sequences of plants, animals and fungi (figure 9). This suggests that the genes in the zygomycete sex-related region have separated from their microsporidian homologs already before the eukaryotic kingdoms emerged. It is thus proposed that an ancestral gene cluster already existed in the common ancestor of plants, animals and fungi (KOESTLER and EBERSBERGER 2011).

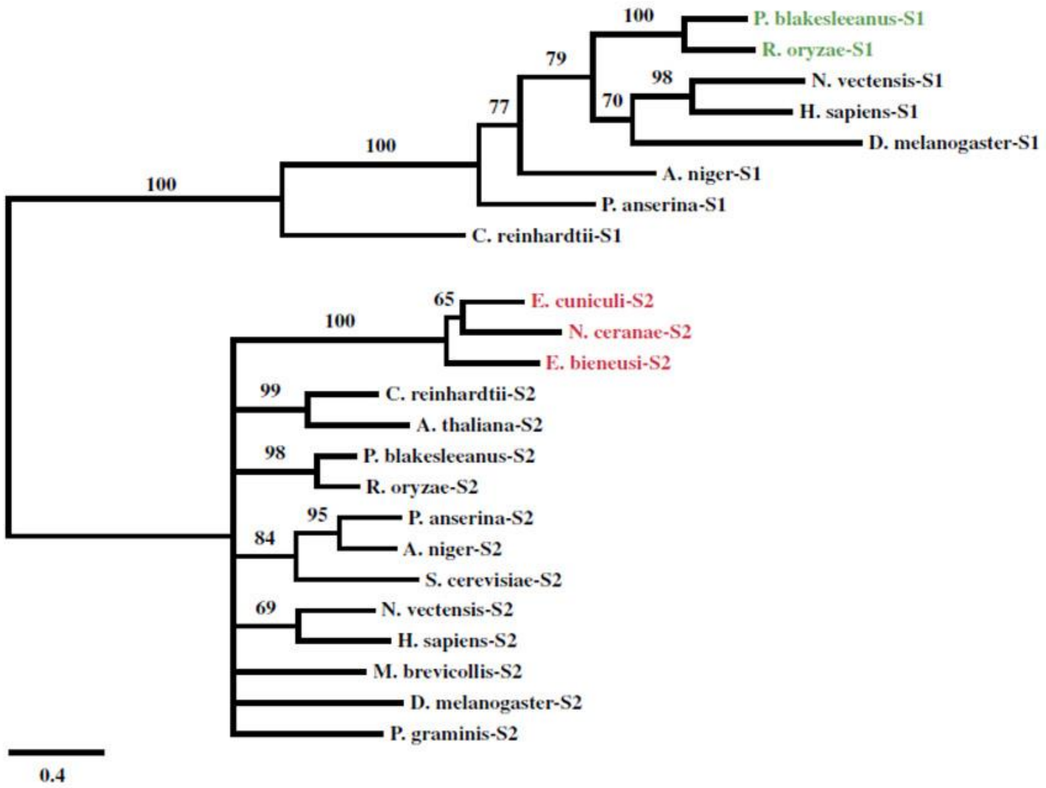


Figure 9. ML tree of the S1 and S2 RNA helicases. Sequences in the zygomycete sex-related region are labeled in red, and sequences in the corresponding region of the microsporidia are labeled in green. Branch labels denote bootstrap support values.

Sex determination in the Plantae lineage

Genetic factors or environmental conditions control sex determination in land plants (CHARLESWORTH 2013). In monoecious (the same gametophyte produces both sperm and eggs) species and in environmental sex determination, genes are, of course, involved in the sex-determining developmental pathway, but there are no sex-determining loci, since all individuals are either monoecious, and capable of developing flowers of either sex, or can develop as either sex, depending on the environment experienced in a given flowering season.

In contrast, many dioecious (gametophytes produce only sperm or eggs) species have a genetic polymorphism involving sex determining genes, or ‘primary sex-determining genes’, which control whether a plant as a whole develops as a male or female (e.g. genes suppressing female functions on a Y chromosome, and loss of function of male fertility factors on the X). Sex determining systems in plants originated independently many times during the evolution. Plants can thus be used to study the time course of events during sex chromosome evolution (CHARLESWORTH 2013).

Green algae seem particularly promising for studying the evolutionary change from isogamy to anisogamy (KIRK 2006) because some taxa, including *Chlamydomonas* species, are isogamous, with no gamete size differences, while others, including *Volvox* species, are anisogamous. In *C. reinhardtii*, the MT locus is within a multi-gene region with suppressed recombination. In *C. reinhardtii*, at least two genes in the MT region are known to be involved in MT determination, although the details of the processes involved remain unknown (CHARLESWORTH and CHARLESWORTH 2010). There is a large non-recombining MT region (more than 200 kilobases (kb)), but this size differs between the two MT alleles. The genes determining gamete size difference are linked to the MT locus (FERRIS *et al.* 2010). Sequencing of *V. carteri* genomic DNA has identified a region containing several genes also known to be present in the *C. reinhardtii* MT region (CHARLESWORTH and CHARLESWORTH 2010).

Sex determination in the Amoebozoa lineage

The most-studied species of social amoeba, *Dictyostelium discoideum*, is notable for having three sexes. Each of the three sexes can pair with each of the other two but not with itself. Single, unrelated genes are sufficient to determine two of the mating types, whereas homologs of both these genes are required in the composite type. The key genes encode polypeptides that possess no recognizable similarity to established protein families. Sex determination in the social amoebae appears to use regulators that are unrelated to any others currently known (BLOOMFIELD *et al.* 2010).

Sex determination in the SAR (Rhizaria, Stramenopila, Alveolata) lineage

The study of sex loci in the SAR lineage is still in its infancy and there are no identified sex loci in this lineage yet. In the brown alga *Ectocarpus siliculosus*, sex is determined by a single locus (COELHO *et al.* 2011). The complicated genetics of mating have been studied in the Oomycetes *Phytophthora infestans* and *P. parasitica*. A MT locus, determined as a single locus, has been identified that exhibits non-Mendelian inheritance in *P. infestans*. This is consistent with a system of balanced lethal loci near the MT locus. Mendelian inheritance was observed in *P. parasitica* (JUDELSON *et al.* 1995; JUDELSON 1996; FABRITIUS and JUDELSON 1997). In diatoms, the rule is that centric diatoms are homothallic while pennates are heterothallic (MANN *et al.* 1999; CHEPURNOV and MANN 2004). The evolution from homothally (centrics) to heterothally (pennates) involves apparently the evolution to genetic MT determination (CHEPURNOV *et al.* 2004a; DAVIDOVICH *et al.* 2010). This evolution is belevieved to be as the normal evolutionary drive (FIGUEROA *et al.* 2010). Experimental evidence suggests also that MT determination in heterothallic pennate diatoms is genetic as Mendelian segregation patterns seem to occur (CHEPURNOV *et al.* 2004a; DAVIDOVICH *et al.* 2010). But how mating type determination is achieved is to date still unknown.

Main techniques used

AFLP technology

The AFLP method is a DNA fingerprinting technique based on selective PCR amplification of restriction fragments from a total digest of genomic DNA (VOS *et al.* 1995; VUYLSTEKE *et al.* 2007).

AFLP markers offer several advantages over other currently used DNA markers, such as simple sequence repeats and single nucleotide polymorphisms. Foremost among these is that the AFLP technology requires no prior sequence information and, hence, has a relatively low start-up cost. In addition, the AFLP technique is very amenable to automation and is highly multiplexed, which offers the potential to improve the efficiency and to increase the throughput of marker data production in organisms that lack the genomics platform (VUYLSTEKE *et al.* 2007).

Different steps are involved during AFLP technology (VUYLSTEKE *et al.* 2007) (figure 10). The first step of the AFLP procedure involves the preparation of templates by restriction digestion of DNA, typically with two different restriction enzymes. The two restriction enzymes used are generally (but not necessarily) a rare cutter and a frequent cutter. The frequent cutter is used to generate fragments that are in the 50–500 bp length range resolvable by electrophoresis. The rare cutter is used to limit the number of fragments that can be amplified and, hence, to define the number of effective AFLP amplicons. Subsequently, double-stranded adapters are ligated to the ends of the restriction fragments. The second step of the AFLP procedure is the PCR amplification of subsets of restriction fragments using selective AFLP primers. The common parts of these primers correspond to the adapter and restriction enzyme recognition sequences, and they have a number of additional bases at the 3'-end extending into the restriction fragments, called the selective nucleotides. These selective nucleotides ensure that only a subset of restriction fragments is amplified to a detectable level. AFLP fingerprinting of relative large genomes (greater than 100 Mb) is usually carried out with two or three selective bases in one or both primers and is generally performed in two consecutive steps: a pre-amplification step, reducing the complexity of the template mixture, and a final selective step. Detection of the AFLP fragments is made possible by radioactive or fluorescent labeling of one of the two AFLP primers used in the final selective amplification reaction. The final step of the AFLP technique is the electrophoretic size fractionation of the fingerprints. For this purpose the labeled reaction products are separated on denaturing polyacrylamide gels. In the case of conventional gel electrophoresis using radio-labeled primers, gels are either dried on paper or fixed on glass plates after electrophoresis, and AFLP images may be generated using either conventional autoradiography or phosphorimaging technology. In the case of gel electrophoresis using infrared dye (IRD) or fluorescently labeled primers, AFLP images may be generated using LI-COR.

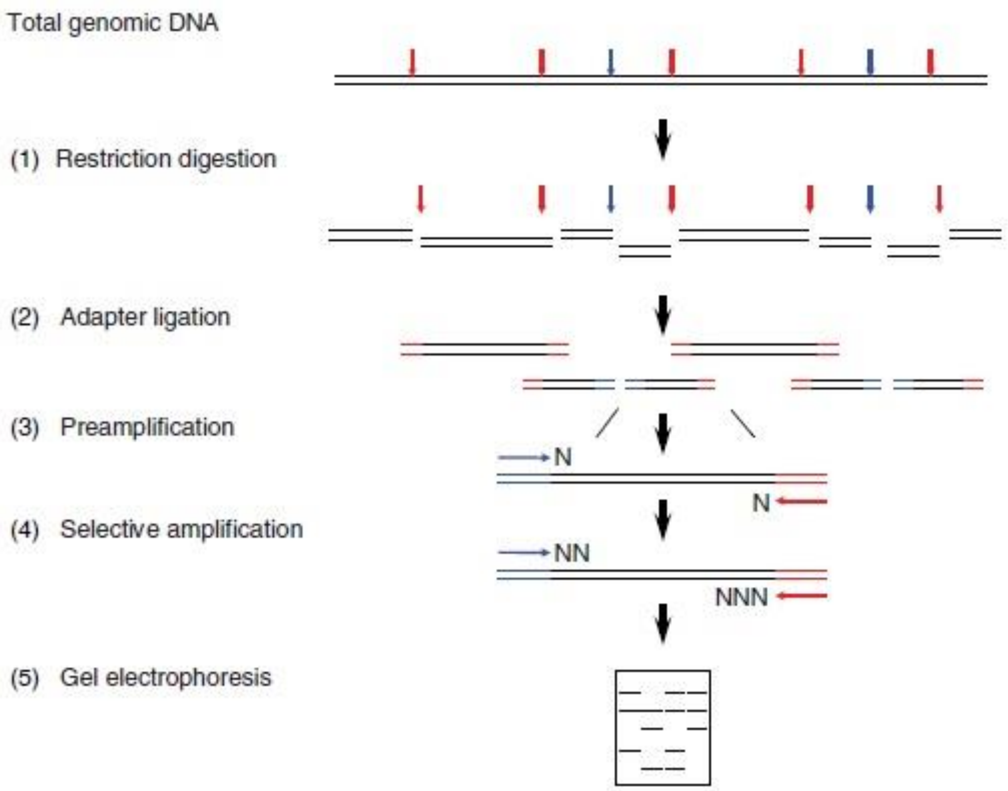


Figure 10. Outline of the AFLP procedure (VUYLSTEKE *et al.* 2007). Template fragments are generated by: (1) digestion of genomic DNA with a combination of the two restriction enzymes *Eco*RI and *Mse*I (blue and red arrows represent *Eco*RI and *Mse*I restriction enzyme sites, respectively); (2) ligation of the double-stranded *Eco*RI- (blue) and *Mse*I- (red) specific adapters to the fragment ends; (3) a pre-amplification step using primers that match the adapter sequences and that carry each one selective nucleotide (represented by N) at their 3' end are used to PCR-amplify subsets of the *Eco*RI/*Mse*I templates; (4) a final selective PCR-amplification step in which additional selective nucleotides are added to the *Eco*RI and *Mse*I primers; and (5) the electrophoretic size fractionation and the display on denaturing polyacrylamide gels of the *Eco*RI/*Mse*I amplification products.

Genetic linkage mapping

A genetic linkage map is a representation of the genome that shows the relative position and distances between markers or genes along chromosomes. It does not show the physical distance between these markers but the genetic distance, defined as a function of the crossover frequency during meiosis. The closer the two genes are on a chromosome, the lower the chances are that crossing over will occur between them. Thus, how often those two markers are inherited together in a population after sexual reproduction reflects their linkage.

Genetic linkage maps have been extensively developed in plants or animals. Such maps have proven to be useful for various applications including the localization of genes or quantitative trait loci (QTL) responsible for economically important trait, map based cloning, marker-assisted selection (breeding), genome organization analysis, or the anchoring of whole-genome sequences (FOULONGNE-ORIOL 2012).

AFLP markers can be used to construct high density genetic maps of genomes or genome segments. In most organisms, AFLP has proven to be a very effective way to construct genetic DNA marker maps.

Whole genome sequencing (WGS)

DNA sequencing technologies ideally should be fast, accurate, easy-to-operate, and cheap. In the past thirty years, DNA sequencing technologies and applications have undergone tremendous development and act as the engine of the genome era which is characterized by vast amount of genome data and subsequently broad range of research areas and multiple applications (LIU *et al.* 2012). Here, we describe two sequencing technologies described in this thesis.

Roche 454 was the first commercially successful next generation system. This sequencer uses pyrosequencing technology (LIU *et al.* 2012). Instead of using dideoxynucleotides to terminate the chain amplification, pyrosequencing technology relies on the detection of pyrophosphate released during nucleotide incorporation. The library DNAs with 454-specific adaptors are denatured into single strand and captured by amplification beads followed by emulsion PCR. Then on a picotiter plate, one of dNTP (dATP, dGTP, dCTP, dTTP) will complement to the bases of the template strand with the help of ATP sulfurylase, luciferase, luciferin, DNA polymerase, and adenosine 5’ phosphosulfate (APS) and release pyrophosphate (PPi) which equals the amount of incorporated nucleotide. The ATP transformed from PPi drives the luciferin into oxyluciferin and generates visible light. At the same time, the unmatched bases are degraded by apyrase . Then another dNTP is added into the reaction system and the pyrosequencing reaction is repeated. In 2008 454 GS FLX Titanium system was launched; through upgrading, its read length could reach 700 bp with accuracy 99.9% after filtering. The most outstanding advantage of Roche is its speed: it takes only 10 hours from sequencing start till completion. The read length is also a distinguished character compared with other NGS systems. But the high cost of reagents remains a challenge for Roche 454 (LIU *et al.* 2012).

The Illumina HiSeq system adopts the technology sequencing by synthesis. The library with fixed adaptors is denatured to single strands and grafted to the flowcell, followed by bridge amplification to form clusters which contains clonal DNA fragments. Before sequencing, the library splices into single strands with the help of linearization enzyme (MARDIS 2008), and then four kinds of nucleotides (ddATP, ddGTP, ddCTP, ddTTP) which contain different cleavable fluorescent dye and a removable blocking group would complement the template one base at a time, and the signal could be captured by a (charge-coupled device) CCD. HiSeq systems features the biggest output and the lowest reagent cost compared with other systems (LIU *et al.* 2012).

Bulked segregant analysis (BSA)

This method involves comparing two pooled DNA samples of individuals from a segregating population originating from a single cross. Within each pool, or bulk, the individuals are identical for the trait or gene of interest but are arbitrary for all other genes. Two pools contrasting for a trait (e.g. resistant and susceptible to a particular disease or different MT's) are analyzed to identify markers that distinguish them. Markers that are polymorphic between the pools will be genetically linked to the loci determining the trait used to construct the pools (MICHELMORE *et al.* 1991). In this thesis, BSA is combined with AFLP and WGS to identify AFLP and SNP (single nucleotide polymorphism) markers that are linked to the MT locus.

Aims and thesis outline

The life cycle of diatoms is not totally clarified yet and a lot of bottlenecks still are present in cell culturing. In this thesis, we study the sexual life cycle of diatoms using the pennate heterothallic model diatom *Seminavis robusta* to study the genetic and molecular basis of MT determination in diatoms. To date, nothing is known about the genetic and molecular nature of the MT locus in diatoms (cf. above). The major aim of the present thesis is to identify the MT locus and the MT gene(s) in the model diatom *Seminavis robusta* and to get insight into the genetic structure of the MT locus. The characterization of the MT locus in

S. robusta will then make it possible to gain insight into the evolution of the MT locus gene(s) in diatoms.

CHAPTER 2 describes the construction of mating type specific linkage maps for *S. robusta* by AFLP technology and the identification of mating type linked AFLP markers. The aim of this chapter is to map and identify the genetic locus underlying the mating type differences segregating in a *S . robusta* mapping population.

CHAPTER 3 describes the characterization of the genetic and molecular structure of the MT locus identified in CHAPTER 2, facilitated by a Bulked Segregant Analysis (BSA) approach in combination with AFLP (amplified fragment length polymorphism) and whole genome sequencing (WGS). This chapter provides the first insight into the molecular nature of the MT locus in diatoms.

CHAPTER 4 describes the evolution of MT loci in diatoms. Here, the relationship between the evolutionary diversification of species and the diversification of the MT locus is studied.

In the GENERAL DISCUSSION (CHAPTER 5), the results obtained in the three “result” chapters are further discussed, and suggestions for future research are provided, in addition to a general conclusion is provided.

Literature cited

Archibald, J. M., 2009 The puzzle of plastid evolution. Curr Biol 19**:** R81-88.

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

Armbrust, E. V., J. A. Berges, C. Bowler, B. R. Green, D. Martinez *et al.*, 2004 The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306**:** 79-86.

Becker, E. W., 2007 Micro-algae as a source of protein. Biotechnol Adv 25**:** 207-210.

Bell, G., 1982 *The Masterpiece of Nature: the Evolution and Genetics of Sexuality*. University of California Press.

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

Bloomfield, G., J. Skelton, A. Ivens, Y. Tanaka and R. R. Kay, 2010 Sex determination in the social amoeba Dictyostelium discoideum. Science 330**:** 1533-1536.

Bolker, J. A., and R. A. Raff, 1997 Beyond Worms, Flies and Mice: It's time to widen the scope of developmental biology. Journal of NIH Research 9**:** 35-39.

Bowler, C., A. E. Allen, J. H. Badger, J. Grimwood, K. Jabbari *et al.*, 2008 The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature 456**:** 239-244.

Bozarth, A., U. G. Maier and S. Zauner, 2009 Diatoms in biotechnology: modern tools and applications. Appl Microbiol Biotechnol 82**:** 195-201.

Brown, J. W., and U. Sorhannus, 2010 A molecular genetic timescale for the diversification of autotrophic stramenopiles (Ochrophyta): substantive underestimation of putative fossil ages. PLoS One 5.

Bruland, K. W., E. L. Rue, G. J. Smith and G. R. DiTullio, 2005 Iron, macronutrients and diatom blooms in the Peru upwelling regime: brown and blue waters of Peru. Marine Chemistry 93**:** 81-103.

Capel, B., 2000 The battle of the sexes. Mechanisms of Development 92**:** 89-103.

Casselton, L. A., and N. S. Olesnicky, 1998 Molecular genetics of mating recognition in basidiomycete fungi. Microbiol Mol Biol Rev 62**:** 55-70.

Charlesworth, D., 2013 Plant sex chromosome evolution. Journal of Experimental Botany 64**:** 405-420.

Charlesworth, D., and B. Charlesworth, 2010 Evolutionary Biology: The Origins of Two Sexes. Current Biology 20**:** R519-R521.

Chepurnov, V. A., and D. G. Mann, 2000 Variation in the sexual behaviour of Achnanthes longipes (Bacillariophyta). III. Progeny of crosses between monoecious and unisexual clones. European Journal of Phycology 35**:** 213-223.

Chepurnov, V. A., and D. G. Mann, 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., D. G. Mann, K. Sabbe and W. Vyverman, 2004 Experimental studies on sexual reproduction in diatoms, pp. 91-154 in *International Review of Cytology*, edited by K. W. Jeon. Elsevier Academic Press, San Diego.

Chepurnov, V. A., D. G. Mann, P. von Dassow, E. V. Armbrust, K. Sabbe *et al.*, 2006 Oogamous reproduction, with two-step auxosporulation, in the centric diatom Thalassiosira punctigera (Bacillariophyta). Journal of Phycology 42**:** 845-858.

Chepurnov, V. A., D. G. Mann, P. von Dassow, P. Vanormelingen, J. Gillard *et al.*, 2008 In search of new tractable diatoms for experimental biology. Bioessays 30**:** 692-702.

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

Church, G. M., and W. Gilbert, 1984 Genomic sequencing. Proc Natl Acad Sci U S A 81**:** 1991-1995.

Coelho, S. M., O. Godfroy, A. Arun, G. Le Corguille, A. F. Peters *et al.*, 2011 Genetic regulation of life cycle transitions in the brown alga Ectocarpus. Plant Signal Behav 6**:** 1858-1860.

Connolly, J. A., M. J. Oliver, J. M. Beaulieu, C. A. Knight, L. Tomanek *et al.*, 2008 Correlated evolution of genome size and cell volume in diatoms (Bacillariophyceae). Journal of Phycology 44**:** 124-131.

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

Deusch, O., G. Landan, M. Roettger, N. Gruenheit, K. V. Kowallik *et al.*, 2008 Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Molecular Biology and Evolution 25**:** 748-761.

Drebes, G., 1977 Sexuality, pp. 250-283 in *The biology of diatoms. Botanical monographs,* edited by D. Werner. Blackwell Scientific Publications, Oxford.

Drum, R. W., and R. Gordon, 2003 Star Trek replicators and diatom nanotechnology. Trends Biotechnol 21**:** 325-328.

Dunahay, T. G., E. E. Jarvis and P. G. Roessler, 1995 Genetic transformation of the diatoms *Cyclotella cryptica* and *Navicula saprophila*. Journal of Phycology 31**:** 1004-1012.

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

Falkowski, P. G., R. T. Barber and V. V. Smetacek, 1998 Biogeochemical Controls and Feedbacks on Ocean Primary Production. Science 281**:** 200-207.

Ferrari, S., V. R. Harley, A. Pontiggia, P. N. Goodfellow, R. Lovell-Badge *et al.*, 1992 SRY, like HMG1, recognizes sharp angles in DNA. EMBO J 11**:** 4497-4506.

Ferris, P., B. J. Olson, P. L. De Hoff, S. Douglass, D. Casero *et al.*, 2010 Evolution of an expanded sex-determining locus in Volvox. Science 328**:** 351-354.

Field, C. B., M. J. Behrenfeld, J. T. Randerson and P. Falkowski, 1998 Primary production of the biosphere: Integrating terrestrial and oceanic components. Science 281**:** 237-240.

Figueroa, R. I., K. Rengefors, I. Bravo and S. Bensch, 2010 From homothally to heterothally: Mating preferences and genetic variation within clones of the dinoflagellate Gymnodinium catenatum. Deep-Sea Research Part Ii-Topical Studies in Oceanography 57**:** 190-198.

Foulongne-Oriol, M., 2012 Genetic linkage mapping in fungi: current state, applications, and future trends. Appl Microbiol Biotechnol 95**:** 891-904.

Fraser, J. A., S. Diezmann, R. L. Subaran, A. Allen, K. B. Lengeler *et al.*, 2004 Convergent evolution of chromosomal sex-determining regions in the animal and fungal kingdoms. PLoS Biol 2**:** e384.

Fraser, J. A., and J. Heitman, 2005 Chromosomal sex-determining regions in animals, plants and fungi. Curr Opin Genet Dev 15**:** 645-651.

Fraser, J. A., J. E. Stajich, E. J. Tarcha, G. T. Cole, D. O. Inglis *et al.*, 2007 Evolution of the mating type locus: insights gained from the dimorphic primary fungal pathogens Histoplasma capsulatum, Coccidioides immitis, and Coccidioides posadasii. Eukaryot Cell 6**:** 622-629.

Geitler, L., 1973 Auxospore Formation and Systematics in Pennate Diatoms and Cytology of Cocconeis Races. Osterreichische Botanische Zeitschrift 122**:** 299-321.

Geitler, L., 1979 Some Peculiarities in the Life-History of Pennate Diatoms Hitherto Overlooked. American Journal of Botany 66**:** 91-97.

Gilbert, S. F., 2000 Chromosomal Sex Determination in *Drosophila* in *Devellopmental Biology*. Sinauer Associates, Sunderland.

Gillard, J., J. Frenkel, V. Devos, K. Sabbe, C. Paul *et al.*, 2013 Metabolomics enables the structure elucidation of a diatom sex pheromone. Angew Chem Int Ed Engl 52**:** 854-857.

Gillard, J., J. Frenkel, V. Devos, K. Sabbe, C. Paul *et al.*, 2012 Metabolomics enabled structure elucidation of the first diatom sex pheromone. Angewandte Chemie-International Edition in press.

Graham, J. M., L. E. Graham, S. B. Zulkifly, B. F. Pfleger, S. W. Hoover *et al.*, 2012 Freshwater diatoms as a source of lipids for biofuels. J Ind Microbiol Biotechnol 39**:** 419-428.

Guarini, J. M., L. Chauvaud and J. Coston-Guarini, 2008 Can the intertidal benthic microalgal primary production account for the 'Missing Carbon Sink'? Journal of Oceanography 1**:** 13-19.

Haag, E. S., and A. V. Doty, 2005 Sex determination across evolution: connecting the dots. PLoS Biol 3**:** e21.

Haber, J. E., 1992 Mating-type gene switching in Saccharomyces cerevisiae. Trends Genet 8**:** 446-452.

Hasle, G. R., 1965 *Nitzschia* and *Fragilariopsis* species studied in the light and electron microscopes, pp. 1-49 in *The genus Fragilariopsis*. SKR Norske Videnk-Akad.

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

Judelson, H. S., L. J. Spielman and R. C. Shattock, 1995 Genetic mapping and non-Mendelian segregation of mating type loci in the oomycete, Phytophthora infestans. Genetics 141**:** 503-512.

Katz, L. A., 2012 Origin and diversification of eukaryotes. Annu Rev Microbiol 66**:** 411-427.

Keeling, P. J., and J. D. Palmer, 2008 Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics 9**:** 605-618.

Kikuchi, K., W. Kai, A. Hosokawa, N. Mizuno, H. Suetake *et al.*, 2007 The sex-determining locus in the tiger pufferfish, Takifugu rubripes. Genetics 175**:** 2039-2042.

Kirk, D. L., 2006 Oogamy: inventing the sexes. Curr Biol 16**:** R1028-1030.

Koestler, T., and I. Ebersberger, 2011 Zygomycetes, microsporidia, and the evolutionary ancestry of sex determination. Genome Biol Evol 3**:** 186-194.

Kooistra, W. H., M. De Stefano, D. G. Mann and L. K. Medlin, 2003 The phylogeny of the diatoms. Progress in Molecular and Subcellular Biology 33**:** 59-97.

Kothe, E., 1996 Tetrapolar fungal mating types: sexes by the thousands. FEMS Microbiol Rev 18**:** 65-87.

Kronstad, J. W., and C. Staben, 1997 Mating type in filamentous fungi. Annu Rev Genet 31**:** 245-276.

Kroth, P., 2007 Molecular biology and the biotechnological potential of diatoms. Adv Exp Med Biol 616**:** 23-33.

Lebeau, T., and J. M. Robert, 2003 Diatom cultivation and biotechnologically relevant products. Part I: cultivation at various scales. Appl Microbiol Biotechnol 60**:** 612-623.

Lee, S. C., N. Corradi, S. Doan, F. S. Dietrich, P. J. Keeling *et al.*, 2010a Evolution of the sex-Related Locus and Genomic Features Shared in Microsporidia and Fungi. Plos One 5.

Lee, S. C., M. Ni, W. Li, C. Shertz and J. Heitman, 2010b The evolution of sex: a perspective from the fungal kingdom. Microbiol Mol Biol Rev 74**:** 298-340.

Liew, W. C., R. Bartfai, Z. Lim, R. Sreenivasan, K. R. Siegfried *et al.*, 2012 Polygenic sex determination system in zebrafish. PLoS One 7**:** e34397.

Liu, L., Y. Li, S. Li, N. Hu, Y. He *et al.*, 2012 Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012**:** 251364.

Lommer, M., M. Specht, A. S. Roy, L. Kraemer, R. Andreson *et al.*, 2012 Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome Biol 13**:** R66.

MacDonald, J. D., 1869 On the structure of the diatomaceous frustule and its genetic cycle. The Annals and magazine of natural history; 4**:** 1-8.

Mann, D. G., 1993 Patterns of Sexual Reproduction in Diatoms. Hydrobiologia 269**:** 11-20.

Mann, D. G., 1994 Auxospore Formation, Reproductive Plasticity and Cell Structure in Navicula-Ulvacea and the Resurrection of the Genus Dickieia (Bacillariophyta). European Journal of Phycology 29**:** 141-157.

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

Mann, D. G., V. A. Chepurnov and S. J. M. Droop, 1999 Sexuality, incompatibility, size variation, and preferential polyandry in natural populations and clones of *Sellaphora pupula* (Bacillariophyceae). Journal of Phycology 35**:** 152-170.

Manolakou, P., R. Angelopoulou and G. Lavranos, 2006 Sex determinants in the genome--lessons from the animal kingdom. Coll Antropol 30**:** 649-652.

Mardis, E. R., 2008 The impact of next-generation sequencing technology on genetics. Trends Genet 24**:** 133-141.

Martens, C., K. Vandepoele and Y. Van de Peer, 2008 Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. Proceedings of the National Academy of Sciences of the United States of America 105**:** 3427-3432.

Matsuda, M., Y. Nagahama, A. Shinomiya, T. Sato, C. Matsuda *et al.*, 2002a DMY is a Y-specific DM-domain gene required for male development in the medaka fish. Nature 417**:** 559-563.

Matsuda, M., Y. Nagahama, A. Shinomiya, T. Sato, C. Matsuda *et al.*, 2002b DMY is a Y-specific DM-domain gene required for male development in the medaka fish. Nature 417**:** 559-563.

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

Medlin, L. K., W. H. C. F. Kooistra, R. Gersonde, P. A. Sims and U. Wellbrock, 1997 Is the origin of the diatoms related to the end-Permian mass extinction? Nova Hedwigia 65**:** 1-11.

Medlin, L. K., W. H. C. F. Kooistra and A.-M. M. Schmid, 2000 A review of the evolution of the diatoms – A total approach using molecules, morphology and geology., pp. 13-35 in *The Origin and Early Evolution of the Diatoms: Fossils, Molecular and Biogreographical Approaches*, edited by A. Witkowski and J. Sieminska.

Meiser, A., U. Schmid-Staiger and W. Trosch, 2004 Optimization of eicosapentaenoic acid production by Phaeodactylum tricornutum in the flat panel airlift (FPA) reactor. Journal of Applied Phycology 16**:** 215-225.

Metin, B., K. Findley and J. Heitman, 2010 The mating type locus (MAT) and sexual reproduction of Cryptococcus heveanensis: insights into the evolution of sex and sex-determining chromosomal regions in fungi. PLoS Genet 6**:** e1000961.

Michelmore, R. W., I. Paran and R. V. Kesseli, 1991 Identification of Markers Linked to Disease-Resistance Genes by Bulked Segregant Analysis - a Rapid Method to Detect Markers in Specific Genomic Regions by Using Segregating Populations. Proceedings of the National Academy of Sciences of the United States of America 88**:** 9828-9832.

Moustafa, A., B. Beszteri, U. G. Maier, C. Bowler, K. Valentin *et al.*, 2009 Genomic footprints of a cryptic plastid endosymbiosis in diatoms. Science 324**:** 1724-1726.

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

Nelson, D. M., P. Treguer, M. A. Brzezinski, A. Leynaert and B. Queguiner, 1995 Production and Dissolution of Biogenic Silica in the Ocean - Revised Global Estimates, Comparison with Regional Data and Relationship to Biogenic Sedimentation. Global Biogeochemical Cycles 9**:** 359-372.

Nieuwenhuis, B. P., and D. K. Aanen, 2012 Sexual selection in fungi. J Evol Biol 25**:** 2397-2411.

Nothiger, R., and M. Steinmannzwicky, 1987 Genetics of Sex Determination - What Can We Learn from Drosophila. Development 101**:** 17-24.

Payre, F., S. Noselli, V. Lefrere and A. Vincent, 1990 The Closely Related Drosophila Sry-Beta and Sry-Delta Zinc Finger Proteins Show Differential Embryonic Expression and Distinct Patterns of Binding-Sites on Polytene Chromosomes. Development 110**:** 141-&.

Pfitzer, E., 1871 Untersuchungen über Bau und Entwicklung der Bacillariaceen (Diatomaceen), pp. 1-189 in *Bot. Abhandl.*, edited by Hanstein.

Poulickova, A., and D. G. Mann, 2008 Autogamous auxosporulation in Pinnularia nodosa (Bacillariophyceae). Journal of Phycology 44**:** 350-363.

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

Poulsen, N., and N. Kroger, 2005 A new molecular tool for transgenic diatoms: control of mRNA and protein biosynthesis by an inducible promoter-terminator cassette. FEBS J 272**:** 3413-3423.

Randerson, J. P., and L. D. Hurst, 2001 The uncertain evolution of the sexes. Trends in Ecology & Evolution 16**:** 571-579.

Reyes-Prieto, A., J. D. Hackett, M. B. Soares, M. F. Bonaldo and D. Bhattacharya, 2006 Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. Curr Biol 16**:** 2320-2325.

Rimini, R., A. Pontiggia, F. Spada, S. Ferrari, V. R. Harley *et al.*, 1995 Interaction of normal and mutant SRY proteins with DNA. Philos Trans R Soc Lond B Biol Sci 350**:** 215-220.

Roshchin, A. M., 1994 *Zhiznennye Tsikly Diatomovykh Vodoroslej*. Naukova Dumka, Kiev, Ukraine.

Round, F. E., R. M. Crawford and D. G. Mann, 1990 Diatoms. Biology and Morphology of the Genera. Cambridge University Press.

Saade, A., and C. Bowler, 2009 Molecular Tools for Discovering the Secrets of Diatoms. Bioscience 59**:** 757-765.

Sabbe, K., V. A. Chepurnov, W. Vyverman and D. G. Mann, 2004 Apomixis in Achnanthes (Bacillariophyceae); development of a model system for diatom reproductive biology. European Journal of Phycology 39**:** 327-341.

Sato, S., G. Beakes, M. Idei, T. Nagumo and D. G. Mann, 2011 Novel sex cells and evidence for sex pheromones in diatoms. PLoS One 6**:** e26923.

Scala, S., and C. Bowler, 2001 Molecular insights into the novel aspects of diatom biology. Cellular and Molecular Life Sciences 58**:** 1666-1673.

Schartl, M., 2004 A comparative view on sex determination in medaka. Mech Dev 121**:** 639-645.

Sheehan, J., T. Dunahay, J. Benemann and P. Roessler, 1998 A look back at the U.S. Department of Energy's aquatic species program: biosiesel from algae.

Smetacek, V., 1999 Diatoms and the ocean carbon cycle. Protist 150**:** 25-32.

Smetacek, V., C. Klaas, V. H. Strass, P. Assmy, M. Montresor *et al.*, 2012 Deep carbon export from a Southern Ocean iron-fertilized diatom bloom. Nature 487**:** 313-319.

Smith, J. M., 1978 *The Evolution of Sex*. Cambridge University Press, Cambridge.

Staats, N., L. J. Stal, B. de Winder and L. R. Mur, 2000 Oxygenic photosynthesis as driving process in exopolysaccharide production of benthic diatoms. Marine Ecology Progress Series 193**:** 261-269.

Theriot, E. C., M. Ashworth, R. K. Jansen, T. Nakov, E. Ruck *et al.*, 2009 A Multigene Approach to Inferring the Diatom Phylogeny: Congruence and Conflict. Journal of Phycology 45**:** 8-8.

Trobajo, R., D. G. Mann, V. A. Chepurnov, E. Clavero and E. J. Cox, 2006 Taxonomy, life cycle, and auxosporulation of Nitzschia fonticola (Bacillariophyta). Journal of Phycology 42**:** 1353-1372.

Van den Hoek, C., D. G. Mann and H. M. Johns, 1997 *Algae, An Introduction to Phycology*. Cambridge Usiversity Press.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee *et al.*, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23**:** 4407-4414.

Vuylsteke, M., J. D. Peleman and M. J. van Eijk, 2007 AFLP technology for DNA fingerprinting. Nat Protoc 2**:** 1387-1398.

Xu, J., 2005 Cost of interacting with sexual partners in a facultative sexual microbe. Genetics 171**:** 1597-1604.

Yoon, H. S., J. D. Hackett, C. Ciniglia, G. Pinto and D. Bhattacharya, 2004 A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol 21**:** 809-818.

Zanetti, S., and A. Puoti, 2013 Sex determination in the Caenorhabditis elegans germline. Adv Exp Med Biol 757**:** 41-69.

2

Linkage mapping identifies the mating type determining region as a single locus in the pennate diatom *Seminavis robusta*

Ives Vanstechelman[1,2,3], Koen Sabbe[1], Wim Vyverman[1], Pieter Vanormelingen[1], and Marnik Vuylsteke[2,3,*]

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

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

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

Key words: diatom, AFLP, linkage map, mating type, *Seminavis robusta*

Modified from published publication: PLoS One (2013;8(3):e60132)

Authors' contributions

IV performed the AFLP and linkage mapping experiments, analyzed the data and wrote the manuscript. PV was involved in the culturing experiments and helped improving the manuscript. KS, WV and MV helped to conceive and design the study, and read and approved the manuscript. MV helped interpreting and processing the data.

Abstract

The pennate diatom *Seminavis robusta*, characterized by an archetypical diatom life cycle including a heterothallic mating system, is emerging as a model system for studying the molecular regulation of the diatom cell and life cycle. One of its main advantages compared with other diatom model systems is that sexual crosses can be made routinely, offering unprecedented possibilities for forward genetics. To date, nothing is known about the genetic basis of mating type determination in diatoms. Here, we report on the construction of mating type-specific linkage maps for *S. robusta*, and use them to identify a single locus mating type determination system in this diatom. We identified 13 mating type plus and 15 mating type minus linkage groups obtained from the analysis of 463 AFLP markers segregating in a full-sib mapping population, covering 963.7 and 972.2 cM, respectively. Five linkage group pairs could be identified as putative homologs. The mating type phenotype mapped as a monogenic trait, disclosing the mating type plus as the heterogametic MT. This study provides the first evidence for a genetic mating type determining mechanism in a diatom.

Introduction

Diatoms (Bacillariophyceae) belong to the Stramenopila, which comprise several microalgal groups dominating primary production in aquatic environments (GRANUM *et al.* 2005). The diatoms are one of the most diverse and productive groups of algae, with an estimated 200,000 species responsible for almost 20% of global primary production (MANN 1999). They are also promising from a biotechnological point of view, and hold great potential for the production of high-value bioproducts such as lipids, pigments and biofuels (CHEPURNOV *et al.* 2008). The available genomic resources for diatoms have grown rapidly over the past few years (ARMBRUST *et al.* 2004; BOWLER *et al.* 2008a). In addition, tools for reverse genetics have been developed (DE RISO *et al.* 2009). However, sexual reproduction, a key feature of most diatom life cycles (CHEPURNOV *et al.* 2004), 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 genetics to link phenotype to genotype, including the use of mutagenesis and QTL mapping (TIERNEY and LAMOUR 2005).

Life cycles, including sexual reproduction, have been studied in detail for only a minute fraction of known diatom species, but these represent most principal diatom lineages. The diatom life cycle comprises two main stages: a prolonged vegetative stage lasting months to years, which is diploid, and a short sexual stage lasting hours to days (LEWIS 1984; CHEPURNOV *et al.* 2004; CHEPURNOV *et al.* 2008). During the vegetative stage of the life cycle, a gradual reduction in cell size takes place, caused by physical constraints imposed by their silica cell wall. This cell wall comprises two parts (or thecae), one of which (the epitheca) is slightly larger than, and overlaps, the other (the hypotheca). During mitosis, each daughter cell inherits one maternal valve (which becomes the new epitheca) and *de novo* synthesizes a smaller hypotheca. Below a species-specific size threshold (the sexual size threshold, SST) cells become capable of sexual reproduction. Restoration of the maximal cell size generally occurs through sexual reproduction. The zygote matures into an auxospore, which expands to on average two to three times the size of the parental cells. After reaching its maximum size, a new so-called initial cell is formed inside the auxospore envelope, initiating a new round of vegetative multiplication. Because of its crucial role in cell size restoration, sexual reproduction is an obligatory stage in the life cycle of most diatoms.

The raphid pennate diatom *S. robusta* has recently been advocated as a model organism to study diatom biology, and in particular life cycle regulation (CHEPURNOV *et al.* 2008). Like most pennate diatoms (ROSHCHIN 1994; MANN *et al.* 1999; CHEPURNOV and MANN 2004; CHEPURNOV *et al.* 2004), *S. robusta* has a heterothallic mating system with two mating types (MT$^+$ and MT$^-$) (CHEPURNOV *et al.* 2002; GILLARD *et al.* 2012). After cell pairing, each of the two cells forms two morphologically and behaviorally identical gametes. Subsequent zygote formation and auxospore expansion finally result in two initial cells with a cell length of 64-73 µm. Sexual reproduction is easily induced in cultured strains with a cell size below the SST by adding a suitable mating partner which allows sex to be reliably controlled in mating experiments. However, as in some other pennate diatoms (e.g. *T. tabulata* and *F. delicatissima*) (ROSHCHIN 1994; DAVIDOVICH *et al.* 2010), intraclonal reproduction (homothally) can sporadically occur (in *S. robusta* in MT$^+$). Additional advantages of *S. robusta* are that the cells are reasonably large (SST ~50 µm apical cell length) and can grow on surfaces allowing easy non-intrusive monitoring of cell and life cycle stages using inverted microscopy.

To date, genetic maps exist for only a small number of species outside the opisthokont and Plantae lineages. Within the stramenopile lineage, linkage maps have been reported for the brown alga *Ectocarpus* siliculosus (HEESCH *et al.* 2010) and several oomycete species (VAN DER LEE *et al.* 1997; MAY *et al.* 2002; SICARD *et al.* 2003; VAN DER LEE *et al.* 2004), while a preliminary linkage map has been published for the kelp *Laminaria japonica* (YANG *et al.* 2009).

Experimental evidence suggests that mating type (MT) determination in heterothallic pennate diatoms is genetic (CHEPURNOV *et al.* 2004; DAVIDOVICH *et al.* 2010), but how MT determination is achieved is unknown. Here, we report on the construction of a MT-specific linkage map for *S. robusta* based on AFLP markers (VOS *et al.* 1995; VUYLSTEKE *et al.* 2007) and its use to identify the MT determining region in this diatom. We present AFLP markers co-segregating with the MT phenotype, the genetic structure of the MT locus and the identification of MT$^+$ as the heterogametic sex in *S. robusta*.

Materials and methods

Production of a F_1 mapping population

The experimental strains were selected from a collection of cryopreserved strains of *S. robusta* publicly available in the BCCM/DCG diatom collection (http://bccm.belspo.be/about/dcg.php). A full-sib (FS) family, containing 116 individual F_1 progeny, was produced from a cross between strains H73A (DCG 0123) and 96A (DCG 0128), having the MT^- and MT^+ mating type, respectively. Their average cell size at the time of crossing was 37.1 ± 0.4 µm and 26.3 ± 0.3 µm respectively. This difference in cell size between the parental strains allowed to distinguish hetero- from homothallic F_1 auxospores and initial cells after mating. Cultures were inoculated for dark synchronization (GILLARD *et al.* 2008) by growing them for 3 days at 18 ºC in a 12:12 h light:dark regime with cool-white fluorescent lamps at approximately 80 µmol photons m^{-2} s^{-1}. At the end of day 3, the dark period was extended for another 18 h, after which cells were mixed by transferring a quarter of the suspended cells of both MTs to a new flask which was then replenished with an equal amount of fresh Guillard's F/2 medium. After another 9 h of darkness, light was switched on again, allowing the mixed cultures to progress synchronously through the different sexual stages. Two days later, 200 F_1 auxospores and initial cells resulting from heterothallic reproduction (derived from two parental gametangia of unequal size) were isolated (by dislodging them from the bottom with a needle and picking them up with a 20 µl pipet) and transferred to 96-well culture plates containing 0.2 ml F/2 medium.

Phenotyping: determining the mating type of the F_1 progeny

After four months of weekly re-inoculating the F_1 cultures in 24-well culture plates (2 ml), cells reached the SST (~50 µm). The MT phenotype of each F_1 progeny was then determined by backcrossing to the parental strains as well as to two other strains (85A and 85B) of known MT. The MT of 116 F_1 progeny could be determined unambiguously, yielding a mapping population of 57 MT^+ and 59 MT^- strains. These strains have been deposited and cryopreserved in the BCCM/DCG culture collection.

DNA extraction of *S. robusta* cultures

For DNA extraction, cultures were grown in 100 ml culture flasks and harvested in exponential phase. Most culture medium was removed and cells were scraped from the surface (cell scrapers; MLS) in 10 ml of remaining culture medium and centrifuged in 15 ml falcon tubes for 15 minutes at 1000 g (4K15(SIGMA)). After removal of the supernatant, the cells were transferred to a 2 ml Eppendorf tube and centrifuged (15 minutes; 1000 g). Eppendorf tubes with cell pellets were frozen at -20° C.

For DNA extraction, 0.5 g Zirconia/Silica Beads 0.1 mm diameter (BioSpec Products), 0.5 ml TE buffer (10 mM TrisHCl (pH 7.6), 1mM EDTA (pH 8.0)) and 0.5 ml buffered phenol were added to each sample. The samples were bead-beaten (3×; 85 s; 30 Hz). Each time, the samples were cooled on ice. The tubes were centrifuged (Centrifuge 5415R(EPPENDORF); 4° C; 5 m; 10,000 rpm). The water phase containing the DNA was then transferred to a new 2 ml Eppendorf tube and 0.5 ml PCI (phenol:chlorophorm:isoamyl alcohol = 25:24:1v/v) was added. The mixture was centrifuged (10,000 rpm; 5 m; 4° C). The water phase was transferred to a new Eppendorf, 0.5 ml PCI was added again and the mixture was centrifuged. The water phase was transferred to a 1.5 ml Eppendorf. 50 µl 3 M NaAc (pH = ± 5), 1 ml 96% ethanol (-20° C) and 2 µl glycogen (-20° C) were added. The samples were incubated over night at -20° C. The next day, the Eppendorfs were centrifuged (13,000 rpm; 30 m; 4° C). The liquid phase was removed and the DNA pellet became visible. 1 ml 70% ethanol was added and the sample was centrifuged (13,000 rpm; 5 m; 4° C). Ethanol was removed and the sample was again centrifuged (13,000 rpm; 5 m; 4° C). The remaining ethanol was removed with a 200 µl pipet and the pellet was dried for at least 20 minutes. Preheated (55° C; 50 µl) TE buffer (pH 8.0) was added to the pellets and the samples were incubated (20 m; 55° C). DNA of the mapping population was stored at -20° C.

Segregation and linkage analysis

The 116 offspring were analyzed with 54 *Eco*RI+2/*Mse*I+3 AFLP primer combinations (PC's; Table S1) as described in (VUYLSTEKE *et al.* 2007). 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 were scored on the basis of relative fragment intensities, using the image analysis software AFLPQuantar*Pro* (http://www.keygene-products.com).

AFLP markers heterozygous in MT$^+$ parent (96A) and homozygous in the MT$^-$, and expected to segregate 1:1 in the F_1 generation, were termed MT$^+$ specific markers. Markers heterozygous in the MT$^-$ parent (H73A) and homozygous in the MT$^+$ were termed MT$^-$ specific markers. Markers heterozygous in both parents, and expected to segregate 1:2:1 in the F_1 generation were termed "biparental markers". Whenever feasible, biparental markers were scored co-dominantly (i.e., following a 1:2:1 segregation pattern). They were scored dominantly (i.e., conforming to a 3:1 segregation pattern) when heterozygosity could not reliably be discriminated from homozygosity for the present AFLP marker allele. Each AFLP marker was identified by a specific code referring to the corresponding PC and the estimated molecular size of the fragment in nucleotides as estimated by AFLP-Quantar*Pro*. Linkage analysis and segregation distortion tests were performed using the software package JoinMap 4.0 (VAN OOIJEN 2006). The appropriate mapping population type was set to option CP, a population resulting from a cross between two heterogeneously heterozygous and homozygous diploid parents, linkage phases originally unknown. Because for population type CP, the segregation type (SEG) might vary across the loci, a code indicating the segregation type has to be given. The SEG was set to <nn×np> for the MT$^+$ markers, <lm×ll> for the MT$^-$ markers and <hk×hk> for the biparental markers. The two characters left and right of the ‘‘×’’ in these codes correspond to the AFLP marker alleles of the first and second parent, respectively. Each different AFLP marker allele is represented by a different character. We first ran through a fairly wide range of logarithm of the odds (LOD) thresholds, from 2.0 to 14.0, to obtain a proper view of what might be the best grouping. In general, we decided to use the grouping obtained with a LOD score of 6.0. In a few cases, the grouping obtained at a LOD score of 8.0 and 14.0 was used. Only linkage groups containing at least three markers were considered for map construction. Maps were constructed in three rounds, each producing a linkage map. In this map-building procedure, each map was calculated by using the pairwise data of loci present on the map, with default settings (recombination frequency (REC) < 0.4; LOD threshold > 1). Once the well-fitting markers (causing a change in goodness of fit smaller than the threshold = 5) were positioned on the map (after two rounds), the remaining markers were forced onto the map by setting the jump threshold to zero. When the markers in the third map caused a jump in goodness of fit larger than an arbitrary threshold of 10, the second map was selected as the final map, otherwise the third map. Single markers with a segregation ratio in discordance with the flanking markers (i.e., markers showing strong segregation distortion flanked by a number of non-distorted markers) were discarded, and the map construction was repeated. A marker order was not forced on any linkage group during map construction. Recombination frequencies were converted to Kosambi centiMorgans (cM) prior to the map estimation. Linkage groups were constructed using the MT^+ and MT^- markers. Biparental markers were included in both the MT^+ and MT^- linkage maps. This way, homologous linkage groups were identified on the presence of identical biparental markers. Editing the linkage groups was done with the software MapChart (VOORRIPS 2002).

Mapping the MT locus

The mapping of the MT phenotype was done in two ways: 1) by including the phenotype as a single marker segregating as a MT^+ or MT^- specific marker, indicated as SEX1 and SEX2, respectively, and 2) by QTL analysis using mixed models as implemented in the QTL menu in GenStat 14 (VSN INTERNATIONAL 2011). Because too few homologous linkage groups are identified, QTL analysis was done for the two mating type specific linkage maps separately. In a preliminary search for QTL, we tested the association of individual marker loci with mating type every 5 cM along the genome, using the commonly known simple interval mapping (SIM) procedure. In a second step, we tested for QTL at particular positions after correcting for QTL elsewhere in the genome, as were identified in the preliminary analysis. This procedure is commonly known as composite interval mapping (CIM). The genome-wide type I error rate was set to $\alpha = 0.05$. The *P* values calculated assume normally distributed errors, when a binomial distribution is more appropriate in the case of mating type. A previous study (COCKRAM *et al.* 2010) has shown that applying the mixed model to binary traits is robust and do not result in an excess of low *P* values (i.e. false discoveries) as long as the minor allele frequency of the response variable and the markers is not too low.

Results

Segregation analysis and linkage mapping

A total of 54 *EcoRI*+2/*MseI*+3 AFLP PCs, generating on average 8.6 markers, resulted in a total of 463 AFLP fragments segregating in the 116 F_1 progeny. In total, 162 MT^+ and 221 MT^- markers were used for the construction of 13 MT^+ and 15 MT^- -specific linkage groups, covering 963.7 cM and 972.2 cM respectively (Figure 1 and 2). Of those marker loci, 28% displayed significantly distorted segregation ratios at the $\alpha = 0.05$ significance level. As segregation distortion is a normal phenomenon in wide crosses, these markers were not a priori excluded, but evaluated after map construction. Although some larger genomic regions did not reveal any markers (e.g. 32.3 cM in the MT^{+}_8 linkage group and 34.2 cM in linkage group MT^{-}_9), the median inter-marker distances were relatively low (4.1 and 2.3 cM for the MT^{+} and the MT^{-} linkage maps respectively) (Table 1).

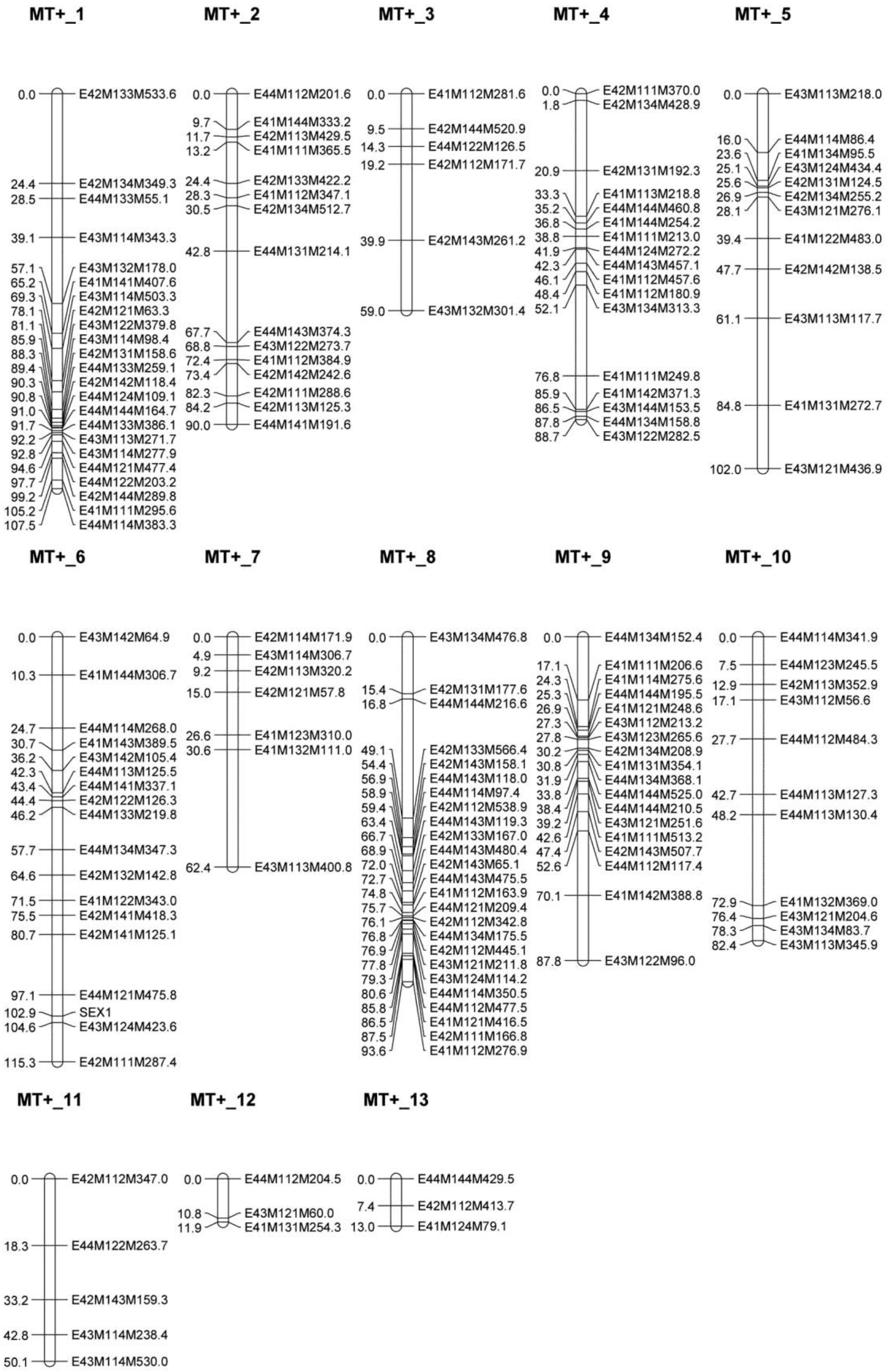


Figure 1. MT$^+$ linkage groups of *S. robusta* containing markers originating from parental strain 96A.

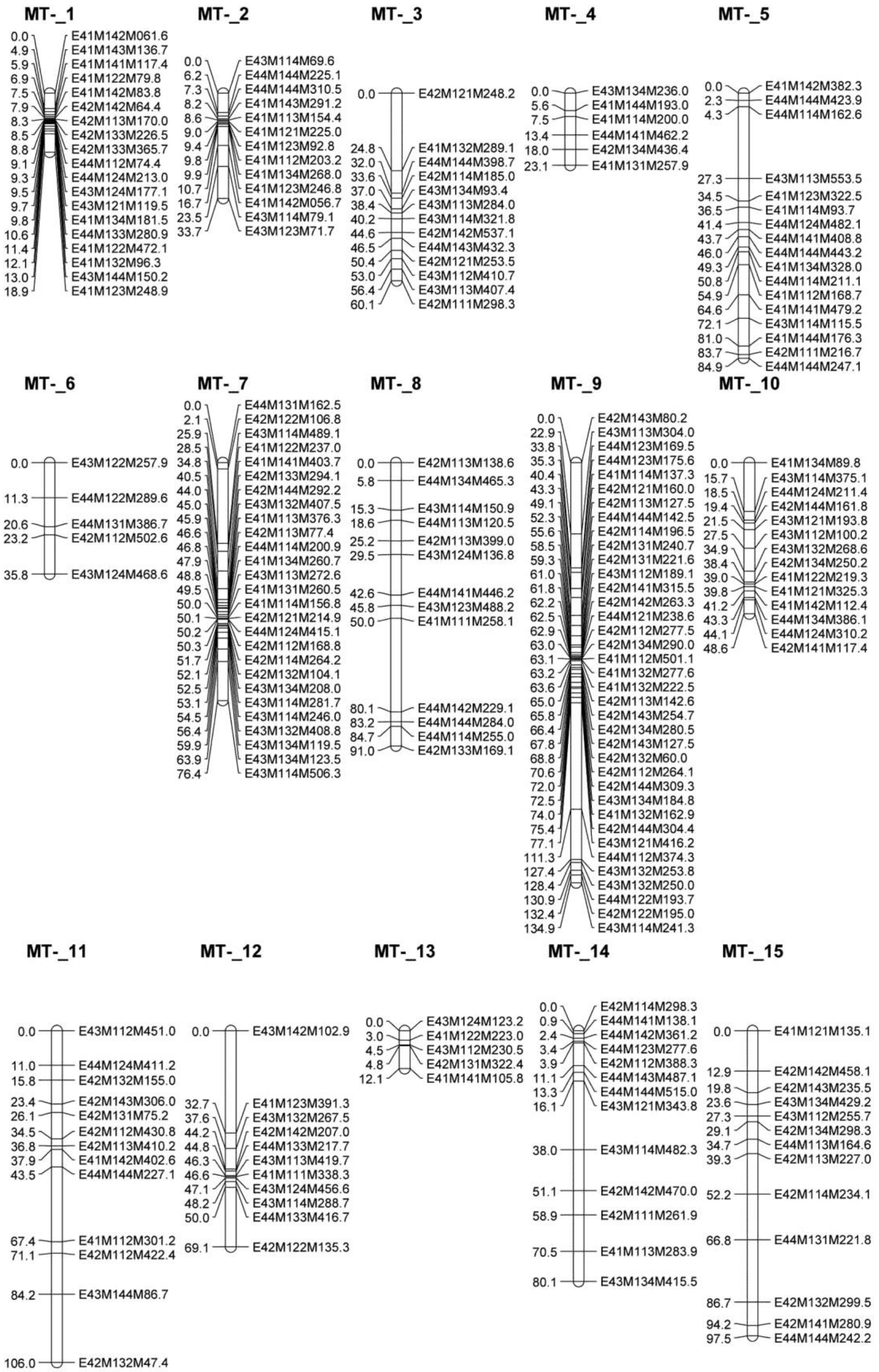


Figure 2. MT$^-$ linkage groups of *S. robusta* containing markers originating from parental strain H73A.

Table 1: Statistics for the integrated MT$^+$ and MT$^-$ linkage maps of *S. robusta*.

	MT$^+$	MT$^-$
No. of linkage groups	13	15
No. of markers per linkage group		
Min	3	5
Max	25	37
Median	12	13
Mean	12.5	14.7
Total	163	221
Size of linkage groups (cM)		
Min	11.9	12.1
Max	115.3	134.9
Median	87.8	69.1
Mean	74.1	64.8
Total	963.7	972.2
Intermarker distance (cM)		
Min	0.1	0
Max	32.3	34.2
Median	4.1	2.3
Mean	6.4	4.7

Sixty-four biparental markers were mapped to the MT$^+$ and MT$^-$ map separately and 20 of those markers showed cosegregation with MT$^+$ and MT$^-$ specific linkage groups. Five homologous maps were identified based on the presence of one or more identical biparental markers (Figure 3). No significant difference was observed between the intervals of the 20 biparental markers in the MT$^+$ and MT$^-$ linkage groups (paired *t*-test, $t = 0.79$; $P = 0.22$), suggesting that recombination frequencies do not differ much between both MTs in *S. robusta*.

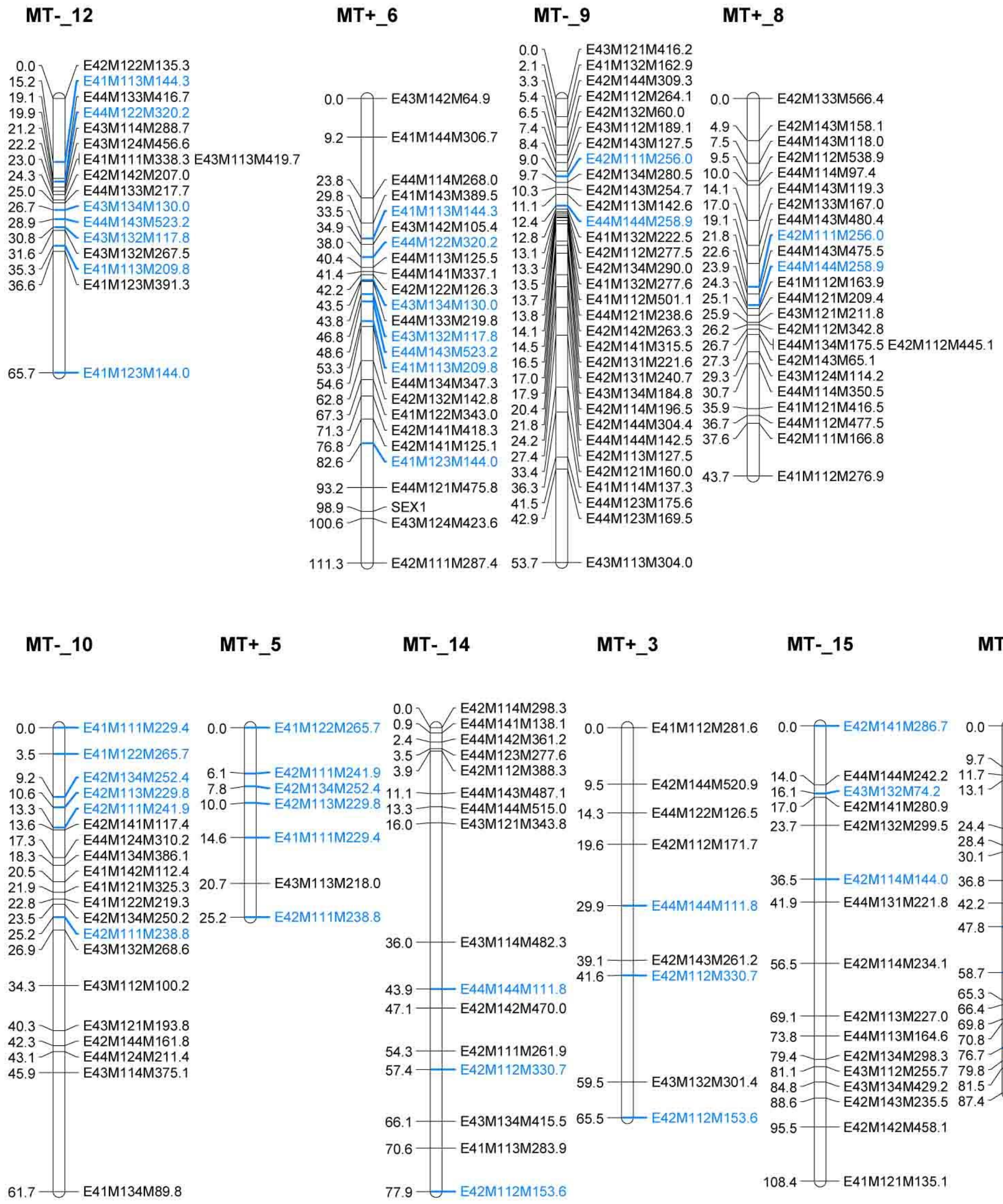


Figure 3. Homologous linkage groups of *S. robusta*. Common biparental markers between the MT^{+} and the MT^{-} linkage groups are shown in blue.

Mapping the MT locus

The mating type phenotype was incorporated as a single marker segregating as either a MT^{+} (SEX1) or MT^{-} specific marker (SEX2). SEX1 could be assigned to the MT^{+}_6 linkage group, including 18 markers and spanning 115.3 cM. SEX1 is flanked by E43M124M423.6 at 1.7 cM and E44M121M475.8 at 5.8 cM distance. In contrast, SEX2 did not cosegregate with any of the markers of the MT^{-} linkage groups (highest LOD = 2).

A QTL analysis of the MT phenotype confirmed the monogenic nature of MT and the identification of MT^+ as the heterogametic MT in *S. robusta* (Figure 4). The genome-wide significance threshold (P = 0.05) for detection of QTL co-segregating with MT was calculated as $-\log_{10}(P)$ = 3.21. The MT phenotype mapped significantly ($-\log_{10}(P)$ = 198.29; E43M124M423.6) to a single locus located on the MT^+_6 linkage group (Figure 4a). In contrast, no significant association was identified between the MT phenotype and MT^- specific markers (Figure 4b).

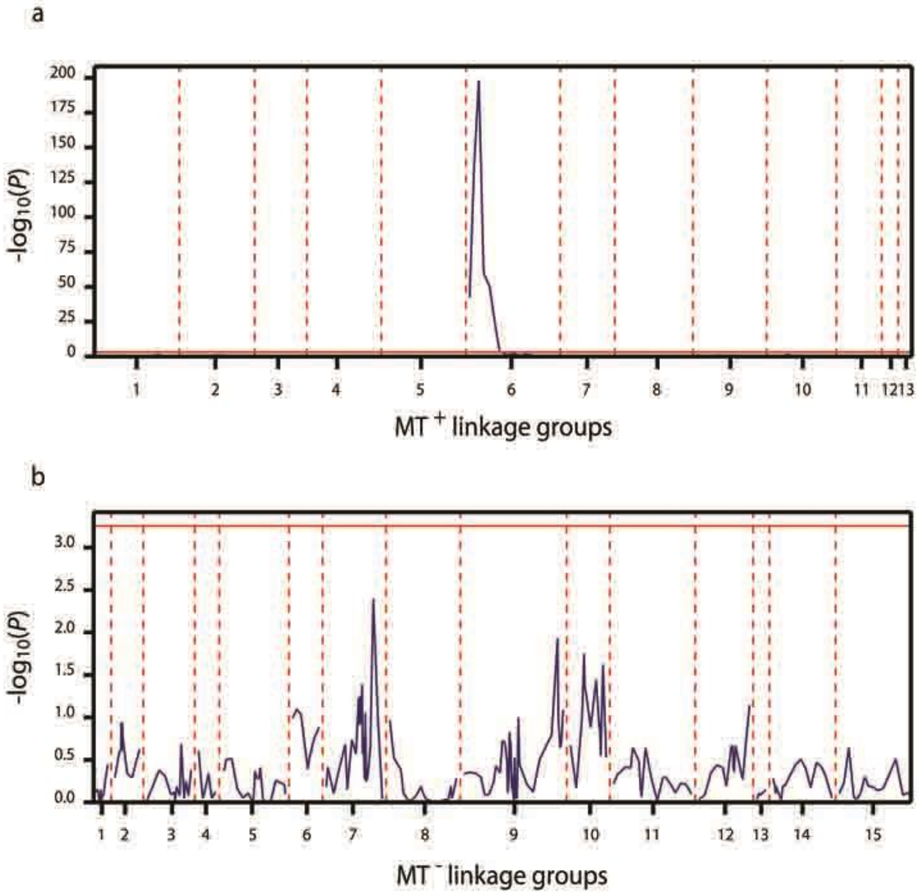


Figure 4. QTL mapping of the MT phenotype in 116 *S. robusta* F_1 progeny. QTL analysis was done for the two MT specific linkage maps separately. Linkage scores (plotted as $-\log_{10}(P)$) for MT^+ (a) and MT^- (b) markers are shown according to genome position. The linkage analysis indicates that a single locus on the MT^+_6 linkage group determines the MT in *S. robusta*.

The two homologous linkage groups MT^-_12 and MT^+_6 (Figure 3) contain seven biparental markers widely spread across the two linkage groups. These identify a relatively large region of recombination between the two MT determining homologous chromosomes.

Discussion

In this study we exploit the high multiplex ratio of AFLP technology to construct the first linkage maps for a diatom species. We applied these maps to demonstrate that MT determination in the heterothallic pennate species *S. robusta* is genetic, and identify the MT determining region as a single locus.

Segregation and linkage analysis of the 463 AFLP markers scored for 116 individuals of an F_1 mapping population resulted in 13 MT^+ and 15 MT^- specific linkage groups. The use of biparental markers, segregating in a 1:2:1 mode and scored co-dominantly, allowed the detection of five putative homologous linkage groups, including those carrying the MT locus.

The MT phenotype cosegregates with markers of a MT^+ specific linkage group, identifying MT^+ as the heterogametic MT and MT^- as the homogametic MT.

Unlike ‘classical’ sex chromosomes (like the XY chromosomes of mammals and WZ chromosomes of birds), which have only a small recombining pseudo-autosomal region (PAR) or do not recombine at all (e.g. the *Drosophila* Y chromosome), the homologous MT determining linkage groups in *S. robusta* appear to have a relatively large region of recombination. It is therefore likely that its X and Y chromosomes are cytologically indistinguishable (or non-heteromorphic, cf. BERGERO and CHARLESWORTH (2009)). Sex chromosomes with extensive recombining PAR regions (in which both homologous chromosomes carry the same gene content) and small non-recombining regions, have been hypothesized to represent recently evolved sex chromosome systems. However, few estimates are available for the age of non-recombining regions (e.g. < 2 MY ago in the papaya plant *Carica papaya*), and some data suggest that such regions may also be much older (e.g. in some bird and snake species) (BERGERO and CHARLESWORTH 2009). Studies on divergence times in the non-recombining regions of the MT determining chromosomes in *S. robusta* and other diatoms can contribute to our understanding of the evolution of sex chromosomes, as the diatoms, having an extensive fossil record, have a relatively well time-calibrated evolutionary record. Heterothally is to date known only in the pennate diatoms, which in the fossil record appeared in the Late Cretaceous (about 75-90 MY ago) (SIMS *et al.* 2006; BOWLER *et al.* 2008b). All studied centric diatoms, from whom the pennate diatoms evolved, appear to be homothallic (i.e. they have no MT differentiation). In some cases, clones are predominantly male or female, but homothally is still present (CHEPURNOV *et al.* 2004). While more studies on mating systems in centric diatoms are still needed to confirm that homothally is the rule in centric diatoms, all available evidence so far suggests that in diatoms the evolution of heterogametic MT determination associated with MT differentiation is a relatively recent event, coinciding with the transition from homothally to heterothally.

The first diatom linkage maps presented in this study will constitute an important resource for future genetic analyses in *S. robusta*. Linkage maps provide important insights into genome organization and can be used for genetic studies of traits of interest (FOULONGNE-ORIOL 2012). A particular advantage is that each individual progeny of the F_1 mapping population can be clonally propagated and they are maintained as cryopreserved strains. A panel of immortalized F_1 individuals has a number of advantages for genetic mapping identical to those of Recombinant Inbred Lines (RILs), often used in plant or rodent genetics: one needs to genotype each progeny only once and can phenotype multiple individuals from each clonal culture to reduce individual, environmental and measurement variability for multiple traits.

The linkage maps will also provide important information about the *S. robusta* genome sequence which is under construction (A. Bones & T. Brembu, pers. comm.). The completion of the genome sequence will also be the opportunity to further progress on the linkage mapping, as it will represent a source of single nucleotide polymorphisms (SNPs) and insertion/deletion (INDEL) markers for mapping. This will provide a framework to solve the position and order of scaffolds during assembly (FOULONGNE-ORIOL 2012). This information can be used to construct pseudo-chromosomes by concatenating adjacent supercontigs, and to carry out broad analyses of genome composition.

Despite the rapidly growing amount of diatom genomic information (LOMMER *et al.* 2012), almost nothing is known about the regulation of the unique diatom life cycle. This is changing fast with the introduction of new model diatoms in which the life cycle and sexual reproduction can be reliably manipulated experimentally, including *S. robusta* (CHEPURNOV *et al.* 2008; GILLARD *et al.* 2012). Further characterization of the mating type locus will prove crucial for our understanding of regulation of the diatom life cycle. The identification of the *S. robusta* genomic region carrying the MT locus in this study provides a starting point for further investigation of the locus and the identification of the gene(s) and sequence polymorphism(s) underlying MT dimorphism in *S. robusta*. In turn, this will pave the road for understanding the mechanisms underlying mating system switches to alternative reproductive modes (homothally, paedogamy, and apomixis), which are regularly observed among closely related pennate diatoms (ROSHCHIN 1994; VANORMELINGEN *et al.* 2008; DAVIDOVICH *et al.* 2010), and more broadly, the evolution of the MT locus in diatoms following the evolution of the pennate lineage from a homothallic centric ancestor.

Acknowledgements

Pieter Vanormelingen is a postdoctoral research fellow with the Research Foundation – Flanders (FWO). The authors acknowledge Belspo for funding the Belgian Co-ordinated Culture Collections of Micro-organisms (BCCM) and Ghent University Research Funding (BOFGOA 01G01911 and BOF BOF09/24J/103).

Literature cited

Armbrust, E. V., J. A. Berges, C. Bowler, B. R. Green, D. Martinez *et al.*, 2004 The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306**:** 79-86.

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

Bowler, C., A. E. Allen, J. H. Badger, J. Grimwood, K. Jabbari *et al.*, 2008a The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature 456**:** 239-244.

Bowler, C., A. E. Allen, J. H. Badger, J. Grimwood, K. Jabbari *et al.*, 2008b The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 456**:** 239-244.

Chepurnov, V. A., and D. G. Mann, 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., D. G. Mann, K. Sabbe and W. Vyverman, 2004 Experimental studies on sexual reproduction in diatoms, pp. 91-154 in *International Review of Cytology*, edited by K. W. Jeon. Elsevier Academic Press, San Diego.

Chepurnov, V. A., D. G. Mann, P. von Dassow, P. Vanormelingen, J. Gillard *et al.*, 2008 In search of new tractable diatoms for experimental biology. Bioessays 30**:** 692-702.

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

Cockram, J., J. White, D. L. Zuluaga, D. Smith, J. Comadran *et al.*, 2010 Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Proc Natl Acad Sci U S A 107**:** 21611-21616.

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

De Riso, V., R. Raniello, F. Maumus, A. Rogato, C. Bowler *et al.*, 2009 Gene silencing in the marine diatom *Phaeodactylum tricornutum*. Nucleic Acids Research 37**:** e96.

Foulongne-Oriol, M., 2012 Genetic linkage mapping in fungi: current state, applications, and future trends. Applied Microbiology and Biotechnology 95**:** 891-904.

Gillard, J., V. Devos, M. J. J. Huysman, L. De Veylder, S. D'Hondt *et al.*, 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., J. Frenkel, V. Devos, K. Sabbe, C. Paul *et al.*, 2012 Metabolomics enabled structure elucidation of the first diatom sex pheromone. Angewandte Chemie-International Edition in press.

Granum, E., J. A. Raven and R. C. Leegood, 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.

Heesch, S., G. Y. Cho, A. F. Peters, G. Le Corguillé, C. Falentin *et al.*, 2010 A sequence-tagged genetic map for the brown alga Ectocarpus siliculosus provides large-scale assembly of the genome sequence. New Phytologist 188**:** 42-51.

Lewis, W. M., Jr., 1984 The diatom sex clock and its evolutionary significance. American Naturalist 123**:** 73-80.

Lommer, M., M. Specht, A. S. Roy, L. Kraemer, R. Andreson *et al.*, 2012 Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome Biol 13**:** R66.

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

Mann, D. G., V. A. Chepurnov and S. J. M. Droop, 1999 Sexuality, incompatibility, size variation, and preferential polyandry in natural populations and clones of *Sellaphora pupula* (Bacillariophyceae). Journal of Phycology 35**:** 152-170.

May, K. J., S. C. Whisson, R. S. Zwart, I. R. Searle, J. A. G. Irwin *et al.*, 2002 Inheritance and mapping of 11 avirulence genes in *Phytophthora sojae*. Fungal Genetics and Biology 37**:** 1-12.

Roshchin, A. M., 1994 *Zhiznennye Tsikly Diatomovykh Vodoroslej*. Naukova Dumka, Kiev, Ukraine.

Sicard, D., E. Legg, S. Brown, N. K. Babu, O. Ochoa *et al.*, 2003 A genetic map of the lettuce downy mildew pathogen, *Bremia lactucae*, constructed from molecular markers and avirulence genes. Fungal Genetics and Biology 39**:** 16-30.

Sims, P. A., D. G. Mann and L. K. Medlin, 2006 Evolution of the diatoms: insights from fossil, biological and molecular data. Phycologia 45**:** 361-402.

Tierney, M. B., and K. H. Lamour, 2005 An introduction to reverse genetic tools for investigating gene function. The Plant Health Instructor.

Van der Lee, T., I. De Witte, A. Drenth, C. Alfonso and F. Govers, 1997 AFLP linkage map of the oomycete *Phytophthora infestans*. Fungal Genetics and Biology 21**:** 278-291.

van der Lee, T., A. Testa, A. Robold, J. van 't Klooster and F. Govers, 2004 High-density genetic linkage maps of *Phytophthora infestans* reveal trisomic progeny and chromosomal rearrangements. Genetics 167**:** 1643-1661.

Van Ooijen, J. W., 2006 *JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations*. Kyazma B.V., Wageningen.

Vanormelingen, P., V. A. Chepurnov, D. G. Mann, K. Sabbe and W. Vyverman, 2008 Genetic divergence and reproductive barriers among morphologically heterogeneous sympatric clones of Eunotia bilunaris sensu lato (Bacillariophyta). Protist 159**:** 73-90.

Voorrips, R. E., 2002 MapChart: software for the graphical presentation of linkage maps and QTLs. Journal of Heredity 93**:** 77-78.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee *et al.*, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research 23**:** 4407-4414.

VSN International, 2011 GenStat *for Windows* 14th Edition, pp. VSN International, Hemel Hempstead, UK.

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

Yang, G., Y. Sun, Y. Shi, L. Zhang, S. Guo *et al.*, 2009 Construction and characterization of a tentative amplified fragment length polymorphism-simple sequence repeat linkage map of *Laminaria* (Laminariales, Phaeophyta). Journal of Phycology 45**:** 873-878.

Supplementary data

Table S1: list of primer combinations used for AFLP analysis; E: *Eco*RI primer with two selective bases; M: *Mse*I primer with three selective bases, selective bases: 1,2,3,4 corresponds to A,C,G,T

E41M111	E42M111	E43M112	E44M112
E41M112	E42M112	E43M113	E44M113
E41M113	E42M113	E43M114	E44M114
E41M114	E42M114	E43M121	E44M121
E41M121	E42M121	E43M122	E44M122
E41M122	E42M122	E43M123	E44M123
E41M123	E42M131	E43M124	E44M124
E41M124	E42M132	E43M132	E44M131
E41M131	E42M133	E43M134	E44M133
E41M132	E42M134	E43M142	E44M134
E41M134	E42M141	E43M144	E44M141
E41M141	E42M142		E44M142
E41M142	E42M143		E44M143
E41M143	E42M144		E44M144
E41M144			

Table S2. AFLP markers underlying the single QTL detected for the MT phenotype

Marker	Linkage group	Position (cM)	-log10(*P*)
E42M111M287.4	MT^+_6	0	42.41
E43M124M423.6	MT^+_6	12.2	198.29
E44M121M475.8	MT^+_6	18.7	60.11
E42M141M125.1	MT^+_6	34.5	21.05
E42M141M418.3	MT^+_6	39.8	3.7

3

Revealing the genetic structure of the mating type determining region in the pennate diatom *Seminavis robusta*

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

* 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

Publication in preparation

Keywords: diatom, *Seminavis robusta,* genome assembly, mating type determination, mating type locus

Authors' contributions

IV performed the experiments, analyzed the data and wrote the manuscript. MJJH an SM were involved in the design of the experiments. TB, PW and AB are involved in the production of sequencing data. KS, WV and MV helped to conceive and design the study, and read and approved the manuscript. MV helped interpreting and processing the data.

Abstract

Using linkage mapping we recently showed that mating type in the pennate heterothallic diatom *Seminavis robusta* is determined by a single locus, with the mating type plus as the heterogametic mating type. In the present study, we used bulked segregant analysis to identify additional genetic markers that are strongly associated with the mating type locus. Two markers enclosing the mating type phenotype mapped on the same scaffold in the *S. robusta* genome assembly. This identified a locus on a scaffold of total length of 24698 bp, containing three protein-coding genes. Differential expression analysis further supports the identification of the SF2-family related helicase/S-adenosylmethionine dependent methyltransferase gene as most likely responsible for mating type determination in *S. robusta*. This gene is the first evidence of a MT determinant in diatoms and in the Stramenopila lineage as a whole.

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 lipids, pigments and biofuels (LEBEAU and ROBERT 2003; BOZARTH *et al.* 2009). In recent years, genomic resources for diatoms are rapidly growing (ARMBRUST *et al.* 2004; BOWLER *et al.* 2008; DE RISO *et al.* 2009; LOMMER *et al.* 2012; HUYSMAN *et al.* 2013). The life cycle of diatoms is diplontic and is accompanied by a typical cell size reduction cycle during vegetative growth (CHEPURNOV *et al.* 2004). Sexual reproduction is a crucial step in the diatom life cycle (CHEPURNOV *et al.* 2004) as it prevents clonal cell death resulting from this gradual size diminution. Sexual reproduction however is only possible once a species-specific cell size threshold (SST) is reached. Cell size restitution is established during sexual reproduction by the expansion of a specialized cell, the auxospore (CHEPURNOV *et al.* 2002). Sexual reproduction has never been demonstrated for the currently 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 (ALONSO and ECKER 2006), and thus also the elucidation of the molecular basis of mating type (MT) determination. The raphid pennate diatom *Seminavis robusta* has recently been put forward as a model organism for studies of the diatom cell and life cycle, as it displays a typical diatom life cycle and sexual reproduction can reliably be controlled (CHEPURNOV *et al.* 2008).

Sexual development is common in eukaryotic organisms. Genetic sex determination is set by a sex chromosome or an autosomal gene, while environmental sex determination is set by temperature, local sex ratio or population density. Though little is known about the molecular mechanisms of 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 homolog chromosomes is smaller (genetic degeneration), or homomorphic (no degeneration in one of 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). In recently evolved sex chromosomes on the other hand, non-recombining regions can be as small as a few megabases and large PAR regions exist (BERGERO and CHARLESWORTH 2009). Separate sexes appear to have arisen independently many times during the evolution of major eukaryotic lineages, witness the differences in sex determination mechanisms in these lineages (BERGERO and CHARLESWORTH 2009). Indeed, sex determination mechanisms appear to be very little conserved. Marked variation exists in both the primary sex determination signal and in the downstream genetic pathways that interpret the signal (HAAG and DOTY 2005). It appears however that across eukaryotic domains, transcription factors are often key determinants regulating sexual development (KOTHE 1996; LEE *et al.* 2010b; KOESTLER and EBERSBERGER 2011; FERRARI *et al.* 1992; RIMINI *et al.* 1995). Transcription factors are cellular components that control gene expression and their activities determine how cells function and respond to the environment (VAQUERIZAS *et al.* 2009). To date, no MT determinants have been identified in the Stramenopila lineage. In the brown alga *Ectocarpus siliculosus*, sex appears to be determined by a single locus (COELHO *et al.* 2011). The complicated genetics of mating have also been studied in the oomycetes *Phytophthora infestans* and *P. parasitica* where Mendelian and non-Mendelian segregation respectively of the MT locus was described. Non-mendelian segregation of the sex locus is triggered by lethal alleles in cosegregation with the sex locus (JUDELSON *et al.* 1995; JUDELSON 1996; FABRITIUS and JUDELSON 1997).

Elucidating the molecular-genetic basis of MT determination and sexual reproduction in diatoms will not only contribute to a better understanding of the regulation and evolution of their life cycles, but will also establish the diatoms as a novel model group to study the evolution of reproductive strategies in eukaryotes. More specifically, they offer unique possibilities to study the evolution of isogamy from anisogamous (TOGASHI *et al.* 2012), and of heterothally from homothallic ancestors, but also to study the molecular basis of switches between reproductive modes in closely related species complexes (VANORMELINGEN *et al.* 2008; DAVIDOVICH *et al.* 2010). In centric diatoms, only homothallic reproduction is observed, while in pennates, heterothallic reproduction is most common (MANN *et al.* 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) and experimental evidence suggests that MT determination in heterothallic pennate diatoms is genetic (CHEPURNOV *et al.* 2004; DAVIDOVICH *et al.* 2010;

VANSTECHELMAN *et al.* 2013). By mapping the MT phenotype to MT specific linkage maps for the heterothallic pennate diatom *S. robusta*, we recently identified a single genomic region harboring the MT determination gene(s) and the MT^+ as the heterogametic MT. The homologous MT determining linkage groups in *S. robusta* appear to have a relatively large region of recombination (VANSTECHELMAN *et al.* 2013), suggesting that the MT determining chromosomes probably only recently evolved, at some point after the appearance of the pennate lineage around 75-90 MY ago (SIMS *et al.* 2006; BOWLER *et al.* 2008).

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 genetic markers strongly linked to the MT locus. In a provisional *S. robusta* genome assembly, this MT locus is located on a ~25 kb scaffold, containing a SF2-family related helicase/S-adenosylmethionine dependent methyltransferase (*HEL-SAM*) gene most likely responsible for the MT determination in *S. robusta*. A first differential expression analysis further supports the potential involvement of the HEL-SAM gene in MT determination.

Material and methods

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

Cell cultures of 25 MT^+ and 25 MT^- progeny from an F_1 mapping population (MP) (publicly available in the BCCM/DCG diatom collection (http://bccm.belspo.be/about/dcg.php), previously used to build MT-specific linkage maps (VANSTECHELMAN *et al.* 2013), 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 18º C with a 12:12-h light:dark period and approximately 85 µmol photons m^{-2} s^{-1} from cool-white fluorescent lights.

For Bulked Segregant Analysis (BSA) using AFLP, genomic DNA was isolated from individual cultures from the F_1 MP in exponential phase as described in Vanstechelman et al. (VANSTECHELMAN *et al.* 2013) followed by selective PCR amplification using AFLP primers (VUYLSTEKE *et al.* 2007b) of restriction fragments of a total digest of the genomic DNA obtained with the restriction enzyme combination *Eco*RI/*Mse*I. Equal quantities of the pre-amplification reactions of the AFLP of five cultures were mixed to form four bulks each of MT$^+$ or MT$^-$ (i.e. 20 MT$^+$ and 20 MT$^-$ strains of the MP were used) for further AFLP analysis 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 patterns were scored for presence/absence of AFLP bands between the two MT's. The polymorphic AFLP loci were further analyzed at the level of all offspring of the MP and integrated in the AFLP-based MT-specific linkage maps (VANSTECHELMAN *et al.* 2013) using the linkage analysis software Joinmap 4.0 (VAN OOIJEN 2006). AFLP markers strongly cosegregating with the MT phenotype were subsequently purified from sequencing gels followed by Sanger sequencing as described in Vuylsteke et al (VUYLSTEKE *et al.* 2007a).

For BSA using high throughput sequencing, we pooled, for each MT, equal amounts of cell material of 25 segregants of the F$_1$ MP. Genomic DNA was isolated as described 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, generating 85 M and 92 M paired end (PE) reads for the MT$^+$ and MT$^-$ bulk, respectively. This results in an initial coverage of 110 X and 120 X based on the genome size of *S. robusta* which was estimated to be 153 MB based on flow cytometry measurements (CHEPURNOV *et al.* 2008).

While allele frequencies should be approximately equal between the two bulks differing for MT in genomic regions without loci affecting the MT, they should differ between the two bulks at the single genomic region containing the MT locus. Consistent with a single sex-determining gene model in which MT$^+$ is heterogametic and MT$^-$ is homogametic, we searched for heterozygosity at SNP loci in the MT$^+$ bulk and homozygosity at SNP loci in the MT$^-$ bulk, using the "somatic option" in VarScan (KOBOLDT *et al.* 2012). The VarScan "somatic option", initially constructed to detect somatic mutations in tumor samples, 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.

The Illumina PE reads of the two bulks differing for MT were mapped to the genome assembly (see below) 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. Using the Fisher's Exact Test in GenStat (VSN INTERNATIONAL 2011), scaffolds were ranked according to their accumulation of LOH events (only scaffolds with minimum 10 SNPs were considered).

Genome assembly

The Illumina PE reads of the two bulks differing for MT were combined with Genome Sequencer (GS)-FLX-based (454 technology) single end (SE) and mate pair (MP) reads (with an insert size of 3, 8 and 20 kb; obtained by NTNU) to assemble a genome sequence for *S. robusta* using CLC Assembly Cell version 4.06beta (www.clcbio.com). The sequencing reads were trimmed for low quality (quality values < 20) using the option quality_trim. Assemblies were built using the quality based trimmed reads with the clc_novo_assemble option. First the word size [w; sequences of a certain length ('word') are built in the assembly] and subsequently the bubble size (b; a bubble is an indicator for allelic variance) were optimized to maximize the N_{50} value (the total length of all contigs of this length or longer is at least 50% of the total length of all contigs) of the obtained contigs. Reads were mapped to the generated assembly using the option clc_ref_assemble and percentage of mapped reads, the insert size for 99% of the paired reads and the coverage were calculated by the assembly_info option.

Scaffolding was performed using the software SSPACE (BOETZER *et al.* 2011). The minimum number of links (read pairs) to compute a scaffold and the maximum number of allowed gaps during mapping with Bowtie (LANGMEAD 2010) was set to three.

The obtained scaffolds were blasted (ALTSCHUL *et al.* 1997) to microbial protein databases (ftp://ftp.ncbi.nih.gov/refseq/release/microbial) to reduce microbial contamination. Scaffolds for blast hits with an E-value smaller than 1e-50, 1e-75, 1e-100 and 1e-200 were filtered out of the scaffolds file and the trimmed genome size was compared to the haploid genome size of *S. robusta* (153 MB) estimated by flow cytometry (CHEPURNOV *et al.* 2008; CONNOLLY *et al.* 2008).

Identification of the MT locus

Identification of conserved DNA domains residing on the 24698 bp scaffold was done using a Conserved Domains Database (CDD) search against the NCBI conserved domains database CDD v3.09-43334 PSSMs (MARCHLER-BAUER *et al.* 2009).

RNA of two strains [85A (MT^{+}) and 85B (MT^{-})] in the S-phase of the cell cycle above and below the SST was extracted and cDNA was synthesized as described in Gillard et al. (GILLARD *et al.* 2008). The analysis of the *S. robusta* transcriptome data set was carried out by the DOE Joint Genome Institute (JGI). Illumina reads of four libraries (table 1) of stranded RNA-seq data were built.

Table 1. Illumina reads of 4 libraries of stranded RNA-seq data.

library	mating type	cell size	#raw reads
I	+	>SST	85,323,814
N		<SST	85,582,710
G	-	>SST	82,053,896
H		<SST	68,113,460

De novo transcript contig assembly was done using Rnnotator (MARTIN *et al.* 2010). The Rnnotator assembly pipeline consists of three major components: preprocessing of reads, assembly, and postprocessing of contigs. Preprocessing consisted of removal of low-quality reads, low-complexity reads, adapter-containing and duplicate reads and read trimming. The assembly was completed with the software Velvet (ZERBINO and BIRNEY 2008) with the minimum contig length set at 100. Redundant contigs were removed using Vmatch (v. 2.1.4; http://www.vmatch.de/) and contigs with significant overlap were further assembled using Minimus2 (SOMMER *et al.* 2007) with a minimum overlap of 40. Contig postprocessing included splitting misassembled contigs, contig extension and polishing using the strand information of the reads. Single base errors were corrected by aligning the reads back to each contig with BWA (LI and DURBIN 2010) to generate a consensus nucleotide sequence. Post-processed contigs were clustered into loci and putative transcript precursors were identified.

The number of RNA-Seq reads which align to each gene model has been shown to be an accurate estimate of gene expression levels. In order to obtain these counts, for each RNA-seq sample, reads were first trimmed to 36bp, to ensure their alignment despite a low quality 3' end, followed by alignment to the reference gene models using BWA (LI and DURBIN 2010) (parameters used: seed length = 25, maximum hits = 1). Next, read counts were RPKM (Reads Per Kilobase of exon model per Million mapped reads) normalized (BULLARD *et al.* 2010), removing technical biases inherent to the sequencing approach, most notably the length of the RNA species and the sequencing depth of a sample. Differentially expressed genes were identified by a pair-wise analysis based on the maximum likelihood ratio statistics and False Discovery Rate (FDR=*q*) statistics.

The genomic scaffold (24698 bp), identified via the BSA procedure was blasted (blastn, (ALTSCHUL *et al.* 1997) against the transcriptome for the identification of the corresponding transcripts. The identified transcripts were used for differential expression analysis as described above.

The software GenomeThreader (GREMME *et al.* 2005) was used to compute gene structure predictions. Gene structure predictions were calculated using a similarity-based approach where cDNA and DNA sequences are used to predict gene structures via spliced alignments.

Results

Genome assembly

The Illumina PE reads and the GS-FLX-based SE and MP reads were used for building a first draft genomic assembly of *Seminavis robusta*. An optimal word size (w) of 28 in function of the N_{50} value was calculated under a default bubble size (b) of 50. Under these conditions of optimal w, the N_{50} value of the contigs was 4770 bp (figure 1). The N_{50} value of the contigs increased when increasing the b-size while keeping w constant at 28 (figure 2).

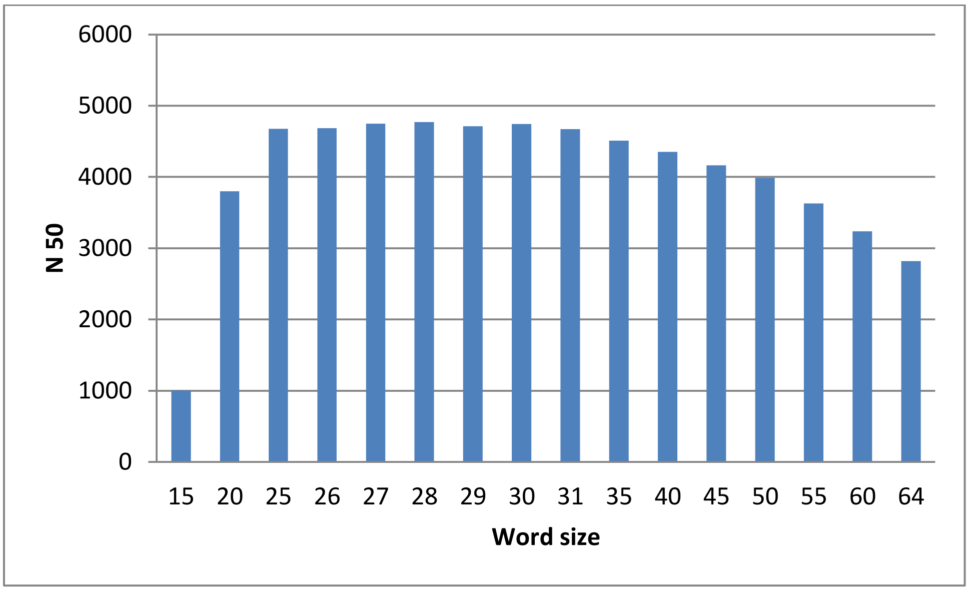


Figure 1. Optimization of the word size in function of the N_{50} value with a default bubble size (b=50).

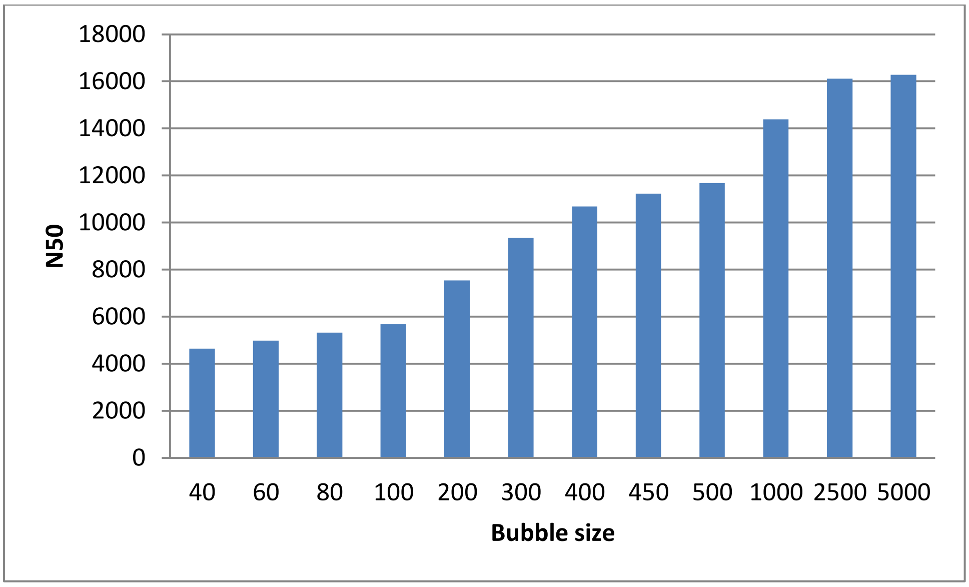


Figure 2. Optimization of the bubble size in function of the N_{50} value of the contigs using a constant optimal word size (w=28).

The assembly corresponding with b=500 was selected for final scaffolding as higher b increased errors in genome assemblies, resulting in N_{50} =11676 bp. The number of trimmed reads, the percentage of reads mapped to the genome sequence, the insert size range for 99% of the paired reads and the coverage of each library are displayed in table 2.

Table 2. Summary statistics for the genome assembly with the number of quality trimmed reads, percentage of reads mapping to the genome assembly, the insert sizes and the coverage of the libraries. The insert size is indicated as a range comprising 99% of the reads.

database	technology	# reads	% mapped reads	insert size (bp)	coverage (X)
MT^+ PE reads	Illumina HiSeq 2000	149,535,396	98.46	75 - 350	66.97
MT^- PE reads	Illumina HiSeq 2000	159,445,500	98.5	72 - 343	71.33
SE reads	GS-FLX	7,322,204	83.02	-	8.11
3kb MP reads	GS-FLX	1,339,458	92.21	608 - 4426	0.77
8kb MP reads	GS-FLX	1,579,222	93.43	2,908 – 13,862	0.94
20 kb MP reads	GS-FLX	1,433,550	90.48	5,082 – 23,805	1.01

Next, the scaffolds were trimmed at various blast E-values to reduce microbial contamination (table 3). Scaffolds with blast hits to the microbial databases smaller than E=1e-100 were trimmed from the assembly. The remaining scaffolds (145 MB, N_{50} = 15268, 72034 sequences) were selected as the *S. robusta* genome assembly for further analysis as this value is comparable to the estimated *S. robusta* genome size (153 MB) (CHEPURNOV *et al.* 2008).

Table 3. The genome size, the N_{50} value and the number of sequences for the genome assembly after trimming for microbial contaminants based on different blast E-values.

Trimmed at E-value of	genome size (MB)	N_{50} (bp)	# scaffolds
no trimming	234	23096	76570
1.e-200	163	15604	74771
1.e-100	145	15268	72034
1.e-75	132	14946	69745
1.e-50	111	13012	65090

Bulked Segregant analysis

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^- F_1 segregants 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 present in the MT^+ bulks only, of which five (E21M442M273.4; E33M132M142.6; E32M231M408.0; E34M221M432.1; E31M321M121.2) strongly cosegregated with the MT locus (LOD ranging from 17.32 to 31.53) which mapped to the previously identified linkage group MT^+_6 (VANSTECHELMAN *et al.* 2013) (figure 3). 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 genome assembly.

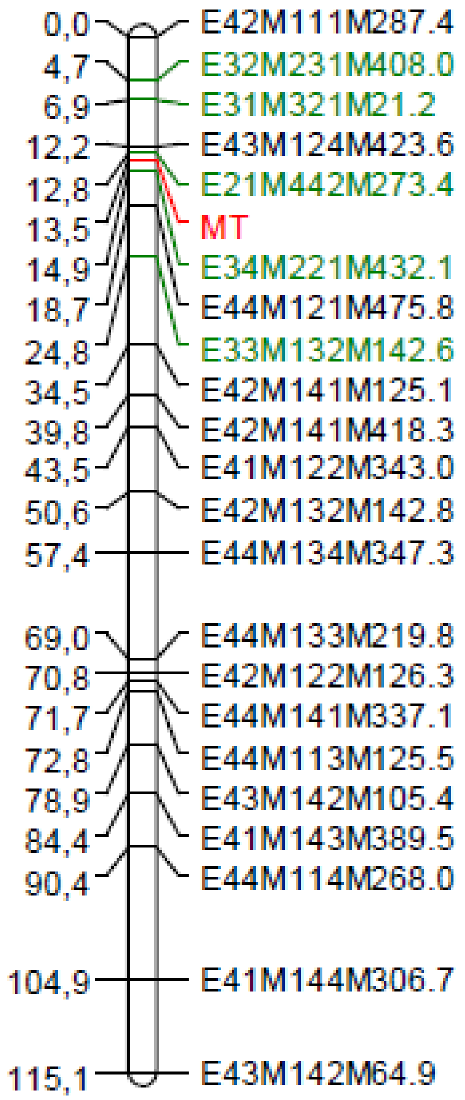


Figure 3. Linkage group MT$^+$_6 of *S. robusta* including the MT locus (red) (VANSTECHELMAN *et al.* 2013) with the five additional AFLP markers (green) identified using AFLP-based BSA.

The two marker sequences both mapped on a single scaffold (scaffold1897, length = 24698 bp) and were separated by ~8kb. This identified scaffold1897 as carrying the *S. robusta* MT locus. We observed a strong discrepancy between the genomic distance of the marker sequences in the genome assembly (7905 bp) and their genetic distance of 2.1 cM. Indeed, given a total map distance of ~970 cM (VANSTECHELMAN *et al.* 2013) and an estimated genome size of about 153 Mb (CHEPURNOV *et al.* 2008), 1 cM should correspond to a genomic length of ~157 kb. The physical to genetic map ratio in this region thus appears to be distorted ~40-fold when compared to the genome wide average.

Consistent with a single MT determining gene model for which MT$^+$ is heterogametic and MT$^-$ is homogametic (VANSTECHELMAN *et al.* 2013), we searched for SNPs showing allele frequency differences reflecting heterozygosity in the MT$^+$ bulk and homozygosity in the MT$^-$ bulk. Scaffolds were then ranked according to their percentages of SNP's 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 72034 scaffolds) in a ranking of scaffolds with the highest relative number of SNPs showing the target allele frequency difference (table S1). These identified scaffolds are covering the MT genomic region (MT locus and surrounding region) of *S. robusta*.

Out of the 405 SNP's residing on scaffold1897, 107 (26.4%) were called as an LOH event indicating that these SNP's (table S2) are in full linkage disequilibrium with the MT locus as heterozygosity occurs in MT$^+$ (~ equal frequencies of the two alleles) and homozygosity occurs in the MT$^-$ (single allele counts). 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 (also tested with Chi-square, *p*<0.001), what suggests that this scaffold covers the MT genomic region.

Identification of the MT locus

Three transcripts for scaffold1897 were found via a blast search of the transcriptome to the genomic sequence. A CDD search of the transcripts corresponding to the genomic region flanked by the two sequenced AFLP markers on scaffold1897 identified three important domain hits: domain PLN00113 is a leucine-rich repeat receptor-like protein kinase (LRR) (E-value = 1.56e-36); domain SF2 is a superfamily of DNA/RNA helicases (E-value=5.21e-03); and domain AdoMet_MTases super family (E-value=1.41e-04) is a family of S-adenosylmethionine dependent methyltransferases (SAM or AdoMet-MTase). Other conserved domains (a protein kinase (PK; E-value=5.47e-32) and a Hedgehog/Intein (Hh; E-value=1.29e-13) could also be identified on scaffold1897. An overview of the domains in scaffold1897 is given in figure 4.

Gene structure (exon-intron) prediction by the software GenomeThreader was done using scaffold1897 and the transcript of the HEL-SAM as cDNA (table S3). The genomic length of the HEL-SAM is 8409 bp with a total exon length of 7866 bp and a total intron length of 543 bp. A schematic overview of the structure of the *HEL-SAM* gene is given in figure 4. An amino acid (AA) sequence (supplementary data 1) of the HEL-SAM transcript was built using the software Expasy and this resulted in a protein length of 1944 AA.

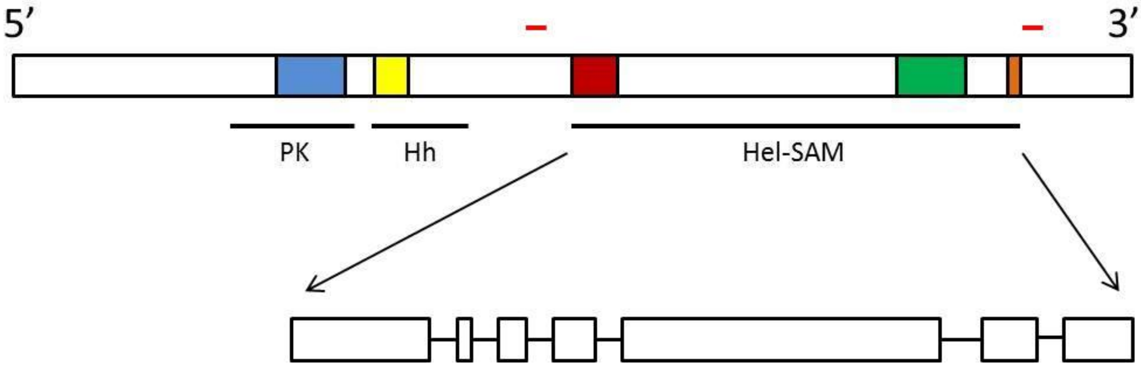


Figure 4. Above: schematic overview of the genes and their corresponding conserved domains on scaffold1897 (24698 bp). Genes are indicated with black lines and conserved domains with colored boxes (blue: protein kinase; yellow: hedgehog/intein; red: SAM; green: Hel; orange: LRR). AFLP markers mapping to this scaffold are shown as red lines. Below: Gene structure of the *HEL-SAM* gene. Exons are indicated as boxes and introns as lines.

Differential expression analysis of the *HEL-SAM* gene identified a significantly (q<0.01) higher expression in MT$^+$ cells compared with MT$^-$ cells, irrespective of size relative to the SST (figure 5).

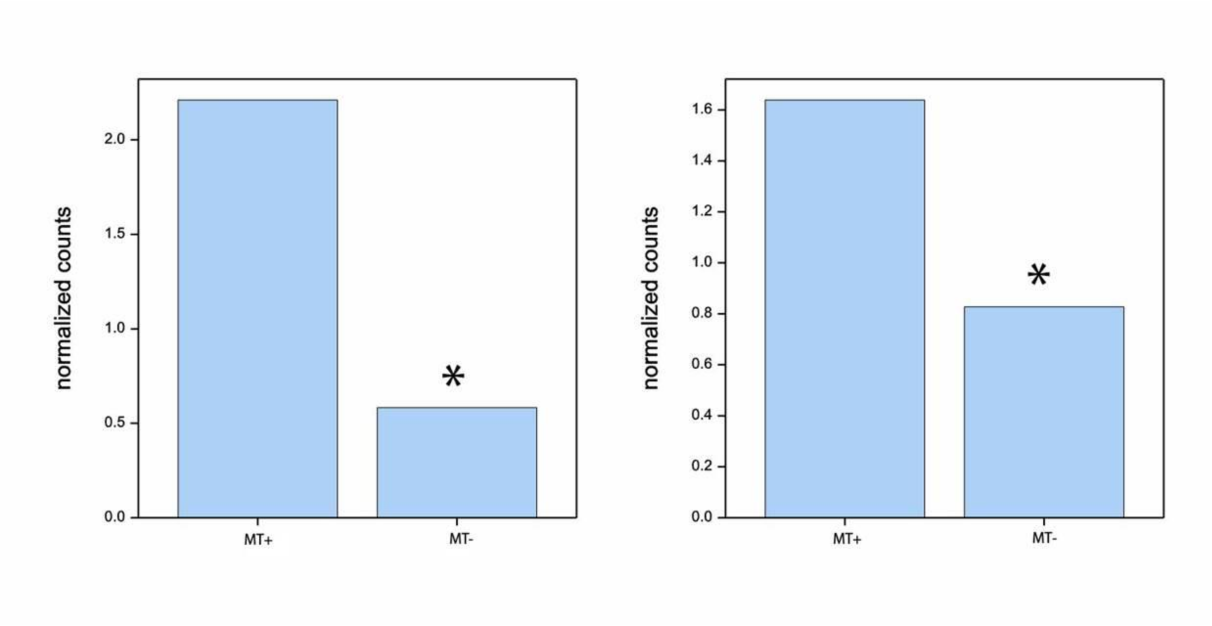


Figure 5: Normalized counts for the transcripts of the *HEL-SAM* for cells larger (left) and smaller (right) compared to the SST for both MT's. Significant differences (*) between MT$^+$ and MT$^-$ for both cell sizes are observed for this gene (q<0.05).

Discussion

We previously showed, using genome-wide linkage mapping, that the MT locus in the heterothallic pennate diatom *Seminavis robusta* segregates as a single locus, disclosing MT^{+} as the heterogametic MT (VANSTECHELMAN *et al.* 2013). Furthermore, we identified the MT locus to be strongly 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 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 (with the highest LOD score for linkage with the MT locus) mapping to a single 24698 bp scaffold (scaffold 1897). The BSA-NGS analysis revealed that this scaffold was highly enriched with SNP's affecting the MT genomic region. Other scaffolds affecting this region were also identified and this shows the need for a better genome assembly and a full annotation of the genome of *S. robusta*. The ~40-fold discrepancy in genomic distance (~ 330 kb) estimated from the 2.1 cM intermarker distance of the two AFLP markers in the linkage map and the genomic distance (7905 bp) between the AFLP markers on the scaffold could be explained by the fact that a ten times higher coverage is observed in the 5' genomic region of this ~8 kb region separating these markers. This higher coverage could be caused by the existence of sequence repeats in the 5' region, as such repeat regions are usually concatenated during assembly. A discrepancy of the physical/genetic map of around 10- to 50-fold compared to the genome wide average is also observed in the fungus *Cryptococcus neoformans*. 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).

Comparing the DNA sequence of scaffold1897 with the *S. robusta* transcriptome made it possible to identify the genetic structure of the MT locus as a SF2-family related Helicase/S adenosylmethionine dependent methyltransferase (*HEL-SAM*) with a transcript length of 7866 bp.

Helicases typically [but not necessarily (DURR *et al.* 2006)] unwind nucleic acid complexes in an ATP-dependent manner (JANKOWSKY 2011). Based on differences in amino acid sequence and protein structures, five classes of helicases (SF1-5) can be distinguished (JANKOWSKY *et al.* 2011). The helicase domain of the *S. robusta* MT locus is a SF2-family related helicase. This family is involved in virtually all aspects of RNA and DNA metabolism (JANKOWSKY *et al.* 2011). RNA helicases have been shown to play a role in sex determination in a few other organisms. A helicase is described as part of the sex chromosome in *Homo sapiens*. The SF2-type RNA helicase *DDX3* has two closely related genes designated DDX3Y and DDX3X, which are localized to the Y and the X chromosomes, respectively. DDX3Y is expressed in male germ cells and is essential for spermatogenesis (ABDELHALEEM 2005). RNA helicase genes are also part of the sex loci in Zygomyceta and Microsporidia where an ancestral sex locus might have spanned a triose phosphate transporter (TPT), a high mobility group protein (HMG) and an RNA helicase (SF2 family) gene. However, only the HMG factor acts as the sex determining factor (LEE *et al.* 2010a). It has been hypothesized based on observations in fungi, animals and plants that an ancestral TPT-HMG-helicase gene cluster existed in a common eukaryotic ancestor but that in most taxa analyzed so far, this ancestral gene cluster has been lost (KOESTLER and EBERSBERGER 2011).

The identification of the HEL-SAM suggests that MT determination may be (partly) epigenetically regulated. A SAM catalyzes the transfer of the methylgroup from S-adenosylmethionine to the target DNA base (CHENG 1995). The combination of a SAM and a SF2-type helicase domain has been described for the DNA methyltransferase 5 (DNMT5) protein family which also has one member in the diatoms *P. tricornutum* and *T. pseudonana* (PONGER and LI 2005; MAUMUS *et al.* 2011). This combination is also described for the *H. sapiens* protein ATRX (2492 AA), which plays a role in establishing and/or maintaining a normal pattern of DNA methylation (ARGENTARO *et al.* 2007). The DNMT5 and ATRX proteins have a size (~2000 AA) comparable to the *S. robusta* HEL-SAM protein. They also consist of an N-terminal methyltransferase and a C-terminal SF2-family related helicase domain, suggesting that they are possible orthologues of the *S. robusta* HEL-SAM. Interestingly, it has been reported that members of the SF2-type helicases are necessary for proper cytosine methylation (JEDDELOH *et al.* 1999) and that they can interact with methyltransferases to repress transcription (MYANT and STANCHEVA 2008).

We speculate that the *S. robusta* MT locus plays a role in MT determination by influencing the expression of other genes, but the exact working mechanism still has to be elucidated. It is possible that the identified SNP's in linkage disequilibrium with MT are responsible for a difference in function between the MT's. The results of our preliminary differential expression analysis of the *HEL-SAM* gene also revealed a significantly higher transcription in the MT^+. However, the difference in expression is only two to four-fold, and transcription in the MT^- is still present. Further transcriptional analyses and a full functional analysis of the *HEL-SAM* gene is necessary to understand the role of this gene as a MT determining factor in *S. robusta* and interaction studies can aid in the identification of downstream regulators of this transcription factor. The upregulation of the *HEL-SAM* in MT^+ is not size dependent, suggesting that other downstream determinants are responsible for the size dependent activation of sexual reproduction in diatoms.

Acknowledgements

The authors acknowledge the Ghent University Research Funding (BOF GOA project 01G01911 and BOF project 09/24J/103). We thank the JGI institute for the transcriptome analysis of *S. robusta* on the project: a deep transcriptomic and genomic investigation of diatom life cycle regulation (JGI proposal#CSP63). Special thanks also go to Tine Verstraete (PAE-Ugent) for helping with the BSA-AFLP analysis. We thank Joachim Jacob for helping with the SNP analysis and Stephane Rombauts and Stephanie De Vos for their support with the genomic assembly.

Literature cited

Abdelhaleem, M., 2005 RNA helicases: regulators of differentiation. Clin Biochem 38**:** 499-503.

Alonso, J. M., and J. R. Ecker, 2006 Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis. Nat Rev Genet 7**:** 524-536.

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

Argentaro, A., J. C. Yang, L. Chapman, M. S. Kowalczyk, R. J. Gibbons *et al.*, 2007 Structural consequences of disease-causing mutations in the ATRX-DNMT3-DNMT3L (ADD) domain of the chromatin-associated protein ATRX. Proc Natl Acad Sci U S A 104**:** 11939-11944.

Armbrust, E. V., J. A. Berges, C. Bowler, B. R. Green, D. Martinez *et al.*, 2004 The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306**:** 79-86.

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

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

Bowler, C., A. E. Allen, J. H. Badger, J. Grimwood, K. Jabbari *et al.*, 2008 The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature 456**:** 239-244.

Bozarth, A., U. G. Maier and S. Zauner, 2009 Diatoms in biotechnology: modern tools and applications. Appl Microbiol Biotechnol 82**:** 195-201.

Bullard, J. H., E. Purdom, K. D. Hansen and S. Dudoit, 2010 Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11**:** 94.

Cheng, X., 1995 DNA modification by methyltransferases. Curr Opin Struct Biol 5**:** 4-10.

Chepurnov, V. A., and D. G. Mann, 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., D. G. Mann, K. Sabbe and W. Vyverman, 2004 Experimental studies on sexual reproduction in diatoms, pp. 91-154 in *International Review of Cytology*, edited by K. W. Jeon. Elsevier Academic Press, San Diego.

Chepurnov, V. A., D. G. Mann, P. von Dassow, P. Vanormelingen, J. Gillard *et al.*, 2008 In search of new tractable diatoms for experimental biology. Bioessays 30**:** 692-702.

Coelho, S. M., O. Godfroy, A. Arun, G. Le Corguille, A. F. Peters *et al.*, 2011 Genetic regulation of life cycle transitions in the brown alga Ectocarpus. Plant Signal Behav 6**:** 1858-1860.

Connolly, J. A., M. J. Oliver, J. M. Beaulieu, C. A. Knight, L. Tomanek *et al.*, 2008 Correlated evolution of genome size and cell volume in diatoms (Bacillariophyceae). Journal of Phycology 44**:** 124-131.

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

De Riso, V., R. Raniello, F. Maumus, A. Rogato, C. Bowler *et al.*, 2009 Gene silencing in the marine diatom Phaeodactylum tricornutum. Nucleic Acids Res 37**:** e96.

Durr, H., A. Flaus, T. Owen-Hughes and K. P. Hopfner, 2006 Snf2 family ATPases and DExx box helicases: differences and unifying concepts from high-resolution crystal structures. Nucleic Acids Res 34**:** 4160-4167.

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

Ferrari, S., V. R. Harley, A. Pontiggia, P. N. Goodfellow, R. Lovell-Badge *et al.*, 1992 SRY, like HMG1, recognizes sharp angles in DNA. EMBO J 11**:** 4497-4506.

Gillard, J., V. Devos, M. J. Huysman, L. De Veylder, S. D'Hondt *et al.*, 2008 Physiological and transcriptomic evidence for a close coupling between chloroplast ontogeny and cell cycle progression in the pennate diatom Seminavis robusta. Plant Physiol 148**:** 1394-1411.

Granum, E., J. A. Raven and R. C. Leegood, 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., V. Brendel, M. E. Sparks and S. Kurtz, 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*. Plenum Press, New York.

Haag, E. S., and A. V. Doty, 2005 Sex determination across evolution: connecting the dots. PLoS Biol 3**:** e21.

Hsueh, Y. P., A. Idnurm and J. Heitman, 2006 Recombination hotspots flank the Cryptococcus mating-type locus: implications for the evolution of a fungal sex chromosome. PLoS Genet 2**:** e184.

Huysman, M. J., A. E. Fortunato, M. Matthijs, B. S. Costa, R. Vanderhaeghen *et al.*, 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.

Jankowsky, A., U. P. Guenther and E. Jankowsky, 2011 The RNA helicase database. Nucleic Acids Res 39**:** D338-341.

Jankowsky, E., 2011 RNA helicases at work: binding and rearranging. Trends in Biochemical Sciences 36**:** 19-29.

Jeddeloh, J. A., T. L. Stokes and E. J. Richards, 1999 Maintenance of genomic methylation requires a SWI2/SNF2-like protein. Nat Genet 22**:** 94-97.

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

Judelson, H. S., L. J. Spielman and R. C. Shattock, 1995 Genetic mapping and non-Mendelian segregation of mating type loci in the oomycete, Phytophthora infestans. Genetics 141**:** 503-512.

Koboldt, D. C., Q. Y. Zhang, D. E. Larson, D. Shen, M. D. McLellan *et al.*, 2012 VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research 22**:** 568-576.

Koestler, T., and I. Ebersberger, 2011 Zygomycetes, microsporidia, and the evolutionary ancestry of sex determination. Genome Biol Evol 3**:** 186-194.

Kothe, E., 1996 Tetrapolar fungal mating types: sexes by the thousands. FEMS Microbiol Rev 18**:** 65-87.

Langmead, B., 2010 Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics Chapter 11**:** Unit 11 17.

Lebeau, T., and J. M. Robert, 2003 Diatom cultivation and biotechnologically relevant products. Part I: cultivation at various scales. Appl Microbiol Biotechnol 60**:** 612-623.

Lee, S. C., N. Corradi, S. Doan, F. S. Dietrich, P. J. Keeling *et al.*, 2010a Evolution of the sex-Related Locus and Genomic Features Shared in Microsporidia and Fungi. Plos One 5.

Lee, S. C., M. Ni, W. Li, C. Shertz and J. Heitman, 2010b The evolution of sex: a perspective from the fungal kingdom. Microbiol Mol Biol Rev 74**:** 298-340.

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

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25**:** 2078-2079.

Lommer, M., M. Specht, A. S. Roy, L. Kraemer, R. Andreson *et al.*, 2012 Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome Biol 13**:** R66.

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

Mann, D. G., V. A. Chepurnov and S. J. M. Droop, 1999 Sexuality, incompatibility, size variation, and preferential polyandry in natural populations and clones of *Sellaphora pupula* (Bacillariophyceae). Journal of Phycology 35**:** 152-170.

Marchler-Bauer, A., J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott *et al.*, 2009 CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Research 37**:** D205-D210.

Martin, J., V. M. Bruno, Z. Fang, X. Meng, M. Blow *et al.*, 2010 Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics 11**:** 663.

Maumus, F., P. Rabinowicz, C. Bowler and M. Rivarola, 2011 Stemming epigenetics in marine stramenopiles. Curr Genomics 12**:** 357-370.

Myant, K., and I. Stancheva, 2008 LSH cooperates with DNA methyltransferases to repress transcription. Mol Cell Biol 28**:** 215-226.

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

Rimini, R., A. Pontiggia, F. Spada, S. Ferrari, V. R. Harley *et al.*, 1995 Interaction of normal and mutant SRY proteins with DNA. Philos Trans R Soc Lond B Biol Sci 350**:** 215-220.

Sims, P. A., D. G. Mann and L. K. Medlin, 2006 Evolution of the diatoms: insights from fossil, biological and molecular data. Phycologia 45**:** 361-402.

Sommer, D. D., A. L. Delcher, S. L. Salzberg and M. Pop, 2007 Minimus: a fast, lightweight genome assembler. BMC Bioinformatics 8**:** 64.

Togashi, T., J. L. Bartelt, J. Yoshimura, K. Tainaka and P. A. Cox, 2012 Evolutionary trajectories explain the diversified evolution of isogamy and anisogamy in marine green algae. Proc Natl Acad Sci U S A 109**:** 13692-13697.

Van Ooijen, J. W., 2006 *JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations*. Kyazma B.V., Wageningen.

Vanormelingen, P., V. A. Chepurnov, D. G. Mann, K. Sabbe and W. Vyverman, 2008 Genetic divergence and reproductive barriers among morphologically heterogeneous sympatric clones of Eunotia bilunaris sensu lato (Bacillariophyta). Protist 159**:** 73-90.

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

Vaquerizas, J. M., S. K. Kummerfeld, S. A. Teichmann and N. M. Luscombe, 2009 A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10**:** 252-263.

VSN International, 2011 GenStat *for Windows* 14th Edition, pp. VSN International, Hemel Hempstead, UK.

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

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

Zerbino, D. R., and E. Birney, 2008 Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18**:** 821-829.

Supplementary information

Supplementary data 1:

>*S. robusta* HEL-SAM

MHESYKPMNHPVYDKLTTAEIQATKQYLDPCGMDATDDIIGRLIGDQIDKVAGLWS

RALAHDKITLGSAETPLRLGTACSGTDAPALAMTMVLEQMELRRRDNDNDNNTLPV

MHYSHEFSCENDPMKQAYLARNFDSKLYPDIARLCDTPAPRDVFGQEQPLPKANMF

VAGTSCKDFSTMRSKKRKDIEDRGCSGETFLAACEVILKEQPTICILENVVGAWWKK

MSEYIVGRIQLSDAGKGKDGETKDATKGPKELKFSFDNKQGNKLVVEEVPKMWGV

HCGSTVAGFLKGDTMGKIHPVEWPTKVKKTEKCTLTDLLGPNKIDKKKDWLVFDRP

VTYQTHWVRVDTKNFGLPQTRQRTYMFVWQTNPQPKDGFFYPDNLGAYWEALVK

HLESPVKHPLEAFLLPLDHDNMRAYREALRGPLGRESTRRKAMYPNFWTTASKNRP

HNKAARTTSGIGDEARHITQWGDNGKKKLPPHYWLEYMNCNSPRCLDVVDILHAQA

LRDAESHDANHASFFWNVSQNVTKERHRTAKPGIASCITPGGDFLLPHQGRNLLGVE

KLLIQGIPYFRLALGNETEVNLADLAGNAMSSTVVCATMLAALMAPQLRKEMQASI

AGDDTTNKAPSLEKIKEFLEKHARLDSTGEQSRKWAGEISMVKTELQSNEADRHTGD

KSCMELFGELAALAPLAVKSSVWCTCESSGNRSRTNKFVKCKVCCVSCCRNCISEHA

GYNLDTHDTVEIEIGNESGADFPVDERNLFEFEKKLRSMLPSSVYLSKEGIGKVAEVE

NDRYRIRGLDKFSFVLYSIKRECGKWLIVYYARDKCGVGEAVAEMRITVGKVRRGL

AGESNLLGLRCELTSFMPARSEPLQYGNLAPCCQLVLLLPKDGSAPNVNDATWQVRL

PEVSSSIQLSGTGDTDSYRVEVGITDEAPEALEQHTKSKKGRYELVVKTKEERRWLYP

KVWKKWPSSIRIRCPHDGIDRRVSGTYLRAACRQTTNQSALWIKQVEGKPSIYLLMK

PNVSRTGGDYAIITTSMDFTDTSCILAELDQYWQPGDAFDATNQRQAVRCQSWGALP

GFECKIPKSTIQVEAPGTENNLVVVQGLKRTEATMLERIDCMNRTSSGDGALVELNV

TSGQRAQQVVRRINAIIAPQILKFAAMGGLRYDLSPTADWEVLESNDDEPFGTCVPVR

PKESWFYDEERERWDRRFDHDASRTYYMNLQGAQKPFEFWLDRERCSLEIKMFPKV

VAHAAAAQLLRGRGILSDDNKHGGAQVSFRLSDVSQQIDPTIDRFRVENCSALDATP

VELQHPYKLYDRQRKVLTKMLAIEEGKTSFLEMEMSESNMPGTVGFSLMAQAKRSR

NICGGVIADAIGAGKTVISIAIIVQGIERARASRATPRKSGATLVVVPPALIDQWESEVQ

KFAPSLVVITVHDFKELDSIKLSSLIEADVVVCPVDILESKKYLANLVQKAGLDKTTSA

EHLPKLPAYSGEREQIEARGVWIPHESADPYNVGSKTNDQQRRNKSAYYTYVYQSAI

QSLREKNFSPSDTGIPLEYFEFERLFVDEIHESLCTTKEELATAAETAKSSNVGFWKEK
NRRAGRELLGITQKDVTKRPLVCRRAVFGLTGTPLLDSNSRVIELANLMGGTYVLGL
SKIWRKLERESCRDIFLHNFLEPKPSRAVRRQGYARCQDYLKVACCRNRSGEEMKGI
ELQEHVEVIQMSEAEKTAYLKSQIGIPADQRCLGIKPTDFDENEGQDMSGLLRQNASC
PSRGAKLVEICQKILSEAPTTKIIVFTDGSIGAGIAAREALCGPDGPDCTWLDTADSAK
LKNEKIGWYQRGDATDEDHQRPRVLVLHYEHAAGLNLQSESHNVILFSPLYVGKGG
CSDDPVSDVSTELQAIGRVYRIGQPSRVVHVYRIEVQGPKGEECLDGKLIDRNSDPEVI
AMATNAE

Table S1: Top 0.1% of scaffolds accumulated with LOH events (based on Fisher’s Exact Test).

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
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
scaffold5375_size11158	49	24	5.23E-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
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 S2. The position of the SNP's identified as LOH event on the scaffold 1897 detected by the software VarScan (KOBOLDT *et al.* 2012), The bases for the 2 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.

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	0.00120402

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
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
10090	A	C	30	13	0	17	3.09E-07
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	0.001612652
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
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	0.002100401
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
20432	T	C	17	15	33	0	2.73E-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 S3. Predicted gene structure of the *HEL-SAM* after spliced alignment performed with the software GenomeThreader (GREMME *et al.* 2005).

exon/intron	start	stop	length (bp)
exon 1	11581	13352	1772
intron 1	13353	13440	88
exon 2	13441	13468	28
intron 2	13469	13552	84
exon 3	13553	13678	126
intron 3	13679	13762	84
exon 4	13763	13915	153
intron 4	13916	14003	88
exon 5	14004	18930	4927
intron 5	18931	19050	120
exon 6	19051	19442	392
intron 6	19443	19521	79
exon 7	19522	19989	468

4

Evolution of the mating type locus in diatoms

Ives Vanstechelman[1,2,3], Marie J.J. Huysman[1,2,3], Sara Moeys[1,2,3] Marina Montresor[4], Mariella Ferrante[4], Remo Sanges[4], Wim Vyverman[1], Marnik Vuylsteke[2,3] and Koen Sabbe[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, 80121 - Naples, Italy

Key words: diatom, evolution, mating type locus, phylogeny

Publication in preparation

Authors' contributions

IV performed the experiments, analyzed the data and wrote the manuscript. MJJH an SM were involved in the design of the experiments. MJJH helped with the phylogenetic analyses and approved the manuscript. MM, MF and RS are involved in the production and in the processing of expression data of *P. Multistriata.* KS, WV and MV helped to conceive and design the study, and read and approved the manuscript.

Abstract

Sexual reproduction is a general feature in diatoms, but mating type determination probably appeared in the predominantly heterothallic pennate diatom lineage, an estimated 75-90 MY ago. The mating type determining region in the pennate heterothallic model diatom *S. robusta* was recently identified as a single locus comprising an SF2-family related helicase/S-adenosylmethionine dependent methyltransferase (*HEL-SAM*) gene (Chapter 3). Here, we study the evolution of this mating type determining factor within the diatom lineage. Sequence homology studies for this gene recovered from genomic data sets of other centric and pennate diatom species revealed the presence of strongly homologous *HEL-SAM* genes in all studied species, implying that an ancestral *HEL-SAM* was already present before the divergence of the centric and the pennate lineage and hence before the origin of mating type determination in diatoms. Phylogenetic analyses of HEL-SAM proteins in different species however revealed strong discrepancies between phylogenies based on the HEL-SAM proteins and trees based on recent phylogenetic studies. This strongly suggest that the HEL-SAM contains little phylogenetic signal, at least within the diatoms as a whole, which may be due to rapid evolution. The absence of mating type differentiation in the centric diatoms suggests that the *HEL-SAM* may only have been recruited in the pennate lineage as a MT determining factor.

Introduction

Diatoms (which belong to the Stramenopila lineage) are a very diverse and productive group of algae (MANN 1999) and often dominate phytoplankton and phytobenthos communities in aquatic environments (GRANUM *et al.* 2005). Genomic information for diatoms has grown rapidly over the past few years (ARMBRUST *et al.* 2004; BOWLER *et al.* 2008; LOMMER *et al.* 2012) and reverse genetic tools have been developed and optimized (SCALA and BOWLER 2001; SIAUT *et al.* 2007; DE RISO *et al.* 2009). Diatoms are recently also increasingly recognized for their potential in biotechnological applications (LEBEAU and ROBERT 2003; BOZARTH *et al.* 2009).

The raphid pennate diatom *Seminavis robusta* is used as a model organism for the study of sexual reproduction and mating type (MT) determination in diatoms as sexual reproduction can be reliably controlled in this heterothallic species and other commonly used model diatoms such as *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* appear to lack a sexual reproduction cycle (CHEPURNOV *et al.* 2008). The construction of MT specific linkage maps recently enabled the identification of a single locus MT determination system in *S. robusta* and revealed the MT$^+$ as the heterogametic MT (VANSTECHELMAN *et al.* 2013). Further investigation of the MT locus in *S. robusta* resulted in the identification of a SF2-family related helicase/S-adenosylmethionine dependent methyltransferase (HEL-SAM), with a protein length of 1944 amino acids (AA), as the MT determining factor in *S. robusta* (Chapter 3). This gene configuration (with a SAM and a SF2-type helicase domain) has been described for the DNA methyltransferase 5 (DNMT5) protein family which has been shown to contain one member in the diatoms *P. tricornutum* and *T. pseudonana* (PONGER and LI 2005; MAUMUS *et al.* 2011).

Diatoms are a relatively young group that appeared in the Mesozoic era (<205 MY) (MEDLIN *et al.* 1997; MEDLIN *et al.* 2000). However, new results based on molecular clock analyses suggest that the diatom lineage evolved sometime near the Devonian-Carboniferous transition (around 350 MY ago), and that the fossils of many diatom groups could be much older than the currently known paleontological record suggests (BROWN and SORHANNUS 2010). The two major clades of diatoms, the centric and the pennates, emerged >180 MY and ~75-90 MY ago respectively (SIMS *et al.* 2006; MEDLIN 2011). Centric diatoms are homothallic (both types of gametes occur in monoclonal cultures) (CHEPURNOV *et al.* 2008) while most pennate diatoms appear to be heterothallic (strains of two different MT's are needed for mating) (ROSHCHIN 1994; MANN *et al.* 1999; CHEPURNOV and MANN 2004). The evolution from homothally (centrics) to heterothally (most pennates) by definition entails the evolution to genetic MT determination (CHEPURNOV *et al.* 2004; DAVIDOVICH *et al.* 2010).

Here we report on the identification of homologous sequences of the *S. robusta* MT determining gene (*HEL-SAM*) in five pennate and two centric diatoms for which whole genome information is available. We then performed preliminary phylogenetic analyses of the identified homologs and studied the synteny of the genomic region of the *HEL-SAM*. Phylogenetic analysis of these homolog HEL-SAMs should give more insight into the evolution of the MT locus in diatoms. Synteny between the MT genomic regions of the different species and different groups should imply that possibly more genes located in the MT determining region are involved in the MT determination and can inform us about the evolution of the MT locus.

Material and methods

Identification of homologs of the *S. robusta* MT locus gene in other diatoms

The genomic, transcriptomic and proteomic sequences of the diatoms *Thalassiosira pseudonana*, *Phaeodactylum tricornutum*, *Fragilariopsis cylindricus*, *Pseudo-nitzschia multiseries* and the transcriptome of *Pseudo-nitzschia multistriata* are available online through the JGI portal (http://genome.jgi-psf.org/). A blastx (ALTSCHUL *et al.* 1998) of the *S. robusta* HEL-SAM transcript was ran to the available proteomic sequences and a tblastx was ran to the *P. multistriata* transcriptome to identify homologs of the *S. robusta* HEL-SAM. A blastx of the *S. robusta* HEL-SAM to the recently published genome sequence of *Thalassiosira oceanica* (LOMMER *et al.* 2012) was also performed. Sequences of *T. oceanica* are available through the NCBI website.

Phylogenetic analysis of the MT locus

An amino acid sequence alignment was constructed for the HEL-SAM proteins of the eight diatom species together with reference sequences selected out of the DNMT phylogenetic tree of Maumus, representing different DNMT groups (MAUMUS *et al.* 2011). The identified DNMT's from diatoms and DNMT proteins from *Arabidopsis thaliana*, *Homo* *sapiens*, *Micromonas pusilla*, *Ostreococcus tauri* and *O. lucimarinus* were selected. Bacterial DNMT's were chosen as outgroup proteins. Separate phylogenetic analyses for the two different domains (HEL and SAM domain) were also performed for the homolog diatom HEL-SAM proteins using the *Homo sapiens* DNMT1 as an outgroup protein. Alignments were performed using the ClustalW (LARKIN *et al.* 2007) application in BioEdit7 (HALL 1999). This was compared with alignments via the MUSCLE software (EDGAR 2004). The alignments were cleaned up with the Gblocks software (http://molevol.cmima.csic.es/castresana/Gblocks) to identify the conserved domains and shown with Multiple Align Show (http://www.bioinformatics.org). Phylogenetic trees were constructed using the software Treecon (VAN DE PEER and DE WACHTER 1997) and the MEGA software (TAMURA *et al.* 2011). The Poisson correction was used for distance calculation, the bootstrap value was set to 500 and the neighborhood joining (NJ) method was compared with the maximum likelihood (ML) method for tree topology.

Analysis of the synteny of the MT genomic region

Synteny of the *HEL-SAM* genomic regions between the available diatom genomes was analyzed. This was performed by a blast search (ALTSCHUL *et al.* 1998) of the *S. robusta* scaffold sequence enclosing the *HEL-SAM* gene (~25 kb) against other diatom genomes. For the identified homolog *HEL-SAM* genes in diatoms, the different protein sequences in the genomic regions (~40 kb) found in the JGI genome browser were compared by blast searches between the different species.

Preliminary expression analysis of the HEL-SAM gene in *P. multistriata*

A preliminary expression analysis with transcriptome data was done in the heterothallic pennate diatom *P. multistriata*. Two strains (strains Sy373(MT$^+$) and Sy379 (MT$^-$)) were grown and cells were collected during the exponential phase (not synchronized) by filtration on 1.2 µm filters. Trizol was used for RNA extraction followed by cleaning up on qiagen columns. RNA quality was checked with a bioanalyzer and cDNA was synthezised. The analysis of the transcriptome data of *P. multistriata* was done by JGI. Data of two libraries (Sy373 (MT$^+$) and Sy379 (MT$^-$)) were made and differences in gene expression levels between the two MT's for the homologous sequence of the *S. robusta HEL-SAM* were compared. The number of RNA-Seq reads which align to each gene model has been shown to be an accurate estimate of gene expression levels. In order to obtain these counts, for each RNA-seq sample, reads were first trimmed to 36bp, to ensure their alignment despite a low quality 3' end, followed by alignment to the reference gene models using BWA (LI and DURBIN 2010) (parameters used: seed length = 25, maximum hits = 1). Next, read counts were RPKM (Reads Per Kilobase of exon model per Million mapped reads) normalized (BULLARD *et al.* 2010), removing technical biases inherent in the sequencing approach, most notably the length of the RNA species and the sequencing depth of a sample and differentially expressed genes were identified by a pair-wise analysis based on the maximum likelihood ratio statistics and False Discovery Rate (FDR) statistics.

Results

Identification of homologs of the *S. robusta* MT locus gene in other diatoms

The results of the blast searches of the *S. robusta* HEL-SAM in different diatom species are shown in table 1.

Table 1. Results of the different blast searches of the *S. robusta* HEL-SAM in the genomes or transcriptomes of different diatom species. The uniquely identified HEL-SAM proteins are listed with their corresponding features (blast E-value, protein ID, protein length and chromosome, scaffold or contig including the identified sequence).

species	E-value	Protein ID	length (AA)	chr/scaff/cont
T. pseudonana	0.0	3158	2271	chr 3
F cylindricus	0.0	212855	2141	scaff 29
P. multiseries	0.0	192957	1935	scaff 134
P. multistriata	0.0	-	2195	-
T. oceanica	0.0	-	2182	cont 397621034
P. tricornutum	0.0*	45071-45072	2369	chr 6

* The *S. robusta* HEL-SAM was blasted on the manually corrected sequence of *P. tricornutum* (see below).

Unique, significant HEL-SAM hits were found in all the examined diatom species. Outside the diatom lineage, unique hits (one hit with E-value=0.0) were only found in the green algae *Ostreococcus tauri* and *O. lucimarinus* (*DNMT5* genes).

The HEL-SAM in *P. tricornutum* was described by Maumus (MAUMUS *et al.* 2011) as a homolog of *T. pseudonana* protein 3158. However, the gene was incorrectly annotated as two different genes with ID 45071 (HEL) and ID 45072 (SAM) as one of the EST's is overlapping a non-transcribed region and no stop codon could be identified at the 3'region of the SAM gene. The manually corrected HEL-SAM of *P. tricornutum* has an AA sequence length of 2369 AA and this protein sequence is provided as supplementary data 1.

Phylogenetic analysis of the MT locus

Eukaryotic DNMTs (DNA methyltransferases) can, on the basis of sequence similarity, be grouped in six functional groups, DNMT1-DNMT6 (PONGER and LI 2005). The combination of a SAM and a SF2-type helicase domain has been described for the DNMT5 protein family which also has one member in the diatoms *P. tricornutum* and *T. pseudonana* (Chapter 3) (PONGER and LI 2005; MAUMUS *et al.* 2011). A phylogenetic analysis was performed to investigate if the diatom HEL-SAM's all cluster in the same DNMT5 clade.

No remarkable differences between the used methods (ClustalW-Muscle and NJ-ML) could be found in the phylogenetic analysis. The evolutionary tree (NJ method) of the conserved domains of the HEL-SAM proteins in the seven diatom species together with the selected DNMT's is shown in figure 1 and the multiple sequence alignment is shown in supplementary data 2. The corresponding evolutionary tree (ML method) is added as supplementary figure 1.

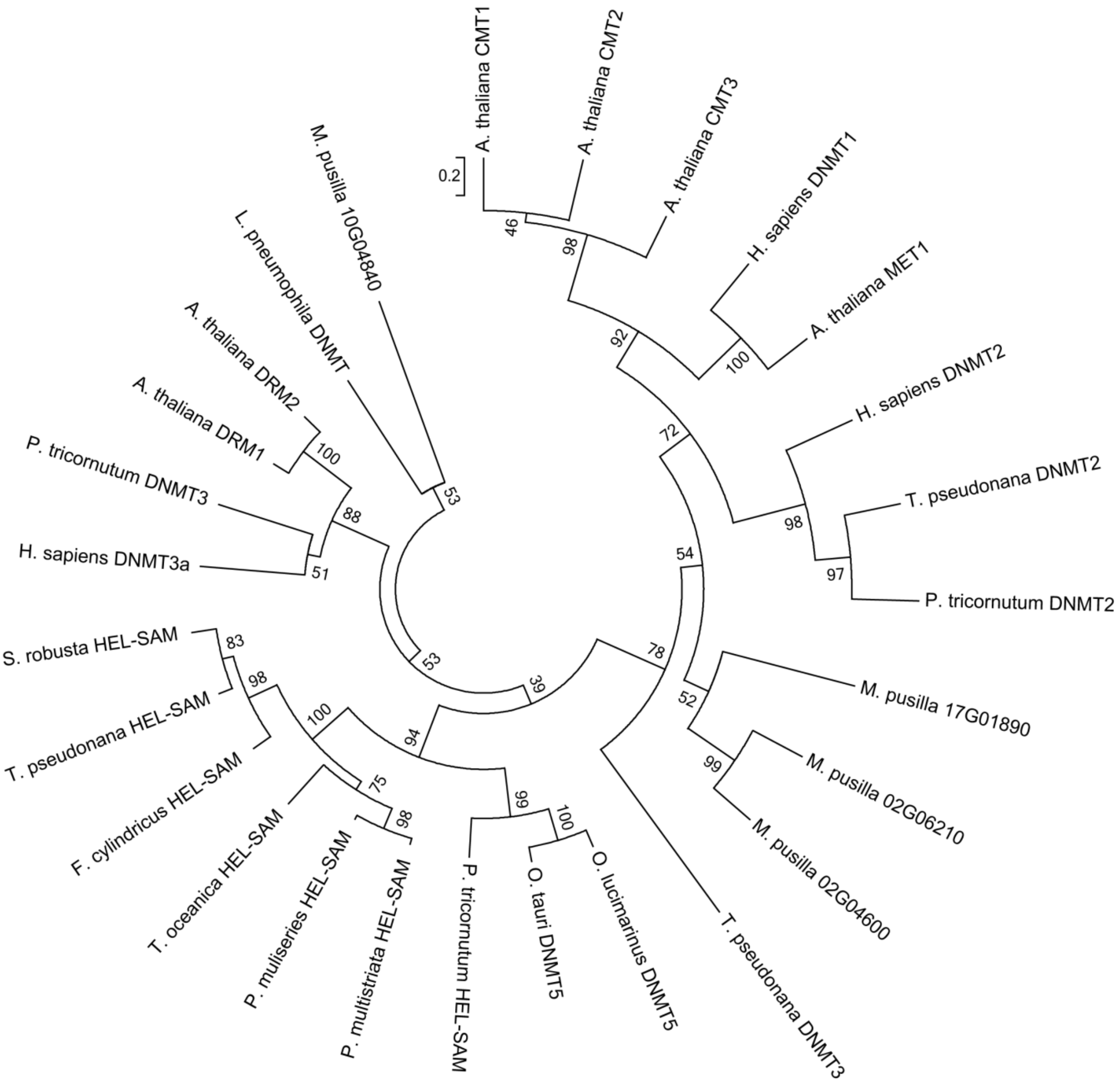


Figure 1. NJ phylogeny for the diatom HEL-SAM proteins and the selected DNMTs based on an alignment of the conserved domains (methyltransferases). Bootstrap values are included. Bacterial DNMTs were chosen as outgroup proteins.

The HEL-SAM proteins of the different diatom species all cluster in the same clade together with DNMT5 proteins (DNMT5 group). DNMT5 proteins in diatoms were known from the diatoms *P. tricornutum* and *T. pseudonana* (PONGER and LI 2005; MAUMUS *et al.* 2011), and were identified as the *S. robusta* HEL-SAM homologs in these species. The *P. tricornutum* HEL-SAM is the least related with the HEL-SAM proteins of the other diatom species as it is clearly differentiated in another clade together with the DNMT5 proteins of the two *Ostreococcus* species.

There seems to be no relationship between the phylogeny of the HEL-SAM proteins and known and generally widely accepted 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). Separate phylogenetic analyses of the SAM and the HEL domains are shown in figure 2 (NJ method). No remarkable differences can be observed between the phylogeny of the diatom HEL-SAM proteins shown in figure 1 and the separate phylogenies of the two domains, and no remarkable differences between the two analyzed domains can be observed.

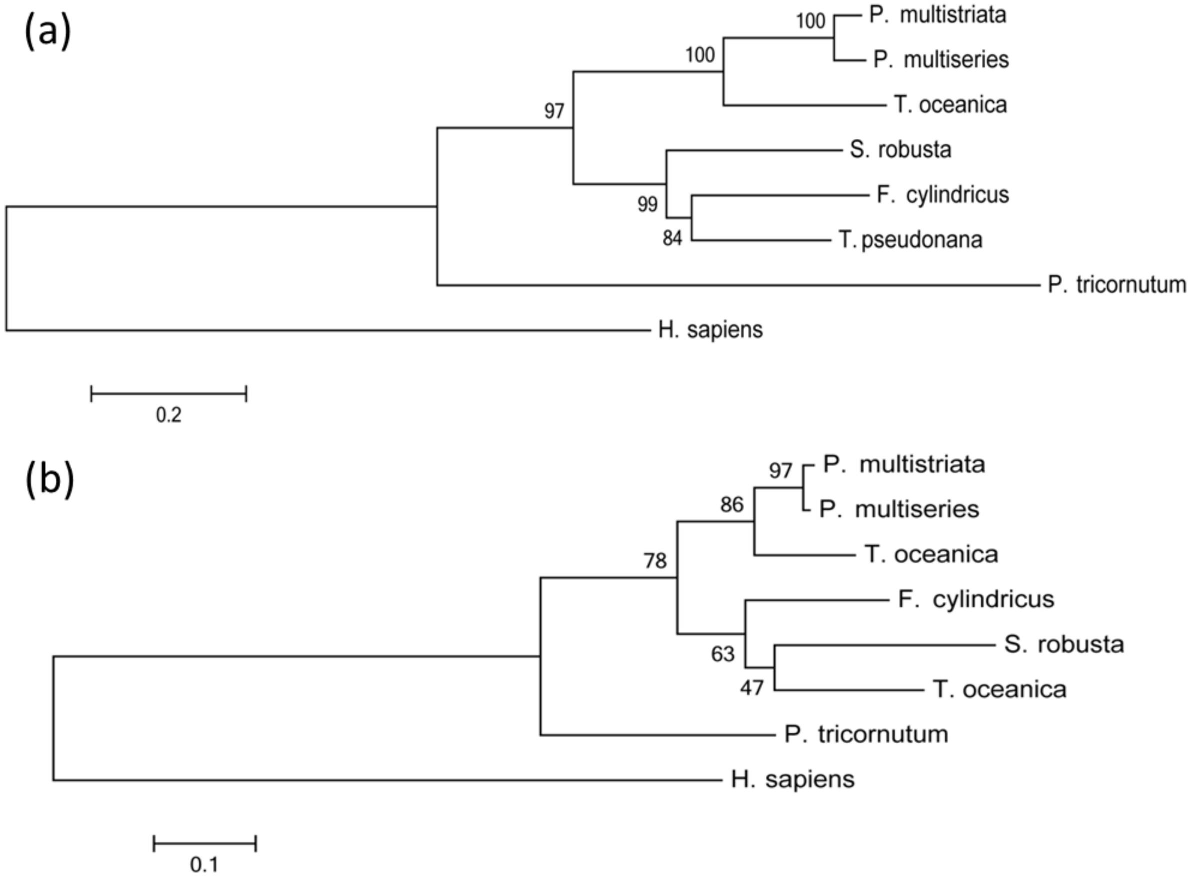


Figure 2. NJ phylogeny of the HEL domain (a) and the SAM domain (b) of the diatom HEL-SAM proteins. Bootstrap values are included. Human DNMT1 was chosen as an outgroup protein.

Analysis of the synteny of the MT genomic region

No significant blast hits could be found between the genomic region (~25 kb scaffold) of the *HEL-SAM* gene in *S. robusta* and the other diatom species. The genomic regions (~ 40 kb) of the identified HEL-SAM's in the JGI genome browser could also not identify any homology of neighboring proteins between different diatom species except between the two *Pseudo-nitzschia* species. The neighboring region of the *HEL-SAM* in *T. oceanica* could not be compared as genomic contigs for this species were too small (N_{50}=3623 bp). The DNA contig (50 kb; draft genome sequence available at Stazione Zoologica Anton Dohrn) mapping the cDNA *HEL-SAM* of *P. multistriata* was provided. This DNA region of *P. multistriata* was used for a tblastx against *P. multiseries*. Three genes 5' of the *HEL-SAM* of *P. multistriata* which are lying in the 50 kb contig including the *HEL-SAM* could be picked up as homologs of neighboring genes of the *P. multiseries HEL-SAM*. The hypothetical protein 5' of the HEL-SAM (protein ID 286440) resulted in a significant blastx hit (E = 4.92 e-167). The neighboring gene 5' of gene 286440 resulted in a significant blastx hit to *P. multistriata* with an E-value of 6.15 e-102. This has the protein ID 192935 and has a haloacid dehalogenase-like hydrolase function. The gene 5'of that protein (protein ID = 286438) (a glycosyltransferase gene) resulted in a significant E-value of 0.0. The regions surrounding the *HEL-SAM* gene for the different diatom species are illustrated in figure 3.

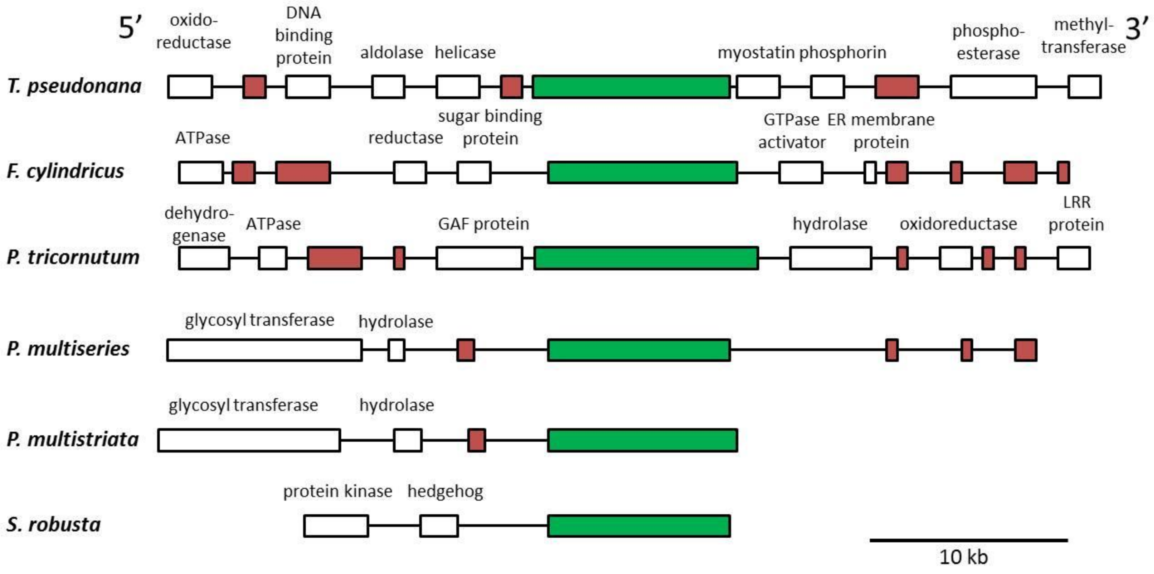


Figure 3. Genomic regions 5' and 3' of the HEL-SAM gene. Boxes indicate genes (green = HEL-SAM; red = hypothetical proteins).

Preliminary expression analysis of the HEL-SAM gene in *P. multistriata*

The normalized transcriptomic counts of the MT^+ and MT^- reads for the *HEL-SAM* homolog of *P. multistriata* shows that transcription is significantly upregulated for the MT^+ strains compared to the MT^- strains (q<0.01) (Figure 4). This phenomenon was also described in *S. robusta* (Figure 5 in Chapter 3).

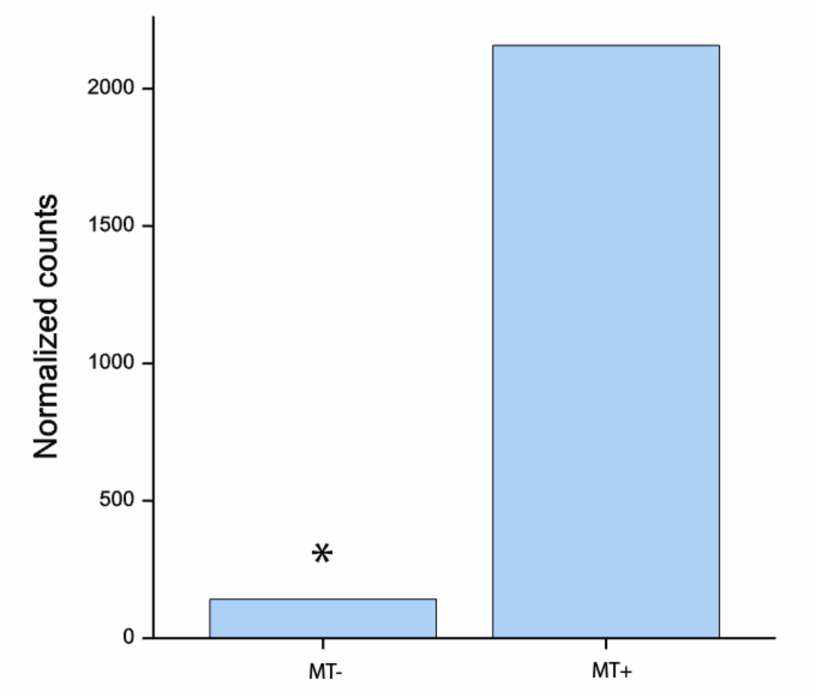


Figure 4. Normalized counts of the transcriptome reads for the HEL-SAM in *P. multistriata*. Comparison between the libraries of the two mating types. (*) Significant difference between two libraries (*q*<0.01).

Discussion

In diatoms, the evolution from homothally (centric diatoms) to heterothally (pennate diatoms) involves the evolution of genetic MT determination (CHEPURNOV *et al.* 2004; DAVIDOVICH *et al.* 2010). Homologs of the MT locus of *S. robusta* (HEL-SAM) could be identified in all examined diatom species for which extensive genomic or transcriptomic information is available: unique blast hits confirmed the existence of this unique HEL-SAM gene in all diatoms sequenced to date. This indicates that an ancestral HEL-SAM originated before MT determination found its origin in diatoms as the HEL-SAM is also present in the centric species where no MT determination is described. Assuming that the HAL-SAM is the MT determinant in the examined species, we can conclude that this ancestral HEL-SAM was probably recruited as the sexual determinant in the pennate lineage.

The phylogenetic analysis shows that all the HEL-SAM proteins of the different diatom species cluster together in the DNMT5 clade. Two green algal (*O. tauri* and *O. lucimarinus*) HEL-SAMs cluster in the same DNMT5 clade as the diatom HEL-SAMs. Green algae and diatoms are not closely related but gene transfers from green algae to diatoms during evolution resulted in genomes of at least some diatom (i.c. *P. tricornutum* and *T. pseudonana*) having almost 16% of their genome of green algal origin (MOUSTAFA *et al.* 2009). It is possible that the diatom *HEL-SAM* genes have a green algal origin. The phylogenetic analysis shows no similarity with the classical phylogeny of diatoms based on e.g. 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). Sex loci are known to evolve relatively rapidly (HAAG and DOTY 2005; BERGERO and CHARLESWORTH 2009; KOESTLER and EBERSBERGER 2011) and so, a discrepancy between the phylogeny of HEL-SAM proteins of the diatom species and the classical phylogeny of diatoms is possible as genes used for classical phylogeny studies (rDNA genes) are more conserved and evolve relatively slow (CAMPO and GARCIA-VAZQUEZ 2012). The Sig proteins (identified in *Thalassiosira weissflogii*) have a role in sexual reproduction as they play a role in mediating sperm-egg recognition during the sexual reproduction phase and are also evolving relatively rapidly (SORHANNUS 2003). The HEL-SAM proteins of the two *Thalassiosira* species are in different clades and *Fragilariopsis cylindricus*, which belongs to the Bacillariales clade, together with *Pseudo-nitzschia*, suggesting loss of phylogenetic signal in the HEL-SAM. The HEL-SAM of *P. tricornutum* is situated in a different clade compared to the proteins of the other diatoms, which can indicate a loss of functionality of the HEL-SAM in *P. tricornutum*, for which sexuality has never been demonstrated. The fact that the two different domains have identical phylogenies suggests that the helicase and the SAM domain are evolving together and are therefore probably both involved in MT determination in diatoms.

No synteny could be observed between the neighboring genomic regions of the *HEL-SAM* genes between the different diatom species, except for the two *Pseudo-nitzschia* species. It seems that synteny is lost between different genera. The loss of the synteny indicates that the *HEL-SAM* gene acts as a sole MT determinant concerning the *HEL-SAM* genes as the real MT determinants in the pennate species.

The results of the first expression analysis of the *HEL-SAM* in *P. multistriata* and *S. robusta* were similar. Transcription in the MT^+ is upregulated in comparison to transcription in MT^- for the *HEL-SAM* in those two species. Notwithstanding the fact that a whole subset of genes was differentially expressed between the two MT's, these results at least do not contradict that the HEL-SAM in *P. multistriata* may also be involved in MT determination in this species.

This chapter described the first preliminary phylogenetic analysis of the HEL-SAM proteins in diatoms. Future research should include a further phylogenetic and functional analysis of the diatom HEL-SAM's which will inform about the origin and the evolution of the MT locus in diatoms and about the evolution from homothallic to heterothallic reproduction.

Acknowledgements

The authors acknowledge the Ghent University Research Funding (BOFGOA 01G01911 and BOF BOF09/24J/103) and the JGI for performing the transcriptome analysis for *P. multistriata*.

Literature cited

Altschul, S., T. Madden, A. Schaffer, J. H. Zhang, Z. Zhang *et al.*, 1998 Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Faseb Journal 12**:** A1326-A1326.

Armbrust, E. V., J. A. Berges, C. Bowler, B. R. Green, D. Martinez *et al.*, 2004 The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306**:** 79-86.

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

Bowler, C., A. E. Allen, J. H. Badger, J. Grimwood, K. Jabbari *et al.*, 2008 The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature 456**:** 239-244.

Bozarth, A., U. G. Maier and S. Zauner, 2009 Diatoms in biotechnology: modern tools and applications. Appl Microbiol Biotechnol 82**:** 195-201.

Brown, J. W., and U. Sorhannus, 2010 A molecular genetic timescale for the diversification of autotrophic stramenopiles (Ochrophyta): substantive underestimation of putative fossil ages. PLoS One 5.

Bullard, J. H., E. Purdom, K. D. Hansen and S. Dudoit, 2010 Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11**:** 94.

Campo, D., and E. Garcia-Vazquez, 2012 Evolution in the block: common elements of 5S rDNA organization and evolutionary patterns in distant fish genera. Genome 55**:** 33-44.

Chepurnov, V. A., and D. G. Mann, 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., D. G. Mann, K. Sabbe and W. Vyverman, 2004 Experimental studies on sexual reproduction in diatoms, pp. 91-154 in *International Review of Cytology*, edited by K. W. Jeon. Elsevier Academic Press, San Diego.

Chepurnov, V. A., D. G. Mann, P. von Dassow, P. Vanormelingen, J. Gillard *et al.*, 2008 In search of new tractable diatoms for experimental biology. Bioessays 30**:** 692-702.

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

De Riso, V., R. Raniello, F. Maumus, A. Rogato, C. Bowler *et al.*, 2009 Gene silencing in the marine diatom *Phaeodactylum tricornutum*. Nucleic Acids Research 37**:** e96.

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

Granum, E., J. A. Raven and R. C. Leegood, 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.

Haag, E. S., and A. V. Doty, 2005 Sex determination across evolution: connecting the dots. PLoS Biol 3**:** e21.

Hall, T. A., 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41**:** 95-98.

Koestler, T., and I. Ebersberger, 2011 Zygomycetes, microsporidia, and the evolutionary ancestry of sex determination. Genome Biol Evol 3**:** 186-194.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan *et al.*, 2007 Clustal W and clustal X version 2.0. Bioinformatics 23**:** 2947-2948.

Lebeau, T., and J. M. Robert, 2003 Diatom cultivation and biotechnologically relevant products. Part I: cultivation at various scales. Appl Microbiol Biotechnol 60**:** 612-623.

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

Lommer, M., M. Specht, A. S. Roy, L. Kraemer, R. Andreson *et al.*, 2012 Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome Biol 13**:** R66.

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

Mann, D. G., V. A. Chepurnov and S. J. M. Droop, 1999 Sexuality, incompatibility, size variation, and preferential polyandry in natural populations and clones of *Sellaphora pupula* (Bacillariophyceae). Journal of Phycology 35**:** 152-170.

Maumus, F., P. Rabinowicz, C. Bowler and M. Rivarola, 2011 Stemming epigenetics in marine stramenopiles. Curr 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, pp. 93-118 in *The Diatom World*.

Medlin, L. K., W. H. C. F. Kooistra, R. Gersonde, P. A. Sims and U. Wellbrock, 1997 Is the origin of the diatoms related to the end-Permian mass extinction? Nova Hedwigia 65**:** 1-11.

Medlin, L. K., W. H. C. F. Kooistra and A.-M. M. Schmid, 2000 A review of the evolution of the diatoms – A total approach using molecules, morphology and geology., pp. 13-35 in *The Origin and Early Evolution of the Diatoms: Fossils, Molecular and Biogreographical Approaches*, edited by A. Witkowski and J. Sieminska.

Moustafa, A., B. Beszteri, U. G. Maier, C. Bowler, K. Valentin *et al.*, 2009 Genomic footprints of a cryptic plastid endosymbiosis in diatoms. Science 324**:** 1724-1726.

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

Roshchin, A. M., 1994 *Zhiznennye Tsikly Diatomovykh Vodoroslej*. Naukova Dumka, Kiev, Ukraine.

Scala, S., and C. Bowler, 2001 Molecular insights into the novel aspects of diatom biology. Cell Mol Life Sci 58**:** 1666-1673.

Siaut, M., M. Heijde, M. Mangogna, A. Montsant, S. Coesel *et al.*, 2007 Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. Gene 406**:** 23-35.

Sims, P. A., D. G. Mann and L. K. Medlin, 2006 Evolution of the diatoms: insights from fossil, biological and molecular data. Phycologia 45**:** 361-402.

Sorhannus, U., 2003 The effect of positive selection on a sexual reproduction gene in Thalassiosira weissflogii (Bacillariophyta): results obtained from maximum-likelihood and parsimony-based methods. Mol Biol Evol 20**:** 1326-1328.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei *et al.*, 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., M. Ashworth, E. Ruck, T. Nakov and R. K. Jansen, 2010 A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. Plant Ecology and Evolution 143**:** 278-296.

Van de Peer, Y., and R. De Wachter, 1997 Construction of evolutionary distance trees with TREECON for Windows: accounting for variation in nucleotide substitution rate among sites. Comput Appl Biosci 13**:** 227-230.

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

Supplementary information

Supplementary data 1.

>P. tricornutum_HEL-SAM

MPPAATRTMASPSNEDRSDQDAFVGEWIDKYFVELDEVYRGRIVSQSAHSTNTHTNS
NTDTHTSALWKVYYPADPTKSYPDSDVEELDWDEVQTGIALYRQRNGKEKENGKT
KTFATRTSPRTPSSNQATSPRTFVSSATSTSISTTSASSTSSTTTTSSTTSLPKVRRTSRR
RSVTQSSYRDEASVESADDETYRTARRSPRTRTSPPRALDAQLAVEPPSKRSRTAVSR
STTRRSSPTSSRSTALPCAFAVLMGARPAPRKARTAKDSVANAPSHHPSSAPEHGPLL
VPPTRTQTQTPTTSHAVEGNPMADGSRFVKKPYTAGDDLPVLAQPHAMFDDLVANL
TDHGRNIHVLWPLVETFRSRTLRVATMCSGTEAPVLALDLLQTSLRHALRRHAADEL
RTRGIDVANVLRIEHVFSCEIEPFKQAYIERNFRPPLLFRDIRELGQPQAYTAYGALRD
VPNRPGCVDVLVAGTSCVDYSNLNNQKKHIDQKGESGQTFHGMMDWVDLAQPPIVI
IENVSGAPWEIKVKMFEERGYAATFLRIDTKEYYIPQTRKRGYLFAIKAQTKNKVVD
RPARWTAAVKSLKRPASAALDDFMLPNDDPRVLRGRARLTAESSAGEGEGRTRTVD
WTKCETNHQQARSMEELGDKRPLTNWSDSGNTAMPSFGWNEWTNAQVHRIHDLM
DINALRLAKLMIDCSHKTMVWNLSQNVHRDTMGVLGLSQCLTPTGVFYVANRGGPL
VGEELLMLQGIPAEDLLLTKESEANLKDFAGNGMSTTVVGTVMLCALLHGHDILRG
GNEREASSAVVPSLVPRSITAPSDASISLEFAHYAKEKLELGPRSSLGREDWSKLLSAA
SSSSKKCITEGKYESLPPGELVTCQECGFTSSKSNAVPPRKYEEHHFVPNAGSNPRVHP
ADLREELLKLLPMRAQAQNFDLDTLSPLDTVSKKLWQEWKVAVKATLYESDATGAI
FLFKEITRTHIWTAHFGSRNGGRLEMRISKSKITWLLFAKPPSEKGELRDILSRPVARL

CVRAPNDSTVPVTLLTGSWELCLPETIACTILVEGSGDTIPSWRNRLGLKGGFETERQF
SRLKVTLESSVLPDLKAVIDGVYAALPQCGGACGSLRKKEKRRDTEPDLFFFLESGRK
SLPGDDTYIFSPTCHRTSNGEYRETLLVVDKAADFDPLHSGVTPPLYRKIVKTFVPGR
WIVLENAFLEANPPCSSRLDLTTMSSCADTQKTLRIGKGSWQAAPKILSCDIPSSDAK
DIQEACQMAGGLVEVNLQKSSNVFQQLSFVLSRLKIPSLFSENCQRWLELGISDTTPL
NKKVNGEIECCQRCAPQKPHVKWRHVWKGKSLKFSPIEDGQEAAVYERAVKQRPQ
AWLVRLETKKSHSLHLQIACNAVSLAYRAFGRFPKNSLSRDSIVGSHINEALKRCSFE
WRVVAHVEKRQPIFPTLVFRSNKRDVQAEQPPTFSKYMLRKEQLRSLAWMLKQEST
MEPYYEEEVTEAILPGLEWRAESRVRRPILARGGIIADEVGYGKTAITLGLIDSSQRIN
GDCPEPPPAYRDRLIQTKATLIIVPEHLMGQWPEEVQKFLGRSKRVIEIKSMASMNKIT
VEDIQKADIVVASFQILSNETYFLNLAEFSGVNASALPSNKNGGRHFDSLYNDCLEGL
LVRVPHLIGDTRKVFSEIRQDAECHESANAEENFRVDGKKSAYKNGSKSSKMKAPET
SKIGRERDPWGLSTSKVKASYRKMSCPPLELFFWNRLVVDEFTYLEDNKRHRALSFIL
GVKSSYRWLLSGTPKHSNFDDIQSLSSLLGIHLGIDESLPGVKMNRSRGVVAEETTGL
ESLSLYLEMHSMQWHERRHVQAQSFLDQFVRQNKAEHDEIPWEEHLIFVDLPPVERA
IYLELETHLKSLDMNKNAQKTKRKSTGDRDNRMQRILQDSASAEEALLKCCSHFNM
SSEAATALETITDIIKLRDTQKKELERDIVVYLASAFRQQHRILQHQSDWLLVSRSEKG
EVASALQQYLREVEKRDSVTHGADDEVHDCILQLVRQAEEAFHADPSRIDSFFDVDE
GDDPQEGSSPKRRRGASPQKSKKEAAEKFTERLFAMKIQLRDHLHLVRSMGKELCGR
VRSLRYVQWIRKFQDASIRFTCGHCRATGLESDQVGVLSSCGHVGCLGCLRVESAAE
KCVEYPSCRARVSSAHVVSSRHLGLHRADSSGGRYGRKLTVLVEKVREIIAMGDRMI
VFCQFDDLKEKIRQVLLENGVPSLEVAGSVHRQIASLRVFQKEIPGPTDPRVLVLKMD
DEQSAGLNLTHLNHALFVHPLLALSRAEYDAYETQAIGRIRRFGQTKTVHLHRFLAR
NTMDMEIWEERTKPRA

Supplementary data 2.

```
  L.pneumophila DNMT  WFDDARTKNLMESEYLNLIVNAIPPK----  26

    H. sapiens DNMT2  TIEGITLEEFDRLSFDMILMSPPCQPFTRI  30

 T. pseudonana DNMT2  PIERITQQELESHSAIVWCMSPPCQPHTRQ  30

P. tricornutum DNMT2  RIEKLTMNQIFEYRADIWMMSPPCQPHTRQ  30
```

```
         H. sapiens DNMT1  TNSRGQRLPQKGDVEM-LCGGPPCQGFSGM 29
        A. thaliana MET1  TEEQKSTLPLPGQVDF-INGGPPCQGFSGM 29
        A. thaliana CMT1  DGFKSHLLPLPGTVYT-VCGGPPCQGISGY 29
        A. thaliana CMT3  LGYKSGILPLPGGVDV-VCGGPPCQGISGH 29
        A. thaliana CMT2  SGFKSKILPLPGRVGV-ICGGPPCQGISGY 29
     T. pseudonana DNMT3  SIGNLVVVEHDSVAEA-VCSSHHNNESVAS 29
       H. sapiens DNMT3a  YGLLRRREDWPSRLQM-FFANNHDQEFDPP 29
    P. tricornutum DNMT3  YNDDSEAHSVS------ILHTHP-DYIQGI 23
        A. thaliana DRM1  WVGKNKLAPLDADEMEKLLGFPRDHTRGGG 30
        A. thaliana DRM2  WVGKNKAAPLEPDEMESILGFPKNHTRGGG 30
     M. pusilla 10G04840  WDTTFRTRPPRVAVEVGCGTGYV------- 23
     M. pusilla 17G01890  PTDPATVECLPPDADV-LAASVVCAEQEGS 29
     M. pusilla 02G06210  -MNPSRPGLRDGMRTA-LCS---------- 18
     M. pusilla 02G04600  ARDVAEVRDLPEETEL-LAAGFPCPDVSTS 29
   P.tricornutum HEL-SAM  LVPPTRTQTQTPTTSHAVEGNPMADGSRFV 30
    O. lucimarinus DNMT5  LAEHLPVIADIGGMFADLVSRVP------- 23
          O. tauri DNMT5  FEDLVSRVP--------------------- 9
     T. oceanica HEL-SAM  FVSPVGIDATDGIVVG-IIRGQV-RKVGKL 28
P. multistriata HEL-SAM  YVDPVGVDPTHGIVER-IVSDQV-RKVGGL 28
 P. multiseries HEL-SAM  YVDPVGVDPTHGIVER-IVSDQV-RKVGGL 28
   T. pseudonana HEL-SAM  FLDPCGMEATDSIIDR-LVGQQL-DKIGGL 28
      S. robusta HEL-SAM  YLDPCGMDATDDIIGR-LIGDQI-DKVAGL 28
 F. cylindricus HEL-SAM  FLDPCGMEATDDIIGG-LVGRQV-DKVGLL 28

      L.pneumophila DNMT  ------------------SSILDLGCGTGE 38
        H. sapiens DNMT2  GRQG--DMTDSRTNS--FLHILDILPRLQP 56
     T. pseudonana DNMT2  HSNQHKELEDPRSKS--FLHLCHVLSVMEP 58
    P. tricornutum DNMT2 HSNQDQELEDPRSRS--FLHLCDLLLELPP 58
       H. sapiens DNMT1 NRFN--SRTYSKFKNSLVVSFLSYCDYYRP 57
       A. thaliana MET1 NRFN--QSSWSKVQCEMILAFLSFADYFRP 57
       A. thaliana CMT1 NRYRNNPLEDQKNQQ--LLVFLDIIDFLKP 57
       A. thaliana CMT3 NRFRNLPLEDQKNKQ--LLVYMNIVEYLKP 57
       A. thaliana CMT2 NRHRNVPLNDERNQQ--IIVFMDIVEYLKP 57
     T. pseudonana DNMT3 YHWMKTELEGSIDAIMAKYGLNAYRQGVHS 59
      H. sapiens DNMT3a KVYP--PVPAEKRKP---IRVLSLFDGIAT 54
    P. tricornutum DNMT3 VQPFRTQQDPSLKHG---LTVLDMFAGIGT 50
       A. thaliana DRM1 ISTTDRGNSFQVDTVAYHINVLSLFTGIGG 60
       A. thaliana DRM2 MSRTERGNSFQVDTVAYHINVLSLFTGIGG 60
     M. pusilla 10G04840 ------IASAALLASASGIVPGDAVPGDG- 46
     M. pusilla 17G01890 WRESWSSVASKLKQVLRLLKAGRGGQGGAV 59
     M. pusilla 02G06210 -------------------HIFRLLRRTRV 29
     M. pusilla 02G04600 NLSR--PGLRHGTQTSLVSHVFRLLERRRV 57
  P.tricornutum HEL-SAM KKPYTANLTDHGRNIHVLLRVATMCSGTEA 60
    O. lucimarinus DNMT5 ------GMENFLTAMKRPLRVATMCSGTES 47
         O. tauri DNMT5 ------GIEAFLNTIKRPLRVATMCSGTES 33
     T. oceanica HEL-SAM LQSQSLDSERELGELPSLVKLNTACSGTDA 58
P. multistriata HEL-SAM LQNVRFQKQKDESELSYPIRLQTACSGTDA 58
  P. multiseries HEL-SAM LQNVRVKSNKDENELSYPIRLQTACSGTDA 58
  T. pseudonana HEL-SAM LQRA--LSGKAIGTNFNPLVLGTACSGTDA 56
      S. robusta HEL-SAM WSRALA--KITLGSAETPLRLGTACSGTDA 56
  F. cylindricus HEL-SAM LQRALSTGTAALGSVHNPLKLGTACSGTDA 58

      L.pneumophila DNMT PIAQFFIEKGFKLTGVDGS----------- 57
       H. sapiens DNMT2 KYILLENVKGFEVSSTRDL---LIQTIENC 83
     T. pseudonana DNMT2  CLILLENVVGFERSWMSACSTQQSQESEPS  88

    P. tricornutum DNMT2  KLIFLENVVGFESSQSCRK---WNTILQSR  85

       H. sapiens DNMT1  RFFLLENVRNFVSFKRSMVLKLTLRCLVRM  87

       A. thaliana MET1  RYFLLENVRTFVSFNKGQTFQLTLASLLEM  87

       A. thaliana CMT1  NYVLMENVVDLLRFSKGFLARHAVASFVAM  87

       A. thaliana CMT3  KFVLMENVVDMLKMAKGYLARFAVGRLLQM  87

       A. thaliana CMT2  SYVLMENVVDILRMDKGSLGRYALSRLVNM  87

     T. pseudonana DNMT3  YMLRFAKLVTLIRKHQQKQQLYFLCENVP-  88

      H. sapiens DNMT3a  GLLVLKDLGIQVDRYIASE----VCEDS--  78

    P. tricornutum DNMT3  ATVCLKRLGLQISKIVRVEDHISTHVY---  77

       A. thaliana DRM1  GEVALHRLQIKMNVVVSVE--ISDANRNIL  88

       A. thaliana DRM2  GEVALHRLQIKMKLVVSVE--ISKVNRNIL  88

     M. pusilla 10G04840  GDGEIDRVQNAAAAAAAAF-AYYATDINP-  74

     M. pusilla 17G01890  PWVLVEASTALLESDGEGVVCDLVSELERL  89

     M. pusilla 02G06210  PWVLLENVPGLLTWHLASDIAYIVQELESL  59

     M. pusilla 02G04600  PWVLLENVPGLLMWHHKDDIAYVADELERL  87

   P.tricornutum HEL-SAM  PVLALDLLQTSLRHALRRHEHVFSCEIEPF  90

    O. lucimarinus DNMT5  PLLALDKIGDATEKLYGTRDHVFSCEIEPF  77

          O. tauri DNMT5  PLLALDKIADATQKLYGTRDHVFSCEIEPF  63

     T. oceanica HEL-SAM  PSIDEEGGKNVHRFD----EHNMSCEIEPF  84

 P. multistriata HEL-SAM  PSIALGLIKESLARISSKSEHMMSCEIEPF  88

  P. multiseries HEL-SAM  PSIALGLIKESLSRLCFKSEHMMSCEIEPF  88

   T. pseudonana HEL-SAM  PALALTLVQEQLEARGLGHDHVFSCEIEPY  86

      S. robusta HEL-SAM  PALAMTMVLEQMELRRRDNSHEFSCENDPM  86

  F. cylindricus HEL-SAM  PSLALGMVHEYLSKRGLPDEHQYSCEVEPF  88


      L.pneumophila DNMT  -----QKMIELCRKRFPDERWIVSDMRDIN  82
```

```
        H. sapiens DNMT2  GFQYQEFLLSPTSLGIPNSRLRYFLLQSEP 113
     T. pseudonana DNMT2  GYQVGHFHLDPTHFRLPNNRPRYYCVAFRR 118
    P. tricornutum DNMT2  QYIIKHFHLNPTQVGVPNDRPRYFCLAVRS 115
        H. sapiens DNMT1  GYQCTFGVLQAGQYGVAQTRRRAIIAPGEK 117
        A. thaliana MET1  GYQVRFGILEAGAYGVSQSRKRAFIAPEEV 117
        A. thaliana CMT1  NYQTRLGMMAAGSYGLPQLRNRVFLQPSEK 117
        A. thaliana CMT3  NYQVRNGMMAAGAYGLAQFRLRFFLLPSEI 117
        A. thaliana CMT2  RYQARLGIMTAGCYGLSQFRSRVFMVPNKN 117
     T. pseudonana DNMT3  -QRWDDQGSIETCFGIPAKRKRSYYIPSNN 117
       H. sapiens DNMT3a  -ITVGMVRHQGKIMYVGDVRSVTQKHIQEW 107
    P. tricornutum DNMT3  -QENHDCSYNPTLADHGDIKHVY-CMAETH 105
        A. thaliana DRM1  RSFWEQTNQKGILREFKDVQKLDDNLMDEY 118
        A. thaliana DRM2  KDFWEQTNQTGELIEFSDIQHLTNDLMEKY 118
     M. pusilla 10G04840  -DALATTRATLRNHGIADERVELTR-GDLL 102
     M. pusilla 17G01890  GYRWAHRTIAAAAFGVPDVKPRVVLLASKH 119
     M. pusilla 02G06210  GYSWAHRVVSLSAFGLPQRRRRVFIVASLH 89
     M. pusilla 02G04600  GYRWAQRVICLTGMGLPQMRRRVFVIASTH 117
  P.tricornutum HEL-SAM  KQAYIERNFRPPLL-FRDIRELGQPAYTAY 119
   O. lucimarinus DNMT5  KQAYIERNFAPPIL-FRDIRELGGDATTAY 106
         O. tauri DNMT5  KQAYIERNFAPPIL-FRDIRELGGDATTAY 92
    T. oceanica HEL-SAM  KQGYIGRNFPGVLL-FPDITKLADGVTDV- 112
P. multistriata HEL-SAM  KQAYIARNFPGTPL-FPDITKLSAIVIDVY 117
 P. multiseries HEL-SAM  KQAYIARNFPGVPL-FPDITKLSAIVVDVY 117
  T. pseudonana HEL-SAM  KQAYLARNFDSVLY--PDIVKLCDEPRDVY 114
      S. robusta HEL-SAM  KQAYLARNFDSKLY--PDIARLCDTPRDVF 114
  F. cylindricus HEL-SAM  KQSYIARNFDGILY--PDIAKLTEEPRDVY 116
```

```
     L.pneumophila DNMT  LQQQ---FDVILAWHSFF-HLDHDSQRNMF  108

       H. sapiens DNMT2  LPFQAPGFPKIESVHPQKYAMDVENKIQEK  143

    T. pseudonana DNMT2  GSLQARLDNIRMGFNVKKHNLFTIESLSNE  148

   P. tricornutum DNMT2  TEIHDSNFHVHAKTKMADSDLRPITPDTNL  145

       H. sapiens DNMT1  LPLFP--EPLHVFAPRAC-QLSVVVDDKKF  144

       A. thaliana MET1  LPEWP--EPMHVFGVPKL-KISLSQGLHYA  144

       A. thaliana CMT1  LPPYPLPHEVAKKFNTPK-EFKDLQVGRIQ  146

       A. thaliana CMT3  IPQFPLPHDLVHRGNIVK-EFQG-----NI  141

       A. thaliana CMT2  LPPFPLPHDVIVRYGLPL-EFE-----RNV  141

    T. pseudonana DNMT3  LPDSHSPVCLADGWMTPS-QMHCYLQGVRT  146

      H. sapiens DNMT3a  GP-----FDLVIGGSPCN-DLSIVNPARKG  131

   P. tricornutum DNMT3  GP-----FDLVIGGPPCV-DYSAVNARRQG  129

       A. thaliana DRM1  GG-----FDLVIGGSPCN-NLAGGNRHHRV  142

       A. thaliana DRM2  GG-----FDLVIGGSPCN-NLAGGNRVSRV  142

    M. pusilla 10G04840  GPLRERLIDLLLFNPPYV-----LTPSEEV  127

    M. pusilla 17G01890  GDPRDVLTEDAGKSANFL-EPPGAETSQTF  148

    M. pusilla 02G06210  GDPRDVISTESICKGQCI-DVSKTDGDKRS  118

    M. pusilla 02G04600  GDPRDVLSAQAVCFGQCI-ELNCRAGNKRQ  146

  P.tricornutum HEL-SAM  GALRDVPVDVLVAGTSCV-DYSNLNNQKKH  148

   O. lucimarinus DNMT5  GAKINVPVDMLVAGTSCV-DYSNLNNERKG  135

         O. tauri DNMT5  GAKVRVPVDMLVAGTSCV-DYSNLNNERKG  121

    T. oceanica HEL-SAM  -------SNLFVAGTSCK-DFSMLKNTKRK  134

P. multistriata HEL-SAM  GRPQSIPGDLFVAGTSCK-DFSMLKTTKRK  146

  P. multiseries HEL-SAM  GRPQNIPGDLFVAGTSCK-DFSMLKTTKRK  146

  T. pseudonana HEL-SAM  GQEKPLPFNMFVAGTSCK-NFSMLMSNKRI  143

     S. robusta HEL-SAM  GQEQPLPANMFVAGTSCK-DFSTMRSKKRK  143

  F. cylindricus HEL-SAM  GRIVPLPINYFVAGTSCK-NFSMLMSKFRL  145
```

```
      L.pneumophila DNMT  KIFESHTKPGGVL--------------AFT  124

       H. sapiens DNMT2  NVEPNISFDGSIQSGKDAILFKLETAEEIH  173

    T. pseudonana DNMT2  PVIH-TEKSIQEYHVVPTIPNIKSFLDSLQ  177

   P. tricornutum DNMT2  PNLNIKGLRDSTVKVSSVAEFLDKDLTPQS  175

       H. sapiens DNMT1  VSNI-TRLSSGPF-RTITVRDTMSDLPEVR  172

       A. thaliana MET1  AVRS-TA-LGAPF-RPITVRDTIGDLPSVE  171

       A. thaliana CMT1  MEFL-KLDNALTL------ADAISDLPPVT  169

       A. thaliana CMT3  VAYD-EGHTVKLA-DKLLLKDVISDLPAVA  169

       A. thaliana CMT2  VAYA-EGQPRKLE-KALVLKDAISDLPHVS  169

    T. pseudonana DNMT3  PV-----CKSLTF--LASSSKIDDERMVKV  169

      H. sapiens DNMT3a  LYEG----TGRLFEFYRLLHDARPKEG---  154

   P. tricornutum DNMT3  AEGV-QGRYTIEFHLIRKLERLQNPHPLFY  158

       A. thaliana DRM1  GL---GGEHSSLF-----------------  152

       A. thaliana DRM2  GL---EGDQSSLF-----------------  152

    M. pusilla 10G04840  RAGG--------------------------  131

    M. pusilla 17G01890  VFNQSAEGEMRVF-----SDFVDGFHP---  170

    M. pusilla 02G06210  AAEC-YDCFMTPPHVTPKISVTCVDLAEIC  147

    M. pusilla 02G04600  QKDVITGNTTTPTAENGGGGDDDEEKD---  173

  P.tricornutum HEL-SAM  IDQ--KGESGQTF--HGMMDWVDLAQPPIV  174

   O. lucimarinus DNMT5  LEA--QGESGQTF--RGMMNWIRKSQPPII  161

         O. tauri DNMT5  LDA--QGESGQTF--RGMMNWIRKSQPSII  147

    T. oceanica HEL-SAM  DIED-KGQSGQTF--IAAVEYLEQELPKFA  161

P. multistriata HEL-SAM  DIQD-KGQSGETF--LAAVEFLDLYQPPFA  173

 P. multiseries HEL-SAM  DIQD-KGQSGETF--LAAVEFLDLYQPPFA  173

  T. pseudonana HEL-SAM  DIED-MGCSGETF--LAACEVLFKEKPQFC  170

     S. robusta HEL-SAM  DIED-RGCSGETF--LAACEVILKEQPTIC  170
 F. cylindricus HEL-SAM DIED-KGCSGETF--MGAVEVLLKTNPKIC 172


     L.pneumophila DNMT SGEE-EGEVW--------------SDNGGQ 139

      H. sapiens DNMT2 RKNQQDSDLSVKMLKDFLEDDTDVKSLLRY 203

    T. pseudonana DNMT2 IPEKVRMSSWCFDLVTPYHLRSSCRGTGSI 207

   P. tricornutum DNMT2 ILQ--RNAAWCFDIVTPESLRSACKGTGSV 203

      H. sapiens DNMT1 NGASALEISYNGEPQSWFQRQLRGRDHICK 202

      A. thaliana MET1 NGDSRTNKEYKEVAVSWFQKEIRGTDHICK 201

      A. thaliana CMT1 NYVANDVMDYDAAPKTEFENFISLFDHQPL 199

      A. thaliana CMT3 NSEKRDEITYDKDPTTPFQKFIRLYDHHPL 199

      A. thaliana CMT2 NDEDREKLPYESLPKTDFQRYIRSHDHRPF 199

    T. pseudonana DNMT3 KIDRN----W--IVKGYYSVADREPGELAL 193

     H. sapiens DNMT3a -----DDRPWENVVAMGVSDKRDI------ 173

   P. tricornutum DNMT3 LVENVFGIDWDPVVLDSLYLSPCRRSCLEE 188

      A. thaliana DRM1 ---------FC-RILEAVRRKARH------ 166

      A. thaliana DRM2 ---------FC-RILEVVRARMRG------ 166

    M. pusilla 10G04840 -----IAAAWGKDGREVVDRLLPDPDGGTM 156

    M. pusilla 17G01890 ----------G-DGGAVMDAAGHLLQGLPP 189

    M. pusilla 02G06210 TLTASNGRRLC-MVEDLGEGKGRAVGWTEP 176

    M. pusilla 02G04600 ------------EGDDAKEEEVMCTGGGGS 191

  P.tricornutum HEL-SAM IIENVSGAPWE-IKVKMFEERGYAPQTRKR 203

   O. lucimarinus DNMT5 ILENVCNAPWD-QVQKKFEDEGYHPHTRTR 190

         O. tauri DNMT5 ILENVCNAPWD-QVQKKFEDEGYHPHTRTR 176

    T. oceanica HEL-SAM IFENVQNAPWN-KMAEYITGRVDLKPDADN 190

P. multistriata HEL-SAM IFENVDGAPWG-KMQEYIQGRIYLKTNADE 202

  P. multiseries HEL-SAM IFENVDGAPWG-KMQEYIQGRIHLKNNADE 202

  T. pseudonana HEL-SAM IFENVTGAPWG-KMGEYITGRIKLRTIKKE 199
```

```
     S. robusta HEL-SAM ILENVVGAWWK-KMSEYIVGRIQLKGPKEL 199

F. cylindricus HEL-SAM IFENVIGAPWK-KMAEYITGRIKIKGGGGD 201


      L.pneumophila DNMT QLYHASLSTKEYESLLKNSSFKVLVHKVCD 169

      H. sapiens DNMT2 ALLLDIVQPTCRRSVCFTKGYGSYIEQTAE 233

   T. pseudonana DNMT2 LYYGQLRLGDSAVGDESNSQWSTV--DLKH 235

  P. tricornutum DNMT2 LYTGPYRDRIRLTNP--------------- 218

      H. sapiens DNMT1 DMSALVAARMRHIPLAPGSDWRDLPNSCVE 232

      A. thaliana MET1 AMNELNLIRCKLIPTRPGADWHDLPKVTLS 231

      A. thaliana CMT1 VLGDDDLERVSYIPKQKGANYRDMPGKAEI 229

      A. thaliana CMT3 NLNINDYERVCQVPKRKGANFRDFPGVKLE 229

      A. thaliana CMT2 HINEDDYARVCQIPNRKGANFRDLPGVCRD 229

   T. pseudonana DNMT3 HLFIVDDVRVLLSLVRTQLIFTFKVNSLFE 223

     H. sapiens DNMT3a --PVMIDAKEVSAAHRARYFWGNLPGTVND 201

  P. tricornutum DNMT3 GFSLPAHIVDGETTAKASCFMASSTRLRMY 218

      A. thaliana DRM1 ------------------------------ 166

      A. thaliana DRM2 ------------------------------ 166

   M. pusilla 10G04840 LMILLEQNKPREVM------------AVLE 174

   M. pusilla 17G01890 GWTLTSRPPAPGSRVDNAPRWKALAACVPA 219

   M. pusilla 02G06210 CFPLKIPGRPWMRAANTQSSYIKRMESVPQ 206

   M. pusilla 02G04600 GGKCSHPPRECYKCFMTPPF------CVPK 215

 P.tricornutum HEL-SAM GYLFAIKAQTKNKVVDRPARWTAAVKAALD 233

  O. lucimarinus DNMT5 VYLFACKASDASLVE----AWKGMVKATLE 216

      O. tauri DNMT5 VYLFACKADDPKLVE----MWKERVKATLE 202

   T. oceanica HEL-SAM KLKFIVDEDGKYVANDVPKQVGIRVGEVVA 220

P. multistriata HEL-SAM DLVFIVNDKGRYEAKAIPPQVGMKAGDIVQ 232

 P. multiseries HEL-SAM DLVFVVNGDGRYEAKSIPPQVGMKAGDIVE 232
```

```
    T. pseudonana HEL-SAM GPLDFVLLDGNIVTNYVPTTVGIRCGAAIA 229

       S. robusta HEL-SAM KFSFDNKQGNKLVVEEVPKMWGVHCGSTVA 229

   F. cylindricus HEL-SAM LELERDTNTNKIQVTKVPNNVGVRCGSIVK 231


       L.pneumophila DNMT PECGEATVWV-------------------- 179

        H. sapiens DNMT2 DYKSLTNLSQFGFPPEFGFPEKITVRYRLL 263

     T. pseudonana DNMT2 DEDWSAAIDWSSFPQSCAMKQ----QWKLL 261

    P. tricornutum DNMT2 -EDRKFDDAWFGFPSTFSFPETITRQWKLI 247

        H. sapiens DNMT1 AAGLYGRLEWVGFPDTYRLFGNILDKHRQV 262

        A. thaliana MET1 DKGLYGRLDWLGFPDSYEFAGNINHKHRQI 261

        A. thaliana CMT1 NKKPFGRLWGLGFPDCYKLCGTIKEKYIQV 259

        A. thaliana CMT3 ECKPFGRLWWLGFPDDYKLFGPPKQKYIQV 259

        A. thaliana CMT2 PKRPFARLWWLGFPDYFQFCGTIKERYCQI 259

     T. pseudonana DNMT3 KSAAFRSEDWVGKPLYYFDREEY--AKRLI 251

       H. sapiens DNMT3a KHGRIAKFSKLGFPVHYTDVSNMSRRQRLL 231

    P. tricornutum DNMT3 VGEREAMMGYLGFEEHILKMTPPQTSKHLI 248

        A. thaliana DRM1 ------------------------------ 166

        A. thaliana DRM2 ------------------------------ 166

     M. pusilla 10G04840 RCDVVRSASA-------------------- 184

     M. pusilla 17G01890 SFEAPVTVPWVGGEDRIGGCVTKETVKALR 249

     M. pusilla 02G06210 AMSPYEIKFSW------------------- 217

     M. pusilla 02G04600 REKRSGPILN-------------------- 225

    P.tricornutum HEL-SAM DEGRTRTVDWMNAQVHRIHDLMDINSHKTM 263

    O. lucimarinus DNMT5 ADSTRANTDWLKAQTDRVLDLMDIDMHKTL 246

           O. tauri DNMT5 ADSTRANTDWLKAQTDRVLDLMDIDMHKTL 232

      T. oceanica HEL-SAM GLGNVSPLSA-------------------- 230

  P. multistriata HEL-SAM GSDVIPLLSE-------------------- 242
```

```
 P. multiseries HEL-SAM GSDVIALLSEI--------------RYCTH 248

   T. pseudonana HEL-SAM GSQRLREIEWM------------------- 240

     S. robusta HEL-SAM GMGKIHPVEWT----DTKNFGLPQTRQRTY 255

F. cylindricus HEL-SAM GDNTVYPVKWSCFETPVTYCCRDVKRQRTY 261


     L.pneumophila DNMT ----                           179

       H. sapiens DNMT2 GNSL                           267

    T. pseudonana DNMT2 GNSL                           265

   P. tricornutum DNMT2 GNSL                           251

       H. sapiens DNMT1 GNAV                           266

       A. thaliana MET1 GNAV                           265

       A. thaliana CMT1 GNAV                           263

       A. thaliana CMT3 GNAV                           263

       A. thaliana CMT2 GNAV                           263

    T. pseudonana DNMT3 GNSY                           255

      H. sapiens DNMT3a GRSW                           235

   P. tricornutum DNMT3 GLAF                           252

       A. thaliana DRM1 ----                           166

       A. thaliana DRM2 ----                           166

    M. pusilla 10G04840 ----                           184

    M. pusilla 17G01890 AAGW                           253

    M. pusilla 02G06210 ----                           217

    M. pusilla 02G04600 ----                           225

  P.tricornutum HEL-SAM VWNL                           267

   O. lucimarinus DNMT5 VWNL                           250

         O. tauri DNMT5 VWNL                           236

    T. oceanica HEL-SAM ----                           230
```

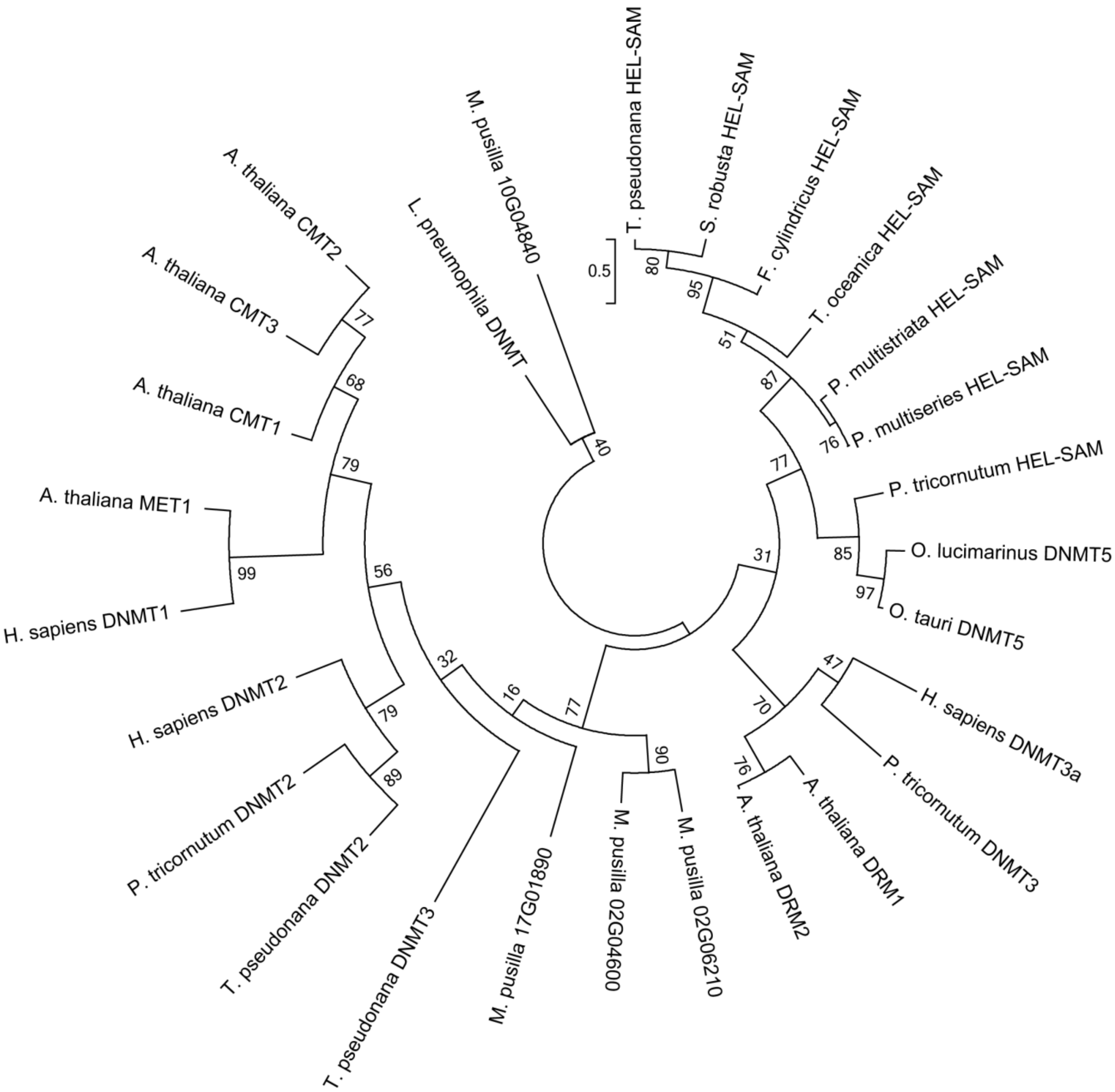
```
P. multistriata HEL-SAM ----                                              242

 P. multiseries HEL-SAM LCKI                                              252

   T. pseudonana HEL-SAM ----                                             240

      S. robusta HEL-SAM MFVW                                             259

 F. cylindricus HEL-SAM MLVW                                              265
```



Supplementary figure 1. ML phylogeny for the diatom HEL-SAM proteins and the selected DNMTs based on an alignment of the conserved domains (methyltransferases). Bacterial DNMTs were chosen as outgroup proteins.

5

General discussion

A better knowledge of the diatom life cycle is crucial for understanding the ecological importance of these major drivers of ocean biogeochemistry and primary production. A better understanding of their life cycle would also improve the use of diatoms in aquaculture and biotechnology, as there are still a lot of bottlenecks in diatom cultivation (LEBEAU and ROBERT 2003). Before the start of this thesis, almost nothing was known about the genetic and molecular nature of the mating type (MT) locus in diatoms. Life cycle traits were studied in only a minority of diatom species (CHEPURNOV *et al.* 2008). The aim of this study therefore was to identify and characterize the MT locus in the model diatom *Seminavis robusta* by generating genetic markers using AFLP (amplified fragment length polymorphism) technology and whole genome sequencing (WGS). In addition, we performed the first preliminary analyses of the evolution of the MT locus. The MT locus was mapped as a single segregating locus and identified as a SF2-family related helicase/S-adenosylmethionine dependent methyltransferase (*HEL-SAM*) gene. Homology and phylogenetic studies suggest that the *HEL-SAM* is the sole MT determinant and that the MT genomic region is evolving relatively rapidly. These findings represent a significant step forward in our understanding of the regulation of the life cycle, and more specifically sexual reproduction, in diatoms. Below, we will first discuss the importance of *S. robusta* as a model for studying sexual reproduction and MT determination. Then, we discuss the results obtained in the three results chapters. Finally, perspectives for future research and a general conclusion are provided.

General importance of *S. robusta* as a model diatom

We used the model diatom *Seminavis robusta* to study the genetic basis of MT determination in diatoms. *S. robusta* was successfully used in sexual reproduction studies in previous studies (CHEPURNOV *et al.* 2002; CHEPURNOV *et al.* 2004), because it is a heterothallic species (however, homothallic reproduction has also been observed very infrequently), and sexual crosses can be easily experimentally manipulated and synchronized, allowing full experimental control of the sexual process (CHEPURNOV *et al.* 2008). In other frequently used model diatoms like *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* (ARMBRUST *et al.* 2004; BOWLER *et al.* 2008a; CHEPURNOV *et al.* 2008), sexual reproduction, which is a key feature in most diatom life cycles (CHEPURNOV *et al.* 2004), has never been demonstrated, and thus, the use of *S. robusta* as a model where the sexual life cycle can be controlled is highly important. This led Chepurnov et al. (CHEPURNOV *et al.* 2008) to propose *S. robusta* as a model for studies of sexual reproduction and life cycle regulation in diatoms. Using *S. robusta*, the first known diatom pheromone was recently identified as a diproline (GILLARD *et al.* 2013). The synchronization and the easy manipulation of the sexual life cycle in this diatom (CHEPURNOV *et al.* 2008) made it possible to construct a full-sib (FS) mapping population (MP) of 116 individuals for *S. robusta* which formed the basis of the linkage maps and the identification of the MT locus (chapters 2 and 3). Other models where sexual reproduction can be controlled (e.g. *Fragilariopsis, Pseudo-nitzschia* and *Cylindrotheca*) are emerging but nothing about MT determination has yet been described. Whole genome information is available for *Fragilariopsis* and two *Pseudo-nitzschia* species and a transformation protocol is available for *Cylindrotheca fusiformis.* This will improve the importance of those species as model organisms (TESSON *et al.* 2013; VANORMELINGEN *et al.* 2013).

It is obvious that genetic studies of traditional and new model systems will make significant progress only if genomic approaches are used. Whole genome sequencing (WGS) and the generation of transgenic organisms, allowing reverse genetic approaches (SCALA and BOWLER 2001), are essential tools for dissecting basic biological processes. The current progress in the genome sequencing of *S. robusta* (a first draft genome sequence of *S. robusta* was constructed as part of the research described in this thesis) combined with the first successful application of a transformation protocol for *S. robusta* (Moeys S., personal communication, April 2013) will greatly improve the usefulness of *S. robusta* as a model organism in genetic studies.

The MT determining system as a single locus in *S. robusta*

In this research, we employed the high multiplex ratio of AFLP technology for the construction of the first linkage maps for a diatom species (chapter 2). These maps were successfully applied to demonstrate that MT determination in *S. robusta* is genetically regulated by a single locus with a Mendelian segregation pattern (VANSTECHELMAN *et al.*

2013). Our findings corroborate studies on most eukaryotic sexual organisms that show that sex is controlled by single loci, sometimes containing multiple genes, and segregate in a Mendelian way (CHARLESWORTH 2002; GOODENOUGH *et al.* 2007; LEE *et al.* 2010b). Only in the Basidiomycete fungi, sex loci can be present as two unlinked multi-allelic loci (METIN *et al.* 2010) and as a consequence, multiple MT's are present in these fungi. The presence of a single sex locus is considered to be the ancestral type of sex locus as other organisms besides the Basidiomycete fungi contain only single sex loci (METIN *et al.* 2010). Our findings confirmed previous hypotheses about the genetic nature of MT determination in diatoms (CHEPURNOV *et al.* 2004; DAVIDOVICH *et al.* 2010). The MT locus forms the basis to trigger the start of heterothallic sexual reproduction. Without a better understanding of the actual functional role of the MT genes, it is at present impossible to explain the infrequent occurrence of homothallic reproduction and hence mating type switching (to date only observed in MT$^+$ strains) in *S. robusta*. It is possible that this phenomenon is driven by a genetic, epigenetic or even by an environmental factor. Mating type switching is very well described in yeast. Here, mobile genetic elements driven by transposases can be captured by the MT locus and this results in MT switching (described in the budding yeast *Kluyveromyces lactis*) (RUSCHE and RINE 2010). In *Saccharomyces cerevisiae*, recombination events occur between the MT gene and other genes resulting in MT switching. The key enzyme enabling this process is a HO (homothallic switching) endonuclease (LEE *et al.* 2010b). It is possible that a similar event occurs during MT switching in *S. robusta*. Further research in *S. robusta* and in other species is needed to unravel the full mechanism of MT switching.

The MT phenotype in *S. robusta* cosegregates with markers of a MT$^+$ specific linkage group, which identifies the MT$^+$ as the heterogametic MT and the MT$^-$ as the homogametic MT in *S. robusta.* We further identified a large recombining region between the two homologous MT locus containing chromosomes. These chromosomes have extensive recombining pseudo autosomal regions (PAR) regions (in which both homologous chromosomes carry the same gene content) and small non-recombining regions. Such chromosomes have been hypothesized to represent recently evolved MT determining chromosome systems, in contrast with ancient MT determining chromosomes where one of the chromosomes is degenerated and large non-recombining regions exist (BERGERO and CHARLESWORTH 2009). Little information is available on the age of such non-recombining regions (BERGERO and CHARLESWORTH 2009). Studies on divergence times between the non-recombining regions of the MT determining chromosomes in *S. robusta* and other diatoms can contribute to our understanding of the evolution of MT determining chromosomes, as the diatoms, having an extensive fossil record, have a relatively well time-calibrated evolutionary record. The evolution from homothally (centrics) to heterothally (so far only reported in pennates (KACZMARSKA *et al.* 2013)) involves the evolution to genetic MT determination (CHEPURNOV *et al.* 2004; DAVIDOVICH *et al.* 2010). Most recent evidence to date suggests that the pennate clade, and therefore heterothally (and MT determining chromosomes) in diatoms, originated in the Late Cretaceous (75-90 MY ago) (SIMS *et al.* 2006; BOWLER *et al.* 2008b). In mammals and birds which have an ancient sex determining system, sex chromosomes evolved for a significant longer time and these chromosomes have small PAR regions and large non-recombining regions (BERGERO and CHARLESWORTH 2009). The origin of some sex chromosomes of these groups can now be dated using molecular clocks and ages of these sex chromosomes vary around 150 MY ago (HANDLEY *et al.* 2004; NAM and ELLEGREN 2008). A consensus is emerging that sexually antagonistic selection is the main selective driver of this reduced recombination (BERGERO and CHARLESWORTH 2009; NATRI *et al.* 2013), and that recombination suppression is driven by chromosome inversion or other chromosomal rearrangements (NATRI *et al.* 2013). Relatively recent MT determining systems like in *S. robusta* do not code for a whole subset of differential effects between the two mating types in comparison with what is been observed in organisms with ancient sex chromosomal systems (like birds and animals). Different sexes in these organisms are characterized by differences in gamete size, behavior, color, etc., and it is an advantage when such differential characteristics are inherited together to maintain antagonistic effects between sexes. As an example, the evolution from isogamy to oogamy (more differential information between MT's is needed because differences in gamete size and structure are present) in green algae involves the evolution to a larger non recombining region (UMEN 2011).

The *S. robusta* MT locus contains the single *HEL-SAM* gene

Bulked segregant analysis (BSA) is routinely and successfully used for the identification of quantitative trait loci (QTL) (MAGWENE *et al.* 2011; PARTS *et al.* 2011) and for the identification of DNA markers linked to a gene of interest (MICHELMORE *et al.* 1991). Here, we applied a BSA approach to MT bulks genetically dissimilar in the selected region (i.e. the MT locus) but genetically similar at all other regions. To further investigate the MT locus in *S. robusta*, BSA, in combination with AFLP and WGS was performed. The genomic sequence was used to identify genomic scaffolds covering the MT locus by (1) mapping

AFLP markers cosegregating with the MT locus and by (2) mapping genomic reads identifying significant differences between allele frequencies between the MT bulks. Two AFLP markers with the highest LOD score of linkage with the MT locus mapped on the same DNA scaffold. The BSA-WGS procedure made it possible to identify SNP's in (full) linkage disequilibrium with the MT locus, allowing the identification of additional genomic scaffolds flanking or potentially covering the MT region. The combination of both BSA methods enabled the identification of the MT locus in *S. robusta* as a SF2-family related helicase/S-adenosylmethionine dependent methyltransferase (*HEL-SAM*) gene with a transcript length of 7866 bp.

The identified helicase is a SF2-family related helicase. This family is involved in virtually all aspects of RNA and DNA metabolism (JANKOWSKY 2011). The helicase function (SF2 type) as part of the MT determining region has been reported in a few other organisms like the Zygomyceta and the Microsporidia (LEE *et al.* 2010a) and even in mammals (ABDELHALEEM 2005). Helicases unwind nucleic acid complexes in an ATP-dependent manner (JANKOWSKY 2011), regulate transcription and start cascade reactions regulating different molecular processes (VAQUERIZAS *et al.* 2009). Sexual development in different lineages is often regulated by transcriptional regulators (KOTHE 1996; KOESTLER and EBERSBERGER 2011; LEE *et al.* 2010b). The identification of the HEL-SAM suggests that MT determination may be (partly) epigenetically regulated (CHENG 1995). The combination of a SAM and a SF2-family related helicase domain has been described for the DNA methyltransferase 5 (DNMT5) protein family that has one member in the diatoms *P. tricornutum* and *T. pseudonana* (PONGER and LI 2005; MAUMUS *et al.* 2011). These genes are comparable in length and all consist of the same configuration as the *HEL-SAM* of *S. robusta*, suggesting that these genes are possibly orthologous.

Although a whole subset of genes was differentially expressed, a significant difference in expression level of the *HEL-SAM* was observed between the two MT's. This difference in expression may trigger the expression of downstream genes [e.g. MT specific pheromones (GILLARD *et al.* 2013)] important for MT determination differently, which would trigger a difference in function between the two MT's. The difference in transcription is independent of the size of the cells relative to the sexual size threshold (SST). It therefore appears that other factors are responsible for triggering the cell size dependent activation of sexual reproduction.

It is tempting to speculate that the *S. robusta* MT locus plays a role in MT determination by up- or downregulating the expression of one or more genes involved in sexual differentiation and development by a combined HEL-SAM action. It is possible that the helicase domain can make DNA more accessible for methylation by the methyltransferase domain. Allelic variation at the MT locus can result in structural differences of the protein between the two MT's and this can differentially affect downstream gene regulation resulting in MT differences in *S. robusta*.

Evolution of the MT locus in diatoms

Homologs of the MT gene of *S. robusta* (*HEL-SAM*) could be identified in all diatom species for which sequence information is available: a unique *HEL-SAM* gene was found in all examined centric and pennate diatoms (chapter 4). The function of this gene in centric diatoms, where MT determination has never been reported and its possible function (if any) in *P. tricornutum*, where sexual reproduction also has never been demonstrated (CHEPURNOV *et al.* 2008), remains unknown. Transcriptome data indicate that transcription of the *HEL-SAM* does occur in *P. tricornutum* (Matthijs M., personal communication, March 2013). However, it is possible that the protein is inactive or that downstream determinants are lacking. No ESTs are available for the *HEL-SAM* gene in *T. pseudonana*, what indicates that the *HEL-SAM* in this species is probably inactive. To further investigate this, a full functional analysis of this gene in these species is needed. We have very strong evidence that the HEL-SAM is the MT determining factor in *S. robusta* (Chapter 3). A preliminary transcriptional analysis showed that transcription of the *HEL-SAM* is higher in the MT^{+} in *S. robusta* and in *P. multistriata*. The expression difference in *P. multistriata* was more pronounced compared to the difference in *S robusta* (chapter 4). However, we need to be cautious when comparing MT determination in these two species. While in *Pseudo-nitzschia* species, active (migrating) and passive gametes are present (anisogamy) (CHEPURNOV *et al.* 2004), it is the MT^{+} (parental) cell in *S. robusta* that migrates to the MT^{-} cell, and no physiological differences between the gametes is present (isogamy) (GILLARD *et al.* 2013). Based on the expression difference in *P. multistriata*, we can speculate that the *HEL-SAM* has a potential MT determining role in this species as well. However, the transcriptional analysis showed that a whole subset of genes was differentially expressed in *S. robusta* and *P. multistriata* and thus that further transcriptional analyses and especially a detailed functional analysis is needed to prove that the HEL-SAM in these species is the real MT determinant. The presence of a unique homolog in the centric diatoms shows that an ancestral *HEL-SAM* gene already originated before MT determination (the evolution from homothally to heterothally) found its origin in diatoms. We can therefore postulate that in the pennate clade, the *HEL-SAM* gene was recruited as the MT determining factor.

The phylogenetic relationship between the diatom HEL-SAM's shows that they all strongly cluster in the same clade and that no similarity with the classical phylogenetic relationships of the diatoms (CHEPURNOV *et al.* 2008; MEDLIN 2011) exists, which also suggests that relatively rapid evolutionary changes occur in the MT determining locus. The *HEL-SAM* of *P. tricornutum* is the most divergent when compared to the *HEL-SAM'*s of the other examined diatoms which can indicate that a potential loss of functionality in this diatom has occurred. It needs to be emphasized that our phylogenetic analysis of the diatom HEL-SAM proteins is as yet very preliminary. The lack of phylogenetic signal needs to be confirmed by the construction of a calibrated tree including a comparison of the evolutionary substitution rate of the HEL-SAM with conserved proteins used in classical phylogenetic studies (e.g. 18S gene). This will inform us about the rate of evolution of the HEL-SAM gene.

Our study proved that synteny in the neighboring genomic regions of the MT locus between the different diatom genera is lost. This suggests that the *HEL-SAM* gene acts as the sole MT determinant in diatoms (provided that the HEL-SAM is indeed the MT determinant in all pennate species). In general, the MT locus is accompanied by suppressed recombination and suppressed recombination often evolves across most of the rest of the MT determining or sex chromosomes, rather than remaining restricted to the sex/MT-determining region (BERGERO and CHARLESWORTH 2009; HOOD *et al.* 2013). However, recombination hotspots can flank the MT region which promotes genetic exchange (HSUEH *et al.* 2006). The loss of synteny between different diatom genera confirms our findings that relatively high recombination frequencies occur in the MT genomic region of *S. robusta* which may be responsible for the distorted physical/genetic map ratio (chapter 3). Synteny studies of homologous chromosomes in wheat also proved that the synteny decreased with an increase in recombination rates along the average chromosome arm (AKHUNOV *et al.* 2003). These high recombination frequencies can also have an impact on MT switching, as has also been observed in *S. robusta*. The presence of recombination hotspots can serve to recombine two loci resulting in self compatibility (HSUEH *et al.* 2006). It is possible that the transition from heterothally to homothally in *S. robusta* occurs via such a mechanism.

Future perspectives

The *HEL-SAM* gene in *S. robusta* is the first MT locus identified in a diatom and in the Stramenopila lineage as a whole. Its identification will be an opportunity to further investigate MT loci in diatoms and in other related organisms. Overall, this will contribute to a better knowledge of the diatom life cycle and it will give insight into the different diatom reproductive strategies and their evolution.

The further improvement of the genome sequence combined with the application of a recently developed successful transformation protocol for *S. robusta* (Moeys S., personal communication, April 2013) will greatly stimulate the use of *S. robusta* as a straightforward genetic model for diatoms. The improvement of the genome sequence involves sequencing of larger insert DNA libraries, a further assembly resulting in larger scaffolds and annotation of the genome. Genetic transformation is essential in leveraging *S. robusta* research to a stage where functional reverse-genetic studies can routinely be performed to study the phenotypic and transcriptomic effects and cellular localization of transgenes (MONTSANT *et al.* 2005; POULSEN *et al.* 2006; SIAUT *et al.* 2007; DE RISO *et al.* 2009).

The next step in this research will be cloning and sequencing the two *HEL-SAM* alleles. These allelic sequence variants can then be expressed in transformable diatoms, e.g. in *P. tricornutum*. The sexual function in this species is lost but is it not clear whether the HEL-SAM or other factors are responsible. Expression of the *HEL-SAM* gene from *S. robusta* can inform about its functionality in this species. Reverse genetic studies with the *HEL-SAM* gene would be important to do a full functional analysis of this gene in *S. robusta* by overexpression and silencing studies. As an alternative to transformation in *S. robusta*, transformation in the heterothallic pennate diatom *Cylindrotheca fusiformis* (VANORMELINGEN *et al.* 2013), where a transformation protocol is already longer available, can be performed.

Analysis of transcriptome data have revealed higher expression of the *HEL-SAM* gene in the MT^+ compared to the expression in MT^- in *S. robusta* and *P. multistriata*. Further analyses into the transcription of the MT gene in *S. robusta* are in progress by a broader transcriptome analysis where cells of the two mating types, above and below the sexual size threshold (SST), are analysed in different cell cycle stages for cells in conditioned medium (medium added from the opposite MT) and in the control medium (Moeys S., personal communication, March 2013). Co-expression studies could be carried out using this database to identify putative regulatory or target genes of the *HEL-SAM*. Interesting expression patterns could be confirmed using RT-qPCR. Putative upstream regulators can be identified using the Yeast One Hybrid (Y1H) technique (REECE-HOYES and MARIAN WALHOUT 2012). Binding reactions of candidate regulators could be confirmed using chromatin immunoprecipitation (ChIP). Given the putative function of the HEL-SAM as a DNA methyltransferase and a chromatin remodelling enzyme, the involvement of epigenetic regulation in MT determination could be assessed. To find those potential DNA target sequences differentially bound by the HEL-SAM in MT^+ cells and MT^- cells, a ChIP assay combined with massively parallel DNA sequencing (ChIPseq) could be performed. Alternatively, since SAM proteins are known to catalyze the transfer of methyl groups on DNA, bisulphite (BS) sequencing of both MT's at different stages of sexual maturation could be performed to look for differentially methylated DNA regions (MEABURN and SCHULZ 2012), and to check whether these are differentially expressed in both MT's. Combining the results of the ChIPseq and BS sequencing could allow creating a shortlist of putative target genes involved in MT determination in diatoms. The progress in sequencing technology will provide more diatom genomes and transcriptomes where homologous sequences can be picked up what will give more information for performing phylogeny studies. Another strategy is to identify the homologs of the *S. robusta* MT locus in other diatoms by RACE technology (HIRANO 2004).

General conclusion

In this study, we further optimized the use of *S. robusta* as a model diatom, in particular to further investigate the role of sexual reproduction in the life cycle of diatoms. We demonstrated that the MT^+ in *S. robusta* is the heterogametic MT and that the MT determining region in this diatom segregates as a single locus. The MT determining chromosomes in *S. robusta* behave as recently evolved MT determining chromosomes where large recombining regions between the chromosomes are present. This recent evolution is in harmony with the fact that genetic MT determination in diatoms evolved around 75-90 MY ago (SIMS *et al.* 2006; BOWLER *et al.* 2008b). Further investigation of the MT locus identified a *HEL-SAM* gene as the MT determining factor in *S. robusta*. Homologs of this gene were found in the other sequenced diatom genomes, even in the diatoms where MT determination is environmental and/or the sexual life cycle is not described. We hypothesize that an ancestral HEL-SAM was already present before MT determination found its origin in diatoms and that this gene evolved as the sole MT determinant in the pennate lineage. It is essential that further research will be done to perform a full functional analysis of this *HEL-SAM* gene and that possible interactors will be identified. Functional analysis in other species will also prove that the identified homologous loci act as the MT determinants in these species. A full phylogenetic study will give further insights into the evolution of MT determination in diatoms.

Literature cited

Abdelhaleem, M., 2005 RNA helicases: regulators of differentiation. Clin Biochem 38**:** 499-503.

Akhunov, E. D., A. R. Akhunova, A. M. Linkiewicz, J. Dubcovsky, D. Hummel *et al.*, 2003 Synteny perturbations between wheat homoeologous chromosomes caused by focus duplications and deletions correlate with recombination rates (vol 100, pg 10836, 2003). Proceedings of the National Academy of Sciences of the United States of America 100**:** 14511-14511.

Armbrust, E. V., J. A. Berges, C. Bowler, B. R. Green, D. Martinez *et al.*, 2004 The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306**:** 79-86.

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

Bowler, C., A. E. Allen, J. H. Badger, J. Grimwood, K. Jabbari *et al.*, 2008a The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature 456**:** 239-244.

Bowler, C., A. E. Allen, J. H. Badger, J. Grimwood, K. Jabbari *et al.*, 2008b The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 456**:** 239-244.

Charlesworth, D., 2002 Plant sex determination and sex chromosomes. Heredity 88**:** 94-101.

Cheng, X., 1995 DNA modification by methyltransferases. Curr Opin Struct Biol 5**:** 4-10.

Chepurnov, V. A., D. G. Mann, K. Sabbe and W. Vyverman, 2004 Experimental studies on sexual reproduction in diatoms, pp. 91-154 in *International Review of Cytology,* edited by K. W. Jeon. Elsevier Academic Press, San Diego.

Chepurnov, V. A., D. G. Mann, P. von Dassow, P. Vanormelingen, J. Gillard *et al.*, 2008 In search of new tractable diatoms for experimental biology. Bioessays 30**:** 692-702.

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

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

De Riso, V., R. Raniello, F. Maumus, A. Rogato, C. Bowler *et al.*, 2009 Gene silencing in the marine diatom *Phaeodactylum tricornutum*. Nucleic Acids Research 37**:** e96.

Gillard, J., J. Frenkel, V. Devos, K. Sabbe, C. Paul *et al.*, 2013 Metabolomics enables the structure elucidation of a diatom sex pheromone. Angew Chem Int Ed Engl 52**:** 854-857.

Goodenough, U., H. Lin and J. H. Lee, 2007 Sex determination in Chlamydomonas. Seminars in Cell & Developmental Biology 18**:** 350-361.

Handley, L. J., H. Ceplitis and H. Ellegren, 2004 Evolutionary strata on the chicken Z chromosome: implications for sex chromosome evolution. Genetics 167**:** 367-376.

Hirano, M., 2004 RACE using only a gene-specific primer - Application of a template-switching model. Molecular Biotechnology 27**:** 179-186.

Hood, M. E., E. Petit and T. Giraud, 2013 Extensive Divergence Between Mating-Type Chromosomes of the Anther-Smut Fungus. Genetics 193**:** 309-315.

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

Jankowsky, E., 2011 RNA helicases at work: binding and rearranging. Trends in Biochemical Sciences 36**:** 19-29.

Kaczmarska, I., A. Poulickova, S. Sato, M. B. Edlund, M. Idei *et al.*, 2013 Proposals for a terminology for diatom sexual reproduction, auxospores and resting stages. Diatom Research 28**:** 263-294.

Koestler, T., and I. Ebersberger, 2011 Zygomycetes, microsporidia, and the evolutionary ancestry of sex determination. Genome Biol Evol 3**:** 186-194.

Kothe, E., 1996 Tetrapolar fungal mating types: sexes by the thousands. FEMS Microbiol Rev 18**:** 65-87.

Lebeau, T., and J. M. Robert, 2003 Diatom cultivation and biotechnologically relevant products. Part I: cultivation at various scales. Appl Microbiol Biotechnol 60**:** 612-623.

Lee, S. C., N. Corradi, S. Doan, F. S. Dietrich, P. J. Keeling *et al.*, 2010a Evolution of the sex-Related Locus and Genomic Features Shared in Microsporidia and Fungi. Plos One 5.

Lee, S. C., M. Ni, W. Li, C. Shertz and J. Heitman, 2010b The evolution of sex: a perspective from the fungal kingdom. Microbiol Mol Biol Rev 74**:** 298-340.

Magwene, P. M., J. H. Willis and J. K. Kelly, 2011 The statistics of bulk segregant analysis using next generation sequencing. PLoS Comput Biol 7**:** e1002255.

Maumus, F., P. Rabinowicz, C. Bowler and M. Rivarola, 2011 Stemming epigenetics in marine stramenopiles. Curr Genomics 12**:** 357-370.

Meaburn, E., and R. Schulz, 2012 Next generation sequencing in epigenetics: insights and challenges. Semin Cell Dev Biol 23**:** 192-199.

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

Metin, B., K. Findley and J. Heitman, 2010 The mating type locus (MAT) and sexual reproduction of Cryptococcus heveanensis: insights into the evolution of sex and sex-determining chromosomal regions in fungi. PLoS Genet 6**:** e1000961.

Michelmore, R. W., I. Paran and R. V. Kesseli, 1991 Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci U S A 88**:** 9828-9832.

Montsant, A., U. Maheswari, C. Bowler and P. J. Lopez, 2005 Diatomics: Toward diatom functional genomics. Journal of Nanoscience and Nanotechnology 5**:** 5-14.

Nam, K., and H. Ellegren, 2008 The chicken (Gallus gallus) Z chromosome contains at least three nonlinear evolutionary strata. Genetics 180**:** 1131-1136.

Natri, H. M., T. Shikano and J. Merila, 2013 Progressive recombination suppression and differentiation in recently evolved neo-sex chromosomes. Mol Biol Evol 30**:** 1131-1144.

Parts, L., F. A. Cubillos, J. Warringer, K. Jain, F. Salinas *et al.*, 2011 Revealing the genetic structure of a trait by sequencing a population under selection. Genome Res 21**:** 1131-1138.

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

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

Reece-Hoyes, J. S., and A. J. Marian Walhout, 2012 Yeast one-hybrid assays: a historical and technical perspective. Methods 57**:** 441-447.

Rusche, L. N., and J. Rine, 2010 Switching the mechanism of mating type switching: a domesticated transposase supplants a domesticated homing endonuclease. Genes Dev 24**:** 10-14.

Scala, S., and C. Bowler, 2001 Molecular insights into the novel aspects of diatom biology. Cellular and Molecular Life Sciences 58**:** 1666-1673.

Siaut, M., M. Heijde, M. Mangogna, A. Montsant, S. Coesel *et al.*, 2007 Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. Gene 406**:** 23-35.

Sims, P. A., D. G. Mann and L. K. Medlin, 2006 Evolution of the diatoms: insights from fossil, biological and molecular data. Phycologia 45**:** 361-402.

Tesson, S. V., C. Legrand, C. van Oosterhout, M. Montresor, W. H. Kooistra *et al.*, 2013 Mendelian inheritance pattern and high mutation rates of microsatellite alleles in the diatom Pseudo-nitzschia multistriata. Protist 164**:** 89-100.

Umen, J. G., 2011 Evolution of sex and mating loci: an expanded view from Volvocine algae. Curr Opin Microbiol 14**:** 634-641.

Vanormelingen, P., B. Vanelslander, S. Sato, J. Gillard, R. Trobajo *et al.*, 2013 Heterothallic sexual reproduction in the model diatom Cylindrotheca. European Journal of Phycology 48**:** 93-105.

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

Vaquerizas, J. M., S. K. Kummerfeld, S. A. Teichmann and N. M. Luscombe, 2009 A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10**:** 252-263.

6

Summary

Diatoms are well-known microbial eukaryotes belonging to the Stramenopila lineage. They are photosynthetic and important members of phytoplankton communities in oceans and lakes. Diatoms are the most speciose group of algae, major players in ocean geochemistry and responsible for about 20% of global primary production. They originated around 200 MY ago and evolved by primary and (most probably) secondary endosymbiosis events where cyanobacteria and red algae respectively invaded the eukaryotic host cell. The best known characteristic of diatoms is their beautiful cell wall made from silica. The two major diatom groups (centrics and pennates) can be distinguished based on their cell wall symmetry and reproductive modes. Besides their ecological importance, diatoms are also important from a biotechnological point of view, and are increasingly used in nanotechnology and lipid production. The unique diplontic life cycle in diatoms consists of a long vegetative multiplication stage, accompanied by gradual cell size reduction, and a short sexual cycle, characterized by cell size restoration via the development of a specialized cell, the auxospore. The capacity of cells to become sexualized is size dependent. The currently used diatom models *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* are characterized by an atypical diatom life cycle which lacks the sexual reproduction cycle and thus, the need of a new model diatom, with the typical diatom life cycle, was urgent. The normal type of reproductive behavior is well documented in the heterothallic pennate diatom *Seminavis robusta* and sexual reproduction in this species can be controlled reliably in laboratorium conditions. This makes *S. robusta* a perfect model organism in the study of sexual reproduction in diatoms. Sexual development is common in eukaryotic organisms and populations consist generally of two genders that are frequently genetically regulated by a pair of sex chromosomes. Separate sexes arose independently many times during the evolution of different organisms and as a consequence, different types of sex determining genes exist in different lineages.

The easy synchronization and manipulation of the sexual cycle in *S. robusta* made it possible to construct a full-sib (FS) mapping population (MP) of 116 individuals. AFLP (amplified fragment length polymorphism) markers segregating in this MP were used for the construction of mating type (MT) specific linkage maps, and these linkage maps were used to identify a single locus MT determination system in this diatom. These findings represent the first evidence for a genetic MT determining mechanism in a diatom. We identified 13 MT$^+$ and 15 MT$^-$ linkage groups obtained from the analysis of 463 AFLP markers segregating in the MP, covering 963.7 and 972.2 cM, respectively. Five linkage group pairs could be identified as putative homologs. The MT phenotype mapped as a monogenic trait, disclosing the MT$^+$ as the heterogametic MT. Homologous MT determining chromosomes identified a large recombining region between the two homologous chromosomes, suggesting the existence of a recently evolved MT determining system in this diatom. The MT determining genomic region was further investigated by bulked segregant analysis (BSA) in combination with AFLP technology and whole genome sequencing (WGS). This BSA procedure identified AFLP markers cosegregating with the mating type phenotype and SNP's (single nucleotide polymorphisms) in (full) linkage disequilibrium with the MT. Mapping of these AFLP markers and genomic reads containing these SNP's on the genome sequence identified a SF2-family related helicase/S-adenosylmethinine dependent methyltransferase (HEL-SAM) most likely responsible for MT determination in *S. robusta*. This *HEL-SAM* gene identified in *S. robusta* is the first MT determinant ever described in diatoms and even in the Stramenopila lineage as a whole. Homolog sequences for this MT determining factor in *S. robusta* are identified in different centric and pennate diatom species indicating that an ancestral *HEL-SAM* gene was already present before the divergence of the centric and the pennate lineage, while MT determination is only known in the pennate lineage. This indicates that the HEL-SAM in the pennate lineage was most probably recruited for a role as a MT determining factor. Expression analysis of the *HEL-SAM* for *S. robusta* and *Pseudo-nitzschia multistriata* showed that expression was upregulated in the MT$^+$, what can indicate that this gene triggers interactor genes responsible for MT differentiation differentially. Analysis of the genomic regions of the *HEL-SAM* in the examined species indicated that synteny of this genomic region is lost between different genera and thus, that the HEL-SAM acts as the sole MT determinant. A phylogenetic analysis indicated that the *HEL-SAM* genes probably evolved relatively rapid as no the HEL-SAM phylogeny is not congruent with classical phylogenies of diatoms.

Despite the fact that this thesis represents a first significant step forward in a better understanding of the sexual cycle of diatoms, it is clear that further research is needed to fully understand the role of the MT gene and the evolution of the MT locus in diatoms.

6

Samenvatting

Diatomeeën zijn microbiële eukaryoten die behoren tot de groep van de Stramenopila. Het zijn fotosynthetische organismen en belangrijke leden van fytoplankton gemeenschappen. Diatomeeën zijn de meest soortenrijke groep van algen. Ze zijn belangrijke spelers in oceaangeochemie en ze zijn verantwoordelijk voor ongeveer 20% van de primaire productie op aarde. Diatomeeën zijn geëvolueerd via primaire en secundaire endosymbiose waar cyanobacteriën en rode algen respectievelijk de eukaryotische cel binnendrongen. Diatomeeën zijn vermoedelijk ongeveer 200 miljoen jaar geleden ontstaan. De best gekende eigenschap van diatomeeën is hun mooie celwand gemaakt van silica en twee belangrijke diatomee groepen (de pennate en de centricate) kunnen worden onderscheiden op basis van hun celwandstructuren en wijze van voortplanting. Naast hun ecologisch belang zijn diatomeeëen belangrijk geworden voor biotechnologie toepassingen gezien ze voornamelijk in nanotechnologie en vetproductie worden gebruikt. De unieke diplontische levenscyclus van diatomeeëen bestaat uit een langdurende vegetatieve cyclus (gekenmerkt door een graduële reductie van de celgrootte) en een kortdurende sexuele levenscyclus (gekenmerkt door een herstel in celgrootte via de ontwikkeling van een gespecialiseerde cel, de auxospore). De capaciteit van cellen om de seksuele levenscyclus in te treden is grootte afhankelijk. De momenteel meest gebruikte modeldiatomeeën, *Phaeodactylum tricornutum* en *Thalassiosira pseudonana*, zijn gekenmerkt door een atypische levenscyclus die de seksuele cyclys ontbreekt. De verdere ontwikkeling van een nieuw model dat de typische levenscyclus van diatomeeën kenmerkte was noodzakelijk. Deze typische levenscyclus is goed gedocumenteerd in de heterothallische pennate diatomee *Seminavis robusta*. De seksuele levenscyclus in *S. robusta* kan makkelijk worden gereproduceerd en gecontroleerd in laboratoriumcondities. Dit maakt van *S. robusta* een perfect modelorganisme in de studie van seksuele reproductie in diatomeeën. Seksuele ontwikkeling komt algemeen voor in eukaryotische organismen en populaties kennen meestal twee verschillende geslachten, die vaak genetisch worden gereguleerd door een paar sekschromosomen. Geslachtsdeterminatie van organismen ontstond verscheidene keren onafhankelijk van elkaar tijdens de evolutie. Dit zorgde ervoor dat verschillende types van seksdeterminerende genen ontstaan zijn in verschillende groepen organismen.

De eenvoudige synchronisatie en manipulatie van de seksuele levenscyclus in *S. robusta* maakte het mogelijk een full-sib (FS) mapping populatie (MP) van 116 individuen te construeren. AFLP (amplified fragment length polymorphism) merkers segregerend in deze MP werden gebruikt voor de constructie van mating type (MT) specifieke koppelingsgroepen. Dit leidde tot de identificatie van een enkel MT locus in deze diatomee. We hebben 13 MT^+ en 15 MT^- specifieke koppelingsgroepen geïdentificeerd gebaseerd op 463 AFLP merkers die uitsplitsten in de MP. Deze koppelingsgroepen behielden 963.7 en 972.2 cM respectievelijk. Vijf paren koppelingsgroepen konden worden geïdentificeerd als mogelijke homologen. Het MT fenotype mapte als een monogenische eigenschap, wat de MT^+ als de het heterogamete MT onthulde. De homologe MT determinerende chromosomen identificeerden een grote recombinerende regio tussen de twee homologe chromosomen wat een recent geëvolueerd MT determinatie mechanisme in deze diatomee doet vermoeden. Deze bevindingen rapporteren over de eerste evidentie van een genetisch MT determinatiesysteem in diatomeeën. De MT determinerende genomische regio was verder onderzocht via bulked segregant analysis (BSA) in combinatie met AFLP technologie en whole genome sequencing (WGS). Deze BSA procedure genereerde AFLP merkers die cosegregeerden met het MT fenotype en SNP's (single nucleotide polymorphisms) in volledig koppelingsonevenwicht met MT. Mapping van deze AFLP merkers en genomische reads die deze SNP's bevatten op de genoomsequentie identificeerde een SF2-familie gerelateerd helicase/S-adenosylmethionine afhankelijk methyltransferase (HEL-SAM). Dit proteine is zeer waarschijnlijk noodzakelijk voor MT determinatie in *S. robusta*. Het *HEL-SAM* gen in *S. robusta* is de eerste MT determinant ooit beschreven in diatomeeën en zelfs breder, in de groep van de Stramenopila. Homologe sequenties voor deze MT determinerende factor in *S. robusta* werden geïdentificeerd in verschillende centricate en pennate soorten wat doet vermoeden dat een voorouderlijk *HEL-SAM* gen al bestond voor de divergentie van de centricate en de pennate soorten, gezien MT determinatie enkel aanwezig is in de pennate diatomeeën. Dit doet ook vermoeden dat het *HEL-SAM* gen in de pennate soorten evolueerde tot de MT determinerende factor. Expressie analyse van het *HEL-SAM* in *S. robusta* en *P. multistriata* toonde aan dat expressie opgereguleerd was in het MT^+, wat doet vermoeden dat dit gen interactors belangrijk in MT determinatie op een verschillende manier beinvloedt. Analyse van de genomische regio's van het *HEL-SAM* in de onderzochte species doet vermoeden dat de syntenie tussen de genomische regio’s van het *HEL-SAM* tussen verschillende genera zijn verloren geraakt. Dit doet veronderstellen dat het *HEL-SAM* als enkele MT determinerende factor ageert. De fylogenetische analyse toonde ook aan dat de HEL-SAM proteïnen van de verschillende soorten relatief snel evolueren en dat geen relatie kon gemaakt worden tussen de fylogenie van de HEL-SAM proteïnes en de klassieke fylogenie van de diatomeeën.

Deze thesis beschrijft een significante stap voorwaarts in het beter begrijpen van de seksuele cyclus van diatomeeën. Meer onderzoek is nodig om de rol van het MT gen en de evolutie ervan volledig te begrijpen.