UNIVERSITY OF CRETE

BIOLOGY DEPARTMENT

Identification and functional characterization of RNA Silencing key-genes in the model pennate diatom species *Phaeodactylum tricornutum*

Doctoral dissertation

Emilia Grypioti

Heraklion, 2019

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

Ανεύρεση και λειτουργική ανάλυση γονιδίων του μηχανισμού της Γονιδιακής Σίγησης στο διάτομο, οργανισμό μοντέλο, *Phaeodactylum tricornutum*

Διδακτορική διατριβή

Αιμιλία Γρυπιώτη

Ηράκλειο, 2019

Supervisor:

Kriton Kalantidis, Associate Professor, University of Crete, IMBB/FORTH Group Leader

Advisory Committee

Kriton Kalantidis, Associate Professor, University of Crete, IMBB/FORTH Group Leader

George Kotoulas, Researcher A, Institute for Marine Biology, Biotechnology and Aquaculture, HCMR

Kiriakos Kotzabasis, Professor, University of Crete

Examination Committee

Kriton Kalantidis, Associate Professor, University of Crete, IMBB/FORTH Group Leader

George Kotoulas, Researcher A, Institute for Marine Biology, Biotechnology and Aquaculture, HCMR

Kiriakos Kotzabasis, Professor, University of Crete

Panagiotis Sarris, Assistant Professor, University of Crete, IMBB/FORTH Group Leader

Paraskevi Pitta, Researcher, Institute of Oceanography, HCMR

Angela Falciatore, Researcher CNRS, Sorbonne Université, Paris

Sotiris Kampranis, Associate Professor, University of Copenhagen

This dissertation was conducted in the facilities of the Biology department, University of Crete under the laboratory supervision of Dr. Frederic Verret.

Part of the present work was conducted in the facilities of IMBB/FORTH, IMBBC/HCMR, Medicine School of University of Crete, LCQB lab of Sorbonne University, Paris and TAKUVIK lab in University of Laval.

This dissertation was funded by the project KRIPIS: "Marine Biology, Biotechnology & Aquaculture", funded by GSRT (General Secretariat for Research & Technology) within the framework of the Action entitled «Proposals for Development of Research Bodies-KRIPIS»-NSRF (Operational Programme II, Competitiveness & Entrepreneurship).

This dissertation was funded by the project “Centre for the study and sustainable exploitation of Marine Biological Resources (CMBR)” (MIS 5002670) which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

Στον πατέρα μου

To my father

Ευχαριστίες

Η διατριβή αυτή δε θα μπορούσε να πραγματοποιηθεί χωρίς την υποστήριξη αρκετών ανθρώπων, τους οποίους σε αυτό το σημείο θα ήθελα να ευχαριστήσω προσωπικά.

Πρώτα από όλους, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου δρ. Κρίτωνα Καλαντίδη αφενός για την εμπιστοσύνη που μου έδειξε δίνοντάς μου τη δυνατότητα να δουλέψω πάνω στο πεδίο που ήθελα τόσο πολύ, και αφετέρου για την υπομονή και την καθοδήγηση του κατά τη διάρκεια υλοποίησης αυτής της εργασίας, το ενδιαφέρον του και για την βοήθειά του όποτε χρειαζόταν.

Θέλω να ευχαριστήσω ιδιαίτερα τον εργαστηριακό επιβλέποντά μου δρ. Frederic Verret για την καθοδήγηση του κατά την εκπόνηση και συγγραφή αυτής της εργασίας και για τη στήριξη του, ακόμα και όταν οι συνθήκες ήταν ιδιαίτερα δύσκολες. Με έμαθε να δουλεύω μεθοδικά, στοχευμένα, με γνώμονα την πληρότητα και εγκυρότητα μεθόδων και αποτελεσμάτων, χωρίς να παύω ποτέ να θέτω νέα ερωτήματα. Το κοινό ενδιαφέρον μας στο πεδίο και η άψογη συνεργασία μας εύχομαι να οδηγήσει σε νέες κοινές επιστημονικές αναζητήσεις.

Τα μέλη της επταμελούς επιτροπής δρ. Γεώργιο Κωτούλα, δρ. Κυριάκο Κοτζαμπάση, δρ. Παναγιώτη Σαρρή, δρ. Σωτήρη Καμπράνη, δρ. Παρασκευή Πήττα και δρ. Angela Falciatore για την κριτική ανάγνωση του κειμένου και τα σχόλιά τους που καθόρησαν και βελτίωσαν την τελική μορφή αυτής της διατριβής.

Επιπλέον θα ήθελα να ευχαριστήσω τον δρ. Γεώργιο Κωτούλα που με υποστήριξε από την αρχή σε αυτή την προσπάθειά μου καθοδηγώντας με στο σωστό δρόμο και εξασφαλίζοντας μέρος της οικονομικής στήριξης που απαιτούσε η παρούσα εργασία. Ευχαριστώ για τις επιστημονικές και φιλοσοφικές συζητήσεις μας, το γνήσιο ενδιαφέρον σου, τη βοήθεια και τη φιλία σου που τόσο απλόχερα μου πρόσφερες.

Ευχαριστώ την δρ. Angela Falciatore που με φιλοξένησε στο εργαστήριο της LQCB, στο Πανεπιστήμιο της Σορβόννης στο Παρίσι. Οι γνώσεις και οι εμπειρίες που αποκόμισα, καθώς και οι σχέσεις που δημιούργησα είναι ανεκτίμητες. Ευχαριστώ όλα τα μέλη του εργαστηρίου και κυρίως τη δρ. Marianne Jaubert που με βοήθησε αμέτρητες φορές με συμβουλές, διευκρινήσεις, πειράματα και ότι άλλο χρειάστηκα.

Τον δρ. Σωτήρη Καμπράνη για την επίβλεψη του κατά τη διερεύνηση της βιοσύνθεσης τερπενίων σε διάτομα, που αποτελεί μέρος της παρούσας εργασίας . Τον ευχαριστώ θερμά για την υπομονή και την καθοδήγησή του στη διεξαγωγή της έρευνας, των πειραμάτων αλλά για τις εποικοδομητικές συζητήσεις μας που διεύρυναν τους ορίζοντές μου. Ευχαριστώ επίσης τα μέλη του εργαστηρίου του Αναστασία Αθανασάκογλου και Codruta Ignea για τις συμβουλές και την πολύτιμη βοήθειά τους στη διεξαγωγή των πειραμάτων.

Ευχαριστώ τον δρ. Guillaume Massé για τη φιλοξενία του στο εργαστήριο της ομάδας TAKUVIK, στο πανεπιστήμιο Laval, Quebec, για την ενδιαφέρουσα εκπαίδευσή μου αλλά και για τη μοναδική ευκαιρία να επισκεφτώ το ερευνητικό παγοθραυστικό πλοίο, Amundsen.

Θέλω να ευχαριστήσω θερμά όλα τα μέλη του εργαστηρίου μας στο πέρασμα αυτών των χρόνων για το όμορφο κλίμα συνεργασίας και αλληλεγγύης που είχαμε. Φοιτητές πτυχιακών, μεταπτυχιακών και διδακτορικών με τους οποίους μοιραστήκαμε γνώσεις, εμπειρίες, κοινά άγχη και φιλοδοξίες. Τη Νάντια Κατσαρού για τις εποικοδομητικές συζητήσεις μας και τη διάθεση που είχε πάντα να βοηθήσει.

Νιώθω πολύ τυχερή που σε αυτή τη διαδρομή εκτός από άψογους συνεργάτες απέκτησα και καλούς φίλους. Μια «σειρά» υποψήφιων διδακτόρων που βοηθήσαμε και στηρίξαμε ο ένας τον άλλον σε κάθε λογίς κατάσταση που αντιμετωπίσαμε, εντός και εκτός εργαστηρίου. Με αυτή τη βοήθεια, πρακτική και ψυχολογική, κατάφερα να ολοκληρώσω αυτή την εργασία και να αποκτήσω πολλές όμορφες εικόνες και εμπειρίες στη διαδρομή. Ρίτα Μέρμηγκα, Νικολέτα Κρυοβρυσανάκη, Γιάννη Βλατάκη και Τάσο Αναστασιάδη δεν ξέρω τι θα έκανα χωρίς εσάς… σας ευχαριστώ πάρα μα πάρα πολύ για όλα!

Να ευχαριστήσω επίσης τις φίλες Βίλυ Μιχαλοπούλου, Ανθή Γεωργοπούλου, Νατάσσα Τοματσίδου, Νατάσα Καμπουράκη, Δήμητρα Τσακιρέλη και Χρύσα Κοκοτίδου, από το ίδιο ή γειτονικά εργαστήρια, που στηρίξαμε και βοηθήσαμε η μία την άλλη.

Δεν βρίσκω αντάξια λόγια για να ευχαριστήσω όσο θα ήθελα την οικογένειά μου και ιδιαιτέρα τους γονείς μου που με στήριζαν πάντα, σε όλες τις επιλογές μου, ακόμα και σε εκείνες που δεν καταλάβαιναν. Χωρίς εσάς δεν θα είχα φτάσει ως εδώ και δε θα ήμουν ο άνθρωπος που έγινα. Σας ευχαριστώ για όλα.

Τέλος, θέλω να ευχαριστήσω το σύντροφο της ζωής μου, καλύτερο φίλο και στήριγμά μου Χρήστο, για την υπομονή, τη φροντίδα και την υποστήριξή του αυτά τα χρόνια. Με τη βοήθεια σου σε όλους τους τομείς, την αισιοδοξία σου και τα ανέκδοτά σου (not!) έκανες πιο εύκολη τη διαδρομή για την ολοκλήρωση αυτής της διατριβής. Στα καλύτερα που έρχονται…

Abstract

Gene silencing, also known as RNA interference (RNAi), is a conserved mechanism of regulation of gene expression mediated by small RNAs (sRNA), (Fire et al., 1998). Silencing of transgenes and endogenous genes following introduction of inverted repeats, antisense constructs and artificial miRNAs has been reported in diatom species, including the model species *P. tricornutum* (De Riso et al., 2009; Kaur and Spillane, 2015). The presence of an endogenous RNAi pathway has been suggested after comprehensive and combinatorial analyses of sRNAs, gene expression and DNA methylation in *P. tricornutum* (Veluchamy et al., 2013; Rogato et al., 2014). This RNAi pathway in *P. tricornutum* may play a role in the regulation of protein coding genes and TEs expression with possible consequences for the acclamatory response to nutrient limitation (Maumus et al., 2009). Homologues of the RNAi-key genes DICER (DCR), ARGONAUTE (AGO) and RNA-Dependent RNA polymerase (RDR) have been previously identified by *in silico* analysis (De Riso et al., 2009). However, the validation of their gene models, the characterization of their functions and the possible physiological role of RNAi in diatoms are still lacking.

In this study, extensive *in silico* analysis of genomic and transcriptomic information available in *P. tricornutum* suggests the presence of a single *PtDCR*, *PtAGO* and *PtRDR* coding gene. Mining and phylogenetic analysis of DCR, AGO and RDR homologues in diatoms from all publically available to date sequence datasets suggest an unanticipated diversification of the RNAi pathway in these organisms. *PtDCR/AGO/RDR* cDNA were cloned and splicing isoforms of *PtDCR* and *PtAGO* were identified. Subcellular localization of PtDCR-/AGO-/RDR-YFP was investigated by confocal microscopy. Functional characterization of *PtDCR* and *PtAGO* was first attempted by heterologous expression in the yeast *Saccharomyces cerevisiae* and the plant *Nicotiana bethamiana* hosts. In a second step, CRISPR/Cas9-mediated mutagenesis approach, recently developed in *P. tricornutum*, was successfully harnessed to generate PtDCR-KO and PtAGO-KO (KnockOut) lines. Growth phenotype of PtDCR-KO lines were investigated under optimal and nitrate depleted culture conditions and during recovery from UV-mediated stress. In parallel, mRNA and small RNAs whole transcriptome analyses were carried out. Culture experiments suggest that PtDCR may play a role in the response to

nitrate starvation. Transcriptomic analysis revealed that both sRNA and mRNA transcriptomes were affected in PtDCR-KO line. At the global scale, sRNA size distribution was found to shift towards larger fragment size in PtDCR-KO line. In addition, the abundance of sRNA mapped to TEs was found dramatically reduced in PtDCR-KO mutant and a concomitant increase in mRNA abundance of some TEs was observed. Interestingly, PtDCR-KO sRNA transcriptome also presented changes in tRNA-derived sRNA populations, suggesting a possible role of DCR in their processing in diatoms. TE mobilization has been proposed to play a pivotal role in diatom species diversification and capacity for adaptation to various environments. Taken together, our results indicate that the single DCR encoding gene present in *P. tricornutum* plays a major role in the production of TE-derived sRNAs and possibly TE mobilization, with important consequences for diatom acclamatory response and evolution.

Contents

1. Introduction

1.1. Diatoms

Diatoms (class Bacillariophyceae**)** are unicellular, photosynthetic phytoplankton that are dominant within both freshwater and seawater ecosystems. They form the basis of many food webs and have played a vital role in the global ecosystem for millions of years (Armbrust et al., 2004). Diatoms represent the largest group of phytoplanktonic organisms responsible for approximately 40% of marine carbon fixation and play a major contribution to the silica geochemical cycle through their cell wall construction (Battarbee, 1988; Sarthou et al., 2005; Seckbach and Kociolek, 2011). They generate a hard outer silica-based cell wall called the frustule which consists of two asymmetrical halves assembled like a Petri dish. The larger half is called the epitheca and the inner one the hypotheca. The physically strong frustule is offering protection since silica cannot be attacked enzymatically. During diatom mitotic division, two daughter cells are produced, with each cell keeping one of the two halves and forming a smaller half within it. This results in a population of continuously decreasing cell size. When a certain minimum size is reached, sexual reproduction involving the combination of male and female gametes produces a diploid auxospore to re-establish the original cell size. Frustules present a distinguish species specific silicon-based pattern design that is used also for their taxonomical classification.

Diatoms are currently classified within the Chromalveolata supergroup of eukaryotes, as a group of heterokonts (or Stramenopiles) which also include non-photosynthetic organisms such as the plant pathogens Phytophthora sp, and Plasmodium sp.

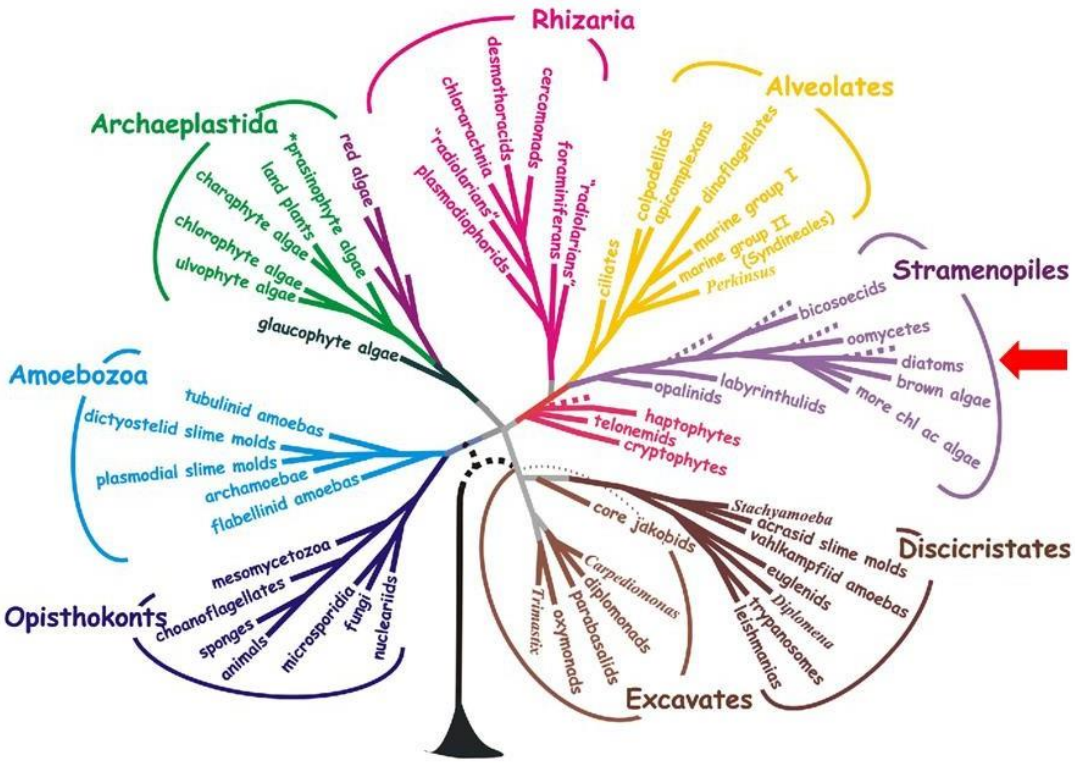

Figure 1. Eukaryote phylogenetic tree, adapted from Baldauf (2003). Red arrow points to Diatom group.

Diatoms are further classified based on their symmetry i.e. centrics (radial symmetry), and pennates (elongated) with or without a raphe involved in cell mobility (raphids or arapids, respectively). Diatom genome evolved via secondary endosynbiotic events between a non-photosynthetic host and red/green algae engulfment. As a consequence, diatoms exhibit a mosaic genome with genes orthologues to both animal and plant lineages (Moustafa et al., 2009; Bowler et al., 2010). Moreover, horizontal gene transfer from bacteria seems common among diatoms, since 5% of *P. tricornutum* genome presents bacterial homologues.

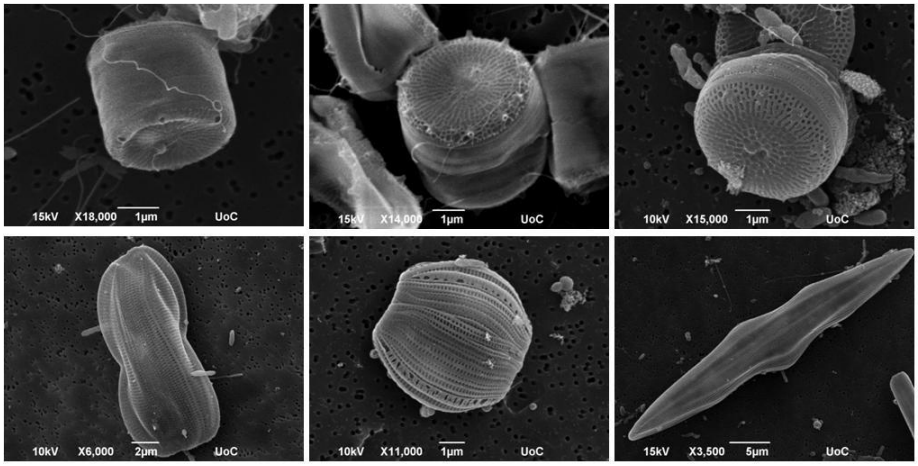

Figure 2. Electron microscopuy pictures (SEM) of diatoms grown at University of Crete within the frame of the PhD. Top row: centric diatoms of genus Thalassiosira. Second row (from left) the araphid pennates *Amphiprora sp., Entomoneis sp.* and the raphid pennate *Haslea ostrearia*.

Diatom plastids are surrounded by four membranes (Falciatore and Bowler, 2002), in contrast to three in the plant lineage. Novel gene combinations enabled the rise of novel metabolic traits. Diatoms can regulate uptake and storage of nitrogen and possess a urea cycle, a feature that they share with animals. Although animals excrete urea to remove excess nitrogen, diatoms possess an active urease and can grow with urea as the unique source of nitrogen. (Armbrust et al., 2004; Allen et al., 2006, 2011). The lipid content in certain diatom species can reach more than 60% of their total dry weight that in combination to their fast growth rates makes them potential candidates for biofuels production. Diatom mitochondria exhibit a partial mitochondrial glycolysis, only found in stramenopiles (Bártulos et al., 2018), and the bacterial Entner–Doudoroff pathway (Fabris et al., 2012). Apart from their conventional role in energy maintenance, diatom mitochondria are apparently tightly coupled metabolically and physically with the plastids by constantly shuttling energy and reducing equivalents between both organelles (Bailleul et al., 2015); Flori et al., 2017).

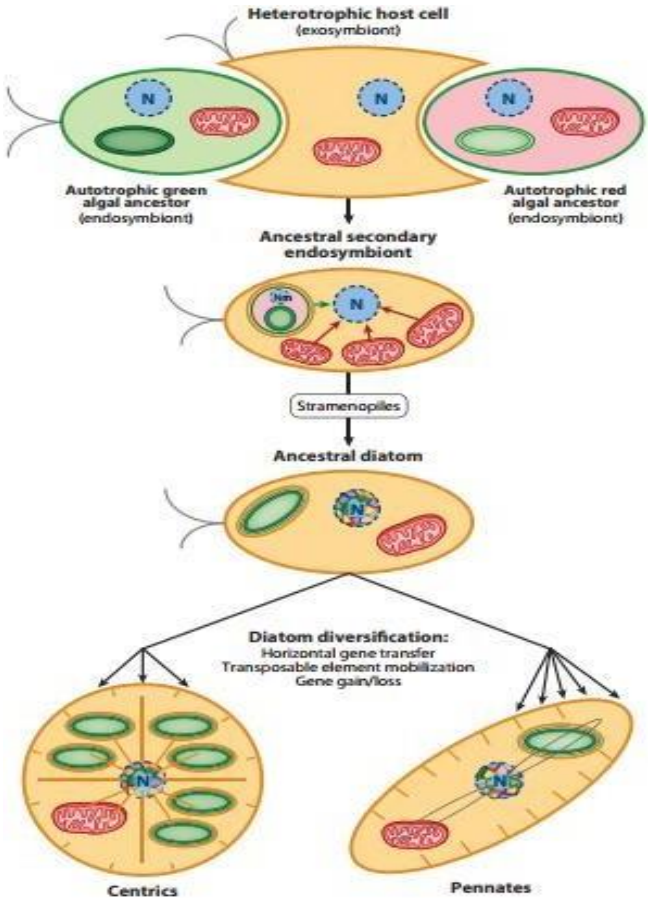

Figure 3. Schematic representation of diatom evolution, adapted from Bowler et al. (2010). The chimeric diatom genome finds its origins in successive gene transfers following endosymbioses between red and green algae and a host heterotrophic cell (Moustafa et al. 2009). Gene transfer from prey nuclei to host nucleus as well as from organelles to nucleus has been proposed. Acquisition of genes through lateral gene transfer both before and after the diversification of pennates and centrics. Abbreviations: N, nucleus; Nm, nucleomorph.

Their chimeric background has also resulted in their high divergence from other eukaryotic lineages, like Archeaplastida (plants, green algae, and red algae) and Opisthokonta (animals and fungi) (Armbrust et al., 2004). In addition, faster rates of divergence seem to occur even among heterokonts in comparison to other lineages. Indeed, Bowler et al. (2008) have shown that the pennate *P. tricornutum* shares only 57% of its genes with the centric diatom *T. pseudonana*. Based on fossil records, it has been proposed that pennate diatoms evolved from the centric forms, and the raphid pennates evolved latter on from the araphid pennates. The pennate diatoms are the most diversified and have habited both benthic (attached to the sea bottom) and pelagic niches (floating in

the water column). The formation of the raphe was probably a determining asset that led to subsequent diversification and although it likely evolved in benthic diatoms to glide on surfaces, many raphid pennates have been shown to be able to colonize planktonic environments.

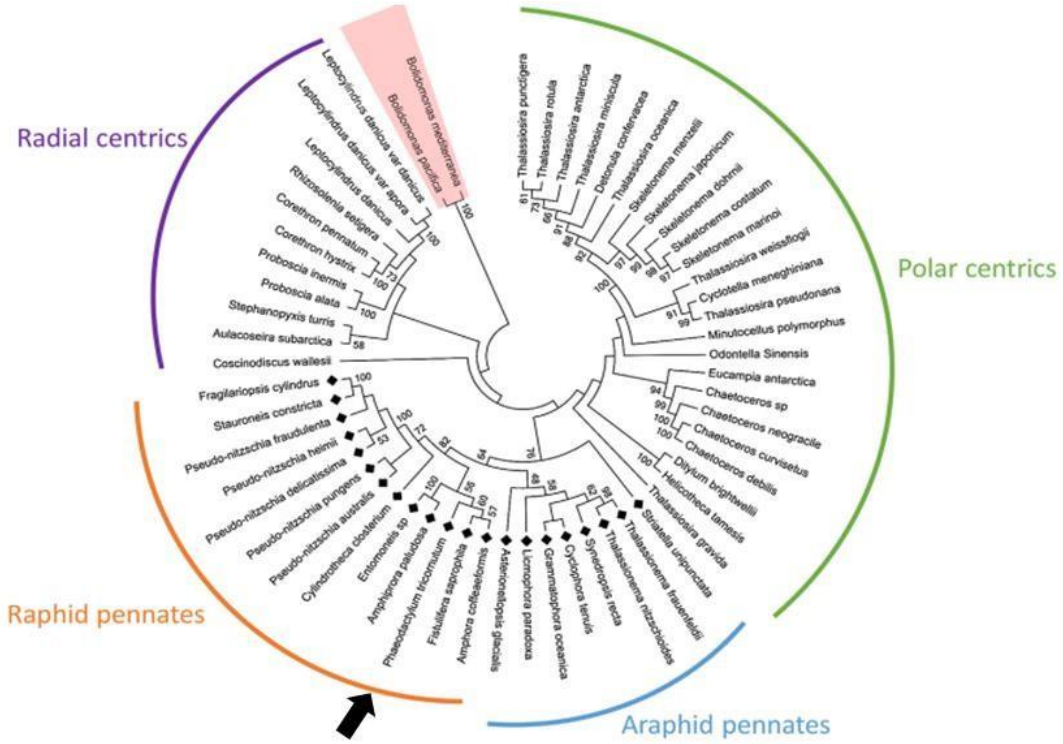

Figure 4. The evolutionary divergence of diatoms, adapted from Tirichine et al. (2017). Sequences from 57 diatom species were used to generate a phylogenetic tree using the Neighbor-Joining method. The tree was generated using 18S rRNA gene. Boxed in red color, the tree is rooted to Bolidophytes, sister lineage to diatoms. Black diamonds indicate pennate diatom species and others are centric diatoms. Black arrow points *P. tricornutum*. The tree is clustered into four subclasses of diatom, based on their valve morphology.

The apparent ecological success and dominance of diatoms has been proposed to result from their capacity to rapidly adapt to the continuously changing microenvironments of the ocean. They produce silicate cell walls, have a very efficient photosynthetic machinery producing lipids and an advanced photoprotection system to avoid UV damage. They possess a urea cycle and can regulate nitrate uptake. Pennate diatoms present a cyanobacterial ferritin-like gene involved in iron storage (Marchetti et al., 2008).

In addition to their ecological success, diatom present a wealth of applications in biotechnology (Bozarth et al., 2009; Athanasakoglou and Kampranis, 2019). Their fossils, known as diatomaceous earth, are used in insulation, abrasives and filtration (Maher et al., 2018; Pytlik and Brunner, 2018). Their ability to produce lipids via photosynthesis makes them a good source of renewable biofuels and of Omega-3 oils as dietary supplements. They are being used in nanotechnology and medicine, both for their silicate frustules and for their bio-active compounds. Diatoms are also used in water quality assessment and forensics based on their tolerance of environmental pressure and their robust exoskeleton that indicates classification and habitant, respectively. For all the above reasons, but also for understanding the distinct molecular mechanisms that evolved in diatoms and armed them with such a successfully adaptive system, studies are focusing on these organisms with a continuously increased interest.

Among diatoms the representative model species are the pennate *Phaeodactylum tricornutum* and the centric *Thalassiosira pseudonana.* A wealth of genomic information is available for these diatoms including sequencing data of their whole genome, messenger and small RNAs transcriptomes (Armbrust et al., 2004; Bowler et al., 2008). Comparative genomic studies have revealed little synteny and no major duplications events between centric *T.pseudonana* and pennate P. tricornutum. This suggests that mobilization of transposable element may be in part responsible for diatom diversification. As sequencing technologies have become more accessible to the scientific community, whole genome sequence information has been gained in additional diatom species including the pennate *Fragilariopsis cylindrus* (Mock et al., 2017) and, *Pseudonitzschia multiserie, Pseudonitzschia multistriata, Thalassiosira oceanica, Cyclotela cryptica, Fistulifera solaris*. A milestone in diatom transcriptomic information has been reached with the Marine Microbial Eukaryotic Transcriptome Sequencing project, MMETS, which has put together over 650 assembled, functionally annotated transcriptomes amongst which 92 diatom species (Keeling et al., 2014). The growing amount of diatom sequences represents a valuable resource for functional and comparative genomics studies that can elucidate evolutionary histories and reveal more metabolic peculiarities.

In this thesis, the diatom under study is the model pennate *Phaeodactylum tricornutum*. *P. tricornutum* has been used in laboratory-based studies of diatom physiology for several decades. *P. tricornutum* genomic and transcriptomic information is available and it is amenable to genetic transformation and genome editing.

1.2. RNA interference

1.2.1. The discovery of RNA interference

RNAi pathway was incidentally discovered in the development of new gene knockdown technique in animals and plants. This phenomenon was initially described in plants (Napoli et al., 1990; Lindbo et al., 1993) as co-suppression and homology induced resistance.

It had been shown that injecting antisense RNA into *Caenorhabditis elegans* could inhibit the expression of complementary mRNA (Fire et al., 1998). Fire et al., (1998) went further showing that double stranded RNA (dsRNA) was able to suppress gene expression even more efficiently than sense or antisense RNA alone. One year later, siRNAs were found for the first time in plants (Hamilton and Baulcombe, 1999). The term RNAi interference (RNAi) stemmed from these pioneering studies. While the initial model proposed that antisense RNA alone could complement mRNA and block translation, Andrew Fire and Craig Mello hypothesized that dsRNA was involved in a more complex and conserved mechanism working at the post-transcriptional level. Their contribution to unveiling the general principle of RNAi was acknowledged globally when they were awarded the Nobel Prize in Physiology or Medicine in 2006. Two years after their hypothesis Hammond et al. (2000) characterised the RNAi core molecular apparatus, the RNA Induced Silencing Complex (RISC). At the same period, plant biologists reported the presence of a phenomenon similar to RNAi which was coined post-transcriptional gene silencing (PTGS) (Jorgensen et al., 1996; Que and Jorgensen, 1998). A similar mechanism called quelling was proven to be present in fungi (Fulci and Macino, 2007; Nicolas et al., 2010). The presence of RNAi in animals, PTGS in plants and quelling, suggested the presence in their eukaryotic ancestor of conserved proteins involved in a RNA-based regulatory mechanism in (Mello and Conte, 2004).

1.2.2. The canonical RNAi pathway

RNAi is triggered by dsRNA processed by RNase-III-type endonucleases. These proteins comprise a family that includes DICER (DCR) and DROSHA which cleave the long dsRNAs at specific points into smaller RNAs. The long dsRNAs are referred to as precursors and small RNAs produced by DCR-like proteins as the mature sequence. Then, one strand of the sRNA is incorporated into a protein complex called the RNAi silencing complex (RISC), along with the Argonaute (AGO) protein. AGO bound sRNA guides the RISC to its mRNA target and prevents its translation either by translational arrest or transcriptional cleveage (Carthew and Sontheimer, 2009).

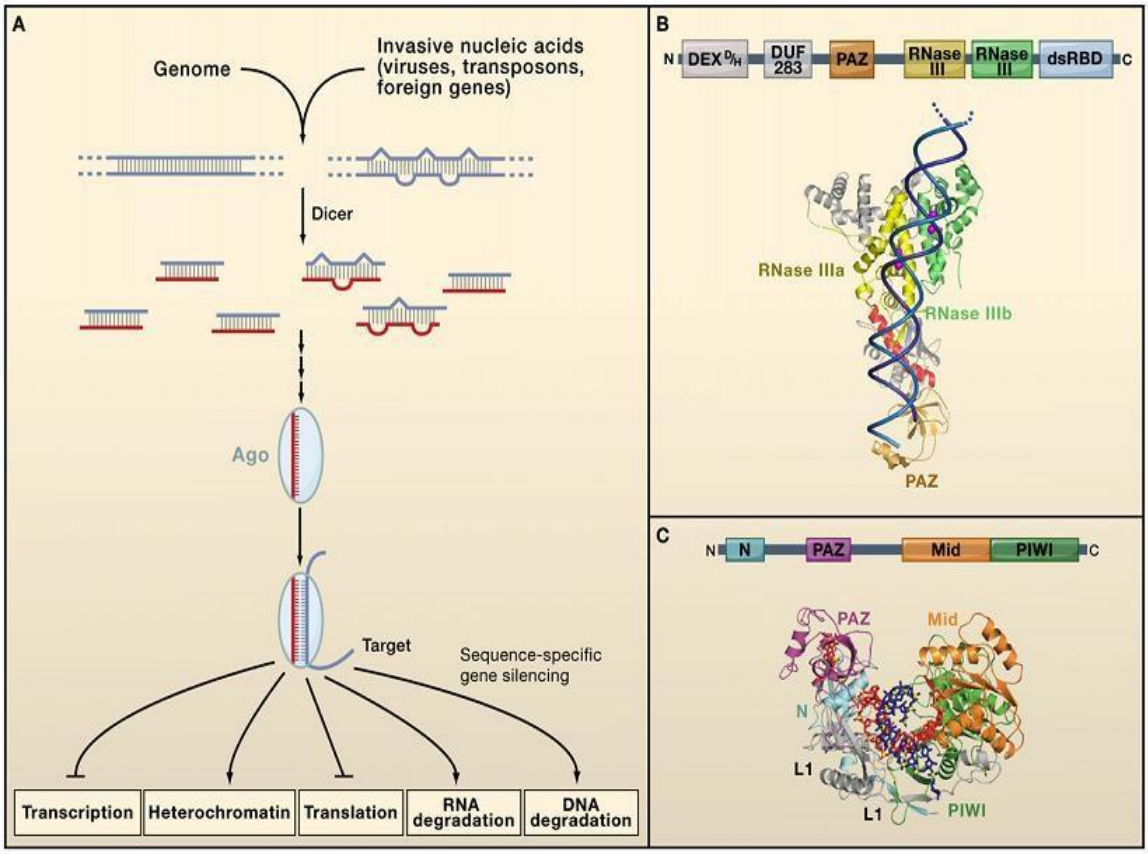

Figure 5. Core Features of miRNA and siRNA Silencing, adapted from (Carthew and Sontheimer, 2009). (A) Common aspects of all miRNA and siRNA pathways. Double-stranded RNA precursors are processed by a DCR protein into short (20−30 nt) fragments. One strand of the processed duplex is loaded into an AGO protein, enabling target RNA recognition based on sequence complementarity. Once the target is recognized, its expression is modulated by one of several distinct mechanisms. (B) The domain arrangement of most DCR enzymes is shown at the top, while the crystal structure

corresponds to *Giardia* DCR that carries only one PAZ and two RNase III domains. (C) The canonical arrangement of AGO domains is shown at the top. Below is the crystal structure of the *Thermus thermophilus* AGO protein, bound to its guide and RNA target.

1.2.3. Sources of small regulatory non-coding RNAs

Small RNAs size range varies between organisms and even different silencing pathways within the same organism (Meister and Tuschl, 2004). Small RNAs are mostly produced from inverted repeats and TE loci. sRNAs are also produced from complementary transcripts (generated in *cis* i.e. convergent transcription but also in *trans*, protein-coding genes, ncRNAs such as tRNAs, and self-complementary transcripts with a hairpin structure). RNA dependent RNA polymerases (RDR) also produce small RNAs using as templates the already RNAi targeted transcripts. RDRs have been characterized in plants but seem to be absent in insects and mammals. The three major classes of small RNAs are micro-RNA (miRNAs), Small interfering RNAs (siRNAs_ and Piwi-interacting RNAs (piRNAs). They present differences in biogenesis, the selection of Argonaute proteins, and the silencing mechanisms.

1.2.4. Small interfering RNAs

siRNAs are generated from long perfectly base-paired dsRNAs which derive from viruses, transgenes, endogenous transcripts of inverted repeat elements (Carthew and Sontheimer, 2009; Borges and Martienssen, 2015). Small interfering RNAs (siRNAs) were originally discovered in plants, where they showed to trigger antiviral post-transcriptional gene silencing (Hamilton and Baulcombe, 1999). siRNAs were also found to derive from repetitive elements, particularly Trasposable Elements (TEs), that are often capable of independent replication across the genome and therefore must be repressed (Malone and Hannon, 2009). The production and maturation of siRNA requires DCR and sometimes the RNA dependent RNA polymerases (RdRs) to amplify secondary siRNAs (Tang *et al.*, 2003, Sijen *et al.*, 2001). Whole genome analyses indicate that DCR, AGO and RDR encoding genes have expended through duplication events in some organisms (e.g. plants, *C.elegans*), or been lost all together in other organisms (eg. *S.cerevisiae*). In organisms encoding more than one DCR, maturation of siRNA and miRNAs may require different DCR homologs (i.e. sub-functionalization). In Drosophila

two DCR homologs, Dicer-1 and Dicer-2, are independently in charge of miRNA and siRNA biogenesis, respectively (Lee et al., 2004). In plants, this phenomenon occurs often as several Dicer-like proteins are usually present (Fukudome and Fukuhara, 2017). In *Arabidopsis*, DCL1 produces miRNA but can also generate siRNAs that repress complementary mRNAs post-transcriptionally (Vazquez et al., 2004; Henderson et al., 2006). DCL2 and DCL4 generate 21 and 22nt siRNAs from transgene RNAs, viruses or endogenous precursors (Xie et al., 2004, 2005; Parent et al., 2015). DCL3 processes mostly endogenous dsRNAs from TEs into 24nt siRNAs, which trigger RNA-directed DNA methylation and transcriptional repression of target TEs (Xie et al., 2004). All DCR-generated siRNAs are perfectly complementary to their target and possess the typical 5'-monophosphate termini and the 2'-*O*methylation as well as the 2nt 3'-overhangs (Li et al., 2005; Horwich et al., 2007).

1.2.5. Micro-RNAs

Mature miRNAs are usually 21-24nt in length. Gene silencing mediated by miRNA was initially discovered in *C. elegans* in 1993 (Lee et al., 1993; Wightman et al., 1993) with the second miRNA identified in 2000 by Reinhart et al. (2000). Both research groups showed that a 22nt ncRNA was negatively regulating several heterochromatic genes through RNA/RNA interactions in their 3'UTRs. Soon, they were shown to represent a large class of ncRNA and were named microRNA (miRNA) (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). Contrarily to siRNAs which are processed from a long dsRNA, miRNA's precursor is always a hairpin structure (Kim, 2005). Many miRNAs were subsequently found in plants and animal with some being conserved between the 2 kingdoms (Lagos-Quintana et al., 2001; Llave et al., 2002; Reinhart et al., 2002). Later on, unicellular organisms including *Chlamydomonas reinhardtii* and *Dictiostellium discoideum* were shown to present miRNAs (Hinas et al., 2007; Molnár et al., 2007; Zhao et al., 2007). MiRNAs play pivotal roles in diverse biological processes, ranging from cell proliferation to organism development, (He and Hannon, 2004; Filipowicz et al., 2008). The miRNA pathways have diverged between plant and animal lineages. In animals, dsRNAs are processed by the two RNAIII-type endonucleases, DROSHA and DCR [Kim, 2005] and are required to present a specific

RNA folding to stem-loop structure, referred to as the primary miRNA (pri-miRNA) (Carthew and Sontheimer, 2009). DROSHA, as part of the microprocessor complex in the nucleus (Carthew and Sontheimer, 2009), cleaves the precursor stem-loop (pre-miRNA) from specific positions so that a 2-nucleotide long 3' overhang is left at the end of the hairpin secondary structure (Meister and Tuschl, 2004). Subsequently, the pre-miRNA is exported from the nucleus where DCR cleaves and produces a short dsRNA with 3'-end 2-nucleotide overhangs. One of the RNA strands is incorporated into the AGO of RISC complex and the other is degraded. In plants, DROSHA is absent and miRNA maturation requires DCR only. DCR alone is processing the pri-miRNA into pre and mature miRNA in the nucleus (Baulcombe, 2004). Because miRNAs have been found so far in only two of the six eukaryotic supergroups, and with distinct biogenesis pathways, miRNAs have been proposed that have independently evolved in animal and plants (Cerutti and Casas-Mollano, 2006).

1.2.6. Piwi-interacting RNA (piRNA)

PiRNAs are 29-30nt small RNAs associated with Piwi (P-element-induced wimpy testis) proteins which comprise a subclade of the large Argonaute protein family (see below) and were initially discovered in the germline of mouse (Lin and Spradling, 1997; Aravin et al., 2006; Grivna et al., 2006; Lau et al., 2006). To date, piRNA have been found only in animals (Saito et al., 2006; Brennecke et al., 2007). They present many differences to miRNAs and siRNAs. First, piRNAs are slightly longer and have a strong bias for uridine at the 5'end and adenine in the tenth position of nucleotides (Brennecke et al., 2007). Second, their biogenesis involved PIWI but not DCR and DROSHA. Third, they originate from piRNA clusters found in discrete intergenic loci and TEs (Brennecke et al., 2007; Goriaux et al., 2014; Rogers et al., 2017). The "ping pong" model proposes that sense transcripts from TEs and antisense transcripts from piRNA master loci fuel an amplification cycle involving a Piwi-containing RISC complex and RDR polymerases (Hartig et al., 2007). They associate only with PIWI proteins in order to silence TEs in germline cells (Brennecke et al., 2007; Kuramochi-Miyagawa et al., 2008; Sienski et al., 2012). However, they carry a 2'-*O*-methylation at their 3'ends like siRNAs and plant miRNAs (Horwich et al., 2007).

1.2.7. TE-derived siRNAs

Transposable elements (TEs) are the major constituent of many eukaryotic genomes, known to create extra copies in the genome. Many TEs are autonomous elements since they encode for the proteins necessary for transposition, like transposase protein (Vitte et al., 2014). The replication cycle of most TEs initiates with transcription of genomic copy by the host's RNA polymerase II. The mRNAs of TEs are subjected to both translation and reverse-transcription (Grandbastien, 1998). As a result of reverse-transcription, the linear and double-stranded DNA is produced and then transported back to the nucleus and integrate to genomic chromosomal DNA by the integrase protein.

Since TE mobilization can be mutagenic, the host genomes have evolved elaborate mechanisms to suppress their activities (Matzke and Mosher, 2014). In plants, TEs are primarily repressed by the epigenetic silencing pathways including histone modification and DNA methylation. In plants, the RNA-directed DNA methylation (RdDM) pathway plays a central role in TE silencing. Genomic regions marked by DNA methylation are recognized by the plant-specific RNA polymerase, RNA PolIV, which transcribes relatively short RNAs (Blevins et al., 2015; Zhai et al., 2015). The transcribed RNAs become dsRNAs by the RNA-dependent RNA polymerase (RDR) 2 avtivity and subsequently sliced to 24 nt small interfering (si) RNAs by DCL3 protein. These 24 nt-siRNAs guide AGO4 proteins to the nascent RNA transcribed by the RNA PolV. AGO4 then recruits multiple proteins including SU(VAR)3-9 HOMOLOG (SUVH) 4/5/6 and DOMAINS REARRANGED METHYLASE (DRM) 1/2 that mediate repressive histone modification (H3K9me2) and DNA methylation, respectively, thus contributing to reinforcement of the silenced state of TE chromatins (Zilberman et al., 2004; Tran et al., 2005; Zhong et al., 2014). TEs escaped from silencing or newly introduced to the genome are recognized by the RDR6-RdDM pathway that post-transcriptionally degrades TE mRNAs. RNA PolII-transcribed TE mRNAs are processed to 21 or 22 nt-siRNAs by the RDR6 and DCL2/4 (Creasey et al., 2014). These 21 or 22 nt-siRNAs associate with AGO1 and target TE mRNAs for degradation. In mammals, PIWI-interacting sRNAs regulate a large number of mRNAs and long non-coding RNAs in testis, suggesting widespread regulatory roles of TE-derived small RNAs in both plants and animals (Watanabe et al., 2015).

1.2.8. tRNA-derived small RNAs (tsRNAs)

In addition to the aforementioned classes of sRNAs, new classes of sRNAs deriving from transfer RNA (tRNAs) (Cole et al., 2009; Lee et al., 2009b; Haussecker et al., 2010), messenger RNA (mRNA), and small nucleolar RNAs (snoRNAs) have been identified. Some of these recently identified sRNAs are considered to be associated at least to some steps of the RNAi pathway.

sRNAs that derive from tRNAs, termed tRNA-fragments (tRFs) have been identified in several studies (Cole et al., 2009; Lee et al., 2009b; Haussecker et al., 2010), (Lee et al., 2009b). tRFs have been found to correspond to either the -3 fragments from matching exactly to the 3' end, tRF-5 fragments mapping exactly to the 5' end of the tRNA, and tRF-1 fragments that map to the 3' end of the pre-tRNA sequence. tRF-3 fragments present the CCA motif that is post- transcriptionally added to the 3' end of tRNAs, confirming that tRF-3 fragments are at least processed from mature tRNAs. tRFs are usually found as 13-22nt RNAs (Tuck and Tollervey, 2011). Others tRFs are tRNA halves that are generated after cleavage at the anticodon loop. tRF-halves have been shown to be produced under stress conditions (Thompson and Parker, 2009). The proteins responsible for this cleavage are stress-activated nucleases, like RNY1 in *Saccharomyces cerevisiae*, part of the RNAse T2 family, and angiogenin in mammals, part of the RNAse A family. It has been shown that tRF-halves can inhibit translation activity in eukaryotes (Zhang et al., 2009). Although the function of tRFs is more elusive, there is evidence suggesting that they compete with other sRNAs for AGO proteins (Haussecker et al., 2010).

1.2.9. DICER

DICERs (DCRs) generate sRNAs including siRNAs and miRNAs, of 21-25 nt and presenting ~2nt 3' overhangs and a 5' phosphate group. In *C.elegans*, mammals, yeast and some other organisms, one DCR is responsible for the production of miRNAs and siRNAs (Tomari and Zamore, 2005), but in other organisms like plants and *Drosophila melanogaster*, different DCRs generate different sRNAs (Lee et al., 2004; Xie et al., 2004).

DCR-like proteins are members of the RNase III family, a group of proteins that catalyze the cleavage of dsRNA (Carthew and Sontheimer, 2009). All members of the RNaseIII family contain a characteristic ribonuclease domain, which has a highly conserved stretch of nine amino acid residues known as the RNaseIII signature motif. RNaseIII proteins vary widely in length, from 200 to 2000 amino acids (Filippov et al., 2000) and have been subdivided into four classes based on domain composition. Class I is the simplest and the smallest. It includes bacterial RNases III and contains a single RNaseIII domain and a dsRNA-binding domain (dsRBD). Class II is identified by the presence of a highly variable N-terminal domain extension and includes the *S. cerevisiae* Rnt1 and *S. pombe* Pac1 proteins (Rotondo and Frendewey, 1996), which are longer than bacterial RNaseIII with an additional ∼100 amino acid fragment at the N-terminus. ClassI and II enzymes function as homodimers (Jaskiewicz and Filipowicz, 2008). Class III, including DROSHA proteins (Filippov et al., 2000), has a dsRBD and two RNaseIII domains. Class IV RNaseIII are the largest and includes typical DCR proteins (Nicholson, 1999; Filippov et al., 2000) that act as monomers (Jaskiewicz and Filipowicz, 2008). Typical (conserved) DCR domains are a) the PAZ domain, shared with AGO proteins, that binds RNA duplex ends based on their characteristic overhangs and b) two RNase III domains that excise the sRNA from its precursor at a specific length based on the distance between PAZ domain and processing centre of the protein (MacRae et al., 2007). At their C' terminus DCRs usually contain a dsRNA binding motif (dsRBD) and at the N'terminus usually an ATPase/helicase domain and a domain of unknown function (DUF283).

PAZ domain binds dsRNA ends and positions it for each RNaseIII domain to cleave one strand, thus generating new ends with 3' 2 nt overhangs. The distance between the PAZ domain and the RNase III domains serves as a molecular ruler to produce sRNAs of specific length (Lingel et al., 2003; Ma et al., 2004; Zhang et al., 2004; Macrae et al., 2006; Macrae and Doudna, 2007). The role of ATPase/helicase and DUF283 domains in sRNA generation is still elusive, since not all DCRs require ATP for endonucleolytic cleavage (Tomari and Zamore, 2005).

DCR domains are generally conserved in eukaryotes but their structure and organization vary (Cerutti and Casas-Mollano, 2006). The greatest variability is the absence of the dsRBD, PAZ and ATPase/helicase with DUF283 domain. The dsDBS of *S.pompe* DCR is not required to process dsRNA in vitro (Colmenares et al., 2007) and deleting the dsRBD from human DCR does not abrogate its catalytic efficiency in vitro (Zhang et al., 2004). *Giardia intestinalis* (excavate) DCR consists of a PAZ followed by two RNase III domains processing 25-27nt sRNAs (Macrae et al., 2006) and *Tetrahymena thermophila* (ciliate protozoan- chromalveolate) DCR only contains two RNase III domains and a dsRBD domain (Mochizuki and Gorovsky, 2005). Finally, the highly divergent *Trypanosoma brucei* (excavate) encodes only two RNaseIII domains (Shi et al., 2006). These findings suggested that the minimal domain requirements for DCR dsRNA cleavage are the two RNaseIII domains (Cerutti and Casas-Mollano, 2006). However, a DCR protein from classI RNase III enzymes has been identified is *Saccharomyces castellii* containing two dsRBDs and acting as homodimer (Drinnenberg et al., 2009). Another protein containing only one RNaseIII domain presents dsRNA cleavage activity and mediates RNAi in a heterologous system has been recently characterized in *Entamoeba hystolytic*a (Pompey et al., 2015a).

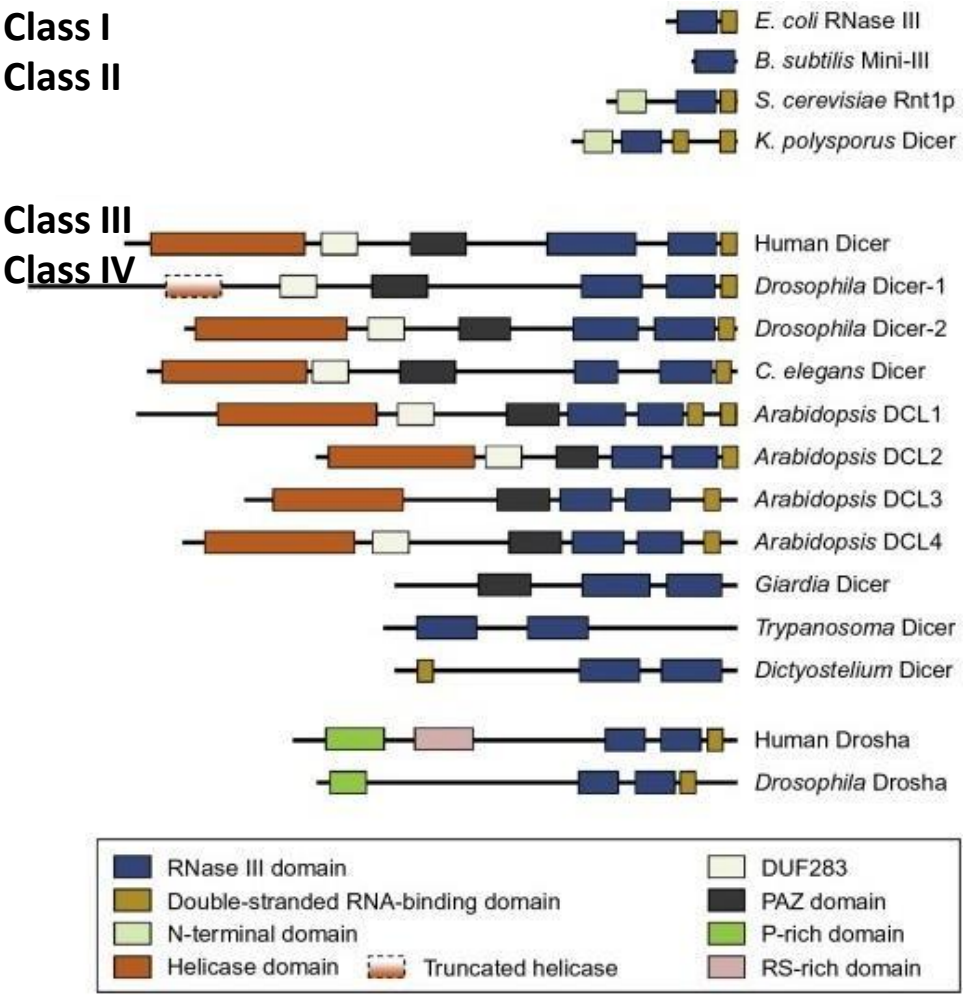

Figure 6. Classification of RNase III family enzymes into four classes, modified from Doyle et al. (2012) Class I contains bacterial RNase III orthologs ClassII contains and fungal yeast Dicer with elongated N' terminus, all containing a single RNase III domain. Class IV contains Dicer and Class III Drosha proteins containing two RNase III domains.

1.2.10. DROSHA

DROSHA proteins have been identified only in metazoan organisms and are thought to have evolved from DCR (Cerutti and Casas-Mollano, 2006). In animals, homologs of DROSHA and the dsRNA binding protein DGCR8 (DiGeorge syndrome critical region 8) are the major components of miRNA maturation (Han et al., 2004). DGCR8 recognizes pri-miRNA and facilitates their binding on the microprocessor, where DROSHA cleaves both strands through its two RNase III domains (RIIIDs) (Han et al., 2004, 2006; Nguyen et al., 2015; Kwon et al., 2016b). The product of this cleavage is a

~70nt hairpin precursor (pre-miRNA), with a 5'-monophosphate and a 3'- hydroxyl (OH), presenting a 2nt-overhang at its 3'-ends. Subsequently, the hairpin precursor is transported to the cytoplasm by Exportin-5 (Exp5) (Yi et al., 2003; Lund et al., 2004) where is further processed by DCR. However, studies during the last decade have revised the functions of DROSHA and DGCR8 concluding that are not limited to pri-miRNA processing (Lee and Shin, 2018). DROSHA can partially regulate RNA metabolism, by post-transcriptional control of RNA stability (Han et al., 2009; Kadener et al., 2009; Chong et al., 2010; MacIas et al., 2012; Heras et al., 2013; Johanson et al., 2015; Kim et al., 2017; Marinaro et al., 2017), transcriptional activation (Gromak et al., 2013), alternative splicing (MacIas et al., 2012; Havens et al., 2014; Lee et al., 2017), 3′-end processing and transcriptional termination (Ballarino et al., 2009; Dhir et al., 2015). DROSHA can promote defense against viruses (Lin and Sullivan, 2011; Shapiro et al., 2014; Aguado and Benjamin, 2017) and regulate the expression of retrotransposons (Heras et al., 2013). Moreover, cleavage-independent functions of DROSHA have been reported, including possible recruitment of positive regulators to promote transcription (Gromak et al., 2013) and alternative splicing (Havens et al., 2014), and binding to its substrates enabling access to other proteins to exert their inhibitory roles (Aguado and Benjamin, 2017; Lee et al., 2017). Finally, indications of cytoplasmic functions of DROSHA are under study as its translocation has been reported under stress condition and a small fraction of DROSHA protein due to alternative splicing is present in the cytoplasm even under normal conditions (Dai et al., 2016; Link et al., 2016; Lee et al., 2017).

1.2.11. Argonaute

sRNAs bind to Argonautes (AGOs) and guide the RISC complex to mediate silencing of its the target. AGO proteins represent a large protein family, highly conserved in many prokaryotic and eukaryotic organisms with a key role in all RNAi-mediated silencing pathways (Swarts et al., 2014). Eukaryotic AGO proteins can be divided in three clades. The AGO clade proteins are widespread and usually load miRNAs and siRNAs (Bohmert et al., 1998), the PIWI clade proteins are restricted in animals, ciliates and amoebozoa and associate with piRNAs (Lin and Spradling, 1997) and a third AGO clade is identified only in *C.elegans* (Yigit et al., 2006).

AGOs are composed of four domains: a N' terminal domain, followed by a PAZ domain, a Mid domain and the finally the PIWI (P-element induced wimpy testis) domain (Cerutti et al., 2000). The 3' end of sRNA binds to the PAZ and MID and PIWI domains bind 5' end of sRNA with strong affinity for uracil or adenine (Ma et al., 2004; Frank et al., 2010, 2012). PAZ domain of PIWI proteins recognizes piRNAs by the methylation at their 3'-ends (Simon et al., 2011; Tian et al., 2011). In many AGO proteins, the RNAse H fold of PIWI contains three conserved residues necessary for cleavage of the target RNA. Other AGOs do not cleave, but recruit other factors to promote silencing of the target (Ma et al., 2004; Carthew and Sontheimer, 2009).

Most eukaryotes have multiple AGO members. *Droshophila melanogaster* expresses 5 AGO, *Arabidopsis thaliana* encodes 10 AGO and *C. elegans* has the larger number of 27 AGO members (Yigit et al., 2006). A few organisms, like *S.pombe* (Volpe et al., 2002), *S.castellii* (Drinnenberg et al., 2009) and *Ectocarpus siliculosus* (Tarver et al., 2015), express only one AGO.

1.2.12. RDR (RNA dependent RNA polymerase)

RDRs were described initially in viruses and were later found in other organisms including plants (Schiebel et al., 1998; Dalmay et al., 2000; Mourrain et al., 2000), yeast Neurospora *crassa* (Cogoni and Macino, 1999) and *S.pombe* (Volpe et al., 2002), *C. elegans* (Smardon et al., 2000), and the protozoan *Tetrahymena thermophila* (Lee and Collins, 2007), RDR homologues, however have not been found in insects and mammals. RDRs are required for efficient RNAi-mediated silencing by amplifying the RNAi response. To do so, they use the RNAi targeted loci as templates to generate dsRNA that are referred to as secondary sRNAs. In organisms containing more than one RDR, like *A. thaliana* and *C.elegans*, different RDR polymerases generate different types of secondary sRNAs that are loaded onto their various AGO proteins in order to amplify and sustain the RNAi response throughout the organism (Sijen et al., 2001; Voinnet, 2008; Carthew and Sontheimer, 2009). *S. pombe* unique RDR generates secondary sRNAs from targets marked with sRNA-loaded AGOs and the newly synthetized DCR-depended sRNAs are required for the formation of centromeric heterochromatin (Motamedi et al., 2004).

1.2.13. The evolution of RNA interference

The presence of conserved RNAi pathways in plants, animals, yeast and other different lineages, indicates that the last common ancestor of eukaryotes had a functional RNAi machinery. Current hypothesis postulate that sRNAs originally evolved in eukaryotes to defend against invading viruses (Baulcombe, 2004). However, it seems that RNAi is not essential for all eukaryotes since it is absent in a few unicellular eukaryotes like *Saccharomyces cerevisiae*, the excavates *Trypanosoma cruzi* and *Leishmania major*, the Archaeplastida *Cyanidioschyzon merolae*, and the malaria-causing pathogen *Plasmodium falciparum* (Cerutti and Casas-Mollano, 2006).

1.3. *Phaeodactylum tricornutum*

P. tricornutum is a pennate diatom belonging to the class Bacillariophyceae and the Phaeodactulaceae family, and is the only species in genus Phaeodactylum. It can exist in three forms: fusiform, a normal cell type for pennate diatom with two arms, oval cells which have no arm, and triradiate which resembles fusiform with three arms. Because these cells are weakly silicified and the oval cell possesses only one theca, they were initially described as new genus of unicellular algae by Bohlin in 1897 (as *Nitzschia closterium*), aside from other diatom genera (Lewin, 1958). Changes in cell shape are induced by environmental conditions, a feature that can be used to explore the molecular basis of cell shape control and morphogenesis. Since *P. tricornutum* silicified frustules is non-essential, it can grow in absence of silicon, is making a good candidate for experimental exploration of silicon-based nanofabrication in diatoms (nanotechnology).

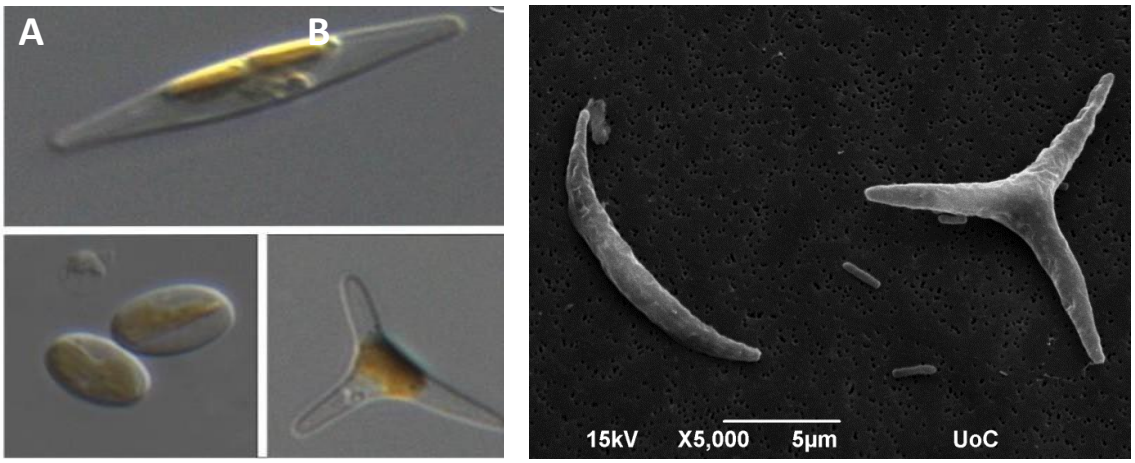

Figure 7. Pictures of *P. tricornutum* under light microscopy (LM) and electron microscopy (SEM).

Although there is no substantial evidence for sexual reproduction of *P. tricornutum*, perhaps its ability to change morphotypes, and even in non-silicified forms, enables it to escape from the inevitable decrease in cell size found in other diatoms. Since diatom sexual reproduction is often inhibited in laboratory cultures, this feature of *P. tricornutum* explains its early investigations in laboratories across the world. Apart its ease of laboratory growth, this diatom can also be genetically transformed.

The establishment of *P. tricornutum* as model led to many studies that brought insights into diatom C, N, and Fe metabolism, as well as cell cycle in diatoms. Other interesting biotechnological applications of this diatom include its use for production of biofuels and high value pharmaceuticals, where it can be used also as a platform for heterologous expression, as well as its cultivation as food in aquaculture (for larval molluscs and fish).

Out of the ten *P. tricornutum* ecotypes characterized, the genome of clone CCAP1055/1 (Pt1) was sequenced and annotated by JGI (DOE Joint Genome Institute, USA), released as Phat2 version (Bowler et al., 2008) and subsequently re-annotated by a collaboration of Ensembl Genomes (EMBL-EBI), the Ecole Normale Superieure (Paris) and the J. Craig Venter Institute (San Diego) in order to provide a refined version (Phatr3 annotation). It was the second model diatom with its genome sequenced following the centric *T.pseudonana* genome, sequenced in 2004 (Armbrust et al., 2004). Diatom expressed sequence tags (ESTs) became also available (Maheswari et al., 2009).

The *P. tricornutum* genome size is 27.4 Mb with a total of 12 233 coding genes and an average of 0.79 introns. Approximately 12.3% of the genome corresponds to repetitive sequences belonging mainly to copia retroelements (~62%), that include also diatom-specific classes (CoDi elements). Regarding the extend of horizontal gene transfer (HGT), 587 genes appear to be most closely related to bacterial genes, accounting for more than 5% of the *P. tricornutum* proteome. The fact that 56% of these genes were also found in *T. pseudonana* genome suggested their acquisition in the common ancestor of central and pennate diatoms. The further acquisition of bacterial origin genes in pennate diatoms including *P. tricornutum* has been exemplified by the recently characterized cyanobacterial ferritin-like gene enabling iron storage, that is absent in centric diatoms (Marchetti et al., 2008). Finally, the genome of *P. tricornutum* contains large numbers of diatom-specific cyclins, heat shock transcription factors (Huysman et al., 2010; Rayko et al., 2010), and far-red light sensors related to phytochromes (Fortunato et al., 2016).

P. tricornutum genome is compact with small gene sizes and intergenic regions, and a significant proportion of transposable elements (TEs) which predominantly belong to class I TEs, in particular the LTR retrotransposon superfamily (Maumus et al., 2009). Analysis of LTR retro-elements of the copia type in both *P. tricornutum* and *T. pseudonana* revealed the existence of seven groups of diatom-specific TEs named CoDi (Copia-like elements from diatoms) (Maumus et al., 2009). In both *P. tricornutu*m and *T. pseudonana* some of the CoDi groups were shown to be expressed under specific conditions suggesting a role of TEs in adaptation and diversification of diatoms (Maumus et al., 2009). The recent re-annotation of *P. tricornutum* (Phatr3 version) confirmed that 12% of the genome corresponds to repeats, with 75% being TEs (Rastogi et al., 2018). Near half of the TEs are associated with repressive epigenetic marks (DNA methylation, H3K27me3, H3K9me2, H3K9me3) suggesting the importance of keeping these TEs under tight control.

Interestingly, a significant number of genes encoding reverse transcriptases with additional protein domains were found in the last annotated version of *P. tricornutum* genome (Rastogi et al., 2018). Their abundance suggests that these proteins are transcriptionally active in diatoms, implicating a possible role in diatom evolution and adaptation to contemporary environments (Lescot et al., 2016). The presence of

additional protein domains supports previous studies (Chuong et al., 2017) which suggest that Rv domain-containing proteins might have originated from domesticated retrotransposons that evolved different functions via acquisition of various N- and C-terminal extensions (Rastogi et al., 2018). Novel classes of TEs were revealed in this annotation including Miniature inverted–repeat transposable elements (MITE), that play a major role in genomes organization and species evolution in plants and animals, and Short interspersed elements (SINE), with a role in mRNA splicing, protein translation and allele expression bias in mammals, plants and some invertebrates but not encountered in unicellular species before (Kramerov and Vassetzky, 2011).

Epigenetic marks

The epigenetic machinery generally found in higher eukaryotes seem to be present in diatoms analyzed *in silico* (Marron et al., 2016), suggesting the ancient origin of this mode of genome regulation. In *P. tricornutum* DNA cytosine methylation has been found over around 5,2% of the genome, localized at genes, intergenic regions, and TEs, and in all contexts (CG, CHH and CHG) but mostly in CpG. A dynamic regulation of DNA methylation was observed under nitrate starvation, with both genes and TEs being differentially methylated, suggesting a role in the acclamatory response to nutrient limitation (Veluchamy et al., 2013) In this study, it was shown that nitrate limitation induced demethylation and upregulation of the LTR-retrotransposon called Blackbeard (Bkb) Veluchamy et al., 2013). Some DNA methylated regions are highly covered by small RNAs (Rogato et al., 2014) suggesting an RNA-directed DNA methylation process (Rastogi et al., 2018). Functional DCR and AGO, along with other proteins, could drive *de novo* methylation.

Post translational modifications (PTMs) of histones in *P. tricornutum* (Veluchamy et al., 2015) were identified and initially seemed to resembled to PTMs specific to plants and animals, thus reflecting the chimeric nature of its genome. However, whole genome mapping of a few key PTMs revealed a conserved histone code more similar to animals (Veluchamy et al., 2015)]. From this analysis came out that histone marks and DNA methylation co-occur to determine chromatin states that either repress or activate the expression of genes and transposable elements. Therefore, epigenetic regulation can provide diatoms the ability to adapt in changing environments.

Alternative splicing

Based on recent findings (Rastogi et al., 2018), extensive alternative splicing has been shown to take place in *P. tricornutum*, including intron retention and exon skipping, thus increasing the diversity of transcripts generated in changing environments. *P. tricornutum* shows a higher rate of intron-retention (IR) which is prominent in plants and unicellular eukaryotes, rather than exon-skipping (ES) which is prominent in metazoans (McGuire et al., 2008). Surprisingly, it was found that genes exhibiting IR were more expressed than genes without exhibiting alternative splicing (Rastogi et al., 2018). This is opposed to the IR found in mammals which down-regulates genes expression. The increased functional diversity, due to the generation of alternative messenger RNAs, along with epigenetic marks-based regulation are likely to play a major role in responses to environmental changes and should also be considered during functional studies in diatoms.

Genetic engineering

Genetic engineering of *P. tricornutum* is enabled due to establishment of tools and methodologies that introduce DNA into diatom cells and ensure its integration into the genome for gene expression. These involve a) the construction of appropriate vectors carrying all the structural elements controlling transgene expression (e.g. promoters and terminators of transcription), b) the identification of selectable markers to isolate transformed cells and c) the development of efficient methods for stable DNA delivery.

Endogenous promoters have been obtained from *P. tricornutum* genes encoding a chlorophyll a/c-binding light-harvesting complex protein, Lhcf, formerly called Fcp, a histone gene (h4) and the elongation factor 2 (ef2) genes. All these promoters are constitutive and drive high levels of transgene expression, but the Fcp promoters were light dependent while the ef2 promoter seems to drive the highest expression levels among them (Seo et al., 2015). The terminators used to date have been those of the Lhcf1, Lhcf9, nr (nitrate reductase), rbcL(rubisco small subunit) and Lhcr14 genes (Apt et al., 1996; Falciatore et al., 1999a; Zaslavskaia et al., 2000a; Poulsen and Kröger, 2005; Xie et al., 2014; Ifuku et al., 2015; Karas et al., 2015). Recently, expression vectors able to express two genes simultaneously have been developed, enabling co-localization studies (Liu et al., 2016) and investigation of metabolic pathways where more than one

gene is involved (Hempel et al., 2011)**.** Inducible promoters switching on and off transgene expression are also being developed. To date, the nitrate reductase (nr) promoter, induced by the presence of nitrate and inactivated in the presence of ammonium ions (Poulsen and Kröger, 2005) exhibits a "leaky" expression (Chu et al., 2016), while iron-starvation induced promoters have also been described (Yoshinaga et al., 2014). Heterologous promoters, like the use of the Lhcf1 promoter from *Cylindrotheca fusiformis,* can be used in order to maintain *P. tricornutum* endogenous regulatory networks (Kadono et al., 2015). Finally, although *P. tricornutum* does not seem to get infected by any known virus, promoters from diatom-infecting viruses, like the ClP1 promoter, have shown that can drive stable expression and in levels higher than those with endogenous diatom promoters (Kadono et al., 2015) ,.

Reporter genes that have been widely used in diatoms are the bacterial b-glucuronidase gene (GUS) for which expression can be monitored on the basis of enzymatic activity (Zaslavskaia et al., 2000b) and the luciferase (LUC) gene (Muto et al., 2013). Monitoring protein localization in vivo was enabled by using as reporters the fluorescent proteins such as the green fluorescent protein (GFP), the cyan fluorescentprotein (CFP) and the yellow fluorescent protein (YFP). The most commonly used selectable markers conferring resistance to antibiotics in diatom transformed cells are nourseothricin (nat), neomycin (nptII), phleomycin/zeocin (sh ble) (Falciatore et al., 1999b) and the most recently added blasticidin (blast) (Buck et al., 2018). A Gateway-based system incorporating the above basic features was constructed for *P. tricornutum* transformation (Siaut et al., 2007) and recently a GoldenGate system based on Type IIS restriction enzymes became available (Pollak et al., 2019).

The initial methodologies for nuclear transformation were based on biolistics (Apt et al., 1996; Falciatore et al., 1999b) but recently transformation procedures via electroporation and bacterial conjugation have also been established (Zhang and Hu, 2014; Karas et al., 2015). In biolistics, also known as 'particle bombardment', particles (microgold or tungsten) coated with DNA are used to deliver transgenes directly into diatom cells. Electroporation is based on the application of a strong electrical field to enhance pore formation in the cell membrane. While these methods induce integration of the introduced plasmid in the genome, the third method is based on episome delivery of

DNA via bacterial conjugation. In this case, the introduced genes can be expressed without the uncontrolled genetic modifications caused during random genomic integration of plasmids. Chloroplast transformation is also developed in *P. tricornutum* (Materna et al., 2009; Xie et al., 2014). The chloroplast presents a high homologous recombination frequency facilitating gene insertion, an "operon" type organization of genes and absence of epigenetic marks.

Most studies on the molecular biology of diatoms use reverse genetics as they are usually interested in linking physiological processes to gene functions, and therefore to investigate the consequences a particular gene's change in order to infer its function. Two approaches are commonly used: (i) the overexpression of endogenous or heterologous genes and (ii) the down-regulation of endogenous genes and genome-editing technologies, based on the use of double-strand-break (DSB) mechanisms efficiently providing targeted modifications (Huang and Daboussi, 2017a). The expression of a specific gene leads a) to study the localization of a protein fused to a fluorescent marker, b) to assess the consequences of this overexpression for metabolism, c) to complement loss of function in a mutant strain, (d) to investigate gene functions by exploring functional conservation in heterologous expression experiments, and (e) to create new functions through the introduction of foreign genes controlling a metabolic pathway (Huang and Daboussi, 2017a).

Targeted genome engineering with site-specific nucleases

The use of double-strand DNA break (DSB) repair mechanisms has emerged as a revolutionary method for highly efficient targeted genome modifications. Homologous recombination in microalgae is not very frequent (less than 10^{-6}), but molecular scissors able to induce DSBs at a specific locus increases this frequency by at least three orders of magnitude (Daboussi et al., 2014). There are three classes of sequence-specific nucleases that have been used to successfully induce targeted modifications in diatom genomes: meganucleases, TALENs and CRISPR/Cas9 enzymes.

Meganucleases (MNs) and Transcriptional Activator-like Effectors (TALENs)

Meganucleases (MNs) derived from the endonucleases involved in the lateral transfer of introns in yeasts (Silva et al., 2011). TALENs derived from the transcriptional activator-like effectors produced by the plant pathogenic bacterium Xanthomonas, which promotes transcription of a plant gene to enable bacterial infection (Christian et al., 2010). The targeted genome modification in *P. tricornutum* by using MNs and TALENs showed a high frequency of both targeted mutagenesis and homologous recombination (Daboussi et al., 2014). Application of these methodologies in metabolic engineering soon led to inactivation of one gene involved in the storage of energy and the creation of a *P. tricornutum* strain producing large amounts of lipids. Since then, TALEN technology in *P. tricornutum* has been used in several studies (Daboussi et al., 2014; Weyman et al., 2015; Fortunato et al., 2016).

CRISPR/Cas9 system

The CRISPR/Cas9 system derived from a bacterial/ archaeal defence system against bacteriophages. It is based on an RNA-guided DNA cleavage by endonuclease Cas9 and storage of the foreign DNA in their genome memory, at the CRISPR locus (Deltcheva et al., 2011; Jinek et al., 2012; Doudna and Charpentier, 2014). The CRISPR/Cas9 system is very efficient, can produce multiple gene modifications and acquires only the expression of a Cas9 protein and a single guide RNA (sgRNA). The custom sgRNA contains a targeting sequence (crRNA sequence) homologous to the genomic region to be modified, and a Cas9 nuclease-recruiting sequence (tracrRNA). The binding specificity is based on the sgRNA and a 3-nucleotide downstream sequence called the protospacer adjacent motif (PAM), which is NGG, in the case of S. *pyogenes* Cas9. The Cas9 nuclease carries two nuclease domains (HNH and RvuC) and cleaves both DNA strands generating DSBs at sites defined by the 20-nucleotide guide sequence. CRISPR/Cas9 system has been successfully used in P. *tricornutum* to knock out the chloroplast signal recognition particle 54 (CpSRP54) and in T. p*seudonana* to knock out the urease gene (Hopes et al., 2016a; Nymark et al., 2016a). For genome editing in *P. tricornutum*, the *Cas9* nuclease and a guide RNA directing the nuclease to a specific DNA sequence were expressed from the same vector in order to increase the probability

of co-delivery during biolistic transformation. The diatom codon-optimized *Cas9* was placed under the control of an Lhcf2 promoter and the guide RNA was placed under the control of the endogenous U6 snRNA promoter (Nymark et al., 2016a).

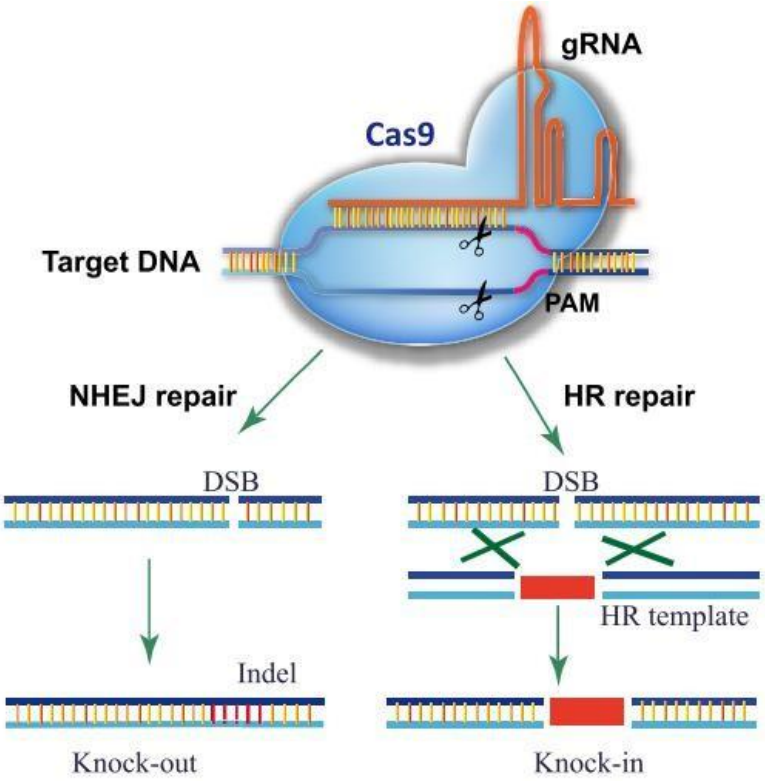

Figure 8. Schematic of Cas9/gRNA genome editing.adapted from (Ding et al., 2016) Cas9 is directed to its DNA target by base pairing between the gRNA and DNA. A PAM motif downstream of the gRNA-binding region is required for Cas9 recognition and cleavage. Cas9/gRNA cuts both strands of the target DNA, triggering endogenous DSB repair. For a knockout experiment, the DSB is repaired via the error-prone NHEJ pathway, which introduces an INDEL at the DSB site that knocks out gene function. In a knock-in experiment, the DSB is repaired by HDR using the donor template present, resulting in the donor DNA sequence integrating into the DSB site.

In order to avoid the random integration and long-term expression of Cas9 nuclease with no ability to perform outcrossing, since *Phaeodactylum* exhibits only asexual reproduction, a DNA-free genome-editing approach was recently developed. This method relies on the simultaneous co-delivery of multiple CRISPR-Cas9 ribonucleoproteins (RNP) by biolistic, one targeting an endogenous gene for which inactivation confers positive selection, and the others targeting genes of interest (Serif et al., 2018).

Genome editing in diatoms can be complicated due to that most DSBs are repaired faithfully. Therefore, the initially transformed cell is not necessarily immediately mutated and each resulting colony is a mixed population of cells with or without mutations and with mutations of different types (mosaicism). Consequently, a subsequent additional streaking step is required (Figure 9)(Huang and Daboussi, 2017).

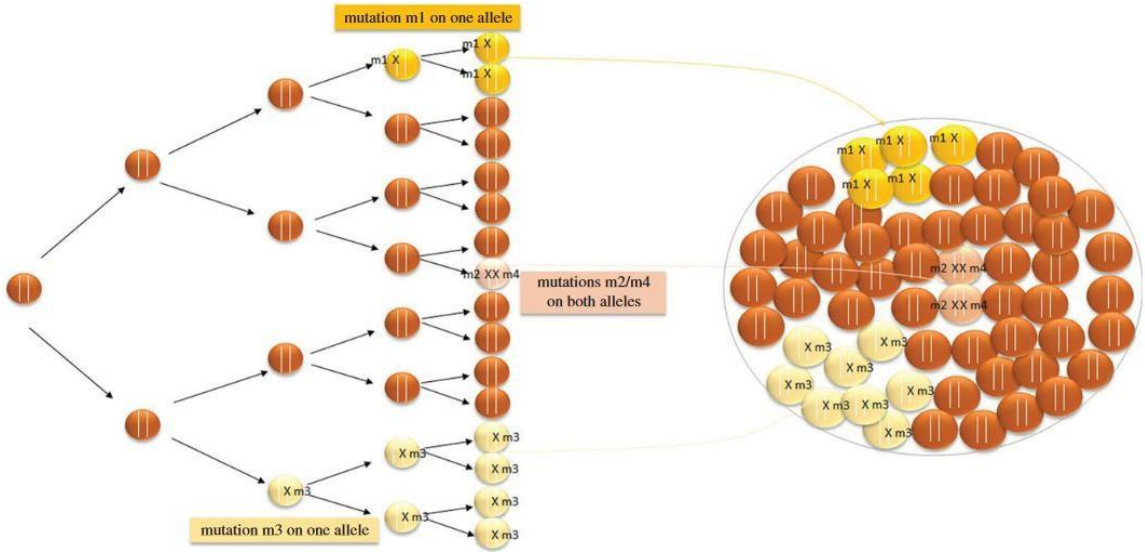

Figure 9. Illustration of the mosaicism concept within a colony obtained from transformation with an engineered nuclease, adapted from (Huang and Daboussi, 2017). The majority of the double-strand breaks induced by the nucleases are repaired by faithful re-ligation such that there is no mutation. However, some double-strand breaks are repaired by the non-homologous end-joining (NHEJ) mechanism, in which the broken chromosomes are rejoined, often imprecisely, thereby introducing nucleotide changes at the break site. Consequently, each colony is a mixed population of cells with or without mutations and with mutations of different types (the mutation m1, m2, m3 and m4 can be different). The inactivation of one gene requires a mutagenic event in both alleles, which is not the most frequent case observed within a colony. In addition, the simultaneous introduction of nuclease and a DNA template with sequences displaying similarity to the targeted sequence leads to the formation of colonies harbouring a mixed population of cells with or without integrating DNA matrix and with or without mutation induced by NHEJ.

1.4. RNAi in *P. tricornutum*

A review of Cerutti et al. (2011) summarized what was currently known about the existence of the core components of the RNAi machinery in algae, while RNA-mediated silencing processes was implicated in defense mechanism against transposon mobilization in several algal species (Maumus et al., 2009). The RNAi pathway is often exploited in experimental biology to down-regulate target genes, and since it does not totally abolish expression of the gene, this technique is usually referred as generating a "knockdown", to distinguish it from "knockout" procedures of genomic mutagenesis.

1.4.1. A functional RNA silencing mechanism

The presence of an efficient RNA silencing mechanism in diatoms was first described in *P. tricornutum* by the group of Angela Falciatore (De Riso et al., 2009). In this seminal work, they showed that a GUS reporter gene expressed in a transgenic line can be successfully silenced following ectopic expression of homologous sequences in inverted repeat (IR) or antisense (AS) orientations. GUS transgene seemed to be silenced by PTGS and/or TGS mechanisms since it presented lower levels of GUS transcripts but also de novo methylation primarily in its targeted loci (targeted by AS and IR constructs). Additionally, they demonstrate that two endogenous *P. tricornutum* genes, encoding phytochrome (Dph1) and cryptochrome/photolyase family 1 (CPF1), can also be silenced by expressing homologous sequences in IR orientation, by PTGS mechanisms including translational arrest and mRNA degradation, respectively. The same study presented an i*n silico* characterization of one putative Dicer-like, one AGO gene and one RDR gene of this species (De Riso et al., 2009). Since then, many groups have used this approach to downregulate genes of interest (Bailleul et al., 2010; Huysman et al., 2013; Trentacoste et al., 2013; Claycomb, 2014; Levitan et al., 2015; Yang et al., 2016) and the successful RNAi-mediated silencing was achieved by artificial miRNAs, too (Kaur and Spillane, 2015).

1.4.2. small RNAs

Analysis of *P. tricornutum* sRNAs repertoire has been performed either exclusively by Next Generation Sequencing data (Lu, Y. Z. and Liu, 2010; Huang et al., 2011), or in combination to experimental validation as well (Rogato et al., 2014). In this last study, the most abundant sRNAs population corresponded to 25-30 nt sRNAs principally mapped to TEs. Although the introduction of artificial miRNA in *P. tricornutum* successfully silences an endogene (Kaur and Spillane, 2015), canonical miRNAs were not found, apart from two putative candidates whose predicted precursors could not be experimentally confirmed (Rogato et al., 2014). The presence of tRNA-derived sRNAs (tsRNA) was also detected. These tsRNA were represented mainly by tRNA fragments (tRFs) of 19nt, while some longer tRFs of 30–35 nt were detected mostly under iron starvation.

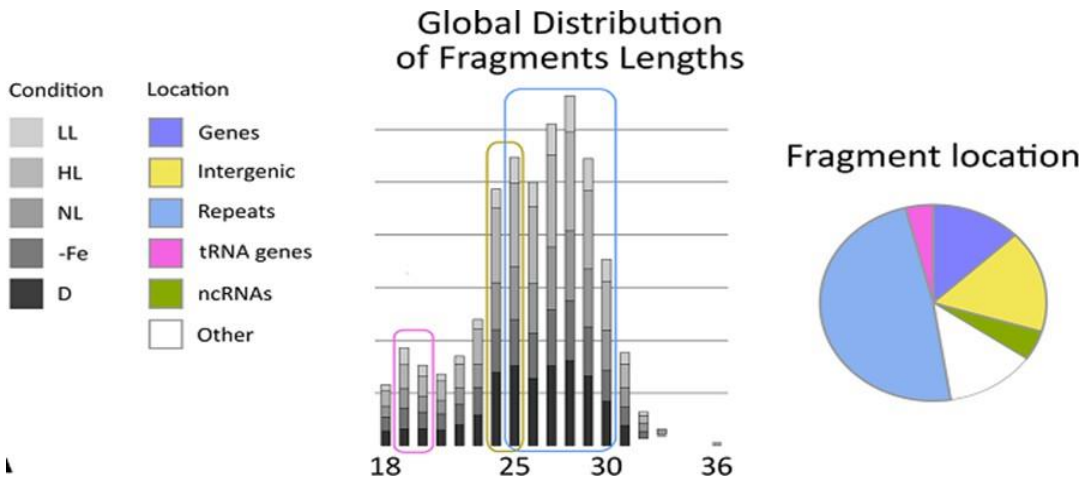

Figure 10. Small RNAs in P. tricornutum, adapted from Rogato et al. (2014)Fragment lengths distribution of reads (histogram, center) is reported in a grey color scale distinguishing the five experimental conditions (LL, HL, NL, −Fe, D). The distribution of fragment location is also reported (pie chart, right) with a color scale indicating genes, intergenic regions, repeat regions, tRNA genes, ncRNAs and other loci.

Interestingly, some of these 25-30 nt sRNAs display a 180 nt-long periodic distribution at several methylated locations in the genome of *P. tricornutum* (Rogato et al., 2014) suggesting a possible nucleosome-related distribution of sRNAs as described

previously for DNA methylation (Veluchamy et al., 2013). The analysis of genome-wide DNA methylome in this diatom revealed also the potential role of DNA methylation in regulating gene expression, transposon mobilization under specific growth conditions (Veluchamy et al., 2013). Thus, these sRNAs may be involved in the repression of genes and TEs by RNA directed DNA methylation (RdDM) and play an important role during the acclimatory response to environmental stress (Rogato et al., 2014; Veluchamy et al., 2013).

Conclusively, the existence of an RNAi mechanism has been proven, but its basic players, their function and the possible RNA silencing pathways in which they contribute, remain to be discovered.

AIM OF THIS THESIS

The aim of this thesis was primarily to identify the direct role of *DCR* and *AGO* genes in the RNAi and RdDM pathways in *P. tricornutum* through a reverse genetic approach, as well as to unravel the RNA silencing physiological roles in this model organism. To this end, *P. tricornutum DCR* and *AGO* gene candidates were cloned and their activity was assessed through heterologous expression in yeast *Saccharomyces cerevisiae* and the plant *Nicotiana bethamiana*. In parallel, we generated *P. tricornutum* knock-out lines for *DCR* and *AGO* genes by performing CRISPR/Cas9-mediated mutagenesis. The phenotypic characterization of the CRISPR/Cas9 mutants included analysis of growth under a) normal conditions, b) nitrate limitation, and c) after UV-induced damage, and analysis of the large and small RNAs in wild type and mutant lines by next generation sequencing. The later analysis provided insights into the effect of DCR presence/absence to small RNA populations and their correlation to differential gene/TE expression between wild type and mutant lines. Furthermore, a phylogenetic analysis of DCR, AGO and RDR proteins from diatoms and other species was contacted, in order to elucidate their placement onto the evolutionary pathway. Finally, their subcellular localization was investigated by introducing DCR, AGO and RDR genes fused to YFP chromophore into *P. tricornutum* and study them with confocal microscopy.

2. Materials and Methods

All the organisms and viroids used in this study:

Diatom

Phaeodactylum tricornutum CCMP623

Yeast

Sacharomyces cereviciae

Plants

Nicotiana bethamiana

Bacteria

Agrobacterium tumefaciens C58C1

Escherichia coli DH10b, JM109, STELAR

Viroid

Potato spindle tuber viroid (PSTVd)

2.1. Media, Cultures, Transformation and other procedures

2.1.1. Diatoms

2.1.1.1. Media, Cultures, Cell counting and Harvesting

Liquid media *F/2* medium: modified from Guillard, 1975

Fresh Sea Water (FSW) was provided by HCMR institute, where natural sea water is filtered through rock and passes through UV radiation before collection. The collected FSW got diluted with DDW in order to reach salinity of 32-33‰ and placed in 1Lt glass bottles before autoclave.

Since *P. tricornutum* is not heavily silicified its cultures do not require the addition of Na_2SiO_3, $9H_2O$. In this case a medium of F/2-Si was prepared. Components of F/2 medium are listed in Tables 1,2 and 3.

Table 1. Composition of F/2 medium.

Reagents	Final concentation	Working Stocks	in 1Lt
Fresh Sea Water (FSW), Salinity ~40‰	80%		800ml
Double Distilled Water (DDW)	20%l		200ml
Autoclave and the add from the working stocks:			
NaNO3	1 mM	1M	1ml
NaH2PO4,H20	36 μM	36mM	1ml
Na2SiO3,9H2O	0,1 mM	106mM	1ml
Trace Metals stock CuSO4,5H2O ZnSO4,7 H2O CoCl2,6 H2O MnCl2,4 H2O Na2MoO4,2 H2O FeCl3 6H2O Na2EDTA 2H2O	 19nM 38.25 nM 21nM 0.65μM 14.4nM 0.117 μM 0.117 μM		1ml
Vitamins stock Biotin B12 Thiamine	 0.4 nM 0.73 nM 0.59 μM	 0.4μM 0.73μM 0.59mM	0,5ml

Table 2. Preparation of trace metals primary stocks x 1000.

Reagents	**Primary Stock**	**1Lt**
DDW	-	Up to 1Lt
$CuSO_4,5H_2O$	1,9 g in 200 ml (=38mM)	1ml
$ZnSO_4,7\ H_2O$	4,4 g in 200 ml (=76.5mM)	1ml
$CoCl_2,6\ H_2O$	2 g in 200 ml (= 42mM)	1ml
$MnCl_2,4\ H_2O$	36 g in 200 ml (=1.30M)	1ml
$Na_2MoO_4,2\ H_2O$	1,26 g in 200 ml (=28.8mM)	1ml

Table 3. Preparation of vitamin working stocks.

	Primary Stock	1Lt
DDW		Up to 1Lt
Biotin	10 mg in100 ml (=0.4 mM)	1ml
B12	10 mg in 10 ml (=0.73mM)	1ml
Thiamine Hcl		200 mg

Primary stocks were stored at -20°C. Working stocks were filter-sterilized, aliquoted in 1.5 ml tubes and stored at -20°C.

Liquid Cultures

Phaeodactylum tricornutum Bohlin CCMP632 strain was obtained from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton. Cultures were grown in glass flasks with F/2-Si medium at 18-20ºC under white fluorescent lights (80 mmol m–1 s–1) and a 12 h:12 h dark–light cycle.

Cell Counting

15 μl of culture were placed on Malassez cell for counting cells under a light microscope. If needed, diluted 1/10 (100 μl culture in 900 μl *f*/2 medium).

Harvesting

Diatom cells were collected simultaneously after reaching the exponential phase of growth and optimally 4 h after the beginning of the light period. Harvesting was performed by:

a) Filtration: used Glass fiber filters, Whatman GF/C (pore 1.2um, diameter 4.7cm)

b) Centrifuge: Cultures were centrifuged in 50ml falcons at specific velocity and time, according to the downstream processes.

Solid media

Media of F/2 agar plates were prepared as described in Table 4.

Table 4. Composition of F/2 agar plates.

***f/2* medium (1Lt)**	**100% Sea Water**	**50% Sea Water**
DDW	200ml	600ml
NaNO3 (stock 1M)	1 ml	1 ml
NaH2PO4,H20 (stock 36mM)	1 ml	1 ml
Na2SiO3X9H2O (stock 106mM)	1 ml	1 ml
Trace Metals stock	1 ml	1 ml
Fresh Sea Water	~800ml	~ 400ml

Autoclaved. 0.5 ml from vitamins stock was added, as well as antibiotics if needed, before pouring the medium in petri dishes (20-25ml/dish).

Cultures on agar plates

From a *P. tricornutum* culture in log phase a new subculture was prepared. When subculture reached log phase (1-2,5 x10^6 cells/ml concentration) cells were centrifuged in 50ml falcon, at 3000 rpm for 15 min (or 25000 rpm, 20 min). Supernatant was discarded, leaving only a drop of medium in the falcon, for the cells to get resuspended. By using a sterile pipette, cells were transferred on a plate of F/2 50% SW or F/2 100% SW and gently spread with a spatula (made by glass Pasteur pipettes). After getting dry under the hood, plates were sealed with parafilm and incubated under optimal conditions (18-20°C, 12:12 light:dark). Growth was obvious after 5-10 days.

2.1.1.2. Contamination Test

1 ml of diatom cultures was added in a 50ml falcon containing 10 ml of the autoclaved medium described in Table 5. Tubes were put in a dark box and checked for bacterial growth after 2-3 days and after 1-2 weeks.

Table 5. Medium F/2 for contamination test (500ml)

Reagents	**Final volume**
DDW	100ml
NaNO3 (stock 1M)	0,5 ml
NaH2PO4, H20 (stock 36mM)	0,5 ml
TRIZMA stock (0.413 M)	2,5 mλ
Peptone	0,5 g
Fresh Sea Water	Up to 500ml

TRIZMA working stock:

5% TRIZMA (25g in 500 ml DDW) titrate to pH 7,8 with H2SO4 (=0.413 M)

2.1.1.3. Cryopreservation

1 to 10 ml of diatoms culture in exponential phase (1-2x10^6 cell/ml) were centrifuged and then resuspended in 900 μl *f/2* medium and 100 μl DMSO (10% final). Tubes were put at 4°C for 1 hour, at -20°C for 1 hour and finally at -80°C.

2.1.1.4. Transformation via Biolistic Bombardement

Biolistic particle delivery is a method of transformation that uses helium pressure to introduce DNA-coated microcarriers into cells. In this study, the Biolistic PDS-1000/He system uses high pressure helium, released by a rupture disk, and partial vacuum to propel a macrocarrier sheet loaded with millions of microscopic tungsten toward target cells at high velocity. The microcarriers are coated with DNA for transformation. The macrocarrier is halted after a short distance by a stopping screen. The DNA-coated microcarriers continue traveling toward the target to penetrate and transform the cells. The launch velocity of microcarriers for each bombardment is dependent upon the helium pressure (rupture disk selection), the amount of vacuum in the bombardment chamber, the distance from the rupture disk to the macrocarrier, the macrocarrier travel distance to the stopping screen, and the distance between the stopping screen and target cells.

Protocol was adapted to transform *P. tricornutum* (Apt et al., 1996; Falciatore et al., 1999a). Rupture disks 1350 psi were used, vacuum in the bombardment chamber was 25 inHg (where 30inHg=100% vacuum), the macrocarrier launch assembly was set by inserting all extra rings between macrocarrier and stopping screen and the target shelf with the cells to be transformed was placed at L2.

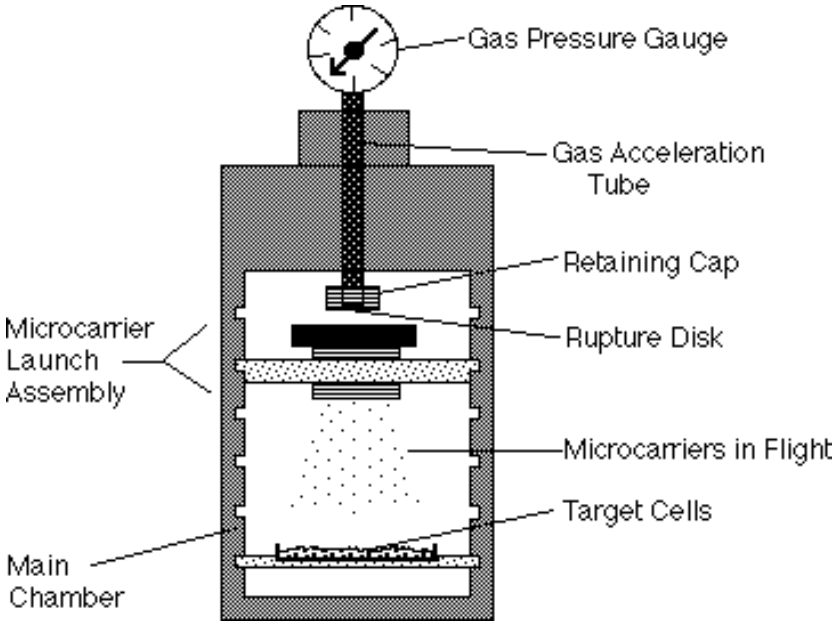

Figure 11. GeneGun chamber used for transformation of *P. tricornutum* via biolistic bombardment.

P. tricornutum was grown in exponential growth phase 1-2,5 x10^6 cells/ml. Preferably cultures with cell concentration of 2 x10^6 cells/ml were used.

In order to collect and transform 50x10^6 cells:

The appropriate volume of culture was centrifuged in 50 ml falcon at 3000 rpm for 15 min (or 2500rpm, 20min). Supernatant was discarded and cells were resuspended in a remaining drop of medium in the falcon. By using a sterile pipette, cells were transferred on a plate of f/2 50% SW and gently spread on the middle of the plate with a spatula (1/3 of the surface, where they were left to dry). Plate was incubated for some hours or overnight in growing chamber, at 18°C, 12:12 light:dark until the time of shooting.

In the present study, co-transformation was employed, with at least two plasmids introduced, one or more carrying the genes of interest and another one carrying the selective marker. Tungsten particles M17 were used for coating with plasmids. Tungsten particles M17 were prepared according to manufacturer's directions, dispensed into aliquots of 3mg in 50μl and stored at -20°C.

For the coating procedure, tungsten aliquots were thawed on ice before the addition of DNA (2μg of each plasmid in 5-20 μl DDW), 50 μl 2.5 M $CaCl_2$ and 20μl of 0.1 M spermidine under continuous vortexing for 3 minutes. Tubes were spin down, supernatant was discarded and coated microparticles were resuspended in 250 μl 100% ethanol. After another round of centrifuge for 1 minute at top speed and removal of supernatant, coated particles were resuspended in 60 μl 100% ethanol and placed on ice until the bombardment (1-2 hours max).

The GeneGun chamber, the Laminar Flow surface and all the tools were sterilized with 70% ethanol and UV radiation. Rupture discs (1350 psi) and macrocarriers were briefly dipped in 100% ethanol and left to dry. Stopping screens and macrocarrier holder were autoclaved.

The bombardment was performed by following the manufacturer's instructions. Every f/2 50% plate was shot 3 times in slightly different places by dividing the coated microparticles in 3 parts and loading 20μl at the center of the macrocarrier each time. After transformation, plates were sealed with parafilm and incubated in the growth chamber with 18°C and 12:12 light:dark for 48 hours, in order to recover.

Finally, cells were collected from the plate by resuspension in 1 ml f/2 liquid

medium and then gently spread on 2 selective plates 50% SW f/2 containing 300 μg/ml NAT (or 8 μg/ml Blasticidin). After 2-3 weeks the first colonies of positive clones started to appear. Transformants were kept on plate or liquid f/2 medium containing 100 μg/ml NAT (or 8 μg/ml Blasticidin).

2.1.2. Yeast

2.1.2.1. Media and Cultures

All yeast media preparation followed the protocols from Treco and Lundblad, (1993).

Yeast cells were cultured in:

1) YPD medium, composed of 1% Yeast extract, 2% Peptone and 2% D-(+)-glucose monohydrate

2) complete minimal medium, composed of 0.13% dropout powder, 0.67% Yeast Nitrogen Base without amino acids and 2% D-(+)-glucose monohydrate or 2% galactose and 1% raffinose.

Dropout powders had all essential amino acids, based on each yeast clone's auxotrophy. pH was set at 5,8-6,2 with addition of NaOH. Solid media had also 2% agar-agar. Yeast cultures were incubated for 2-3 days at 30ºC and liquid cultures were also under orbital shaking at 200-250 rpm.

2.1.2.2. Yeast Trasnformation with Lithium Acetate

In the present study, yeast was transformed by using the lithium acetate method, which is based on the fact that alkali cations make yeast competent to take up DNA. Protocol modified from Becker and Lundblad (1994).

Grow and prepare yeast cells

Two days before the experiment, 5 ml YPD medium were inoculated with a single yeast colony of the strain to be transformed and grown overnight at 30°C to saturation. 50 ml YPD medium in a 250ml sterile flask was inoculated with the appropriate amount of the saturated culture in order to have an OD600= 0,2. Grew for ~3 hours (depending on the strain) in order to reach OD600= 0.5-0.6. Cells were centrifuged for 5 min at 4000 × *g* in room temperature. Supernatant was discarded, pellet was resuspended in 10 ml DDW and cells were centrifuged again for 5 min at 4000 × *g* in room temperature.

Supernatant was discarded and pellet was rinsed with solution1 (1:1:8, 1 TE buffer 10x, pH 7.5, 1 lithium acetate stock solution 10x, 8 sterile water). Cells were centrifuged for 5 min at 4000 × *g* in room temperature, supernatant was discarded and cells to be transformed were resuspended in 400μl solution1 per 50ml culture (cells can be stored in solution 1 for ~3 hours).

Transform yeast cells

For each transformation, 200μg denaturated carrier DNA were mixed with ≥5 μg transforming DNA in a sterile 1.5-ml microcentrifuge tube. Total volume of DNA was kept ≤20μl. To each tube 80μl yeast suspension were added along with 350μl PEG solution, freshly prepared (8:1:1, 8 PEG 50%, 1 TE buffer 10x, pH 7.5, 1 lithium acetate stock solution 10x). Mix was incubated for 45 min at 30 ^{o}C with shaking and then heat shocked exactly for 15 min at 42°C. After a centrifugation for 2 min at 2000rpm at room temperature, supernatant was discarded and yeast cells were resuspended in 150μl of TE buffer 1x. Finally, cells were spread onto selective agar plates and incubated at 30°C for 2-3 days.

2.1.3. Bacteria

2.2.3.1. Media and Cultures

The bacterial strains used in this study (*E. coli*: DH10b, JM109, STELAR and *A.tumefaciens*: C59C1) were grown in/on LB medium (1% Tryptone, 1% NaCl, 0,5% Yeast extract and for solid media 1.5 % agar-agar) with antibiotics was added as selective markers when appropriate. Cultures were incubated overnight at 37 °C for *E.coli* and for at least 48 hours at 28 °C for *A.tumefaciens.* Liquid cultures were incubated with orbital shaking at 200-250 rpm.

2.1.3.2. E.coli

2.1.3.2.1 Transformation of chemically competent E.coli cells

An aliquot of 100μl chemically competent cells was thawed on ice and mixed with the appropriate amount of plasmid to be transformed. The bacteria/DNA mixture was incubated on ice for 20 minutes, incubated in a 42 °C for 30-45 seconds, and quickly chilled on ice for 2 minutes. Then, 900 μl of room temperature LB was added, followed

by 1 hour incubation at 37°C with shaking at 250 rpm. The transformed bacteria were centrifuged at 2000g for 1-2 minutes, 900μl of supernatant was discarded and the remaining 100μl with all resuspended cells were spread on selective LB-agar plates and incubated over night at 37 °C.

2.1.3.2.2 Preparation of chemically competent of E.coli cells

A single colony of a chemically competent *E.coli* strain that has grown on LB-agar plate was inoculated in a 3 ml LB medium and incubated at 37°C overnight with shaking. From this preculture, 2 ml were inoculated in 200 mL of LB medium in a 1Lt flask and incubated at 37°C with shaking until O.D.600 reached 0.6-0.8 (2-3 hours). Then, culture was centrifuged for 10 min at 210g and at 4°C. After discarding the supernatant, the cell pellet was resuspended in 60 ml cold solution 1 (CH3COOK 30mM, RbCl 100mM, MnCl24H2O 50mM, CaCl22H2O 10mM, glycerol 15%, pH=5,8 with CH3COOH) and incubated in ice for 30 min. Next, the mixture was centrifuged for 10 min at 210g and at 4°C, supernatant was discarded and the cell pellet was resuspended in 8 ml solution2 (MOPS 10mM, CacL22H20 15mM, RbCl 10mM, glycerol 15%, pH=6.5 with KOH). Cells were dispensed into aliquots of 100 μL in 1.5 mL sterile eppendorfs, flash-frozen in liquid nitrogen and stored at -80 °C.

2.1.3.3. Agrobacterium tumefaciens

2.1.3.3.1. Transformation of A.tumefaciens cells

An aliquot of 100μl chemically competent cells was thawed on ice and after the addition of the appropriate amount of plasmid (preferably 5μg of plasmid), it was flash-frozen in liquid nitrogen for some seconds. Then, the cells were incubated at 37°C for 30 minutes. After the addition of 900 mL LB medium, the cells were incubated at 28 °C with shaking for 2 hours and then centrifuged for 2 min at 2000g. Supernatant was discarded and the remaining 100μl was spread on LB/rifampicin plates containing appropriate selective antibiotics and incubated for 2 days at 28 °C.

2.1.3.3.2. Preparation of chemically competent of A.tumefaciens cells

A single colony of a chemically competent *A.tumefaciens* strain that has grown on LB-agar plate with rifampicin (100 μg/ml) was inoculated in 3ml LB liquid with rifampicin (100 μg/ml) and incubated at 280C overnight with shaking. 2ml of this

preculture were inoculated in 50ml LB with rifampicin (100 μg/ml) and incubated at 28ºC with shaking until O.D.600 reaches ~0.6 (6-7 hours). Cells were centrifuged for 10 min at 210g and 4ºC, supernatant was discarded and the cell pellet was resuspended in 1 ml of cold Solution1 (20 mM CaCl2, pH=5,7). Finally, cells were dispensed into 100 μl aliquots, flash-frozen in liquid nitrogen and stored at -80 °C.

2.1.4. Plants

2.1.4.1. Media and Cultures

N. benthamiana plants were cultivated under greenhouse conditions. Seeds were sown on potting soil and covered with plastic bags to create a high humidity environment for germination. At the two-leaves stage, seedlings were separated to individual pots and again covered with plastic bags. Over the course of 1-2 weeks the bags were slowly opened to facilitate the acclimation of the plants to normal humidity levels. Once fully accustomed, individual plants were re-potted to fresh soil (consisting of 2 parts potting soil: 1 part peat moss: 0,5 part Perlite and fertiliser) and grown to maturity.

2.1.4.2. Agroinfiltration

Agroinfiltration is a technique used to ectopically express a gene or a pathogen of choice in plants, via *A. tumefaciens* cells. Agro-infiltration-induced expression is confined to the infiltrated space(s) and does not spread. In this way, specific regions can be exposed to expression of the transgene, while adjacent regions in the same leaf will not be affected and may serve as controls.

A liquid culture of transformed agrobacteria that contain the sequences of interest was incubated at 28ºC with shaking for ~2 days. Cells were centrifuged for 10 minutes at 2.700 rpm and 4ºC and then supernatant was discarded. The cell pellet was resuspended in an equal volume of MMA (MS 1X, 10 mM MES pH=5.7 and 200 μM AcS) and incubated for at least 1 hour at 28ºC with shaking. Then, cells were centrifuged for 10 minutes at 2.700 rpm and 4ºC, supernatant was discarded and pellet was resuspended in half volume of 10 mM MgCl2. This step was repeated two more times. Finally, cells were resuspended in the appropriate volume of 10 mM MgCl2 in order to have an $O.D_{600}$=0.2-0.5, depending on the experimental requirements.

This working suspension of *A.tumefaciens* was injected into leaves of *Nicotiana benthamiana* using sterile 1 mL syringes. Optimally, plants that were used for agroinfiltration had not been watered the previous 6-8 hours and they were watered right after agroinfiltration. When transformed agrobacteria carried GFP sequences alone or fused to the genes of interest, GFP expression was visible after ~2 days under UV radiation and was captured with a Nikon D5100 camera.

2.1.4.3. PSTVd (viroid) infection in Nicotiana bethamiana

Infection with *Potato spindle tuber viroid* (PSTVd) strain PH106 was induced in young *N.benthamiana* plants that were agroinfiltrated at the stage of 4-6 leaves, in order to achieve greater efficiency of infection and the presence of characteristic phenotype. Full systemic infection of the plants established approximately 3-7 weeks post infection and persisted until senescence.

2.2. Basic Molecular Biology techniques

2.2.1. Polymerase chain reaction

The enzyme used for PCR amplification was Taq polymerase (Minotech) and, when needed, High Fidelity Polymerase Kapa Hi-Fi (Kapa Biosystems) according manufacturer's instructions. PCR primers were designed using the Oligo Analyser.

2.2.2 cDNA synthesis by reverse transcription

2-5 μg RNA was used a template to synthesize cDNA by using the reverse transcriptase PrimeScript™ RT-PCR kit (TaKaRa) according to the manufacturer's protocol. Primers were random exams or oligod(T) (Invitrogen) .

2.2.3. Agarose gel electrophoresis

Gels with an agarose concentration of 0.6-2.5 % (w/v) were prepared from 1x TAE and a final concentration of ethidium bromide 0,01% (w/v). Gels were run in 1x TAE running buffer at 20-100 V depending on gel size and application. DNA/RNA bands were visualized using a UV-based documentation system.

2.2.4. DNA Isolation from agarose gel

DNA was extracted from agarose gels by using the kit "NucleoSpin®Gel and PCR Clean-up" (Macherey-Nagel).

2.2.5 Ligation of DNA fragments

Standard ligation reactions were performed by incubating insert DNA fragments and restriction enzyme-cut vector at a ratio of 3:1 with addition of ligase (T4 DNA ligase, Promega) and buffer at 1x concentration at 4 °C overnight.

2.2.6. Plasmid preparation-mini scale

Plasmid preparations (mini scale) were performed using a NID prep protocol (Lezin et al., 2011) or the kit "NucleoSpin® Plasmid" by Machery-Nagel according to the manufacturer's instructions.

2.2.7. Plasmid preparation-midi scale

For larger scale plasmid preparations the NucleoBond® Xtra Midi Kit by Macherey-Nagel was used according to the manufacturer's instructions.

2.2.8. Cleavage of dsDNA with restriction endonucleases

Endonucleases from Minotech or New England Biolabs were used for site-specific cleavage of dsDNA, according to manufacturer's protocol. Typically 1 μg of DNA was incubated with 1 u of a restriction enzyme for 2h at the enzyme's activity optimal temperature.

2.2.9. Phenol extraction and Ethanol Precipitation of nucleic acids

At the DNA solution to be purified an equal volume of phenol:chloroform:isoamyl alcohol (PCI, 25:24:1) is added and then it gets centrifuged for 5 min at 12000g. The top aqueous phase containing the DNA is transferred in a new tube, gets mixed with an equal volume of chloroform:isoamyl alcohol (CI, 24:1) and gets centrifuged for 5 min at 12000g. 1/10 volume of NaAc 3M, pH 5,2 (sodium acetate) and 2,5 volumes of 100%

ethanol (-20oC) is added in a new tube containing the top aqueous phase and then placed at -20oC for 30 minutes to overnight (or in -80oC for 30 min) . Tubes were then centrifuged for 20 min at maximum speed at 4oC and supernatant was removed. Pellet was washed with 70% ethanol, briefly centrifuged to remove supernatant and then let to dry before resuspension in the appropriate volume of water.

2.2.10. Gateway (Invitrogen) Cloning system

The Gateway system was used for cloning DNA constructs in vector pENRT3C (Invitrogen) and then transfer them in specific destination (expression) vectors through recombination mediated by LR Clonase II (Invitrogen), according to manufacturer's instructions.

2.3. Diatom DNA extraction protocol with CTAB

Frozen cell pellets (up to 300mg) were resuspended in 750μl of CTAB lysis buffer (2% CTAB, 100 mM Trsi pH=8,0, 20 mM EDTA, 1.4M NaCl, 1% PVP-40) after the addition of 2-mercaptoethanol and Proteinase K at final concentration of 0,2% and 100μg/ml, respectively. Tubes were incubated at 60°C for 1hour and then briefly cooled on bench. An equal volume of phenol: chloroform: isoamylalcohol (PCI, 25:24:1) was added, tube was centrifuged for 10 min at 13000g and the upper aqueous phase was transferred to a new tube. This extraction step with phenol was repeated and then followed the addition of an equal volume of chloroform: isoamylalcohol, centrifugion for 10 min at 13000g and transfer of the upper aqueous phase to a new tube. 2/3 volume of -20°C isopropanol was added, mixed and put to -20°C overnight for precipitation of DNA. Then, tubes were centrifuged at 13000 g for 20 min at 4°C and supernatant was removed. The DNA pellet was washed once with 70 % ethanol and subsequently resuspended in 500 μL water containing 100 μg/mL RNase A. The DNA/RNase A solution was incubated at 37 °C for 30-60 minutes, followed by standard phenol/chloroform extraction and isopropanol precipitation. The RNase A-treated DNA pellet was finally resuspended in 100 μL water and the DNA concentration was determined using the NanoDrop® spectrophotometer.

2.4. RNA extraction

2.4.1 *Nicotiana bethaniana* RNA extraction with TRIZOL

Total RNA from plant material was extracted following TRIzol method (Chomczynski and Sacchi, 1987). 100 mg of grounded tissue was resuspended in 1 mL Trizol reagent (phenol pH:7 38%, Guanidine thiocyanate 0.8M, Ammonium thiocyanate 0.4M, Sodium acetate pH:5 0.1M, glycerol 5%) by vortexing for 30 seconds. After incubation for 10 minutes, cellular debris was pelleted by centrifugation at 12.000g for 10 minutes and supernatant was transferred in a new tube. Supernatant was mixed with 200 μL chloroform by vortex, incubated for 10 minutes, and centrifuged at 12.000g for 10 minutes. The clear supernatant was mixed with 500 μL isopropanol and incubated at -80 for 1 hour, in order to precipitate the small RNAs, too. After centrifugation at 12.000 g for 10 minutes at 4°C, the RNA pellet was washed once in 70 % ethanol, centrifuged at 12.000 g for 5 minutes and subsequently resuspended in 40 μL of water. RNA's concentration was estimated by a spectrophotometer (ND-1000, NanoDrop). In cases where higher clarity of RNA was needed, the purification protocol of phenol and chloroform was used (described above).

2.4.2 Phaeodactylum tricornutum RNA extraction

a) Total RNA for cDNA synthesis and Northern blot analysis was extracted from 100mg cell pellet (from culture in log phase, ~2x106 cells/ml) according to the Trizol method, as described above.
b) Total RNA for preraration of DNA libraries of large and small RNAs was extracted from 35-45mg grounded cells with the "Total RNA Purification" kit from Norgen. DNaseI treatment was performed on column (Norgen).

2.4.3 *Sacharomyces cereviciae* RNA extraction with Hot Acidic Phenol

Total RNA from yeast for Northern blot analysis was isolated from intact cells by extraction with acidic phenol (pH 5) and SDS at 65°C, as described from Collart and Oliviero, 1993.

2.5. Northern analysis

Northern blot technique was employed for comparative gene expression analysis. RNA samples were size-fractionated on denaturing gels, transferred to nylon membranes, hybridized with specific radioactively labeled probes and imprinted in film.

2.5.1. Denaturing agarose gel electrophoresis for mRNA adetection

10-20 μg total RNA with 5x RNA loading dye containing bromophenol blue, boiled for 5 min and cooled on ice, were size-fractionated on 1.2 % agarose gels (Table 6). Gels were run at approximately 110 W in 1x MOPS with formaldehyde. For the capillary blotting procedure 10x SSC (1.5 M NaCl, 150 mM Sodium citrate, pH=7) was used while the gel and (soaked in 2xSSC membrane and appropriately sized whatman were assembled. RNAs transferred in nylon membrane (pore 0.45 μm) were crosslinked with Stratalinker device (1200 mJ/cm², radiation 254 nm, UV crosslinker, Stratagene).

Table 6. Solution composition for electrophoresis of denaturating agarose gel

Solution	Composition
5X loading dye	40 μl 0,5 M EDTA, pH 8.0, 360 μl 37% formaldehyde, 1 ml 100% glycerol, 1542 μl formamide, ·2 ml 10 x MOPS, 0,25% β/ο xylene cyanol και 0,25% w/v bromophenol blue, ddH2O up to 5 ml
10X MOPS	200 mM MOPS, 50 mM CH3COONa, 10 mM EDTA, pH=7 with NaOH
Denaturating agarose gel	1X MOPS, 0.7% formaldehyde, 0,035μg ethidium bromide/ml gel
Running buffer	1X MOPS, 0.7% formaldehyde
20X SSC	3 M NaCl, 0,3 M citric acid, pH=7,0

2.5.2. Denaturing polyacrylamide gel electrophoresis (PAGE) for small RNAs detection

20-60 μg total RNA mixed with 2x RNA sample loading dye (PAGE), boiled for 5 min and cooled on ice, was size-fractionated on denaturating polyacrylamide gels (14%

acrylamide: bisacrylamide 38:2, 7 M Urea, 1x TBE (0.2 M Tris, 0.2 M boric acid, 4 mM EDTA), 250 μL APS, 25 μL TEMED). Gels were run initially at 22 W and after 40 min at 35V in 1x TBE, keeping the gel temperature at 50 °C. RNA was blotted onto nylon membranes (pore 0.2μm) by semi-dry blotting in an SD20 Semi Dry Midi unit. Gel, membrane and appropriately sized whatman paper soaked in 1x TBE were assembled and RNA transfer was achieved by applying conductive surface of 2 mA/cm2 for 35 min at 4ºC. Finally, membranes were crosslinked with Stratalinker device (1200 mJ/cm², radiation 254 nm, UV crosslinker, Stratagene).

The RNA transfer was validated by incubation for a few minutes in a methylene blue solution (0,03% methylene blue, 0,3 M acetic acid ph=5,2) and then membrane was ready for hybridization with the appropriate probe.

Table 7. Solution composition for electrophoresis of denaturating polyacrylamide

Solution	**Composition**
2X loading dye	80% formamide, 10 mM EDTA pH=8, 1 mg/ml bromophenol blue, 1 mg/ ml xylene cyanol
5X TBE	1,1 M Tris, 900mM Boric acid, 25 mM EDTA, pH=8.0

2.5.3. Radioactive labeling of nucleic acid probes

For the detection of RNA sequences during Northern analysis, specific probes were labeled radioactively with ^{32}P. Template of 100 ng PCR product mixed with 3μg random primers (Invitrogen) was denaturated at 100ºC for 5 min briefly cooled down on ice. 1 μl of each nucleotide (0.5 M GTP, TTP and ATP), 20 units Klenow enzyme (Minotech), 1x Klenow buffer (125 mM Tris-HCl pH 6.8, 12,5 mM MgCl2, 25 mM β-mercaptoethanol) and 5 μl [α-32P]CTP was added to the mixture with 50 μl final volume. They reaction was incubated for 1 hour at 37oC and then got purified with Microspin, G-sephadex 25, (700ul) (Biorad) in order to discard non incorporated nucleotides. Purified DNA probes were denatured at 95°C for 5 minutes and quick-chilled on ice before being added to the hybridization buffer.

2.5.4. Hybridization, washing, and exposure of northern membranes

Membranes were pre-hybridized in pre-warmed Church hybridization buffer (Table 8) for 1 hour at temperatures dependent on the size of the RNA to be detected and/or the length of the probe used. Typically, mRNA northerns were hybridized at 65°C, while small RNA northerns were hybridized at 48-50 °C. After addition of denaturated probes, the hybridization took place for 14-16 hours and then membranes were washed twice with Washing solution 1 for 30min and 20 min and once with Washing solution 2 for 10 min. All washing steps were performed at hybridization temperature. Washed membranes were rinsed in 2x SSC, sealed in plastic bags while still wet and then exposed to X-Ray films. Exposure times were determined empirically and X-Ray films were developed automatically using a Curix 60 developer (Agfa).

Table 8. Composition of Hybridization and Washing solutions

Solutions	**for mRNA**	**for small RNA**
Hybridization	5X SSC, 1%SDS, 1XDenhardt's, 2.5mg/t RNA	5 X SSC, 7% SDS, 20 mM NαPi pH=6.8, 1x Denhardt's
Wash 1	2x SSC/0.3 % SDS	2x SSC/0.1 % SDS
Wash 2	1x SSC/0.3 % SDS	1x SSC/0.1 % SDS
10 X Denhardt's	1% BSA, 1% PVP-40, 1% Ficoll	
10 x tRNAs	10 mg/ml tRNAs	

2.6. Diatoms under optical, electron and confocal microscopy

Diatom cells were observed and counted alive with a Nikon Eclipse E800 microscope.

Diatom cells were fixed by Microscopy Lab's personnel and were analyzed with a JEOL JEM-100C Electron Microscope operating at 80 kV.

To examine the expression and subcellular localization of the proteins under study, *P. tricornutum DICER* (DCRa), *AGO* and *RDR* genes were amplified from cDNA without their stop codon, they were fused in frame with eYFP at their C-terminus and cloned in pKS_FcpB_eYFP_At vectors. WT cells were co-transformed with pKS_FcpB_(DCRa/AGO/RDR)_eYFP_At and pNptII vectors via microparticle bombardment. Clones grown on selective nourseothricin plates were validated for the presence of DCR/AGO/RDR-YFP constructs by PCR amplification of the fused area (800 bp). Cultures from positive clones were grown under optimal conditions with low light that enhances expression under FcpB promoter. 1 ml of cells at early log phase was collected after 8-9 hours from the initiation of photoperiod. Without further dilution or washes, the cells were stained 5 min before observation with 0,5μg/ml Hoechst 33342 (Invitrogen) that is permeable in live cells and stains the nuclei by biding on adenine-thymine. Cells were observed alive in the inverted Confocal Laser Scanning Microscope Leica *TCS SP8* (lasers: UV 405nm Diode, DPSS 561nm, He-Ne 633nm, Argon 458-476-488-496-514nm) and software *LAS AF 3*. Images were captured using a 63X oil objective with sequential scanning and excitation/emission settings (in nm): 405/412-490 for Hoechst, 405/650-750 for chlorophyll and 514/521-575 for YFP. Images were edited by ImageJ and were all adjusted identically.

2.7. Phylogenetic analysis

2.7.1. Selection of *DCR*, *AGO* and *RDR* genes and proteins in *P. tricornutum* and other organisms

P. tricornutum DCR, AGO and RDR proteins were searched in NCBI database (www.ncbi.nlm.nih.gov), DOE JGI database (genome.jgi-psf.org) and Ensemble database by using keywords, TBLASTN and BLASTP searches. Various queries were used, including full-length sequences or RNase III, PIWI, PAZ and RDRP domain sequences from species such as *Arabidopsis thaliana, Homo sapiens, Chlamydomonas reinhardtii, Ectocarpus siliculosus, saccharomyces castellii, Aquifex aerolicus* and more. Homologues of these proteins from other organisms were selected from the aforementioned databases and included in phylogenetic analysis.

More diatom species were searched for RNAi key-genes from the Marine Microbial Eukaryote Transcriptome Sequencing Project, *MMETSP*, (Keeling et al., 2014), which provided their predicted proteome and more recently updated versions of their transcriptome, too.

Data from MMETS are in “fasta” format and cannot be handled online. So, fasta files were downloaded and formed a “local database” in BioEdit tool. Queries through BLASTP and TBLASTN were performed in BioEdit’s interface for MMETS data and best hits were manually found and transferred from the original MMETS fasta files to new smaller ones. All aminoacid sequences were separately tested for their protein’s domain prediction at Pfam and CDD tool (Conserved Domain Database, NCBI), online.

2.7.2. Alignments of DCR, AGO and RDR proteins

Predicted domains were isolated manually form each sequence peptide and transferred in a new fasta file. Alignmens were performed with MUSCLE from MEGA version 7 (Kumar et al., 2016) in default parameters and MAFFT online (www.ebi.ac.uk/Tools/msa/mafft/) in default parameters. Sequences were manually trimmed at the N’ and C’ ends, while amino acid positions which were absent in a majority of the analyzed species were removed from the alignment.

2.7.3. Phylogenetic Trees of DCR, AGO and RDR proteins based on conserved domains

Condensed domain alignments were used for Bayesian phylogenetic analyses using MrBayes-3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). All analyses shared common settings: mixed amino acid model of evolution with invariant gamma rates, unconstrained branch lengths, sample frequency of 100, 2 runs with 4 chains each, temperature of 0.2, and diagnostic frequency of 1000. The concatenated RNase IIIa/b analysis ran for 32 million generations with a final burn-in of 25%, the PIWI analysis ran for 32 million generations with a final burn-in of 25%, and the RDRP analysis ran for 32 million generations with a final burn-in of 25%.

2.7.4. Phylogenetic analysis of diatom 18S rRNA gene and their RNAi-gene sets

Extra diatom Dicer-like, AGO-like and RDR-containing peptides were grouped and analyzed separately from non-diatom organisms, in order to investigate the different set of RNAi-genes, their distribution and diversion among diatom taxa. Sequences were aligned with MUSCLE and phylogenetic trees were created with Neighbour Joining (NJ) and Maximum Likelihood (ML) analysis in MEGA. Results of the above analysis were summarized in a table indicating the presence/absence and type of each RNAi component.

Transcriptome-derived nucleotide sequences of the 18S gene of every diatom species in this study were downloaded from MMETSP, except from diatoms with their whole genome sequenced and their 18S sequence available in NCBI, JGI or Ensemble.

Manually isolated sequences containing V4-V7 variable regions (775bp-1435bp) of diatom 18S gene (1.7Kb) provide enough diversity to taxonomically separate diatom species and provide complementary information regarding the evolution of their RNAi-gene sets. Haptophyte species *Emiliania huxleyi* and *Gephyrocapsa oceanica* were used as an out group. Sequences were aligned with MAFFT and a phylogenetic tree was created with Maximum Likelihood analysis in MEGA with 1000 x Bootstrap replications, Tamura-Nei model, uniform rates and complete deletion of missing data.

2.8. Heterologous expression systems: *Sacharomyces cereviciae* and *Nicotiana benthamiana*

2.8.1. Reconstitution of a functional RNA silencing pathway in *S.cerevisiae* by introducing *PtDCR* and *PtAGO* genes.

Yeast strains and integrative plasmids in Tables 9 and 10 were kindly provided by Dr. Bartel ((Drinnenberg et al., 2009).

Table 9. Yeast strains *S. cerevisiae* W303-1B.

Strain	Genotype	Phenotype
DPB249	*MAT α leu2-3,112 trp1-1 can1-100 ura3::EGFP(S65T)-KanMX6 ade2-1 his3-11,15*	• GFP expressed
DPB258	*MAT α LEU2::pTEF-Dcr1 TRP1::pTEF-Ago1 can1-100 ura3::EGFP(S65T)-KanMX6 ade2-1 his3-11,15*	• ScAGO+ScDCR silence GFP transgene
DPB271	*MAT α leu2-3,112 trp1-1 can1-100 ade2-1 his3-11,15*	• URA3 expressed
DPB272	*MAT α leu2-3,112 trp1-1 can1-100 ade2-1 HIS3::pGAL1-hpSC_URA3*	• URA3 expressed
DPB276	*MAT α LEU2::pTEF-Dcr1 TRP1::pTEF-Ago1 can1-100 ade2-1 HIS3::pGAL1-hpSC_URA3*	• ScAGO+ScDCR silence URA3 endogenous gene

Table 10. Yeast integration plasmids

Plasmid	Description
pRS404-P*TEF* -Ago1	*S. cerevisiae* integrating plasmid, *S. castellii AGO1* under *TEF* promoter
pRS405-P*TEF* -Dcr1	*S. cerevisiae* integrating plasmid, *S. castellii DCR1* under *TEF* promoter
pRS403-P*GAL1* -weakSC_GFP	*S. cerevisiae* integrating plasmid, weak *GFP* silencing construct under *GAL1* promoter
pRS403-P*GAL1* -strongSC_GFP	*S. cerevisiae* integrating plasmid, strong *GFP* silencing construct under *GAL1* promoter
pRS403-P*GAL1* -hpSC_URA3	*S. cerevisiae* integrating plasmid, hairpin *URA3* silencing construct under *GAL1* promoter

The weak GFP silencing construct represents 275 bp of *GFP* sequence from pFA6a was then cloned in the sense orientation, while the strong GFP silencing construct represents the same 275 bp of *GFP* sequence cloned both in the sense and antisense orientation with a 73-bp sequence spanning intron from *S. pombe rad9.* In the same way, hairpin URA3 silencing construct was made from 339 bp of *URA3* sequence cloned in sense and antisense orientation. For this study *PtDCR* and *PtAGO* were cloned in order to replace *ScDCR* and *ScAGO* in the integrative vectors pRS405-PTEF -Dcr1 and pRS404-

PTEF -Ago1, respectively. Genes of *S.castellii DCR* and *AGO* and *P. tricornutum DCR* and *AGO* were also clone in the double episomal vector p2WGT (carrying both genes) kindly provided by Dr.Kampranis.

2.8.1.1. Silencing the endogenous URA3 gene in S.cerevisiae

For this study the same DPB271 strain (containing URA3) and DPB272 strain (containing URA3 and URA hairpin) were transformed with pRS405-PTEF –PtDCR, pRS404-PTEF –PtAGO and either the Ura hairpin construct (Hp) or no silencing construct (Ø). Strain DPB276 was used as positive control, but was also recreated by transforming the initial DPB272 strain with all pRS404-PTEF –Ago1 and pRS405-PTEF -Dcr1 or by transforming the DPB271 strain with all pRS404-PTEF -Ago1, pRS405-PTEF -Dcr1 and pRS403-PGAL1 -hpSC_URA3, in order to validate our transformation process. Transformation procedures, cultures and media preparation with the appropriate selection are described above.

2.8.1.2 Silencing the GFP transgene in S.cerevisiae

For this study the same DPB249 strain (GFP expressing) was transformed with pRS405-PTEF –PtDCR, pRS404-PTEF –PtAGO and the strong GFP hairpin construct pRS403-PGAL1 –strongSC_GFP (Hp) or no silencing construct (Ø). Negative control was DPB271 strain (URA3, without GFP). Strains were generally grown in glucose media and then transferred in galactose media in order to induce hairpin production and get tested for GFP expression. During experiments, all yeast strains were grown in both liquid glucose and galactose media, in order to compare induced (galactose) and not induced (glucose) hairpin expression.

GFP fluorescence measured at spectrophotometer:

From recently streaked plates, colonies were picked up and inoculated 3ml cultures of glucose for each strain before grown overnight (recipes and conditions described above). From these precultures, new cultures were started in 3ml galactose and glucose media Cells were collected by centrifugation, and resuspended in TE before measuring their OD, that was around 4 or more. All strains we brought to an OD:4 and also prepared three dilutions: 1/10, 1/100, 1/1000 in an ELISA plate. Absorbance (cell density) and GFP fluorescence was measured by Spectrophotometer Systems with parameters (565T): excitation 485/20 and emission 530/20.

GFP fluorescence measured Live-during yeast growth at spectrophotometer:

From recently streaked plates, colonies were picked up and inoculated 3ml cultures of glucose for each strain before grown overnight. From these precultures, new cultures of 100μl galactose and glucose media were started with an OD: 0.01 in Elisa 96 well plate. Two biological repetitions for each sample were included. Cells were incubated at 30°C and orbital shaking for 48 hours. Measurements of absorbance (cell OD) and GFP fluorescence were taken every 30 min. Systems parameters (565T) for GFP documentation were: excitation 485/20 and emission 530/20.

Fluorescence-activated cell sorting (FACS)

From recently streaked plates, colonies were picked up and inoculated 3ml cultures of glucose for each strain before grown at 30°C and orbital shaking overnight. Their OD was measured and then 1 ml of each culture was centrifuged at 3min for 1-2 min and cells were resuspended in the galactose medium. From these glucose or galactose resuspended precultures, new cultures of 4 ml and 20 ml galactose and glucose media (four in total) were started with an OD: 0.1. Cultures 4ml galactose for strains with ScDCR/AGO, PtDCR/AGO and PtDCR+ScAGO were made in duplicates in order to test different periods of galactose induction. Cells were incubated at 30°C and orbital shaking overnight. After 13 hours the aforementioned duplicates of 4ml galactose cultures with OD: 1.5-2 were collected, while all the rest were left to grow for 20 hours reaching OD: 3-5. Yeast cells in TE with OD: ~1 were counted to correspond to ~2x 10^7 cells/ml for the strains in this study. So, cell pellet from 0.5 ml culture (4ml cultures) was washed twice and resuspended in TE before further analysis in FACS, performed by qualified personnel in IMBB.

Northern blot analysis of GFP small RNAs

The 20ml galactose and glucose cultures after 20 hours of incubation with OD: 3-5, were centrifuged in 50ml falcons at 3000 rpm for 3 min and resuspended in 2ml sterile H_2O to wash out remaining media. Then, they were centrifuged at 3000 rpm for 2 min, supernatant was discarded and cell pellets were flash freezed in liquid nitrogen.

RNA extraction was performed by following the hot acidic phenol extraction, as described before.

2.8.2. RNA silencing in *N.benthamiana* DCL2/3/4i knockdown by PtDCR complementation

Nicotiana benthamiana transgenic lines with downregulation of DCL2, DCL3, DCL4, were generated in our lab after crossing DCL2/4.5(x)3.10i lines (Katsarou et al., 2016, 2019). *S. castellii* DCR, as positive control, and *P. tricornutum* DCRb and DCRc spliceoforms were cloned into pENTR3C vector and then into GATEWAY vectors pB2WG7 after LR reaction. *A. tumefaciens* C58C1 strain was transformed with these constructs was then agroinfiltated in *N. bethamiana* WT and DCL2/3/4i healthy or PSTVd infected leaves. GATEWAY cloning, agroinfiltration and PSTVd infection are described above.

2.9. Generation of PtDCR KO mutants via CRISPR/Cas9 technology

Recent development of CRISPR-Cas9 mediated mutagenesis methodology in *P. tricornutum* was described in (Nymark et al., 2016a).

2.9.1. Construction of CRISPR-Cas9 vectors for DCR/AGO/RDR-KO mutants

Single guide RNAs (sgRNA) targeting DCR, AGO and RDRP (sgRNA-DCR/AGO/RDR) were designed using the PhytoCRISP-Ex application tailored for analysis in algae including diatoms (Rastogi et al., 2016). Results were double checked with CRISPOR (http://crispor.tefor.net/) and CHOPCHOP (chopchop.cbu.uib.no) applications and the best common candidates were used (Table 11). After identifying the best candidates, an amplicon of the targeted area from genomic DNA was sent for Sanger sequencing, in order to verify the absence of polymorphisms.

Table 11. Table of sgRNA used in this study with parameters given from PhytoCRISPR-EX.

Targeted gene	Orientation	Position gDNA	Guide seq +PAM	Specificity Score	Predicted Efficiency	ut of frame score	Off-targets for 0-1-2-3-4 mismaches
Dicer	rev	345-367	GAGTGAACTCGACCCGAAAT TGG	100	30/41	9	0
Dicer	fw	887-909	GTTTAGGACGGGACGCCGTG CGG	100	68/65	79	0
AGO	fw	1327-1340	TCCGTGGAACTAGATTCTTG TGG	100	71/19	70	0
AGO	fw	1662-1684	TGTAGACTTGTTCAAACTCG TGG	100	51/53	60	0
AGO	rev	1992-1972	ACGGTTGCTTGCATTTGCAG CGG	100	62/55	54	0
RDR	Fw	2590	TCGAACGAGCCGATCCTATTGTGG	100	67/62	54	0
RDR	Fw	2262	CCATGTGGGATCGCACCCGT TGG	100	67/18	73	0

Each sgRNA was cloned in the pKSdiaCas9 vector, kindly provided by Dr. A.Falciatore, in which expression of the Cas9 gene and sgRNA are placed under the control of the pFcpB and ubiquitin 6 promoters, respectively. sgRNAs were initially designed and prepared as 5’phosphorylated primers with four extra specific nucleotides at 5’end, that were used as complementary overhangs in cloning. Complementary ssDNA “primers” were denaturized and then slowly annealed together in ds DNA with overhangs. PKS diaCas9 vector was digested with BsaI restriction enzyme, leaving complementary overhangs for ligation with the sgRNA constructs (protocol described in Nymark et al. 2016). Positive colonies were screened by colony PCR using the Fw sgRNA primers and M13Rv primer: *CAGGAAACAGCTATGAC*, producing a 548 bp amplicon.

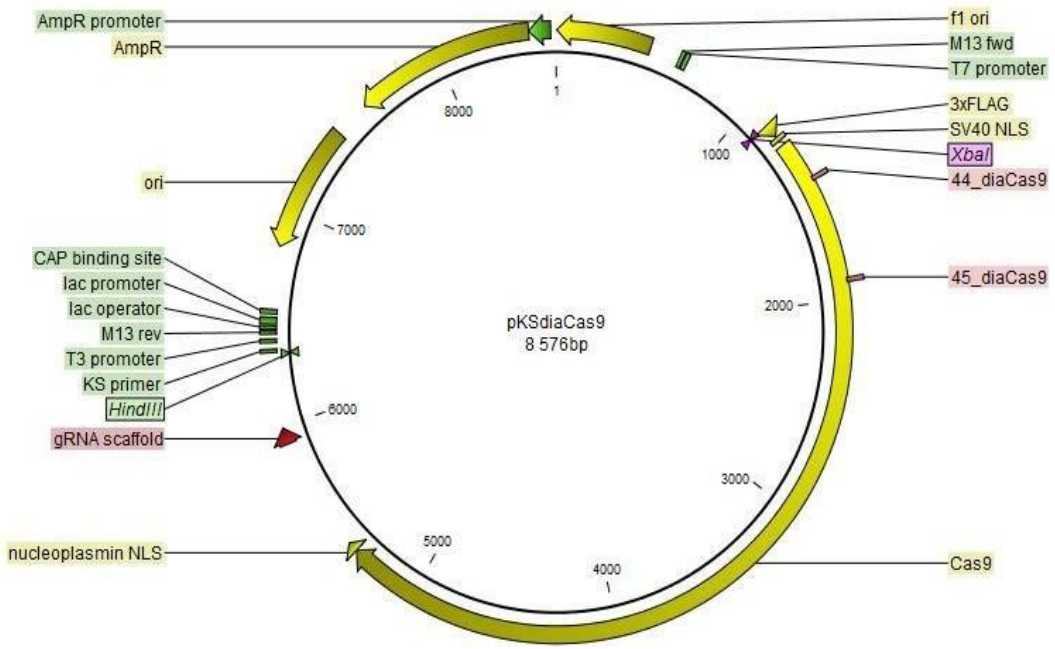

Figure 12. Cloning of sgRNA construct in pKSdiaCas9 with BsaI enzyme.

2.9.2. Transformation and molecular characterization of KO lines

pKSdiaCas9_sgRNA-DCR/AGO/RDR vectors and a vector carrying the selection marker were co-delivered by biolistic transformation in *P. tricornutum* cells. The vectors with selection markers were a) pNptII, containing the nptII gene under the control of pFcpB to transform the WT line, and b) pZeo, containing the zeocin gene under the control of pFcpB to transform the GUS and GUS/RNAi line. Transformants were selected on solid media with the appropriate antibiotics (transformation protocol described above) and individual colonies were recovered after 2-3 weeks.

Single transformant colonies were transferred on less stringent selective plates, in order to grow faster (1 week) and provide enough material for analysis. Lysates were prepared in 96 well plates, as follows:

P. tricornutum transformant cells were picked up with a sterile tip and transferred in 20μl of fresh lysate buffer solution (Table 12) in a 96 well plate. Incubation at 4°C for 15 min, then at 95 for 10 min and finally cool down to room temperature (a PCR machine was used). Lysates were diluted with addition of 80 μl and stored at -80°C before use.

Table 12. Lysate Buffer composition

Reagents	Final volume (1ml)
Triton x100, 10%dilluted	100μl
TrisHCL 1M, pH=8	20μl
EDTA 0.5M	4μl
H2O	up to1ml

Lysates were used as templates for colony PCR, in order to amplify a 496bp piece of Cas9 gene with primers Fw_diaCas9: CGAACCGCGAGACGAAGATA and Rv_diaCas9: CGGGAAGTTGCGCAATCAAA.

The lysates from Cas9 positive clones in previous analysis were used as templates for colony PCR amplification of the DCR/AGO/RDR targeted locus. Forward primers upstream the 1st sgRNA target and reverse primers downstream the 2nd sgRNA target, were used in candidate KOs and WT lysates as well.

Table 13. Primers used for screening Pt DCR/AGO KO mutants.

Primer Name (Tm)	Primer Sequence
PtDCR sgRNA-all INS Fw (Tm:55,7)	ACCAGGAGCTTGGTATTTCCAG
PtDCR sgRNA-all INS Rev (Tm:57,6)	ACCATAATCCAGTCAATCGCTGG
PtDCR sgRNA-all OUT Fw (Tm:59,6)	AGACTGGCATGAATTCGAGGAACTC
PtDCR sgRNA-all OUT Rev (Tm:57,6)	AGCGATTGCGCAAGATGTCG
DCL target ins Fw (Tm:57,8)	ACGGTCAATTTGATGGTCGCAG
DCL target ins Rv (Tm:57,4)	GAGCTTTGTATAATTGCTTTTCTCG
PtAGO sgRNA-all INS Fw (Tm:57,5)	ATGGATTTGCTCCCGCACG
PtAGO sgRNA-all INS Rev (Tm:56,6)	ATGTCTTTCTCCGTTGTTGATTCC
PtAGO sgRNA-all OUT Fw (Tm:58)	AATTACAGGGCTGAACGTAAGTGC
PtAGO sgRNA-all OUT Rev (Tm:57,7)	AATCTCATCCTTTGACATATCCAACAC

HRMC, High Resolution Melting Curve analysis

P. tricornutum mutant candidates demonstrating a WT-like size amplicon in agarose gel electrophoresis were tested with high resolution melting curve (HRMC) analysis for small single point INDELs. HRMC analysis detects the release of a fluorescent dye bound to dsDNA when it reaches the melting point, which is sequence

specific and very sensitive. Primers were designed to amplify smaller amplicons (100-200bp) surrounding the targeted locus excluding any single nucleotide polymorphisms (SNPs), that are frequent in *P. tricornutum* genes. Templates from WT clones were always included as controls and three technical replicates per sample were performed.

2.9.3. Phenotypic analysis of KO lines

2.9.3.1. Growth under normal conditions or under nutrient starvation

Growth of *P. tricornutum* lines under normal conditions or nutrient starvation was observed in order to highlight any differences in growth rate and morphology between WT and mutants. Cultures were made in triplicates, starting with 10 x10^4 cells/ml and grown at 19°C and 12:12 light:dark (80 µmol m-2 s-1). Cell density was measured daily under a light microscope, enabling the observation of cell size, morphotypes and general health of each culture.

Liquid media used were a) F/2 without NaNO3 and b) F/2 complete.

2.9.3.2. Growth after UV induced damage

P. tricornutum lines were tested for their ability to recover after exposure to UV radiation on agar plates, as described in De Riso et al. 2009. Cells were collected from cultures in exponential phase ~2x10^6 cells/ml and were concentrated to 30 x 10^6 cells/ml. Dilutions were prepared in order to make spots of 10µl on F/2 agar plate containing 5x10^5, 3x10^5, 2x10^5, 1x10^5, 0.5x10^5 cells of each line. An identical pair of plates was prepared in order for one plate to be exposed to UV radiation (1200 mJ/cm², radiation 254 nm, UV crosslinker, Stratagene) and the other one to be kept as control. Then, plates were sealed with parafilm and placed in chamber at 19°C and 12:12 light:dark (80 µmol m-2 s-1) for 2 weeks. Photographs were taken every five days and three independent repeats were performed.

2.9.3.3. Sequencing (NGS) of large and small RNA of P. tricornutum DCR KO mutants

2.7.3.3.1. RNA extraction

Axenic cultures 300 ml of *P. tricornutum* in exponential phase 1,8-2,4x10^6 cells/ml grown under optimal conditions in glass flasks were harvested with centrifugation, flash

frozen in liquid nitrogen and grinded in fine powder. RNA was extracted with 'Total RNA Purification" kit by applying DNaseI on column (Norgen). RNAs were quantified in NanoDrop, run in a 1,2 % MOPS/formaldehyde gel to validate the absence of RNA degradation and finally run in a Bioanalyser (Agilent) gel to define more accurately their quality and RIN numbers.

2.9.3.3.2. cDNA libraries preparation

Small RNA cDNAlibraries were prepared from 1μg total RNA with "NEBNext Multiplex Small RNA Library Prep Set for Illumina" (New England Biolabs). All ligation steps, reverse transcription and PCR amplification were executed by following manufacturer's basic instructions. Adaptors were not diluted; 14 cycles of PCR were performed (proved to be optimal for *P. tricornutum* after standardizing the protocol) and cDNAs were eluted in 30μl H2O. cDNAs were size fractionated in a 6%PAGE gel, where pieces of 140-160bp (containing 20-30bp sRNAs with 120-127bp adaptors) were clearly separated from adaptor/primer dimmers. Small RNA cDNA libraries were extracted from polyacrylamide gel, purified with "NucleoSpin®Gel and PCR Clean-up" (Macherey-Nagel) and eluted in 15μl H2O. cDNA template before and after size selection was kept for analysis in Bianalyser Agilent. Expected results after size selection would be a single peak of ~155bp without any contaminations of different sized fragments.

3prime quant RNA-seq libraries were prepared by IMBB sequencing facility personnel, according to manufacturer's protocol (Lexogen 3prime quant forward for Illumina library kit). This protocol needs 500ngr total RNA input and generates highly strand-specific next-generation sequencing (NGS) libraries close to the 3' end of polyadenylated RNAs. Only one fragment per transcript is generated, directly linking the number of reads mapping to a gene to its expression. Library generation is initiated by oligo-dT priming and then first-strand synthesis and RNA removal is followed by random-primed synthesis of the complementary strand (second-strand synthesis). Illumina-specific linker sequences are introduced by the primers and the resulting double-stranded cDNA is purified with magnetic beads. The libraries were generated with 14 cycles of PCR, as 14 were the optimal cycles for *P. tricornutum* (after standardizing the protocol specifically to diatoms RNA) and validated in Bioanalyser.

2.9.3.3.3. *Illumina Sequencing*

All samples were sequenced with Illumina NextSeq500 Sequencer technology, in IMBB, FORTH. 3prime quant RNA-sequencing was performed with single 1 x 100 bp run and aimed for 10 million reads. Small RNA-sequencing was performed with single 1 x 55 bp run and aimed for 5 million reads.

2.9.3.3.4. Bioinformatic Analysis

Bioinformatic analysis of both 3prime quant-RNA sequencing and small RNA sequencing was performed by Dr. Hugues Richard (as. Professor, Computational and Quantitative Biology Laboratory, University Pierre & Marie Curie, Paris).

The reads were aligned to the *P. tricornutum* genome version 2 (Phatr2) with BWA version 0.7.17-r1188 (Li and Durbin, 2009). The command line used to align the small RNA (sRNA) libraries was bwa mem -k 12 -t 7 -A 1 -L 2 -T 13 <Phatr2 genome> <reads_file>. The RNA-Seq sequences were aligned with default parameters. For the sRNA libraries, the reads were subsequently filtered by keeping reads with a PHRED mapping quality of at least 5.

Read coverage and fragment length was computed using the set of quality filtered reads. For each aligned read, the length of the fragment is the length of the alignment from end to end (e.g. taking out soft clipped ones on both ends).

For the RNA-Seq sequencing experiment, due to the protocol, the sequences were enriched in the 3’ UTR regions of the genes. However, those regions are rarely annotated in the gene annotation, so we could not simply count the number of reads aligning within a gene. Thus, a simple strategy to estimate gene expression was devised. Strand specific contiguous regions that are covered by at least 10 reads in one of the conditions were extracted. Then, these regions were merged across experiments and the number of reads overlapping them, were counted. Each region was associated to its nearest 5’ gene. Differential expression analysis was carried out with DESeq2 (Love et al., 2014).

As genome reference, the Phatr v2.0 from JGI was used with the corresponding gene annotation (Bowler et al., 2008), and non-coding RNA annotation from RFam v11.0 (Gardner et al., 2011). Additionally, tRNA genes were all reannotated using tRNAscan(Lowe and Eddy, 1997). All data analyses were done using a combination of python, R and shell scripts, and the bedtools suite (Quinlan and Hall, 2010).

2.9.4 GUS histochemical assay

Protocol was followed as described in Falciatore et al., 1999. Briefly, diatom cultures of 10^6 cells/mL were concentrated to $5x10^6$ cells/mL, spotted (5μl) on F/2 agar plate without selection and let to grow in optimal conditions for 3-5 days. Then, pieces of agar containing one culture spot were extracted from petri dish by using a cryotube (for uniform cuts between spots) and placed in wells of a 24 well plate. 100 μl X-Gluc solution (100 mM Na-phosphate, pH 7.0, 10mM Na_2 ethylenediaminetetracetic acid (EDTA), 5 mM K 4Fe(CN)6, 5 mM K3Fe(CN)6, 0.1% Triton X-100, 0.3% X-Gluc (Clontech)) were added at the bottom of each agar piece, in order to slowly absorb it, and the plate was incubated at 37°C in dark, overnight. Blue spots were obvious and blue cells were also detected by microscopy.

3. Results

3.1. Cloning of *P. tricornutum* RNAi-key genes

P. tricornutum RNAi-key genes were identified and cloned, in order to validate their sequence and a) investigate their function by expressing them in heterologous expression systems and by generating *P. tricornutum* Knock-Out (KO) mutants via CRISPR/Cas0 genome editing, and b) investigate their subcellular localization by expressing them fused to a fluorescent protein in *P. tricornutum*.

A single *P. tricornutum DCR* gene (*PtDCR*), an *AGO* gene (*PtAGO)* and a single RDR domain-containing gene (*PtRDR*) were initially identified *in silico* by De Riso et al. (2009) and the absence of any extra gene candidate was validated after thorough search in *P. tricornutum* genome. This search utilized specific keywords (e.g.DCR, RNseIII, PIWI) and known conserved sequences from related and unrelated species using the BLASTP and TBLASTN algorithms initially in the Genome Portal of JGI (genome.jgi-psf.org) that provides the Phat2 annotation and subsequently in the Ensembl (http://protists.ensembl.org) database that provides the Phat3-annotation.

Primers were designed in order to amplify the coding region (first ATG to stop codon) of *PtDCR, PtAGO* and *PtRDR* genes. For the amplification process the following three templates were used: genomic DNA and cDNA from *P. tricornutum* culture strain CCAP 1055/1 grown in the lab and a cDNA library synthesized from mRNA of the same strain by Frederic Verret at the MBA- Plymouth.

3.1.1. *PtDCR* gene

The *PtDCR* gene was amplified from genomic DNA, cloned in TOPOII vector and validated by Sanger sequencing, which confirmed the sequence proposed by JGI's Genome Portal except from some polymorphisms that produce only silent mutations. *PtDCR* gene amplified from cDNA library was also cloned in TOPOII and sequenced, revealing a smaller length of the first intron (95bp) than the one predicted in Phat2 (212bp); this discrepancy was corrected in the Phat3_annotation in Ensembl. Moreover, Sanger sequencing of multiple cDNA clones revealed the presence of two distinct *PtDCR* spliceoforms: *PtDCRb* (2541 bp) and *PtDCRc* (2556 bp) (Figure 13). The proposed spliceoform (*PtDCRa*) could not be amplified. The *PtDCRb* spliceoform was produced

when alternative splicing resulted in retention of the second intron towards the 3' end of the gene and introduction of a stop codon. Spliceoform *PtDCRc* was produced when alternative splicing resulted in removal of the second intron along with some nucleotides of the third exon that changed the frame and introduced a stop codon. None of the sequenced cDNA clones corresponded to the predicted-spliced full length gene model and *PtDCRa* could not be retrieved (Figure 13). A recent publication shed some light on the extensive alternative splicing events in this organism (Rastogi et al., 2018).

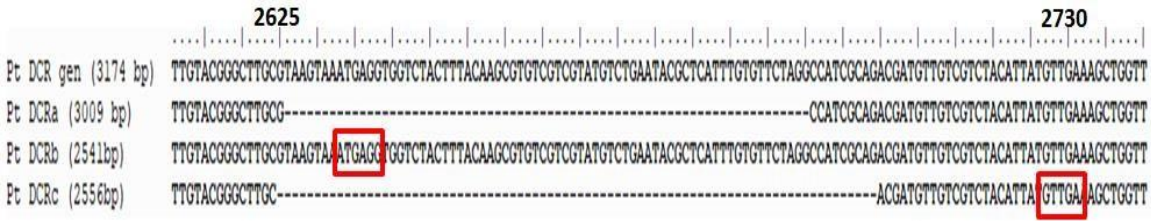

Figure 13. Alignment of *PtDCR* genomic DNA sequence *and PtDCR* spliceoforms a, b, c at the alternatively spliced intron area. Introduced stop codons are indicated in red boxes.

PtDCRa-c functional domains were screened in *silico* with Pfam (EMBL-EBI) (pfam.xfam.org) and CDD (NCBI) (www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), and only a single RNaseIII domain was predicted in all three spliceoforms. A second RNaseIII domain could be predicted in *DCRa* when the E-value threshold was increased to 1. This second less conserved domain, however, was interrupted in PtDCRb and PtDCRc predicted proteins. The high E-value associated with the prediction of the second RNase III domain in *DCRa*, however, questioned its presence. In addition, independent RT-PCR and sequencing of both full-length cDNA clones and cDNA fragments including the intron-exon junction encoding the predicted second RNase III domain could not experimentally confirm its presence. Moreover functional DCRs presenting a single RNase III domain have been characterized in the single cell eukaryotes *S.castellii* (Drinnenberg et al., 2009) and *Entamoeba hystolytica* (Pompey et al., 2015b). It was concluded that *P. tricornutum* presents a single *DCR* gene, transcribed into two spliceoforms, both encoding a single RNase III domain. At that stage of the PhD, experiments on the *DCRb* and c functional characterization in yeast and plants were carried out.

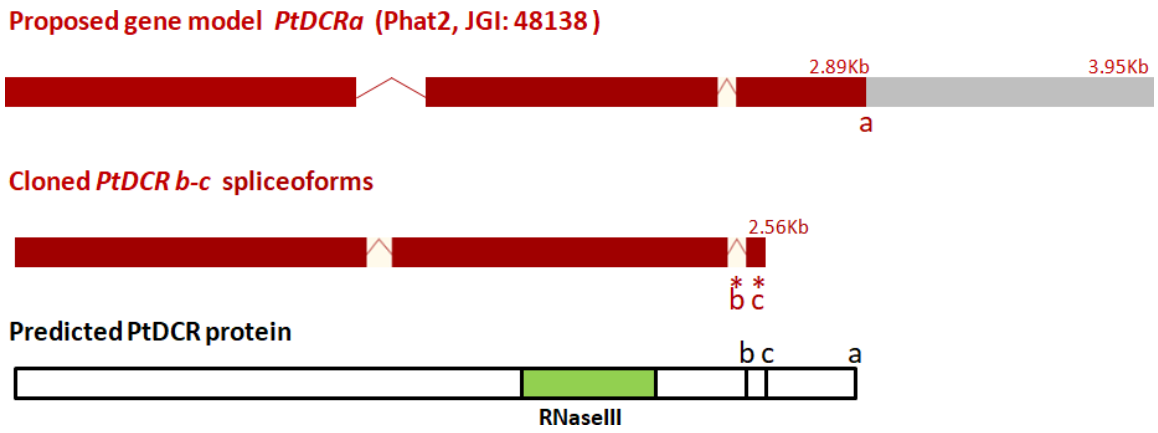

Figure 14. *PtDCRa-c* gene models. Asterisks indicate stop codons introduced by alternative splicing of the second intron. Predicted 3'UTR is indicated in grey color.

During an EMBO short fellowship visit in Dr A. Falciatore's lab in UPMC, Paris, where the effect of diatom circadian clock on gene expression was studied and analogous alternative splicing patterns were previously observed, we decided to test the PCR amplification of *PtDCR* spliceoforms during the circadian clock.

The template used was a mix of seven cDNAs that were derived from RNAs collected every ~4 hours from the same *P. tricornutum* culture during 24 hours (representative of the circadian clock). In this case, the proposed-spliced gene (*PtDCRa*) was amplified, validated through Sanger sequencing and cloned into a pUC19 vector (Figure 15). In addition, *PtDCRb* was amplified and found equally abundant to *PtDCRa*, while *PtDCRc* was not found (Figure 15). However, when traces of DNA between *PtDCRa* and *PtDCRb* bands were extracted and analyzed by Sanger sequencing, the chromatogram showed a mix of *PtDCRa* and *PtDCRb* sequences. The chromatogram sequence was uniform until the slicing point and then a mixture of the two sequences followed, whose biggest peaks could be read as the PtDCRc sequence.

Although it seems as a common PCR artifact, the sequence of *PtDCRc* from the cDNA library was clear (not a mixture) and found in six clones. This finding suggests either a PCR artifact introduced in cDNA library, or indeed spliceoform *PtDCRc* was produced from *P. tricornutum* strain in Plymouth.

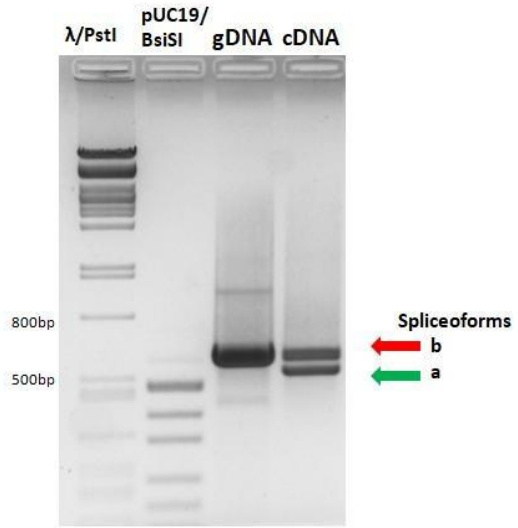

Figure 15. *PtDCRa* and *PtDCRb* spliceoforms amplified from the circadian cDNA mix in equal level.

MMETSP available transcriptomes and predicted proteomes from diatoms (described below) were queried for *DCR* genes that validated the presence a DCR-type protein with a conserved RNAseIIIa domain and a less conserved RNaseIIIb domain. Interestingly, a diverged PAZ domain not predicted using common tools, except when a much higher E-value is applied, was identified that appeared conserved among these DCR proteins. Based on these findings, the *PtDCR* gene model is illustrated in Figure 16. Interestingly, Pt DCR protein architecture resembles the protist-like DCRs that are characterized by a shorter N' terminal domain and the absence of some domains usually present in multicellular eukaryotic DCRs.

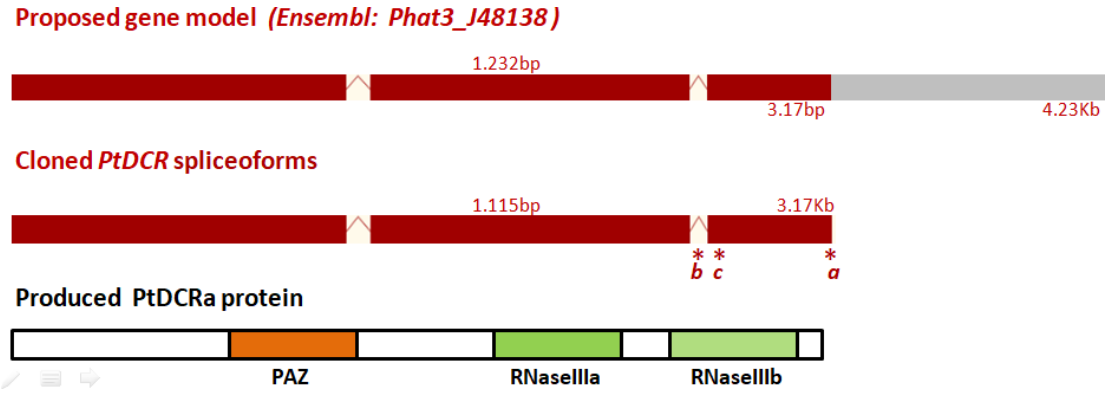

Figure 16. The identified *PtDCR* gene and protein models. Predicted 3'UTR is indicated in grey color.

In the present study, *PtDCRb* and *PtDCRc* spliceoforms were expressed in yeast and plant heterologous expression systems, while all other experiments were performed with *PtDCRa*.

3.1.1. *PtAGO* gene

The PtAGO gene (ATG to stop codon) was amplified from genomic DNA, cloned in TOPOII vector and validated by Sanger sequencing that confirmed the proposed sequence with some polymorphisms producing only silent mutations. However, the amplification of the proposed *PtAGO (PtAGOa)* from cDNA library was not successful, even after testing with either different primers upstream/downstream of the proposed ATG, or under various annealing temperatures. Downstream, in frame ATGs were identified and forward primers were designed to hybridize at these positions. Finally, *PtAGO* gene was efficiently amplified from cDNA library with the forward primer hybridizing at the ATG in position 150bp (which is the fourth ATG in frame) of the predicted coding gene (position 223bp in genomic sequence, after the first intron). At the time, 13 introns were identified in contrast to 16 introns predicted in Genome Portal, a finding that was later confirmed in the Phat3 Ensembl database. Since multiple tests and repeats produced the same results and no UTR regions were annotated, the truncated *PtAGOb* was studied assuming the *PtAGO* cDNA transcription initiated downstream to its prediction.

As described before for the *PtDCR* spliceoforms, the amplification of the full proposed *PtAGOa* gene was achieved when the circadian cDNA mix was used as template (Figure 17). Morover, *PtAGOa* was cloned in PUC19 vectors and new *PtAGO* sliceoforms were identified after validating multiple clones with Sanger sequencing, all of which demonstrated truncated forms of PtAGO protein caused by alternative splicing. These findings underline once again the importance of considering the circadian clock effect on gene expression. It would be interesting to analyze each time point sample individually in order to investigate the presence/absence of each spliceoform.

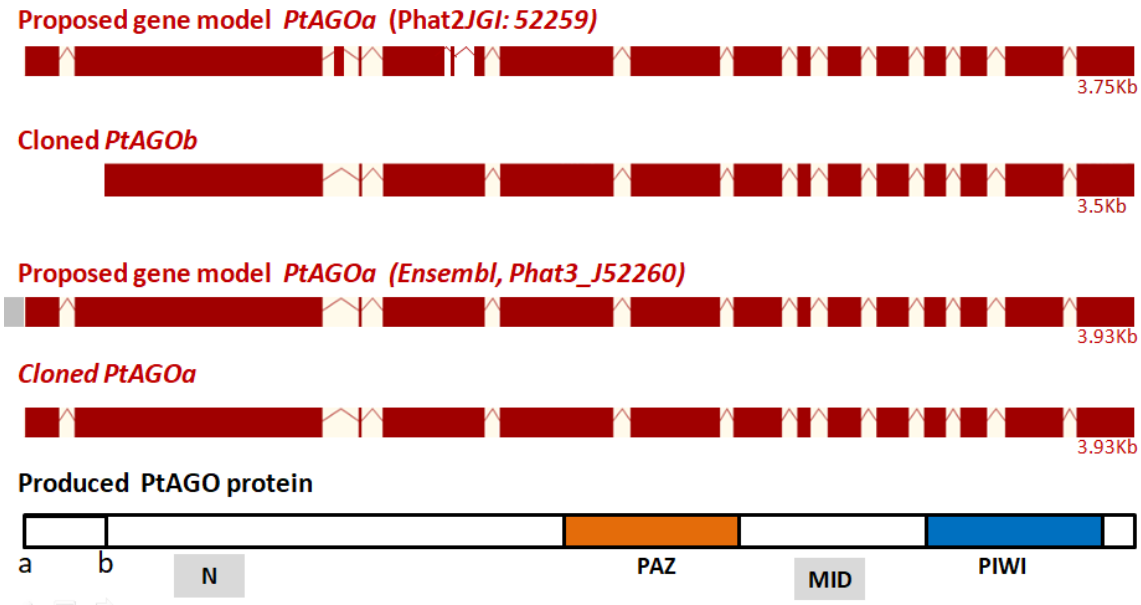

Figure 17. The identified *PtAGO* gene and protein model. Predicted 5'UTR is indicated in grey color.

In the present study, *PtAGOb* was expressed in yeast heterologous expression system, while all other experiments were performed with *PtAGOa*.

3.1.1. *PtRDR* gene

PtRDR gene was amplified from genomic DNA and circadian mixed cDNA and cloned in TOPOII vectors. Validation by Sanger sequencing confirmed the proposed sequence (Ensemble, Phat3) and the absence of introns.

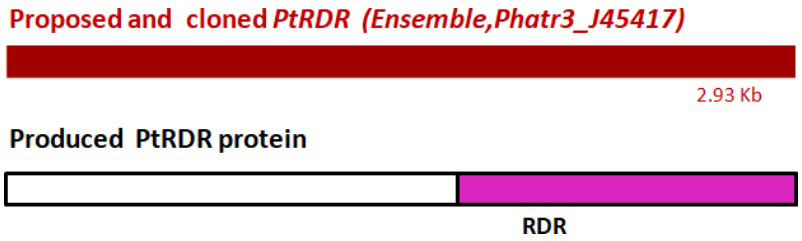

Figure 18. The identified *PtRDR* gene and protein model.

In summary, cloning of *PtDCRa-c*, *PtAGOa-b* and *PtRDR* genes enabled us to proceed to their functional characterization and identification of their subcellular localization, but also provided some original data in the following aspects:

- Based on extensive search in updated genomic and transcriptomic information, that became only recently available, the findings of De Riso et al (2009) on the probable presence of a single *DCR/AGO/RDR* gene in *P. tricornutum* were confirmed.
- Since *Pt DCR/AGO/RDR* genes were cloned from cDNA they are probably producing active proteins.
- Identification of alternative splicing in *PtDCR* and *PtAGO* genes, that was later confirmed by findings on the extensive alternative splicing events taking place in *P. tricornutum* (Rastogi et al., 2018).
- Alternative splicing of *PtDCR* and *PtAGO* seem to be influenced by circadian clock.
- Pt DCR presents a protist-like DCR architecture with shorter N' terminal domain and the less functional domains compared to multicellular eukaryotic DCRs.

3.2. Phylogenetic Analysis of *P. tricornutum* RNAi-key genes

In order to support a representative phylogenetic analysis of these genes, other organisms, from close diatom relatives to some well-studied distant species (like plants and animals), were searched for the basic three RNAi-key genes. Apart from the aforementioned databases that contain a large number of usually well studied organisms, the Marine Microbial Eukaryote Transcriptome Sequencing Project, MMETSP (Keeling et al., 2014), provided a vast amount of transcriptomic information and predicted proteome of unicellular marine eukaryotes and among them, diatoms. This created an opportunity to make a deeper search on RNAi proteins among diatom taxa, accompanied with their 18S phylogenetic analysis for clarification of their evolutionary distances. Data from MMETS were downloaded and formed a "local database", where queries through BLASTP and TBLASTN were performed. The best hits were validated by prediction of their protein's domains. Predicted amino acid sequences from MMETS were usually incomplete but still included in phylogenetic analysis, as long as the aligned domain was almost intact. Sequence with partial domains were removed from further phylogenetic analysis, but still taken into account for a presence/absence analysis of these proteins among diatom taxa. For broad species analyses, a few representative diatom proteins were selected, in order to avoid bias and keep a basic amount of data in the process. Phylogenetic analyses of DCR, AGO and RDR contained sequences that derived from the same set of species, with exceptions where species lacking specific genes were replaced by close relatives. In order to reduce errors introduced by the presence of partial sequences and to moderate the computational load, amino acid positions which were absent in a majority of the analyzed species were removed from the alignment (Baurain et al., 2010).

Nomenclature of genes followed (Zong et al., 2009) the proposed format of "Gens" (the first three letters of the genus name followed by the first letter of the species name).

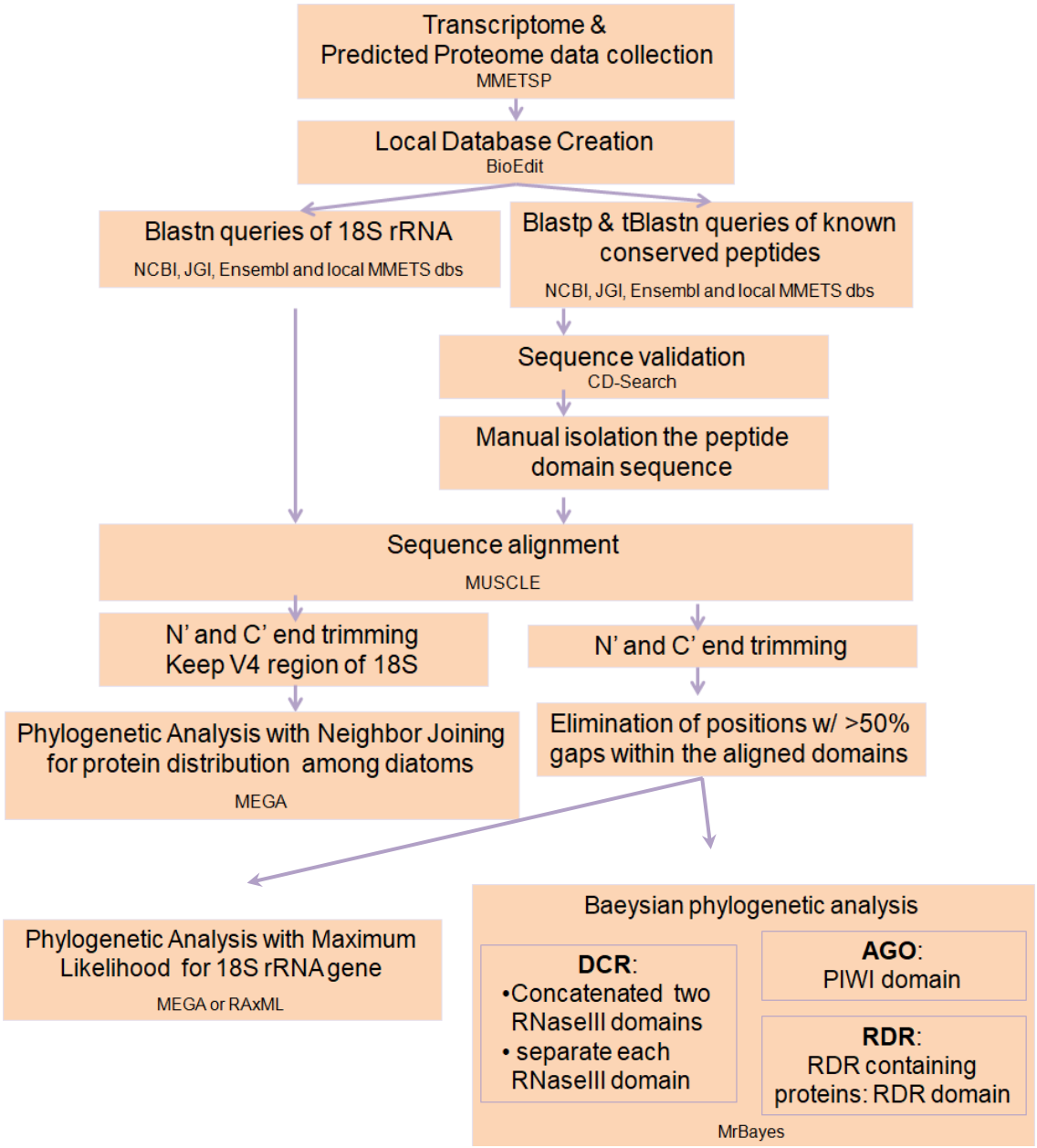

Figure 19. Pipeline of DCR/AGO/RDR Phylogenetic Analysis.

3.2.1. Phylogenetic analysis of *P. tricornutum* DCR protein

DCR homologs minimally had to contain two RNase III domains and DROSHA homologues were also included in this analysis. Cases with one RNaseIII were also kept for further analysis, including bacteria and the functional *S. castellii* DCR1. Research in different organisms resulted in different naming conventions (DCR and DCL). The annotated provided name was retained, while for the new identified proteins we used the DCR designation. Validation of diatom DCR candidates by domain prediction showed relatively low conservation of their domains. In some cases, the E-value had to be very loose in order to predict specific domains, especially some RNaseIIIb, or PAZ domains.

For analysis including organisms other than diatoms, the conserved concatenated RNase IIIa and RNase IIIb domains were aligned following the deletion of their linking sequences. DROSHA homologues were included in this analysis, as they have the same double RNaseIII domain structure, while proteins containing only one RNaseIII domain were excluded. Condensed domain alignments were used for Bayesian phylogenetic analysis and the produced tree is illustrated in Figure 20.

The phylogenetic analysis of DCR proteins clearly leads to three conclusions:

1) DCR proteins in general have greatly diverged, since only closely related organisms or previously duplicated DCRs are clustered together with confidence. Thus, the origin of diatom DCRs from phosynthetic algae or animal-like ancestor cell cannot be clarified. However, the extent of diatom DCR divergence from other stramenopiles was not expected.
2) Diatom DCRs form two distinct groups with different types of DCR protein. In Group A, diatoms have a typical “domain architecture” of DCR found in multicellular eukaryotes, that contains two conserved RNaseIII domains, a Helicase-C domain, a dicer dimerization domain and a PAZ. In Group B, diatoms have shorter, protist-type Dicer that contains only two RNaseIII domains and a PAZ.
3) There are diatom species that have one or the other type of DCR, DCRa or DCRb, and then there are diatoms that possess both types of DCRs.

Diatoms with their whole genome sequenced were included in this analysis, to ensure the liability of these genes being present/absent in the genome. Out of these, *Fragilariopsis cylindrus*, *Pseudonitzschia multiseries* and *Thalassiosira oceanica* have DCRa, *Phaeodactylum tricornutum and Thalassiosira pseudonana* have DCRb and *Cyclotela cryptica* and *Fistulifera solaris* have both DCRa and DCRb.

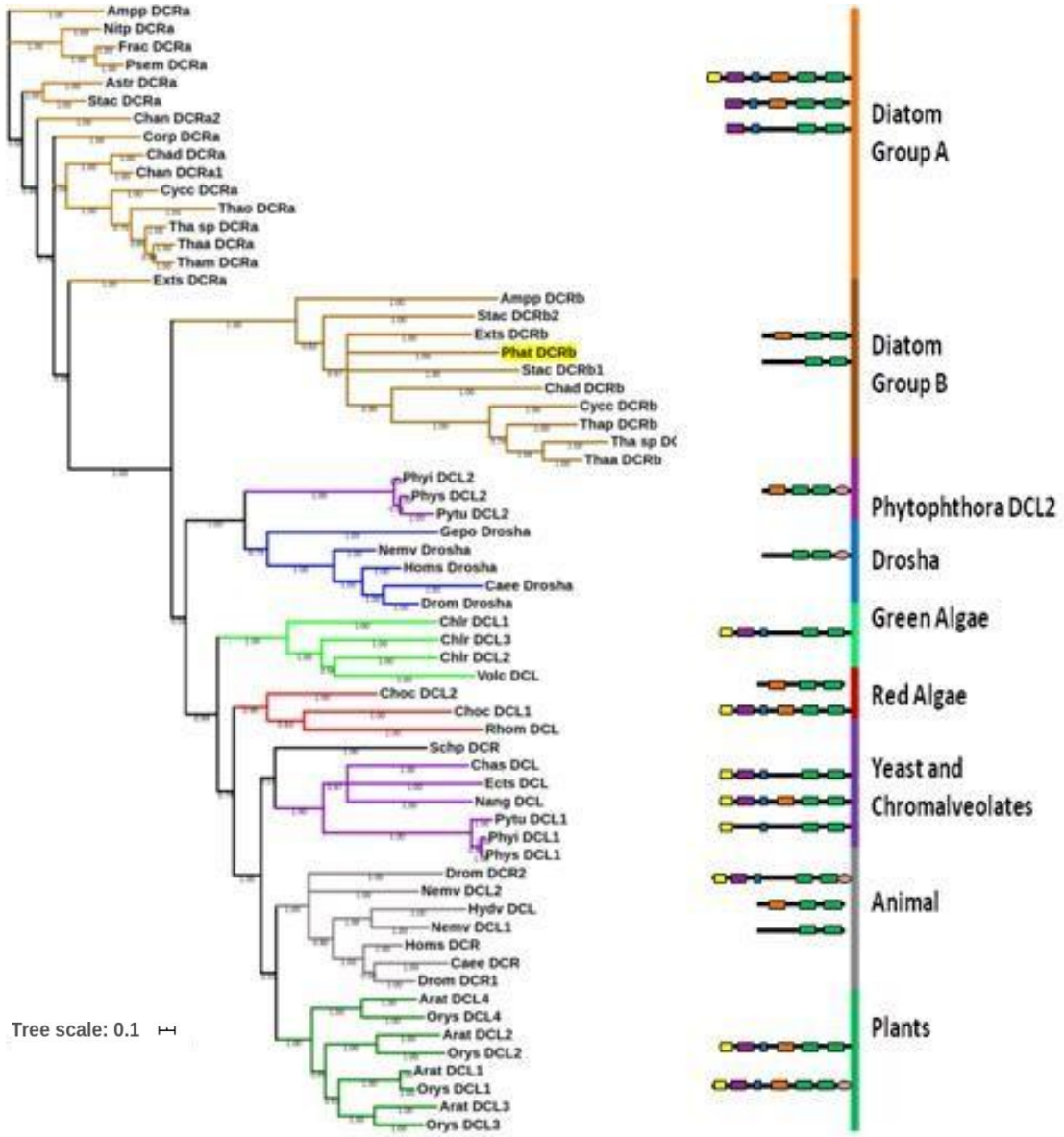

Figure 20. Bayesian phylogenetic analysis of DCR proteins. Colored taxonomical groups are indicated on the right along with their representative domain architecture.

Based on the Bayesian analysis of the RNaseIII domains, no conclusions could be drawn regarding the progeny of each diatom group or about the evolutionary relationship between DCRs in diatoms and other organisms. However, Group A DCRs contained all typical DCR domains, which were more conserved and thus easily predicted with currently available tools. On the other hand, Group B only had the RNaseIIIa domain clearly conserved and predicted with traditional tools. In this case RNaseIIIb and PAZ domains were predicted only when E-value limit was set to 1 and they were still predicted as partial. Based on these observations, Group A DCRs would be expected to be more closely related to canonical plant and animal DCRs than those of Group B. The lower conservation of Group B DCRs laid on their more diverged RNaseIIIb and the level of differentiation was clearly shown in alignment of only separate RNaseIIIb domains (Figure 21).

Figure 21. Alignment of RNaseIIIb domains from various DCRs, DROSHAs and bacterial RNaseIIIs (70% similarity highlighted). RNaseIIIb domains from Diatom Group B (including *P. tricornutum*, red arrow) and are less conserved.

3.2.2. Phylogenetic analysis of *P. tricornutum* AGO protein

All selected AGO homologs featured a PIWI domain and preferably also a PAZ domain. For broad species analysis, only the PIWI domains were aligned and a Bayesian analysis and phylogenetic tree were produced (Figure 22).

The phylogenetic analysis of AGO proteins lead to three conclusions:

1) Diatom AGOs formed two distinct groups. In Group B, diatoms had a typical AGO protein that was grouped with all other eukaryotic AGOs (with posterior probability 0.92). In Group A, diatoms had an AGO that contained the necessary PAZ and PIWI domains but had diverged from other known AGOs and did not relate to any other group. A posterior probability of 0.82 in this phylogenetic tree was not high enough to support a meaningful relation with the PIWI clade proteins.
2) Diatom species that only had DCRa had also AGO type A (AGOa), those who had only DCRb had only AGO type B (AGOb) and those who had both types of DCR had also both types of AGOs. Thus, a “set” of DCR/AGO RNAi key-genes presenting the same phylogeny seems to be variably distributed among diatom taxa.

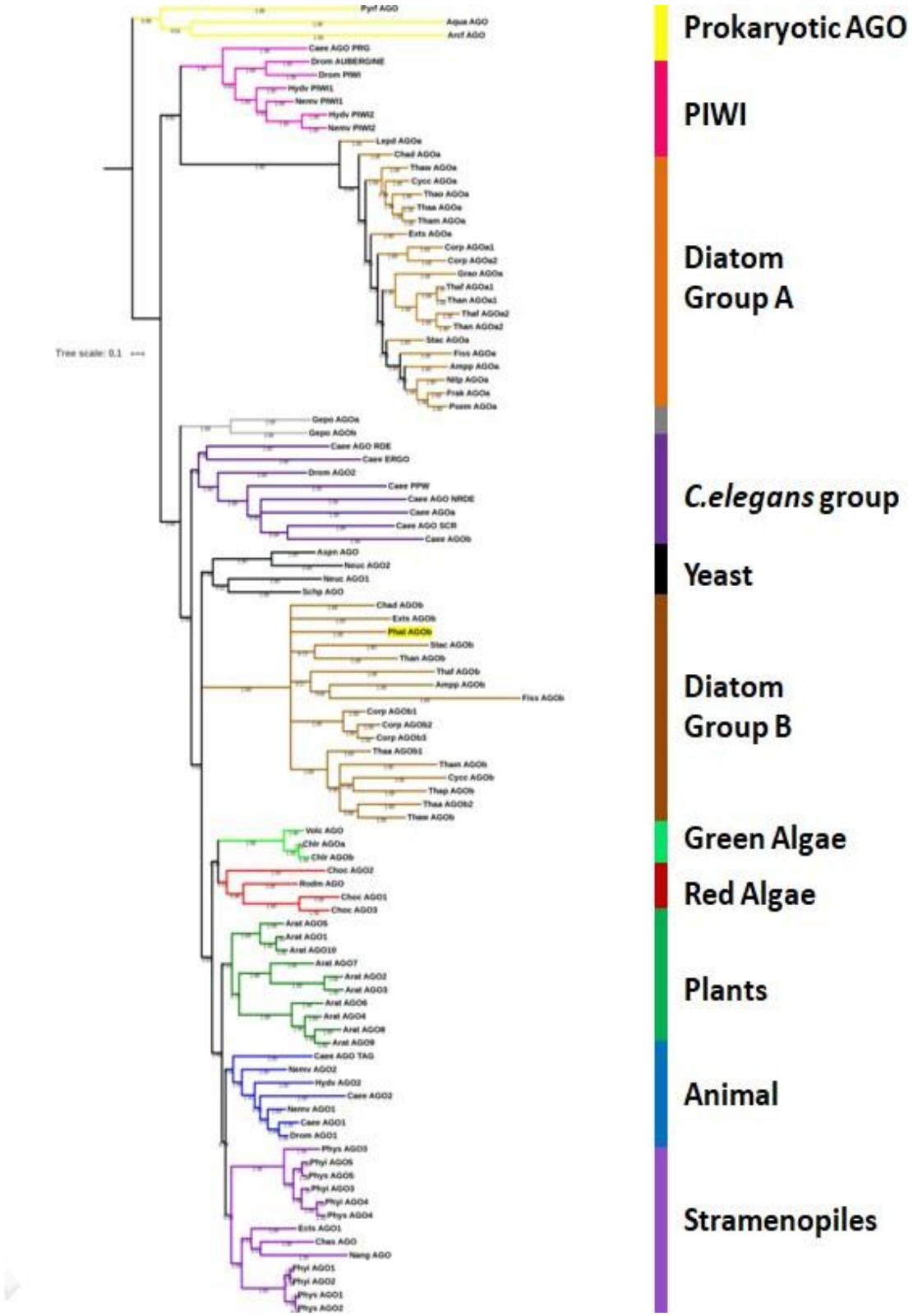

Figure 22. Bayesian phylogenetic analysis of AGO proteins (rooted). Colored taxonomical groups are indicated on the right.

3.2.3. Phylogenetic analysis of *P. tricornutum* RDR protein

RDR proteins and all RDR-containing proteins were included in this analysis, but only the RDR domains were aligned. RDR proteins are usually absent in animals, except from some basal animals, that have been included in this analysis. Some diatom and red algae species, that were used in DCR and AGO phylogenetic analysis, lack RDRs and have been excluded or replaced with close relatives, for this analysis.

A scattered distribution of RDR has been previously noticed among different organisms (citation needed) and it was anticipated in this study, too. The classification of RDRs in three types: α, β and γ was followed, based on Zong et al. (2009). The RDR domains were aligned and Bayesian analysis produced the phylogenetic tree in Figure 23.

The phylogenetic analysis showed that:

1) The majority of diatom RDRPs is grouped together and they belong to type γ RDR proteins, independently of their DCR-AGO type protein set as described above.
2) A few diatoms, specifically *Leptocylindrus danicus, Cyclotela cryptica* and *Thalassiosira oceanica* have also RDR proteins that belong to type α. All of them have, exclusively or not, DCRa and AGOa proteins.
3) Diatoms *Fragilariopsis cylindrus* and *Pseudonitzschia multiseries* have exclusively DCRa and AGOa proteins and lack an RDR protein.

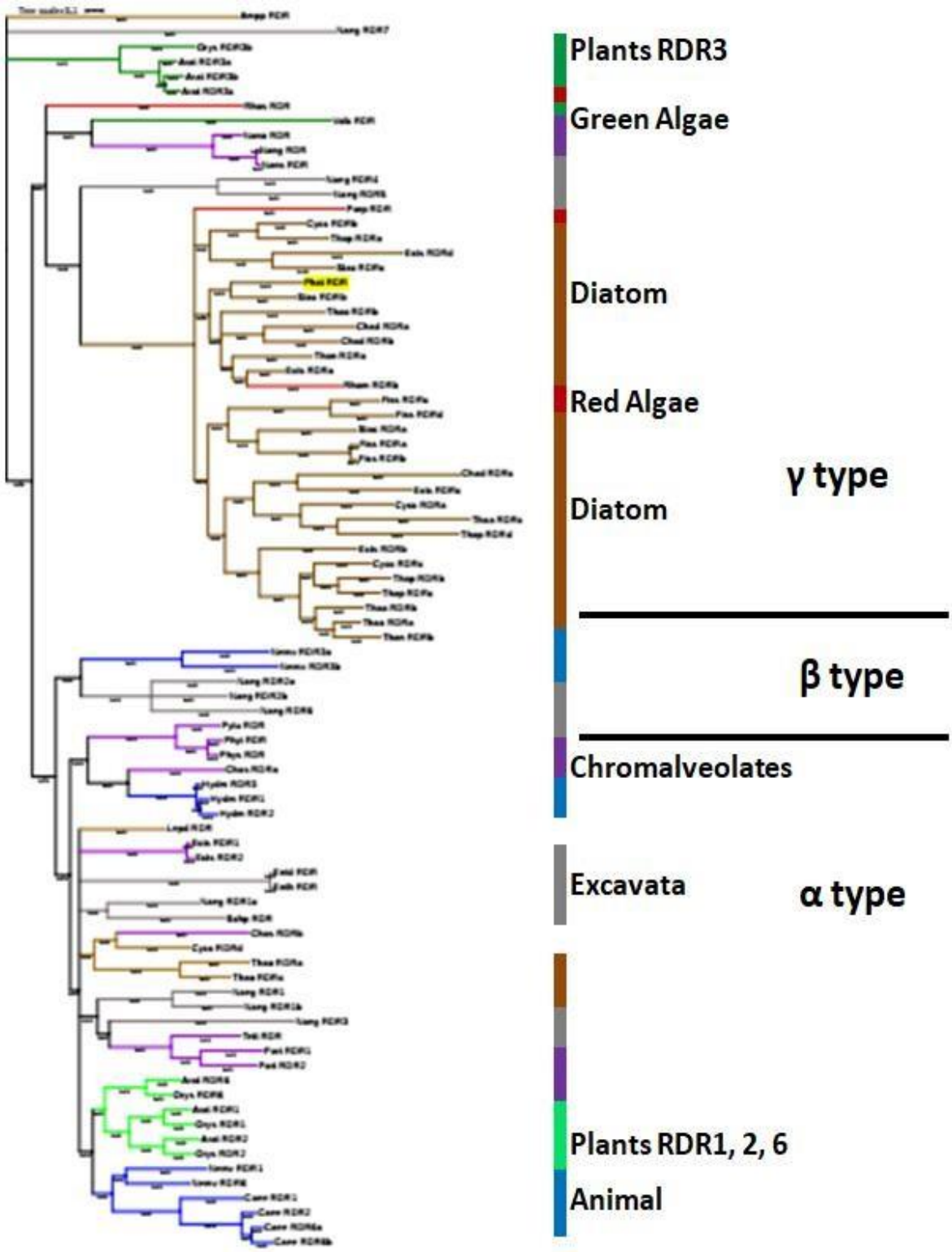

Figure 23. Bayesian phylogenetic analysis of RDRP proteins (unrooted). Colored taxonomical groups are indicated on the right along with their RDR type classification.

3.2.4. Phylogenetic analysis of RNAi-key proteins among diatoms

Based on the DCR and AGO broad species phylogenetic analyses, a larger number of diatoms from different taxonomical groups were analyzed. Considering the phylogenetic diversity among diatom taxa, the analysis of more diatom species covering a range of evolutionary diversified representatives would provide insights into the evolution of RNAi key-genes among these organisms. Predicted peptides from MMETS were identified and the two RNaseIII domains (DCR) the PIWI domain (AGO) and RDR domains were isolated and aligned as described above. Phylogenetic analysis was performed only among diatoms species with Neighbor Joining method, as the more reliable Bayesian analyses had previously validated the two diatom groups for DCR and AGO proteins. The RNAi-key protein distribution among diatoms is summarized in (Table 14).

A Maximum Likelihood phylogenetic analysis of their diatomaceous partial 18S gene was performed in order to correlate their taxonomical place with their acquired RNAi protein set and uncover the evolutionary pathway of this mechanism in diatoms (Figure 24).

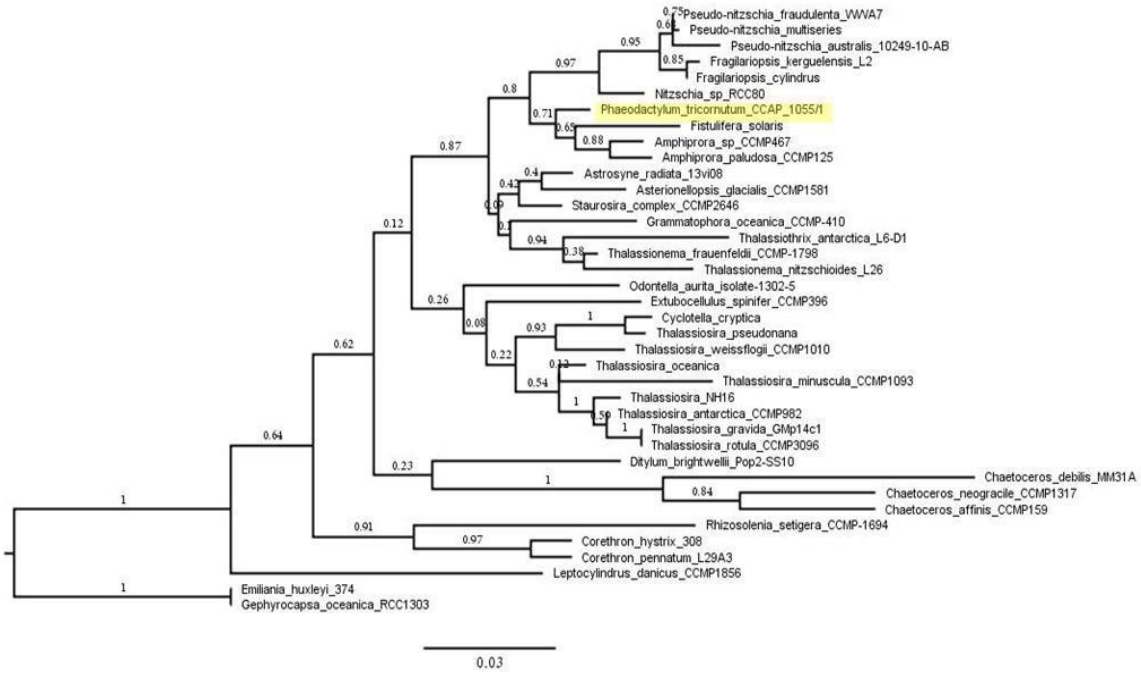

Figure 24. Phylogenetic tree of diatom partial 18S, generated with Maximum Likelihood method.

Table 14.Summary of DCR, AGO and RDR protein distribution among diatom taxa. DCR peptides with a single RNaseIII domain intact are indicated as: GramoDCRa(b), Thaf DCRb(b), TharDCRb(b), Tham DCRb(b), AstrDCRb(b). DCR peptide with only one broken RNaseIIIb domain is designated as ThagDCRa*(b).

Group		Subclass	Species	DCR-A	DCR-B	AGO-A	AGO-B	RDR-α	RDR-γ
Centric	Coscinodiscophyceae	Corethrophycidae	*Corethron hystrix*	1(b)		2	1	2	1
			Corethron pennatum	1		2	3		2
		Rhizosolenianae	*Rhizosolenia setigera*	2		2	1		1
	Mediophyceae	Chaetocerotophycidae	*Chaetoceros affinis*	2		2	1	2	3
			Chaetoceros debilis	1	1	1	1		3
			Chaetoceros neogracile	2		1			4
			Leptocylindrus danicus var. danicus	1		1		1	1
		Cymatosirophycidae	*Extubocellulus spinifer*	1	1	1	1		4
		Thalassiosirophycidae	*Ditylum brightwellii*	2		2	1	1	2
			Cyclotella cryptica	1	1	1	1	1	4
			Odontella sp	2	1	2	1	1	3
			Thalassiosira antarctica	1	1	2	2		3
			Thalassiosira gravida	*1(b)	1	1	1		2
			Thalassiosira miniscula	1	(b)	1	1		3
			Thalassiosira oceanica	1		1		2	1
			Thalassiosira pseudonana		1		1		4
			Thalassiosira rotula	1	1(b)	1	1		2
			Thalassiosira sp StrainNH16	1	1	1	1		5
			Thalassiosira weissflogii	1	1	1	1		5
Pennate araphid	Bacillariophyceae	Bacillariophycidae	*Amphiprora paludosa*	1	1	1	1		1
			Amphiprora sp	1		1			1
		Urneidophycidae	*Asterionellopsis glacialis*		1(b)		1		2
		Bacillariophycidae	*Fistulifera solaris*	1	1	1	1		4
		Fragilariophycidae	*Thalassionema nitzschioides*	1		2	1		2
			Thalassionema frauenfeldii	1	(b)	2	1		1
			Thalassiothrix antarctica	1	(b)		2	1	2
			Astrosyne radiata	1	(b)		1	1	
			Grammatophora oceanica	1(b)	1	1			2
			Staurosira complex sp	1	2	1	1		3
Pennate raphid	Bacillariophyceae	Bacillariophycidae	*Fragilariopsis cylindrus*	1		1			
			Fragilariopsis kerguelensis	1		2			
			Nitzschia punctata	1		1		1	
			Phaeodactylum tricornutum		1		1		1
			Pseudonitzschia australis	1		1			
			Pseudonitzschia fradulenta	1		1			
			Pseudonitzschia multiseries	1		1			

Diatoms arose 250 million years ago and diverged in centric and pennate clades 90 million ago. Centric diatoms are the ancestral group, while the raphid pennates have evolved more recently. From the data summarized in Table 14 we can propose the following evolutionary pathway:

Diatoms originally possessed both A and B type of DCR/AGO genes and lost type B genes along the way, with the most ancestral ones having lost type B and the most recent ones having evolved only with type A genes.

This hypothesis can be further supported by the presence of type B AGO, but not DCR, in the more ancestral centric species, as evidence of the type B RNAi genes "traces", while both type B genes are entirely lost in the most recent raphid pennates.

Moreover, the more recent Bacillariophycidae, with exclusively type A RNAi gene set (except *P. tricornutum*), seem to have lost or never obtain a RDR protein, while the later is present in ancestral lines and kept even in higher numbers in species with an active type B RNAi gene set. Gene duplication has taken place in all taxonomical groups, perhaps more extensively in Coscinodiscophyceae.

Interestingly, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*, both model diatoms on which most studies have focused and most genomic data (including small RNAs, Transposable Elements and methylome analysis) are available, represent exceptions among their taxonomic groups possessing only type B DCR/AGO proteins.

Conclusively two RNA silencing pathways (type A and B) seem to be present among diatoms that are differentially lost or conserved between species. *Phaeodactylum tricornutum*, the species under study in this thesis, although it is the best-studied model diatom, it represents an exception when it comes to RNAi among raphid pennates and highlights the necessity of future studies on RNAi in more diatom species.

3.3. *PtDCR* and *PtAGO* functional characterization in heterologous expression systems.

At the time of intitiation of this study, targeted mutagenesis for *P. tricornutum* was not feasible. Therefore, it was envisioned to functionally characterize *PtDCR* and *PtAGO* candidates in heterologous expression systems. Yeast *S. cerevisiae* offered a "clear" RNAi lacking system. In parallel, *N. bethamiana* provided a well controlled system for DCR complementation analysis.

3.3.1. Reconstitution of a functional RNA silencing pathway in *S. cerevisiae* by introducing *PtDCR* and *PtAGO* genes.

S. cerevisiae is an RNA silencing deficient yeast species. but previous studies have demonstrated a RNA silencing pathway can be established by ectopic expression of *DCR* and *AGO* homologue from yeast species *S. castellii* and *Homo sapiens* (Drinnenberg et al., 2009). The same approach was used, in order to express *P. tricornutum DCRb* and *AGOb* genes simultaneously in *S. cerevisiae* and assess the ability of the introduced genes to replace *S. castellii* genes in RNA silencing targeting a *GFP* transgene and the *URA3* endogenous gene.

3.3.1.1. Silencing the endogenous URA3 gene in S. cerevisiae

S. cerevisiae strains that contain functional *URA3* genes (URA3) are autotroph for uracil and can grow on plates with media SC-Ura (lacking uracil), while mutant lines for URA3 genes (ura3) are auxotrophes. Conversely, URA3 strains are sensitive to 5FOA while ura3 are not. So, uracil and 5FOA allow the selection/counter-selection for the expression of URA3 gene in yeast.

As previously described by Drinnenberg et al (2009), the system of reconstituting RNA silencing in *S. cerevisiae* is constructed in the following way: *S. cerevisiae* strains expressing *S. castellii AGO* and *Dicer* genes and either the galactose induced Ura hairpin construct (Hp) or no silencing construct (Ø), were tested for Ura3p expression by plating serial dilutions on galactose complete medium (SC), medium lacking uracil (SC–Ura) and medium containing 5-FOA (Figure 25).

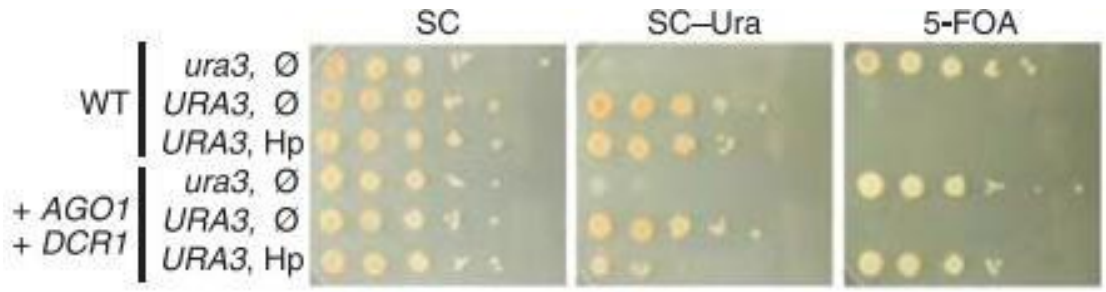

Figure 25. *S. cerevisiae* strains with nonfunctional ura3 or functional URA3 genes, expressing *S. castellii DCR/AGO* genes and either the hairpin construct (Hp) or not (Ø), were tested for Ura3p expression, based on their uracil autotrophy or sensitivity to 5-FOA, on galactose complete medium (SC), medium without uracil (SC–Ura) and medium containing 5-FOA (Drinnenberg et al., 2009).

As expected, all strains grew on complete medium. Strains with ura3 gene were not able to grow on SC-U medium but were able to grow on complete medium containing 5-FOA. Since, the URA3 strain expressing *DCR* and *AGO* managed to grow on complete medium containing 5-FOA but not on SC-U medium when the hairpin was expressed, indicating that the URA3 gene was silenced. The small amount of growth of the two first colonies on SC-U can be explained as a lag in growth due to the amount of time needed by RNA silencing to take place and due to traces of already produced URAp in the cells. The opposite effect was observed when the hairpin was absent, validating the presence of a functional URA 3 gene.

In this study, the yeast strains and vectors were kindly provided by Dr. Bartel lab in order to test the function of *PtDCRb* and *PtAGO*b. As a control, we attempted to recapitulate RNAi with *ScDCR* and *ScAGO* by introducing them in the URA3+Hp strain (DRB72 strain). However, after multiple transformation attempts to introduce *S. castellii DCR* and *AGO* with integrating vectors, the same results could not reproduced (Figure 26). To overcome this issue, it was decided that *S. castelli* genes would be cloned together in a binary episomal vector with a tryptophan marker and transform the URA+Hp strain. However, the silencing of the URA3 gene failed to take place. The problem appeared to be the hairpin construct (with HIS3 selection) that was unstable and was lost after new rounds of transformation. The issue was resolved by performing all

rounds of transformation in the URA3 strain without the integrated hairpin (strain DPB271) sequentially, starting with the introduction of *ScDCR*, continuing with *ScAGO* and finally with the URA hairpin. Both the integrative and episomal vectors carrying *ScDCR* and *ScAGO* were tested (Figure 26). When URA3 strain was transformed with either integrative or episomal vectors carrying *ScDCR*, *ScAGO* and then the integrative URA hairpin, the RNA silencing pathway was reconstituted and URA3 gene was silenced as shown in (Figure 26).

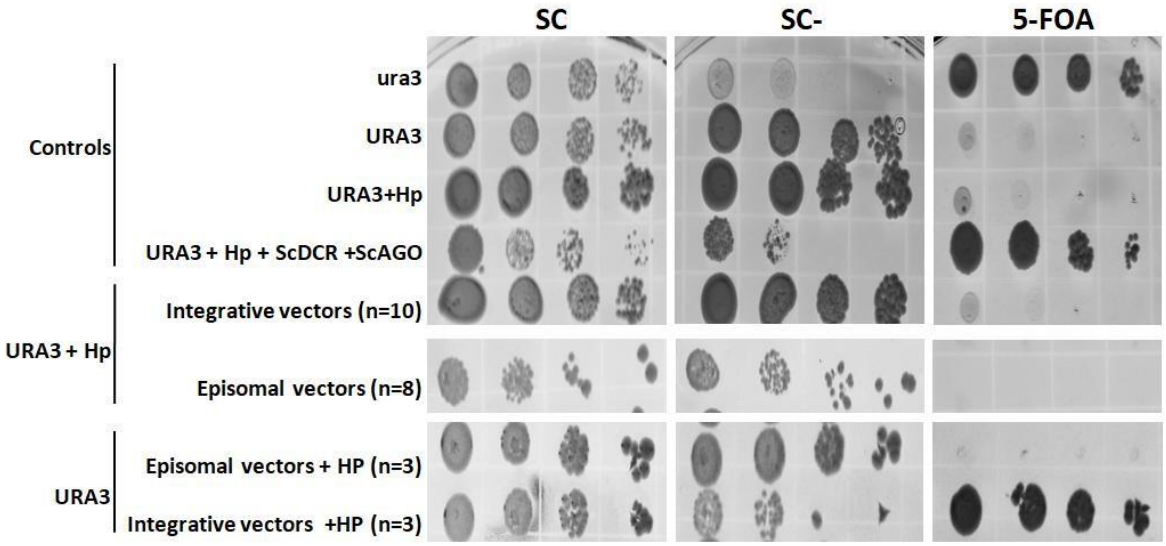

Figure 26. *S. cerevisiae* URA3 and ura3 strains, expressing hairpin construct (Hp) and *ScDCR/ScAGO* were tested for Ura3p expression, based on their uracil autotrophy or sensitivity to 5-FOA, on galactose complete medium (SC), medium without uracil (SC–Ura) and medium containing 5-FOA. Episomal and integrative vectors carry *ScDCR/ScAGO*.
***S. cerevisiae* URA3 strains with the Hp integrated before *ScDCR/ScAGO* introduction didn't demonstrate RNAi. Only when Hp was introduced after the *ScDCR/ScAGO* the RNA silencing of URA3 took place.**

During the process of setting this system up and recreating the previously published RNAi efficient strains, a variety of technical issues were observed: a) inconsistency in the growth of yeast strains, b) inconsistency in the silencing efficiency of the transformants, c) inability to reproduce positive results, d) reversion of the tryptophan mutation and therefore loss of auxotrophy and most importantly e) loss of the silencing ability that the RNAi positive clones had over time and during storage at -80 ^{o}C. Thus, the exploitation of the transgene GFP silencing system in strains carrying GFP+Hp was decided.

3.3.1.2. Silencing the GFP transgene in S. cerevisiae

In the system of Drinnenberg et al (2009), a *S. cerevisiae* GFP-transgenic strain that expressed either a strong (St) or a weak (Wk) silencing construct under galactose induction, generated GFP siRNAs when ScDcr1 was introduced (Figure 27). When both *ScAGO* and *ScDCR* were present, silencing of GFP could also be measured by fluorescence-activated cell sorting (FACS), with the strong GFP construct repressing fluorescence to background autofluorescence (Figure 27).

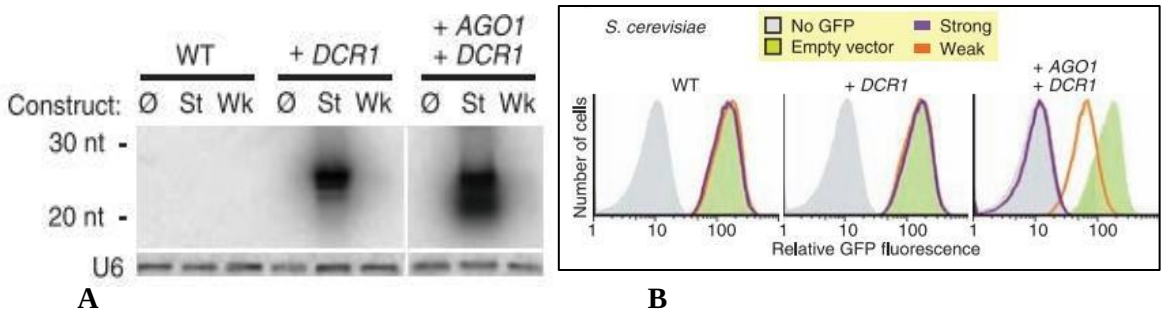

Figure 27. A. RNA blot demonstrates production of GFP siRNAs in *S. cerevisiae* strains expressing either no *S. castellii* genes (WT) or the indicated integrated *S. castellii* genes, and either a GFP hairpin (St), a GFP antisense construct (Wk), or no (Ø) silencing construct. B. FACS histograms showing GFP fluorescence in the *S. cerevisiae* galactose-induced strains (WT, GFP, GFP+DCR, GFP+DCR+AGO) expressing the indicated silencing constructs (color boxes) (Drinnenberg et al., 2009).

For this study, the same DPB249 strain (expressing GFP) was transformed with pRS405-PTEF –ScDCR, pRS404-PTEF –ScAGO and the strong GFP hairpin construct pRS403-PGAL1 -strongSC_GFP (Hp) or no silencing construct (Ø). Strains grew in galactose medium to induce hairpin production and were subsequently tested for GFP expression by fluorescence-activated cell sorting (FACS) and for GFP siRNAs generation by Northern blot analysis. Due to the growth, transformation and efficiency issues described above it was decided to test complementation by both integrative and episomal vector based expression. Initially, it was attempted to introduce the integrative GFP hairpin before *ScDCR/ScAGO* genes without successful establishment of RNAi. Subsequently, the integrative GFP hairpin was introduced last and in linear form.

Because the GFP transgenic lines were silenced at the same level with both integrative and episomal vectors (data not shown), the episomal plasmids p2WG-*ScDCR/ScAGO* and p2WG-*PtDCR/PtAGO* were preferred for further analysis. A p2WGT-*PtDCR/ScAGO* plasmid was additionally prepared in order to test for PtDCR activity separately. The positive RNAi strains with integrative vectors were stored at -80ºC. Cultures grown for different time periods under galactose induction were prepared before the FACS analysis of GFP producing cells versus GFP silenced cells (Figure 28).

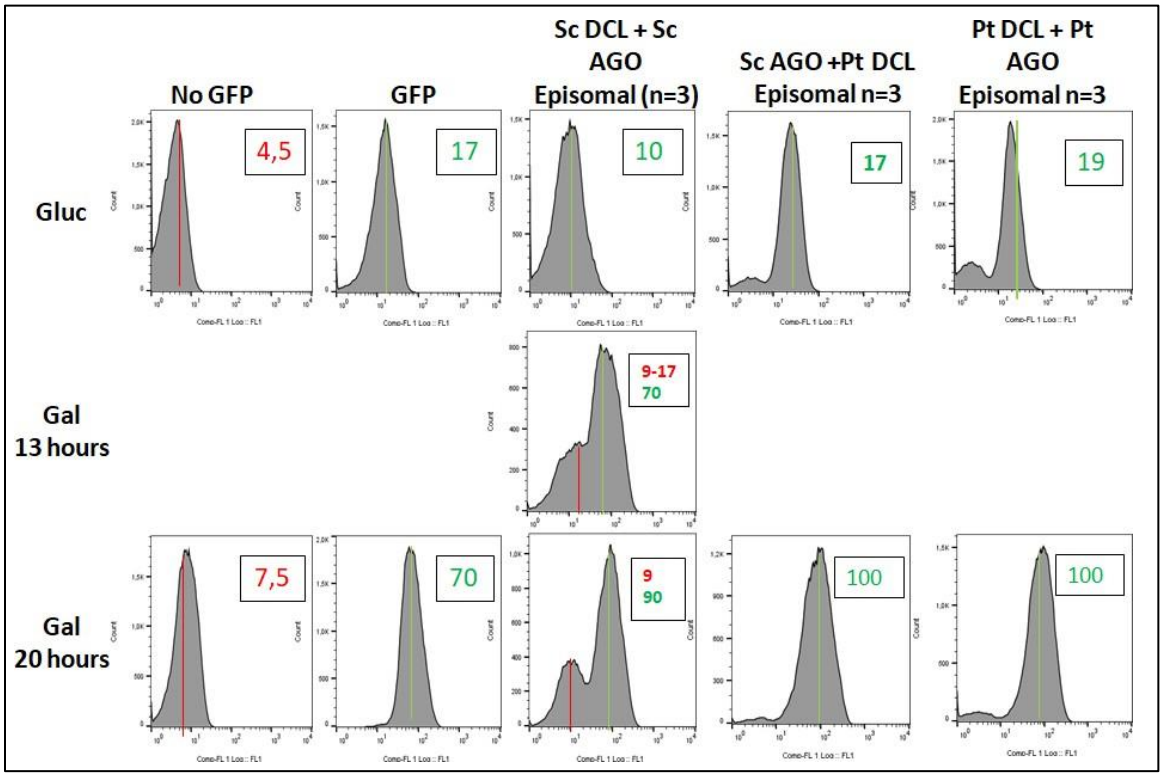

Figure 28. FACS analysis of *S. cerevisiae* strains): WT, expressing GFP, expressing GFP +Hp +*ScDCR/ScAGO*, expressing GFP+Hp +*PtDCRb/ScAGO* and GFP+Hp +*PtDCRb/PtAGOb*. Levels of autofluoresence are indicated in WT and represent the 100% silencing of GFP transgene. So, efficient silencing would "swift" levels of GFP fluorescence in GFP strains growing in galactose to the WT-fluoresence levels when strains grew in galactose (inducing Hp expression). Boxed numbers indicate the higher level of GFP fluorescence (green) and the lower levels close to cell autofluoresence (red).

GFP strain transformed with *ScDCR, ScAGO* and the hairpin presented incomplete GFP silencing after 20 hours of galactose induction. Expression of *PtDCRb/ PtAGOb* or *PtDCRb/ScAGO* did not show any silencing effect, as the small peak with lower fluorescence corresponded to dead cells. The partial silencing efficiency demonstrated by the positive RNAi strains with *ScDCR/SCAGO* is in contrast to the almost 100% silencing efficiency reported by Drinnenberg et al, 2009.

During this experiment, analyzed transformants were always "fresh" after transformation procedures, different sequential transformations were tested introducing first or last the hairpin construct and immediate cryopreservation was performed in order for positive clones to maintain RNAi activity. However, the system appears to be unstable. This observation was confirmed by the Roth lab working on human RNAi in yeast that encountered the same problems.

The transcription of *PtDCR* and *ScDCR* in *S. cerevisiae* strains was validated by Northern blot analysis (Figure 29). In order to explore the possibility that PtDCRb can produce sRNAs, but truncated PtAGOb is not functional and ScAGO is unable to use the specific PtDCRb produced sRNAs, total RNA was extracted from the same strains before FACS analysis (Figure 28) and Northern blots were performed to detect the presence of GFP sRNA (Figure 30).

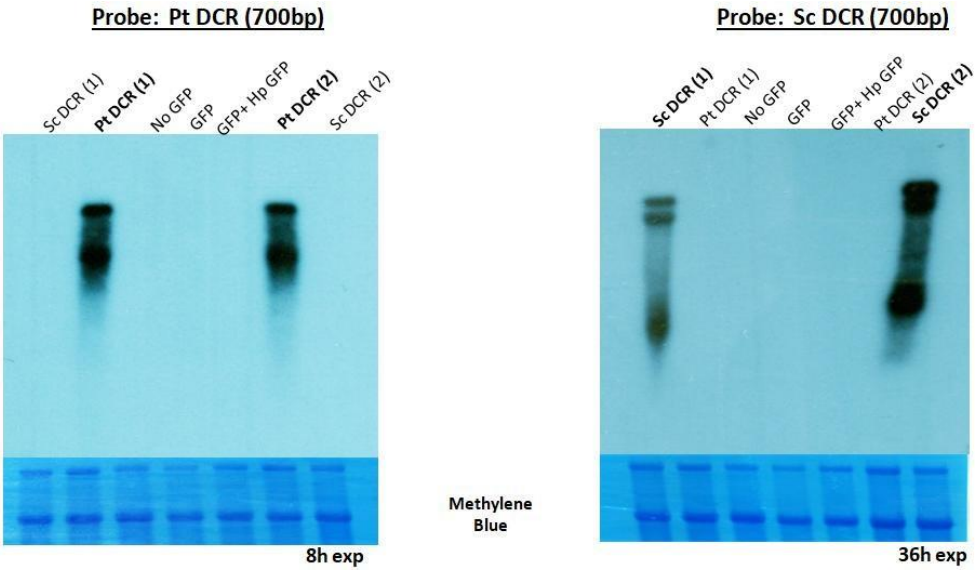

Figure 29. Northen Blot for *PtDCRb* and *ScDCR*. *S. cerevisiae* GFP transgenic strains were transformed with *PtDCRb* or *ScDCR* genes and GFP Hp. Two different transformed strains were tested (numbers 1 and 2).

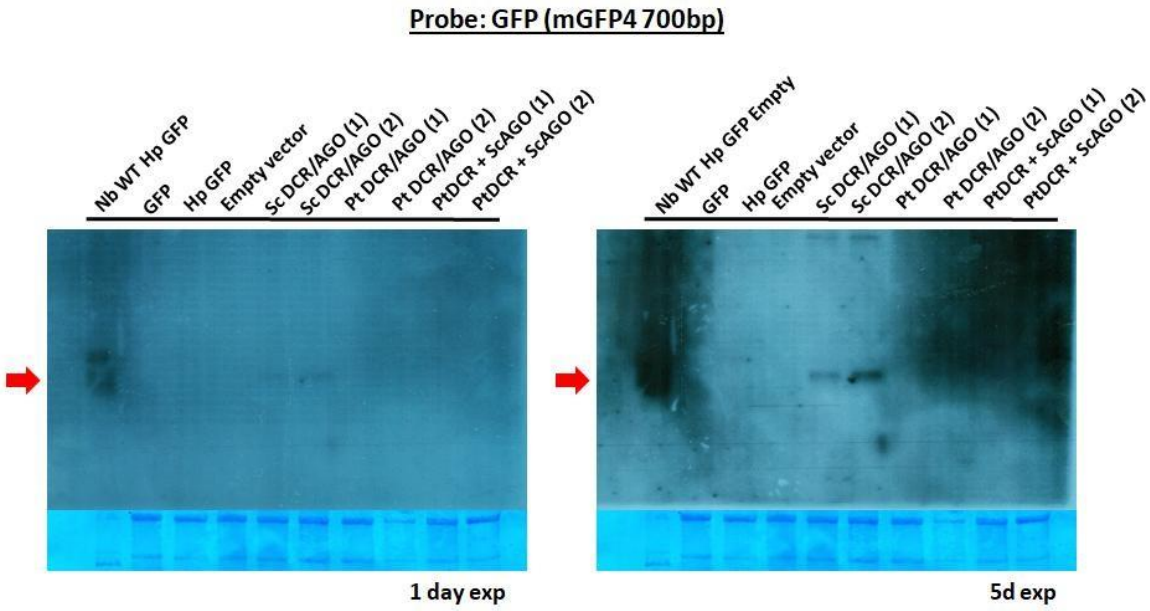

Figure 30. Northern Blot analysis of GFP small RNAs in the transgenic GFP yeast strains transformed with GFP HP and either *PtDCR/AGO* or *ScDCR/AGO* or *PtDCR+ScAGO* (two different strains of each transformation, 1 and 2). *N. bethamiana* (Nb) transgenic GFP line (16C) transiently expressing GFP Hp was used as positive control for the detection of GFP sRNAs (red arrow).

Northern blot analysis (Figure 30) demonstrated that although *PtDCR* was transcribed, it did not produce small RNAs like ScDCR, either in combination with *PtAGO* or with *ScAGO*. Analysis was conducted using 23µg RNA and the film was overexposed (5 days) in order to maximize the detection of GFP sRNA. Together, these results in *S. cereviciae* system indicate that possibly either the *PtDCRb* spliceoform was not functional or it needs other cofactors to efficiently produce small RNAs in this host.

3.3.2. *PtDCR* expression in plant *N.bethamiana*

RNA silencing is relatively well known in plants. Functional complementation assay of a transgene RNAi that is not targeted by the endogenous system would be instrumental in deciphering the function of key RNAi components.

N.benthamiana transgenic lines Nb DCL2/3/4i with downregulation of DCL2, DCL3, DCL4 have been previously generated and validated in the lab. Since *PtDCR* is so divergent from *NbDCLs*, the silencing constructs used for *NbDCLs* downregulation could not target *PtDCR* based on sequence similarity. Thus, this plant heterologous expression

system seemed optimal to explore the possibility of having a functional PtDCR. In addition, eventual cofactor needed for PtDCR function may also be present in *N.bethamiana* in order to drive an efficient RNAi pathway. NbDCL1 is normally expressed (as its downregulation is proven lethal) along with NbAGO set of proteins and every other component participating in *N.bethamiana* RNA silencing mechanisms. Nb DCL2/3/4i line was shown to produce only 21 nt small RNAs from DCL1 activity while 22 and 24 nt were absent (Figure 31b) and had a suppressed RNAi mechanism.

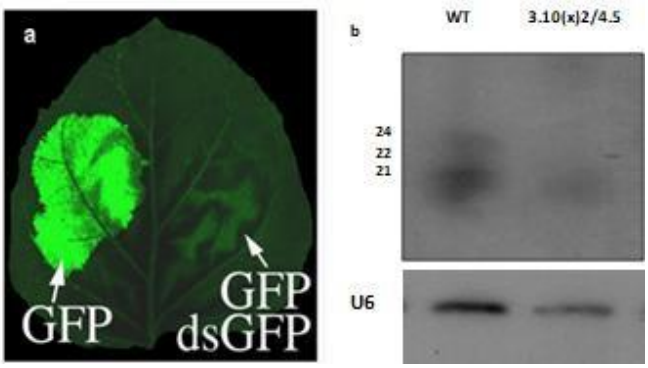

Figure 31. a) Fluorescence of transiently expressed GFP without (left) and with (right) silencing construct in wild type *N.bethamiana* (Bazzini et al., 2007), b) PSTVd small RNA population in *N.bethamiana* wild type and DCL2/3/4i lines after PSTVd infection (Katsarou et al., 2016).

Based on the possible compatibility between the two photosynthetic organisms, PtDCR was tested to check whether it could complement the function of another downregulated NbDCL, when acting in a RNAi efficient environment. In this case, only the *P. tricornutum* DCR function could be evaluated and *S. castellii* DCR was used as a positive control, since it has an experimentally validated Dicer activity.

PtDCR and *ScDCR* were transiently expressed by agroinfiltration in *N.benthamiana* wild type and the mutant line DCL2/3/4i. Together with *ScDCR* or *PtDCR*, a GFP transgene and a GFP hairpin were transiently co-expressed (Figure 32). GFP fluorescence was monitored over time under UV lamp (Figure 31a). In parallel Northern blot analysis was conducted to test for the presence of GFP sRNAs. Importantly, GFP sRNA could still be present even in absence of observable GFP silencing because of PtDCR produced sRNAs being non-compatible with NbAGOs. This analysis would also provide the size estimation of small RNA population produced by PtDCR.

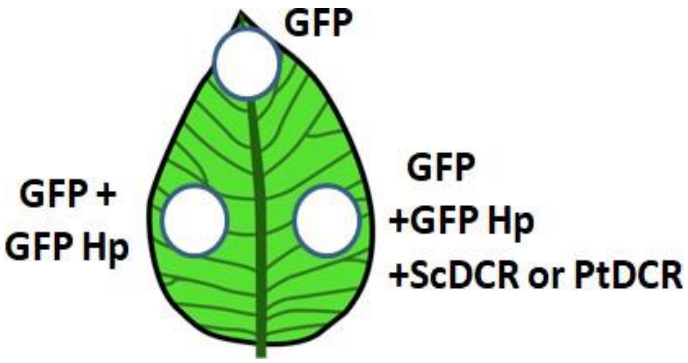

Figure 32. Representation of the agroinfiltrated *DCR* genes and GFP constructs. On top, there is GFP only as a control of tissue quality and plant's growth conditions. All spots were agroinfiltrated with the same OD (0,2max)of agrobacteria by adding bacteria carrying empty vectors, when needed.

Expression of *ScDCR* or *PtDCR* did not complement DCL2/3. In the WT line a GFP transgene and a GFP hairpin did not show any effect of DCR complementation and GFP silencing activation in the mutant line. PtDCR did not alter GFP fluorescence at all in dcl2/3/4i line and probably delayed GFP silencing in the WT line (Figure 33).

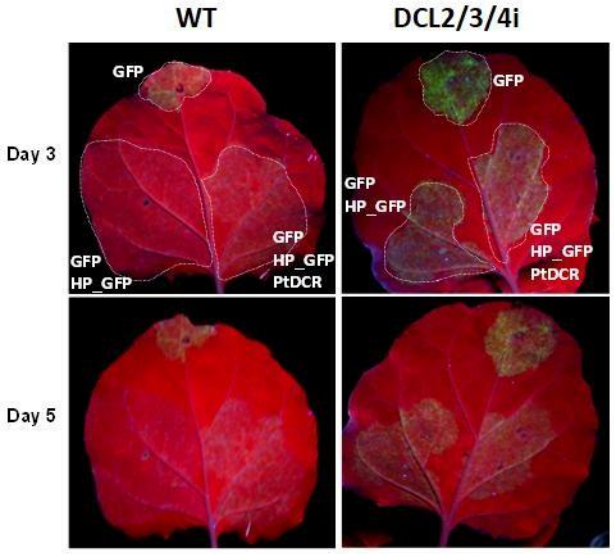

Figure 33. *N. bethamiana* WT (left) and dcl2/3/4i (right) lines transiently expressing GFP transgene, GFP hairpin and *PtDCR*. Photos were captured under UV radiation, at 3rd and 5th day after agroinfiltration.

In this agroinfiltration experiment, levels of GFP fluorescence were relatively low and the WT line was silencing the GFP transgene too quickly for observation. After repeating the experiments, the agrobacterium carrying the GFP construct was revealed to be problematic and was subsequently replaced. The Northern blot analysis of Hp GFP small RNAs was carried in Nb WT and Nb dcl2/3/4i lines transiently expressing a GFP hairpin and either *ScDCR* or *PtDCR* or empty vector. No increase of small RNA populations in WT or any new population of small RNAs in Nb dcl2/3/4i mutants after expressing *Pt DCR* (b-c spliceoforms) or *ScDCR* was observed (Figure 34).

A second experiment in the same direction, that was based on RNA silencing of the plant-infecting viroid PSTVd (*Potato Spindle Tuber Viroid*) transcript, validated these results. Small RNAs produced in PSDVd infected *N. bethamiana* wild type and DCL2/3/4i, after transient expression of *ScDCR* or *PtDCR* were analyzed with Northern blot and the population of PSTVd small RNAs was investigated. In this case, a successful PSTVd infection was observed in *N. bethamiana* plants, ensuring the presence of a stable high level of viroid RNA, so only the transient expression of *ScDCR* or *PtDCR* or empty vector was required (agroinfiltration). Again, *ScDCR* and *PtDCR* spliceoforms did not affect the populations of sRNA in any Nb line (Figure 35). *N. bethamiana* was concluded to be an unsuitable heterologous expression system for this case and it was not used in any further experimental approaches.

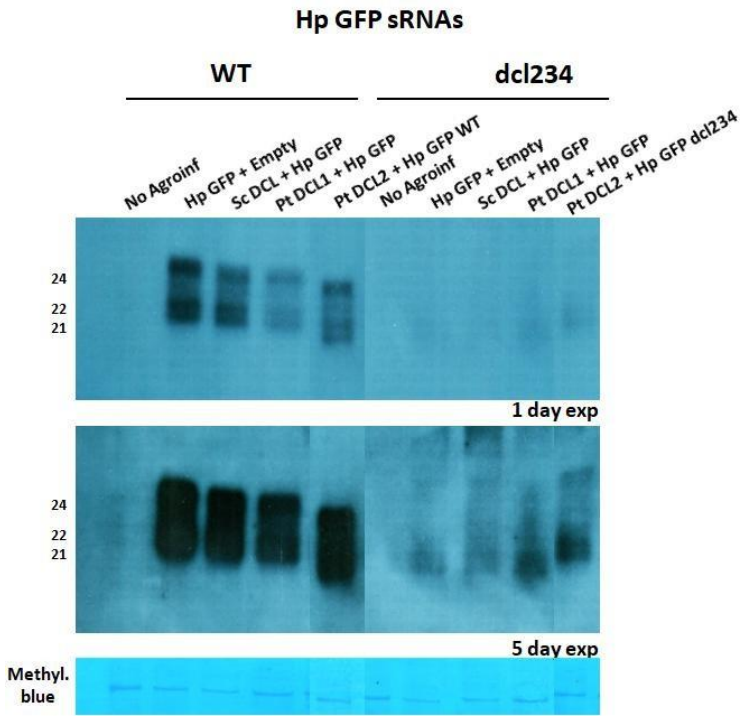

Figure 34. Northern blot analysis of Hp GFP small RNAs produced in *N. bethamiana* WT and dcl2/3/4i lines after transient expression (agroinfiltration) of Hp GFP and either *ScDCR* or *PtDCR* or empty vector.

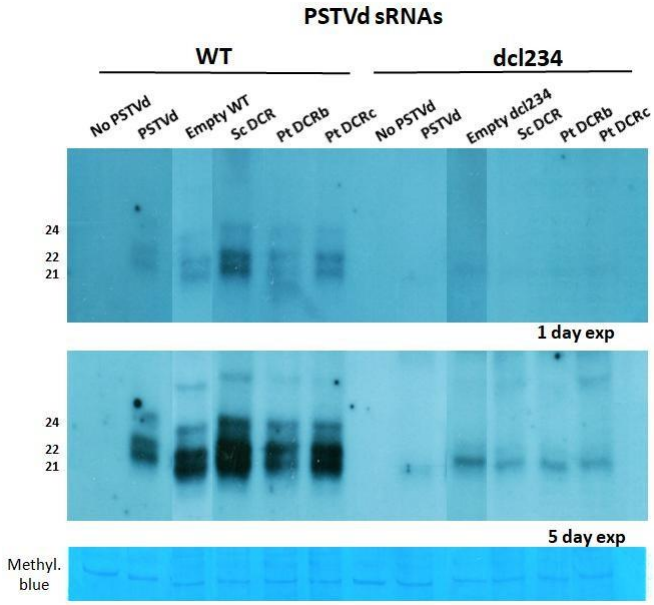

Figure 35. Northern blot analysis of PSTVd small RNAs produced in *N. bethamiana* PSTVD infected WT and dcl2/3/4i lines after transient expression (agroinfiltration) of either *ScDCR* or *PtDCR* or empty vector.

3.4. Subcellular localization of PtDCR, PtAGO and PtRDR proteins

P. tricornutum WT cells were co-transformed with a vector carrying *DCRa/AGOa/RDR* genes fused with YFP at their C-terminus under FcpB promoter. Clones grown on selective plates were validated for the presence of -YFP constructs by PCR amplification of the fused area (800 bp). Cultures from positive clones were grown under optimal conditions until early log phase and cells were collected after 8-9 hours from initiation of photoperiod. Cells were stained with Hoechst dye and observed with confocal microscopy.

Although numerous positive clones were identified by PCR, only the YFP control and two RDR-YFP clones exhibited YFP fluorescence. Therefore, all putative positive clones were tested further by amplifying the whole fused construct from *DCRa/AGOa/RDR* ATG to the stop codon of the YFP sequence. Electrophoresis in agarose gel validated the presence of many broken constructs, while only the two RDR-YFP clones exhibiting YFP fluorescence contained the whole length construct (Figure 36). This effect is probably caused by the introduction of long constructs via biolistic transformation. *P. tricornutum* RDR protein is localized in the nucleus, as expected, and specifically at nucleoli regions where transcription is very active (Figure 37).

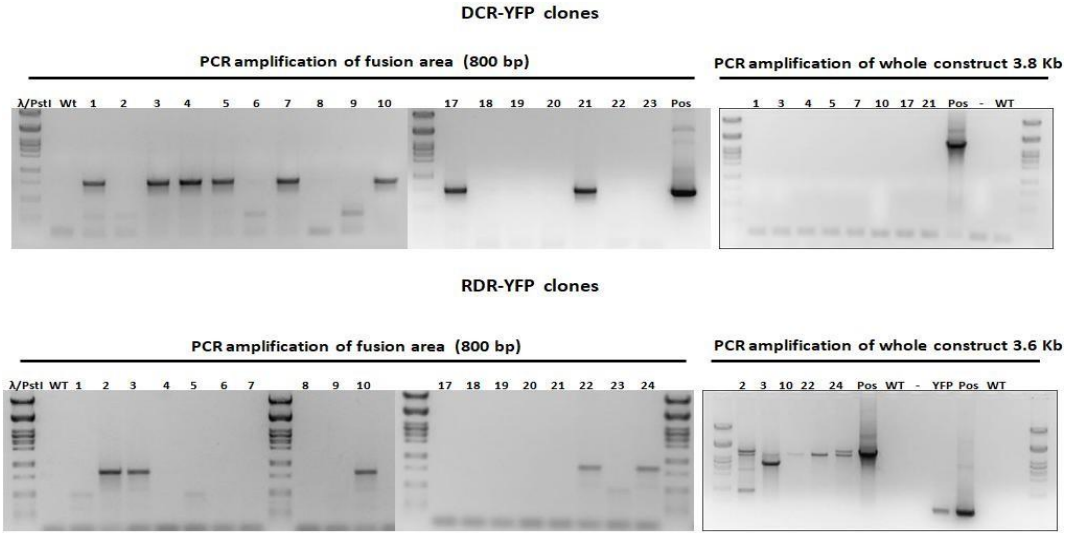

Figure 36. PCR amplification of the *DCR/RDR -YFP* fusion area (on the left) and of the whole *DCR/RDR-YFP* construct (on the right) from clones grown on selective plates. The plasmids used for transformation were also used as positive controls (Pos).

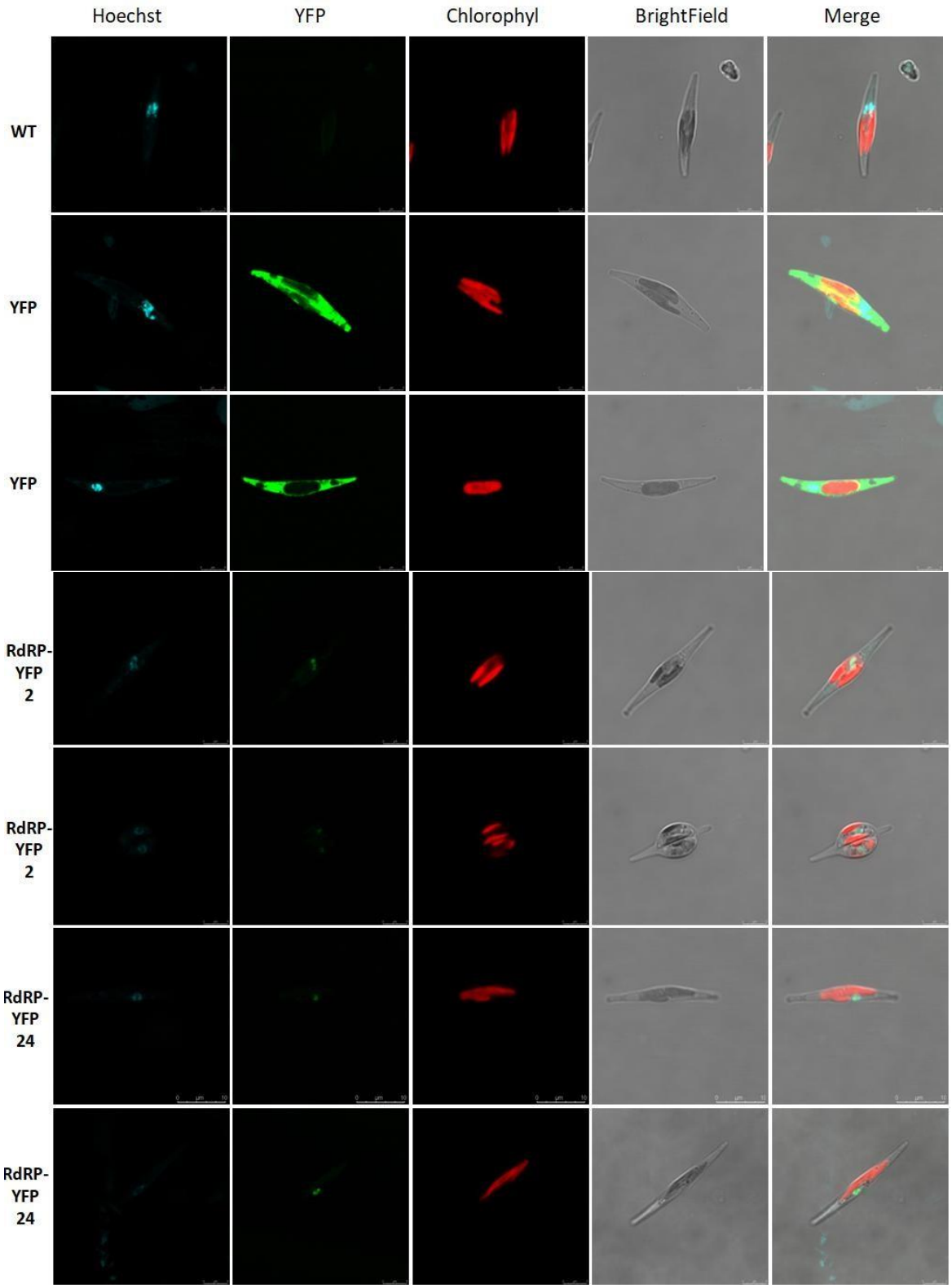

Figure 37. *P. tricornutum* RDR protein is localized in the nucleus and more specifically at the nucleoli. YFP expressing strain is used for positive control and WT strain as negative control.

3.5. PtDCR and PtAGO functional characterization by CRISPR/Cas9 generated DCR and AGO Knock-Out mutants.

CRISPR-Cas9 mediated mutagenesis methodology in *P. tricornutum* was developed during the time of this PhD (Nymark et al., 2016a). It provided the unprecedented opportunity to conduct reverse genetic approach through the generation and phenotypic analysis of DCR, AGO and RDR knock out mutant lines. During the EMBO short-term fellowship (3 months) in Dr. A. Falciatore's lab, UPMC, Paris, the generation of DCR/AGO KO mutants was conducted in WT lines and GUS or GUS/RNAi lines, previously generated in that lab (Riso et al., 2009).

3.5.1. CRISPR/Cas9 experimental design

Single guide RNAs (sgRNA) targeting *DCR*, *AGO* and *RDR* (sgRNA-DCR/AGO/RDR) were designed. Mutagenesis was targeted upstream of the sequence predicted to encode functional domains (e.g. upstream PAZ domain of DCR) or within the functional domain (e.g. RDR) to promote a loss-of-function mutation, resulting from frameshift insertions/deletions (INDELs). For a more efficient mutagenesis and an easy screening method, two sgRNAs were used simultaneously per gene target, in order to create deletions of 300-600 bp that could be visible in an agarose gel following PCR amplification. Each sgRNA was cloned in a Cas9 containing vector (pKS diaCas9), generating pKS diaCas9_sgRNA-DCR/AGO/RDR vectors (Figure 38).

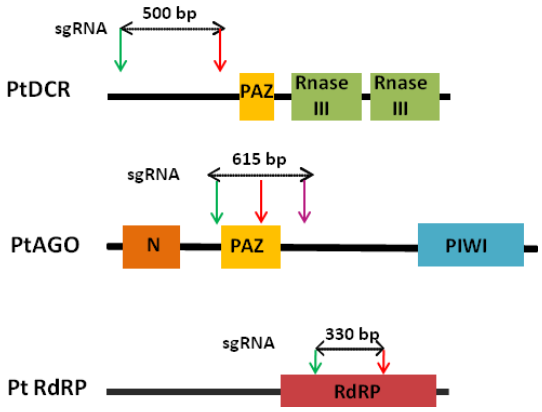

Figure 38. Illustration of sgRNA targets in *P. tricornutum DCR, AGO* and *RDR* gene.

3.5.2. Generation of PtDCR KOs and PtAGO KOs.

Two vectors, each carrying Cas9 and a sgRNA were used per transformation, along with the pNAT vector for the WT line or pBlast vector for the GUS and GUS/RNAi line. For PtAGO gene, three sgRNAs were designed and tested in different combinations (Figure 39), but only the first two (Figure 38) were efficient in mutagenesis. Moreover, the generation of double mutants Pt DCR/AGO KOs was attempted by transforming cells with 5 plasmids: 2 targeting DCR, 2 targeting AGO and the pNAT vector.

Vectors were co-delivered by biolistic transformation in WT and GUS, GUS/RNAi background cell lines and transformants were selected on nourseothricin containing solid media. Individual colonies were recovered after 2-3 weeks and screened for frameshift INDELs.

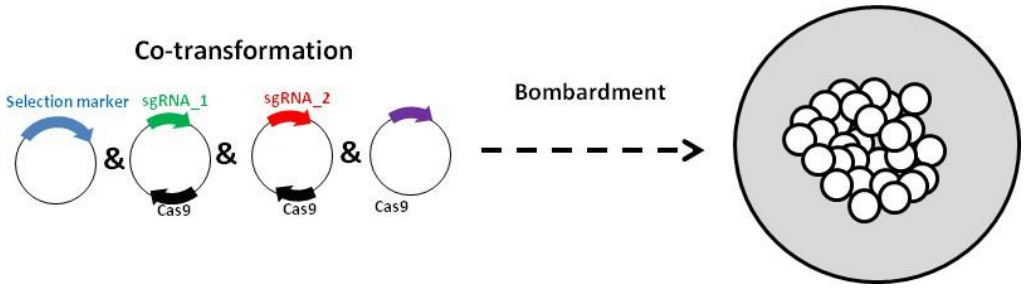

Figure 39. Biolistic transformation, where tungsten microparticles coated with the plasmids of interest are bombarded into *P. tricornutum* cells.

3.5.3. Screening and validation of mutations in PtDCR KOs and PtAGO KOs

Initially, nourseothricin resistant clones from WT background (containing the NAT gene) were screened by PCR for the presence of Cas9 gene that would indicate the presence of at least one sgRNA, too. Then, PCR amplicons containing the targeted locus of DCR/AGO from Cas9 positive mutant lysates were screened in simple agarose gels (Figure 40). Amplicons of mixed variously sized bands (appearing as a smear) were the product of mixed-clone lysate amplification; their corresponding primary colonies were re-streaked, in order to isolate single clones. Bands with smaller sizes in comparison to the WT-like size-band were extracted and sent for Sanger sequencing. The resulted chromatographs were mostly mixed sequences. The CRISP-ID tool

(http://crispid.gbiomed.kuleuven.be/) enables the identification of two sequences, usually deriving from CRISPR/Cas9 differentially mutated alleles. However, more than two sequences were routinely present, indicating the presence of more than one clones and the necessity of serial re-streaking rounds in order to isolate monoclonal biallelic mutants.

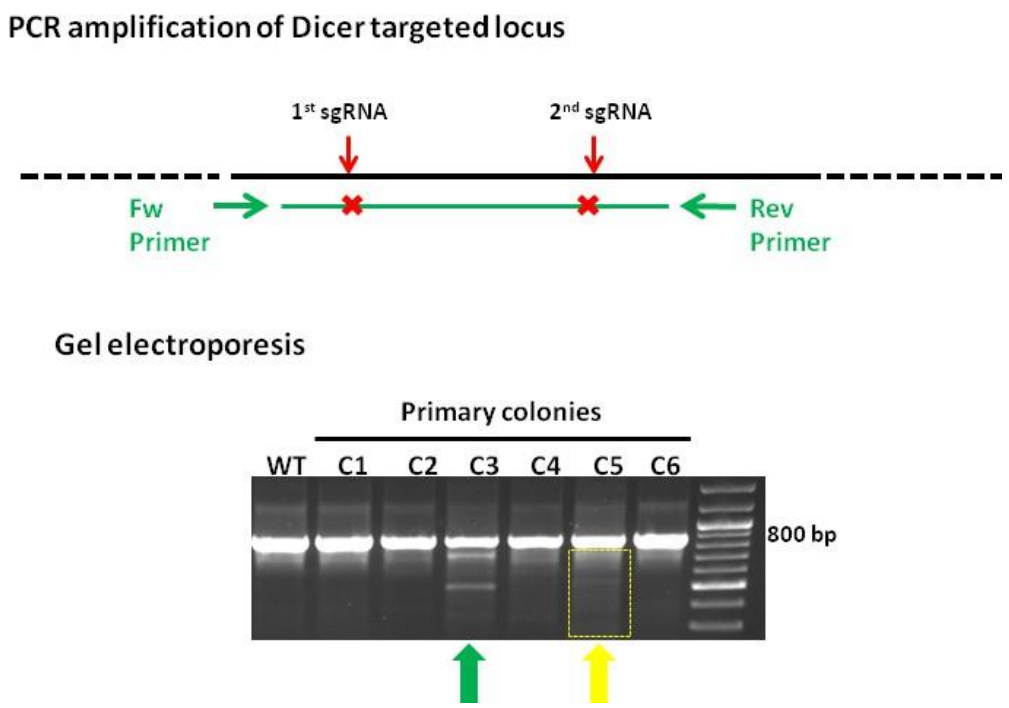

Figure 40. Screening PtDCR KO mutants by amplification of the targeted locus (primers hybridizing upstream and downstream the targeted sequence) and agarose gel electrophoresis. Mutant primary colony C3 showed a clear pattern of two bands with smaller size in comparison to WT (green arrow), while C5 showed a smeary pattern of multiple smaller-sized bands (yellow arrow) that indicates a mixture of multiple mutated clones in this colony.

Subsequently, PCR amplicons of the DCR/AGO targeted locus from the isolated monoclonal lysates run in agarose gels (Figure 41) and a) samples with obvious deletion size-patterns were sent for Sanger sequencing and b) samples with a WT-like size of band were tested with high resolution melting (HRM) curve analysis for small single point INDELs.

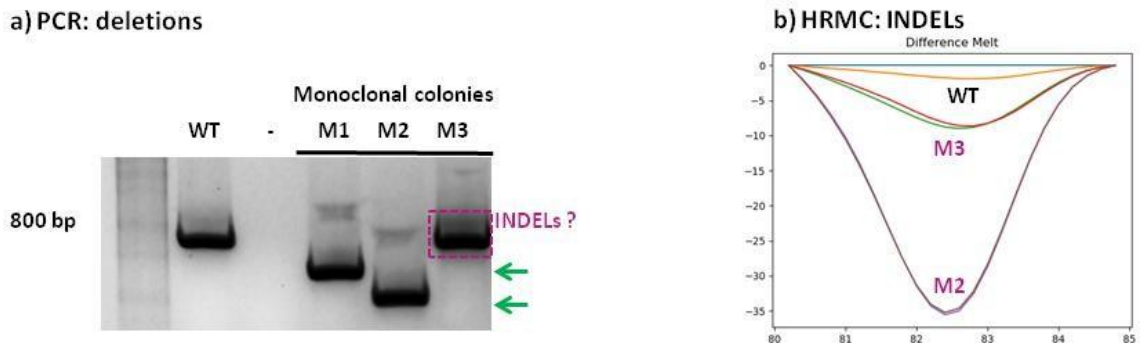

Figure 41. Amplicons of *PtDCR* targeted locus from monoclonal colonies, that derived from the C5 primary colony, were analyzed in agarose gel electrophoresis (a). Mutant colonies M1 and M2 showed a clear pattern of single smaller-sized band in comparison to WT band, indicating the presence of a big INDEL or deletion. Mutant colony M3 demonstrates the WT-like size of band and needs to be screened through HRMC analysis (b) in order to determine the presence of a small INDEL or a WT clone. In this case a WT line, M3 mutant and M2 mutant were analyzed to validate the sensitivity of HRMC method.

Validation of biallelic mutations with a frameshift (preferably homozygotes) was performed with extraction of genomic DNA, amplification of the targeted locus, and cloning in vectors that were sent for Sanger sequencing (Table 15).

Table 15. Generated and validated PtDCR and PtAGO KO mutants in WT background.

	Primary Colonies			Mutants					
						In Frame		Out Of Frame	
	Total	Cas9^{+}	Deletion^{+}	Total	Bi-allelic	INDEL	Deletion	INDEL	Deletion
DCR	45	21	8	8	8	0	0	6	2
AGO	13	9	3	14	14	2	3	9	0
DCR&AGO	48	21	4	1	1	0	0	1	0
1st sgRNA						2	0	13	na
2nd sgRNA						0	0	1	na
1st&2nd sgRNA						0	3	2	2

The majority of mutations in both DCR and AGO genes were INDELs introduced by the 1st sgRNA, with only two cases per gene having a deletion caused by both sgRNAs. However, some mutants had INDELs of up to 550 bp missing. Biallelic mutants were either homozygotes or heterozygotes. In this system, Cas9 gene and sgRNA constructs are integrated in the genome under constitutive expression, while repair mechanisms of the cell are constantly functioning. Thus, we preferred to proceed by analyzing biallelic homozygote KO mutants, when possible.

In order to validate that the presence of WT alleles was not overlooked due to bias during PCR amplification of smaller pieces towards the longer WT, we examined the presence/absence of the deleted locus in genomic DNA of DCR KO mutants. Targeted gene sequences were amplified by using a primer hybridizing only within the deleted locus, upstream and downstream of the deleted locus, as well as combinations of hybridization outside-inside the deleted locus (Figure 42). A PCR of 40 cycles in genomic DNA template of WT and DCR KO mutants (1-8) was performed, in order to distinguish specific amplicons from enhanced nonspecific PCR products. Subclones from the same colony with a clear deletion pattern (primers hybridizing upstream and downstream targeted locus, clones 1, 2, 4, 6, 7, 8) indeed presented faint/weak amplifications of WT-sized bands in some cases (clones 1, 2, 6). Moreover, in clones with only one clear smaller band (clones 4, 7, 8) the deleted sequence could still get amplified in the genome (primers hybridizing inside the targeted locus). To clarify if this deleted sequence was integrated in another genomic locus or still present at the same location, amplification with Fw primer upstream and Rv inside the deleted locus was performed, as well as the reverse orientation. Indeed, in the latter case clones 4, 7, 8 demonstrated a lower level of amplification of non-specific products, suggesting the transposition of the deleted sequence in the genome. Clone 7 was considered as the best KO candidate and was used for further analysis (mutant dcr1 below).

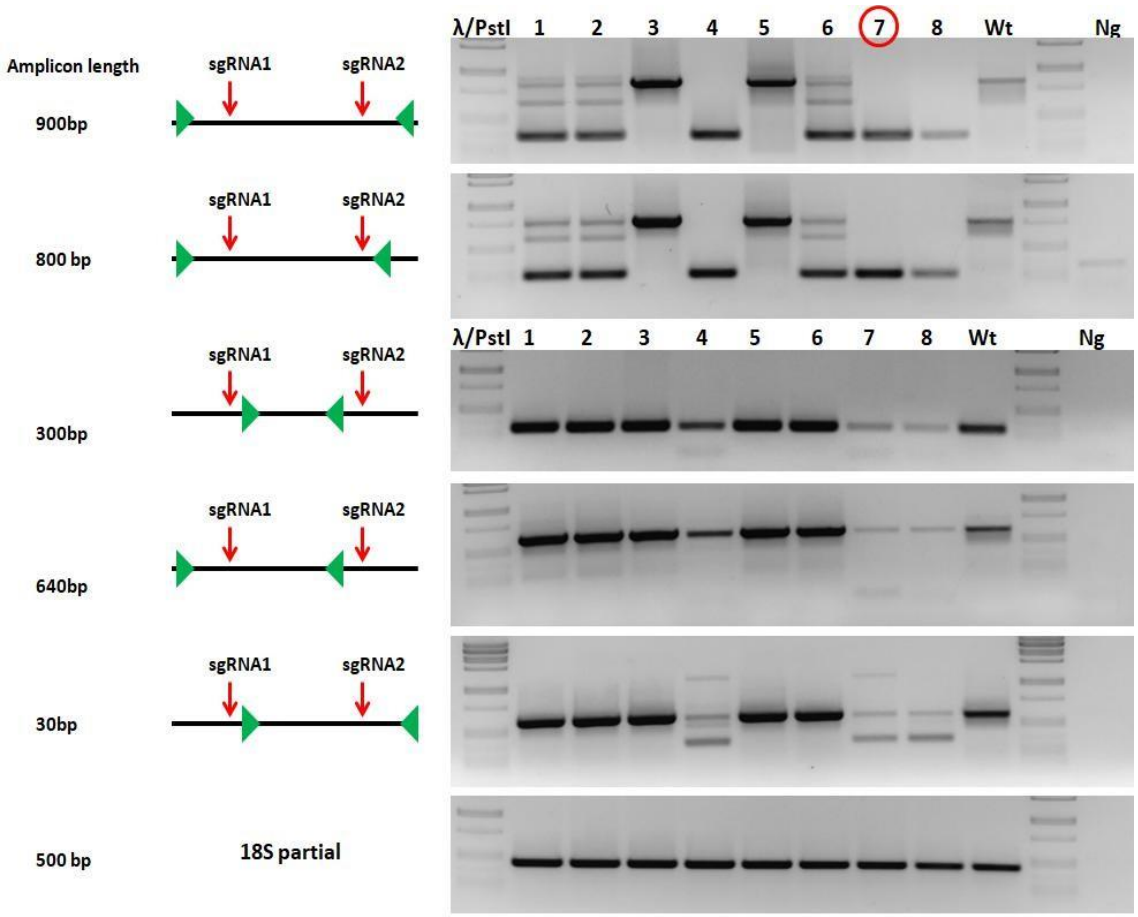

Figure 42. PCR amplification of CRISPR/Cas9 targeted Pt DICER sequence from WT and Pt DCR KO mutant genomic DNA. Subclones 1-8 come from the same colony. Primer hybridization sites are indicated with green arrows and the extent of deletion is indicated by the sgRNA target (red arrows). WT and WT-like clones 3 and 5 were used as negative controls and amplification of partial 18S sequence was used for quality and quantity control. Clone 7 was considered the best candidate for further analysis (in red circle).

All INDELs and deletions were introducing frameshift resulting in premature stop codon. Biallelic heterozygote mutants are considered to be generated from non-homologous end Joining (NHEJ) mechanism that repairs DNA double stranded brakes caused by Cas9 activity, while biallelic homozygote mutants are probably generated due to repairing mechanisms that use gene conversion. When NHEJ mechanism acts, the polymorphisms upstream and downstream of the breakage are conserved in each allele, but in the case of gene conversion only the polymorphisms of one allele are present around this locus, actually defining the expansion of the event.

After genomic DNA extraction of WT, **m4** and **m7** mutants (described below), a longer piece of Pt *DICER* gene (2Kb), starting from ATG, was amplified, cloned and sent for Sanger sequencing. This analysis validated the presence of the same INDEL/deletion in both alleles and revealed the expansion of gene conversion per case (Figure 43).

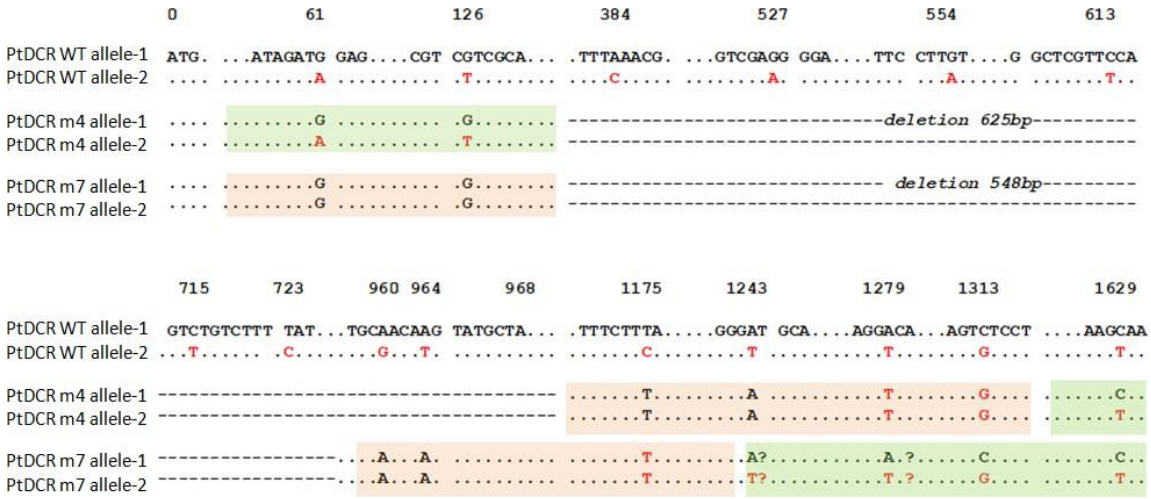

Figure 43. Alignment of partial PtDCR alleles in WT (DCR) and m4, m7 DCR KO mutants illustrating the SNPs (sinle nucleotide polymorphism) location while identical nucleotides were replaced by dots. SNPs corresponding to WT allele 1 are colored black and the SNPs of WT allele 2 are marked red. SNP retention after gene conversion is indicated with pink and the original SNPs per allele are indicated with green.

Finally, Pt m4 and m7 mutants were validated for the absence of WT contamination by more rounds of re-streaking and then PCR amplification of the targeted locus from genomic DNA and cDNA with HF polymerase. As demonstrated in Figure 44, amplicons from genomic DNA and cDNA are single bands without traces of contamination or background.

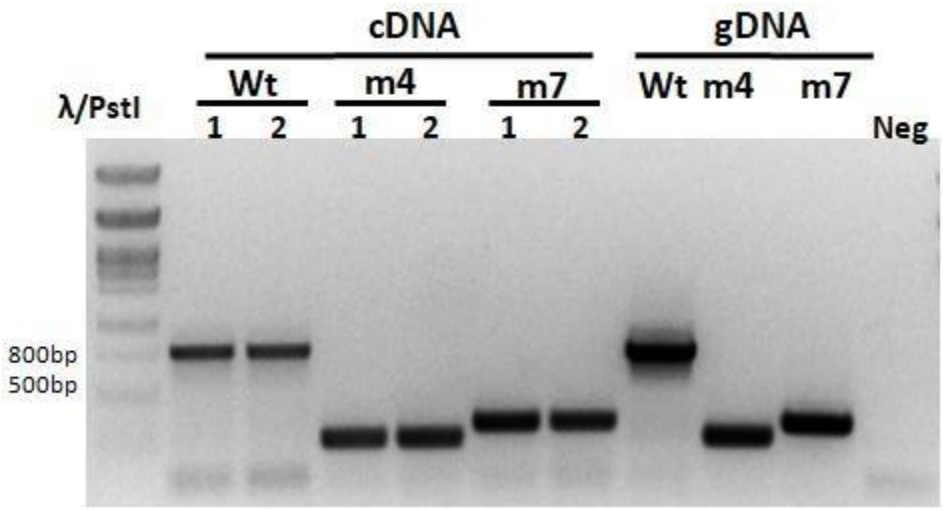

Figure 44. PCR amplification of the *DCR* targeted locus with HF polymerase from WT, m4 and m7 genomic DNA and cDNA. Two biological repeats of cDNA were tested.

The phenotypic analysis of Pt DCR KO mutants towards the functional characterization of *PtDCR* gene was prioritized and designated the scope of this thesis.

The following biallelic Pt DCR KO mutants were used for further analysis:

- **Pt m4:** Ptdcr_344Δ625 gene is homozygote and has a deletion of 635 bp, starting at position 344bp. Deletion was produced by two sgRNAs.
- **Pt m7:** Ptdcr_344Δ527 gene is homozygote and has a deletion of 527 bp, starting at position 344bp. INDEL was produced by probably the 1st sgRNA (the sequence of the 2nd sgRNA is almost intact).
- **Pt m8:** Ptdcr_345Δ53/352Δ1 gene is heterozygote and has a deletion of 53 bp, starting at position 3445bp in allele1 and a deletion of 1 bp at position 352 in allele2. In both cases INDELs were produced by the 1st sgRNA.
- **Pt m9:** Ptdcr_341Δ512 gene is homozygote and has a deletion of 512bp, starting at position 341bp. INDELs was produced by the 1st sgRNA.

3.5.4. Recapitulation of GUS RNAi system in PtDCR KOs

The importance of DCR and AGO in the initiation and maintenance of RNAi was assessed by comparing RNAi efficiency in WT versus DCR/AGO-KO lines mutagenized before and after recapitulation of the RNAi systems, respectively.

- *P. tricornutum* GUS transgenic lines and GUS RNAi silenced lines (generated by De Riso et al, 2009) were targeted in their *DCR* and *AGO* genes via CRISPR/Cas9 system, as described before. Biolistic co-transformation was performed with pKS_Blast vector, offering blasticidin resistance. Positive GUS or GUS/RNAi DCR/AGO KO mutants were screened and identified as described above.
- Generated *P. tricornutum* DCR KO and AGO KO lines were co-transformed with pKS_FcpB_GUS_At vector and pKS_Blast vector in order to integrate the GUS transgene. Then, vectors expressing GUS antisense or hairpin constructs and the Sh ble gene would be delivered by biolistic transformation and transformants would be selected on media containing phleomycin. GUS expressing cell lines were validated by PCR amplification of a 500bp GUS sequence from genomic DNA.

The efficiency of GUS RNAi was analyzed at protein level by histochemical assay (Figure 45). Results demonstrate that only one positive transformant (sample 17) out of sixteen PCR validated clones was expressing GUS, as it turned blue. Moreover, a GUS/RNAi strain had lost the RNA silencing capacity (sample 2). PCR amplification of the whole GUS gene from genomic DNA of the analyzed strains confirmed its presence only in sample 17 (data not shown). The most troubling part was that the GUS transgene could not get fragmented only due to its integration via biolistic transformation, as in some cases mutagenesis was performed onto validated GUS and GUS/RNAi lines that were actually used as controls in this analysis (samples 2 and 3). Thus, it appears that GUS transgene gets fragmented or lost after subsequent rounds of biolistic transformation.

Taken into consideration all the above, it was decided that this system did not provide the necessary stability for this type of experimentation and further analysis should not continue.

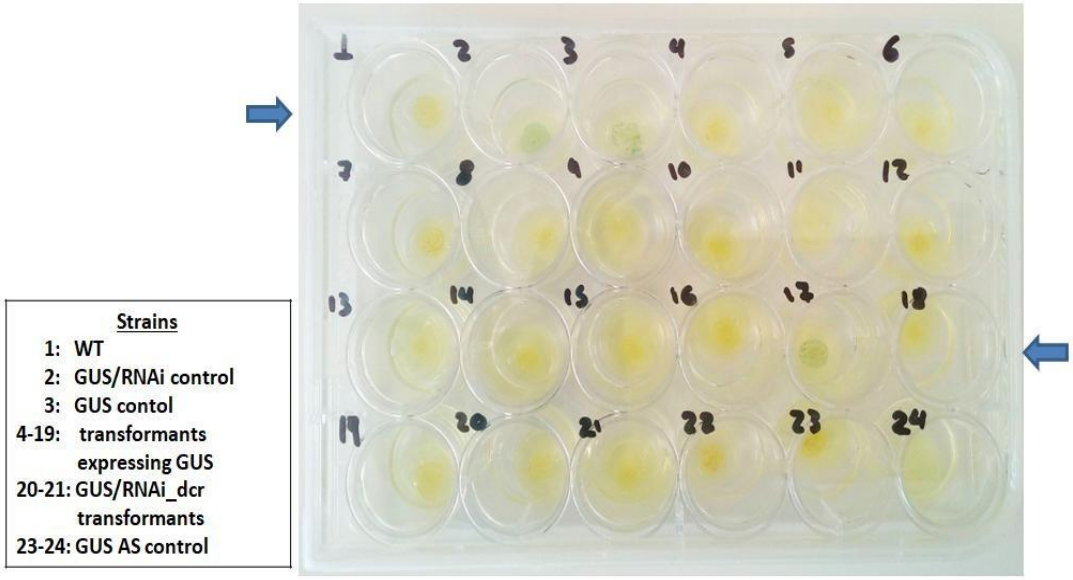

Figure 45. GUS histochemical assay of *P. tricornutum* WT, GUS, GUS/RNAi , DCR KO/AGO KO_GUS, GUS_DCR KO and GUS/RNAi_dcr lines. GUS expressing cells are stained blue.

3.5.5. Phenotypic characterization of PtDCR KO mutants

DCR-KO mutants in multicellular organisms usually exhibit obvious phenotypes. In animals, where DROSHA is also present, elimination of DCR induces problems in cell differentiation/maturation, while in plants elimination of DCL1 is lethal. In protists there is no information regarding the effect of DCR abrogation, except the reported inability of functional sRNAs to silence DCR and AGO genes in *Entamoeba hystolitica* (Pompey et al., 2014). The first diatom DCR KOs were generated in this study and since mutant lines were obtained, DCR does not seem to be vital in diatom growth and survival under normal conditions.

Growth tests under standard conditions, nutrient starvation and after induced UV damage in Pt WT and DCR KO lines were performed. Finally, a deep sequencing analysis of small and large RNAs from the aforementioned lines was performed for the exploration of alterations in gene expression and in small RNA populations.

3.5.4.1. Growth under optimal conditions

Triplicate cultures of *P. tricornutum* WT and DCR KO lines were grown in F/2 liquid media under optimal light and temperature conditions. Cell divisions per day in each *P. tricornutum* line do not indicate a different phenotype in DCR KO mutants (Table 16). Cell size and morphology was observed under light microscopy. No differences in comparison to WT were found.

Table 16. Cell divisions per day during log phase of *P. tricornutum* WT and DCR KO mutant lines.

***P. tricornutum* line**	**WT**	**m4**	**m7**	**m8**	**m9**
Cell divisions/day	**1.92**	**1.60**	**1.64**	**1.83**	**1.86**

3.5.4.2. Growth under stress conditions

Adaptation to NO3 Starvation

Nitrogen (N) metabolism in diatoms is vitally linked to photosynthesis and carbon (C) flux. In NO_3^--limited conditions the reduction of photosynthetic capacity, C-fixation, N-assimilation and suspension of growth is observed (Hockin et al., 2012; Bender et al., 2014; Alipanah et al., 2015), while proteins and nucleic acids are also affected (Mock and Kroon, 2002; Bertozzini et al., 2013; Mus et al., 2013). Diatoms are capable of rapid uptake and storage of NO_3 in vacuoles (RAVEN, 1987), managing to survive in absence of NO_3 for short periods of time. Nitrate starvation in diatoms has been well studied and enables the observation of diatom growth in limiting but not lethal conditions.

Triplicate cultures of *P. tricornutum* WT and DCR KO lines were grown under optimal light and temperature conditions in liquid media of F/2 (control) and F/2 lacking NO_3 nutrients. All lines were expected to get affected by NO_3 starvation but a phenotype presenting greater difficulty or delay in growth was anticipated in DCR KO mutants. Indeed, all *P. tricornutum* lines grew slower in comparison to their growth in F/2 medium (Figure 46). m4 and m8 mutants had the slower growth, while m7 and m9 mutants demonstrated an intermediate growth rate (Figure 47Figure 46).

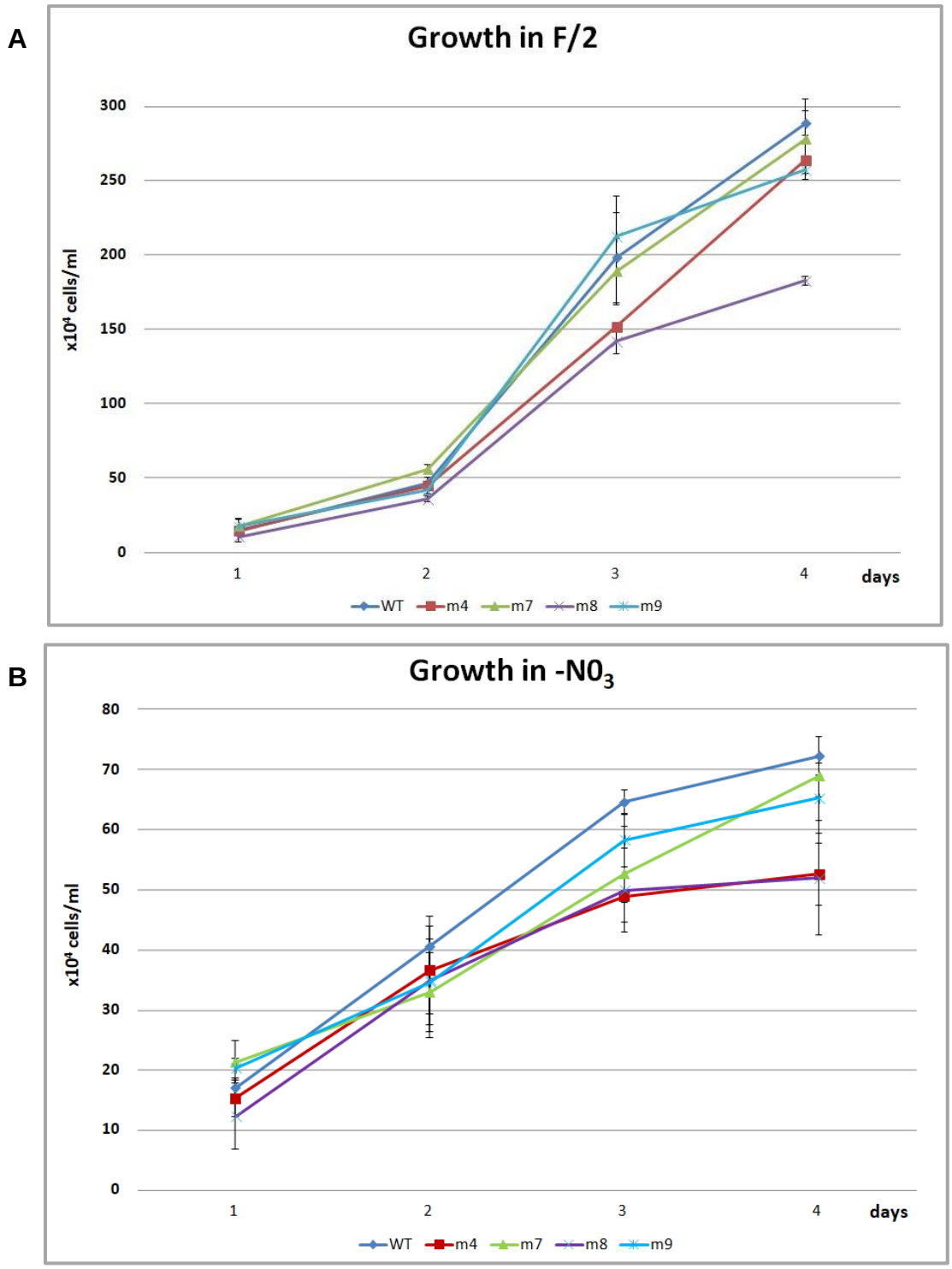

Figure 46. ***P. tricornutum*** **WT and DCR KO mutant lines grown in normal conditions (A) and under NO_3 starvation (B). Average cell counts per day were calculated from three replicate cultures. Standard deviation is denoted in error bars.**

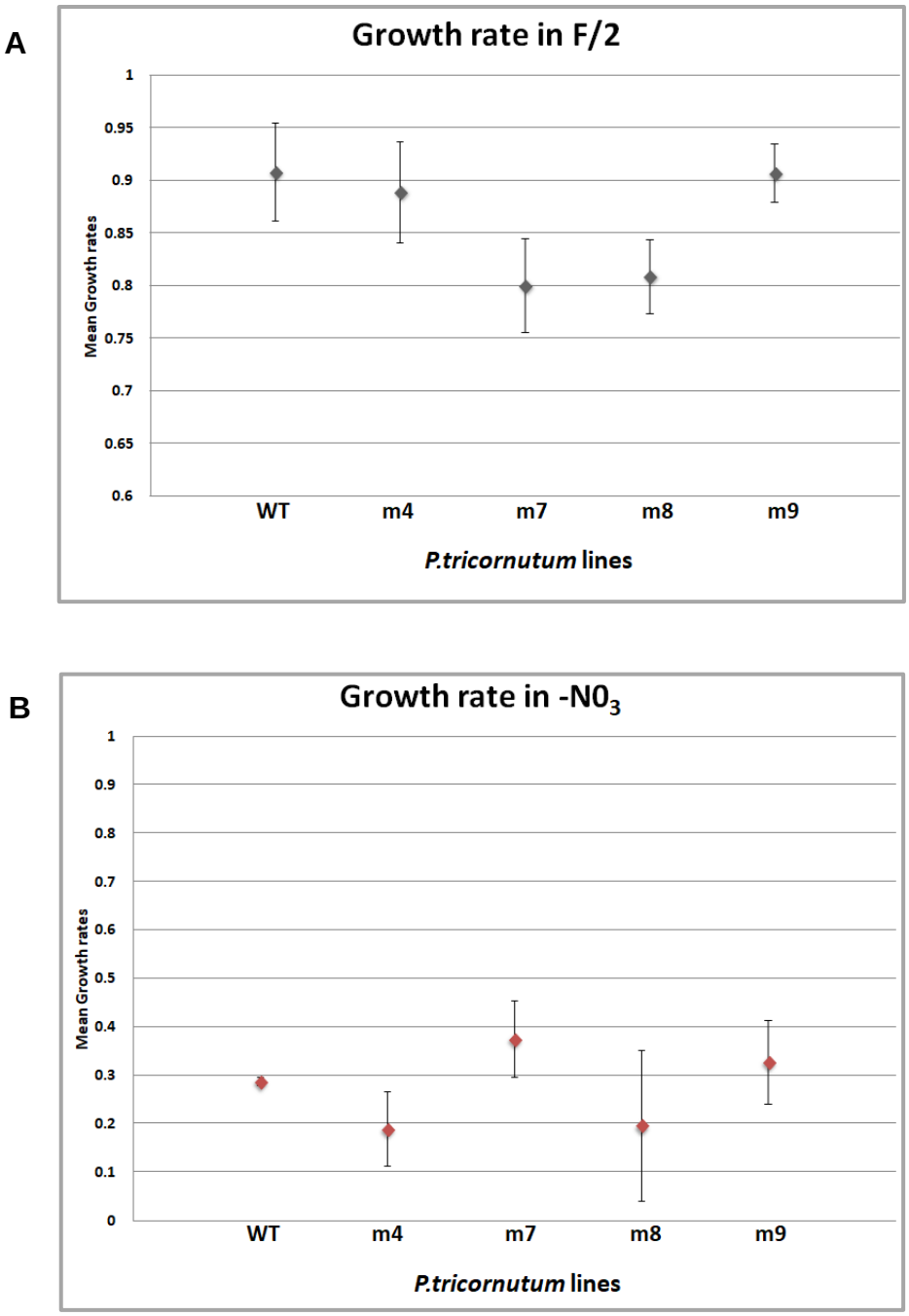

Figure 47. Average growth rates of *P. tricornutum* WT and DCR KO mutant lines in normal conditions (A) and under NO_3 starvation (B). Average growth rate was calculated from three replicate cultures. Standard deviation is denoted in error bars.

In conclusion, DCR mutants and especially m4 and m8, present a phenotype of slower growth in comparison to WT, when are subjected to NO_3 starvation.

Recovery after UV-induced damage

Ultraviolet (UV) radiation kills cells mostly by damaging their DNA (Pattison and Davies, 2006), but also affects other biological processes like photosynthesis (Tedetti and Sempéré, 2006; Häder et al., 2007; Williamson et al., 2014) while inducing the synthesis of fatty acids in photosynthetic microalgae (Skerratt et al., 1998). Based on the role of DCR protein in DNA repair mechanisms (Tang and Ren, 2012), that would be activated as a response to UV-mediated DNA damage, we decided to test the efficiency of recovery following exposure to UV radiation in the DCR KO mutants. A phenotype of delay or incapability to grow would be expected in the absence of a functional DCR.

P. tricornutum WT and DCR KO mutant cells were spotted in serial dilutions on F/2 agar plates, in duplicates. One plate was kept as control (no damage induced) and the second one was exposed to UV radiation. Then, cells in both plates were grown under optimal conditions for 10 days. A common phenotype presenting greater difficulty or delay in growth after UV-induced damage in comparison to WT was anticipated in all DCR KO mutants. Instead, only m4 and m8 mutants showed such phenotype (Figure 48).

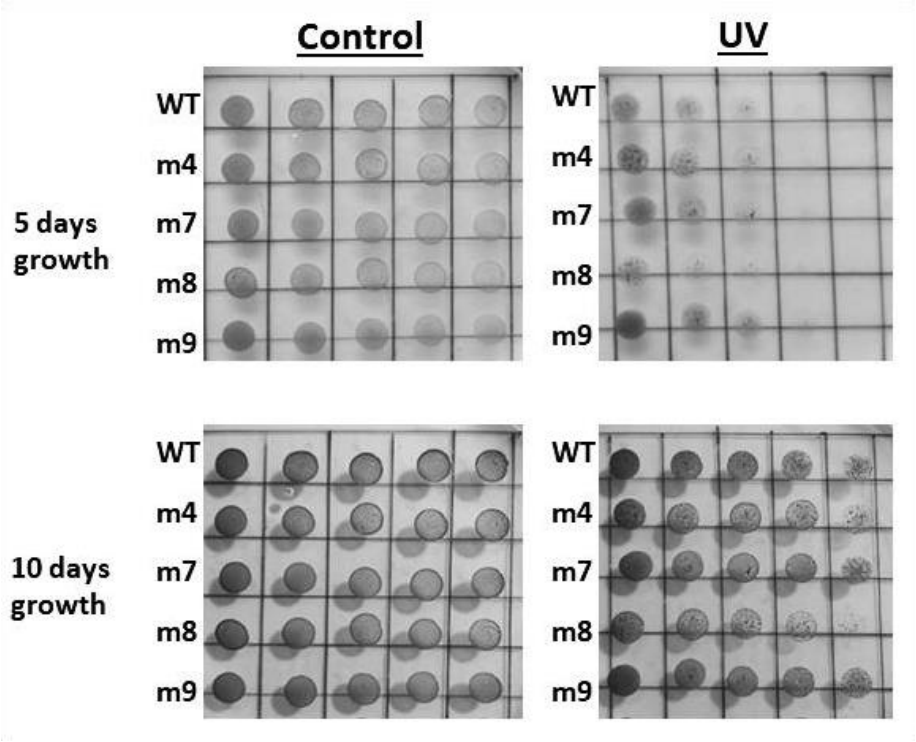

Figure 48.*P. tricornutum* WT and DCR KO mutants growth under normal conditions (control) and after exposure to UV radiation (UV).

In conclusion, recovery after UV-induced damage seems to be slightly compromised in two out of four mutants, showing a delay in growth more obvious after 10 days. However, these results are not convincing enough to support the presence of a strong phenotype in DCR KO mutants with compromised repair mechanisms.

3.4.3.3. Proteomic analysis in PtDCR-KO mutants

P. tricornutum WT, m4 and m7 mutant cells from cultures in log phase were collected for proteomic analysis. A preliminary proteomic analysis aiming to optimize the protocol for diatoms was conducted with pooled biological triplicates. Total protein extraction was analyzed at the IMBB Proteomic facility for their protein composition. Deregulation of DCR-repressed protein coding genes was expected in m4 and m7. In addition, proteomic analysis was expected to support the absence of DCR in m4 and m7 mutants.

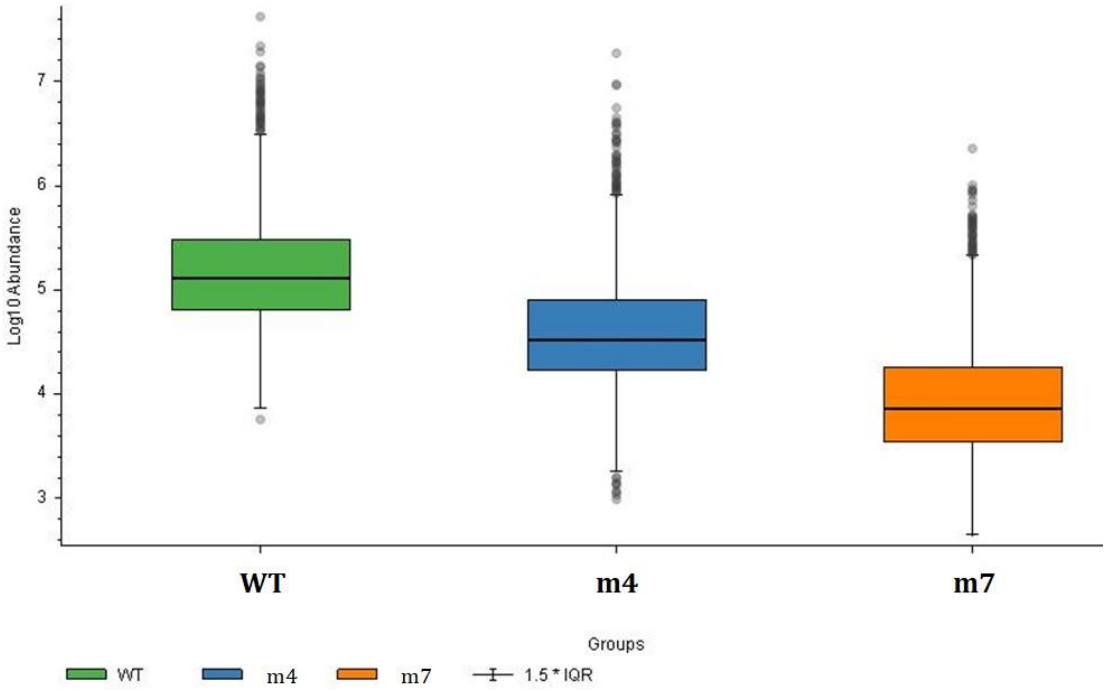

Figure 49. Quantification of total protein based on label-free quantitation from Proteome Discoverer.

Overall, more proteins were detected in WT than in DCR KO mutants suggesting a global transcriptional and/or translational decrease in m4 and m7 (**Figure 49**). Subsequently, peptides were identified and the portion of shared proteins between sample was analyzed (**Figure 50**). Protein diversity was larger in m7 than in WT and m4 (**Figure 51**). This result suggests that possible genomic rearrangements induced by biolistic transformation may have affected m7 global proteome. Pt DCR peptides however, were not detected in any line, suggesting that DCR is present in low amounts in WT. Further analysis aiming to detect DCR peptide hence should be carried out by preliminary protein size-specific fractions.

.

Label	Exclusive	Total
WT	142	764
m4	32	628
m7	16	488
WT, m4	162	584
m4, m7	12	434
WT, m7	38	460
WT, m4, m7	422	422

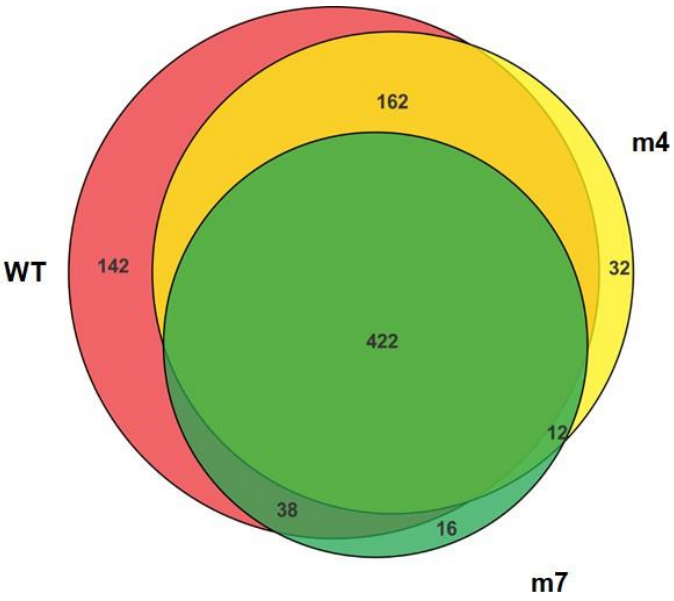

Figure 50. Specific and shared proteins in WT, m4 and m7 mutants.

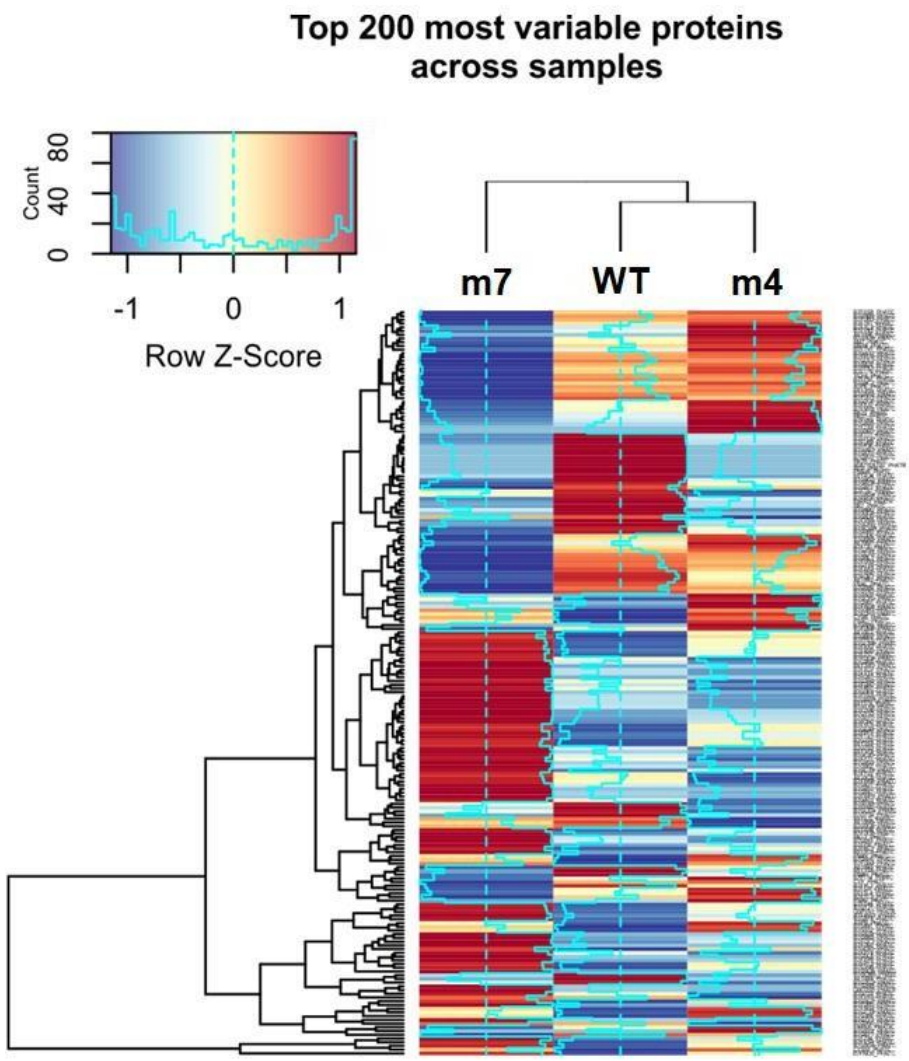

Figure 51. Heatmap of Top variable proteins (NSAF) across WT, m4 and m7 mutants underlying the high divergence of m7 mutant.

3.4.3.4. Analysis of DCR KO mRNA and small RNA transcriptomes

Total RNAs were extracted from *P. tricornutum* grown in liquid culture under optimal culture condition. Small RNA cDNA libraries were synthesized in our lab and cDNA libraries from messenger RNA were synthesized at the IMBB sequencing facility. 3prime Quant sequencing was performed for large RNAs, where only one 3' end fragment per transcript is generated, sequenced and counted.

Differential gene expression analysis in WT, m4 and m7 DCR KO mutants.

The differential gene expression between WT and DCR-deficient mutant line was expected to follow a downregulation of genes involved in central metabolic processes e.g. cell cycle, as a response to stress, and an upregulation of genes involved in stress responses, transcription and TE mobilization. 3prime Quant sequencing of mRNA cDNA libraries yielded 7-10 million reads across samples (Table 17). Preliminary data of gene expression in m7 mutant showed a non-canonical diversification that mirrors the proteomic analysis results described above. Therefore m7 mutant was excluded from further analysis.

Table 17. Yield of total reads from 3prime Quant sequencing of mRNA cDNA libraries (Q30=85.92).

Samples	Total reads (fastq) per technical repeat			
	1	2	3	4
WT	8148002	8029416	7737706	7354815
m4	10105384	7915315	8219441	7032100
m7	8639740	8799093	8704920	9346096

In m4 mutant 658 genes were differentially expressed in comparison to WT line, with 285 genes upregulated (UP) and 373 genes downregulated (DOWN). Surprisingly, the majority of the DOWN genes was localized in chromosome 18 (Figure 52).

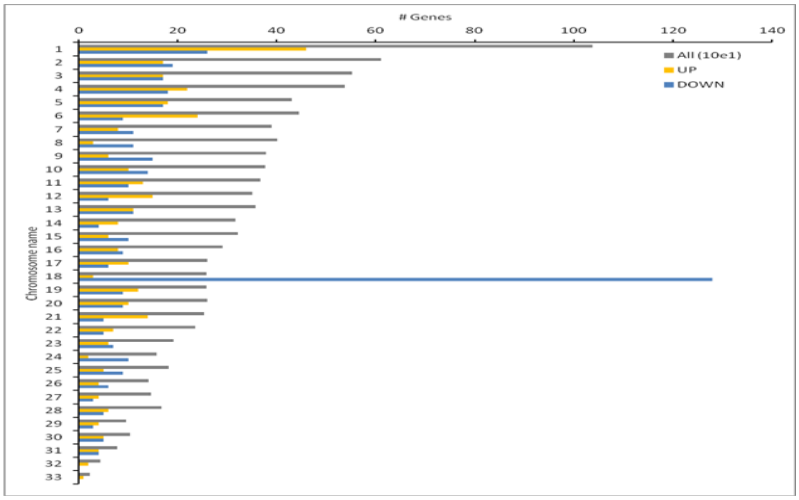

Figure 52. Chromosome localization of total (grey) and UP (yellow) and DOWN (blue) genes.

A large fraction of the differentially expressed genes, both UP and DOWN, lacked GO annotation. In some cases however, prediction of functional domains like CLADE (Clade-centered models), DAMA (Domain Annotation by a Multi-objective Approach) and PFAM (Protein Families) would give an indication about the function of each gene (Table 18). Genes were filtered by keeping those with Pvalue<0.05 and log2ChangeFold>1 (upregulated) or log2ChaneFold<-1 (downregulated), respectively.

Table 18. The 10 most upregulated (green box) and downregulated (red box) genes in m4 mutant in comparison to WT. Manually inserted biological functions are indicated with grey letters.

Phatr3	Chromosome	Biological Process	PFAM Domains	Functional domains (CLADE, DAMA)	log2FoldChange
2278	chr_14	Virus-associated	NA	Herpesvirus large structural phosphoprotein UL32	3.71
2350	chr_5	Growth arrest and DNA-damage-inducible proteins	NA	Growth arrest and DNA-damage-inducible proteins	3.47
47572	chr_14	NA	NA	Ribosome receptor lysine/proline rich region	3.40
49025	chr_20	protein phosphorylation	Protein kinase domain	Protein kinase domain	3.22
1910	chr_25	NA	NA	NADH-ubiquinone oxidoreductase ASHI subunit	2.95
46703	chr_11	Senescence	Senescence-associated protein	Senescence-associated protein	2.91
41302	chr_29	NA	NA	Origin recognition complex subunit 6 (ORC6)	2.86
55138	chr_26	NA	NA	TMEM154 protein family	2.76
48225	chr_16	NA	NA	Leucine Rich repeat;Leucine Rich Repeat	2.72
48225	chr_16	NA	NA	Leucine Rich repeat;Leucine Rich Repeat	2.72
29666	chr_18	NA	Domain of unknown function (DUF389)	Domain of unknown function (DUF389)	-6.62
48538	chr_18	NA	Domain of unknown function (DUF2NA2);VTC domain	VTC domain;Domain of unknown function (DUF202)	-4.18
39066	chr_18		S1 RNA binding domain	S1 RNA binding domain	-3.78
48655	chr_18	NA	NA	Protein of unknown function (DUF722)	-3.69
39046	chr_18	NA	NAD(P)-binding Rossmann-like domain	Flavin containing amine oxidoreductase;NAD(P)-binding Rossmann-like domain;	-3.19
33459	chr_3	DNA integration	Reverse transcriptase (RNA-dependent DNA polymerase)	Integrase core domain;RNase H;Reverse transcriptase (RNA-dependent DNA polymerase)	-3.00
48667	chr_18	regulation of transcription	HSF-type DNA-binding	HSF-type DNA-binding	-2.88
48392	chr_17	NA	NA	Lamin-B receptor of TUDOR domain	-2.81
39432	chr_19	NA	PhoD-like phosphatase	PhoD-like phosphatase;	-2.55
47347	chr_13	NA	NA	Bcl2-/adenovirus E1B nineteen kDa-interacting protein 2	-2.52

Many UP genes were predicted to be implicated in TE transposition, DNA integration, transcription, DNA repair as well as tRNA synthesis and modification-associated genes. Most of DOWN genes were predicted to be involved in metabolic processes and oxidation reduction processes. In both cases there were still a lot of genes with unpredicted biological function (Figure 53). The well-known suppression of TE expression and mobilization by the RNA silencing mechanism was expected to be shifted in the absence of DCR protein in m4 mutant.

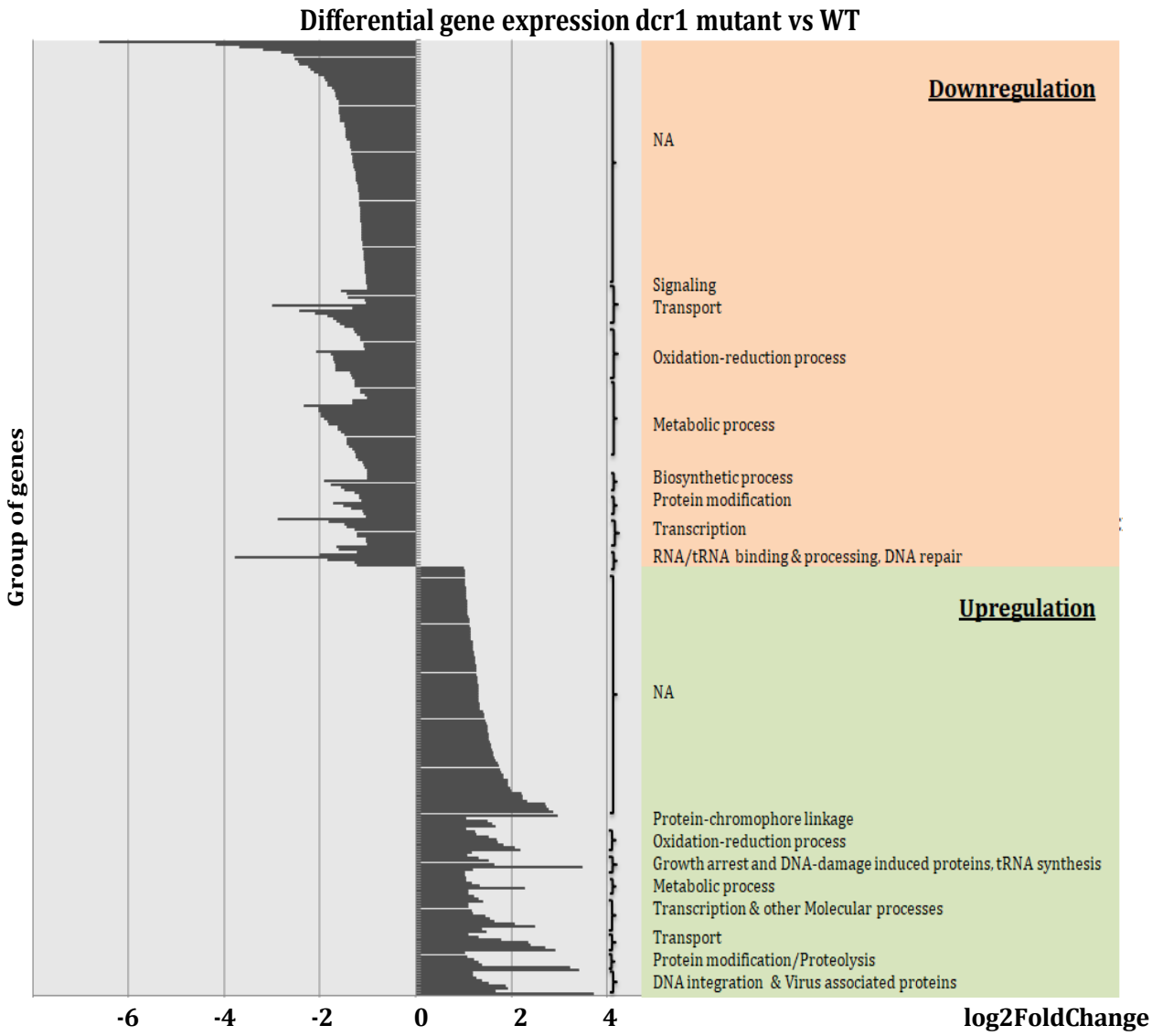

Figure 53. Differential gene expression between m4 mutant and WT line. Only genes with Pvalue<0.05 and log2FoldChange>1 or <-1 were analyzed. The basic functional groups of genes are indicated on the right.

Analysis of Small RNA transcriptomes in WT and m4 DCR KO mutant

Analysis of the sRNAs in WT and m4 would provide seminal information on the role of DCR in the production of the sRNAs populations described in Rogato et al. (2014) deriving from mono- or dual-strand, tRNAs, TE and gene loci. sRNAs deriving from dual strand TE/repeat loci were the first expected to drop in m4 mutant compared to WT. Sequencing of WT and dcr1 small RNA cDNA libraries aimed for five million reads. Two samples yield more reads and were subsequently normalized (Table 19).

Table 19. Yield of total reads from sequencing of small RNA cDNA libraries

	Total reads (fastq) per technical repeat			
Samples	**1**	**2**	**3**	**4**
WT	4399767	5357492	10952738	5634119
m4	23410480	5531566	6658374	4150199

Strikingly, the origin (loci/strand bias) of our sRNA dataset overlapped with the data reported in Rogato et al (2014) despite the different protocol used for sRNA library preparation and sequencing instrument, validating our experimental procedure. In this study, further analysis was focused on the differences of sRNA populations between the WT line and DCR deficient mutant m4.

Global analysis of size distribution

DCRs are expected to generate discrete size range of sRNAs. In plants and animals DCR-dependent sRNAs are usually 21-26nt long, but among unicellular organisms species specific DCR-like proteins may produce different sizes of sRNAs. In *P. tricornutum* the majority of sRNAs have 25-30 nt length and map to repetitive sequences. Here, global analysis of sRNA size distribution revealed a shift towards larger size in m4 mutant from 25-30 to 33-40 nt (Figure 54).

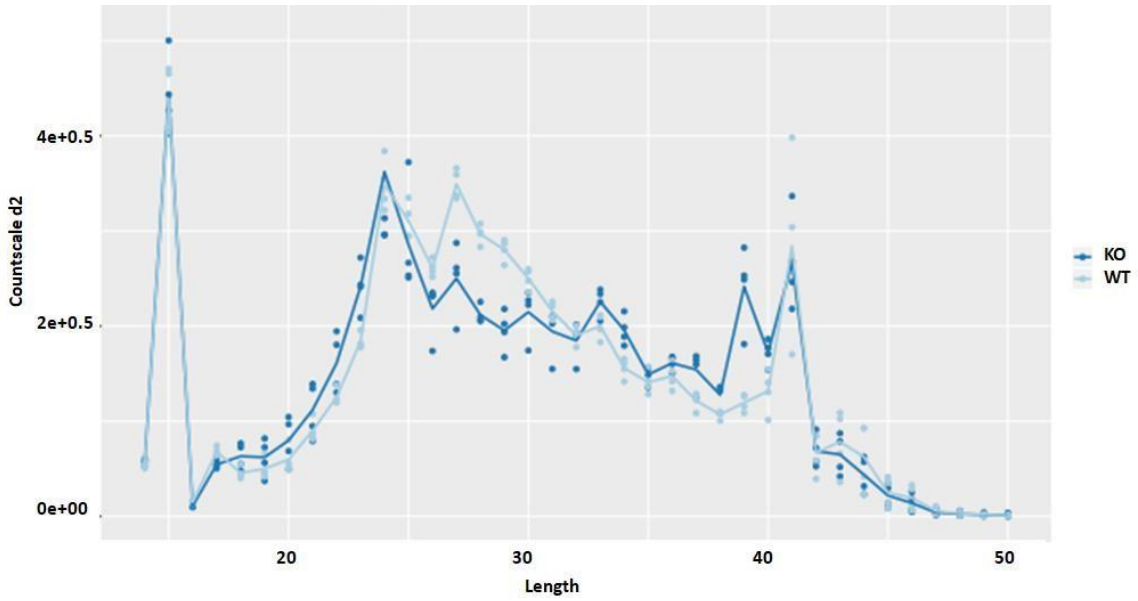

Figure 54. Global length distribution of fragments illustrates the overall shift from smaller to larger sRNA in m4. Y axis corresponds to the number of mapped fragments and x axis to the sRNA length.

By focusing on the length distribution of fragments per chromosome, some cases with high coverage but no change in fragment length were identified, like chr11 and chr21 (Figure 55). These represent mostly sRNAs with single strand specific peaks mapping to intergenic regions and nuclear RNA loci that are considered to be DCR-independent (described below in functional groups).

A decrease in small fragments that could be DCR-dependent or a swift towards longer fragments in m4 mutant is present in many chromosomes (Figure 55). In most cases a decrease in abundance is observed while only two chromosomes, chr3 and chr22 show a shift toward longer fragments.

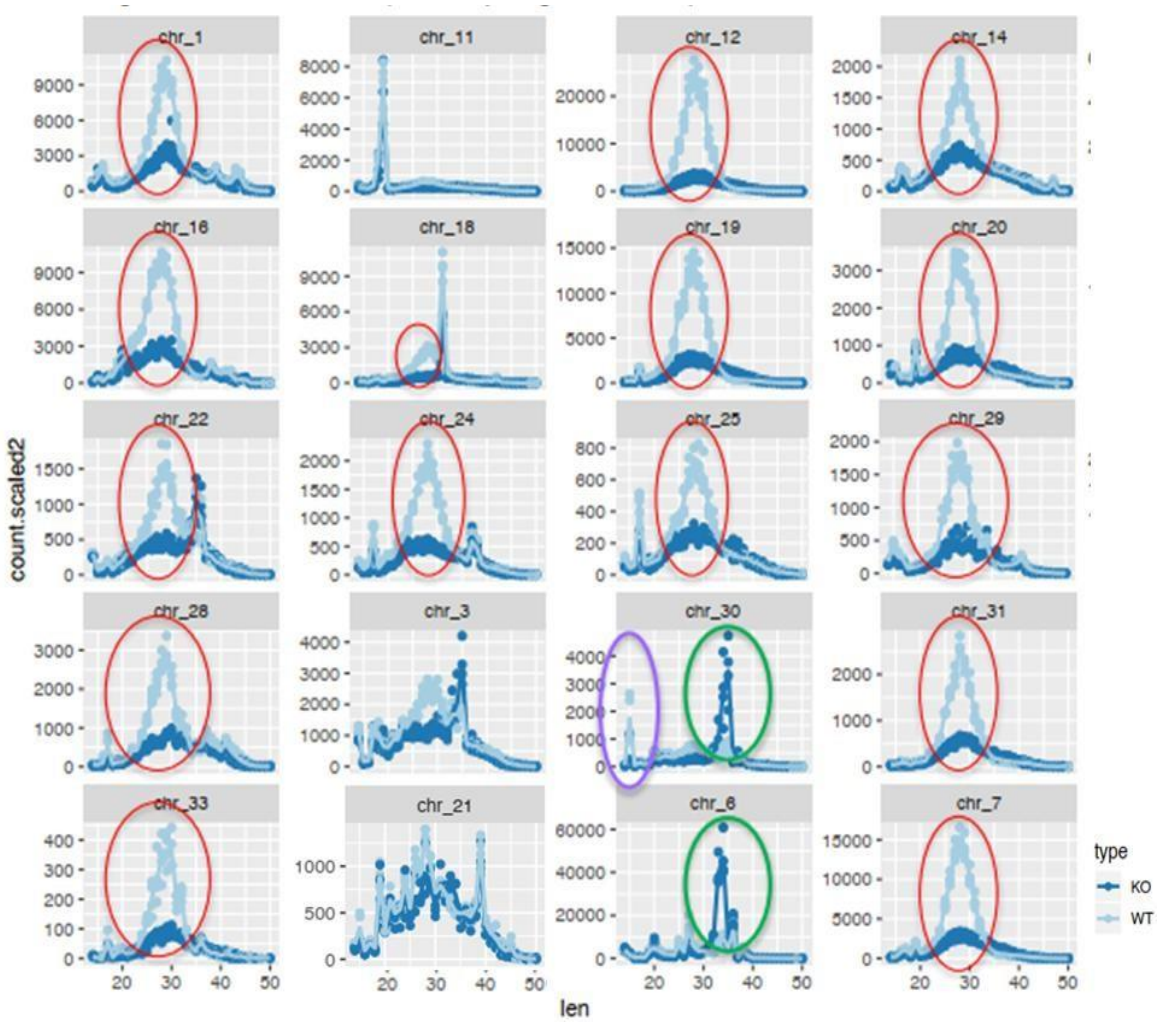

Figure 55. sRNAs length and abundance at each chromosome. Cases with high coverage but no change in fragment length in m4 mutant (chr11 and 21), cases with decrease of 25-23nt fragments in m4 mutant (red cycle), cases with increase of 35nt fragment in m4 mutant (green cycle) and a swift from 28nt towards 32-35nt in dcr1 mutant at chromosome 3 are demonstrated. Peak in purple cycle at chromosome 30 is an artifact.

Functional categories

Looking more into details, the genomic loci with mapped sRNAs were analyzed separately, based on their functional characterization. The functional categories revealed which small RNA producing loci were the most affected.

a) <u>TEs</u>

The largest effect was observed in Transposable Element (TE) regions where the abundance of specific size 25-31nt small RNAs was dropped dramatically in m4 mutant. These sRNAs are mapped in both strands of TE loci and are DCR-dependent.

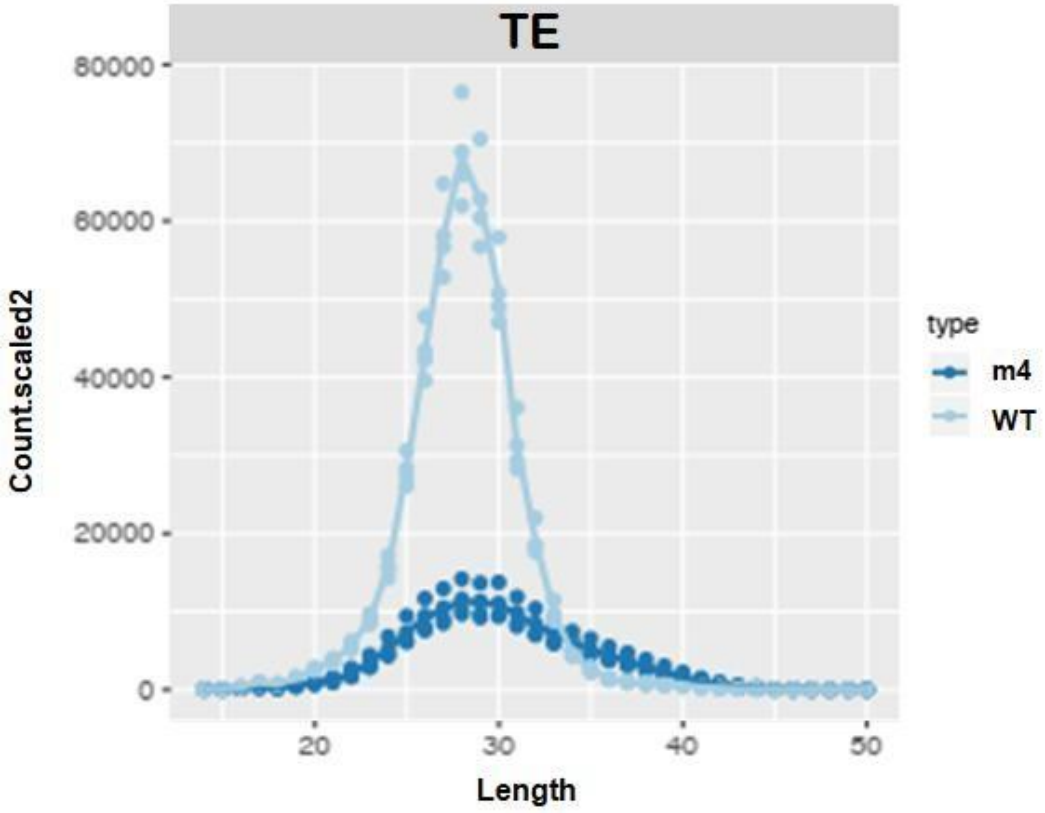

Figure 56. TE loci demonstrate a major decrease of 25-31nt sRNA coverage.

b) Intergenic regions and ncRNAs

sRNAs with single strand specific peaks mapping to intergenic regions and ncRNA loci were equally abundant in m4 and WT mutant and are considered to be DCR-independent (Figure 57).

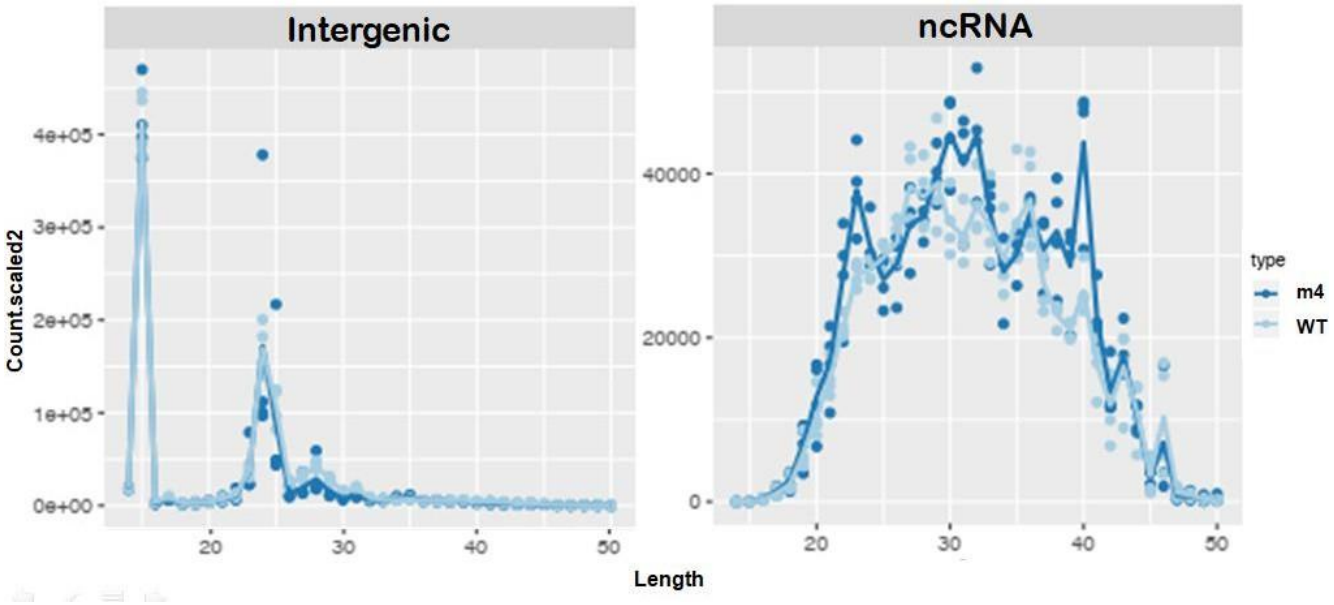

Figure 57. Intergenic regions and ncRNA loci demonstrate a non-affected coverage of sRNAs.

c) tRNAs

While tRNA-derived sRNA abundance did not seem highly affected, they shifted from small fragments of 25-28nt towards larger fragments of 33-35nt, with the later showing a high coverage in the m4 mutant compared to WT line (Figure 58). The peak of 33-35nt in tRNAs is also in agreement with fragments experimentally detected in WT line by Northern blot analysis in Rogato et al. (2014). Fragments mapped at the 3' end of tRNAs were manually validated for the presence of CCA trinucleotide.

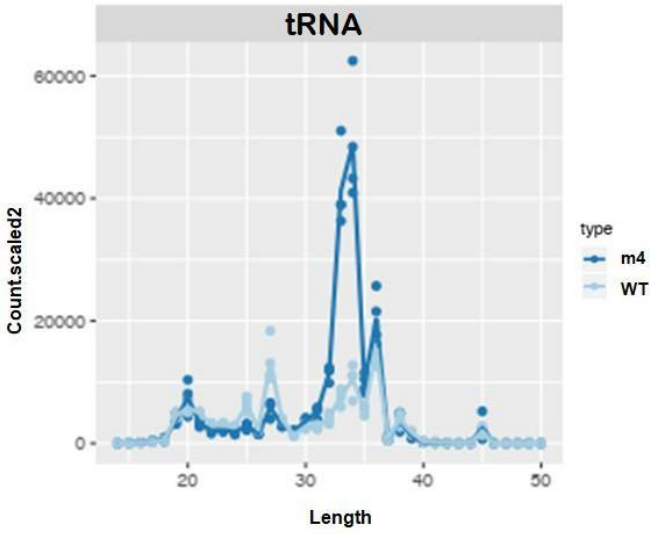

Figure 58. sRNA mapped to tRNA loci present a shift from 25-28nt towards 33-35nt fragments, with the latter demonstrating a higher abundance in m4 mutant.

A more detailed analysis, localizing the tRNA-derived fragments (tRFs) on matured tRNA, revealed the shift towards longer tRNA fragments in m4 mutant from the same tRNA present in WT (Figure 59). Interestingly the fragments seem to be cut mostly from an anchor on their 3‘ end.

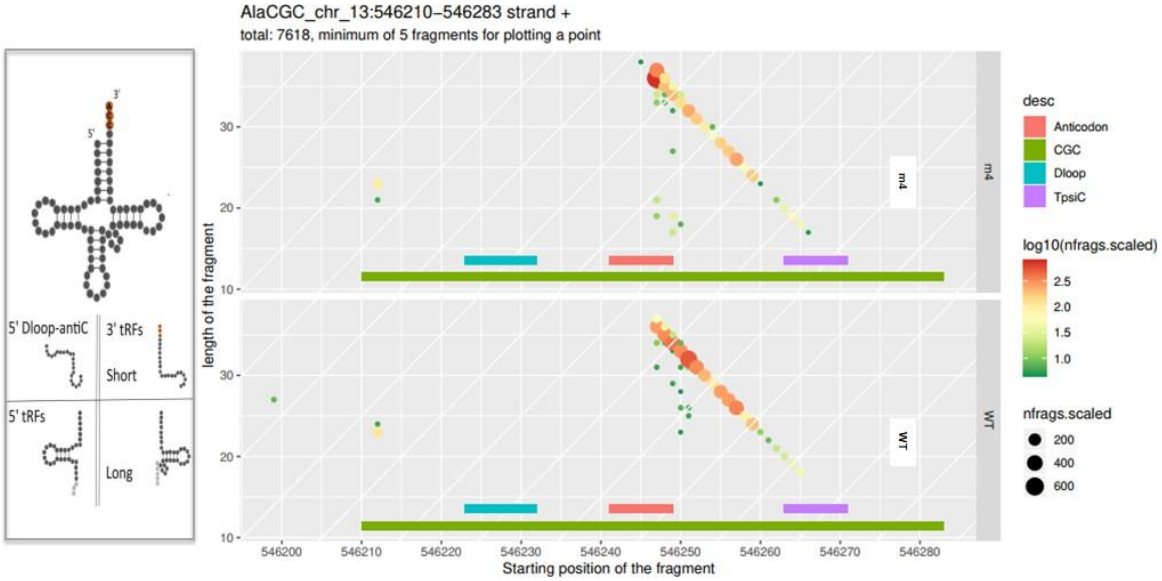

Figure 59. Plotted tRFs based on their length and their localization on tRNA AlaCGC. The region of D-loop, anticodon and the TpsC loop are indicated in colored boxes. Size and color of circles represent the abundance of specific-sized fragments. Illustration of a tRNA and its derived fragments on the left.

TE sRNA and mRNA correlation

sRNA coverage in TEs was dropped significantly in m4 mutant. Analysis was focused on 119 autonomous TEs and the vast majority demonstrated a 2fold decrease in sRNA coverage (Figure 60).

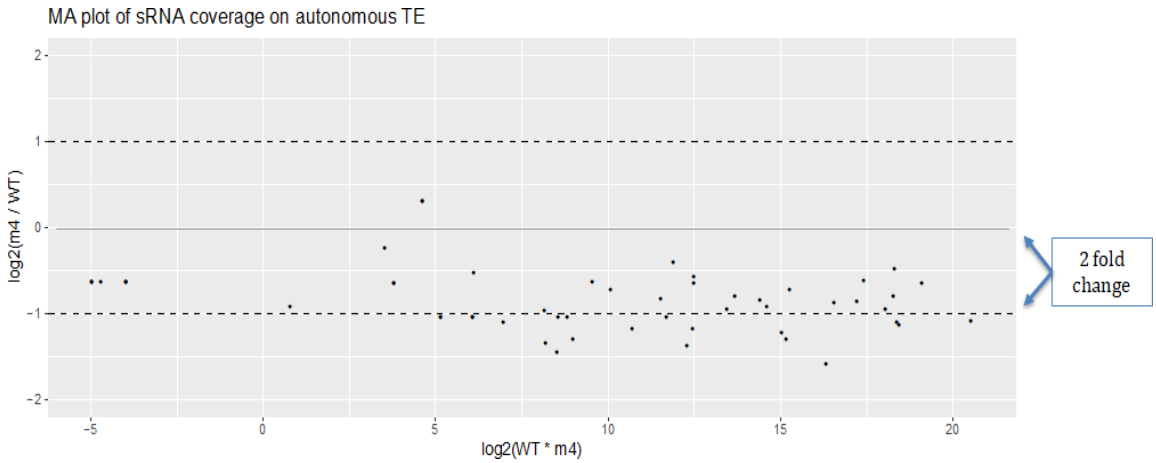

Figure 60. MA plot illustrates the consistent 2fold decrease of sRNA coverage on most TEs of m4 mutant.

Autonomous TEs are transposons still carrying their Long Terminal Repeat; thus they have the potential to translocate and so should be under stringent surveillance. Since small RNA coverage is expected to be related to RNA silencing, the next step was to investigate the differential expression of genes located in these autonomous TE regions. Indeed, autonomous TEs with lower coverage of small RNAs were generally upregulated demonstrationg a 2fold or 1,5foldChange (Figure 61).

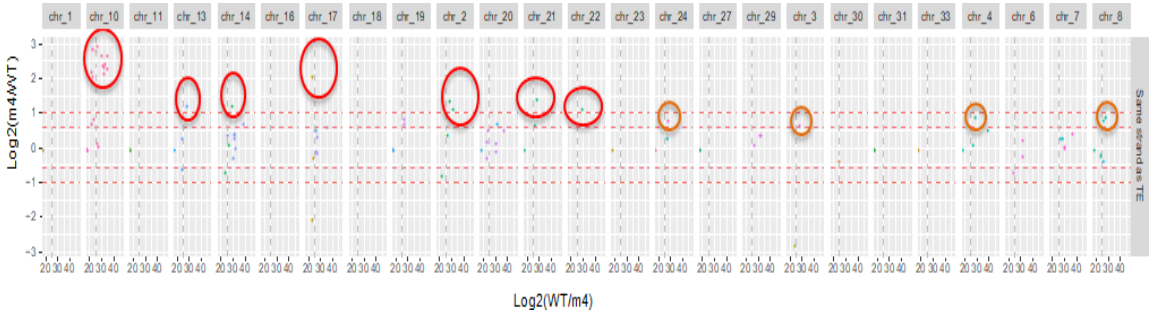

Figure 61. MA plot of mRNA probes mapped on autonomous TEs per chromosome in m4 mutant, illustrates TE upregulation. Inner red dotted lines represent 1fold change and outer red dotted lines a 2fold change. TEs exhibiting 2foldChange (red cycles) and 1,5foldChange (orange circles) are indicated.

The correlation between lower coverage of small RNAs and upregulation of TEs can be visualized in IGV program, where all tracks of small RNA and large RNA from this study but also available gene models, TEs, highly methylated regions and small RNA tracks from previous works (Maumus et al., 2009; Veluchamy et al., 2014; Rogato et al., 2014) can be mapped on *P. tricornutum* genome (Phatr2) and visualised (Figure 62).

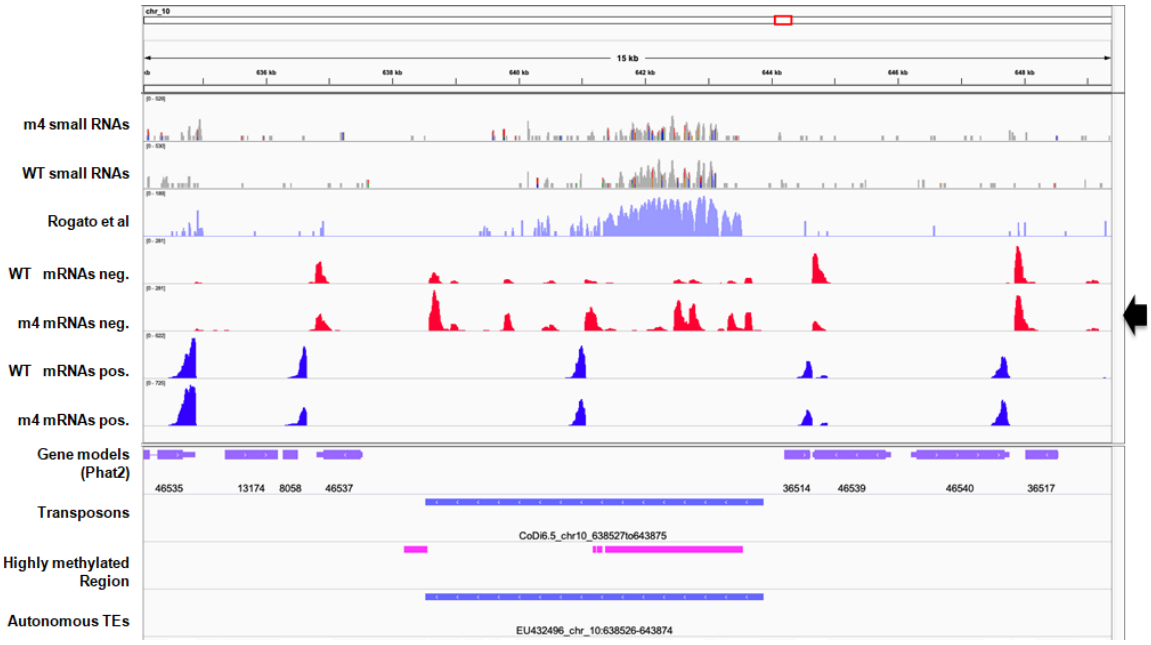

Figure 62. Example of a TE with decreased sRNA and concomitant increase in mRNA abundance (arrow) in m4 mutant compared to WT. Tracks (from top to bottom) of sRNAs analyzed in this study and in Rogato et al. (2014), large RNAs negative and positive strands analyzed in this study, Gene models from Phath2 annotation, transposons, Highly methylated regions and autonomous TEs were loaded and visualized in IGV program.

In the representative example illustrated above, an autonomous TE (CoDi6.5) appears to be highly methylated and covered with small RNAs in WT line leading to its suppression. In m4 mutant line the sRNA coverage in this region is decreased and the upregulation of TE expression is reflected by its increased mRNA level. Overall, these data suggest a role for DCR in the production of sRNA targetting and repressing TEs.

4. Discussion

Although RNAi is well conserved in higher eukaryotic organisms, unicellular eukaryotes, and among them photosynthetic microalgae, being evolutionary distant to plants, animals and fungi may exhibit significant differences in key RNA silencing genes and pathways. Subsequently, small RNA diversity and specialized RNA silencing mechanisms have evolved in these organisms (Cerutti and Casas-Mollano, 2006; Cerutti et al., 2011). The presence of a *PtDCR*, *PtAGO* and *PtRDR* was initially proposed by the seminal work of De Riso et al. (2009) revealing the existence of a functional RNA silencing mechanism in *P. tricornutum*. In this study, extensive *in silico* analysis of a recently re-annotated version of *P. tricornutum* genome confirmed the presence of a single *PtDCR*, *PtAGO* and *PtRDR* putative coding genes in this diatom. This is in contrast to most multicellular eukaryotes, like plants and animals which present multiple DCR, AGO and RDR paralogues arising from gene duplication events followed by subfunctionalization (e.g. DCR1-4 in *A. thaliana*, and DCR and DROSHA in animals). Effective RNAi, however, was reported in eukaryotes presenting a single DCR, AGO and RDR genes, like *Chlorela variabilis*, *Ectocarpus siliculosus* (Cerutti et al., 2011), *Volvox carteri* (Dueck et al., 2016) and *Entamoeba hystolitica* (Pompey et al., 2015a).

A peculiar RNAi machinery in Diatoms

This study revealed the presence of two different sets of RNAi key-genes in diatoms. Phylogenetic analysis of DCR RNase III and AGO PIWI domains resulted in the formation of two distinct groups (type A and type B) for both proteins. Group A-DCRs present domain architecture and catalytic sites generally found in multicellular DCRs. Group A-AGOs are also well conserved resembling known eukaryotic AGOs in terms of domain composition and catalytic sites. Conversely, group B-DCRs, such as PtDCR are non-canonical DCRs, in which the PAZ and second RNaseIII domains have diverged. Finally group B AGOs present well conserved PAZ and PIWI domains, although residues critical for PIWI splicing activity are absent. Despite their relative good conservation with known eukaryotic DCRs and AGOs, diatom type A DCRs and AGOs are clustered

apart from most known eukaryotic homologues. On the other hand, type B DCR is clearly divergent but type B AGOs, with a probably non-slicer PIWI domain, cluster closer to known AGOs indicating evolutionary forces for its differential function and conservation. The fact that diatom type A-DCRs are positioned in such phylogenetic distance from other known DCRs, while other stramenopile DCRs cluster inside the group of typical DCRs, highlights the rapid evolution and divergence of these proteins in diatoms. This means that attempts to investigate a putative relation to plant or animal DCRs would likely be unfruitful.

All diatoms present at least one type of DCR and AGO protein underlining the importance and conservation of an RNAi machinery in these organisms. DCRs and AGOs of the same group are generally present together, indicating their specialized co-function. Probably one type A DCR or AGO cannot substitute for the type B-DCR/AGO and *vice versa*, which suggest there are two distinct RNAi pathways at the protein level. Interestingly, there are diatoms having both type A and B proteins, indicating an ancestral origin of both sets of RNAi key-genes in diatoms and the subsequent loss of type B proteins in the most recently evolved raphid pennates.

Moreover, at least one RDR protein, classified as type γ RDR (Zong et al. 2009), seems to be an indispensable partner of type B DCRs and AGOs, while missing in diatoms that possess only type A proteins. *A. thaliana* RDR1, RDR2 and RDR6 belong to type α RDRs involved in the antiviral response by converting viral-derived ss RNAs into ds RNAs (Leibman et al., 2018), the biogenesis of repeat-associated siRNAs (Matzke et al., 2009) and the production of phased short interfering RNAs (siRNAs) (Bouché et al., 2006; Howell et al., 2007). *A. thaliana* RDR3, RDR4 and RDR5 belong to type γ RDRs of unknown function (Willmann et al., 2011).

A more thorough phylogenetic investigation of each RNAseIII domain separately may shed some light on the origin of diatom DCRs. Preliminary analysis indicates that DCR-A RNaseIIIa and RNaseIIIb and DCR-B RNaseIIIa domains are more closely related between each other than with the DCR-B RNase IIIb domain. Hence, DCR-B seems specific to diatoms and may have evolved either from the duplication of one DCR-A RNaseIII domain followed by its fusion with an RNAseIII-like domain of unknown origin, or by duplication of two DCR-A RNaseIII domains with the second one being

highly diversified for functional specialization. Among other organisms evolutionary closer to diatoms, like Phytophthora species belonging to Stamenopiles, the only proteins resembling the diatom type B DCR domain architecture was Phytophthora DCR2. However, Phytophtora DCR2 is clustering closely to DROSHA proteins. Regarding the overall resemblance of DCR-B to DROSHA domain architecture, the phylogenetic analysis does not support a higher similarity between these proteins. Morover, recent studies on DROSHA structure suggest that although all RNaseIII classes may have evolved from ClassI bacterial RNaseIIIs, DROSHAs have most likely emerged from class III (DICER) in an early metazoan ancestor, allowing the generation of animal microRNAs (Kwon et al., 2016a).

It is worth mentioning that besides DCR, the only other RNaseIII domain-containing protein we found in diatoms was miniIII, which was clearly clustered separately with other plant miniIII. Bacterial miniIII have been shown to present dsRNA substrate specificity (Redko et al., 2008; Głów et al., 2016). In plants, miniIII have been shown to participate in rRNA maturation and intron recycling in the chloroplast (Hotto et al 2015).

Strikingly, both model diatoms *P. tricornutum* and *T.pseudonana,* from which derives most of our current knowledge about RNAi in diatoms, represent exceptions among their taxonomical groups regarding their RNAi key-genes repertoire. This finding highlights the need to expand the study of RNAi in more diatom species presenting either type A-DCR and type A-AGO, or DCR and AGO of both A and B types. Comparative analysis of their epigenome, sRNA transcriptome and TE landscape will shed new light on the evolution and importance of the RNAi machinery to their acclamatory responses and adaptation to contrasted environments.

Alternative Splicing in *P. tricornutum*

Cloning and sequencing of PtDCR and PtAGO cDNA revealed the presence of alternative spliceoforms. PtDCR and PtAGO alternative spliceoforms may encode non-functional proteins as they present an altered reading frame introducing a premature stop codon upstream of functional domains. Surprisingly, the R-PCR experiment indicated that both expected and alternative DCR spliceoforms were present at comparable levels.

This finding suggests that PtDCR alternative spliceoforms may be functional and indicates a possible regulatory mechanism of these proteins production or the generation of alternative functional peptides.

A recent publication (Rastogi et al., 2018) unveiled the prevalence of alternative splicing in *P. tricornutum.* Genes in eukaryotic protists, including diatoms, present few and small introns (Armbrust et al., 2004; Stajich et al., 2007; Mock et al., 2017). This pattern of intron size and density has been proposed to mirror genome size, or alternatively to enhance transcriptional efficiency and splicing accuracy in metabolically fluctuating environments (Zhang and Edwards, 2012). *P. tricornutum* has ~53% intron-less genes and ~33% have one intron. In this study, the presence of 2 introns in *PtDCR* and 13 introns in *PtAGO* was validated. While intron-retention (IR) is the main alternative splicing code in plants and unicellular eukaryotes and exon-skipping (ES) is prominent within metazoans (McGuire et al., 2008), both types of alternative splicing are equally frequent in this diatom. IR and ES were observed in 24% and 20% of *P. tricornutum* genes, respectively (Rastogi et al., 2018). Here, Pt *DCRb* represented an IR spliceoform of the second intron and *PtDCRc* could represent a partially ES spliceform skipping a few bp of the third exon, (or represent a second intron donor site). *PtAGOb*, however, represented a transcript with alternative start of transcription, since both first and partially second exon were skipped. More PtAGO spliceoforms were identified in this study, but were not further investigated since their reading frames would not encode any functional protein domain.

Another interesting outcome of Rastogi et al. (2018) is that genes in *P. tricornutum* that can undergo IR are more highly expressed than genes that do not show alternative splicing, which is in contrast to intron-retention in mammals that down-regulates the genes that are physiologically less relevant (Braunschweig et al., 2014). They show that under nitrogen starvation, the genes that are involved in this stress response and they have an upregulated expression also express more IR spliceoforms (Rastogi et al., 2018). Moreover, in *C. elegans* it has been shown an alternative function of a truncated DCR reduced to its C-terminus and containing only a single RNase III domain and dsRBD. This shortened DCR, product of the pre-apoptotic caspase CED-1 activity, can bind to DNA where it cleaves or promotes cleavage of one strand inducing chromosome

fragmentation (Nakagawa et al., 2010). Although this truncated DCR does not result from alternative splicing, it is tempting to speculate possible alternative functions of a similarly domain-composed PtDCRb-c protein. *In vitro* studies on the enzymatic activity of DCR alternative spliceoforms and the identification of their subcellular localization by confocal microscopy may reveal their eventual function(s).

Investigation of *PtDCRb-c* spliceforms and *PtAGOb* function in heterologous expression systems

The functional roles of the alternative spliceoform *PtDCRb* and *PtAGOb* were investigated by heterologous expression in *S. cerevisiae. S. cereviciae* is a unicellular eukaryote, lacking RNAi, but being able to sustain an introduced RNA silencing mechanism based on *S. castellii* (yeast) DCR and AGO transgenic expression (Drinnenberg et al., 2009), or the human DCR, AGO and TRBP transgenic expression (Suk et al., 2011). Since at the time, the transformation procedure for *P. tricornutum* was not accessible, the use of the unicellular, experimentally validated, yeast heterologous expression system seemed ideal. However, the yeast system proved to be unstable by either losing integrated constructs after new rounds of transformation, or by losing RNAi capacity in positive clones. These observations are in agreement with the same stability problems encountered in other research groups (personal communication). Attempts to show DCR activity by hairpin-induced silencing of a homologous GFP transgene were unsuccessful. PtDCRb activity tested in combination with PtAGOb and ScAGO, did not result in the production of GFP sRNAs. These results suggest that either *PtDCRb* and *PtAGOb* a) are non-functional spliceoforms, or b) achieve other functions than RNA cleavage, or c) need co-factors absent in *S. cerevisiae*., or d) *S. cerevisiae* is an incompatible system to sustain diatom RNA silencing.

In 2015 Pompey and coworkers reported the DCR-like activity of the single RNaseIII domain-containing DCR-like protein from *E. hystolytica* (*EhRNS*) in *S. cerevisiae* (Pompey et al., 2015a). In the same study, however, co-expression of EhRNS and *E. hystolytica AGO* (*EhAGO*) failed to trigger silencing. Silencing was effective only when *EhRNS* was co-expressed with *S. castellii* functional AGO (ScAGO). Yet, GFP silencing was incomplete and GFP derived sRNAs were found to be between 50-60nt.

Since ScAGO is expected to bind ScDCR-produced 23-34nt sRNAs, it is plausible that the longer EhRNS produced sRNA may be poorly effective in triggering ScAGO mediated silencing. Moreover, *E. histolytica* endogenous sRNAs have been shown to be around 27nt in length which suggest that EhDCR activity in yeast is altered.

In addition to the yeast model, phenotypic complementation assays in *N. benthamiana* transgenic line KD for key endogenous *DCR* and ectopically expressing PtDCRb-c were unsuccessful. Based on our phylogenetic analysis, it is plausible that diatom DCR is too evolutionary distant to plants and thus its silencing genes are not able to complement a plant DCR homologue loss. Future studies employing heterologous expression systems should be addressed with host taxonomically closer to diatoms. Another and likely more straightforward approach would consist to analyze if the ectopic expression of PtDCR and PtAGO alternative spliceoforms in Pt DCR/AGO-KO lines could rescue the WT phenotype.

Efficient generation of Pt KO mutants by inducing large deletions in *PtDCR*, *PtAGO* and *PtRDR* genes

In 2016, two research groups independently developed CRISPR/Cas9 tools for targeted genome-mutagenesis in *P. tricornutum* (Hopes et al., 2016b; Nymark et al., 2016b). In this study, for the first time, two pKSdiaCas9 vectors carrying each a different sgRNA were co-transformed in order to generate two distant dsDNA breaks. The targets of the two sgRNAs were in a distance of ~300 and ~500 bp, inducing the biggest so far genomic deletions in diatoms, whereas Hopes and coworkers (2016) were introducing a deletion of 37bp. Since the transformation method used was biolistic bombardment with a co-vector carrying antibiotic resistance, the number of introduced plasmids in *P. tricornutum* was raised at three integrative vectors per targeted gene.

The results in this study show a high efficiency in the generation of biallelic homozygous KO mutants with deletions of up to 625 bp. In addition, numerous biallelic KO mutants either homozygous for smaller INDELs or heterozygous for various INDELS were produced. We went further and introduced five integrative vectors in *P. tricornutum*, targeting both *PtDCR* and *PtAGO* genes and generating the first PtDCR/AGO KO double mutants via this methodology. These KO mutant lines will

greatly facilitate further studies in diatom RNAi pathways, used as platforms for functional complementation assays and to investigate their transcriptome, epigenome and acclamatory response under environmental stress. Furthermore, PtDCR KO and PtAGO KO lines, in which transgene silencing is likely reduced, may represent stable and efficient algal chassis for the expression of heterologous protein of interest.

A dispensable DCR gene in *P. tricornutum?*

Elimination of PtDCR, PtAGO or PtRDR proteins doesn't seem to be lethal in *P. tricornutum* under standard laboratory culture conditions. Thus, RNAi seems dispensable in *P. tricornutum*, at least under the conditions employed. This observation comes as a surprise considering that *P. tricornutum* presents a single *DCR*, A*GO* and *RDR* gene

PtDCR KO mutants were also tested under NO_3 starvation. Compared to WT, PtDCR-KO mutants presented a slower growth and reached a lower maximum cell density in stationary phase. It has been previously suggested that epigenetic modification and reactivation of TE may play a role in the acclamatory response to nitrate starvation (Maumus et al., 2009; Veluchamy et al., 2013, 2015). Since nitrate limitation often occurs in the ocean, it is likely that DCR is indispensable for diatom cells in their natural environment.

On the other hand, exposure to UV radiation, known to induce DNA damage (Pattison and Davies, 2006) did not seem to affect PtDCR KO mutants much more that WT line. In the fungus *Neurospora crassa*, DNA damage has been shown to promote the generation of DCR-depended sRNA that interact with other proteins in order to inhibit rRNA biogenesis and protein synthesis (Lee et al., 2009a). In animals and plants dsDNA breaks induce DCR/DROSHA-dependent generation of sRNAs (Francia et al., 2012; Wei et al., 2012) that are assumed to guide a DNA repair machinery to the compromised DNA locus (Johanson et al., 2013). In diatoms, DCR function seems not likely to be involved in DNA repair.

Modification of other sRNAs populations in the absence of PtDCR

PtDCR KO mutant presents a drop in sRNA abundance and a shift in sRNA size distribution towards larger sizes. TE-derived sRNA, which are found on both strands of

the TE loci, dropped dramatically in DCR-KO. In contrast, the abundance sRNAs mapped to intergenic regions was comparable in WT and DCR-KO lines. This result is in agreement with the known enzymatic product of DCR, which is double stranded sRNAs. The enzyme(s) involved in the generation of the intergenic mono-strand sRNA await characterization.

PtDCR KO presented a different size distribution in tRFs with an increase of the longer fragments ~31nt and a decrease of the smaller ~18 nt fragments derived from the same mature tRNA. Documentation on the role of DCR in the processing of tRFs in animals and plants is limited. We propose that the observed changes in tRFs size distribution in PtDCR-KO is more likely a stress response to Cas9 integration and activity, DNA damage induced during biolistic or the DCR absence. This type of response could include decreased levels of proteinosynthesis, as indicated from the altered tRNA modification and the downregulation of genes that are involved in translation, metabolic processes and oxidation reduction processes identified in the PtDCR-KO mRNA transcriptome. Accordingly, the proteomic analysis suggests that PtDCR KO mutant presents a lower protein abundance per cell compare to WT.

TE expression in DCR KO mutant

In the *P. tricornutum* genome, 75% of the repeat sequences correspond to TEs out of which 75% is suppressed carrying epigenetic marks of methylation and histone modifications (Maumus et al., 2009, 2011; Veluchamy et al., 2015). Analysis of *P. tricornutum* sRNA populations showed that the majority of sRNAs were 25-30nt long that map to repetitive and silenced TEs marked by DNA methylation, while some of them also target DNA methylated protein-coding genes (Rogato et al., 2014). The functional RNA silencing machinery in *P. tricornutum* (De Riso et al., 2009), driven by only one identified DCR protein, could therefore generate small RNAs capable of guiding DNA methylation in an RNA-dependent DNA methylation (RdDM) fashion (Wassenegger et al., 1994; Teixeira et al., 2009). Furthermore, it has been shown that more than half of the highly methylated genes that are differentially expressed under nitrate starvation are targeted by sRNAs (Veluchamy et al., 2013; Rogato et al., 2014), suggesting that RdDM may have a role in the regulation of transcription of a subset of genes in diatoms. In this

study, whole analysis of Pt DCR KO sRNA indicates that PtDCR plays a pivotal role in the generation of TE targeted sRNAs.

Analysis of mRNA transcriptome in Pt DCR KO revealed an upregulation of a limited number of protein coding genes. Among the most upregulated transcripts were genes involved in TE expression and mobilization, like Reverse transcriptases, DNA integrases, Chomo-domain (Chromatin Organization Modifier) containing proteins and genes involved in transcription, DNA repair, as well as tRNA synthesis and modification. These results reinforce that DCR-dependent sRNAs are involved in TEs silencing. A recent analysis reported that reverse transcriptase encoding genes are prevalent in *P. tricornutum* genome (Rastogi et al., 2018). Their high abundance and active transcription suggest they may play an important role in diatom evolution and adaptation to contemporary environments (Lescot et al., 2016).

In DCR KO mutant the total number of TE-derived sRNAs, mapping to both DNA strands, dropped dramatically. The fact that a small number of sRNAs still mapped to TEs in DCR KO mutant remains enigmatic. The presence of a MiniIII protein has not been reported so far to complement a DCR activity in other organisms. It would be interesting to analyze the level of DNA methylation and histone modification at TEs and highly methylated genes in PtDCR KO. The major decrease in TE-derived sRNA abundance was reflected in the highly upregulated expression of specific autonomous TEs. These results underline the major role of PtDCR in the production of sRNAs to maintain TE repression. Future analysis in functionally complemented DCR-KO mutant i.e. ectopically expressing a functional PtDCR copy would allow to pinpoint the possible role of DCR in the initiation of TE silencing via the RdDM pathway.

Taken together, the results in this work indicate that the single and atypical DCR encoding gene present in *P. tricornutum* is involved in the generation of sRNAs repressing TE expression and may play a role in response to nitrate starvation. More work is needed to decipher the role of AGO and RDR in *P. tricornutum* and in other diatom species, in order to better understand the possible roles of RNAi in the evolution and acclamatory responses of these pivotal primary producers.

CHAPTER II:

Molecular and Functional characterization of HBI-Biosynthesis genes in the pennate diatom *Haslea ostrearia*

Supervisors:

Kriton Kalantidis, Associate Professor, University of Crete, IMBB/FORTH Group Leader

Sotiris Kampranis, Associate Professor, University of Copenhagen

Frederic Verret, Postdoc researcher, IMBBC/HCMR

Introduction

The ecological success and distinctive evolution of diatoms has driven scientific interest towards the study of their metabolism unveiling unique features that can be exploited in many potential biotechnological applications (Damsté et al., 2004; Kooistra et al., 2007; Bozarth et al., 2009; Allen et al., 2011; Obata et al., 2013). Their chimeric genetic background, product of endosymbiotic events and horizontal gene transfer, has provided diatoms a complex and flexible metabolism, exemplified by a broad diversity of metabolites including isoprenoids (Stonik and Stonik, 2015).

Isoprenoids are secondary metabolites well-known for their biotechnological applications in pharmaceutical, cosmetics, food and biofuels. The anti-cancer drug taxol and the antimalarial agent artemisinin are two of the well known and well studied molecules (Tippmann et al., 2013; George et al., 2015). Due to their beneficial characteristics, studies are focusing on engineering heterologous isoprenoid production such as in microbial hosts (Ro et al., 2006; Zhou et al., 2015; Meadows et al., 2016; Vickers et al., 2017; Ignea et al., 2018) and other platforms (Vavitsas et al., 2018; Lauersen, 2019).

Isoprenoid biosynthesis

All isoprenoids are produced by combinations of the same five-carbon atom precursors (C5), the two five-carbon isomers isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). IPP and DMAP are synthesized of these C5 via two biosynthetic pathways: the mevalonate (MVA) and the methylerythritol phosphate (MEP) pathway. Combination of IPP and DMAPP precursors forms prenyl diphosphate molecules which are subsequently used as substrates for the synthesis of the different isoprenoid classes (geranyl diphosphate (GPP) for C10 monoterpenoids, farnesyl diphosphate (FPP) for C15 sesquiterpenoids, C30 triterpenoids and sterols; geranylgeranyl diphosphate (GGPP) for C20 diterpenoids and C40 carotenoids; etc.). Prenyltransferase-type enzymes are responsible for the catalysis of these condensation reactions that consist the central steps of the isoprenoid biosynthetic pathway (Figure 63).

The next step comprises either the synthesis of squalene by FPP and GGPP leading to formation of sterols, or the synthesis of phytoene by FPP and GGPP leading to

formation of carotenoids, catalyzed by squalene/phytoene synthase-type enzymes (Vranová et al., 2012). In diatoms the presence of both MVA and MEP pathways has been reported (Cvejić and Rohmer, 2000; Massé et al., 2004b).

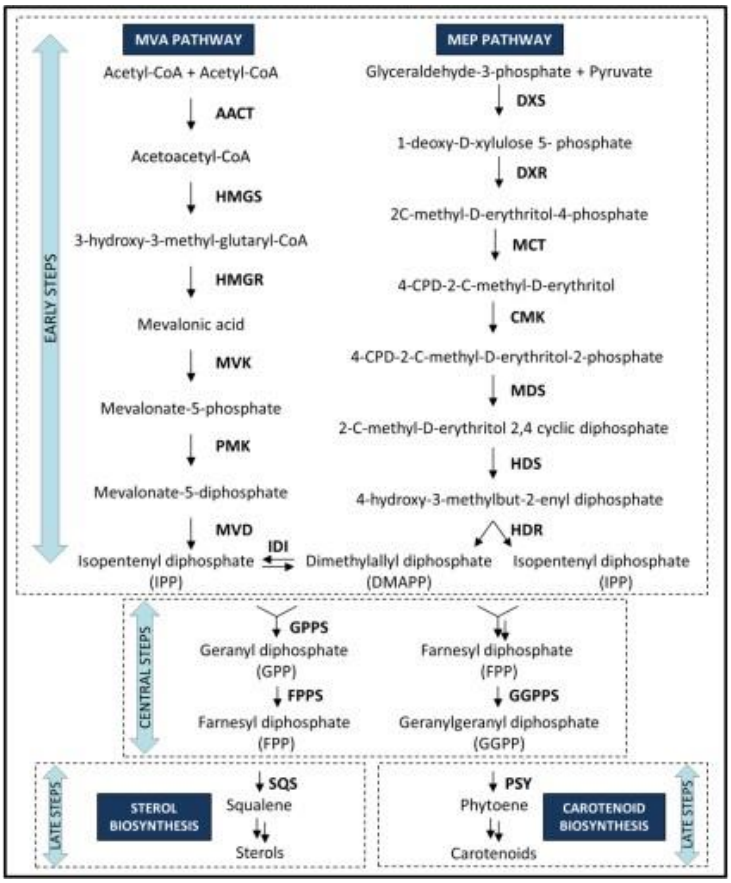

Figure 63. Basic isoprenoid biosynthesis pathway. AACT; Acetyl-coa c-acetyltransferase, HMGS; hydroxy-methylglutaryl-CoA synthase, HMGR; hydroxyl-methylglutaryl-CoA reductase, MVK; mevalonate kinase, PMK; phosphomevalonate kinase, MVD; mevalonate disphosphate decarboxylase, DXS; 1-deoxy-D-xylulose 5-phosphate synthase, DXR; 1-deoxy-Dxylulose 5-phosphate reductoisomerase, MCT; 2-C-methyl-D-erythritol-4-phosphate-cytidylyltransferase, CMK; 4-diphosphocytidyl-2c-methyl-d-erythritol kinase, MDS; 2-C-methyl-D-erythritol 2,4- cyclodiphosphate synthase, HDS; (E)-4-hydroxy-3-methylbut-2-enyl diphosphate synthase, HDR; hydroxymethylbutenyl diphosphate reductase, IDI; ispentenyl diphosphate isomerase, GPPS; geranyl diphosphate synthase, FPP; farnesyl diphosphate synthase, GGPP; geranylgeranyl diphosphate synthase, SQS; squalene synthase, PSY; phytoene synthase (adapted by Athanasakoglou et al., 2019).

Isoprenoids in diatoms

Compared to isoprenoids from terrestrial organisms, marine isoprenoids often present distinct characteristics (Stonik and Stonik, 2015). Among marine organisms, diatoms produce many isoprenoids with either conserved structures or consisting species-specific molecules.

Fucoxanthin has a distinctive structure contributing to its anti-oxidant (Sachindra et al., 2007), anti-obesity (Maeda, 2015) and anti-inflammatory effects (Kim et al., 2010; Tan and Hou, 2014). Fucoxanthinol (deacetylated derivative of fucoxanthin), has been reported to show anti-cancer activity (Martin, 2015) and the other two diatom specific xanthophylls, diatoxanthin and diadinoxanthin may have potential bioactivities (Sathasivam and Ki, 2018). Sterols are present in all eukaryotes as structural membrane components (Dufourc, 2008) but diatoms present a high diversity of sterol content. *Phaeodactylum tricornutum* produces brassicasterol as the main sterol, but also ergosterol (typically found in fungi) as an intermediate compound (Fabris et al., 2014a), *Cylindrotheca fusiformis* and *Nitzschia closterium* produce cholesterol (typically found in animals) as their major sterol and other species produce different phytosterols, underlining the effect of their complex genetic background. Pseudo-nitzschia diatoms release the neurotoxin domoic acid when they form algal blooms. Studies on domoic acid biosynthesis aim to provide a better understanding, monitoring and prevention of those harmful blooms (Brunson et al., 2018). Specific diatom genera, like Haslea, Rhizosolenia, Pleurosigma, Navicula and Berkeleya have representatives species that produce Highly Branched Isoprenoids (HBIs) with unique chemical structure and potential for biotechnological applications.

Recent progresses in genomic and transcriptomic NGS data (Keeling et al., 2014) has enabled the study of more diatom species. Considering the high diversity in diatoms, the identification of new isoprenoids is increasingly expected.

Highly Branched Isoprenoids (HBIs) in diatoms

HBIs are usually encountered in structures of 25 or 30 carbon atoms and 1–6 double bonds (Belt et al., 2000), but other monocyclic compounds, in addition to epoxides, alcohols, thiolanes and thiophenes, have also been found (Belt et al., 2000, 2006; Massé et al., 2004b, 2004a). Based on their structure, their role as membrane components has been proposed but their exact biological role remains uknown.

Applications

HBIs have been extensively used in reconstructing paleoenvironmental climate. They are used as geochemical markers which, in combination to development of climate models, can reconstruct changes in sea-ice coverage, interpret past climate conditions and predict future climate states (Belt and Müller, 2013; Collins et al., 2013). Moreover, HBI branched structure and degree of saturation makes them good candidates for biofuel production. Some HBIs C25 (25 carbon atoms) have demonstrated in vitro cytostatic effects against human lung cancer cell lines (Rowland et al., 2001).

Biosynthesis

Diatoms have both MEP pathway and MVA route for IPP and DMAPP synthesis (Cvejić and Rohmer, 2000; Massé et al., 2004b, 2004a; Fabris et al., 2014b; Di Dato et al., 2015). By using isotopic precursor labelling experiments and inhibitors of the MVA and MEP pathway Masse and coworkers found that centric HBI-producing diatom species *Rhizosolenia setigera* synthesizes sterols and HBIs via the MVA pathway. Conversely, *H. ostrearia* synthesizes HBIs and β-sitosterol using precursors from the MEP pathway (Massé et al., 2004b, 2004a). Ferriols and coworkers proposed that in *R. setigera* two condensed FPP (C15) molecules in a headto-middle (1′-6′) orientation synthesize C30 HBIs, while the condensation of one FPP (C15) and one GPP (C10) molecule likely produce C25 isomers (Ferriols et al., 2015). Since other functionalized HBIs like epoxides and alcohols had been reported in *H. ostrearia* (Belt et al., 2006), other isoprenoid precursors may also be substrates (e.g. C10 linalool and C15 farnesol) or the addition of the functional groups could take place later. By in vivo inhibition of a characterized FPP synthase (AKH49589.1) in *R. setigera* the HBI synthesis was reduced, indicating that FPP would be one of the HBIs precursors (Ferriols et al., 2015). More

recently, identification of IDI and SQS transcripts in C30 HBI producing *R. setigera* strain and an IDI-SQS fusion in C25 HBI producing *H. ostrearia* and *R. setigera* strains has suggested that IDI-SQS gene fusion may be involved in the regulation of the GPP supply required for C25 but not for C30 HBIs (Ferriols et al., 2017).

Although the enzymes involved in the initial steps of the MVA and MEP pathways in HBI-producing diatoms have been studied, the function, subcellular localization and regulation of prenyltransferases involved in the next steps remains elusive. Considering that synthesis of prenyl diphosphates determine the flux towards different isoprenoids, elucidation of the enzymes responsible for their synthesis and modifications would improve our understanding on HBI synthesis in diatoms.

The aim of this project was to identify genes involved in isoprenoid and HBI biosynthesis in *Haslea ostrearia*. *H. ostrearia* is a cosmopolitan species known for the production of C25 HBIs. Haslea species also produce the blue pigment marennine, which is a water-soluble, blue-green pigment responsible for the greening of oysters. Genome and transcriptome sequencing analysis enabled the identification of six gene candidates. Functional characterization was carried out by heterologous expression in *S. cerevisiae.* Large scale cultures of *H. ostrearia* were attempted to provide sufficient biomass for a Nuclear Magnetic Resonance *(*NMR*)* spectroscopy analysis of the produced HBIs.

Materials and Methods

Diatom cultures

The *H. ostrearia* clone NCC 153.8 used in this study was a product of heterothallic reproduction between clones NCC 141 and NCC 171, both of which were isolated in 2003 from natural populations of the oyster pond Lainard (La Barre de Monts). Clone NCC 153.8 was kindly provided by Dr Vona Medeler (University of Nantes) and cultured at the University of Crete Greece, in f/2 medium (Guillard, 1975) at 20℃ under an irradiance of 50 μmol m-2 s-1 (12 h : 12 h light : dark cycle).

HBI analysis in *H. ostrearia* (in University of Laval, Quebec)

Chemical extraction

The production HBIs in the strain NCC 153.8 was confirmed by chemical extraction and gas chromatography–mass spectrometry (GC–MS) analysis according to a previously described mehodology (Goutte et al., 2013). *H. ostrearia* cells in log phase were harvested by filtration and were then saponified by incubation in a KOH solution (MeOH/ H2O, 80/20) for 2 hours at 80 °C. The non-saponifiable lipids were extracted 3 times in hexane and dried over anhydrous Na2SO4. Open column chromatography (50:1 SiO2:NSLs (w/w)) were used for purification of the extract.

GC/MS analysis

The fraction containing HBIs was analyzed using an Agilent (Santa Clara, CA, USA) 7890A gas chromatograph (GC) fitted with a 30 m fused silica Agilent J&C GC column (0.25 mm i.d.,0.25 mm film) coupled to an Agilent 5975C Series mass selective detector (MSD). The GC oven temperature was programmed from 40 to 300 ºC at 10 ºC min-1 and held at the final temperature for 10 min.

Genome sequencing

Genome sequencing of *H. ostrearia* had already been performed in sequencing facilility of IMBB gDNA was extracted following CTAB based protocol.

RNA sequencing

RNA extraction, transcriptome sequencing and analysis were performed at the Institute of Applied Biosciences of the Centre for Research and Technology Hellas,

CERTH, Thessaloniki. Total RNA was extracted with SpectrumTM Plant Total RNA Kit (Sigma-Aldrich). Messenger RNA was isolated from total RNA using the NEBNext® Poly(A) mRNA Magnetic Isolation Module (New England Biolabs). Complementary DNA (cDNA) library was constructed using the NEBNext Ultra Directional RNA Library Kit for Illumina (New England Biolabs), according to the manufacturer's instructions. Sequencing was performed on an Illumina NextSeq500 platform using the NextSeqTM 500/550 Mid Output Kit (2 9 150 cycles) (Illumina). Bioinformatics analysis involved removal of adaptors and low-quality sequence and de novo assembly of the reads using the TRINITY software suite (Grabherr et al., 2011) with default parameters.

Gene identification and signal peptides prediction in their protein products

Candidate genes involved in isopropanoid biosynthesis in *H.ostrearia* were identified based on homology with characterized and annotated homologues retrieved from the Genbank protein database (NCBI) and DiatomCyc database (Fabris et al., 2012). Protein subcellular localization was predicted by SignalP-5.0 (www.cbs.dtu.dk/services/SignalP) and TargetP-2.0 (www.cbs.dtu.dk/services/TargetP/) for signal peptide and transmembrane domain prediction, respectively.

Gene amplification and cloning

Genomic DNA was extracted using a modified CTAB protocol (described in chapter I). In order to optimize yield and quality of extracted DNA from *H. ostrearia*, the step of phenol:chloroform: isoamyl (25:24:1) was omitted, as it appeared to interfere with other compounds that induced DNA degradation. Gene candidates were initially amplified from genomic DNA with primers designed on the genome sequencing results. PCR products were gel-purified, quantified and cloned in pCRII-TOPO vectors (Thermo Fisher Scientific) or in pGEM-Teasy vectors (Promega) after A-tailing with Taq (Minotech). Gene sequences were confirmed by Sanger sequencing.

Total RNA was extracted using Trizol, and cDNA was synthesized using SuperScript III RT (Thermo Fisher Scientific) and DNAseI (Roche Applied Biosystems) treatment. Both full-length and truncated variants of the selected genes were PCR amplified using KAPA HiFi (HighFidelity) DNA polymerase (KAPABiosystems) and

cDNA as template. PCR products were gel-purified, quantified and cloned in pCRII-TOPO vectors (Thermo Fisher Scientific) or in pGEM-Teasy vectors (Promega) after A-tailing with Taq (Minotech). Gene sequence was validated with Sanger sequencing.

For yeast expression, candidate genes were cloned in yeast expression vectors pUTDH3myc and pWTDHmyc (Ignea et al., 2012), after digestion with the appropriate restriction enzymes and ligation with T4 DNA ligase (Invitrogen). Final constructs were verified by sequencing. Primers used for cloning are listed in Table 20.

Table 20. Primers used in this study.

Primer name	Sequence
Primers for basic cloning of TpS genes	
HoTp 982 Fw	AGATTGAATCCTTTTGTTTGACGCC
HoTp 982 R	AATTGGGGAACAATTGAGAGGCT
HoTp 706 Fw	AACTACGATGGTTCAGAGAGATG
HoTp 706 R	ATGCGACCTAAAGTGGCTATTG
TpS 697 Fw2	GCTCGATAATGAGTTCGTTGGT
TpS 697 R2	AGCTACATATTCCTGTCCCTAGA
TpS 727 Fw2	AGTGAGATAATGTCAGGGAAGATT
TpS 727 R2	TGATCCAGAAAATGTCCTCTTCTAT
TpS 784 Fw2	TGTGTGTCATGTCAATGAATTCCAA
TpS 784 R2	TCCCACCGAATGCCTTGATT
TpS 144 Fw2	AGAAAATGACGGAGCCAAGC
TpS 144 R2	TCACAGAGTCTTTTTCTTCAAACG
Primers with restriction sites for cloning TpS genes in yeast vectors	
TpS 697-EcoRI Fw	AGAATTCATGAGTTCGTTGGTCATCCCTG
TpS 697-Sali Rv	AAAGTCGACTTATTTGGATCGCTTGTAAATTTTG
TpS 706-EcoRI Fw	AGAATTCATGGTTCAGAGAGATGAATTGAAAC
TpS 706 -SalI Rv	AAAGTCGACTCAACTAAAGACAAGACTTGGAAAAG
TpS 727-BamHI Fw	AGGATCCATGTCAGGGAAGATTCTAGAGGG
TpS 727-SalI Rv	AAAGTCGACCTATATGTGCAATTTCGACACAGG
TpS 982-BamHI Fw	AGGATCCATGTTGAGAAAATGCAGCATGCAC
TpS 982-NotI Rv	AAAAGCGGCCGCACTATTGGTTTGTTCGAGATACAACTTTATG
TpS 166-BglII Fw	AAAAGATCTCAATTGATGTCAATGAATTCCAATAATACTAATAA
TpS 166 - XhoI Rv	ACTCGAGTTAGGTGGATTCTATACGCTCGG
TpS 451-BamHI Fw	AGGATCCATGCACTCTTTCAGCAAGGC
TpS 451 -XhoI Rv	ACTCGAGTTAGTTTTTACGGTTGATAATGTAG
Primers for cloning TpS genes in yeast vectors without Signal Peptides	
TpS 451 pos21 -BamHI Fw	AGGATCCATGTCCGTGAGTCCAACCTACCAACGCA
TpS 697 pos16 -EcoRI Fw	AGAATTCATGGCTAACTTGACCGACAAATTGGATGTGTTTG
TpS 982 pos35 -BamHI Fw	AGGATCCATGAGGTCAACATCCCCTGCAGGC
TpS 706 pos37-EcoRI Fw	AGAATTCATGTTCACCCACATCAACAATTCTGG
TpS 706 pos44 -EcoRI Fw	AGAATTCATGGGCGCCCAAACGCACTTG

Expression and product extraction in yeast

The yeast strain AM94 (Ignea et al., 2012) was transformed with yeast expression vectors carrying the candidate genes. Yeast cells were cultured in complete minimal medium, composed of 0.13% (w/v) dropout powder (all essential amino acids minus auxotrophy markers) 0.67% (w/v) Yeast Nitrogen Base without amino acids (Y2025, US Biological Life Sciences) and 2% D-(+)-glucose monohydrate (16301, Sigma-Aldrich Co. St. Louis). Cultures and media preparation are described in chapter I.

HBI analysis in *S.cerevisiae* (in University of Crete and HCMR)

Extraction

For the detection of HBI intermediate products, yeast cells grown for three days in 10 ml cultures with addition of 1ml dodecane were centrifuged at 3.400 rpm. The upper phase of dodecane was collected and cells pellets were disrupted with glass beads and extracted three times in hexane after saponification (described above). Samples were dried with Na2SO4 powder, the upper phase was collected after centrifugion, evaporated and finally eluted in 50 μλ Hexane. Internal control sclareol (diluted x1000 in hexane) was added in dodecane extracts and in cell pellet extract during the first hexane extraction. Both dodecane extracts and hexane extracts were analyzed in GC/MS. Individual compounds were identified by comparing their GC retention indices and mass spectra with those of internal standard sclareol.

GC/MS analysis

Dodecane extracts and hexane fraction wer analyzed using a Hewlett–Packard(HP) 6890 Series GC coupled to a HP 5972 A MS. The GC oven temperature was programmed from 40 to 300 oC at 10 oC min-1 and held at the final temperature for 10 min. GC/MS analysis was performed by Dr.Mandalakis (HCMR).

Results and Discussion

Production of HBIs in *Haslea ostrearia*

The presence of C25 HBIs in *H. ostrearia* NCC 153.8 strain was confirmed during a visit at the laboratory of Dr. Massé TAKUVIK, in University of Laval, Quebec, within the frame of the Biovadia project (RISE) (Figure 64). Training in chemical extraction, characterization and quantification of HBIs by GC/MS in diatom was obtained.

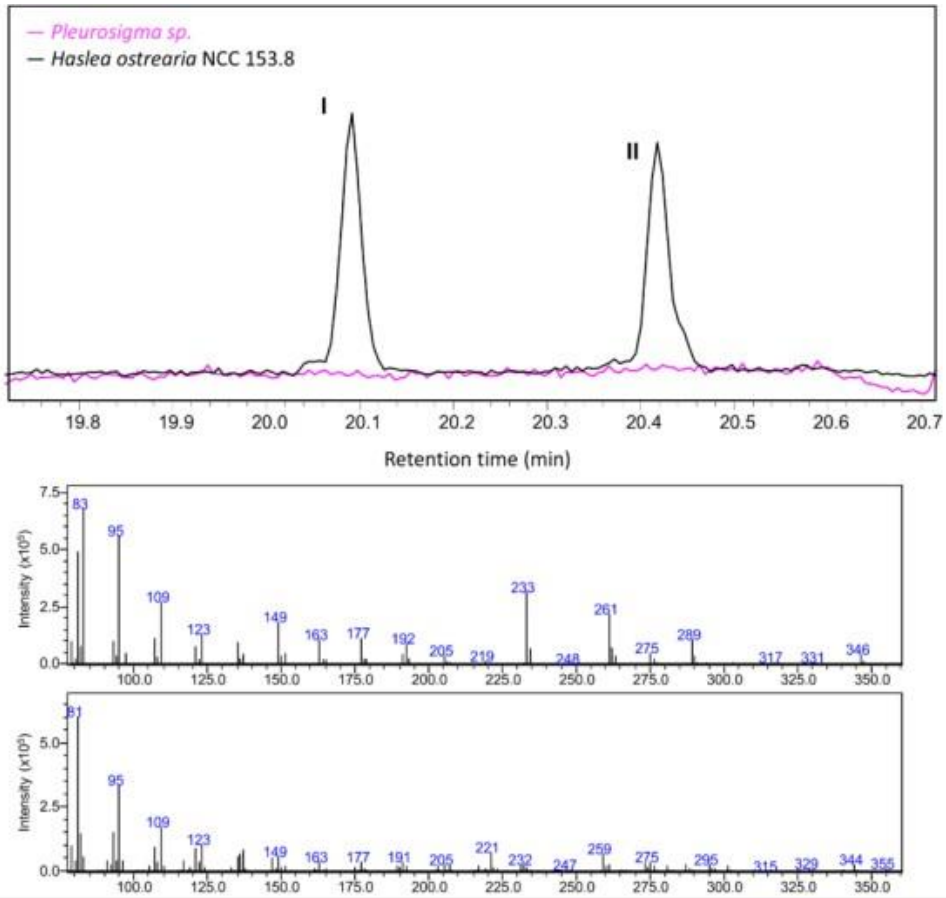

Figure 64. GC/MS profile of HBI analysis in *Haslea ostrearia* NCC 153.8 strain in comparison to a non-HBI-producing strain of Pleurosigma species (negative control).

Large-scale cultures of *H. ostrearia*

The establishment of large-scale cultures of *H. ostrearia* was attempted in order to produce large quantities of biomass for extraction and purification of HBIs, aiming at their structural characterization. Large-scale cultures in flow-bioreactors at the Institute for Marine Biology, Biotechnology and Aquaculture, HCMR) proved to be unproductive due to the exclusively benthic life style of this organism. Thus, multiple medium-scale cultures in large square bioassay dishes were established in order to provide the larger possible surface for diatom growth. Cultures were let to grow for short periods of time due to the limited volume of media and the cells were harvested by centrifugion. The produced biomass is being currently analyzed for the chemical characterization of HBIs at dr. V. Roussis lab (in Department of Pharmacognosy and Natural Product Chemistry, Athens), within the frame of the ongoing CMBR project (ESFRI).

Cloning of candidate biosynthetic genes

HBI biosynthesis gene candidate genes were initially identified and cloned based on the data produced from genome sequencing. Names of gene candidates TpS*** correspond to “Terpene Synthases” and the last three numbers of the contig they were identified on. Amplified genes from genomic DNA were cloned and their sequences were validated by Sanger sequencing. Subsequently, gene amplification from cDNA followed with cloning and Sanger sequencing validation as well. Out of the seven gene candidates, six were amplified from cDNA and four of them were also amplified from genomic DNA (Table 21). A truncated sequence was amplified from genomic DNA of candidate TpS784 (indicated with an asterisk at Table 21) due to low quality sequencing data at this region in contig, while its trasncriptomic full length sequence was later identified. Gene candidate TpS144 could not be amplified from any template and was not analyzed further.

RNA sequencing data validated the expression of all gene candidates and due to the high degree of sequence conservation among isoprenoid biosynthetic genes enabled the identification of new genes that putatively encode enzymes of the MVA and MEP pathways. Search for an IDI gene led to the identification a contig corresponding to a fusion of an IDI with a squalene synthase (HoIDI-SQS) that was initially reported as TpS144.

Sequence analysis of the identified gene candidates led to identification of two conserved aspartic acid-rich motifs DDxx(xx)D (First Aspartic acid-Rich Motif – FARM; and Second Aspartic acid-Rich Motif – SARM), that are involved in the binding of magnesium ions and are essential for prenyltransferase activity (PTS). The identification of intact motifs in the five candidate genes TpS697/706/982/784/451 suggested they likely encode PTSs active enzymes.

In the present study, the functional characterization of the six aforementioned cloned candidates was attempted using *S. cerevisiae* strain AM94 as heterologous expression system (Ignea et al., 2012).

Table 21. Genes identified and cloned in this study.

Gene name	Size in genDNA	Size in cDNA	Scaffold Genomic ID	Transcriptome ID	Putative name (best NCBI Hit)
TpS 697	1.395bp	1.395bp	scaffold17697	c18049_g1_i1	farnesyl pyrophosphate synthase
TpS 706	1.718bp	1.718bp	scaffold38706	c18223_g1_i1	farnesyl pyrophosphate/diphosphate synthase
TpS 727	1.336 Kb	1.279bp	scaffold37727	c15731_g1_i1	phytoene synthase
TpS 982	1.519bp	1.519bp	scaffold34982	c9848_g1_i2	geranylgeranyl diphosphate/pyrophosphate
TpS 784	1.272bp*	1.017bp	scaffold37784	c19166_g1_i1	geranylgeranyl diphosphate/pyrophosphate synthase
TpS 451	-	1014bp	-	c22451_g1_i1	geranylgeranyl pyrophosphate synthase-farnesyltranstransferase
TpS 144	x	x	scaffold9144	c14333_g1_i1	Isopentenyl-diphosphate delta isomerase fused to squalene synthase

Expression of candidate biosynthetic genes in *S. cerevisiae*

TpS cDNA were sub-cloned in the yeast expression vectors pUTDH3myc and pWTDHmyc. Cloning each gene candidate to both vectors enabled the simultaneous transformation of two different constructs in yeast and the analysis of their combined expression. cDNA were expressed either alone or in all possible pairs.

Cultures were grown in liquid media with an overlay of dodecane in order to trap secreted products. Cell pellet-derived fraction eluted in hexane and dodecane from each culture were analyzed for the presence of HBIs or compounds that could correspond to HBI intermediate products. Since yeast cells produce isoprenoids, yeast strains

transformed with empty pUTDH3myc and/or pWTDHmyc were used as negative controls to distinguish this background from the products of candidate gene expression. Expression of one or combinations of two candidate genes was expected to produce new peaks in the GC/MS chromatogram, in comparison to the profile of negative control strains. Identification of compounds was based on comparison of accurate mass and retention time between reaction substrates/products and known standards (profile of known compounds).

Dodecane extract analysis

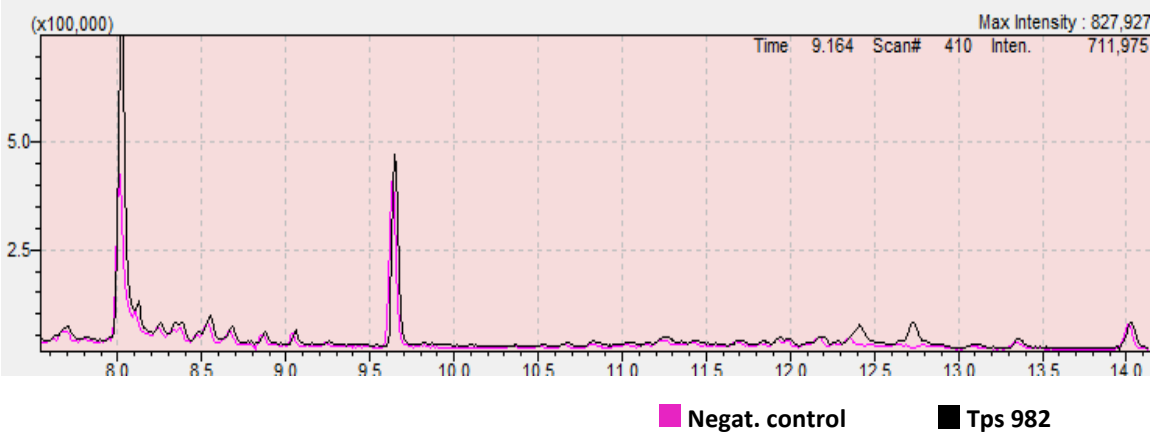

Figure 65. Gas Chromatography profile of yeast extract in dodecane. Yeast strains transformed with candidate gene TpS982 or with empty vector (Negat.control) produce the same compounds.

Yeast cells transformed with each gene candidate separately or in any possible combination of two genes did not produce any new compounds compared to the negative control from either dodecane supernatant or their cell extract (Figure 65). The only difference observed was the higher yield in farnesol production when TpS727 was present (Figure 66), indicating its efficient expression in yeast. The expression of only TpS727 gene candidate in yeast was validated also by western blot analysis (data not shown).

Cell pellet extract analysis

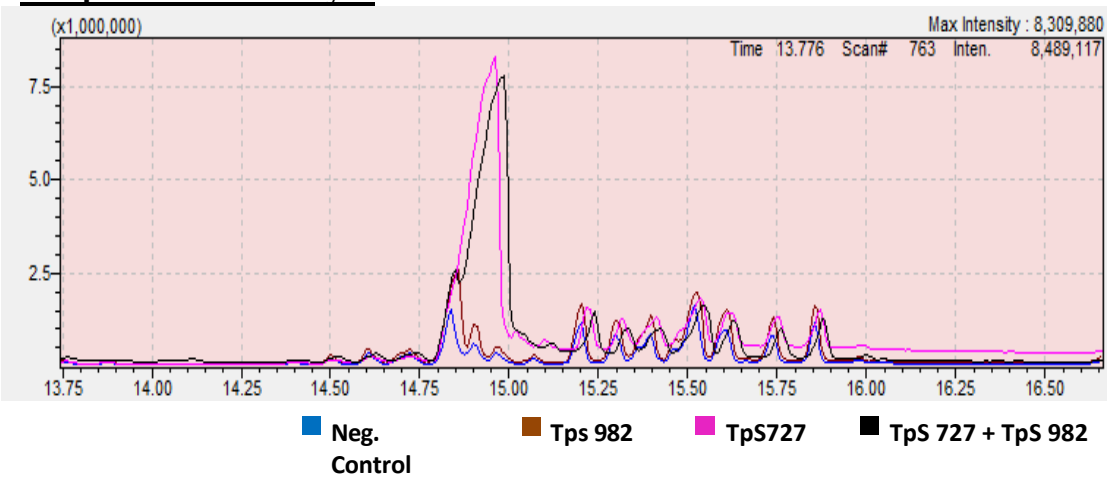

Figure 66. Gas Chromatography profile of yeast extract in hexane. Yeast strains transformed with candidate gene TpS982, TpS727, a combination of both genes or with empty vector (Negat.control) produce the same compounds, but a higher yield of farnesol is achieved when TpS727 gene is expressed.

To investigate the possibility that HBI candidates carry signal/target peptides preventing their correct localization in yeast, a prediction of signal/target peptides was performed with online tools. Five gene candidates were predicted to have subcellular localization signal peptides (Table 22). Truncated genes lacking the predicted signal peptides were PCR amplified and cloned in pUTDH3myc and pWTDHmyc.

Table 22. Prediction of signal peptides in produced proteins by SignalP.

Gene name	SignalP prediction of localization	Position of target peptide in protein
TpS 697	secretory	16aa
TpS 706	plastid	37aa, 44aa
TpS 727	-	-
TpS 982	mitochondrion	35aa
TpS 784	secretory	32aa
TpS 451	plastid	21aa

From this point on, the project was continued from members of Dr.Kampranis lab (Anastasia Athanasakoglou and Condruta Ignea) who performed expression experiments of the aforementioned and new gene candidates in yeast, bacteria and plants, along with a more advanced signal/target peptide prediction and phylogenetic analysis of the genes under study. These experiments confirmed that only TpS727 and the IDI-SQSfusion TpS144 could be expressed in yeast. The expression of the other five gene candidates was successful only in bacteria. The outcome of the functional characterization of the HBI candidate genes has been published in *New Phytologist, 2018* (appendix). A summary of these results is presented below (Table 23). As mentioned before the chemical characterization of HBIs and other compounds from medium scale cultures of *Haslea osteraria* are currently analyzed with Nuclear Magnetic Resonance *(*NMR*)* spectroscopy.

Table 23. Final functional characterization of gene candidates summarized, from Athanasakoglou et al, 2018.

Gene name in this study	Characterized Gene name	Characterized gene function
Mevalonate pathway		
Tps144	HoIDISQ	Isopentenyl-diphosphate delta-isomerase fused to squalene synthase
Centarl steps		
Tps697	HoPTS1	Farnesyl diphosphate synthase
Tps706	HoPTS2	Polyprenyl diphosphate synthase
Tps784	HoPTS3	Geranylgeranyl diphosphate synthase
Tps982	HoPTS4	Putative polyprenyl synthase
Tps451	HoPTS5	Geranylgeranyl diphosphate synthase
Tps727	HoPSY	Phytoene synthase

REFERENCES

Aguado, L. C., and Benjamin, R. (2017). RNase III Nucleases and the Evolution of Antiviral Systems. 1700173, 1–6. doi:10.1002/bies.201700173.

Alipanah, L., Rohloff, J., Winge, P., Bones, A. M., and Brembu, T. (2015). Whole-cell response to nitrogen deprivation in the diatom Phaeodactylum tricornutum. *J. Exp. Bot.* doi:10.1093/jxb/erv340.

Allen, A. E., Dupont, C. L., Oborník, M., Horák, A., Nunes-Nesi, A., McCrow, J. P., et al. (2011). Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature*. doi:10.1038/nature10074.

Allen, A. E., Vardi, A., and Bowler, C. (2006). An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. *Curr. Opin. Plant Biol.* 9, 264–273. doi:10.1016/j.pbi.2006.03.013.

Apt, K. E., Kroth-Pancic, P. G., and Grossman, A. R. (1996). Stable nuclear transformation of the diatom Phaeodactylum tricornutum. *Mol. Gen. Genet.* 252, 572–579. doi:10.1007/s004380050264.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*. doi:10.1038/nature04916.

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., et al. (2004). The Genome of the Diatom Thalassiosira Pseudonana: Ecology, Evolution, and Metabolism. *Science (80-.).* 306, 79–86. doi:10.1126/science.1101156.

Athanasakoglou, A., Grypioti, E., Michailidou, S., Ignea, C., Makris, A. M., Kalantidis, K., et al. (2019). Isoprenoid biosynthesis in the diatom *Haslea ostrearia*. *New Phytol.* 222, 230–243. doi:10.1111/nph.15586.

Athanasakoglou, A., and Kampranis, S. C. (2019). Diatom isoprenoids: Advances and biotechnological potential. *Biotechnol. Adv.* 37. doi:10.1016/j.biotechadv.2019.107417.

Bailleul, B., Berne, N., Murik, O., Petroutsos, D., Prihoda, J., Tanaka, A., et al. (2015).

Energetic coupling between plastids and mitochondria drives CO 2 assimilation in diatoms. *Nature* 524, 366–369. doi:10.1038/nature14599.

Bailleul, B., Rogato, A., De Martino, A., Coesel, S., Cardol, P., Bowler, C., et al. (2010). An atypical member of the light-harvesting complex stress-related protein family modulates diatom responses to light. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1007703107.

Baldauf, S. L. (2003). The deep roots of eukaryotes. *Science* 300, 1703–6. doi:10.1126/science.1085544.

Ballarino, M., Pagano, F., Girardi, E., Morlando, M., Cacchiarelli, D., Marchioni, M., et al. (2009). Coupled RNA Processing and Transcription of Intergenic Primary MicroRNAs. *Mol. Cell. Biol.* doi:10.1128/mcb.00664-09.

Bártulos, C. R., Rogers, M. B., Williams, T. A., Gentekaki, E., Brinkmann, H., Cerff, R., et al. (2018). Mitochondrial glycolysis in a major lineage of eukaryotes. *Genome Biol. Evol.* doi:10.1093/gbe/evy164.

Battarbee, R. W. (1988). The use of diatom analysis in archaeology: A review. *J. Archaeol. Sci.* doi:10.1016/0305-4403(88)90057-X.

Baulcombe, D. (2004). RNA silencing in plants. *Nature*. doi:10.1038/nature02874.

Baurain, D., Brinkmann, H., Petersen, J., Rodríguez-Ezpeleta, N., Stechmann, A., Demoulin, V., et al. (2010). Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Mol. Biol. Evol.* 27, 1698–1709. doi:10.1093/molbev/msq059.

Bazzini, A. A., Mongelli, V. C., Hopp, H. E., Del Vas, M., and Asurmendi, S. (2007). A practical approach to the understanding and teaching of RNA silencing in plants. *Electron. J. Biotechnol.* 10, 178–190. doi:10.2225/vol10-issue2-fulltext-11.

Becker, D. M., and Lundblad, V. (1994). Introduction of DNA into Yeast Cells. *Curr. Protoc. Mol. Biol.* 27, 13.7.1-13.7.10. doi:10.1002/0471142727.mb1307s27.

Belt, S. T., Allard, W. G., Massé, G., Robert, J. M., and Rowland, S. J. (2000). Highly branched isoprenoids (HBIs): Identification of the most common and abundant sedimentary isomers. *Geochim. Cosmochim. Acta*. doi:10.1016/S0016-7037(00)00464-6.

Belt, S. T., Massé, G., Rowland, S. J., and Rohmer, M. (2006). Highly branched

isoprenoid alcohols and epoxides in the diatom Haslea ostrearia Simonsen. *Org. Geochem.* doi:10.1016/j.orggeochem.2005.10.005.

Belt, S. T., and Müller, J. (2013). The Arctic sea ice biomarker IP 25 : a review of current understanding , recommendations for future research and applications in palaeo sea ice reconstructions. *Quat. Sci. Rev.* 79, 9–25. doi:10.1016/j.quascirev.2012.12.001.

Bender, S. J., Durkin, C. A., Berthiaume, C. T., Morales, R. L., and Armbrust, E. V. (2014). Transcriptional responses of three model diatoms to nitrate limitation of growth. *Front. Mar. Sci.* doi:10.3389/fmars.2014.00003.

Bertozzini, E., Galluzzi, L., Ricci, F., Penna, A., and Magnani, M. (2013). Neutral lipid content and biomass production in Skeletonema marinoi (Bacillariophyceae) culture in response to nitrate limitation. *Appl. Biochem. Biotechnol.* doi:10.1007/s12010-013-0290-3.

Blevins, T., Podicheti, R., Mishra, V., Marasco, M., Wang, J., Rusch, D., et al. (2015). Identification of pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding de novo DNA methylation in arabidopsis. *Elife*. doi:10.7554/eLife.09591.

Bohmert, K., Camus, I., Bellini, C., Bouchez, D., Caboche, M., and Banning, C. (1998). AGO1 defines a novel locus of Arabidopsis controlling leaf development. *EMBO J.* doi:10.1093/emboj/17.1.170.

Borges, F., and Martienssen, R. A. (2015). The expanding world of small RNAs in plants. *Nat. Publ. Gr.* 16, 1–15. doi:10.1038/nrm4085.

Bouché, N., Lauressergues, D., Gasciolli, V., and Vaucheret, H. (2006). An antagonistic function for Arabidopsis DCL2 in development and a new function for DCL4 in generating viral siRNAs. *EMBO J.* 25, 3347–56. doi:10.1038/sj.emboj.7601217.

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., et al. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. 456, 239–244. doi:10.1038/nature07410.

Bowler, C., Vardi, A., and Allen, A. E. (2010). Oceanographic and Biogeochemical Insights from Diatom Genomes. *Ann. Rev. Mar. Sci.* 2, 333–365. doi:10.1146/annurev-marine-120308-081051.

Bozarth, A., Maier, U. G., and Zauner, S. (2009a). Diatoms in biotechnology: Modern tools and applications. *Appl. Microbiol. Biotechnol.* doi:10.1007/s00253-008-1804-

8.

Bozarth, A., Maier, U., and Zauner, S. (2009b). Diatoms in biotechnology : modern tools and applications. 195–201. doi:10.1007/s00253-008-1804-8.

Braunschweig, U., Barbosa-Morais, N. L., Pan, Q., Nachman, E. N., Alipanahi, B., Gonatopoulos-Pournatzis, T., et al. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* doi:10.1101/gr.177790.114.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., et al. (2007). Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell*. doi:10.1016/j.cell.2007.01.043.

Brunson, J. K., McKinnie, S. M. K., Chekan, J. R., McCrow, J. P., Miles, Z. D., Bertrand, E. M., et al. (2018). Biosynthesis of the neurotoxin domoic acid in a bloom-forming diatom. *Science (80-.).* doi:10.1126/science.aau0382.

Buck, J. M., Bártulos, C. R., Gruber, A., and Kroth, P. G. (2018). Blasticidin-S deaminase , a new selection marker for genetic transformation of the diatom Phaeodactylum tricornutum. 1–13. doi:10.7717/peerj.5884.

Carthew, R. W., and Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*. doi:10.1016/j.cell.2009.01.035.

Cerutti, H., and Casas-Mollano, J. A. (2006). On the origin and functions of RNA-mediated silencing: From protists to man. *Curr. Genet.* doi:10.1007/s00294-006-0078-x.

Cerutti, H., Ma, X., Msanne, J., Cerutti, H., Ma, X., Msanne, J., et al. (2011). RNA-Mediated Silencing in Algae : Biological Roles and Tools for Analysis of Gene Function RNA-Mediated Silencing in Algae : Biological Roles and Tools for Analysis of Gene Function □. doi:10.1128/EC.05106-11.

Cerutti, L., Mian, N., and Bateman, A. (2000). Domains in gene silencing and cell differentiation proteins: tTe novel PAZ domain and redefinition of the Piwi domain. *Trends Biochem. Sci.* doi:10.1016/S0968-0004(00)01641-8.

Chomczynski, P., and Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* 162, 156–159. doi:10.1016/0003-2697(87)90021-2.

Chong, M. M. W., Zhang, G., Cheloufi, S., Neubert, T. A., Hannon, G. J., and Littman,

D. R. (2010). Canonical and alternate functions of the microRNA biogenesis machinery. *Genes Dev.* doi:10.1101/gad.1953310.

Christian, M., Cermak, T., Doyle, E. L., Schmidt, C., Zhang, F., Hummel, A., et al. (2010). Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics*. doi:10.1534/genetics.110.120717.

Chu, L., Ewe, D., Río Bártulos, C., Kroth, P. G., and Gruber, A. (2016). Rapid induction of GFP expression by the nitrate reductase promoter in the diatom *Phaeodactylum tricornutum*. *PeerJ* 4, e2344. doi:10.7717/peerj.2344.

Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* doi:10.1038/nrg.2016.139.

Claycomb, J. M. (2014). Ancient Endo-siRNA Pathways Reveal New Tricks. *Curr. Biol.* 24, R703–R715. doi:10.1016/j.cub.2014.06.009.

Cogoni, C., and Macino, G. (1999). Gene silencing in Neurospora crassa requires a protein homologous to RNA-dependent RNA polymerase. *Nature*. doi:10.1038/20215.

Cole, C., Sobala, A., Lu, C., Cole, C., Sobala, A., Lu, C., et al. (2009). Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. 2147–2160. doi:10.1261/rna.1738409.

Collart, M. A., and Oliviero, S. (1993). Preparation of Yeast RNA. *Curr. Protoc. Mol. Biol.* 23, 13.12.1-13.12.5. doi:10.1002/0471142727.mb1312s23.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J., Fichefet, T., Friedlingstein, P., et al. (2013). “Long-term Climate Change: Projections, Commitments and Irreversibility. In: Climate Change 2013: The Physical Science,” in *Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* doi:10.1017/CBO9781107415324.024.

Colmenares, S. U., Buker, S. M., Buhler, M., Dlakić, M., and Moazed, D. (2007). Coupling of Double-Stranded RNA Synthesis and siRNA Generation in Fission

Yeast RNAi. *Mol. Cell*. doi:10.1016/j.molcel.2007.07.007.

Creasey, K. M., Zhai, J., Borges, F., Van Ex, F., Regulski, M., Meyers, B. C., et al. (2014). MiRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. *Nature*. doi:10.1038/nature13069.

Cvejić, J. H., and Rohmer, M. (2000). CO2 as main carbon source for isoprenoid biosynthesis via the mevalonate-independent methylerythritol 4-phosphate route in the marine diatoms Phaeodactylum tricornutum and Nitzschia ovalis. *Phytochemistry*. doi:10.1016/S0031-9422(99)00465-3.

Daboussi, F., Leduc, S., Maréchal, A., Dubois, G., Guyot, V., Perez-Michaut, C., et al. (2014). Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology. *Nat. Commun.* doi:10.1038/ncomms4831.

Dai, L., Chen, K., Youngren, B., Kulina, J., Yang, A., Guo, Z., et al. (2016). Cytoplasmic Drosha activity generated by alternative splicing. *Nucleic Acids Res.* doi:10.1093/nar/gkw668.

Dalmay, T., Hamilton, A., Rudd, S., Angell, S., and Baulcombe, D. C. (2000). An RNA-dependent RNA polymerase gene in arabidopsis is required for posttranscriptional gene silencing mediated by a transgene but not by a virus. *Cell*. doi:10.1016/S0092-8674(00)80864-8.

De Riso, V., Raniello, R., Maumus, F., Rogato, A., Bowler, C., and Falciatore, A. (2009). Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic Acids Res.* 37, e96–e96. doi:10.1093/nar/gkp448.

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., et al. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*. doi:10.1038/nature09886.

Dhir, A., Dhir, S., Proudfoot, N. J., and Jopling, C. L. (2015). Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs. *Nat. Struct. Mol. Biol.* doi:10.1038/nsmb.2982.

Di Dato, V., Musacchia, F., Petrosino, G., Patil, S., Montresor, M., Sanges, R., et al. (2015). Transcriptome sequencing of three Pseudo-nitzschia species reveals comparable gene sets and the presence of Nitric Oxide Synthase genes in diatoms. *Sci. Rep.* doi:10.1038/srep12329.

Ding, Y., Li, H., Chen, L.-L., and Xie, K. (2016). Recent Advances in Genome Editing Using CRISPR/Cas9. *Front. Plant Sci.* 7, 703. doi:10.3389/fpls.2016.00703.

Doudna, J. A., and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science (80-.).* 346, 1258096. doi:10.1126/science.1258096.

Doyle, M., Jaskiewicz, L., and Filipowicz, W. (2012). *Dicer Proteins and Their Role in Gene Silencing Pathways*. 1st ed. Elsevier Inc. doi:10.1016/B978-0-12-404741-9.00001-5.

Drinnenberg, I. A., Weinberg, D. E., Xie, K. T., Mower, J. P., Wolfe, K. H., Fink, G. R., et al. (2009). RNAi in budding yeast. *Science (80-.).* 326, 544–550. doi:10.1126/science.1176945.RNAi.

Dueck, A., Evers, M., Henz, S. R., Unger, K., Eichner, N., Merkl, R., et al. (2016). Gene silencing pathways found in the green alga Volvox carteri reveal insights into evolution and origins of small RNA systems in plants. *BMC Genomics* 17. doi:10.1186/s12864-016-3202-4.

Dufourc, E. J. (2008). Sterols and membrane dynamics. *J. Chem. Biol.* doi:10.1007/s12154-008-0010-6.

Fabris, M., Matthijs, M., Carbonelle, S., Moses, T., Pollier, J., Dasseville, R., et al. (2014a). Tracking the sterol biosynthesis pathway of the diatom Phaeodactylum tricornutum. *New Phytol.* doi:10.1111/nph.12917.

Fabris, M., Matthijs, M., Carbonelle, S., Moses, T., Pollier, J., Dasseville, R., et al. (2014b). Tracking the sterol biosynthesis pathway of the diatom Phaeodactylum tricornutum. *New Phytol.* 204, 521–535. doi:10.1111/nph.12917.

Fabris, M., Matthijs, M., Rombauts, S., Vyverman, W., Goossens, A., and Baart, G. J. E. (2012). The metabolic blueprint of Phaeodactylum tricornutum reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *Plant J.* doi:10.1111/j.1365-313X.2012.04941.x.

Falciatore, A., and Bowler, C. (2002). R Evealing the M Olecular S Ecrets of M Arine D Iatoms . *Annu. Rev. Plant Biol.* 53, 109–130. doi:10.1146/annurev.arplant.53.091701.153921.

Falciatore, A., Casotti, R., Leblanc, C., Abrescia, C., and Bowler, C. (1999a). Transformation of Nonselectable Reporter Genes in Marine Diatoms. *Mar.*

Biotechnol. (NY). 1, 239–251. doi:10.1007/PL00011773.

Falciatore, A., Casotti, R., Leblanc, C., Abrescia, C., and Bowler, C. (1999b). Transformation of Nonselectable Reporter Genes in Marine Diatoms. 239–251.

Ferriols, V. M. E. N., Yaginuma-Suzuki, R., Fukunaga, K., Kadono, T., Adachi, M., Matsunaga, S., et al. (2017). An exception among diatoms: unique organization of genes involved in isoprenoid biosynthesis in Rhizosolenia setigera CCMP 1694. *Plant J.* doi:10.1111/tpj.13719.

Ferriols, V. M. E. N., Yaginuma, R., Adachi, M., Takada, K., Matsunaga, S., and Okada, S. (2015). Cloning and characterization of farnesyl pyrophosphate synthase from the highly branched isoprenoid producing diatom Rhizosolenia setigera. *Sci. Rep.* doi:10.1038/srep10246.

Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: Are the answers in sight? *Nat. Rev. Genet.* doi:10.1038/nrg2290.

Filippov, V., Solovyev, V., Filippova, M., and Gill, S. S. (2000). A novel type of RNase III family proteins in eukaryotes. *Gene* 245, 213–221. doi:10.1016/S0378-1119(99)00571-5.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans. *Nature*. doi:10.1038/35888.

Flori, S., Jouneau, P.-H., Bailleul, B., Gallet, B., Estrozi, L. F., Moriscot, C., et al. (2017). Plastid thylakoid architecture optimizes photosynthesis in diatoms. *Nat. Commun.* 8, 15885. doi:10.1038/ncomms15885.

Fortunato, A. E., Jaubert, M., Enomoto, G., Bouly, J. P., Raniello, R., Thaler, M., et al. (2016). Diatom phytochromes reveal the existence of far-red-light-based sensing in the ocean. *Plant Cell*. doi:10.1105/tpc.15.00928.

Francia, S., Michelini, F., Saxena, A., Tang, D., De Hoon, M., Anelli, V., et al. (2012). Site-specific DICER and DROSHA RNA products control the DNA-damage response. *Nature*. doi:10.1038/nature11179.

Frank, F., Hauver, J., Sonenberg, N., and Nagar, B. (2012). Arabidopsis Argonaute MID domains use their nucleotide specificity loop to sort small RNAs. *EMBO J.*

doi:10.1038/emboj.2012.204.

Frank, F., Sonenberg, N., and Nagar, B. (2010). Structural basis for 5′-nucleotide base-specific recognition of guide RNA by human AGO2. *Nature*. doi:10.1038/nature09039.

Fukudome, A., and Fukuhara, T. (2017). Plant dicer-like proteins: double-stranded RNA-cleaving enzymes for small RNA biogenesis. *J. Plant Res.* doi:10.1007/s10265-016-0877-1.

Fulci, V., and Macino, G. (2007). Quelling: post-transcriptional gene silencing guided by small RNAs in Neurospora crassa. *Curr. Opin. Microbiol.* doi:10.1016/j.mib.2007.03.016.

Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., et al. (2011). Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.* 39, D141–D145. doi:10.1093/nar/gkq1129.

George, K. W., Alonso-Gutierrez, J., Keasling, J. D., and Lee, T. S. (2015). Isoprenoid drugs, biofuels, and chemicals—artemisinin, farnesene, and beyond. *Adv. Biochem. Eng. Biotechnol.* doi:10.1007/10_2014_288.

Głów, D., Kurkowska, M., Czarnecka, J., Szczepaniak, K., Pianka, D., Kappert, V., et al. (2016). Identification of protein structural elements responsible for the diversity of sequence preferences among Mini-III RNases. *Sci. Rep.* 6, 38612. doi:10.1038/srep38612.

Goriaux, C., Desset, S., Renaud, Y., Vaury, C., and Brasset, E. (2014). Transcriptional properties and splicing of the flamenco piRNA cluster. *EMBO Rep.* doi:10.1002/embr.201337898.

Goutte, A., Cherel, Y., Houssais, M.-N., Klein, V., Ozouf-Costaz, C., Raccurt, M., et al. (2013). Diatom-Specific Highly Branched Isoprenoids as Biomarkers in Antarctic Consumers. *PLoS One* 8, e56504. doi:10.1371/journal.pone.0056504.

Grandbastien, M. A. (1998). Activation of plant retrotransposons under stress conditions. *Trends Plant Sci.* doi:10.1016/S1360-1385(98)01232-1.

Grivna, S. T., Beyret, E., Wang, Z., and Lin, H. (2006). A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev.* doi:10.1101/gad.1434406.

Gromak, N., Dienstbier, M., Macias, S., Plass, M., Eyras, E., Cáceres, J. F., et al. (2013).

Drosha regulates gene expression independently of RNA cleavage function. *Cell Rep.* doi:10.1016/j.celrep.2013.11.032.

Guillard, R. R. L. (1975). "Culture of Phytoplankton for Feeding Marine Invertebrates," in *Culture of Marine Invertebrate Animals* (Boston, MA: Springer US), 29–60. doi:10.1007/978-1-4615-8714-9_3.

Häder, D. P., Kumar, H. D., Smith, R. C., and Worrest, R. C. (2007). Effects of solar UV radiation on aquatic ecosystems and interactions with climate change. *Photochem. Photobiol. Sci.* doi:10.1039/b700020k.

Hamilton, A. J., and Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science (80-.).* doi:10.1126/science.286.5441.950.

Hammond, S. M., Bernstein, E., Beach, D., and Hannon, G. J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*. doi:10.1038/35005107.

Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H., and Kim, V. N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev.* doi:10.1101/gad.1262504.

Han, J., Lee, Y., Yeom, K. H., Nam, J. W., Heo, I., Rhee, J. K., et al. (2006). Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex. *Cell*. doi:10.1016/j.cell.2006.03.043.

Han, J., Pedersen, J. S., Kwon, S. C., Belair, C. D., Kim, Y. K., Yeom, K. H., et al. (2009). Posttranscriptional Crossregulation between Drosha and DGCR8. *Cell*. doi:10.1016/j.cell.2008.10.053.

Hartig, J. V., Tomari, Y., and Förstemann, K. (2007). piRNAs--the ancient hunters of genome invaders. *Genes Dev.* 21, 1707–13. doi:10.1101/gad.1567007.

Haussecker, D., Huang, Y., Lau, A., Parameswaran, P., Fire, A. Z., and Kay, M. A. (2010). Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA*. doi:10.1261/rna.2000810.

Havens, M. A., Reich, A. A., and Hastings, M. L. (2014). Drosha Promotes Splicing of a Pre-microRNA-like Alternative Exon. *PLoS Genet.* doi:10.1371/journal.pgen.1004312.

He, L., and Hannon, G. J. (2004). MicroRNAs: Small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* doi:10.1038/nrg1379.

Hempel, F., Bozarth, A. S., Lindenkamp, N., Klingl, A., Zauner, S., Linne, U., et al. (2011). Microalgae as bioreactors for bioplastic production. *Microb. Cell Fact.* doi:10.1186/1475-2859-10-81.

Henderson, I. R., Zhang, X., Lu, C., Johnson, L., Meyers, B. C., Green, P. J., et al. (2006). Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nat. Genet.* 38, 721–725. doi:10.1038/ng1804.

Heras, S. R., MacIas, S., Plass, M., Fernandez, N., Cano, D., Eyras, E., et al. (2013). The Microprocessor controls the activity of mammalian retrotransposons. *Nat. Struct. Mol. Biol.* doi:10.1038/nsmb.2658.

Hinas, A., Reimegård, J., Wagner, E. G. H., Nellen, W., Ambros, V. R., and Söderbom, F. (2007). The small RNA repertoire of Dictyostelium discoideum and its regulation by components of the RNAi pathway. *Nucleic Acids Res.* doi:10.1093/nar/gkm707.

Hockin, N. L., Mock, T., Mulholland, F., Kopriva, S., and Malin, G. (2012). The response of diatom central carbon metabolism to nitrogen starvation is different from that of green algae and higher plants. *Plant Physiol.* doi:10.1104/pp.111.184333.

Hopes, A., Nekrasov, V., Kamoun, S., and Mock, T. (2016a). Editing of the urease gene by CRISPR---Cas in the diatom Thalassiosira pseudonana. doi:10.1101/062026.

Hopes, A., Nekrasov, V., Kamoun, S., and Mock, T. (2016b). Editing of the urease gene by CRISPR-Cas in the diatom Thalassiosira pseudonana. doi:10.1101/062026.

Horwich, M. D., Li, C., Matranga, C., Vagin, V., Farley, G., Wang, P., et al. (2007). The Drosophila RNA Methyltransferase, DmHen1, Modifies Germline piRNAs and Single-Stranded siRNAs in RISC. *Curr. Biol.* doi:10.1016/j.cub.2007.06.030.

Howell, M. D., Fahlgren, N., Chapman, E. J., Cumbie, J. S., Sullivan, C. M., Givan, S. A., et al. (2007). Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell* 19, 926–942. doi:10.1105/tpc.107.050062.

Huang, A., He, L., and Wang, G. (2011). Identification and characterization of microRNAs from Phaeodactylum tricornutum by high-throughput sequencing and bioinformatics analysis. *BMC Genomics*. doi:10.1186/1471-2164-12-337.

Huang, W., and Daboussi, F. (2017a). Genetic and metabolic engineering in diatoms.

Huang, W., and Daboussi, F. (2017b). Genetic and metabolic engineering in diatoms. *Philos. Trans. R. Soc. B Biol. Sci.* 372. doi:10.1098/rstb.2016.0411.

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755. doi:10.1093/bioinformatics/17.8.754.

Huysman, M. J. J., Fortunato, A. E., Matthijs, M., Costa, B. S., Vanderhaeghen, R., van den Daele, H., et al. (2013). AUREOCHROME1a-mediated induction of the diatom-specific cyclin dsCYC2 controls the onset of cell division in diatoms (Phaeodactylum tricornutum). *Plant Cell*. doi:10.1105/tpc.112.106377.

Huysman, M. J. J., Martens, C., Vandepoele, K., Gillard, J., Rayko, E., Heijde, M., et al. (2010). Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome Biol.* doi:10.1186/gb-2010-11-2-r17.

Ifuku, K., Yan, D., Miyahara, M., Inoue-Kashino, N., Yamamoto, Y. Y., and Kashino, Y. (2015). A stable and efficient nuclear transformation system for the diatom Chaetoceros gracilis. *Photosynth. Res.* doi:10.1007/s11120-014-0048-y.

Ignea, C., Pontini, M., Motawia, M. S., Maffei, M. E., Makris, A. M., and Kampranis, S. C. (2018). Synthesis of 11-carbon terpenoids in yeast using protein and metabolic engineering. *Nat. Chem. Biol.* doi:10.1038/s41589-018-0166-5.

Ignea, C., Trikka, F. A., Kourtzelis, I., Argiriou, A., Kanellis, A. K., Kampranis, S. C., et al. (2012). Positive genetic interactors of HMG2 identify a new set of genetic perturbations for improving sesquiterpene production in Saccharomyces cerevisiae. *Microb. Cell Fact.* 11, 162. doi:10.1186/1475-2859-11-162.

Jaskiewicz, L., and Filipowicz, W. (2008). “Role of Dicer in Posttranscriptional RNA Silencing,” in (Springer, Berlin, Heidelberg), 77–97. doi:10.1007/978-3-540-75157-1_4.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E.

(2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (80-.).* doi:10.1126/science.1225829.

Johanson, T. M., Keown, A. A., Cmero, M., Yeo, J. H. C., Kumar, A., Lew, A. M., et al. (2015). Drosha controls dendritic cell development by cleaving messenger RNAs encoding inhibitors of myelopoiesis. *Nat. Immunol.* doi:10.1038/ni.3293.

Johanson, T. M., Lew, A. M., and Chong, M. M. W. (2013). MicroRNA-independent roles of the RNase III enzymes Drosha and Dicer. *Open Biol.* doi:10.1098/rsob.130144.

Jorgensen, R. A., Cluster, P. D., English, J., Que, Q., and Napoli, C. A. (1996). Chalcone synthase cosuppression phenotypes in petunia flowers: Comparison of sense vs. antisense constructs and single-copy vs. complex T-DNA sequences. *Plant Mol. Biol.* doi:10.1007/BF00040715.

Kadener, S., Rodriguez, J., Abruzzi, K. C., Khodor, Y. L., Sugino, K., Marr, M. T., et al. (2009). Genome-wide identification of targets of the drosha-pasha/DGCR8 complex. *RNA.* doi:10.1261/rna.1319309.

Kadono, T., Miyagawa-Yamaguchi, A., Kira, N., Tomaru, Y., Okami, T., Yoshimatsu, T., et al. (2015). Characterization of marine diatom-infecting virus promoters in the model diatom Phaeodactylum tricornutum. *Sci. Rep.* 5, 1–13. doi:10.1038/srep18708.

Karas, B. J., Diner, R. E., Lefebvre, S. C., Mcquaid, J., Phillips, A. P. R., Noddings, C. M., et al. (2015). conjugation. *Nat. Commun.* 6, 1–10. doi:10.1038/ncomms7925.

Katsarou, K., Mavrothalassiti, E., Dermauw, W., Van Leeuwen, T., and Kalantidis, K. (2016). Combined Activity of DCL2 and DCL3 Is Crucial in the Defense against Potato Spindle Tuber Viroid. *PLOS Pathog.* 12, e1005936. doi:10.1371/journal.ppat.1005936.

Katsarou, K., Mitta, E., Bardani, E., Oulas, A., Dadami, E., and Kalantidis, K. (2019). DCL-suppressed *Nicotiana benthamiana* plants: valuable tools in research and biotechnology. *Mol. Plant Pathol.* 20, 432–446. doi:10.1111/mpp.12761.

Kaur, S., and Spillane, C. (2015). Reduction in carotenoid levels in the marine diatom Phaeodactylum tricornutum by artificial microRNAs targeted against the endogenous phytoene synthase gene. *Mar. Biotechnol. (NY).* 17, 1–7.

doi:10.1007/s10126-014-9593-9.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., et al. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol.* 12. doi:10.1371/journal.pbio.1001889.

Kim, B., Jeong, K., and Kim, V. N. (2017). Genome-wide Mapping of DROSHA Cleavage Sites on Primary MicroRNAs and Noncanonical Substrates. *Mol. Cell*. doi:10.1016/j.molcel.2017.03.013.

Kim, K. N., Heo, S. J., Yoon, W. J., Kang, S. M., Ahn, G., Yi, T. H., et al. (2010). Fucoxanthin inhibits the inflammatory response by suppressing the activation of NF-κB and MAPKs in lipopolysaccharide-induced RAW 264.7 macrophages. *Eur. J. Pharmacol.* doi:10.1016/j.ejphar.2010.09.032.

Kim, V. N. (2005). Small RNAs: Classification, biogenesis, and function. *Mol. Cells*.

Kooistra, W. H. C. F., Gersonde, R., Medlin, L. K., and Mann, D. G. (2007). "The Origin and Evolution of the Diatoms. Their Adaptation to a Planktonic Existence.," in *Evolution of Primary Producers in the Sea* doi:10.1016/B978 012370518-1/50012-6.

Kramerov, D. A., and Vassetzky, N. S. (2011). Origin and evolution of SINEs in eukaryotic genomes. *Heredity (Edinb).* doi:10.1038/hdy.2011.43.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi:10.1093/molbev/msw054.

Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., Ikawa, M., et al. (2008). DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes Dev.* doi:10.1101/gad.1640708.

Kwon, S. C., Nguyen, T. A., Choi, Y.-G., Jo, M. H., Hohng, S., Kim, V. N., et al. (2016a). Structure of Human DROSHA. *Cell* 164, 81–90. doi:10.1016/J.CELL.2015.12.019.

Kwon, S. C., Nguyen, T. A., Choi, Y. G., Jo, M. H., Hohng, S., Kim, V. N., et al.

(2016b). Structure of Human DROSHA. *Cell*. doi:10.1016/j.cell.2015.12.019.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science (80-.).* doi:10.1126/science.1064921.

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science (80-.).* doi:10.1126/science.1065062.

Lau, N. C., Seto, A. G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D. P., et al. (2006). Characterization of the piRNA complex from rat testes. *Science (80-.).* doi:10.1126/science.1130164.

Lauersen, K. J. (2019). Eukaryotic microalgae as hosts for light-driven heterologous isoprenoid production. *Planta*. doi:10.1007/s00425-018-3048-x.

Lee, D., Nam, J. W., and Shin, C. (2017). DROSHA targets its own transcript to modulate alternative splicing. *RNA*. doi:10.1261/rna.059808.116.

Lee, D., and Shin, C. (2018). Emerging roles of DROSHA beyond primary microRNA processing. *RNA Biol.* 15, 186–193. doi:10.1080/15476286.2017.1405210.

Lee, H. C., Chang, S. S., Choudhary, S., Aalto, A. P., Maiti, M., Bamford, D. H., et al. (2009a). QiRNA is a new type of small interfering RNA induced by DNA damage. *Nature*. doi:10.1038/nature08041.

Lee, R. C., and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science (80-.).* doi:10.1126/science.1065329.

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*. doi:10.1016/0092-8674(93)90529-Y.

Lee, S. R., and Collins, K. (2007). Physical and functional coupling of RNA-dependent RNA polymerase and Dicer in the biogenesis of endogenous siRNAs. *Nat. Struct. Mol. Biol.* doi:10.1038/nsmb1262.

Lee, Y. S., Nakahara, K., Pham, J. W., Kim, K., He, Z., Sontheimer, E. J., et al. (2004). Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell*. doi:10.1016/S0092-8674(04)00261-2.

Lee, Y. S., Shibata, Y., Malhotra, A., and Dutta, A. (2009b). A novel class of small

RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.* doi:10.1101/gad.1837609.

Leibman, D., Kravchik, M., Wolf, D., Haviv, S., Weissberg, M., Ophir, R., et al. (2018). Differential expression of cucumber RNA-dependent RNA polymerase 1 genes during antiviral defence and resistance. *Mol. Plant Pathol.* doi:10.1111/mpp.12518.

Leonardo, S., Prieto-Simón, B., and Campàs, M. (2016). Past, present and future of diatoms in biosensing. *TrAC - Trends Anal. Chem.* doi:10.1016/j.trac.2015.11.022.

Lescot, M., Hingamp, P., Kojima, K. K., Villar, E., Romac, S., Veluchamy, A., et al. (2016). Reverse transcriptase genes are highly abundant and transcriptionally active in marine plankton assemblages. *ISME J.* doi:10.1038/ismej.2015.192.

Levitan, O., Dinamarca, J., Zelzion, E., Gorbunov, M. Y., and Falkowski, P. G. (2015). An RNA interference knock-down of nitrate reductase enhances lipid biosynthesis in the diatom Phaeodactylum tricornutum. *Plant J.* 84, 963–973. doi:10.1111/tpj.13052.

LEWIN, J. C. (1958). The taxonomic position of Phaeodactylum tricornutum. *J. Gen. Microbiol.* doi:10.1099/00221287-18-2-427.

Lezin, G., Kosaka, Y., Yost, H. J., Kuehn, M. R., and Brunelli, L. (2011). A One-Step Miniprep for the Isolation of Plasmid DNA and Lambda Phage Particles. 6. doi:10.1371/journal.pone.0023457.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324.

Li, J., Yang, Z., Yu, B., Liu, J., and Chen, X. (2005). Methylation protects miRNAs and siRNAs from a 3′-end uridylation activity in Arabidopsis. *Curr. Biol.* doi:10.1016/j.cub.2005.07.029.

Lin, H., and Spradling, A. C. (1997). A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. *Development*.

Lin, Y. T., and Sullivan, C. S. (2011). Expanding the role of Drosha to the regulation of viral gene expression. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1105799108.

Lindbo, J. A., Silva-Rosales, L., Proebsting, W. M., and Dougherty, W. G. (1993). Induction of a highly specific antiviral state in transgenic plants: Implications for

regulation of gene expression and virus resistance. *Plant Cell*. doi:10.2307/3869691.

Lingel, A., Simon, B., Izaurralde, E., and Sattler, M. (2003). Structure and nucleic-acid binding of the Drosophila Argonaute 2 PAZ domain. *Nature*. doi:10.1038/nature02123.

Link, S., Grund, S. E., and Diederichs, S. (2016). Alternative splicing affects the subcellular localization of Drosha. *Nucleic Acids Res.* doi:10.1093/nar/gkw400.

Liu, X., Hempel, F., Stork, S., Bolte, K., Moog, D., Heimerl, T., et al. (2016). Addressing various compartments of the diatom model organism Phaeodactylum tricornutum via sub-cellular marker proteins. *Algal Res.* 20, 249–257. doi:10.1016/j.algal.2016.10.018.

Llave, C., Kasschau, K. D., Rector, M. A., and Carrington, J. C. (2002). Endogenous and silencing-associated small RNAs in plants. *Plant Cell*. doi:10.1105/tpc.003210.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8.

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.* 25, 955–964. doi:10.1093/nar/25.5.955.

Lu, Y. Z. and Liu, J. (2010). In silico identification of MicroRNAs and their targets in diatoms. *African J. Microbiol. Res.*

Lund, E., Güttinger, S., Calado, A., Dahlberg, J. E., and Kutay, U. (2004). Nuclear Export of MicroRNA Precursors. *Science (80-.).* doi:10.1126/science.1090599.

Ma, J. B., Ye, K., and Patel, D. J. (2004). Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature*. doi:10.1038/nature02519.

MacIas, S., Plass, M., Stajuda, A., Michlewski, G., Eyras, E., and Cáceres, J. F. (2012). DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. *Nat. Struct. Mol. Biol.* doi:10.1038/nsmb.2344.

Macrae, I. J., and Doudna, J. A. (2007). Ribonuclease revisited : structural insights into ribonuclease III family enzymes. 138–145. doi:10.1016/j.sbi.2006.12.002.

Macrae, I. J., Macrae, I. J., Zhou, K., Li, F., Repic, A., Brooks, A. N., et al. (2006). Structural Basis for Double-Stranded RNA Processing by Dicer. 311.

doi:10.1126/science.1121638.

MacRae, I. J., Zhou, K., and Doudna, J. A. (2007). Structural determinants of RNA recognition and cleavage by Dicer. *Nat. Struct. Mol. Biol.* doi:10.1038/nsmb1293.

Maeda, H. (2015). Nutraceutical effects of fucoxanthin for obesity and diabetes therapy: A review. *J. Oleo Sci.* doi:10.5650/jos.ess14226.

Maher, S., Kumeria, T., Aw, M. S., and Losic, D. (2018). Diatom Silica for Biomedical Applications: Recent Progress and Advances. *Adv. Healthc. Mater.* doi:10.1002/adhm.201800552.

Maheswari, U., Mock, T., Armbrust, E. V., and Bowler, C. (2009). Update of the Diatom EST Database : a new tool for digital transcriptomics. 37, 1001–1005. doi:10.1093/nar/gkn905.

Malone, C. D., and Hannon, G. J. (2009). Small RNAs as Guardians of the Genome. *Cell*. doi:10.1016/j.cell.2009.01.045.

Marchetti, A., Lundholm, N., Kotaki, Y., Hubbard, K., Harrison, P. J., and Virginia Armbrust, E. (2008). IDENTIFICATION AND ASSESSMENT OF DOMOIC ACID PRODUCTION IN OCEANIC *PSEUDO-NITZSCHIA* (BACILLARIOPHYCEAE) FROM IRON-LIMITED WATERS IN THE NORTHEAST SUBARCTIC PACIFIC. *J. Phycol.* 44, 650–661. doi:10.1111/j.1529-8817.2008.00526.x.

Marinaro, F., Marzi, M. J., Hoffmann, N., Amin, H., Pelizzoli, R., Niola, F., et al. (2017). MicroRNA-independent functions of DGCR8 are essential for neocortical development and TBR1 expression. *EMBO Rep.* doi:10.15252/embr.201642800.

Marron, A. O., Ratcliffe, S., Wheeler, G. L., Goldstein, R. E., King, N., Not, F., et al. (2016). The evolution of silicon transport in eukaryotes. *Mol. Biol. Evol.* doi:10.1093/molbev/msw209.

Martin, L. J. (2015). Fucoxanthin and its metabolite fucoxanthinol in cancer prevention and treatment. *Mar. Drugs*. doi:10.3390/md13084784.

Massé, G., Belt, S. T., Allard, W. G., Lewis, C. A., Wakeham, S. G., and Rowland, S. J. (2004a). Occurrence of novel monocyclic alkenes from diatoms in marine particulate matter and sediments. *Org. Geochem.* doi:10.1016/j.orggeochem.2004.03.004.

Massé, G., Belt, S. T., Rowland, S. J., and Rohmer, M. (2004b). Isoprenoid biosynthesis in the diatoms Rhizosolenia setigera (Brightwell) and Haslea ostrearia (Simonsen). *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.0400902101.

Materna, A. C., Sturm, S., Kroth, P. G., and Lavaud, J. (2009). First induced plastid genome mutations in an alga with secondary plastids: psbA mutations in the diatom phaeodactylum tricornutum (bacillariophyceae) reveal consequences on the regulation of photosynthesis 1. *J. Phycol.* doi:10.1111/j.1529-8817.2009.00711.x.

Matzke, M. A., and Mosher, R. A. (2014). RNA-directed DNA methylation : an epigenetic pathway of increasing complexity. *Nat. Publ. Gr.* doi:10.1038/nrg3683.

Matzke, M., Kanno, T., Daxinger, L., Huettel, B., and Matzke, A. J. (2009). RNA-mediated chromatin-based silencing in plants. *Curr. Opin. Cell Biol.* doi:10.1016/j.ceb.2009.01.025.

Maumus, F., Allen, A. E., Mhiri, C., Hu, H., Jabbari, K., Vardi, A., et al. (2009). Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics* 10, 1–19. doi:10.1186/1471-2164-10-624.

Maumus, F., Rabinowicz, P., Bowler, C., Rivarola, M., Inserm, U., Nacional, I., et al. (2011). Stemming Epigenetics in Marine Stramenopiles. 357–370.

McGuire, A. M., Pearson, M. D., Neafsey, D. E., and Galagan, J. E. (2008). Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol.* doi:10.1186/gb-2008-9-3-r50.

Meadows, A. L., Hawkins, K. M., Tsegaye, Y., Antipov, E., Kim, Y., Raetz, L., et al. (2016). Rewriting yeast central carbon metabolism for industrial isoprenoid production. *Nature*. doi:10.1038/nature19769.

Meister, G., and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. *Nature*. doi:10.1038/nature02873.

Mello, C. C., and Conte, D. (2004). Revealing the world of RNA interference. *Nature*. doi:10.1038/nature02872.

Mochizuki, K., and Gorovsky, M. A. (2005). A Dicer-like protein in Tetrahymena has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes Dev.* doi:10.1101/gad.1265105.

Mock, T., and Kroon, B. M. A. (2002). Photosynthetic energy conversion under extreme

conditions - I: Important role of lipids as structural modulators and energy sink under N-limited growth in Antarctic sea ice diatoms. *Phytochemistry*. doi:10.1016/S0031-9422(02)00216-9.

Mock, T., Otillar, R. P., Strauss, J., Mcmullan, M., Paajanen, P., Schmutz, J., et al. (2017). Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus The Southern Ocean houses a diverse and productive community of organisms. *Nat. Publ. Gr*. doi:10.1038/nature20803.

Molnár, A., Schwach, F., Studholme, D. J., Thuenemann, E. C., and Baulcombe, D. C. (2007). miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature*. doi:10.1038/nature05903.

Motamedi, M. R., Verdel, A., Colmenares, S. U., Gerber, S. A., Gygi, S. P., and Moazed, D. (2004). Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. *Cell*. doi:10.1016/j.cell.2004.11.034.

Mourrain, P., Béclin, C., Elmayan, T., Feuerbach, F., Godon, C., Morel, J. B., et al. (2000). Arabidopsis SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. *Cell*. doi:10.1016/S0092-8674(00)80863-6.

Moustafa, A., Beszteri, B., Maier, U. G., Bowler, C., Valentin, K., and Bhattacharya, D. (2009). Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* 324, 1724–6. doi:10.1126/science.1172983.

Mus, F., Toussaint, J. P., Cooksey, K. E., Fields, M. W., Gerlach, R., Peyton, B. M., et al. (2013). Physiological and molecular analysis of carbon source supplementation and pH stress-induced lipid accumulation in the marine diatom Phaeodactylum tricornutum. *Appl. Microbiol. Biotechnol.* doi:10.1007/s00253-013-4747-7.

Muto, M., Fukuda, Y., Nemoto, M., Yoshino, T., Matsunaga, T., and Tanaka, T. (2013). Establishment of a Genetic Transformation System for the Marine Pennate Diatom Fistulifera sp. Strain JPCC DA0580-A High Triglyceride Producer. *Mar. Biotechnol.* doi:10.1007/s10126-012-9457-0.

Napoli, C., Lemieux, C., and Jorgensen, R. (1990). Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. *Plant Cell*. doi:10.2307/3869076.

Nguyen, T. A., Jo, M. H., Choi, Y. G., Park, J., Kwon, S. C., Hohng, S., et al. (2015).

Functional anatomy of the human microprocessor. *Cell*. doi:10.1016/j.cell.2015.05.010.

Nicholson, A. W. (1999). Function, mechanism and regulation of bacterial ribonucleases. *FEMS Microbiol. Rev.* doi:10.1111/j.1574-6976.1999.tb00405.x.

Nicolas, F. E., Moxon, S., de Haro, J. P., Calo, S., Grigoriev, I. V., Torres-MartÍnez, S., et al. (2010). Endogenous short RNAs generated by Dicer 2 and RNA-dependent RNA polymerase 1 regulate mRNAs in the basal fungus Mucor circinelloides. *Nucleic Acids Res.* doi:10.1093/nar/gkq301.

Nymark, M., Sharma, A. K., Sparstad, T., Bones, A. M., and Winge, P. (2016a). A CRISPR/Cas9 system adapted for gene editing in marine algae. *Sci. Rep.* 6, 24951. doi:10.1038/srep24951.

Nymark, M., Sharma, A. K., Sparstad, T., Bones, A. M., Winge, P., Field, C. B., et al. (2016b). A CRISPR/Cas9 system adapted for gene editing in marine algae. *Sci. Rep.* 6, 24951. doi:10.1038/srep24951.

Obata, T., Fernie, A. R., and Nunes-Nesi, A. (2013). The central carbon and energy metabolism of marine diatoms. *Metabolites*. doi:10.3390/metabo3020325.

Parent, J. S., Bouteiller, N., Elmayan, T., and Vaucheret, H. (2015). Respective contributions of Arabidopsis DCL2 and DCL4 to RNA silencing. *Plant J.* doi:10.1111/tpj.12720.

Pattison, D. I., and Davies, M. J. (2006). Actions of ultraviolet light on cellular structures. *EXS*, 131–157. doi:10.1007/3-7643-7378-4_6.

Pollak, B., Matute, T., Nunez, I., Cerda, A., Lopez, C., Vargas, V., et al. (2019). Universal Loop assembly (uLoop): open, efficient, and species-agnostic DNA fabrication. *bioRxiv*, 744854. doi:10.1101/744854.

Pompey, J. M., Foda, B., and Singh, U. (2015a). A Single RNaseIII Domain Protein from Entamoeba histolytica Has dsRNA Cleavage Activity and Can Help Mediate RNAi Gene Silencing in a Heterologous System. 1–21. doi:10.1371/journal.pone.0133740.

Pompey, J. M., Foda, B., and Singh, U. (2015b). A Single RNaseIII Domain Protein from Entamoeba histolytica Has dsRNA Cleavage Activity and Can Help Mediate RNAi Gene Silencing in a Heterologous System. 1–21. doi:10.1371/journal.pone.0133740.

Pompey, J. M., Morf, L., and Singh, U. (2014). RNAi pathway genes are resistant to

small RNA mediated gene silencing in the protozoan parasite Entamoeba histolytica. *PLoS One* 9, 1–9. doi:10.1371/journal.pone.0106477.

Poulsen, N., and Kröger, N. (2005). A new molecular tool for transgenic diatoms: Control of mRNA and protein biosynthesis by an inducible promoter-terminator cassette. *FEBS J.* doi:10.1111/j.1742-4658.2005.04760.x.

Pytlik, N., and Brunner, E. (2018). Diatoms as potential green nanocomposite and nanoparticle synthesizers: Challenges, prospects, and future materials applications. *MRS Commun.* doi:10.1557/mrc.2018.34.

Que, Q., and Jorgensen, R. A. (1998). Homology-based control of gene expression patterns in transgenic petunia flowers. *Dev. Genet.* doi:10.1002/(SICI)1520-6408(1998)22:1<100::AID-DVG10>3.0.CO;2-D.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi:10.1093/bioinformatics/btq033.

Rastogi, A., Maheswari, U., Dorrell, R. G., Rocha, F., Vieira, J., Kustka, A., et al. (2018). Integrative analysis of large scale transcriptome data draws a comprehensive landscape of Phaeodactylum tricornutum genome and evolutionary origin of diatoms. *Sci. Rep.*, 1–14. doi:10.1038/s41598-018-23106-x.

Rastogi, A., Murik, O., Bowler, C., Tirichine, L., Chari, R., Mali, P., et al. (2016). PhytoCRISP-Ex: a web-based and stand-alone application to find specific target sequences for CRISPR/CAS editing. *BMC Bioinformatics* 17, 261. doi:10.1186/s12859-016-1143-1.

RAVEN, J. A. (1987). THE ROLE OF VACUOLES. *New Phytol.* doi:10.1111/j.1469-8137.1987.tb00149.x.

Rayko, E., Maumus, F., Maheswari, U., Jabbari, K., and Bowler, C. (2010). Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. *New Phytol.* doi:10.1111/j.1469-8137.2010.03371.x.

Redko, Y., Bechhofer, D. H., and Condon, C. (2008). Mini-III , an unusual member of the RNase III family of enzymes , catalyses 23S ribosomal RNA maturation in B . subtilis. 68, 1096–1106. doi:10.1111/j.1365-2958.2008.06207.x.

Reinhart, B. J., Slack, F. J., Basson, M., Pasquienelll, A. E., Bettlnger, J. C., Rougvle, A.

E., et al. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*. doi:10.1038/35002607.

Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., and Bartel, D. P. (2002). MicroRNAs in plants. *Genes Dev.* doi:10.1101/gad.1004402.

Riso, V. De, Raniello, R., Maumus, F., Rogato, A., Bowler, C., and Falciatore, A. (2009). Gene silencing in the marine diatom Phaeodactylum tricornutum. 1–12. doi:10.1093/nar/gkp448.

Ro, D. K., Paradise, E. M., Quellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., et al. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*. doi:10.1038/nature04640.

Rogato, A., Richard, H., Sarazin, A., Voss, B., and Navarro, S. C. (2014). The diversity of small non-coding RNAs in the diatom Phaeodactylum tricornutum The diversity of small non-coding RNAs in the diatom Phaeodactylum tricornutum. doi:10.1186/1471-2164-15-698.

Rogers, A. K., Situ, K., Perkins, E. M., and Toth, K. F. (2017). Zucchini-dependent piRNA processing is triggered by recruitment to the cytoplasmic processing machinery. *Genes Dev.* doi:10.1101/gad.303214.117.

Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi:10.1093/bioinformatics/btg180.

Rotondo, G., and Frendewey, D. (1996). Purification and Characterization of the Pac1 Ribonuclease of Schizosaccharomyces Pombe. *Nucleic Acids Res.* 24, 2377–2386. doi:10.1093/nar/24.12.2377.

Rowland, S. J., Belt, S. T., Wraige, E. J., Massé, G., Roussakis, C., and Robert, J. M. (2001). Effects of temperature on polyunsaturation in cytostatic lipids of Haslea ostrearia. *Phytochemistry*. doi:10.1016/S0031-9422(00)00434-9.

Sachindra, N. M., Sato, E., Maeda, H., Hosokawa, M., Niwano, Y., Kohno, M., et al. (2007). Radical scavenging and singlet oxygen quenching activity of marine carotenoid fucoxanthin and its metabolites. *J. Agric. Food Chem.* doi:10.1021/jf071848a.

Saito, K., Nishida, K. M., Mori, T., Kawamura, Y., Miyoshi, K., Nagami, T., et al.

(2006). Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. *Genes Dev.* doi:10.1101/gad.1454806.

Sarthou, G., Timmermans, K. R., Blain, S., and Tréguer, P. (2005). Growth physiology and fate of diatoms in the ocean: A review. *J. Sea Res.* doi:10.1016/j.seares.2004.01.007.

Sathasivam, R., and Ki, J. S. (2018). A review of the biological activities of microalgal carotenoids and their potential use in healthcare and cosmetic industries. *Mar. Drugs*. doi:10.3390/md16010026.

Schiebel, W., Pélissier, T., Riedel, L., Thalmeir, S., Schiebel, R., Kempe, D., et al. (1998). Isolation of an RNA-directed RNA polymerasespecific cDNA clone from tomato. *Plant Cell*. doi:10.1105/tpc.10.12.2087.

Seckbach, J. (Joseph), and Kociolek, J. P. (2011). *The diatom world*. Springer Science + Business Media Available at: https://books.google.gr/books?hl=el&lr=&id=Va35Mtn9VGYC&oi=fnd&pg=PR3&dq=Seckbach,+J.%3B+Kociolek,+P.+The+diatom+world&ots=g8iLiFWsvk&sig=bfhT-SBktsa1aC8SwCuJLqsJQRA&redir_esc=y#v=onepage&q=Seckbach%2C J.%3B Kociolek%2C P. The diatom world&f=false [Accessed December 8, 2019].

Seo, S., Jeon, H., Hwang, S., Jin, E. S., and Chang, K. S. (2015). Development of a new constitutive expression system for the transformation of the diatom Phaeodactylum tricornutum. *Algal Res.* doi:10.1016/j.algal.2015.05.012.

Serif, M., Dubois, G., Finoux, A.-L., Teste, M.-A., Jallet, D., and Daboussi, F. (2018). One-step generation of multiple gene knock-outs in the diatom Phaeodactylum tricornutum by DNA-free genome editing. *Nat. Commun.* 9, 3924. doi:10.1038/s41467-018-06378-9.

Shapiro, J. S., Schmi, S., Aguado, L. C., Sabin, L. R., Yasunaga, A., Shim, J. V., et al. (2014). Drosha as an interferon-independent antiviral factor. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1319635111.

Shi, H., Tschudi, C., and Ullu, E. (2006). An unusual Dicer-like1 protein fuels the RNA interference pathway in Trypanosoma brucei. *RNA*. doi:10.1261/rna.246906.

Siaut, M., Heijde, M., Mangogna, M., Montsant, A., Coesel, S., Allen, A., et al. (2007).

Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. 406, 23–35. doi:10.1016/j.gene.2007.05.022.

Sienski, G., Dönertas, D., and Brennecke, J. (2012). Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell*. doi:10.1016/j.cell.2012.10.040.

Sijen, T., Fleenor, J., Simmer, F., Thijssen, K. L., Parrish, S., Timmons, L., et al. (2001). On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell*. doi:10.1016/S0092-8674(01)00576-1.

Silva, G., Poirot, L., Galetto, R., Smith, J., Montoya, G., Duchateau, P., et al. (2011). Meganucleases and Other Tools for Targeted Genome Engineering: Perspectives and Challenges for Gene Therapy. *Curr. Gene Ther.* doi:10.2174/156652311794520111.

Simon, B., Kirkpatrick, J. P., Eckhardt, S., Reuter, M., Rocha, E. A., Andrade-Navarro, M. A., et al. (2011). Recognition of 2′-o-methylated 3′-end of piRNA by the PAZ domain of a Piwi protein. *Structure*. doi:10.1016/j.str.2010.11.015.

Sinnighe Damsté, J. S., Rampen, S. W., Masse, G., Allard, G. W., Belt, S. T., Robert, J., et al. (2004). The Rise of the Rhizosolenid Diatoms. *Science (80-.).*

Skerratt, J. H., Davidson, A. D., Nichols, P. D., and Mcmeekin, T. A. (1998). Effect of UV-B on lipid content of three antarctic marine phytoplankton. *Phytochemistry* 49, 999–1007. doi:10.1016/S0031-9422(97)01068-6.

Smardon, A., Spoerke, J. M., Stacey, S. C., Klein, M. E., MacKin, N., and Maine, E. M. (2000). EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in C. elegans. *Curr. Biol.* doi:10.1016/S0960-9822(00)00323-7.

Stajich, J. E., Dietrich, F. S., and Roy, S. W. (2007). Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biol.* doi:10.1186/gb-2007-8-10-r223.

Stonik, V. S., and Stonik, I. (2015). Low-molecular-weight metabolites from diatoms: Structures, biological roles and biosynthesis. *Mar. Drugs*. doi:10.3390/md13063672.

Suk, K., Choi, J., Suzuki, Y., Ozturk, S. B., Mellor, J. C., Wong, K. H., et al. (2011). Reconstitution of human RNA interference in budding yeast. *Nucleic Acids Res.* 39.

doi:10.1093/nar/gkq1321.

Swarts, D. C., Makarova, K., Wang, Y., Nakanishi, K., Ketting, R. F., Koonin, E. V., et al. (2014). The evolutionary journey of Argonaute proteins. *Nat. Struct. Mol. Biol.* 21, 743–753. doi:10.1038/nsmb.2879.

Tan, C. P., and Hou, Y. H. (2014). First evidence for the anti-inflammatory activity of fucoxanthin in high-fat-diet-induced obesity in mice and the antioxidant functions in PC12 cells. *Inflammation*. doi:10.1007/s10753-013-9757-1.

Tang, K. F., and Ren, H. (2012). The role of dicer in DNA damage repair. *Int. J. Mol. Sci.* 13, 16769–16778. doi:10.3390/ijms131216769.

Tarver, J. E., Cormier, A., Pinzón, N., Taylor, R. S., Carré, W., Strittmatter, M., et al. (2015). microRNAs and the evolution of complex multicellularity: identification of a large, diverse complement of microRNAs in the brown alga *Ectocarpus*. *Nucleic Acids Res.* 43, 6384–6398. doi:10.1093/nar/gkv578.

Tedetti, M., and Sempéré, R. (2006). Penetration of Ultraviolet Radiation in the Marine Environment. A Review. *Photochem. Photobiol.* doi:10.1562/2005-11-09-ir-733.

Teixeira, F. K., Heredia, F., Sarazin, A., Roudier, F., Boccara, M., Ciaudo, C., et al. (2009). A role for RNAi in the selective correction of DNA methylation defects. *Science (80-.).* doi:10.1126/science.1165313.

Tian, Y., Simanshu, D. K., Ma, J. B., and Patel, D. J. (2011). Structural basis for piRNA 2′-O-methylated 3′-end recognition by Piwi PAZ (Piwi/Argonaute/Zwille) domains. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1017762108.

Tippmann, S., Chen, Y., Siewers, V., and Nielsen, J. (2013). From flavors and pharmaceuticals to advanced biofuels: Production of isoprenoids in Saccharomyces cerevisiae. *Biotechnol. J.* doi:10.1002/biot.201300028.

Tirichine, L., Rastogi, A., and Bowler, C. (2017). ScienceDirect Recent progress in diatom genomics and epigenomics. *Curr. Opin. Plant Biol.* 36, 46–55. doi:10.1016/j.pbi.2017.02.001.

Tomari, Y., and Zamore, P. D. (2005). Perspective: Machines for RNAi. *Genes Dev.* doi:10.1101/gad.1284105.

Tran, R. K., Zilberman, D., de Bustos, C., Ditt, R. F., Henikoff, J. G., Lindroth, A. M., et al. (2005). Chromatin and siRNA pathways cooperate to maintain DNA methylation

of small transposable elements in Arabidopsis. *Genome Biol.* doi:10.1186/gb-2005-6-11-r90.

Treco, D. A., and Lundblad, V. (1993). Preparation of Yeast Media. *Curr. Protoc. Mol. Biol.* 23, 13.1.1-13.1.7. doi:10.1002/0471142727.mb1301s23.

Trentacoste, E. M., Shrestha, R. P., Smith, S. R., Glé, C., Hartmann, A. C., and Hildebrand, M. (2013). Metabolic engineering of lipid catabolism increases microalgal lipid accumulation without compromising growth. doi:10.1073/pnas.1309299110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1309299110.

Tuck, A. C., and Tollervey, D. (2011). RNA in pieces. *Trends Genet.* doi:10.1016/j.tig.2011.06.001.

Vavitsas, K., Fabris, M., and Vickers, C. E. (2018). Terpenoid metabolic engineering in photosynthetic microorganisms. *Genes (Basel).* doi:10.3390/genes9110520.

Vazquez, F., Vaucheret, H., Rajagopalan, R., Lepers, C., Gasciolli, V., Mallory, A. C., et al. (2004). Endogenous trans-acting siRNAs regulate the accumulation of arabidopsis mRNAs. *Mol. Cell*. doi:10.1016/j.molcel.2004.09.028.

Veluchamy, A., Lin, X., Maumus, F., Rivarola, M., Bhavsar, J., Creasy, T., et al. (2013). Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum. *Nat. Commun.* 4. doi:10.1038/ncomms3091.

Veluchamy, A., Lin, X., Maumus, F., Rivarola, M., Bhavsar, J., Creasy, T., et al. (2014). Insights into the role of DNA methylation. doi:10.1038/ncomms3091.

Veluchamy, A., Rastogi, A., Lin, X., Lombard, B., Murik, O., Thomas, Y., et al. (2015). An integrative analysis of post-translational histone modifications in the marine diatom Phaeodactylum tricornutum. *Genome Biol.*, 1–18. doi:10.1186/s13059-015-0671-8.

Vickers, C. E., Williams, T. C., Peng, B., and Cherry, J. (2017). Recent advances in synthetic biology for engineering isoprenoid production in yeast. *Curr. Opin. Chem. Biol.* doi:10.1016/j.cbpa.2017.05.017.

Vitte, C., Fustier, M. A., Alix, K., and Tenaillon, M. I. (2014). The bright side of transposons in crop evolution. *Briefings Funct. Genomics Proteomics*. doi:10.1093/bfgp/elu002.

Voinnet, O. (2008). Use, tolerance and avoidance of amplified RNA silencing by plants. *Trends Plant Sci.* doi:10.1016/j.tplants.2008.05.004.

Volpe, T. A., Kidner, C., Hall, I. M., Teng, G., Grewal, S. I. S., and Martienssen, R. A. (2002). Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science (80-.).* doi:10.1126/science.1074973.

Vranová, E., Coman, D., and Gruissem, W. (2012). Structure and dynamics of the isoprenoid pathway network. in *Molecular Plant* doi:10.1093/mp/sss015.

Wassenegger, M., Heimes, S., Riedel, L., and Sänger, H. L. (1994). RNA-directed de novo methylation of genomic sequences in plants. *Cell*. doi:10.1016/0092-8674(94)90119-8.

Watanabe, T., Cheng, E. C., Zhong, M., and Lin, H. (2015). Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline. *Genome Res.* doi:10.1101/gr.180802.114.

Wei, W., Ba, Z., Gao, M., Wu, Y., Ma, Y., Amiard, S., et al. (2012). A role for small RNAs in DNA double-strand break repair. *Cell*. doi:10.1016/j.cell.2012.03.002.

Weyman, P. D., Beeri, K., Lefebvre, S. C., Rivera, J., Mccarthy, J. K., Heuberger, A. L., et al. (2015). Inactivation of Phaeodactylum tricornutum urease gene using transcription activator-like effector nuclease-based targeted mutagenesis. *Plant Biotechnol. J.* 13, 460–470. doi:10.1111/pbi.12254.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*. doi:10.1016/0092-8674(93)90530-4.

Williamson, C. E., Zepp, R. G., Lucas, R. M., Madronich, S., Austin, A. T., Ballaré, C. L., et al. (2014). Solar ultraviolet radiation in a changing climate. *Nat. Clim. Chang.* doi:10.1038/nclimate2225.

Willmann, M. R., Endres, M. W., Cook, R. T., and Gregory, B. D. (2011). The Functions of RNA-Dependent RNA Polymerases in Arabidopsis. *Arab. B.* doi:10.1199/tab.0146.

Xie, W. H., Zhu, C. C., Zhang, N. S., Li, D. W., Yang, W. D., Liu, J. S., et al. (2014). Construction of Novel Chloroplast Expression Vector and Development of an Efficient Transformation System for the Diatom Phaeodactylum tricornutum. *Mar.*

Biotechnol. doi:10.1007/s10126-014-9570-3.

Xie, Z., Allen, E., Wilken, A., and Carrington, J. C. (2005). DICER-LIKE 4 functions in trans-acting small interfering RNA biogenesis and vegetative phase change in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.0506426102.

Xie, Z., Johansen, L. K., Gustafson, A. M., Kasschau, K. D., Lellis, A. D., Zilberman, D., et al. (2004). Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol.* doi:10.1371/journal.pbio.0020104.

Yang, J., Pan, Y., Bowler, C., Zhang, L., and Hu, H. (2016). Knockdown of phosphoenolpyruvate carboxykinase increases carbon flux to lipid synthesis in Phaeodactylum tricornutum. *Algal Res.* 15, 50–58. doi:10.1016/j.algal.2016.02.004.

Yi, R., Qin, Y., Macara, I. G., and Cullen, B. R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev.* doi:10.1101/gad.1158803.

Yigit, E., Batista, P. J., Bei, Y., Pang, K. M., Chen, C. C. G., Tolia, N. H., et al. (2006). Analysis of the C. elegans Argonaute Family Reveals that Distinct Argonautes Act Sequentially during RNAi. *Cell*. doi:10.1016/j.cell.2006.09.033.

Yoshinaga, R., Niwa-Kubota, M., Matsui, H., and Matsuda, Y. (2014). Characterization of iron-responsive promoters in the marine diatom Phaeodactylum tricornutum. *Mar. Genomics*. doi:10.1016/j.margen.2014.01.005.

Zaslavskaia, L. A., Casey Lippmeier, J., Kroth, P. G., Grossman, A. R., and Apt, K. E. (2000a). Transformation of the diatom Phaeodactylum tricornutum (Bacillariophyceae) with a variety of selectable marker and reporter genes. *J. Phycol.* 36, 379–386. doi:10.1046/j.1529-8817.2000.99164.x.

Zaslavskaia, L. A., Casey Lippmeier, J., Kroth, P. G., Grossman, A. R., and Apt, K. E. (2000b). Transformation of the diatom Phaeodactylum tricornutum (Bacillariophyceae) with a variety of selectable marker and reporter genes. *J. Phycol.* doi:10.1046/j.1529-8817.2000.99164.x.

Zhai, J., Bischof, S., Wang, H., Feng, S., Lee, T. F., Teng, C., et al. (2015). A one precursor one siRNA model for pol IV-dependent siRNA biogenesis. *Cell*. doi:10.1016/j.cell.2015.09.032.

Zhang, C., and Hu, H. (2014). High-efficiency nuclear transformation of the diatom

Phaeodactylum tricornutum by electroporation. *Mar. Genomics* 16, 63–66. doi:10.1016/j.margen.2013.10.003.

Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E., and Filipowicz, W. (2004). Single processing center models for human Dicer and bacterial RNase III. *Cell*. doi:10.1016/j.cell.2004.06.017.

Zhang, Q., and Edwards, S. V. (2012). The evolution of intron size in amniotes: A role for powered flight? *Genome Biol. Evol.* doi:10.1093/gbe/evs070.

Zhang, S., Sun, L., and Kragler, F. (2009). The phloem-delivered RNA pool contains small noncoding RNAs and interferes with translation1[W][OA]. *Plant Physiol.* doi:10.1104/pp.108.134767.

Zhao, T., Li, G., Mi, S., Li, S., Hannon, G. J., Wang, X. J., et al. (2007). A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. *Genes Dev.* doi:10.1101/gad.1543507.

Zhong, X., Du, J., Hale, C. J., Gallego-Bartolome, J., Feng, S., Vashisht, A. A., et al. (2014). Molecular mechanism of action of plant DRM de novo DNA methyltransferases. *Cell*. doi:10.1016/j.cell.2014.03.056.

Zhou, K., Qiao, K., Edgar, S., and Stephanopoulos, G. (2015). Distributing a metabolic pathway among a microbial consortium enhances production of natural products. *Nat. Biotechnol.* doi:10.1038/nbt.3095.

Zilberman, D., Cao, X., Johansen, L. K., Xie, Z., Carrington, J. C., and Jacobsen, S. E. (2004). Role of Arabidopsis ARGONAUTE4 in RNA-directed DNA methylation triggered by inverted repeats. *Curr. Biol.* doi:10.1016/j.cub.2004.06.055.

Zong, J., Yao, X., Yin, J., Zhang, D., and Ma, H. (2009). Evolution of the RNA-dependent RNA polymerase (RdRP) genes : Duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene* 447, 29–39. doi:10.1016/j.gene.2009.07.004.

APPENDIX

Isoprenoid biosynthesis in the diatom *Haslea ostrearia*

Anastasia Athanasakoglou[1], Emilia Grypioti[2], Sofia Michailidou[3], Codruta Ignea[1], Antonios M. Makris[3], Kriton Kalantidis[2,4], Guillaume Massé[5,6], Anagnostis Argiriou[3], Frederic Verret[2] and Sotirios C. Kampranis[1]

[1]Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, Frederiksberg C 1871, Denmark; [2]Department of Biology, University of Crete, PO Box 2208, Heraklion 71003, Greece; [3]Institute of Applied Biosciences – Centre for Research and Technology Hellas (INAB-CERTH), 6th km. Charilaou – Thermi Road, PO Box 60361, Thermi, Thessaloniki 57001, Greece; [4]Institute of Molecular Biology and Biotechnology – Foundation of Research and Technology Hellas (IMBB-FORTH), Nikolaou Plastira 100, Heraklion, Crete GR-70013, Greece; [5]UMI 3376 TAKUVIK, Centre national de la recherche scientifique (CNRS), Paris, France; [6]D´epartement de Biologie, Universit´e Laval, Qu´ebec, QC, Canada

Authors for correspondence:
Sotirios C. Kampranis
Tel: +45 353329 89
Email: soka@plen.ku.dk

Frederic Verret
Tel: +30 6938876848
Email: fverret@imbb.forth.gr

Received: *20 August 2018*
Accepted: *28 October 2018*

New Phytologist (2018)
doi: 10.1111/nph.15586

Key words: diatoms, *Haslea ostrearia*, heterokonts, isoprenoids, phytoene synthase, prenyltransferase, squalene synthase.

Summary

- Diatoms are eukaryotic, unicellular algae that are responsible for *c*. 20% of the Earth's primary production. Their dominance and success in contemporary oceans have prompted investigations on their distinctive metabolism and physiology. One metabolic pathway that remains largely unexplored in diatoms is isoprenoid biosynthesis, which is responsible for the production of numerous molecules with unique features.
- We selected the diatom species *Haslea ostrearia* because of its characteristic isoprenoid content and carried out a comprehensive transcriptomic analysis and functional characterization of the genes identified.
- We functionally characterized one farnesyl diphosphate synthase, two geranylgeranyl diphosphate synthases, one short-chain polyprenyl synthase, one bifunctional isopentenyl diphosphate isomerase – squalene synthase, and one phytoene synthase. We inferred the phylogenetic origin of these genes and used a combination of functional analysis and subcellular localization predictions to propose their physiological roles.
- Our results provide insight into isoprenoid biosynthesis in *H. ostrearia* and propose a model of the central steps of the pathway. This model will facilitate the study of metabolic pathways of important isoprenoids in diatoms, including carotenoids, sterols and highly branched isoprenoids.

Introduction

Diatoms (phylum Heterokontophyta, class Bacillariophyceae) are one of the most diverse and ecologically important groups of phytoplankton. With > 100 000 species that are widely distributed in aquatic environments, it is estimated that they contribute to *c.* 20% of the global primary production. Consequently, they play central roles in aquatic food webs and in the biogeochemical cycling of nutrients (Nelson *et al.*, 1995; Field *et al.*, 1998; Falkowski, 2002). This profound ecological success has created great interest in distinctive physiological features of diatoms. Even though several studies have already investigated unique facets of their metabolism (Kroth *et al.*, 2008; Ast *et al.*, 2009; Gruber *et al.*, 2009; Allen *et al.*, 2011; Fabris *et al.*, 2012; Smith *et al.*, 2012; Obata *et al.*, 2013), there are still important gaps in our understanding of key biochemical pathways. One such unexplored area is the biosynthesis of isoprenoids, a large class of metabolites that have vital biological functions in all domains of life (Holstein & Hohl, 2004).

Despite their structural diversity, all isoprenoids are assembled from the same five-carbon atom precursors, isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP). Two distinct biosynthetic routes are responsible for the synthesis of these five-carbon precursors: the mevalonate (MVA) pathway and the methylerythritol phosphate (MEP) pathway. Beyond these early steps, IPP and DMAPP are condensed to form prenyl diphosphate molecules of various lengths that serve as substrates for the synthesis of the different isoprenoid classes (geranyl diphosphate (GPP) for C_{10} monoterpenoids; farnesyl diphosphate (FPP) for C_{15} sesquiterpenoids, C_{30} triterpenoids and sterols; geranylgeranyl diphosphate (GGPP) for C_{20} diterpenoids and C_{40} carotenoids; etc.). This set of condensation reactions comprises the central steps of the pathway and they are catalyzed by prenyltransferase-type enzymes. After this central part, synthesis of carotenoids and sterols proceeds through the commitment of FPP and GGPP to squalene and phytoene, respectively, by squalene/phytoene synthase-type enzymes (Supporting Information Fig. S1) (Vranov´a *et al.*, 2012).

Initial work using feeding experiments with labelled precursors and specific pathway inhibitors has shown that both the MVA and MEP pathways are present in diatoms (Cveji´c & Rohmer,

2000; Mass´e *et al.*, 2004). Following the genomic sequencing of model diatom species *Thalassiosira pseudonana* (Armbrust *et al.*, 2004) and *Phaeodactylum tricornutum* (Bowler *et al.*, 2008), the focus of investigation has moved towards the final steps of the pathway; that is, the carotenoid biosynthetic branch that provides important light-harvesting and photoprotective molecules (Coesel *et al.*, 2008; Dambek *et al.*, 2012; Eilers *et al.*, 2016) and the synthesis of sterols, which serve as membrane structural components (Fabris *et al.*, 2014). An additional branch of isoprenoid biosynthesis that has attracted special interest is that of the highly branched isoprenoids (HBIs). HBIs are only synthesized by a limited number of diatom species. They are extensively used as geochemical and paleoenvironmental markers (Mass´e *et al.*, 2011; Belt & Müller, 2013) and show potential for use as pharmaceuticals and as an alternative form of fuels (Rowland *et al.*, 2001; Ferriols *et al.*, 2015, 2017).

Nevertheless, the central steps of the isoprenoid pathway are less well explored, and there is limited understanding on the function, subcellular localization and regulation of prenyltransferases from diatoms. Considering that the synthesis of prenyl diphosphates is a key regulatory step in the pathway that determines the flux towards different branches, a thorough investigation of the central steps will significantly improve our overall understanding of isoprenoid biosynthesis in diatoms. To this end, we selected the diatom species *Haslea ostrearia* for further studies, as this species is also able to synthesize HBIs (Wraige *et al.*, 1999) (Fig. S2), thus providing a more comprehensive model to study isoprenoid biosynthesis. Through transcriptomic analysis, we *in silico* reconstructed the MVA and MEP pathways and confirmed the expression of five putative prenyltransferases and two putative phytoene/squalene synthases. We cloned six of these genes and characterized their function in a heterologous host or *in vitro*. By the combination of phylogenetic analysis, subcellular localization prediction and functional characterization, we proposed a model of the isoprenoid pathway in *H. ostrearia*. This model will serve as a basis for the elucidation of the biosynthesis of HBIs or other useful isoprenoids from diatoms.

Materials and Methods

Diatom cultures

The *H. ostrearia* clone NCC 153.8 used in this study was a product of heterothallic reproduction between clones NCC 141 and NCC 171, both of which were isolated in 2003 from natural populations of the oyster pond Lainard (La Barre de Monts 46°53'33''N, 02°07'59''W). Clone NCC 153.8 was kindly provided by Dr Vona Medeler (University of Nantes) and cultured at the University of Crete Greece, in f/2 medium (Guillard, 1975) at 20°C under an irradiance of 50 μmol m^{-2} s^{-1} (12 h : 12 h light : dark cycle) before analysis. The ability of the strain NCC 153.8 to synthesize HBIs was confirmed by extraction and gas chromatography–mass spectrometry (GC–MS) analysis (Methods S1; Fig. S3).

RNA extraction and transcriptome sequencing

Total RNA was extracted with Spectrum™ Plant Total RNA Kit (Sigma-Aldrich) and quantified using the Qubit™ RNA BR Assay Kit (Invitrogen). Messenger RNA (mRNA) was isolated from total RNA using the NEBNext® Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, Ipswich, MA, USA). Complementary DNA (cDNA) library was constructed using the NEBNext Ultra Directional RNA Library Kit for Illumina (New England Biolabs), according to the manufacturer's instructions. Library quantification was conducted with the Kapa Library Quantification kit for Illumina sequencing platforms (Kapa Biosystems, Wilmington, MA, USA) on a Rotor-Gene Q thermocycler (Qiagen, Hilden, Germany). Sequencing was performed at the Institute of Applied Biosciences of the Centre for Research and Technology Hellas, on an Illumina NextSeq500 platform (Illumina Inc., San Diego, CA, USA) using the NextSeq™ 500/550 Mid Output Kit (2 × 150 cycles) (Illumina).

Transcriptome analysis

The overall bioinformatics strategy included the following steps: first, trim and clean-up of the sequencing reads using the trim galore wrapper (https://github.com/FelixKrueger/TrimGalore) with default parameters, except for -length 40 and the -fastqc option, so as to remove adaptors and low-quality sequences; second, *de novo* assembly of the read using the Trinity software suite (Grabherr *et al.*, 2011) with default parameters. All the analysis was implemented on a Linux-based HPC cluster assigning one node with 32 cores and 256 GB RAM.

Gene identification, sequence and phylogenetic analysis

Identification of candidate biosynthetic genes from *H. ostrearia* was based on homology with characterized and annotated homologues from red and green algae and heterokonts retrieved from the Genbank protein database (NCBI Resource Coordinators, 2017) and DiatomCyc database (Fabris *et al.*, 2012). For the phylogenetic analysis, all diatom homologues were retrieved from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP; Keeling *et al.*, 2014) database using Blast search. The species names and corresponding MMETSP ID numbers are listed in Table S1. Sequences from other heterokonts, red algae, green algae, land plants, cyanobacteria, bacteria, archaea, fungi and metazoa were retrieved from GenBank protein database (NCBI Resource Coordinators, 2017). Sequences were aligned using ClustalW (Larkin *et al.*, 2007), and alignments were manually edited by exclusion of ambiguously aligned regions. Phylogenies were inferred using the maximum likelihood method (Whelan & Goldman, 2001) and JTT matrix-based model in Mega v.7 (Kumar *et al.*, 2016). All positions with < 85% site coverage were eliminated. Branch support was generated using nonparametric bootstrap analysis with 100 replicates. Conserved motifs in the selected sequences were identified using an NCBI conserved domain search

(Marchler-Bauer *et al.*, 2015). Prediction of subcellular localization was carried out through ASAFind (Gruber *et al.*, 2015) and Hectar (Gschloessl *et al.*, 2008) in combination with SignalP 3.0 (Bendtsen *et al.*, 2004) and Tmhmm v.2.0 (Krogh *et al.*, 2001) for signal peptide and transmembrane domain prediction, respectively.

Gene amplification and cloning

Basic sequence analysis and design of primers (Table S2) was performed using CLC Workbench (Qiagen). Total RNA was extracted using Trizol, and cDNA was synthesized using SuperScript III RT (Thermo Fisher Scientific, Bremen, Germany) and DNAseI (Roche Applied Biosystems, Nutley, NJ, USA) treatment. Both full-length and truncated variants of the selected genes were PCR amplified using Phusion High-Fidelity DNA Polymerase (New England Biolabs) and cDNA as template. PCR products were gel purified, A-overhangs were added using MyTaq Polymerase (Bioline Reagents Ltd, London, UK) and the products were subsequently cloned into vector pCRII-TOPO using a TOPO TA cloning kit (Thermo Fisher Scientific). After digestion with the appropriate restriction enzymes, the digests were ligated to bacterial and yeast expression vectors. For bacterial expression, pRSET (N-terminal 6xHis tag) or pET102-His (C-terminal 6xHis tag) or pET102-Trx-His (N-terminal thireodoxin and C-terminal 6xHis tag) plasmid backbones, digested with the same restriction enzymes, were used. For yeast expression, vectors pUTDH3myc, pWTDHmyc and pHTDH3myc (Ignea *et al.*, 2012) were used. Final constructs were verified by sequencing. Gene expression in bacteria and yeast and protein purification protocols are described in Methods S1.

In vitro enzymatic assays

For the characterization of the prenyltransferases, 200 ll enzymatic assays were carried out in glass vials. The reaction mixture contained 10 mM MOPS buffer (pH 7.0), 5 mM magnesium chloride, 1 mM dithiothreitol, 1 mg ml^{-1} BSA, 100 lM prenyl diphosphate substrate. The substrates used were dimethylallyl diphosphate (D4287; Sigma-Aldrich), isopentenyl diphosphate (I0503; Sigma-Aldrich), geranyl diphosphate (G6772; Sigma-Aldrich), farnesyl diphosphate (F6892; Sigma-Aldrich), and geranylgeranyl diphosphate (G6025; Sigma-Aldrich). The reactions were initiated by addition of 50 ng of purified enzyme. After 16 h incubation at 25°C, the reactions were terminated by addition of an equal volume of 2 M hydrochloric acid in 83% ethanol and after 20 min incubation they were neutralized with 0.14 ml of 10% sodium hydroxide. The hydrolyzed diphosphates were extracted three times using 300 ll of hexane. The hexane extracts were concentrated to a final volume of 50 ll, and 1 ll of each reaction was used for GC–MS analysis (Methods S1). Individual compounds were identified by comparing their GC retention indices and mass spectra with those of authentic standards.

Data availability

All sequences from *H. ostrearia* mentioned in this study have been submitted to the GenBank database (www.ncbi.nlm.nih.gov); accession numbers are provided in Table 1. Transcriptomic data have been submitted to the European Nucleotide Archive database under experiment accession no. ERX2834706 and Run accession no. ERR2827962.

Results

Transcriptomic analysis and identification of candidate biosynthetic genes

Haslea ostrearia NCC 153.8 strain was cultured and its ability to produce HBIs was confirmed (Fig. S3) before RNA extraction, library construction and sequencing on an Illumina platform. After quality filtering and trimming, a total of 38 631 556 pair-end reads were *de novo* assembled into contigs using the Trinity suite (Haas *et al.*, 2013). A total of 45 508 contigs were obtained. We screened the assembled transcriptome to identify genes that putatively encode enzymes of the MVA and MEP pathways. Exploiting the high degree of sequence conservation among genes of isoprenoid biosynthesis, we based our screening on other characterized or annotated algal sequences (see the Materials and Methods section). A pairwise alignment between the queries and the contigs obtained from the transcriptome assembly was performed with the Massblast tool (Ver´ıssimo *et al.*, 2017). Through this analysis, we identified seven contigs corresponding to full-length protein sequences with high similarities to enzymes that catalyze the seven steps of the MEP pathway. We additionally identified six contigs with homology to genes involved in the conversion of acetyl coenzyme A (CoA) to IPP, through the MVA route. The last step of the MVA pathway involves the isomerization of IPP to DMAPP by isopentenyl diphosphate isomerase (IDI) (Berthelot *et al.*, 2012). Searching for an IDI, we were only able to identify a contig that corresponds to a fusion of an IDI with a squalene synthase (*Ho*IDI-SQS). Squalene synthases catalyze dimerization of two farnesyl diphosphate molecules to form squalene, the precursor of all sterols (Spanova & Daum, 2011). It was not possible to identify independent transcripts of IDI or SQS in our sequencing data, and we concluded that only the fusion of the two genes was expressed under the specific conditions.

We continued our analysis focusing on the central steps, which are catalyzed by prenyltransferase-type enzymes. Mining our assembled transcriptome, we were able to retrieve five contigs with similarity to annotated *trans*-prenyltransferases (PTSs). These included one putative farnesyl diphosphate synthase, sharing 57% sequence similarity at the amino acid level, with a functionally characterized homologue from *Rhizosolenia setigera* (Ferriols *et al.*, 2015) (from now on referred to as *Ho*PTS1) and four other putative polyprenyl diphosphate synthases (named *Ho*PTS2–5) that exhibited similarities to other diatom and algal prenyltransferases (Fig. 1a; Table 1). We investigated whether this set of the five PTSs is conserved among different diatom

Table 1 Candidate isoprenoid biosynthetic genes, their closest homologues, their evolutionary origin, and subcellular localization as indicated by analysis in this study.

Gene name	GenBank ID	Closest homologue (percentage similarity)	Evolutionary origin	Subcellular localization prediction
Methylerythritol phosphate pathway				
1-Deoxy-D-xylulose 5-phosphate synthase (DXS)	MH731010	*Phaeodactylum tricornutum* XP_002176386.1 (83%)	Red algae	Chloroplast
1-Deoxy-D-xylulose 5-phosphate reductoisomerase (DXR)	MH731011	*Thalassiosira pseudonana* XP_002295597.1 (84%)	Red algae	Chloroplast
2-*C*-Methyl-D-erythritol-4-phosphate-cytidylyltransferase (MCT)	MH731012	*Fistulifera solaris* GAX13480.1 (78%)	Red algae	Chloroplast
4-Diphosphocytidyl-2c-methyl-d-erythritol kinase (CMK)	MH731013	*Phaeodactylum tricornutum* XP_002178363.1 (77%)	Green algae	Chloroplast
2-*C*-Methyl-D-erythritol 2,4-cyclodiphosphate synthase (MDS)	MH731014	*Phaeodactylum tricornutum* XP_002180038.1 (76%)	Algae	Chloroplast
(*E*)-4-Hydroxy-3-methylbut-2-enyl diphosphate synthase (HDS)	MH731015	*Fragilariopsis cylindrus* OEU20628.1 (77%)	Green algae	Chloroplast
4-Hydroxy-3-methylbut-2-enyl diphosphate reductase (HDR)	MH731016	*Phaeodactylum tricornutum* XP_002178617 (73%)	Algae	Chloroplast
Mevalonate pathway				
Acetyl-CoA *C*-acetyltransferase 1 (AACT)	MH731017	*Phaeodactylum tricornutum* XP_002185228.1 (69%)		Cytosol
Hydroxy-methylglutaryl-CoA synthase (HMGS)	MH731018	*Fragilariopsis cylindrus* OEU16767.1 (73%)		Cytosol
Hydroxyl-methylglutaryl-CoA reductase (HMGR)	MH731019	*Fragilariopsis cylindrus* OEU16221.1 (82%)		Cytosol
Mevalonate kinase (MVK)	MH731020	*Thalassiosira pseudonana* XP_002287787.1 (73%)		Cytosol
Phospho-mevalonate kinase (PMK)	MH731021	*Fragilariopsis cylindrus* OEU13939.1 (49%)		Cytosol
Mevalonate diphosphate decarboxylase (MVD)	MH731022	*Fragilariopsis cylindrus* OEU17781.1 (69%)		Cytosol
Isopentenyl-diphosphate delta-isomerase fused to squalene synthase (*Ho*IDISQS)	MH720297	*Fistulifera solaris* GAX27897.1 (63%)	Green algae	Cytosol
Central steps				
Farnesyl diphosphate synthase (*Ho*PTS1)	MH720291	*Rhizosolenia setigera* AKH49589.1 (56%)	Not traced	ER or PPC
Polyprenyl diphosphate synthase (*Ho*PTS2)	MH720292	*Thalassiosira oceanica* EJK71722.1 (75%)	Algae	Chloroplast
Geranylgeranyl diphosphate synthase (*Ho*PTS3)	MH720293	*Phaeodactylum tricornutum* XP_002181666.1 (74%)	Heterotrophic host	Cytosol
Putative polyprenyl synthase (*Ho*PTS4)	MH720294	*Phaeodactylum tricornutum* XP_002185039.1 (70%)	Algae	Mitochondria
Geranylgeranyl diphosphate synthase (*Ho*PTS5)	MH720295	*Phaeodactylum tricornutum* XP_002178555.1 (75%)	Red algae	Chloroplast
Phytoene synthase (*Ho*PSY)	MH720296	*Phaeodactylum tricornutum* XP_002178776.1 (69%)	Red algae	Chloroplast

ER, endoplasmic reticulum; PPC, periplastidial compartment.

species by examining the publicly available diatom transcriptomes in the MMETSP database (Keeling *et al.*, 2014). Based on homology searches, *Ho*PTS1, *Ho*PTS2, and *Ho*PTS4 orthologues are present in all diatom species investigated (a total of 26 species, representing both centric and pennate diatoms). An *Ho*PTS5 orthologue seems to be missing only in one species (*Skeletonema menzelii*), while a *Ho*PTS3 orthologue seems to be absent from seven diatom species (Table S3; Fig. S4). Among the selected transcripts, *Ho*PTS5 showed significantly high expression levels (Fig. S5).

Sequence analysis of the identified prenyltranserases, led to identification of conserved polyprenyl synthase domains (Fig. 1b) and conserved motifs. Among them, the two aspartic acid-rich motifs DDxx(xx)D (First Aspartic acid-Rich Motif – FARM; and Second Aspartic acid-Rich Motif – SARM), found in characterized prenyltransferases from all domains of life. These motifs are involved in the binding of magnesium ions and are essential for prenyltransferase activity. The identification of intact motifs in all five PTSs suggested that these genes likely encode active enzymes. Conserved amino acids were also observed at positions 4 and 5 upstream of FARM. It has previously been shown that these residues are involved in the regulation of the product chain length (Tarshis *et al.*, 1996; Wang & Ohnuma, 1999; Liang *et al.*, 2002). In *Ho*PTS1, like in other characterized FPP

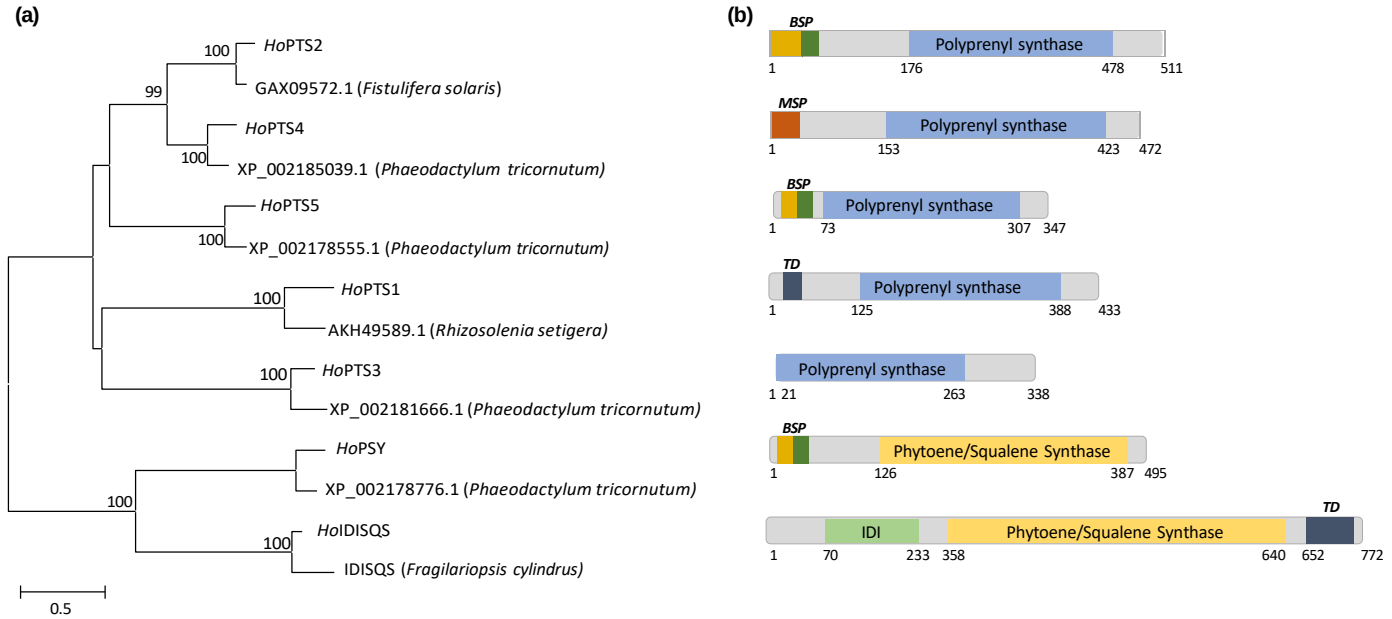

Fig. 1 (a) Neighbor-joining phylogenetic tree of the selected sequences and their closest homologues. Numbers on the branches indicate bootstrap support values from 1000 trees; values under 70 were removed. (b) Protein domain structure of selected sequences. BSP, bipartite signal peptide; MSP, mitochondrial signal peptide; TD, transmembrane domain.

synthases, this region contains aromatic residues that are bulkier and block further chain elongation (Fig. S6a). By contrast, smaller residues are observed at this region in the remaining four PTSs, suggesting that these likely synthesize longer chain products (Fig. S6b).

The prenyl diphosphates synthesized in the central steps are allocated to different branches of the pathway for the synthesis of final isoprenoid products, with sterol and carotenoid pathways being the main ones. Specific enzymes commit FPP and GGPP to these pathways. As already described, we identified a squalene synthase, N-terminally fused to an IDI (*Ho*IDI-SQS). In the absence of additional SQS transcripts, this bifunctional protein could be involved in committing FPP to the synthesis of sterols. We also identified a single transcript, termed *Ho*PSY, that shares 72% similarity with a characterized phytoene synthase from *P. tricornutum* (Dambek *et al.*, 2012) and could catalyze the first committed step of carotenoid biosynthesis in *H. ostrearia*. Both *Ho*IDI-SQS and *Ho*PSY were found to contain the conserved phytoene/squalene synthase domain (Fig. 1b).

Phylogenetic analysis

Diatoms have a distinctive evolutionary history. Red and green algae, as well as glaucophytes, evolved after a primary endosymbiotic event, when a heterotrophic eukaryote engulfed a cyanobacterial cell (McFadden, 2001; Rodr´ıguez-Ezpeleta *et al.*, 2005). Diatoms arose from secondary endosymbiosis, during which a different heterotrophic eukaryote captured a cell of red algal origin (Yoon *et al.*, 2004; Prihoda *et al.*, 2012). After incorporation, the engulfed cell was transformed into the plastid organelle (Bhattacharya *et al.*, 2007), lost its mitochondria and nucleus, and genetic information was transferred into the nucleus of the heterotrophic host, in a process termed endosymbiotic gene transfer (Timmis *et al.*, 2004). The evolutionary linkage between diatoms and red algae is usually observed in phylogenetic surveys (Armbrust *et al.*, 2004; Bowler *et al.*, 2008; Frommolt *et al.*, 2008). However, such surveys have also revealed a green phylogenetic signal. It has been proposed that this is due to a cryptic endosymbiotic event that involved a green algal cell, which was later replaced by the red algal endosymbiont (Moustafa *et al.*, 2009; Chan *et al.*, 2011). Alternatively, the green related genes could have been acquired through repeated horizontal gene transfer events from green algae, early after the evolution of the first diatoms (Oborn´ık & Green, 2005; Dorrell & Smith, 2011). This multisourced genetic background has created unique, chimeric metabolic pathways that combine features from multiple lineages.

To investigate how this is reflected on isoprenoid biosynthesis, we inferred the phylogenetic origin for each of the candidate biosynthetic genes. Three different possibilities were distinguished: (1) genes that originated from the secondary heterotrophic host, (2) genes that originated from the red algal endosymbiont, and (3) genes that originated from a green algal cell, and have been acquired either through endosymbiotic or horizontal gene transfer.

It is evident that the MEP pathway genes can only have algal origin, as only the autotrophic cells involved in the evolution of diatoms had the plastidial route of isoprenoid biosynthesis. Taking this into account, we set out to investigate whether it was a red or a green algal cell that mediated each of the MEP pathway gene transfers to diatoms (Fig. S7). According to our analysis, 1-deoxy-D-xylulose 5-phosphate synthase, 1-deoxy-D-xylulose 5-phosphate reductoisomerase, and 2-*C*-methyl-D-erythritol-4-phosphate-cytidylyltransferase (Fig. S7a–c) were transferred from the red algal endosymbiont. This is evident from the highly

supported clustering of diatom and red algal proteins. By contrast, the 4-diphosphocytidyl-2-*C*-methyl-D-erythritol kinase and (*E*)-4-hydroxy-3-methylbut-2-enyl diphosphate synthase (Fig. S7d,f) genes were transferred from a green algal related cell. For the remaining two, 2-*C*-methyl-D-erythritol 2,4-cyclodiphosphate synthase and 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (Fig. S7e,g), diatom proteins form well-separated clades, in distance from their algal homologues. This topology suggests that the corresponding genes have diverged, so that their origin cannot be traced.

Prenyltransferases and phytoene/squalene synthases could have been acquired from any of the cells involved. The poorly supported branching patterns obtained for *Ho*PTS1 and its homologues did not allow us to trace its exact evolutionary origin (Fig. S8a). *Ho*PTS3 clusters closely to prenyltransferases from opisthokonts (fungi/metazoa), whereas other algal and cyanobacterial homologues are branching separately (Fig. S8c). This topology suggests heterotrophic host origin. By contrast, the topologies obtained for the rest of the prenyltransferases, as well as the phytoene/squalene synthases, are suggestive of algal origin. For *Ho*PTS5 and *Ho*PSY this is the red algal endosymbiont, as diatom and red algal proteins form well-supported branches together (Fig. S8e,f). The *Ho*IDI-SQS gene is only conserved among heterokonts, haptophytes, and dinoflagellates, so it is likely that the fusion occurred in the common ancestor of these groups (Davies *et al.*, 2015). We investigated the evolutionary origin of each domain separately, and both for IDI and SQS the inferred phylogenies indicate green algal origin. Whereas for SQS the corresponding cluster is poorly supported, branching is more robust for IDI (Fig. S8g,h). *Ho*PTS2 and *Ho*PTS4 have also been acquired from an algal cell; however, owing to divergence, it was not possible to pinpoint their exact origin (Fig. S8b,d).

This divergence could alternatively be explained by gene duplication events. To resolve this, we examined three other diatom species (*P. tricornutum*, *T. pseudonana* and *Fragilariopsis cylindrus*) that had their genomes sequenced. We obtained the homologues of the *H. ostrearia* genes with nonresolved origin and we identified their localization in the respective genome. In all cases, those genes are localized in different genomic regions (Table S4), indicating that they have been likely independently acquired.

For genes of the MVA pathway, see the extended discussion on the phylogenetic analysis that is included in Notes S1.

Prediction of subcellular localization

To facilitate analysis of the physiological roles of the genes identified, we investigated the presence of signal/target peptides and obtained insight into the subcellular localization of the encoded proteins. The different endosymbiotic events provided diatoms with secondary plastids surrounded by four membranes (Kroth & Strotmann, 1999). As a result, nuclear encoded diatom proteins targeted to the chloroplast contain a bipartite N-terminal presequence, consisting of a signal peptide with a conserved motif at its cleavage site (ASAFAP motif), followed by a chloroplast transit peptide. These presequences are essential for efficient transportation through all of the four membranes (Gruber *et al.*, 2007). Two independent computational tools, ASAFind (Gruber *et al.*, 2015) and Hectar (Gschloessl *et al.*, 2008), developed for identification of such presequences, were used in our study. A different tool that is used to predict transmembrane domains in protein sequences (Tmhmm Server v.2.0; Krogh *et al.*, 2001) was included to further support this analysis. The results are summarized in Table S5. As expected, enzymes of the MEP pathway are predicted to be targeted to the chloroplast and those of the MVA to the cytosol. *Ho*PTS1 was found to contain a type II signal anchor, and analysis with Tmhmm suggested a strong possibility for the presence of a transmembrane domain at the N-terminal region. Taken together, these results suggest that *Ho*PTS1 is likely localized on the endoplasmic reticulum (ER) membrane. The ER membrane is continuous with the outermost chloroplastic membrane, so the possibility that *Ho*PTS1 is related to the periplastidial compartment (PPC; the space between the second and third outermost membranes) cannot be ruled out. For *Ho*PTS3 and *Ho*IDI-SQS, no target or signal peptides were predicted, indicating cytosolic localization. *Ho*PTS2, *Ho*PTS5, and *Ho*PSY were predicted to be targeted to the chloroplast, and *Ho*PTS4 to the mitochondria.

Functional characterization of prenyltransferases

Since the subcellular compartmentalization analysis revealed the presence of signal/target peptides for the majority of the

Table 2 Functionally characterized enzymes in this study.

	Full-length ORF	Variants studied	Accepted substrates	Products
*Ho*PTS1	1299 bp	Full length	No activity	—
		Ser73–end	DMAPP + IPP	GPP, FPP
			GPP + IPP	FPP
		Val89–end	DMAPP + IPP	GPP, FPP
			GPP + IPP	FPP
*Ho*PTS2	1584 bp	Full length	No activity	—
		Arg45–end	FPP + IPP	C_{20}–C_{30} PPP
			GGPP + IPP	C_{25}–C_{30} PPP
		Gly140–end	No activity	No activity
*Ho*PTS3	1014 bp	Full length	GPP + IPP	FPP, GGPP
			FPP + IPP	GGPP
*Ho*PTS4	1416 bp	Full length	No activity	—
		Leu121–end	No activity	
*Ho*TPS5	1011 bp	Full length	No activity	—
		Ser36–end	DMAPP + IPP	GPP, FPP, GGPP
			GPP + IPP	FPP, GGPP
			FPP + IPP	GGPP
*Ho*PSY	1485 bp	Full length	Not tested	—
		Ser78–end	GGPP + GGPP	Phytoene
*Ho*IDISQS	2277 bp	Full length	FPP + FPP	Squalene

ORF, open reading frame; DMAPP, dimethylallyl diphosphate; IPP, isopentenyl diphosphate; GPP, geranyl diphosphate; FPP, farnesyl diphosphate; GGPP, geranylgeranyl diphosphate; PPP, polyprenyl diphosphate.

prenyltransferases identified, both the full-length and different truncated variants of the enzymes (Table 2) were cloned into bacterial vectors, in-frame with a C- or N-terminal 6xHis-tag. This allowed the purification of the expressed proteins and the assessment of their activity in *in vitro* enzymatic reactions with prenyl diphosphate substrates. IPP was always used as the homoallylic substrate, whereas DMAPP, GPP, FPP, and GGPP were used as the allylic substrates. Control reactions with cells transformed with empty vectors and subjected to the same purification steps were run in parallel. Following acid hydrolysis that facilitated the removal of diphosphates from the substrates and products, all reactions were extracted and analyzed by GC–MS. Compound identification was based on combination of accurate mass and retention time comparisons between reaction products and acid hydrolysis products of authentic standards.

To characterize *Ho*PTS1, the full-length enzyme was initially incubated with DMAPP or GPP and IPP. As no activity was detected, two truncated variants, *Ho*PTS1(Ser73–end) and *Ho*PTS1(Val89–end), were tested with the same substrates. Analysis of the reaction extracts resulted in the identification of linalool (LOH) and nerolidol (NOH), the acid hydrolysis products of GPP and FPP, respectively (Fig. 2a). In the presence of DMAPP and IPP, both alcohols were detected, whereas in the presence of GPP and IPP the only product identified was NOH. No other substrate combination resulted in new product formation, demonstrating that *Ho*PTS1 is an FPP synthase.

Initial expression of full-length and truncated versions of *Ho*PTS2 resulted in formation of inclusion bodies. To obtain soluble proteins, the variants were subcloned in frame with thioredoxin at the N-terminus and a C-terminal 6xHis-tag. Soluble *Ho*PTS2 fusions were obtained and purified. The activities of the full-length enzyme and *Ho*PTS2(Gly140–end) variant were tested with all substrate combinations but no new product was formed. However, variant *Ho*PTS2(Arg45–end) was active and gave a range of C_{20}–C_{30} prenyl diphosphates in the presence of FPP and IPP and C_{25}–C_{30} prenyl diphosphates when GGPP and IPP were used as reaction substrates. We were able to identify these products only after hydrolysis treatment and formation of the corresponding prenyl alcohols (Fig. 2b). The fact that the variant *Ho*PTS2(Gly140–end) was inactive indicates the presence of a functionally important region between amino acids 45 and 140. Taken together, these results suggest that *Ho*PTS2 is likely a short-chain polyprenyl synthase.

As there was no prediction for the presence of any signal/target peptide in *Ho*PTS3, the full-length enzyme was expressed, purified, and tested with different substrates. In the presence of DMAPP and IPP, there was no formation of new products. However, when GPP and IPP were used as reaction substrates, GGPP was produced, detected as its acid hydrolysis product geranyl-LOH. The same product was formed when the enzyme was incubated with FPP and IPP (Fig. 2c). These results, in combination with the lack of a signal peptide predicted for *Ho*PTS3, suggest that *Ho*PTS3 is possibly a cytosolic GGPP synthase in *H. ostrearia.* Efforts to characterize *Ho*PTS4 were unsuccessful, as there was no activity detected in any reaction extract analyzed, either when using the full-length protein or the truncated variant *Ho*PTS4 (Leu121–end). The prediction of a mitochondrial target peptide at the N-terminal region suggests that this, possibly, affects the catalytic activity of *Ho*PTS4, and more truncated variants should be tested to identify the optimal cleavage position for obtaining an active enzyme.

Finally, full-length *Ho*PTS5 and one truncated variant, *Ho*PTS5(Ser36–end), were tested with different substrate combinations. Even though the full-length enzyme was inactive, the truncated variant *Ho*PTS5(Ser36–end) produced GGPP as the main product. This was detectable by the formation of geranyl-LOH after acid hydrolysis of the reaction products. Minor peaks of LOH and NOH were detected when DMAPP and IPP were used as substrates, whereas NOH could be detected when GPP and IPP were used instead, indicating that the formation of GGPP proceeds via GPP and FPP (Fig. 2d). Taken together with the prediction of a chloroplastic target peptide in its N-terminus, these results suggest that *Ho*PTS5 is possibly a GGPP synthase, likely functioning in the chloroplast providing the substrates for carotenoid and phytol biosynthesis.

Functional characterization of the squalene/phytoene synthase family members

IDI and SQS are enzymes that catalyze two nonconsecutive reactions. IDI isomerizes IPP to DMAPP, whereas SQS catalyzes the formation of squalene using two FPP molecules. In order to characterize the *Ho*IDISQS fusion, we studied the two domains separately. Initially, we evaluated the functionality of the SQS domain by introducing the full-length gene into a *Saccharomyces cerevisiae* strain engineered to produce high amounts of FPP (Ignea *et al.*, 2012). Analysis of the nonsaponifiable lipid extract of yeast cells expressing the fusion showed production of higher amounts of squalene when *Ho*IDISQS was expressed, compared with the levels of squalene produced by the endogenous yeast squalene synthase (Fig. 3a). This suggests that the SQS domain of *Ho*IDISQS likely contains a functional synthase.

The functionality of the IDI domain was assessed both in the fusion (*Ho*IDISQS) and as a separately cloned IDI domain (*Ho*IDI). These two enzymes were coexpressed with a monoterpene synthase from *Salvia fruticosa*, 1,8-cineole synthase (*Sf*CinS1) (Kampranis *et al.*, 2007), in a yeast strain engineered to produce GPP (Ignea *et al.*, 2012). By catalyzing the isomerization of IPP to DMAPP, IDI provides substrates to the endogenous yeast enzyme, Erg20p, which produces GPP. It has been shown that functional IDI expression in this strain increased synthesis of 1,8-cineole by *Sf*CinS1 (Ignea *et al.*, 2011). We used the endogenous yeast IDI (*Sc*IDI) as positive control. We sampled the head space of all yeast cell cultures expressing the different combinations with solid-phase microextraction. Their analysis showed that, when *Ho*IDISQS, *Ho*IDI, or *Sc*IDI was expressed, significantly higher amounts of 1,8-cineole were produced (Fig. 3a), indicating that the IDI domain is functional both individually and in the fusion.

We also employed the yeast expression system for the characterization of the candidate phytoene synthase, *Ho*PSY. During our signal/target peptide analysis, *Ho*PSY was predicted to have a

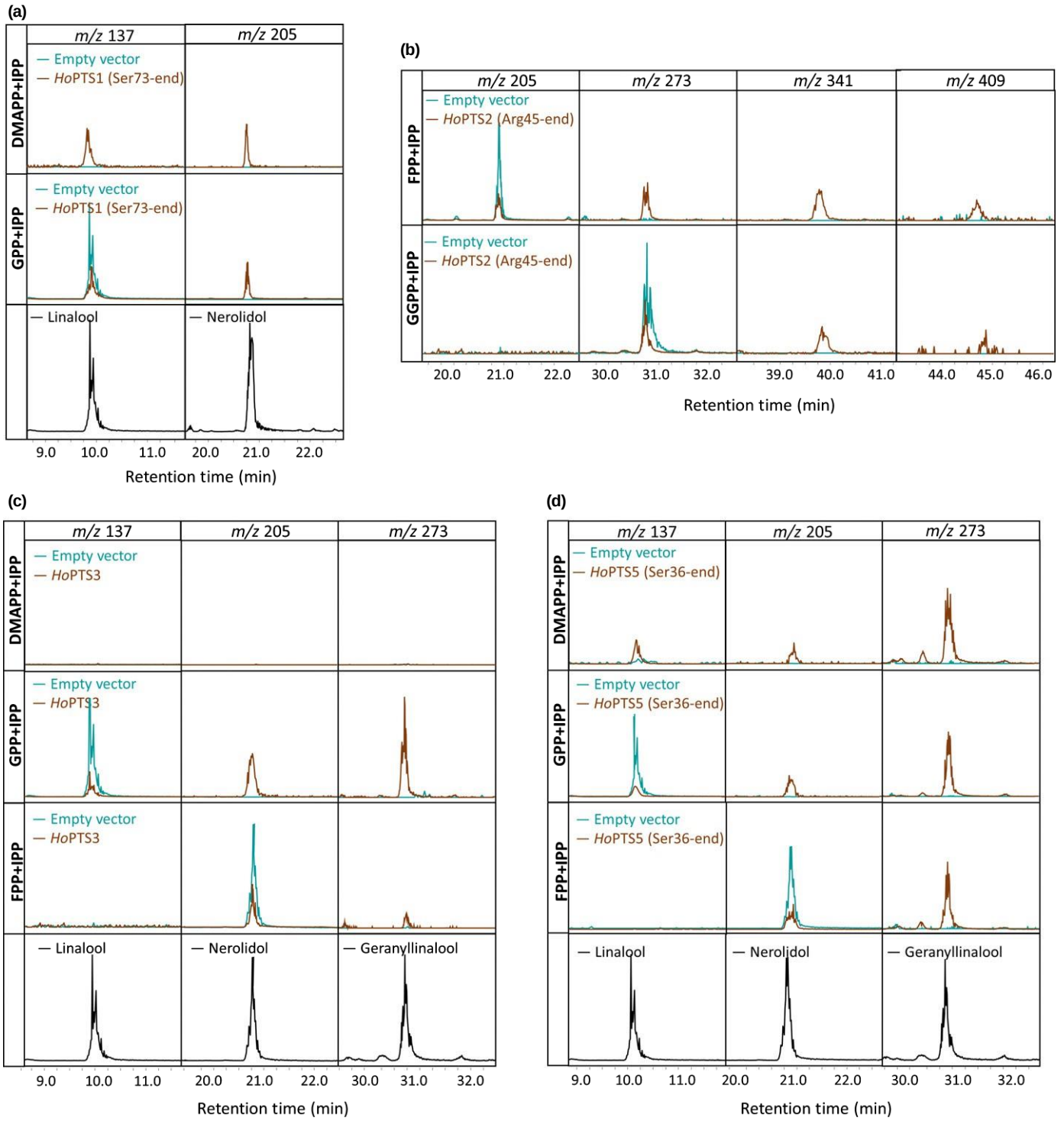

Fig. 2 Gas chromatography–mass spectrometry profile of (a) *Ho*PTS1(Ser73–end), (b) *Ho*PTS2(Arg45–end), (c) *Ho*PTS3, (d) *Ho*PTS5(Ser36–end) *in vitro* reaction products using different prenyl diphosphates as substrates. Identification of compounds was based on comparison of accurate mass and retention time between reaction substrates/products and acid hydrolysis products of authentic standards (bottom panels within a, c, d). Ions *m/z* 137, 205, 273, 341 and 409 are derived from precursor ions (*m/z* 154, 222, 290, 358 and 426, respectively) by loss of water in positive ion mode.

chloroplastic signal peptide. Taking into account the cleavage position predictions, we introduced the truncated variant *Ho*PSY (Ser78–end) into a yeast expression vector. In order to evaluate the functionality of the variant, we employed an *in vivo* chromogenic assay. In this assay, the candidate phytoene synthase is coexpressed with a GGPP synthase and a phytoene desaturase. The GGPP synthase provides the substrate for phytoene biosynthesis. If the examined enzyme is functional and produces phytoene, the desaturase will take it up to produce lycopene (Fig. 3c). As yeast colonies that accumulate lycopene become

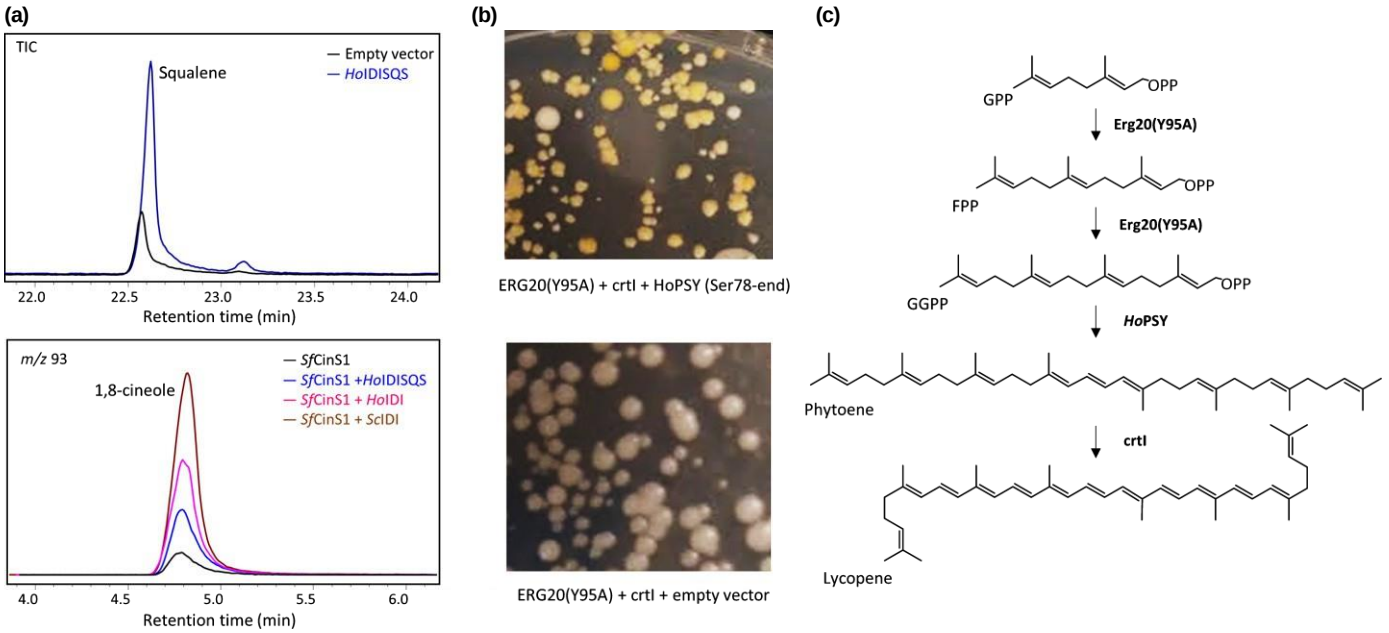

Fig. 3 Functional characterization of squalene/phytoene synthase family members. (a) Expression of *Ho*IDISQS in the yeast strain AM94 resulted in high accumulation of squalene (upper panel). TIC, total ion chromatogram. Coexpression of *Ho*IDISQS with *Sf*CinS1 resulted in high accumulation of 1,8-cineole (lower panel). *Ho*, *Haslea ostrearia*; *Sf*, *Salvia fruticosa*; *Sc*, *Saccharomyces cerevisiae*. (b) Yeast colonies coexpressing Erg20p(Y95A), crtI, and *Ho*PSY(Ser78–end) (upper panel) in comparison with the control colonies that carry an empty vector instead of a phytoene synthase (lower panel). (c) Lycopene biosynthetic pathway.

orange colored, this assay provides a quick and reliable means for the evaluation of phytoene synthase activity. We used the yeast strain AM94, which is engineered for the efficient production of isoprenoids (Ignea *et al.*, 2012), and introduced *Ho*PSY(Ser78–end) together with Erg20p(Y95A), an engineered GGPP synthase (Ignea *et al.*, 2015), and crtI, a phytoene desaturase from *Xanthophyllomyces dendrorhous* (Verwaal *et al.*, 2007). Whereas control cells that were transformed with an empty vector, instead of *Ho*PSY(Ser78–end), did not develop any color, coexpression of the three genes resulted in orange-colored yeast colonies, confirming the ability of *Ho*PSY to synthesize phytoene (Fig. 3b).

Discussion

Aiming to shed light on isoprenoid biosynthesis in diatoms, we investigated different aspects of the pathway in the species *H. ostrearia*. By combining transcriptomic analysis with functional characterization of enzyme activities and predictions for the subcellular localization of the corresponding proteins, we can propose a model of isoprenoid biosynthesis in *H. ostrearia* that is summarized in Fig. 4 and discussed in the following paragraphs.

Haslea ostrearia retains a functional cytosolic MVA pathway and a functional plastidial MEP pathway. The majority of the sequenced diatom species investigated to date have both routes for IPP and DMAPP synthesis. However, these precursors seem to be differentially allocated towards the final isoprenoid products across different species. A plant-like dichotomy has been observed in *P. tricornutum*, *Nitzschia ovalis* and *R. setigera.* Accordingly, sterols are synthesized using precursors from the MVA route, whereas the biosynthesis of carotenoids and the diterpenoid phytol proceeds via MEP-generated precursors in the chloroplast (Cveji´c & Rohmer, 2000; Mass´e *et al.*, 2004). On the contrary, *H. ostrearia* uses chloroplast-derived precursors to synthesize its main sterol (24-ethylcholest-5-en-3-ol) (Mass´e *et al.*, 2004). Contribution of the MEP pathway to sterol (24-methylcholesta-5,24(28)-dien-3b-ol) biosynthesis is also observed in the centric species *T. pseudonana* under fast-growing, nitrogen-replete culture conditions (Zhang *et al.*, 2009). Similar differentiations were shown for HBI biosynthesis, with *R. setigera* incorporating C_5 precursors derived from the MVA route and *H. ostrearia* synthesizing HBIs with precursors generated via the MEP pathway (Mass´e *et al.*, 2004). These different patterns on precursor partitioning indicate different regulation mechanisms that, in addition to being responsive to external conditions (e.g. nutrient availability), also possibly involve precursor transportation between cytosol and chloroplast. MVA and MEP pathway crosstalk has already been reported in some plants (Bick & Lange, 2003; Hemmerlin *et al.*, 2003a). Even though there are substantial differences between the organization of the primary plastids of plants and the secondary plastids of diatoms, the four membranes of diatoms' plastids were previously shown to be permeable through specific transporters (Ast *et al.*, 2009).

The first key step towards prenyl diphosphate synthesis is the isomerization of IPP to DMAPP by IDI. Photosynthetic heterokonts, including diatoms and brown algae, haptophytes, and dinoflagellates, are characterized by the expression of a bifunctional protein fusion between IDI and SQS that catalyzes the dimerization of FPP towards sterol synthesis. In *H. ostrearia* this enzyme fusion does not appear to have any target peptide, whereas it is predicted to contain a pair of transmembrane helices

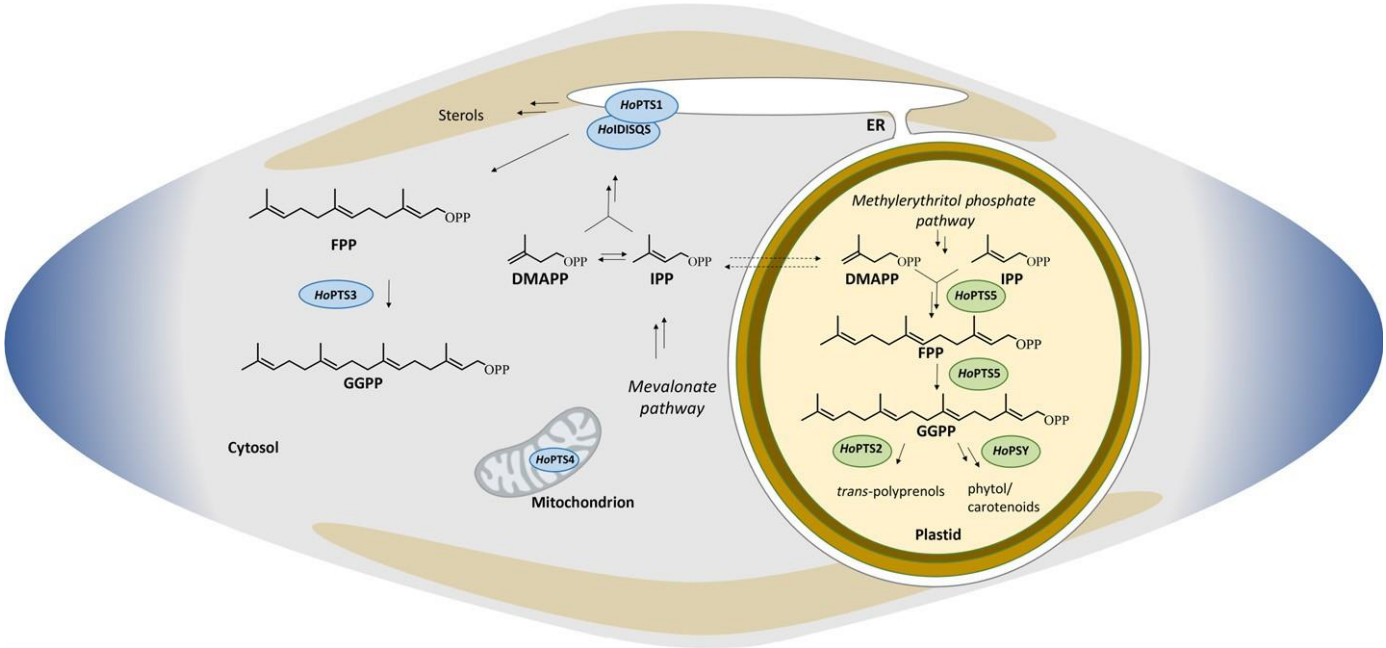

Fig. 4 A model for isoprenoid biosynthesis in *Haslea ostrearia*. Isopentenyl diphosphate (IPP) and dimethyl allyl diphosphate (DMAPP) are synthesized via the mevalonate and methylerythritol phosphate pathways in the cytosol and plastid. Each of these precursor pools is used for the synthesis of prenyl diphosphates by prenyltransferases in different subcellular compartments. *Ho*PTS1 is responsible for farnesyl diphosphate (FPP) synthesis at the endoplasmic reticulum (ER), in close proximity to *Ho*IDISQS. *Ho*PTS3 mediates synthesis of geranylgeranyl diphosphate (GGPP) in the cytosol. In the chloroplast *Ho*PTS2 and *Ho*PTS5 synthesize precursors for polyprenol and carotenoid synthesis. The first committed step of carotenoid biosynthesis is catalyzed by *Ho*PSY. *Ho*PTS4 likely mediates prenyl diphosphate synthesis in mitochondria. Dashed arrows indicate possible precursor transportation between cytosol and plastids. OPP, diphosphate.

at the C-terminal part, suggesting ER localization. No other contig encoding for an IDI could be identified in our transcriptomic data. Analysis of 33 diatom species by Ferriols *et al.* (2017) showed that all but one specific strain of *R. setigera* express an IDI-SQS fusion, and 19 of them additionally express an independent IDI. Many of these fall within the same phylogenetic clade with the independent IDI from *R. setigera* that contains a putative chloroplast-targeting peptide. A growing body of evidence recently supports the occurrence of alternative splicing in diatoms (Rastogi *et al.*, 2018). Thus, the possibility that diatoms that only contain the IDI-SQS fusion gene may also produce an alternatively spliced form producing only the IDI protein cannot be ruled out.

The intermediate step between the reactions catalyzed by IDI and SQS is FPP synthesis, and according to our functional characterization is likely catalyzed by *Ho*PTS1. Sequence analysis of *Ho*PTS1 suggested the presence of a type II signal anchor. This is responsible to anchor the enzyme to a membrane. It is reasonable to assume that this is the ER membrane, which is continuous with the outermost chloroplastic membrane in diatoms. Since *Ho*PTS1 and *Ho*IDISQS catalyze consecutive reactions, we can speculate that they might also physically interact, forming an enzymatic complex that is localized at the ER. Previous studies showed that *H. ostrearia*'s main sterol, 24-ethylcholest-5-en-3-ol, incorporates precursors generated from the plastid localized MEP pathway (Mass´e *et al.*, 2004). Taken together, these results indicate again precursor transportation from the chloroplastic stroma to the cytosol. The mechanism and the exact regulation of such events are unknown. Recently, a novel isoprenoid regulatory mechanism that possibly involves precursor transportation between subcellular compartments has been described in plants. According to this, a cytosolic isopentenyl phosphate kinase and specific Nudix hydrolases regulate IPP supply to the pathway by active phosphorylation–dephosphorylation of IP/IPP. Perturbation of IP supply in *Nicotiana tabacum* was shown to affect both MVA- and MEP-derived isopenoids, revealing a connection with precursor transportation (Henry *et al.*, 2015, 2018). Mining the transcriptome of *H. ostrearia* for similar genes, we were able to identify homologues of isopentenyl phosphate kinase and Nudix hydrolase (candidate sequences can be found in Notes S2) that are also present in transcriptomes of other sequenced diatom species. Whether a similar regulatory mechanism is present in diatoms and how this controls IP/IPP supply and/or transportation are open questions that remain to be answered.

Our model suggests that *Ho*PTS5 acts in the plastids to generate the GGPP that is essential for carotenoid synthesis. The high relative expression levels of *Ho*PTS5 (Fig. S5) probably reflect the high demand for carotenoids and phytol under the specific growth conditions. The first committed step towards carotenoids is catalyzed by the phytoene synthase *Ho*PSY. Both of these enzymes were acquired from the red algal endosymbiont, a fact that corroborates their predicted targeting to the chloroplast. *Ho*PTS2, which was also predicted to have chloroplastic localization, showed activity as a short-chain polyprenyl synthase. The lack of detailed information on the isoprenoid content of diatoms prevents us from assigning a role to *Ho*PTS2. It is possible that

this enzyme synthesizes polyprenyl diphosphates or the corresponding alcohols *in vivo.* In plants these compounds have been shown to be incorporated in thylakoid membranes and modulate their fluidity, influencing the overall photosynthetic performance (Bajda *et al.*, 2009; Surmacz & Swiezewska, 2011; Akhtar *et al.*, 2017).

We propose that cytosolic isoprenoid biosynthesis is supported by precursors generated by the MVA pathway and prenyl diphosphates synthesized by *Ho*PTS3. Our analysis showed that this is probably the only enzyme that is not conserved among different diatom species. In *H. ostrearia*, the *Ho*PTS3 gene likely originates from the heterotrophic host of secondary endosymbiosis.

Although it was not possible to characterize the activity of *Ho*PTS4, we presume that this enzyme likely acts as a prenyl transferase in the mitochondria. As there was no other prenyltransferase predicted to be targeted to mitochondria, *Ho*PTS4 probably uses substrates transported from the cytosol or chloroplast. Crosstalk between mitochondria and chloroplast in diatoms has been previously proposed for other metabolic processes (Prihoda *et al.*, 2012).

Even though the majority of prenyltransferases catalyze the linear, head-to-tail or head-to-head, condensation of prenyl diphosphates, there are examples where these enzymes are involved in the synthesis of irregular isoprenoids (Rivera *et al.*, 2001; Demissie *et al.*, 2013). A characteristic example is an FPP synthase in *Artemisia tridentata* that, among other reactions, catalyzes the head-to-middle linkage of two DMAPP molecules to produce the branched isoprenoid lavandulyl diphosphate (Hemmerlin *et al.*, 2003b). It is thus likely that HBI biosynthesis also involves prenyltransferase-type enzymes. Furthermore, different HBI-producing diatom species were shown to use different precursors for the synthesis of these molecules (Mass´e *et al.*, 2004), suggesting that the pathway is regulated differently among species. By providing here a model of the isoprenoid pathway in *H. ostrearia* that illustrates the subcellular distribution of different prenyltransferases, we believe that we provide a basis for future HBI biosynthetic studies that will test the currently identified or different enzymes, substrates, and/or conditions.

Conclusions

In conclusion, our investigation has significantly contributed to our understanding of isoprenoid biosynthesis in diatoms. At least five prenyltransferases mediate the prenyl diphosphate synthesis and provide substrates to downstream branches of the pathway. It is likely that precursors are transported from the plastids to the cytosol. Further studies that will confirm the localization of the enzymes *in vivo* and explore the regulatory mechanisms and crosstalk between the MVA and MEP pathways are essential for a thorough elucidation of the mechanisms involved. Our phylogenetic analysis provided an insight into the multisourced genetic background of diatoms that is reflected clearly on isoprenoid biosynthesis. Although there is still a general debate on the extent of the contribution of each lineage (red and green) to the genomes of diatoms (Dagan & Martin, 2009; Burki *et al.*, 2012; Deschamps & Moreira, 2012), it is commonly accepted that these events have armed diatoms with the genetic potential and the metabolic plasticity to succeed in contemporary oceans. Finally, our results provide the blueprint for the elucidation of the biosynthetic pathways leading to unique diatom isoprenoids, such as HBIs.

Acknowledgements

We wish to thank Dr David Ian Pattison for assistance in GC–MS sample running and analysis and Dr Vona Medeler, the scientific curator of the Nantes Culture Collection (NCC), for providing the NCC 153.8 strain. We kindly acknowledge the financial support of the European Commission through projects PIRSES-GA-2010-269294-BIOVADIA and RISE 734708 – GHaNA (to SCK, KK and GM) and the Greek Secretariat of Research and Technology through project 11SYN_3_770 (to SCK, AMM and A Argiriou) that was co-financed by the European Regional Development Fund.

Author contributions

AAthanasakoglou, EG, SM, FV and CI carried out experiments. SCK, FV, KK, AMM, AArgiriou and GM designed experiments. AAthanasakoglou, EG, SM, CI, FV, SCK, AMM, GM, KK and AArgiriou analyzed results. AAthanasakoglou and SCK wrote the manuscript. All authors have read and commented on the manuscript.

ORCID

Anagnostis Argiriou http://orcid.org/0000-0002-9164-632X
Anastasia Athanasakoglou http://orcid.org/0000-0003-2607-6292
Emilia Grypioti http://orcid.org/0000-0002-0144-8160
Codruta Ignea http://orcid.org/0000-0002-6304-9413
Kriton Kalantidis http://orcid.org/0000-0002-6698-8563
Sotirios C. Kampranis http://orcid.org/0000-0001-6208-1684
Antonios M. Makris http://orcid.org/0000-0002-0661-7795
Sofia Michailidou http://orcid.org/0000-0002-8250-0513
Frederic Verret http://orcid.org/0000-0002-8175-7325

References

Akhtar TA, Surowiecki P, Siekierska H, Kania M, Van Gelder K, Rea KA, Virta LKA, Vatta M, Gawarecka K, Wojcik J *et al.* 2017. Polyprenols are synthesized by a plastidial *cis*-prenyltransferase and influence photosynthetic performance. *Plant Cell* 29: 1709–1725.

Allen AE, Dupont CL, Oborn´ık M, Hor´ak A, Nunes-Nesi A, McCrow JP, Zheng H, Johnson DA, Hu H, Fernie AR *et al.* 2011. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* 473: 203–207.

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M *et al.* 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306: 79–86.

Ast M, Gruber A, Schmitz-Esser S, Neuhaus HE, Kroth PG, Horn M, Haferkamp I. 2009. Diatom plastids depend on nucleotide import from the cytosol. *Proceedings of the National Academy of Sciences, USA* 106: 3621–3626.

Bajda A, Konopka-Postupolska D, Krzymowska M, Hennig J, Skorupinska-Tudek K, Surmacz L, W´ojcik J, Matysiak Z, Chojnacki T, Skorzynska-Polit E *et al.* 2009. Role of polyisoprenoids in tobacco resistance against biotic stresses. *Physiologia Plantarum* 135: 351–364.

Belt ST, Mu€ller J. 2013. The Arctic sea ice biomarker IP 25: a review of current understanding, recommendations for future research and applications in palaeo sea ice reconstructions. *Quaternary Science Reviews* 79: 1–17.

Bendtsen JD, Nielsen H, Von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology* 340: 783–795.

Berthelot K, Estevez Y, Deffieux A, Peruch F. 2012. Isopentenyl diphosphate isomerase: a checkpoint to isoprenoid biosynthesis. *Biochimie* 94: 1621–1634.

Bhattacharya D, Archibald JM, Weber APM, Reyes-Prieto A. 2007. How do endosymbionts become organelles? Understanding early events in plastid evolution. *BioEssays* 29: 1239–1246.

Bick JA, Lange BM. 2003. Metabolic cross talk between cytosolic and plastidial pathways of isoprenoid biosynthesis: unidirectional transport of intermediates across the chloroplast envelope membrane. *Archives of Biochemistry and Biophysics* 415: 146–154.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP *et al.* 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456: 239–244.

Burki F, Flegontov P, Oborn´ık M, Cihl´aˇr J, Pain A, Lukeˇs J, Keeling PJ. 2012. Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin. *Genome Biology and Evolution* 4: 626–635.

Chan CX, Reyes-Prieto A, Bhattacharya D. 2011. Red and green algal origin of diatom membrane transporters: insights into environmental adaptation and cell evolution. *PLoS ONE* 6: e29138.

Coesel S, Oborn´ık M, Varela J, Falciatore A, Bowler C. 2008. Evolutionary origins and functions of the carotenoid biosynthetic pathway in marine diatoms. *PLoS ONE* 3: e2896.

Cveji´c JH, Rohmer M. 2000. CO_2 as main carbon source for isoprenoid biosynthesis via the mevalonate-independent methylerythritol 4-phosphate route in the marine diatoms *Phaeodactylum tricornutum* and *Nitzschia ovalis*. *Phytochemistry* 53: 21–28.

Dagan T, Martin W. 2009. Seeing green and red in diatom genomes. *Science* 324: 1651–1652.

Dambek M, Eilers U, Breitenbach J, Steiger S, Bu€chel C, Sandmann G. 2012. Biosynthesis of fucoxanthin and diadinoxanthin and function of initial pathway genes in *Phaeodactylum tricornutum*. *Journal of Experimental Botany* 63: 5607–5612.

Davies FK, Jinkerson RE, Posewitz MC. 2015. Toward a photosynthetic microbial platform for terpenoid engineering. *Photosynthesis Research* 123: 265–284.

Demissie ZA, Erland LAE, Rheault MR, Mahmoud SS. 2013. The biosynthetic origin of irregular monoterpenes in lavandula: isolation and biochemical characterization of a novel *cis*-prenyl diphosphate synthase gene, lavandulyl diphosphate synthase. *Journal of Biological Chemistry* 288: 6333–6341.

Deschamps P, Moreira D. 2012. Reevaluating the green contribution to diatom genomes. *Genome Biology and Evolution* 4: 683–688.

Dorrell RG, Smith AG. 2011. Do red and green make brown? Perspectives on plastid acquisitions within chromalveolates. *Eukaryotic Cell* 10: 856–868.

Eilers U, Dietzel L, Breitenbach J, Bu€chel C, Sandmann G. 2016. Identification of genes coding for functional zeaxanthin epoxidases in the diatom *Phaeodactylum tricornutum*. *Journal of Plant Physiology* 192: 64–70.

Fabris M, Matthijs M, Carbonelle S, Moses T, Pollier J, Dasseville R, Baart GJE, Vyverman W, Goossens A. 2014. Tracking the sterol biosynthesis pathway of the diatom *Phaeodactylum tricornutum*. *New Phytologist* 204: 521–535.

Fabris M, Matthijs M, Rombauts S, Vyverman W, Goossens A, Baart GJE. 2012. The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner–Doudoroff glycolytic pathway. *The Plant Journal* 70: 1004–1014.

Falkowski PG. 2002. The ocean's invisible forest. *Scientific American* 287: 54–61.

Ferriols VMEN, Yaginuma R, Adachi M, Takada K, Matsunaga S, Okada S. 2015. Cloning and characterization of farnesyl pyrophosphate synthase from the highly branched isoprenoid producing diatom *Rhizosolenia setigera*. *Scientific Reports* 5: e10246.

Ferriols VMEN, Yaginuma-Suzuki R, Fukunaga K, Kadono T, Adachi M, Matsunaga S, Okada S. 2017. An exception among diatoms: unique organization of genes involved in isoprenoid biosynthesis in *Rhizosolenia setigera* CCMP 1694. *The Plant Journal* 92: 822–833.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281: 237–240.

Frommolt R, Werner S, Paulsen H, Goss R, Wilhelm C, Zauner S, Maier UG, Grossman AR, Bhattacharya D, Lohr M. 2008. Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Molecular Biology and Evolution* 25: 2653–2667.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29: 644–652.

Gruber A, Rocap G, Kroth PG, Armbrust EV, Mock T. 2015. Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *The Plant Journal* 81: 519–528.

Gruber A, Vugrinec S, Hempel F, Gould SB, Maier UG, Kroth PG. 2007. Protein targeting into complex diatom plastids: functional characterisation of a specific targeting motif. *Plant Molecular Biology* 64: 519–530.

Gruber A, Weber T, Bartulos CR, Vugrinec S, Kroth PG. 2009. Intracellular distribution of the reductive and oxidative pentose phosphate pathways in two diatoms. *Journal of Basic Microbiology* 49: 58–72.

Gschloessl B, Guermeur Y, Cock JM. 2008. Hectar: a method to predict subcellular targeting in heterokonts. *BMC Bioinformatics* 9: e393.

Guillard RRL. 1975. Culture of phytoplankton for feeding marine invertebrates. In: Smith WL, Chanley MH, eds. *Culture of marine invertebrate animals*. New York, NY, USA: Plenum Publishing, 29–60.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M *et al.* 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8: e1494.

Hemmerlin A, Hoeffler JF, Meyer O, Tritsch D, Kagan IA, Grosdemange-Billiard C, Rohmer M, Bach TJ. 2003a. Cross-talk between the cytosolic mevalonate and the plastidial methylerythritol phosphate pathways in tobacco bright yellow-2 cells. *Journal of Biological Chemistry* 278: 26666–26676.

Hemmerlin A, Rivera SB, Erickson HK, Poulter CD. 2003b. Enzymes encoded by the farnesyl diphosphate synthase gene family in the big sagebrush *Artemisia tridentata* ssp. *spiciformis*. *Journal of Biological Chemistry* 278: 32132–32140.

Henry LK, Gutensohn M, Thomas ST, Noel JP, Dudareva N. 2015. Orthologs of the archaeal isopentenyl phosphate kinase regulate terpenoid production in plants. *Proceedings of the National Academy of Sciences, USA* 112: 10050–10055.

Henry LK, Thomas ST, Widhalm JR, Lynch JH, Davis TC, Kessler SA, Bohlmann J, Noel JP, Dudareva N. 2018. Contribution of isopentenyl phosphate to plant terpenoid metabolism. *Nature Plants* 4: 721–729.

Holstein SA, Hohl RJ. 2004. Isoprenoids: remarkable diversity of form and function. *Lipids* 39: 293–309.

Ignea C, Cvetkovic I, Loupassaki S, Kefalas P, Johnson CB, Kampranis SC, Makris AM. 2011. Improving yeast strains using recyclable integration cassettes, for the production of plant terpenoids. *Microbial Cell Factories* 10: e4.

Ignea C, Trikka FA, Kourtzelis I, Argiriou A, Kanellis AK, Kampranis SC, Makris AM. 2012. Positive genetic interactors of *HMG2* identify a new set of genetic perturbations for improving sesquiterpene production in *Saccharomyces cerevisiae*. *Microbial Cell Factories* 11: e162.

Ignea C, Trikka FA, Nikolaidis AK, Georgantea P, Ioannou E, Loupassaki S, Kefalas P, Kanellis AK, Roussis V, Makris AM *et al.* 2015. Efficient diterpene production in yeast by engineering Erg20p into a geranylgeranyl diphosphate synthase. *Metabolic Engineering* 27: 65–75.

Kampranis SC, Ioannidis D, Purvis A, Mahrez W, Ninga E, Katerelos NA, Anssour S, Dunwell JM, Degenhardt J, Makris AM *et al.* 2007. Rational conversion of substrate and product specificity in a *Salvia* monoterpene synthase: structural insights into the evolution of terpene synthase function. *Plant Cell* 19: 1994–2005.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ *et al.* 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biology* 12: e1001889.

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* 305: 567–580.

Kroth PG, Chiovitti A, Gruber A, Martin-Jezequel V, Mock T, Parker MS, Stanley MS, Kaplan A, Caron L, Weber T *et al.* 2008. A model for carbohydrate metabolism in the diatom *Phaeodactylum tricornutum* deduced from comparative whole genome analysis. *PLoS ONE* 3: e1426.

Kroth P, Strotmann H. 1999. Diatom plastids: secondary endocytobiosis, plastid genome and protein import. *Physiologia Plantarum* 107: 136–141.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 33: 1870–1874.

Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, Valentin F, Wallace I, Wilm A, Lopez R *et al.* 2007. CLUSTALW and CLUSTALX version 2. *Bioinformatics* 23: 2947–2948.

Liang PH, Ko TP, Wang AHJ. 2002. Structure, mechanism and function of prenyltransferases. *European Journal of Biochemistry* 269: 3339–3354.

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI *et al.* 2015. CDD: NCBI's conserved domain database. *Nucleic Acids Research* 43: D222–D226.

Mass´e G, Belt ST, Crosta X, Schmidt S, Snape I, Thomas DN, Rowland SJ. 2011. Highly branched isoprenoids as proxies for variable sea ice conditions in the Southern Ocean. *Antarctic Science* 23: 487–498.

Mass´e G, Belt ST, Rowland SJ, Rohmer M. 2004. Isoprenoid biosynthesis in the diatoms *Rhizosolenia setigera* (Brightwell) and *Haslea ostrearia* (Simonsen). *Proceedings of the National Academy of Sciences, USA* 101: 4413–4418.

McFadden GI. 2001. Primary and secondary endosymbiosis and the origin of plastids. *Journal of Phycology* 37: 951–959.

Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D. 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* 324: 1724–1726.

NCBI Resource Coordinators. 2017. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 45: D12–D17.

Nelson DM, Tr´eguer P, Brzezinski MA, Leynaert A, Qu´eguiner B. 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.

Obata T, Fernie AR, Nunes-Nesi A. 2013. The central carbon and energy metabolism of marine diatoms. *Metabolites* 3: 325–346.

Oborn´ık M, Green BR. 2005. Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. *Molecular Biology and Evolution* 22: 2343–2353.

Prihoda J, Tanaka A, De Paula WBM, Allen JF, Tirichine L, Bowler C. 2012. Chloroplast–mitochondria cross-talk in diatoms. *Journal of Experimental Botany* 63: 1543–1557.

Rastogi A, Maheswari U, Dorrell RG, Vieira FRJ, Maumus F, Kustka A, McCarthy J, Allen AE, Kersey P, Bowler C *et al.* 2018. Integrative analysis of large scale transcriptome data draws a comprehensive landscape of *Phaeodactylum tricornutum* genome and evolutionary origin of diatoms. *Scientific Reports* 8: e4834.

Rivera SB, Swedlund BD, King GJ, Bell RN, Hussey CE, Shattuck-Eidens DM, Wrobel WM, Peiser GD, Poulter CD. 2001. Chrysanthemyl diphosphate synthase: isolation of the gene and characterization of the recombinant non-head-to-tail monoterpene synthase from *Chrysanthemum cinerariaefolium*. *Proceedings of the National Academy of Sciences, USA* 98: 4373–4378.

Rodr´ıguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, L€offelhardt W, Bohnert HJ, Philippe H, Lang BF. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Current Biology* 15: 1325–1330.

Rowland SJ, Belt ST, Wraige EJ, Mass´e G, Roussakis C, Robert JM. 2001. Effects of temperature on polyunsaturation in cytostatic lipids of *Haslea ostrearia*. *Phytochemistry* 56: 597–602.

Smith SR, Abbriano RM, Hildebrand M. 2012. Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways. *Algal Research* 1: 2–16.

Spanova M, Daum G. 2011. Squalene – biochemistry, molecular biology, process biotechnology, and applications. *European Journal of Lipid Science and Technology* 113: 1299–1320.

Surmacz L, Swiezewska E. 2011. Polyisoprenoids – secondary metabolites or physiologically important superlipids? *Biochemical and Biophysical Research Communications* 407: 627–632.

Tarshis LC, Proteau PJ, Kellogg BA, Sacchettini JC, Poulter CD. 1996. Regulation of product chain length by isoprenyl diphosphate synthases. *Proceedings of the National Academy of Sciences, USA* 93: 15018–15023.

Timmis JN, Ayliff MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics* 5: 123–135.

Ver´ıssimo A, Bassard J-E, Julien-Laferri`ere A, Sagot M-F, Vinga S. 2017. MASSBLAST: a workflow to accelerate RNA-seq and DNA database analysis. *bioRxiv*. doi: 10.1101/131953.

Verwaal R, Wang J, Meijnen JP, Visser H, Sandmann G, Van Den Berg JA, Van Ooyen AJJ. 2007. High-level production of beta-carotene in *Saccharomyces cerevisiae* by successive transformation with carotenogenic genes from *Xanthophyllomyces dendrorhous*. *Applied and Environmental Microbiology* 73: 4342–4350.

Vranov´a E, Coman D, Gruissem W. 2012. Structure and dynamics of the isoprenoid pathway network. *Molecular Plant* 5: 318–333.

Wang K, Ohnuma S. 1999. Chain-length determination mechanism of isoprenyl diphosphate synthases and implications for molecular evolution. *Trends in Biochemical Sciences* 24: 445–451.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* 18: 691–699.

Wraige EJ, Johns L, Belt ST, Mass´e G, Robert J-M, Rowland S. 1999. Highly branched C25 isoprenoids in axenic cultures of *Haslea ostrearia*. *Phytochemistry* 51: 69–73.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular Biology and Evolution* 21: 809–818.

Zhang Z, Sachs JP, Marchetti A. 2009. Hydrogen isotope fractionation in freshwater and marine algae: II. Temperature and nitrogen limited growth rate effects. *Organic Geochemistry* 40: 428–439.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Overview of isoprenoid biosynthesis highlighting the different stages of the pathway (early, central, final steps).

Fig. S2 Chemical structures of some of the isoprenoids synthesized by *H. ostrearia.*

Fig. S3 HBI profile of *H. ostrearia* NCC 153.8 strain.

Fig. S4 Phylogenetic relationship of prenyltransferases from diatoms.

Fig. S5 Expression patterns of genes involved in isoprenoid biosynthesis as detected by RNA-seq.

Fig. S6 Multiple sequence alignment of *Ho*PTS1 with known farnesyl diphosphate synthases from other species, and multiple

sequence alignment of *Ho*PTS2-5 with known polyprenyl diphosphate synthases from other species.

Fig. S7 Phylogenetic trees of MEP pathway proteins.

Fig. S8 Phylogenetic trees of prenyltransferases and squalene/phytoene synthase family members.

Methods S1 Supplementary methods.

Notes S1 Extended discussion on the phylogenetic analysis.

Notes S2 Diatom sequences.

Table S1 IDs of the transcriptomes from the MMETSP used in the phylogenetic analysis.

Table S2 List of primers used in this study.

Table S3 Identification of *Ho*PTS1-*Ho*PTS5 homologues in other diatom species.

Table S4 Genomic location of isoprenoid biosynthetic genes in the *P. tricornutum, T. pseudonana* and *F. cylindrus.*

Table S5 Summarized results from predictions of protein subcellular localization.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.