

Ghent University - Faculty of Sciences

Department of Plant Biotechnology and Bioinformatics

VIB - Department of Plant Systems Biology

**A first look at the genetic control mechanisms
shaping metabolic changes during nitrogen
starvation in *Phaeodactylum tricornutum***

Michiel Matthijs

Thesis submitted in fulfilment of the requirements
for the degree of Doctor (PhD) in Sciences:
Biotechnology and Biochemistry

Academic year: 2013-2014

Promoter: Prof. Dr. Alain Goossens

Co-promoter: Prof. Dr. Wim Vyverman

This work was conducted at the VIB Department of Plant Systems Biology, Ghent University and at the Department of Biology – Laboratory of Protistology and Aquatic Ecology, Ghent University.



侘寂

(Wabi-sabi) Japanese term describing the aesthetic of things that are imperfect, impermanent and incomplete

Members of the Examination Board

Prof. Dr. Sofie Goormachtig (chairwoman)

VIB Department of Plant Systems Biology, Ghent University

Prof. Dr. Alain Goossens (promotor)

VIB Department of Plant Systems Biology, Ghent University

Prof. Dr. Wim Vyverman (co-promotor)

Department of Biology, Ghent University

Prof. Dr. Bart Devreese*

Department of Biochemistry and microbiology, Ghent University

Dr. Marc Heijde*

VIB Department of Plant Systems Biology, Ghent University

Prof. Dr. Peter Kroth*

Faculty of Biology, University of Konstanz

Prof. Dr. Bartel Vanholme*

VIB Department of Plant Systems Biology, Ghent University

Prof. Dr. Filip Rolland*

Department of Biology, University of Leuven

**Members of the Reading Committee*

Table of Contents

List of abbreviations	7
Chapter 1	
Summary	9
Chapter 2	
Scope & objectives	15
Chapter 3	
Introduction	18
Chapter 4	
RNA sequencing during nitrogen starvation reveals NMB1 as a new type of transcription factor in <i>Phaeodactylum tricornutum</i>	45
Chapter 5	
Coordinated expression of the TCA cycle enzymes during nitrogen starvation is guided by the transcription factor bZIP14	72
Chapter 6	
Materials and methods	100
Chapter 7	
Conclusions and perspectives	108
Chapter 8	
Acknowledgements	119
Chapter 9	
Curriculum vitae, primer list and other papers by the author	123

List of abbreviations

2-OG	2-Oxoglutarate
3-AT	3-Amino-1,2,4-Triazole
AMP	Adenosine monophosphate
AMPK	AMP-activated Protein Kinase
APS	5'-adenylsulphate
ATP	Adenosine triphosphate
bHLH	basic Helix Loop Helix
bZIP	Basic Leucine Zipper
CDK	Cyclin Dependent Kinases
CI	Confidence Interval
CoA	Coenzyme-A
CV	Coefficient of Variation
DHLP	Dihydrolipoamide dehydrogenase
DMS	Dimethylsulphide
DMSP	Dimethylsulphonioacetate
EMSA	Electrophoretic Mobility Shift Assay
EST	Expressed Sequence Tag
FCP	Fucoxanthine Binding Proteins
FPKM	Fragments Per Kilobase Of Exon Per Million Fragments Mapped
GOGAT	Glutamine oxoglutarate aminotransferase
HSF	Heat Shock Factor
IPTG	Isopropyl β -D-1-thiogalactopyranoside
kDa	Kilodalton
LUC	Luciferase
N	Nitrogen
NADH	Nicotinamide adenine dinucleotide
NADPH	Nicotinamide Adenine Dinucleotide Phosphate
NiR	Nitrite Reductase
NMB	Nitrogen Binding Motif
NR	Nitrate Reductase
OD	Optical Density
QPCR	Quantitative PCR
TCA	Tricarboxylic acid cycle
TF	Transcription Factor
TOR	Target of Rapamycin
Ub	Ubiquitin
Y1H	Yeast One Hybrid

Chapter 1:

Summary

NEDERLANDSE SAMENVATTING

Diatomeeën zijn unicellulaire algen die ecologisch succesvol zijn in aquatische omgevingen. Ze vormen één van de basisschakels van mariene voedselnetwerken en leveren een aanzienlijke bijdrage aan de zuurstofproductie op deze planeet. De meeste diatomeeën hebben een glazen celwand die buitengewoon symmetrisch is en reeds in de 19^{de} eeuw microscopisten kon fascineren. Diatomeeën bezitten een potentieel commercieel belang, onder andere als producent van lipiden rijk aan poly-onverzadigde omega-3-vetzuren. Diatomeeën accumuleren vooral lipiden onder condities die hen niet toelaten zich verder te vermenigvuldigen. Eén van de best bestudeerde omstandigheden waarbij dit gebeurt, is wanneer er onvoldoende nutriënten aanwezig zijn, zoals een tekort aan stikstof. In hoofdstuk drie worden deze reacties op nutriëntdeprivatie samengevat en gecontrasteerd met groene algen. De moleculaire studie van diatomeeën kwam in een stroomversnelling na het publiceren van het genoom van de twee modeldiatomeeën, waaronder de pennaat *Phaeodactylum tricornutum*. Deze diatomee kan getransformeerd worden en er is een beperkte, maar groeiende hoeveelheid moleculaire hulpmiddelen beschikbaar. In dit werk werd gebruik gemaakt van *P. tricornutum* om met behulp van RNA-sequencing te bepalen welke veranderingen de cel doormaakt op transcriptoomniveau als er onvoldoende stikstof beschikbaar is, waarbij onderzocht werd welke metabolische aanpassingen plaatsvinden. Hierbij vergeleken we drie verschillende stresscondities.

De veranderingen in de cel worden gestuurd door controlegenen. Op RNA-niveau wordt deze controle voornamelijk uitgevoerd door transcriptiefactoren. In hoofdstuk vier wordt de zoektocht beschreven naar transcriptiefactoren die de betrokken genen aansturen. Bij een *de novo* zoektocht naar cis-regulatorische elementen in genen responsief tijdens stikstofgebrek werden verscheidene motieven opgepikt. Door gebruik te maken van een yeast-one-hybridscreening werd de transcriptiefactor NMB1 geïdentificeerd, die één van deze motieven bindt. Deze factor behoort tot een voordien onbekende familie die betrokken is bij stikstofgebrek en een expressiepatroon bezit dat sterk geïnduceerd wordt bij stikstofgebrek.

In hoofdstuk vijf gaan we in op de opvallendste verandering in onze transcriptoomdataset: de gecoördineerde toename van genen betrokken bij de citroenzuurcyclus tijdens stikstofstarvatie. Deze cyclus betreft een centrale rol bij het primaire metabolisme. Door de werking van de citroenzuurcyclus kunnen enerzijds suikers, aminozuren en vetten worden omgezet in reducerende equivalenten, die op hun beurt in energie kunnen leveren. Anderzijds zijn veel intermediären van de cyclus eveneens voorlopers van aminozuren en ander moleculen. In dit werk werd aangetoond dat er meerdere intermediären van de citroenzuurcyclus in hoge concentraties aanwezig zijn en dat de koolstof in deze moleculen voornamelijk bestaat uit eerder opgeslagen producten.

Aangezien het expressiepatroon van de genen betrokken in de citroenzuurcyclus gelijkaardig was, werd verondersteld dat een actie van een beperkte groep transcriptiefactoren dit veroorzaakt. Door co-expressieanalyse werd de transcriptiefactor bZIP14 geïdentificeerd. Deze transcriptiefactor vertoonde een

gelijkaardig patroon als de genen van de citroenzuurcyclus. Door gebruik te maken van een proteïnebindende microarray, werd het motief gebonden door deze factor achterhaald. Dit motief bleek aanwezig te zijn in verscheidene promotoren van citroenzuurcyclusgenen. Bij overexpressie van bZIP14 bleken de genen van de citroenzuurcyclus ook toe te nemen, een duidelijke indicatie dat de transcriptiefactor betrokken is in dit proces.

ENGLISH SUMMARY

Diatoms are unicellular algae that are successful in both fresh and marine environments. Their photosynthesis supplies a fifth of all the oxygen we breathe and forms a large part of the primary oceanic productivity that feeds our seas and ultimately us. Diatoms have exquisitely symmetric cell walls that have fascinated microscopists since the 19th century. Besides their ecological importance, they also have a commercial application since, among other things, they produce lipids rich in polyunsaturated omega-3 fatty acids. This accumulation mainly occurs during conditions adverse for growth like when a nutrient such as nitrogen is no longer present in sufficient quantities to support growth. In chapter three, the known reactions to macronutrient starvation are summarized and they contrast to those of green algae.

The arrival of next generation sequencing has accelerated the molecular study of diatoms by making the sequencing of genomes and transcriptomes fast and affordable. The molecular toolbox for *P. tricornutum* is limited but it can be transformed and the genome is publically available. In this study, the model diatom *Phaeodactylum tricornutum* is used to further the understanding of metabolic and transcriptomic changes during nitrogen starvation. These findings were contrasted with three other stress conditions.

The aforementioned changes are steered by a set of controlling genes that was virtually unknown. On the RNA level, these changes are mainly steered by transcription factors (TF's). In chapter four, the search is described for transcription factors that steer adaptations during N starvation. The problem was approached by identifying overrepresented cis-regulatory motifs in genes responsive to N starvation. Using a yeast one hybrid screening, a transcription factor termed NMB1 was identified which binds one of these motifs and is induced during N scarcity. This gene is part of a previously unknown transcription factor family conserved in heterokonts.

In chapter five, the most striking find of our RNA-seq dataset is analyzed: a coordinated upregulation of the genes involved in the Krebs or tricarboxylic acid (TCA) cycle. This cycle plays a central role in the primary metabolism as it is able to turn sugars, lipids and amino acid into reducing equivalents that can be turned into energy. The TCA cycle can also move in the reverse reaction providing the carbon skeletons for synthesizing these molecules.

In this work it was shown that the intermediates of the TCA cycle mainly consist out of carbon that was already present in the cell.

Since the pattern of transcription for these enzymes was so similar, it was hypothesized that there was a common transcription factor steering the expression all these genes. Chapter five also details the identification of the conserved transcription factor bZIP14 by co-expression analysis. This transcription factor had a similar expression pattern as several TCA enzymes and was induced during N starvation. Using a protein binding microarray, the motif bound by bZIP14 was identified, which was present in

several promoters of TCA genes. Upon overexpression of this TF several genes involved in the TCA cycle were also upregulated, a clear indication that this factor is involved in the regulation of the Krebs cycle.

Chapter 2:

Scope & Objectives

When observing the landmasses of earth from space several large areas of green rainforest or fields can be seen, but when we turn to the oceans there appears to be only blue. Looks can be deceiving as marine photosynthesis contributes almost half of the atmospheric oxygen and is crucial for our ecosystem and economy. Nevertheless all marine algae combined would make up only 0.2% of the biomass of land plants (Falkowski, Barber, & Smetacek, 1998). However, when we zoom in a bit closer on the oceans there do appear small patches of color, often only temporary. These patches are massive microalgal blooms that are almost exclusively caused by unicellular chromalveolates. The reason why these chromalveolates, often diatoms, outcompete all others has could be linked with their capacity to efficiently uptake nutrients.

The genome projects of the past five years start to shed light on the metabolic adaptations that make diatoms so successful. Because most enzymes are highly conserved on the primary sequence level it is often possible to reliably assign functions based solely on the coding sequence. Conversely, the regulation of metabolic processes is usually highly variable between species. Enzyme activities are strongly regulated at the transcriptional and post-transcriptional level in complex and robust signaling networks within where timing and dosage of gene expression is also important. As a consequence, the ectopic expression of a single enzyme is often insufficient to increase levels of a metabolite; the enzyme must work in a cellular context that promotes the production of the metabolite. Commonly several metabolic genes are under the control of a limited number of transcription factors and targeting these factors for overexpression or silencing can lead to large increases in the accumulation of the target product. With the advent of RNA-sequencing it has become easier than ever before to characterize the transcriptome of model and non model species. This enables us to identify which genes are altered at the transcriptional level under a particular condition, to search for genes under common regulation and, ultimately, to give hints on the regulators themselves.

The aim of this PhD work, was to understand the metabolic reprogramming the model diatom *Phaeodactylum tricornutum* undergoes during periods of nutrient starvation. It has been well established that during nutrient deprivation lipids and carbohydrates are accumulated in the diatom cell. The focus of the work was on nitrogen starvation as it is both ecologically and economically relevant. At the start of this project only tentative modeling of the carbohydrate metabolism and the canonical lipid biosynthesis were known. Within the SUNLIGHT SBO project, a database was made with all likely enzymatic reactions occurring in *Phaeodactylum tricornutum*. The regulation and signaling within the cell was still completely unknown. It was our aim to identify transcription factors that steer the reprogramming of the diatom cell from exponential growth to lipid accumulation. To achieve this goal RNA-sequencing was coupled to metabolic profiling. By identifying the patterns in expression and metabolism, the regulatory genes steering this would become apparent as we expected them to move in a similar manner as the pathways they regulate. The original focus was to find the transcriptional regulators of lipid biosynthesis, along the way this has been extended to primary metabolism in general.

Falkowski, P. G., Barber, R. T., & Smetacek, V. V. (1998). Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science*, 281(5374), 200-207.

Chapter 3:

The agile algae: adaptations of diatoms during nutrient starvation

Manuscript in preparation for publication

Author contribution:

MM wrote this chapter

ABSTRACT

Diatoms are unicellular, heterokont algae that are important contributors to aquatic primary biomass production. When there are sufficient light and nutrients, they outcompete other species of algae, but in nutrient poor conditions, they are not as numerous but able to rapidly increase in numbers when conditions improve. In this review, we look at how diatoms adapt their metabolism to these unfavourable conditions and what properties help them to rapidly resume growth when a nutrient influx occurs. The analysis of both DNA- and RNA-sequencing data indicates that mechanisms employed by these organisms during nutrient deprivation are markedly different from those employed by green plants or algae. Understanding the starvation response of diatoms has industrial applications, since diatoms store lipids under adverse conditions that can be used for food or fuel. Insight in the nutrient starvation program is therefore key to improve the production of these storage products. To date, the signalling pathways involved remain elusive, but the unique metabolic reprogramming has become apparent and is markedly different from green algae. Here, the current knowledge on the signalling pathways likely to be involved in nutrient stress responses is summarized and contrasted with the response of the green lineage of photosynthetic organisms.

INTRODUCTION

NUTRIENT RESTRICTION LIMITS CARBON FIXATION IN AQUATIC ENVIROMENTS

In contrast to terrestrial environments where primary production is dominated by green lineage organisms, photoautotrophic carbon fixation in oceanic environments is carried out by photosynthetic organisms with a diverse evolutionary background (fig. 1). Besides green algae, prokaryotic cyanobacteria, red algae and a large number of chlorophyll *c* containing protists can all be found in marine environments. These groups derive from several independent secondary endosymbiosis events, in which a heterotrophic host took up a red alga. In most species, this red algal cell has been reduced to a chloroplast with a small genome, with remnants of its origin being the four membranes that surround it instead of the two membranes of green chloroplasts. Besides diatoms, these mainly marine organisms also comprise other well-known algae lineages such as dinoflagellates, cryptophytes and haptophytes. Originally all these phyla were grouped in the chromalveolata, because it was assumed that the engulfment of the red symbiont happened only once. This hypothesis has been contested and recently a paraphyletic origin was proposed, which is gaining more and more acceptance (Baurain et al., 2010). Diatoms are usually grouped in the line of heterokonts that also contains brown seaweeds and the parasitic water moulds or oomycetes of Irish potato blight infamy. Of all mentioned photosynthetic plankton species, diatoms are among the most important, since a fifth of the primary productivity of the ocean is produced by diatoms (Armbrust, 2009). This accounts for 20% of the oxygen we breathe.

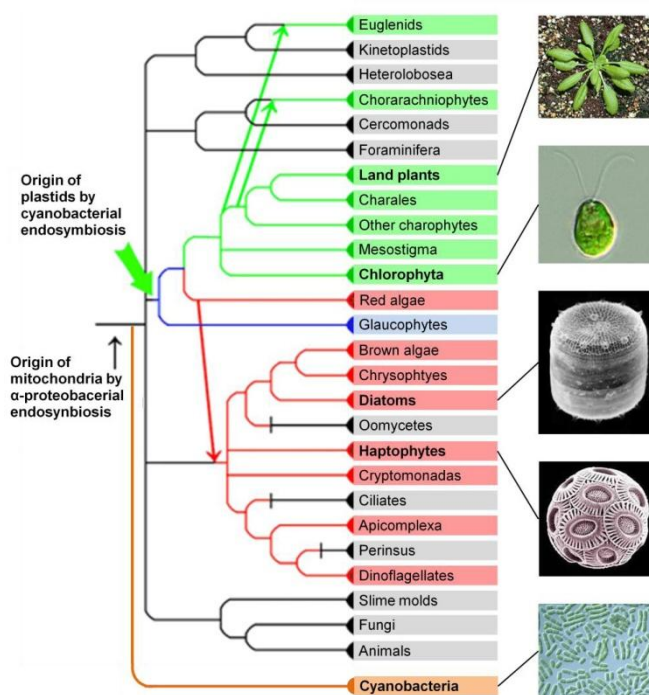


Figure 1: Schematic phylogenetic tree of the major photosynthetic phyla described in this paper, branch length has no relation with evolutionary distance (Hockin, 2011)

The primary production of the oceans depends on the availability of light and nutrients. In this review, we focus on the latter and summarize what molecular knowledge has been gained in the genomic age. In most marine and freshwater environments the limiting factors are nitrogen and phosphorous, whereas iron is limiting in 30-40% of the open ocean (Armbrust, 2009). A unique feature of diatoms is that they also require silica for their cell walls and this often becomes the limiting nutrient during blooms (Dortch & Whitley, 1992). Organisms that prosper in these environments must therefore be able to cope with long periods of nutrient starvation and to rapidly assimilate any available nutrients, often while in competition with other organisms. Diatoms thrive in nutrient rich water, where they are able to outcompete most other species of algae (Furnas, 1990).

Because of their abundance, diatoms are important components in several geochemical cycles. One of the most visible differences with other algae is their silica cell wall. Since dissolved silica is mainly consumed by diatoms, they have a dominant impact on the geochemical silicon cycle (Ragueneau et al., 2000). Due to their heavy silica cell wall, diatoms sediment to the ocean floor after death, taking assimilated carbon with them. In fact, petroleum reserves were formed out of these carbon deposits and burning them releases carbon stored over millions of years, changing global CO₂ levels. Increasing diatom growth, by fertilizing the open ocean with iron, would sequester carbon and is one of the possible strategies to mitigate climate change (De Baar et al., 2005).

Besides carbon, the geochemical cycle of nitrogen has also been altered anthropogenically by the fixation of nitrogen in the Haber-Bosch process. The open ocean is still largely unaffected, but coastal waters and

freshwater habitats have been enriched by phosphorous and nitrogen run-off from agriculture to the extent that the nutrient limitation has been lifted and massive seasonal algal blooms are both increasing in duration and size(Heisler et al., 2008). Some bloom forming organisms produce toxins that affect fisheries, aquaculture and tourism. Additionally, the oxygen demands for the decomposition of these blooms can result in oxygen deprived ‘dead zones’. Blooms can be formed by dinoflagellates, haptophytes, as well as diatoms such as *Chaetoceros* and *Pseudo-Nitzschia*(Riegman, Noordeloos, & Cadée, 1992). This review will not focus on dinoflagellatenutrient responses, because their genomic organization and evolutionary history appears to be markedly different and this has been covered elsewhere(Steve Dagenais-Bellefeuille & David Morse, 2013).

Aside from carbon capture, the other main driver in diatom nutrient deprivation research is the potential for biofuel production. The aquatic species program of the 1970’s concluded that diatoms were among the most promising organisms for economical biofuels(Sheehan, Dunahay, Benemann, & Roessler, 1998). Interest declined with dropping oil prices, but in the last decade rising prices brought algal oil once more to the forefront. This renewed interest combined with advances in ‘omics’ techniques has provided an increasing momentum for diatom research.

While algal research has a long history, e.g. the Calvin-Benson cycle was for a large part elucidated in the *Chlorella* algae, most studies have been done on green algae, such as *Chlamydomonas reinhardtii*, and comparatively little attention had been paid to non-green algae (Benson, 2002). While a lot of work has been performed on diatom photosynthesis, most of the acquired knowledge is still extrapolated from green algae. Besides obvious differences in pigment composition, the unique metabolism of diatoms during nutrient starvation will be illustrated by a number of examples in this review. Recent genomic, proteomic and metabolomic projects have shown that many aspects of photoauxotrophic metabolism are different in diatoms, e.g. the likely presence of a carbon concentrating mechanism in diatoms that often cannot be fitted to the classical C3 or C4 pathways(Hopkinson, Dupont, Allen, & Morel, 2011). Furthermore, the localization of enzymes seems to differ in many pathways, including nucleotide synthesis, glycolysis and the pentose phosphate pathway(Ast et al., 2009; Peter G Kroth et al., 2008).

Outlining the features of diatom primary metabolism and nutrient acquisition strategies in general is still difficult, since it is dangerous to draw general conclusions from the scarce number of species that have been studied using ‘omics’ approaches, especially taking into account the enormous species diversity and the wide range of habitats in which they live. An example is the differences in auxotrophy for vitamins or the ability to grow on a external carbon source in the absence of light(Lewin & Lewin, 1960).

Additionally, the environmental heterogeneity of the oceans is not represented proportionally in the sequence databases, although this gap is closing. Compared to photosynthetic protists, far more plant species have sequenced genomes. The sequenced diatoms of today represent mainly coastal species (*T. pseudonana*, *P. tricornutum*) with one arctic diatom (*F. cylindrus*) and one centricate of the open ocean (*T.*

oceanica)(Armbrust et al., 2004; Bowler et al., 2008b; Markus Lommer et al., 2012). Important gaps are tropical and freshwater diatoms as they have no complete genomes to date. Sequencing data on other heterokonts besides diatoms are scarce with the exceptions of the pelagophyte *Aureococcus anophagefferens* and the eustigmatophyte *Nannochloropsis gaditana*(Christopher J. Gobler et al., 2011; Radakovits et al., 2012).

Even before the arrival of sequencing data, it was clear that diatoms have distinct properties from green algae. For example, several diatoms survive long periods of darkness relatively unscathed, whereas green algae succumb relatively rapidly(Berges & Falkowski, 1998). Furthermore, resting spores can be formed when conditions are unfavourable, after sinking to a sediment for example, and metabolism can resume when the situation improves(Smetacek, 1985). Recently it has been shown that a diatom was able to use nitrate as an electron acceptor in anoxic conditions, something that was seen previously only in the realm of prokaryotes(Kamp, de Beer, Nitsch, Lavik, & Stief, 2011).

Many algae, with very different evolutionary origins, store lipids or carbohydrates during adverse growing conditions(Sheehan et al., 1998). One of the main triggers for the accumulation of storage components is the depletion of nutrients such as silicon, nitrogen or phosphate. Much of the recent research in algae has been spurred by the need to increase oil yields and strategies to do this have been extensively described elsewhere(Mühlroth et al., 2013). Next to the macro-nutrients, many diatoms have a requirement for vitamins, such as B12, which they usually obtain from bacteria(Bertrand et al., 2012). Optimal growth also requires several trace metals, such as copper and even cadmium, however this review will focus on the macro-nutrients(Lane & Morel, 2000).

THE OMICS SHIFT HAS GIVEN A NEW DRIVE TO DIATOM MOLECULAR RESEARCH

With rapidly decreasing sequencing costs, the number of sequenced non-green algae has steadily increased over the past few years. Novel techniques, such as RNA-sequencing and shotgun proteomics, give insights on a scale that is unprecedented, since they rapidly reveal the genetic make-up of the organism(Koboldt, Steinberg, Larson, Wilson, & Mardis, 2013). It has never been easier to study non-model organisms on a molecular and systems level with the arrival of ‘omics’ techniques, because they allow the collection of large datasets, even when no prior genetic information was available, as exemplified by a study in *Chlorella vulgaris*(Guarnieri et al., 2011). It can be more cost-effective to use transcriptomic studies in a phylum as extensive, but underexplored as the heterokonts because it requires less sequencing depth. This approach has been successfully applied by the CAMERA and TARA oceans projects(Karsenti et al., 2011; Keeling et al., 2014). While RNA-sequencing will only show actively transcribed genes, it requires less effort and is able to provide a snapshot of an organism during a particular state or growth condition(Rismani-Yazdi, Haznedaroglu, Hsin, & Peccia, 2012). The sensitivity of RNA-seq has made it possible to take large-scale sequencing outside of the lab. Metagenomics or -transcriptomics is the process of sequencing organisms

in the wild without culturing. Until now, it has mostly been focused on prokaryotes, but there is no technical reason why eukaryotes cannot be investigated in the same way. Even metaproteomic approaches are feasible (Morris et al., 2010). One of the major advantages of metatranscriptomics and –proteomics over metagenomics is that they give insight into which processes are currently taking place in a certain environment, at a certain time. While we are still far from a complete picture but it will soon be possible to take a snapshot of the metabolism of a certain environment by looking at which genes are actively transcribed or translated.

In-depth sequencing of diverse species combined with metagenomics will shed light on the currently unclear evolutionary relationships between diatoms, haptophytes and other algal species that arose from secondary endosymbiosis. The role of horizontal gene transfer, which is common in eukaryotic algae, is still unclear (Bowler et al., 2008a). For example: a recently sequenced early green alga and a chromalveolate, to which diatoms also belong, demonstrate that genes that were thought to be distinctively red or green are combined in unexpected ways. This suggests that what was thought to be horizontal gene transfer might only be due to the undersampling of genomic diversity in photosynthetic organisms (Curtis et al., 2012).

A sequenced genome, however, remains an advantage, since it provides an overview of all possible reactions, and many genes might not be detected or incorrectly assembled on the basis of transcriptomic data alone. Most sequencing experiments have been done on *Thalassiosira pseudonana* or *Phaeodactylum tricornutum* because of their small genomes (34 and 30 Mb resp.) and the established body of literature (over 8000 and 13,000 works resp., in Google Scholar). Both contain approximately ten thousand genes but the arctic diatom *Fragiliariopsis cylindrus* contains an estimated 27000 genes. Because of the evolutionary distance between organisms, it remains to be seen if the observations made in the literature cited below are applicable to most species and provide just a glimpse of the nutrient starvation response. Even with the current high throughput techniques, it will take several years to sample even a fraction of algal diversity, since in diatoms alone there are over 200,000 expected species (Armbrust, 2009). The task will be especially daunting since *P. tricornutum* shares only half of its genes with *T. pseudonana*, illustrating the fact that a large fraction of genes appears to be species specific (Bowler et al., 2008a). It should also be noted that *T. pseudonana* and *P. tricornutum* are both atypical diatoms, for which sexual reproduction has not been observed. Both models can maintain a constant valve size during asexual cell division, while other diatoms grow progressively smaller. Because of their constant size, they are not forced to undergo sexual reproduction and are easy to maintain in culture. Silicification of the cell wall is even optional in *P. tricornutum* and it does not have a significant ecological occurrence compared to other genera such as *Chaetoceros* or *Pseudo-Nitzschia* (Ishii, Iwataki, Matsuoka, & Imai, 2011; Trainer et al., 2012). The number of diatom genomes is sure to increase in the coming years, since sequencing costs drop, but it should not be forgotten that annotation of gene models is also important and the time required for this is not likely to decrease.

The genomic data of five sequenced heterokonts is available together with a wide range of green algal and plant genomes comprehensively bundled in the pico-PLAZA website (Klaas Vandepoele et al., 2013). It is especially useful for orthology relationships, but *T. oceanica* data is lacking, as well as that of non-green species such as *Nannochloropsis*. The Small Read Archive (SRA) of NCBI is a collection of RNA-sequencing data currently holding 292 transcriptomic datasets from diatoms. Most of these were generated by the Microbial Eukaryote Sequencing Project (Keeling et al., 2014). Gathering the data is only the first step, since making comparisons is equally important. Regrettably, there is no specialized portal for obtaining diatom genomic and transcriptomic data and combining available data has been non-trivial until now. Hopefully, this gap will be filled soon by the community. Given the large number of unknown genes, this is unfortunate, since sufficiently large co-expression clusters across different conditions often hint at the gene function (Vandepoele, Quimbaya, Casneuf, De Veylder, & Van de Peer, 2009). For ease of comparison, these co-expression studies should be done in a single species. A successful example of such a co-expression database is the *Arabidopsis thaliana* focused ATTED-II which contains over a thousand datasets from both micro-arrays and sequencing experiments. This is a staggering number, and since sequencing costs continue to drop, it might only be a few years until over a hundred RNA-seq experiments have been performed on either *P. tricornutum* or *T. pseudonana*. While this review focuses on transcriptomic and genomic data, it should not be forgotten that both proteomics and metabolomics have also advanced tremendously. Both of these techniques give insight on levels closer to the observed phenotypes and have provided important clues in diatom specific metabolism and protein responses.

GENES RESPONSIVE TO NUTRIENT STARVATION

The available proteome and transcriptome surveying techniques allow a quick overview in the adaptations during stress. Available studies show that nutrient deprivation results in large-scale reprogramming of the cell. A comparison of different stresses shows that for diatoms a third of all transcripts change upon nitrogen starvation, 10% during iron starvation and this number is even more limited for phosphorous (A. E. Allen et al., 2008b; Kimberlee Thamatrakoln, Olga Korenovska, A Kalani Niheu, & Kay D Bidle, 2012; Valenzuela et al., 2012). Regardless of which nutrient is limited, some responses are shared among most stresses: 1) a reduction in photosynthetic efficiency, 2) increased expression of assimilation and transport genes for the limiting nutrients, and 3) a halt in the cell cycle. The most atypical response is that to silicon, which will be discussed later in the text.

All proteomic and transcriptomic studies to date show that there is a large decrease in photosynthetic activity upon nutrient stress, being the least pronounced during phosphate and silicon starvation. Since photosynthesis is a process extensively depending on protein synthesis, this is not unexpected for nitrogen and neither for iron starvation, since many proteins rely on reducing equivalents provided by the iron containing ferredoxin or require iron to function like nitrite reductase. However carbon assimilation during starvation continues, because the C:N ratio continues to increase. The role of alternative carbon

fixating pathways in storage product accumulation is still unclear, since some of the carbon concentrating mechanisms seem to increase in activity while photosynthesis is reduced, hinting that this carbon might end up in lipids or sugars.

In all investigated diatoms, genes are packed closely together on the genome and are often organized in clusters that participate in the same process and respond in a coordinated manner (A. E. Allen et al., 2008b; Sapriel et al., 2009). While the number of transcriptomic studies is rapidly increasing, the overlap between different studies is relatively small, as shown in figure 2. Several RNA-sequencing studies have been published for nitrogen starvation in *P. tricornutum* and there was only a 15% overlap (Valenzuela et al., 2012; Z. K. Yang et al., 2013). In the case of two iron starvation experiments in *T. pseudonana*, there was an extensive overlap, but the amount of differentially expressed genes differed tenfold, while silicon deprivation studies showed that less than 10 genes responded in a similar fashion (R. P. Shrestha et al., 2012).

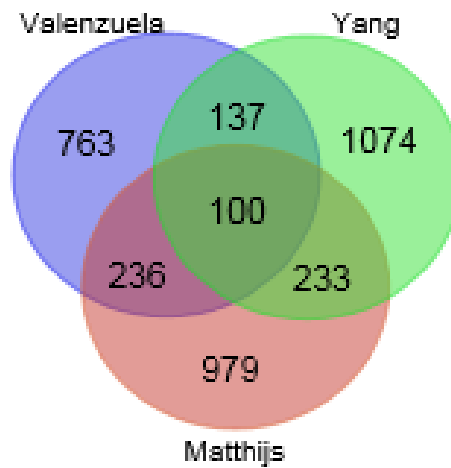


Figure 2: Overlap between three nitrogen deprivation transcriptomic studies on *P. tricornutum*, data was manually curated from the lists reported by the authors

Partial explanations may be the different culture media and growth conditions, and sampling and sample processing protocols that were used. Furthermore, the used strains were not always identical. Diatoms are known to mutate rapidly and in *P. tricornutum* there appears to be a large amount of point mutations in the RNA-sequencing datasets compared to the reference genome (Matthijs et al., in preparation). Interestingly, it has been shown that many of the genes under positive selection in *Thalassiosira sp.* are related to nutrient deprivation (Koester, Swanson, & Armbrust, 2012).

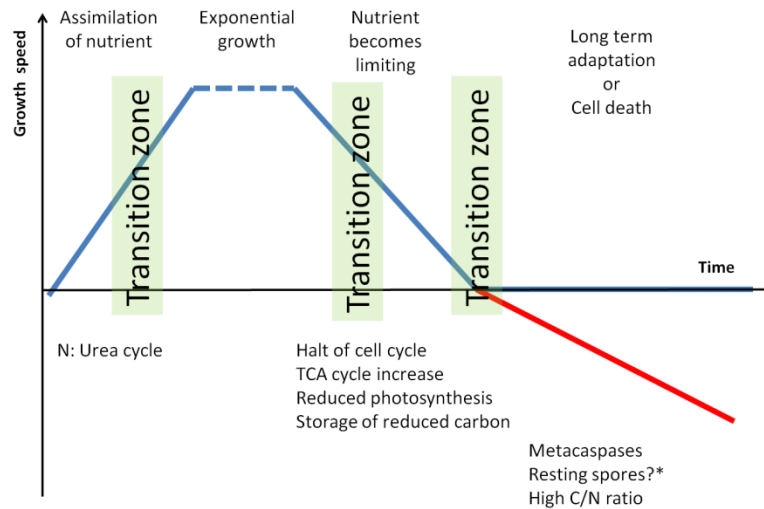


Figure 3: Course of nutrient stress response

This discrepancy in observations is not unique to algae, because similar observations have been made in *Arabidopsis thaliana* (Massonnet et al., 2010). In this study, different knockout lines with altered drought resistance showed phenotypic variations depending on the labs where they were grown, and transcriptome profiles differed even more, indicating that care must be taken when interpreting results.

The method of sample collection can also impact results, e.g. centrifugation will deprive the cells of light and create unwanted changes. The speed of these responses should not be underestimated as was demonstrated by the *P. tricornutum* *dsCyc2* gene that a 5 minute exposure to the light was enough to upregulate gene over a hundred fold (M. J. Huysman et al., 2010). Cultures can unwittingly be synchronized when they are transferred to new, specific medium in a manner independent of circadian rhythms. In *Saccharomyces cerevisiae*, populations of cells that were limited by nitrogen or glucose showed metabolic rhythms once the limitation was lifted. These four hour periods separate reductive and oxidative metabolism and give rise to radically different gene expression patterns (Xu & Tsurugi, 2006). Washing cells in nutrient free medium or applying the aforementioned dark period during culture collection might induce such changes. Finally, the light regime during preculture will be entrained on the cell, circadian rhythms and light availability will affect the different sampling times. With the decreasing cost of sequencing, more in depth time series will be possible to give insight into the time component of the response.

Complicating matters further, it is impossible to immediately remove a nutrient from the culture medium, since cells store some nutrients internally, such as phosphate in polyphosphate nodules or iron bound by ferritin in some pennate species (A. Marchetti et al., 2009). This can make it hard to pinpoint when a nutrient becomes limiting for cell growth, however, if after addition of a single nutrient the cell resumes division, it can be said that the cells were starved of that particular nutrient. The onset of nutrient starvation is therefore a gradual transition or, when there are no large internal stores, an artificially created

shock after transferring harvested cells to the desired medium. The transcriptome and proteome can show completely different processes dependant on which phase of nutrient deprivation sampling takes place as illustrated by figure 3. Chemical inhibitors are easier to add, but it has to be determined to what extent these results mimic the desired stress. Treatment of *P. tricornutum* with an inhibitor of nitrate uptake compared to actual N starvation showed significant differences in metabolite levels (Guerra et al., 2013). Despite these caveats, RNA-sequencing and proteomics are still an enormous step forward from the previous low throughput methods. In fact, it is only because the techniques are so sensitive that these changes can be observed.

METABOLIC CHANGES DURING NITROGEN STARVATION AS THE ARCHETYPE

The most common form of nitrogen in aquatic environments is nitrate. In green algae, the incorporation into biomolecules starts with the reduction of nitrate to ammonia by the NADH dependent nitrate reductase (NR) and ferredoxin dependent nitrite reductase (NiR). Ammonia can then be coupled to alpha-ketoglutarate to form glutamate by the enzyme glutamine oxoglutarate aminotransferase (GOGAT). In diatoms, the process follows the same steps, but in addition to the ferredoxin nitrite dependent reductase, they also have a NAD(P)H dependent nitrite reductase which can use oxidative phosphorylation as an energy source, allowing assimilation in the dark. Glutamate can be conjugated to a second ammonium by the ATP dependent glutamine synthase (GS), turning it into glutamine. Diatoms contain both a cytosol located glutamine synthase and a chloroplast localized version. It has been shown in *T. pseudonana* that the chloroplast localized enzyme is more actively transcribed when light is available and vice versa for the cytosol localized enzymes (Brown, Twing, & Robertson, 2009).

Increasing the capacity for importing and assimilating a limiting nutrient is often the first response of algae to starvation. Many of the highly induced genes are transporters or enzymes involved in the uptake or breakdown of nitrogenous compounds. Besides nitrate, ammonia and urea, many heterokont algae are able to use a variety of complex nitrogenous compounds (S. Dagenais-Bellefeuille & D. Morse, 2013; C. J. Gobler et al., 2011). The pelagophyte and facultative heterotrophic *Aureococcus anophagefferens* is exceptionally well equipped for heterotrophy and can even break down peptides (C. J. Gobler et al., 2011). The capacity to assimilate complex nitrogenous sources is likely to be species dependent, but many amidase, are upregulated during N starvation (Valenzuela et al., 2012).

Diatoms are often associated with bacteria, e.g. the production of toxins in *P. multiseriis* increases when bacteria are present (Foster & Zehr, 2006). In tropical waters, diatoms partake in symbiotic relationships with diazotrophic cyanobacteria when the diatom trades sugars for nutrients (Bates, Gaudet, Kaczmarzka, & Ehrman, 2004; Carpenter et al., 1999). Recently, it was shown that a diatom responds to auxin like compounds in the presence of bacteria, so exchanging carbon for nitrogen could be widespread (Shady A Amin, Parker, & Armbrust, 2012).

Diatoms excrete portions of carbon as extracellular polysaccharides and these provide attachment and/or energy for these bacteria(Shady A Amin et al., 2012). Besides the release of EPS, diatoms can also store carbohydrates under the form of chrysolaminaran, but conflicting reports about sugar accumulation exist. Some authors have observed an increase upon nitrogen deprivation, while others report a decrease(Espen Granum, KIRKVOLD, & Myklestad, 2002; P. G. Kroth et al., 2008; Valenzuela et al., 2012). Recent evidence shows that in *P. tricornutum* a knockout of the UDP glucose pyrophosphorylase gene, involved in a committing step in chrysolaminaran synthesis, increases lipid levels. This increase was seen even during nitrogen starvation, indicating that a significant proportion of the fixed carbon must go into carbohydrate polymers(Fayza Daboussi et al., 2014). Yet most studies show that transcripts for gluconeogenesis enzymes were unaltered in *P. tricornutum*.

Reusing existing nitrogen is a key process in N starved cells. Ribosomes contain large amounts of nitrogen both in their proteins and in the associated RNAs. Amino acids become scarce during N limitation, repressing protein synthesis. Recycling the unused protein synthesis capacity will allow the cell to keep the essential processes running(Klaas Vandepoele et al., 2013). The process transpires with the proteins of the photosynthetic apparatus and the size of chloroplasts is reduced under N starvation and chlorophyll production halts[58]. Like the ribosomal apparatus it is also protein intensive. Levels of photosystem II (PSII) decrease while photosystem I (PSI) levels are less affected, likely because maintaining PSI still allows cyclic electron flow and this imbalance is likely to prevent the formation of reactive oxygen species at PSII(Berges, Charlebois, Mauzerall, & Falkowski, 1996). Nevertheless, reactive oxygen species are still produced as catalases and glutaredoxins are still upregulated(Matthijs, et al. in preparation). This might be a signal for other metabolic pathways, since nitrogen assimilating enzymes and the urea cycle are under redox control(Rosenwasser et al., 2014).

The presence of a complete urea cycle provides a method to trap nitrogen within the cell and its discovery in diatoms was one of the biggest surprises of diatom genome projects(Andrew E Allen et al., 2011). This cycle has been shown to be functional and is suggested to increase nitrogen reuse within the cell. It does not appear to be upregulated during starvation, but transcripts respond when there is a large need for protein synthesis, such as after resupplementation with nitrogen or iron, or upon re-illumination after a period of darkness(A. Marchetti et al., 2012)(Matthijs et al. in preparation).

Both protein and transcript levels of the TCA cycle enzymes increase during N starvation in diatoms, which is similar to cyanobacteria, but does not happen in land plants(Nicola Louise Hockin, Thomas Mock, Francis Mulholland, Stanislav Kopriva, & Gill Malin, 2012; Valenzuela et al., 2012). This is also reflected on the metabolite level, since oxoglutarate levels were significantly increased(Guerra et al., 2013). The TCA cycle can both create the carbon skeletons needed for amino acid biosynthesis and reuse the carbon for energy generation after deamination, thus serving as a carbon redistribution hub. There is a potential link provided between the carbon entering TCA cycle and lipid biosynthesis through citrate lyase, but transcripts for this enzyme were not upregulated in any of the examined datasets. The

breakdown of branched chain amino acids has recently been linked to lipid biosynthesis, although the lipids still consist out of newly fixed carbon hinting that the generated acetyl CoA is burned up by the TCA cycle (Ge et al., 2014b). These responses are similar to those seen in non N₂ fixing cyanobacteria under nitrogen starvation. In recent studies, an increase in the transcripts and metabolites of the TCA cycle has been seen and glycogen, the prime carbon storage pool, was synthesized from the breakdown products of amino acids (Deschoenmaecker et al., 2014; Osanai et al., 2014).

The ability of microalgae to accumulate high levels of lipids during nitrogen starvation has been a crucial stimulus for research since the first oil crisis (Sheehan et al., 1998). Although diatoms are rich in omega-3 fatty acids, these are mainly present in membranes (Mühlroth et al., 2013), while the accumulating intracellular lipid droplets consist mainly of fully saturated triglycerides made from palmitic acid. This accumulation of lipids is by no means unique to diatoms, because green algae, such as *Chlorella vulgaris* and *Nannochloropsis gaditana*, also show this trait and have sequencing data available during N starvation (Carpinelli et al., 2014; Guarnieri et al., 2011). It appears that the storage of reduced carbon is a crucial adaptation in rapidly shifting aquatic environments, because the storage products can be used rapidly to assimilate available nutrients.

The metabolic adaptations that result in lipid accumulation, however, appear to differ between the diatoms and green algae. In diatoms most enzymes of lipid biosynthesis are not markedly increased during nitrogen starvation and control seems to occur on the post transcriptional level (Ge et al., 2014b).

In contrast, *N. gaditana* enzymes performing the committing steps were highly expressed during N starvation (Carpinelli et al., 2014). Similarly, in green algae coordinated changes were seen in lipid synthesis enzymes (Guarnieri et al., 2011). Likewise, the upregulation of ACC-carboxylase, performing the rate limiting step for fatty acid biosynthesis, was seen in coccolithophores (Carrier et al., 2014). In diatoms, high levels of ACC-carboxylase were only seen in the starting phases of nitrogen starvation, even though lipid accumulation continued after this peak in expression and ectopic expression of this enzyme had no effect in the diatom *Cylindrotheca fusiformis* (Valenzuela et al., 2012). Furthermore, the redistribution of nitrogen and carbon in green algae must take a different path as neither the TCA cycle is upregulated nor is there a urea cycle present.

IRON LIMITATION

Diatoms tend to be rare in the iron poor regions of the ocean (high nutrient, low chlorophyll) where they are outcompeted by green algae and cyanobacteria. However, open ocean iron fertilization results in a bloom dominated by diatoms and molecular insight in these artificial blooms has been gained using a metatranscriptomic experiment (A. Marchetti et al., 2012). Iron requirements vary widely between species. *T. pseudonana* for example needs 50-fold higher iron concentrations for its steady-state growth rate compared to *P. tricornutum* or *T. oceanica* (Kustka, Allen, & Morel, 2007; Adrian Marchetti, Maldonado,

Lane, & Harrison, 2006). Examining the genomes of these diatoms shows that only the latter contains ferritin which has significant implications during iron starvation periods(A. E. Allen et al., 2008a). Given this different genetic toolbox, it is no surprise that only a third of iron upregulated responsive genes was shared between *T. pseudonana* and *P. tricornutum*(K. Thamtrakoln, O. Korenovska, A. K. Niheu, & K. D. Bidle, 2012).

A decline in photosynthesis is the most apparent response of severe iron limitation. The immediately visible symptoms of iron stress are a reduced chlorophyll content and an increased rate of photorespiration(Geider, La Roche, Greene, & Olaizola, 1993). This response is shared with nitrogen limitation, but the underlying cause is different. Nitrogen starvation hampers the production of proteins for the photosynthetic apparatus, while Fe deficiency impedes the electron transport chains of both oxidative phosphorylation and photosynthesis. This imbalance in redox proteins results in oxidative stress and an upregulation of photoprotective carotenoid biosynthesis and non-photochemical quenching, since the energy from captured photons needs to be dissipated(A. E. Allen et al., 2008b). In line with the decrease of photosynthesis, cells during Fe starvation do not accumulate carbohydrates or lipids(Allen J. Milligan & Paul J. Harrison, 2000), but try to compensate this energy shortage by oxidizing stored carbon. Carbohydrate reserves are mobilized as exemplified by the upregulation of several transcripts involved in glycolysis(A. E. Allen et al., 2008b). The TCA cycle is also involved, because some metabolites are upregulated, such as malate and alpha-ketoglutarate, while levels of succinyl-CoA decreased and those of citrate were unaltered.

When iron-sulfur containing electron donors, such as ferredoxin and cytochrome c6 cannot be synthesized, some algae can exchange them for less efficient alternative carriers, for example by the iron-free flavodoxin and the copper-containing plastocyanin. Most green algae contain plastocyanin as do diatoms adapted to the open ocean such as *T. oceanica* which has even lost its cytochrome c6 gene(Gupta, He, & Luan, 2002; G. Peers & N. M. Price, 2006). A further adaptation is swapping the standard superoxide dismutase for a copper dependent version and the use of bacterial type rhodopsin protein pumps(A. Marchetti et al., 2012). Horizontal gene transfer likely lies at the basis of these adaptations. In general, open ocean diatoms require less iron than green algae, because they use alternative electron carriers, since they have lost their more efficient iron containing genes(Graham Peers & Neil M. Price, 2006).

When cells are limited by iron availability, N assimilation is also affected, since the conversion of nitrate to ammonia requires a significant input of reductive equivalents. This becomes a problem during iron starvation, because the supply of electrons decreases together with photosynthesis, but also because diatoms use a bacterial type nitrate reductase that employs ferredoxin as the electron donor. However no difference in growth rates was seen in iron limited cultures grown either with ammonia or nitrate(Maldonado & Price, 1996).

Conflicting reports exist on the C:N ratio of iron deficient diatoms, some authors see a decline in assimilated nitrogen per unit of carbon, while others do not(Allen J Milligan & Paul J Harrison, 2000). If there is a decrease in N assimilation, it is more likely the result of a lack of energy or carbon skeletons rather than of a missing cofactor, since nitric reductase activity does not decrease enough to explain this drop in ratio(Allen J. Milligan & Paul J. Harrison, 2000). As with nitrogen fixing bacteria, the interaction between iron capturing bacteria and coccolithophores has been reported(Shady A. Amin et al., 2009). It is likely that similar exchanges of carbon for iron occur with diatoms.

PHOSPHOROUS STARVATION

Phosphorous is often a limiting nutrient in fresh water and it can also limit productivity in coastal waters when there is a large influx of fresh water(Nausch, Nausch, & Wasmund, 2004). Like green algae, diatoms are able to store phosphate when it is abundant, making it difficult to determine when the nutrient becomes limited in the medium(Omelon, Ariganello, Bonucci, Grynpas, & Nanci, 2013; Ruiz, Marchesini, Seufferheld, & Docampo, 2001). Storage of polyphosphate is mediated by vacuolar transporter chaperone 4 (VTC4) like genes which are upregulated during phosphorous starvation in both *T. pseudonana* (Thaps43150) and *P. tricornutum* (Phatr50019)((Ruiz et al., 2001), Matthijs et al., in preparation). The stored phosphate is a likely explanation for the less severe phenotype seen during the first days of phosphate starvation. During this period, cells continue to grow at nearly the same rate as when replete. Unlike the other stresses mentioned above, there is no immediate reduction in the photosynthetic apparatus reported when phosphorous becomes limiting. It presumably takes a long time for the photosynthetic apparatus to be affected, but oxidative processes, such as the TCA cycle, appear to be upregulated even during the early stages of phosphorous limitation(Yang et al., 2014). The excess energy fixed by photosynthesis under P-limitation is stored in the form of lipids and carbohydrates, leading to an increase in C:N ratio.

Changes are seen in glycolysis and translation machinery, but these are relatively mild compared to iron or nitrogen starvation. The cells respond mainly by increasing their uptake capacity as the expression of phosphate transporters rises (Phatr40433 and Thaps24435). In addition, diatoms are able to release phosphate bound to organic sources using a diverse set of alkaline phosphatases, of which many are upregulated during phosphate deficiency(Reynolds, 2006). Turnover of intracellular phosphorous is likely to increase due to the increased activity of intracellular phosphatases. Phospholipids for example are a large reservoir for phosphorous and these can be partly replaced by sulphur or betaine containing lipids. Overall, the response of *P. tricornutum* appears very similar to that of *T. pseudonana*. Interestingly, phosphate starvation was the subject of one of the few studies in which transcriptomics and proteomics was combined, and transcript and protein levels were relatively well coordinated for upregulated genes (60%)(Dyhrman et al., 2012).

SILICON AND SULPHUR STARVATION

Sulphur is an essential element for life, since it is present in two amino acids (methionine and cysteine), incorporated in sulpholipids and part of several cofactors such as Coenzyme A. Sulphur assimilation starts with the action of ATP-sulphurylase, which converts ATP and sulphate in 5'-adenylsulphate (APS). This compound is stepwise reduced to sulphide, after which it is incorporated into cysteine. The reduction of APS to sulphite is the key regulation point of the pathway in land plants, but recent findings indicate that regulation of sulphur assimilation in diatoms differs radically (Kettles, Kopriva, & Malin, 2014; H. Takahashi, S. Kopriva, M. Giordano, K. Saito, & R. Hell, 2011). In fresh water, sulphate can be limiting, but with a concentration of 29 mM in dissolved seawater, it is virtually in endless supply for marine organisms. Interestingly, diatoms contain, on average, much more sulphur per unit of carbon than green algae (Quigg et al., 2003). Because fresh water can be limiting in sulphate, green algal sulfur deficiency has been well studied, but studies for diatoms are scarce (Hideki Takahashi, Stanislav Kopriva, Mario Giordano, Kazuki Saito, & Rüdiger Hell, 2011). The observation that *C. reinhardtii* produces hydrogen during sulphate starvation is promising for biofuel production and has therefore been extensively studied (González-Ballester et al., 2010). Heterokont algae, to which diatoms belong, play an important role in sulphur cycling, since they produce substantial amounts of dimethylsulphoniopropionate (DMSP), which is converted into the volatile compound dimethylsulphide (DMS), implicated in rain cloud formation and global warming (Kettle & Andreae, 2000).

To our knowledge no 'omics' data is available at the time of writing for diatoms under sulphur limitation. However, recently a study was published in which the coccolithophore *Emiliania huxleyi* was grown in a medium with a lowered sulphate content. Approximately 5% of the transcripts were altered in expression (Bochenek et al., 2013). The genes for fatty acid and carbohydrate biosynthesis were upregulated and there was a general decline in S containing metabolites, such as DMSP and cysteine. Several TCA cycle genes were also upregulated and photosynthesis was not severely impacted, likely because there is a large internal pool of sulphur in the form of DMSP. In general the sulphur response appeared similar in green plants, but it would be interesting to see a study looking at a fresh water diatom.

Almost all diatoms are encased in a silica cell wall. This means that, compared to other types of algae, they require an extra nutrient. This makes their success even more remarkable because silica as a nutrient is not at all that common in marine environments with an average concentration of 70 μM (V Martin-Jézéquel, Daoud, & Quéguiner, 1997). While land plants also require silica for healthy growth, green algae have no use for this element (Cooke & Leishman, 2011). Why diatoms thrive despite this extra requirement is not fully understood, but it likely has to do with the scarcity of nitrogen and iron. Energetically it seems more efficient to build a silica cell wall than a carbohydrate one, which could mean that diatoms can devote more energy to the uptake of nutrients (Véronique Martin-Jézéquel, Hildebrand, & Brzezinski, 2000). Compared to the other nutrients, silicon is unique because the energy used for its incorporation, and its deposition is derived solely from aerobic respiration and not linked to the

photosynthetic process such as the ferredoxin (or equivalent) required for nitrate reduction. Because cellular energy metabolism is not impaired, proteins of the transcription and translation machinery remain highly expressed to allow the cell to resume division as soon as conditions improve (Roshan P Shrestha et al., 2012). There is also putative crosstalk with phosphate starvation, since there is an increased production of alkaline phosphatases during silicon starvation (Fuentes, Wikfors, & Meseck, 2013).

When a nutrient becomes limiting, it can pay to remove the competition by the production of toxins. During silicon and phosphorous starvation, the synthesis of domoic acid, a toxin of *Pseudo-nitzschia multiseries*, increases markedly (Pan, Bates, & Cembella, 1998). While allelopathy in dinoflagellates has been well studied, there seems to be little known in diatoms in this context (Ianora et al., 2011). Large bloom forming organisms, such as *Aureococcus*, probably release compounds that inhibit the growth of other algae.

CELL CYCLE AND CONTROL MECHANISMS

The large transcriptomic changes seen during nutrient limitation are steered by transcription factors (TFs), of which many are upregulated but functional validation is lacking at the moment. A recent report by Rayko and co-authors lists all possible factors found on the basis of homology (E. Rayko, F. Maumus, U. Maheswari, K. Jabbari, & C. Bowler, 2010), but currently the only TFs characterized are involved in light and CO₂ sensing (M. J. J. Huysman et al., 2013; Ohno et al., 2012). Of course many transcription factors are upregulated during nutrient stress but functional validation is lacking at the moment.

When cells are deprived of nutrients, their division stops, but the halting point of cell division differs depending on the condition. During nitrogen or phosphorous starvation, diatoms arrest their cell cycle at the G1/S phase, whereas in dinoflagellates G2/M blocks are also reported (Lin et al., 2004). Silicon starvation is atypical and the cell cycle can halt in both G2/M or G1/S (Brzezinski, Olson, & Chisholm, 1990). To the best of our knowledge, no studies have been performed on iron limited diatoms. The regulation of the cell cycle in eukaryotes is to a large extent dependent on cyclins and their interactions with cyclin dependent kinases (CDKs). The sequenced diatom genomes show that the gene family of the former is greatly expanded and many of these cyclins are diatom specific. Interestingly, some of these cyclins responded to the addition of nutrients such as silica or phosphate, hinting that these might convey the nutrient status to the cell cycle machinery (M. J. Huysman et al., 2010). Further corroborating this is the observation that *T. pseudonana* does not take up nitrogen during mitosis, showing that there is crosstalk between nutrient assimilation and cell division (Mocquet, Sciandra, Talec, & Bernard, 2013). So far no direct links have been reported, but in *P. tricornutum* *dsCYC7* was shown to be upregulated upon addition of phosphate, whereas Valenzuela et al. reported that this cyclin was downregulated during nitrogen starvation [37]. However, in house generated data showed that *dsCYC7* expression levels steadily increase during the first 20 hours of nitrogen starvation (Matthijs et al., in preparation). The exact role and importance of diatom specific cyclins remains to be determined.

Diatoms can adapt to the lack of nutrients and shift their expression profiles to compensate for it. This process requires a signalling pathway from metabolism and nutrient availability to the transcription and translation apparatuses. In most eukaryotes signalling energy and nutrient shortage happens through two kinase complexes, i.e. AMP-activated Protein Kinase (AMPK, or SNF1 in yeast), and Target of Rapamycin (TOR). AMPK signals a lack of energy by sensing the AMP/ATP ratio and slowing anabolism in response, whereas TOR senses when insufficient amino acids are present (Ghillebert et al., 2011; Kim & Guan, 2011). Both are strongly evolutionary conserved and are apparently present in all eukaryotic microalgae.

All three subunits of AMPK are present in the diatom genome but the gene families did not go through the dramatic gene expansion seen in green algae and land plants (Baena-Gonzalez, Rolland, Thevelein, & Sheen, 2007). The assimilation of nutrients such as sulfur and nitrogen, demands extensive energy inputs from the cell, necessitating crosstalk between energy availability and nutrient status. In green algae and land plants, there is a strong coupling of nitrogen uptake with light, likely due to the availability of photosynthetic energy to assimilate nitrogen. In contrast to green algae, diatoms are able to take up nitrogen in the absence of light, indicating that there is an uncoupling between energy status and nitrogen uptake. This rapid uptake of resources is possibly one of the reasons for the ecological success of diatoms. When cells can no longer divide, but photosynthesis is still active, there will be excess energy that can be funnelled into carbon storage. In most eukaryotic organisms this is signalled by AMPK [89]. When the ratio between AMP and ATP drops when the cell has abundant energy, the complex becomes less active. In humans, this lifts the inhibitory phosphorylation of ACC, the committing step in fatty acid synthesis, and glycogen synthase. All sequenced diatoms contain the necessary components of the AMPK complex and the gene family has undergone expansion in the green lineage. In the green algae *Chlorella vulgaris*, AMPK is downregulated under nitrogen starvation concomitant with lipid increase. This pattern was not seen in any heterokonts in the published studies. However, in a chemical screen in several green algae and *P. tricornutum*, it was found that the AMPK inhibitor AICAR has a positive impact on *P. tricornutum* oil productivity (Franz, Danielewicz, Wong, Anderson, & Boothe, 2013). Another compound from the screen, EGCG, had a similar effect and is closely related to a class of AMPK inhibitors. It is striking that in this study of Franz et al. there is no overlap between the compounds increasing lipid productivity in green algae and those in diatoms, underscoring the differences in regulatory pathways, although the signalling pathways seem related. Similarly, a SNRK like gene encoding a kinase closely related to AMPK is involved in sulphur adaptation in green plants but does not show regulation in *E. huxleyi* (Bochenek et al., 2013). In yeast it has been shown that there is a direct link with nitrogen sensing but the nitrogen signalling pathways remain, for the moment, unknown in heterokonts (Orlova, Ozcetin, Barrett, & Kuchin, 2010).

In heterokonts and dinoflagellates, the TOR complex (TORC) lacks Tsc1, Tsc2 and Rictor, which form the basis of distinct signalling complexes in mammals, but otherwise seem to contain the same components as most other eukaryotes (Serfontein, Nisbet, Howe, & de Vries, 2010). TORC determines

rates of translation by phosphorylating S6K. As with AMPK, none of the components of the TOR pathway vary greatly in expression during nitrogen starvation. Nevertheless, since these kinases have been conserved, it is unlikely that energy and nutrient status would not pass through this pathway. Besides these conserved complexes, there are over a hundred serine/threonine kinases in each sequenced diatom, of which, for most, we do not know any function.

The conserved signalling pathways outlined above are only the distributors of nutrient starvation signals. The sensors for these processes are still unknown and will likely involve signalling cascades such as calcium fluxes or the notch pathway, of which several components are upregulated during nitrogen starvation (Matthijs et al., in preparation). In *Chlamydomonas*, the PII system that has a bacterial origin and is also present in land plants, was shown to be one of the main signalling systems for nitrogen starvation (Fokina, Chellamuthu, Forchhammer, & Zeth, 2010). PII senses the abundance of α -ketoglutarate, which also accumulates in diatoms during nitrogen starvation, but the protein appears to be absent in heterokonts (Ermilova et al., 2013). Besides intracellular communication, algae also communicate with their neighbours. Nitric oxide has also been involved in attracting diazotroph cyanobacteria and may signal other cells (Shady A Amin et al., 2012). Cell-cell communication must exist in bloom forming organisms, since the collapses of these blooms can be triggered by nutrient starvation and are characterised by massive cell death. Metacaspases are present in the heterokonts and are presumably involved (Nedelcu, 2009).

OUTLOOK

The pace with which marine organisms are being sequenced will only increase in the coming years and this wave of data will give insight in the conservation and uniqueness of diatom metabolic pathways. While the enzymes of the primary metabolism are easy to identify, comparatively little work has been done on the regulation of these pathways. Transcriptomic and proteomic data add a dynamic layer to the genomic level, and visualize how enzyme levels respond to conditions such as nutrient deprivation. Although there have been only a handful of transcriptomic or proteomic experiments performed for most nutrient stresses, the general metabolic changes are becoming clear. A drawback is the substantial variation in experimental setup and sampling conditions, which likely intertwine with the effect of the treatment. This problem is likely to solve itself as the number of RNA-sequencing datasets will likely increase dramatically in the foreseeable future and the signal of nutrient starvation will rise above the experimental noise.

While green algae and diatoms often inhabit similar environments, their responses to the same stresses differ. It appears that diatoms use different solutions for some of the same problems. Perhaps this knowledge can even be transferred to land crops, for which increased nitrogen efficiency would decrease the damaging eutrophication of waterways. A similar approach has been tried successfully with cyanobacterial proteins and the recent use of a diatom nucleotide transporter further illustrates this (Malyshev et al., 2014).

What is completely lacking however is insight in how these changes in growth conditions are perceived and how this signal is translated into metabolic flux changes. In other organisms, it has been shown that these control mechanisms are less evolutionary conserved than those of central regulators such as TOR. It is a trivial matter to identify the kinases and TFs in a transcriptome or genome, but finding their targets will require dedicated experiments. High throughput experiments, such as yeast one- or two-hybrid, are feasible especially when the predicted number of TFs for *P. tricornutum* is only 200. Understanding the regulation will create a useful tool to create better strains for oil production. If we understand the switches that trigger the metabolic reprogramming from growth to storage, we can engineer strains that turn this program at will.

Additionally, it is to be expected that there is substantial intra-species variation in diatoms, since many species inhabit specific niches with unique restrictions and challenges. It will be particularly informative to compare more open ocean species adapted to low iron conditions with more coastal species that are struggling with nitrate or phosphate limitation. At the moment is not yet clear if there is a transcriptome or proteome set of genes that is shared across all types of nutrient limitation. While most stresses halt growth and decrease photosynthesis, there are connections between the different nutrients, as exemplified by the upregulation of silicon transporters and phosphatases during nitrogen starvation.

Finally, diatoms do not live alone in their habitats, they interact with prokaryotes and other photosynthetic eukaryotes and compete for limited nutrients. Meta-projects will show which species groups co-exist and begin to outline species-species interactions. Many land plants exchange photosynthetically fixed carbon for nutrients such as nitrogen or phosphate (Smith, Smith, & Jakobsen, 2003; Vitousek, 1984). It will be interesting to see what interactions exist between diatoms and other unicellular organisms. Other metabolic adaptations to nutrient deficiency, such as the enzymes responsible for the production of allelopathic secondary metabolites, remain to be investigated.

Most of the above described is based on what we have learned from other model species with well-characterized genes, but there are limits to homology based approaches. In short, the 'omics' age has gifted us tools that are exceptionally well suited to catalogue which genes are present in which species and habitat. While it is possible to assume functions based on the expression patterns of genes, this does not suffice as proof. But help is on the way, since new tools are becoming available for the first time in diatoms: targeted nucleases. The ease with which homologous recombination can generate gene knockouts in *Saccharomyces cerevisiae* and *Escherichia coli* have undoubtedly contributed much to their usefulness as model systems. With the possibility to tailor TALEN or CRISPR nucleases, these tools are becoming available to each transformable organism. This will allow the validation of metabolic models and predictions when certain nutrients such as amino acids become scarce (Long & Antoniewicz, 2014). In short, we now have both the tools to identify the genes shifting during nutrient limitation and the means to investigate their function.

REFERENCES

1. Baurain, D., et al., *Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles*. Mol Biol Evol, 2010. **27**(7): p. 1698-709.
2. Armbrust, E.V., *The life of diatoms in the world's oceans*. Nature, 2009. **459**(7244): p. 185-92.
3. Dortch, Q. and T.E. Whitledge, *Does nitrogen or silicon limit phytoplankton production in the Mississippi River plume and nearby regions?* Continental Shelf Research, 1992. **12**(11): p. 1293-1309.
4. Furnas, M.J., *In situ growth rates of marine phytoplankton: approaches to measurement, community and species growth rates*. Journal of Plankton Research, 1990. **12**(6): p. 1117-1151.
5. Ragueneau, O., et al., *A review of the Si cycle in the modern ocean: recent progress and missing gaps in the application of biogenic opal as a paleoproductivity proxy*. Global and Planetary Change, 2000. **26**(4): p. 317-365.
6. De Baar, H.J., et al., *Synthesis of iron fertilization experiments: from the iron age in the age of enlightenment*. Journal of Geophysical Research: Oceans (1978–2012), 2005. **110**(C9).
7. Heisler, J., et al., *Eutrophication and harmful algal blooms: a scientific consensus*. Harmful algae, 2008. **8**(1): p. 3-13.
8. Riegman, R., A.A. Noordeloos, and G.C. Cadée, *Phaeocystis blooms and eutrophication of the continental coastal zones of the North Sea*. Marine Biology, 1992. **112**(3): p. 479-484.
9. Dagenais-Bellefeuille, S. and D. Morse, *Putting the N in dinoflagellates*. Frontiers in microbiology, 2013. **4**.
10. Sheehan, J., et al., *A look back at the US Department of Energy's aquatic species program: biodiesel from algae*. Vol. 328. 1998: National Renewable Energy Laboratory Golden.
11. Benson, A.A., *Paving the path*. Annu Rev Plant Biol, 2002. **53**: p. 1-25.
12. Hopkinson, B.M., et al., *Efficiency of the CO₂-concentrating mechanism of diatoms*. Proceedings of the National Academy of Sciences, 2011. **108**(10): p. 3830-3837.
13. Ast, M., et al., *Diatom plastids depend on nucleotide import from the cytosol*. Proceedings of the National Academy of Sciences, 2009. **106**(9): p. 3621-3626.
14. Kroth, P.G., et al., *A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis*. PloS one, 2008. **3**(1): p. e1426.
15. Lewin, J.C. and R.A. Lewin, *AUXOTROPHY AND HETEROTROPHY IN MARINE LITTORAL DIATOMS*. Canadian Journal of Microbiology, 1960. **6**(2): p. 127-134.
16. Armbrust, E.V., et al., *The genome of the diatom Thalassiosira pseudonana: Ecology, evolution, and metabolism*. Science, 2004. **306**(5693): p. 79-86.
17. Bowler, C., et al., *The Phaeodactylum genome reveals the evolutionary history of diatom genomes*. Nature, 2008. **456**(7219): p. 239-244.

18. Lommer, M., et al., *Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation*. *Genome biology*, 2012. **13**(7): p. R66.
19. Gobler, C.J., et al., *Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics*. *Proceedings of the National Academy of Sciences*, 2011. **108**(11): p. 4352-4357.
20. Radakovits, R., et al., *Draft genome sequence and genetic transformation of the oleaginous alga *Nannochloropsis gaditana**. *Nature communications*, 2012. **3**: p. 686.
21. Berges, J.A. and P.G. Falkowski, *Physiological stress and cell death in marine phytoplankton: induction of proteases in response to nitrogen or light limitation*. *Limnology and Oceanography*, 1998. **43**(1): p. 129-135.
22. Smetacek, V., *Role of sinking in diatom life-history cycles: ecological, evolutionary and geological significance*. *Marine biology*, 1985. **84**(3): p. 239-251.
23. Kamp, A., et al., *Diatoms respire nitrate to survive dark and anoxic conditions*. *Proceedings of the National Academy of Sciences*, 2011. **108**(14): p. 5649-5654.
24. Mühlroth, A., et al., *Pathways of lipid metabolism in marine algae, co-expression network, bottlenecks and candidate genes for enhanced production of EPA and DHA in species of *Chromista**. *Marine drugs*, 2013. **11**(11): p. 4662-4697.
25. Bertrand, E.M., et al., *Influence of cobalamin scarcity on diatom molecular physiology and identification of a cobalamin acquisition protein*. *Proc Natl Acad Sci U S A*, 2012. **109**(26): p. E1762-71.
26. Lane, T.W. and F.M. Morel, *A biological function for cadmium in marine diatoms*. *Proc Natl Acad Sci U S A*, 2000. **97**(9): p. 4627-31.
27. Koboldt, D.C., et al., *The next-generation sequencing revolution and its impact on genomics*. *Cell*, 2013. **155**(1): p. 27-38.
28. Guarnieri, M.T., et al., *Examination of triacylglycerol biosynthetic pathways via de novo transcriptomic and proteomic analyses in an unsequenced microalga*. *PLoS One*, 2011. **6**(10): p. e25851.
29. Karsenti, E., et al., *A holistic approach to marine eco-systems biology*. *PLoS Biol*, 2011. **9**(10): p. e1001177.
30. Rismani-Yazdi, H., et al., *Transcriptomic analysis of the oleaginous microalga *Neochloris oleoabundans* reveals metabolic insights into triacylglyceride accumulation*. *Biotechnol Biofuels*, 2012. **5**(1): p. 74.
31. Morris, R.M., et al., *Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction*. *Isme j*, 2010. **4**(5): p. 673-85.
32. Bowler, C., et al., *The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes*. *Nature*, 2008. **456**(7219): p. 239-44.
33. Curtis, B.A., et al., *Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs*. *Nature*, 2012. **492**(7427): p. 59-65.
34. Vandepoele, K., et al., *pico-PLAZA, a genome database of microbial photosynthetic eukaryotes*. *Environmental microbiology*, 2013. **15**(8): p. 2147-2153.

35. Vandepoele, K., et al., *Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks*. Plant physiology, 2009. **150**(2): p. 535-546.
36. Thamatrakoln, K., et al., *Whole-genome expression analysis reveals a role for death-related genes in stress acclimation of the diatom Thalassiosira pseudonana*. Environmental microbiology, 2012. **14**(1): p. 67-81.
37. Valenzuela, J., et al., *Potential role of multiple carbon fixation pathways during lipid accumulation in Phaeodactylum tricornutum*. Biotechnol Biofuels, 2012. **5**(1): p. 40.
38. Allen, A.E., et al., *Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation*. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(30): p. 10438-10443.
39. Sapriel, G., et al., *Genome-wide transcriptome analyses of silicon metabolism in Phaeodactylum tricornutum reveal the multilevel regulation of silicic acid transporters*. PLoS One, 2009. **4**(10): p. e7458.
40. Yang, Z.K., et al., *Molecular and cellular mechanisms of neutral lipid accumulation in diatom following nitrogen deprivation*. Biotechnol Biofuels, 2013. **6**(1): p. 67.
41. Shrestha, R.P., et al., *Whole transcriptome analysis of the silicon response of the diatom Thalassiosira pseudonana*. BMC Genomics, 2012. **13**: p. 499.
42. Koester, J.A., W.J. Swanson, and E.V. Armbrust, *Positive selection within a diatom species acts on putative protein interactions and transcriptional regulation*. Molecular biology and evolution, 2012: p. mss242.
43. Massonnet, C., et al., *Probing the reproducibility of leaf growth and molecular phenotypes: a comparison of three Arabidopsis accessions cultivated in ten laboratories*. Plant Physiol, 2010. **152**(4): p. 2142-57.
44. Huysman, M.J., et al., *Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling*. Genome Biol, 2010. **11**(2): p. R17.
45. Xu, Z. and K. Tsurugi, *A potential mechanism of energy-metabolism oscillation in an aerobic chemostat culture of the yeast Saccharomyces cerevisiae*. FEBS Journal, 2006. **273**(8): p. 1696-1709.
46. Marchetti, A., et al., *Ferritin is used for iron storage in bloom-forming marine pennate diatoms*. Nature, 2009. **457**(7228): p. 467-70.
47. Guerra, L.T., et al., *Regulatory branch points affecting protein and lipid biosynthesis in the diatom Phaeodactylum tricornutum*. Biomass and Bioenergy, 2013. **59**: p. 306-315.
48. Brown, K.L., K.I. Twing, and D.L. Robertson, *UNRAVELING THE REGULATION OF NITROGEN ASSIMILATION IN THE MARINE DIATOM THALASSIOSIRA PSEUDONANA (BACILLARIOPHYCEAE): DIURNAL VARIATIONS IN TRANSCRIPT LEVELS FOR FIVE GENES INVOLVED IN NITROGEN ASSIMILATION*. Journal of Phycology, 2009. **45**(2): p. 413-426.
49. Dagenais-Bellefeuille, S. and D. Morse, *Putting the N in dinoflagellates*. Front Microbiol, 2013. **4**: p. 369.
50. Gobler, C.J., et al., *Niche of harmful alga Aureococcus anophagefferens revealed through ecogenomics*. Proc Natl Acad Sci U S A, 2011. **108**(11): p. 4352-7.

51. Foster, R.A. and J.P. Zehr, *Characterization of diatom–cyanobacteria symbioses on the basis of nifH, hetR and 16S rRNA sequences*. Environmental microbiology, 2006. **8**(11): p. 1913-1925.
52. Bates, S.S., et al., *Interaction between bacteria and the domoic-acid-producing diatom *Pseudo-nitzschia multiseries* (Hasle) Hasle; can bacteria produce domoic acid autonomously?* Harmful Algae, 2004. **3**(1): p. 11-20.
53. Carpenter, E.J., et al., *Extensive bloom of a N₂-fixing diatom/cyanobacterial association in the tropical Atlantic Ocean*. 1999.
54. Amin, S.A., M.S. Parker, and E.V. Armbrust, *Interactions between diatoms and bacteria*. Microbiology and Molecular Biology Reviews, 2012. **76**(3): p. 667-684.
55. Granum, E., S. KIRKVOLD, and S.M. Myklestad, *Cellular and extracellular production of carbohydrates and amino acids by the marine diatom *Skeletonema costatum*: diel variations and effects of N depletion*. Marine ecology. Progress series, 2002. **242**: p. 83-94.
56. Kroth, P.G., et al., *A model for carbohydrate metabolism in the diatom *Phaeodactylum tricornutum* deduced from comparative whole genome analysis*. PLoS One, 2008. **3**(1): p. e1426.
57. Daboussi, F., et al., *Genome engineering empowers the diatom *Phaeodactylum tricornutum* for biotechnology*. Nature communications, 2014. **5**.
58. Yang, Z.-K., et al., *Molecular and cellular mechanisms of neutral lipid accumulation in diatom following nitrogen deprivation*. Biotechnol. Biofuels, 2013. **6**(67): p. 1-67.
59. Berges, J.A., et al., *Differential effects of nitrogen limitation on photosynthetic efficiency of photosystems I and II in microalgae*. Plant Physiology, 1996. **110**(2): p. 689-696.
60. Rosenwasser, S., et al., *Mapping the diatom redox-sensitive proteome provides insight into response to nitrogen stress in the marine environment*. Proc Natl Acad Sci U S A, 2014. **111**(7): p. 2740-5.
61. Allen, A.E., et al., *Evolution and metabolic significance of the urea cycle in photosynthetic diatoms*. Nature, 2011. **473**(7346): p. 203-207.
62. Marchetti, A., et al., *Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability*. Proc Natl Acad Sci U S A, 2012. **109**(6): p. E317-25.
63. Hockin, N.L., et al., *The response of diatom central carbon metabolism to nitrogen starvation is different from that of green algae and higher plants*. Plant physiology, 2012. **158**(1): p. 299-312.
64. Ge, F., et al., *Methylcrotonyl-CoA Carboxylase Regulates Triacylglycerol Accumulation in the Model Diatom *Phaeodactylum tricornutum**. Plant Cell, 2014.
65. Deschoenmaeker, F., et al., *Proteomic and cellular views of *Arthrospira* sp. PCC 8005 adaptation to nitrate depletion*. Microbiology, 2014: p. mic. 0.074641-0.
66. Osanai, T., et al., *Capillary electrophoresis–mass spectrometry reveals the distribution of carbon metabolites during nitrogen starvation in *Synechocystis* sp. PCC 6803*. Environmental Microbiology, 2014. **16**(2): p. 512-524.

67. Carpinelli, E.C., et al., *Chromosome scale genome assembly and transcriptome profiling of Nannochloropsis gaditana in nitrogen depletion*. Molecular plant, 2014. **7**(2): p. 323-335.
68. Carrier, G., et al., *Comparative transcriptome of wild type and selected strains of the microalgae Tisochrysis lutea provides insights into the genetic basis, lipid metabolism and the life cycle*. PloS one, 2014. **9**(1): p. e86889.
69. Kustka, A.B., A.E. Allen, and F.M.M. Morel, *SEQUENCE ANALYSIS AND TRANSCRIPTIONAL REGULATION OF IRON ACQUISITION GENES IN TWO MARINE DIATOMS*¹. Journal of Phycology, 2007. **43**(4): p. 715-729.
70. Marchetti, A., et al., *Iron requirements of the pennate diatom Pseudonitzschia: Comparison of oceanic (high-nitrate, low-chlorophyll waters) and coastal species*. Limnology and oceanography, 2006. **51**(5): p. 2092-2101.
71. Allen, A.E., et al., *Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation*. Proc Natl Acad Sci U S A, 2008. **105**(30): p. 10438-43.
72. Thamatrakoln, K., et al., *Whole-genome expression analysis reveals a role for death-related genes in stress acclimation of the diatom Thalassiosira pseudonana*. Environ Microbiol, 2012. **14**(1): p. 67-81.
73. Geider, R.J., et al., *RESPONSE OF THE PHOTOSYNTHETIC APPARATUS OF PHAEODACTYLUM TRICORNUTUM (BACILLARIOPHYCEAE) TO NITRATE, PHOSPHATE, OR IRON STARVATION*¹. Journal of Phycology, 1993. **29**(6): p. 755-766.
74. Milligan, A.J. and P.J. Harrison, *Effects of non-steady-state iron limitation on nitrogen assimilatory enzymes in the marine diatom thalassiosira weissflogii (BACILLARIOPHYCEAE)*. Journal of Phycology, 2000. **36**(1): p. 78-86.
75. Peers, G. and N.M. Price, *Copper-containing plastocyanin used for electron transport by an oceanic diatom*. Nature, 2006. **441**(7091): p. 341-344.
76. !!! INVALID CITATION !!!
77. Maldonado, M. and N. Price, *Influence of N substrate on Fe requirements of marine centric diatoms*. Marine ecology progress series. Oldendorf, 1996. **141**(1): p. 161-172.
78. Milligan, A.J. and P.J. Harrison, *Effects of non-steady-state iron limitation on nitrogen assimilatory enzymes in the marine diatom thalassiosira weissflogii (BACILLARIOPHYCEAE)*. Journal of Phycology, 2000. **36**(1): p. 78-86.
79. Amin, S.A., et al., *Photolysis of iron–siderophore chelates promotes bacterial–algal mutualism*. Proceedings of the National Academy of Sciences, 2009. **106**(40): p. 17071-17076.
80. Nausch, M., G. Nausch, and N. Wasmund, *Phosphorus dynamics during the transition from nitrogen to phosphate limitation in the central Baltic Sea*. Marine Ecology Progress Series, 2004. **266**: p. 15-25.
81. Yang, Z.K., et al., *Systems-level analysis of the metabolic responses of the diatom Phaeodactylum tricornutum to phosphorus stress*. Environmental microbiology, 2014. **16**(6): p. 1793-1807.

82. Omelon, S., et al., *A review of phosphate mineral nucleation in biology and geobiology*. *Calcif Tissue Int*, 2013. **93**(4): p. 382-96.
83. Ruiz, F.A., et al., *The polyphosphate bodies of Chlamydomonas reinhardtii possess a proton-pumping pyrophosphatase and are similar to acidocalcisomes*. *Journal of Biological Chemistry*, 2001. **276**(49): p. 46196-46203.
84. Reynolds, C.S., *The Ecology of Phytoplankton*. 2006: Cambridge University Press.
85. Dyhrman, S.T., et al., *The transcriptome and proteome of the diatom Thalassiosira pseudonana reveal a diverse phosphorus stress response*. *PLoS One*, 2012. **7**(3): p. e33768.
86. Quigg, A., et al., *The evolutionary inheritance of elemental stoichiometry in marine phytoplankton*. *Nature*, 2003. **425**(6955): p. 291-4.
87. Takahashi, H., et al., *Sulfur assimilation in photosynthetic organisms: molecular functions and regulations of transporters and assimilatory enzymes*. *Annual review of plant biology*, 2011. **62**: p. 157-184.
88. González-Ballester, D., et al., *RNA-seq analysis of sulfur-deprived Chlamydomonas cells reveals aspects of acclimation critical for cell survival*. *The Plant Cell Online*, 2010. **22**(6): p. 2058-2084.
89. Bochenek, M., et al., *Transcriptome analysis of the sulfate deficiency response in the marine microalga Emiliania huxleyi*. *New Phytol*, 2013. **199**(3): p. 650-62.
90. Martin-Jézéquel, V., N. Daoud, and B. Quéguiner, *Coupling of silicon, carbon and nitrogen metabolisms in marine diatoms*. *Integrated Marine System Analysis*. Vrije University of Brussels, 1997: p. 65-83.
91. Martin-Jézéquel, V., M. Hildebrand, and M.A. Brzezinski, *SILICON METABOLISM IN DIATOMS: IMPLICATIONS FOR GROWTH* *Journal of Phycology*, 2000. **36**(5): p. 821-840.
92. Shrestha, R.P., et al., *Whole transcriptome analysis of the silicon response of the diatom Thalassiosira pseudonana*. *BMC genomics*, 2012. **13**(1): p. 499.
93. Fuentes, S., G.H. Wikfors, and S. Meseck, *Silicon Deficiency Induces Alkaline Phosphatase Enzyme Activity in Cultures of Four Marine Diatoms*. *Estuaries and Coasts*, 2013: p. 1-13.
94. Pan, Y., S.S. Bates, and A.D. Cembella, *Environmental stress and domoic acid production by Pseudo-nitzschia: a physiological perspective*. *Nat Toxins*, 1998. **6**(3-4): p. 127-35.
95. Ianora, A., et al., *The Relevance of Marine Chemical Ecology to Plankton and Ecosystem Function: An Emerging Field*. *Marine Drugs*, 2011. **9**(9): p. 1625-1648.
96. Rayko, E., et al., *Transcription factor families inferred from genome sequences of photosynthetic stramenopiles*. *New Phytol*, 2010. **188**(1): p. 52-66.
97. Huysman, M.J.J., et al., *AUREOCHROME1a-mediated induction of the diatom-specific cyclin dsCYC2 controls the onset of cell division in diatoms (Phaeodactylum tricornutum)*. *Plant Cell*, 2013. **25**(1): p. 215-228.
98. Ohno, N., et al., *CO₂-CAMP-responsive cis-elements targeted by a transcription factor with CREB/ATF-like basic zipper domain in the marine diatom Phaeodactylum tricornutum*. *Plant physiology*, 2012. **158**(1): p. 499-513.

99. Lin, S., et al., *INTENSE GRAZING AND PREY-DEPENDENT GROWTH OF PFIESTERIA PISCICIDA (DINOPHYCEAE) 1*. Journal of phycology, 2004. **40**(6): p. 1062-1073.
100. Brzezinski, M., R. Olson, and S. Chisholm, *Silicon availability and cell-cycle progression in marine diatoms*. Marine ecology progress series. Oldendorf, 1990. **67**(1): p. 83-96.
101. Mocquet, C., et al., *Cell cycle implication on nitrogen acquisition and synchronization in Thalassiosira weissflogii (Bacillariophyceae)*. Journal of Phycology, 2013. **49**(2): p. 371-380.
102. Franz, A.K., et al., *Phenotypic screening with oleaginous microalgae reveals modulators of lipid productivity*. ACS Chem Biol, 2013. **8**(5): p. 1053-62.
103. Orlova, M., et al., *Roles of the Snf1-activating kinases during nitrogen limitation and pseudohyphal differentiation in Saccharomyces cerevisiae*. Eukaryotic cell, 2010. **9**(1): p. 208-214.
104. Serfontein, J., et al., *Evolution of the TSC1/TSC2-TOR signaling pathway*. Sci Signal, 2010. **3**(128): p. ra49.
105. Fokina, O., et al., *Mechanism of 2-oxoglutarate signaling by the Synechococcus elongatus PII signal transduction protein*. Proc Natl Acad Sci U S A, 2010. **107**(46): p. 19760-5.
106. Ermilova, E., et al., *PII signal transduction protein in Chlamydomonas reinhardtii: localization and expression pattern*. Protist, 2013. **164**(1): p. 49-59.
107. Nedelcu, A., *Comparative Genomics of Phylogenetically Diverse Unicellular Eukaryotes Provide New Insights into the Genetic Basis for the Evolution of the Programmed Cell Death Machinery*. Journal of Molecular Evolution, 2009. **68**(3): p. 256-268.
108. Malyshev, D.A., et al., *A semi-synthetic organism with an expanded genetic alphabet*. Nature, 2014.
109. Vitousek, P.M., *Litterfall, nutrient cycling, and nutrient limitation in tropical forests*. Ecology, 1984. **65**(1): p. 285-298.
110. Smith, S.E., F.A. Smith, and I. Jakobsen, *Mycorrhizal fungi can dominate phosphate supply to plants irrespective of growth responses*. Plant physiology, 2003. **133**(1): p. 16-20.
111. Long, C.P. and M.R. Antoniewicz, *Metabolic flux analysis of Escherichia coli knockouts: lessons from the Keio collection and future outlook*. Curr Opin Biotechnol, 2014. **28c**: p. 127-133.

Chapter 4:

RNA sequencing during nitrogen starvation reveals NMB1 as a new type of transcription factor in *Phaeodactylum tricornutum*

Manuscript in preparation for publication

Matthijs M, Fabris M, Baart G, Carbonelle S, Vanden Bossche R, Vyverman W, Goossens A

Author contributions:

MM wrote the manuscript, analyzed the data and designed and performed most of the experiments

ABSTRACT

Unicellular algae often inhabit highly variable habitats where they are confronted with a wide variety of stresses. Diatoms thrive in high nutrient conditions with relatively low light intensity. One of the main challenges to algal growth is the shortage of nutrients, of which nitrogen deficiency is one of the most common. However, very little is known about the genetic regulation of the switch from growth to cell division arrest during nitrogen starvation. In this study we profiled the transcriptome of the model diatom *Phaeodactylum tricornutum* in the early phases of nitrogen starvation and used a bioinformatics approach to identify overrepresented motifs in the promoters of upregulated genes during this process. One of the identified motifs was shown to be bound by a transcription factor termed *Nitrogen Motif Binding 1* (NMB1). NMB1 is part of a previously uncharacterised family that is conserved in heterokont algae. This transcription factor is one of the first to be linked with nitrogen starvation and shows the feasibility of identifying unknown transcription factors using a bio-informatics motif discovery approach combined with a high throughput DNA binding assay.

INTRODUCTION

In contrast to terrestrial habitats, photosynthesis in the ocean is performed by organisms with a varied phylogenetic background. Green, red and brown eukaryotic algae all contribute significantly to oceanic primary productivity (Falkowski et al. 2004). With the exception of green algae, the study on the genetics of these algae has only just begun. Diatoms are important components of marine phytoplankton that thrive in well mixed waters (Tozzi et al. 2004). The heterokonts, the superphylum to which diatoms belong, arose after a secondary endosymbiosis event between a heterotroph and a red algae. The red algae was reduced to a symbiont, but the four membranes surrounding the chloroplast and many genes of a red algal origin are remnants of this origin. Diatoms are an important part of the marine phytoplankton and over 200.000 species are expected to exist (Armbrust 2009). Besides their pigmentation and origin they are also unique because of their silica cell wall. It has been hypothesized that because the energy requirements for a glass cell wall are lower than that of carbohydrates polymers, they can devote more of their energy to nutrient acquisition, a process where they often outcompete other algae (Martin-Jézéquel et al. 2000; Ingall et al. 2013). Because of their evolutionary distance to land plants, diatoms genomes contain metabolic pathways that were unexpected for a photosynthetic organism, such as a complete urea cycle and a not fully understood carbon concentrating mechanism (Giordano et al. 2005).

Diatoms become particularly abundant when there is an influx of nutrients, either seasonal or when the ocean is artificially fertilized with iron, resulting in large blooms (Marchetti et al. 2012). Besides light availability, lack of nutrients is the main limit to marine primary productivity (Elser et al. 2007). The limiting nutrient is not the same in every location. Ocean tracts near coastal waters are usually deficient in nitrogen, while the open ocean commonly lacks sufficient iron for algal growth. One possible explanation

for the success of brown photosynthetic organisms in marine habitats is their ability to tolerate long term nutrient deprivation and rapidly assimilate available nutrients when conditions change.

The adaptations of diatoms to their ecological niche must be contained within their diverse genomes. There are four completed diatom genome projects, and several more nearing completion (Bowler et al. 2008; Armbrust et al. 2004; Lommer et al. 2012). Among the sequenced diatoms are the two main model systems: the centricate *Thalassiosira pseudonana* and the pennate *Phaeodactylum tricornutum*. Most of the molecular data has been derived from these two species. Two others are also pennates: the polar *Fragilariopsis cylindrus* while the centricate *Thalassiosira oceanica* was chosen for sequencing, as unlike the other mentioned diatoms it inhabits the iron poor regions of the open ocean. The genomes show that there is a large fraction of diatom specific genes and that they have taken up genes from a variety of other organisms through horizontal gene transfer (Bowler et al. 2008). The amount of overlap between diatom genomes is not very high: for example *P. tricornutum* and *T. pseudonana* only have 40% of genes in common. This results in very different responses to nutrient deprivation, e.g. the lack of the ferritin and plastocyanin genes in *T. pseudonana* makes it less adapted to iron starvation compared to those who have alternative electron carriers such as *P. tricornutum* or *T. oceanica* (Lommer et al. 2012).

Because diatoms are only distantly related to better-studied groups including metazoans, fungi or land plants, a large number of diatom genes do not have assigned functions since many genes have no orthologs in other model species. This high proportion of genes with unknown functions has not hampered the exploration of diatom metabolism through orthology based approaches. To date, these have resulted in a number of metabolic models that catalog all likely reactions within the diatom cell (Kroth et al. 2008; Fabris et al. 2012a). Using this approach, it was shown that diatoms contain a number of unusual pathways for photosynthetic organisms such as a complete urea cycle, the Entner-Doudoroff alternative glycolysis and several carbon concentrating mechanisms aiding photosynthesis. These are all evidence that diatom metabolism is uniquely adapted to aquatic life (Fabris et al. 2012b; Allen et al. 2011). However many questions still remain especially in terms of the dynamics of the metabolism (Zheng et al. 2013).

It has become clear that diatoms differ in many respects from green algae. Uptake of nitrogen in green algae for example only happens in the light while diatoms are able to use stored carbon as an energy source and a building block for amino acids (Bender et al. 2012).

Besides the ecological importance, most nutrient depletions also stimulate lipid accumulation in algae with nitrogen starvation as one of the most studied triggers for lipid production in micro-algae (Mühlroth et al. 2013). Because of the ecological importance and economic potential, several studies have looked at nitrogen starvation in *P. tricornutum* and *T. pseudonana* (Hockin et al. 2012a; Valenzuela et al. 2012; Yang et al. 2013). In contrast to the previous studies we decided to focus on the transition of exponential growth to full cell cycle arrest (Hockin et al. 2012b; Valenzuela et al. 2012). In this study we exclusively used *P. tricornutum* as a model even though it is an atypical diatom. The used strain has different morphotypes and

unlike most diatoms it maintains its cell size after division and therefore does not need to undergo sexual division. Despite these atypical features, it has many benefits as a diatom physiological model species such as a rapid generation time and a wide body of available literature.

In this study RNA-sequencing was used to search for genes under transcriptional regulation during nitrogen starvation. Starting from this list of upregulated genes, a bioinformatics approach was used to identify overrepresented motifs and the degree of evolutionary conservation in this response. The main aim of this study was to find regulators that steer the transition from growth to survival under nitrogen deficiency.

RESULTS

GENE EXPRESSION PROFILING OF NITROGEN STARVED CELLS

Nitrogen starvation is well known to halt cell division and increases lipid production in diatoms (Yongmanitchai and Ward 1991). For all performed experiments exponentially growing cells were harvested and transferred to nitrogen free medium or replete medium. Samples for RNA were taken at 4, 8 and 20 hours after medium transfer. Within this timeframe cell density doubled in the control medium while only 75% OD increase was seen in the nitrogen starved cells. We focused our transcriptomic profiling on the first 20 hours after the start of the stress conditions because it roughly corresponds to the time required for one cell division in nutrient-replete conditions. During this period, the cells under N starvation started to accumulate lipids, began bleaching and halted their cell cycle (Chapter 5). Exponential growth in the control sample continued for at least another 12 hours after the end of the experiment. It must be noted that several genes related to iron starvation were beginning to increase in the control sample after hour 20, indicating that iron concentrations are starting to become limited although growth continued.

RNA-seq was performed on the extracted RNA using an Illumina Hi-Seq. After the removal of low quality reads, we obtained on average 16 ± 2.5 million reads (95% CI), of which 93% \pm 1% of reads could be mapped to the genome, indicating that the reference genome is close to complete. However gene models were less accurate, as over 20% of the models reported in the original genome paper lack a start or stop codon. Consequently, start codons, stop codons and splice junctions of transcripts differed to a large degree from the gene models. The amount of point mutations was also high compared to the reference sequence, a phenomenon also seen in other labs (Chris Bowler, personal communication).

After Cuffdiff analysis, using default settings, 4770 genes showed statistically significant differential expression at some point during the experiment. This large number was not unexpected as nitrogen starvation halts photosynthesis and protein synthesis while simultaneously the cell division halts. This necessitates major adaptations to cell metabolism and the cell cycle. Beside responses caused directly by the

lack of nitrogen, such as the decreased protein synthesis, secondary effects are also expected e.g. due to decreased photosynthesis.

Of the genes that differed over four fold compared to the control, 771 genes were shared across all nitrogen starved samples. To get a general overview of the expression patterns a cluster analysis was performed using the TMEV program. Transcripts were grouped into 11 clusters using the self organizing tree algorithm (fig. 1). Remarkably, downregulated genes are divided in only three clusters while upregulated genes took up the eight remaining categories. This is suggestive of a smaller set of transcriptional regulators controlling this pattern. In the following sections the most apparent changes are described in more detail.

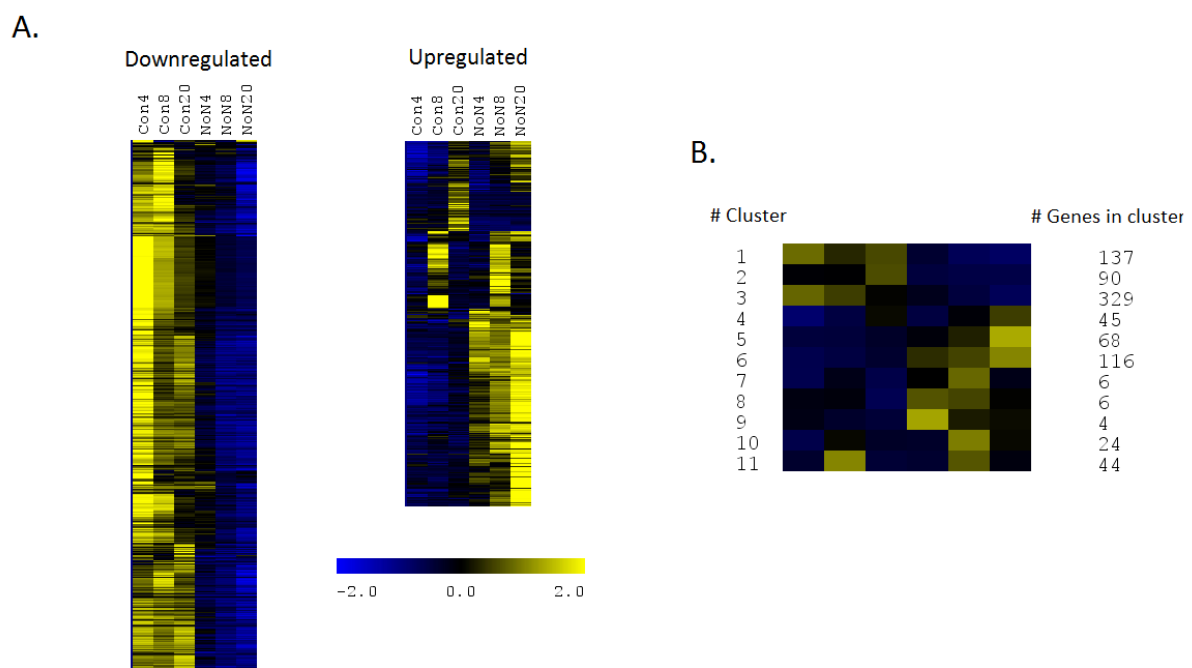


Figure 1: A. Overview of differentially expressed genes that differ more than $\log_{1.5}$ fold from the relevant control sample at the same timepoint. Normalized FPKM values, ordered using hierarchical clustering as implemented in the TMEV4 tool B. Self organizing tree of the same data

Photosynthesis

Nitrogen deficiency reduces the efficiency of the photosynthetic apparatus, in particular that of photosystem II (Kolber et al. 1988). Many genes encoding proteins involved in photosynthesis like the oxygen enhancer complex or the photosystem stability factor (Phatr13895) decreased markedly in expression level. As reported previously, roughly half of all Calvin cycle genes annotated in the Diatomcyc pathway annotation database are downregulated. While most fucoxanthine binding proteins (FCP) genes had decreased transcription, some were upregulated such as Phatr38720 and especially Phatr16481. Likely

these served to dissipate excess energy. The cell appears to be under oxidative stress since two glutathione peroxidases were also highly induced (Phatr50084 & Phatr48636). A peptide methionine sulfoxide reductase Phatr12243, which reduces oxidized methionine residues gradually increased to peak after four hours and gradually reduced in transcript abundance afterwards. Correspondingly, the cells were visually bleached after 20 hours and chlorophyll biosynthesis genes dropped severely in expression.

Protein synthesis

In our dataset protein synthesis showed a very abrupt halt. The earliest time point, four hours after medium change, showed that most ribosomal components have already decreased dramatically in expression (cluster1). Of the 77 genes that were annotated as ribosomal, 52 have their lowest expression values during nitrogen starvation timepoints. The eukaryotic 40S and 60S components show a very similar transcription pattern: after four hours of N starvation the ribosomal components were already down regulated and this decline continued during the subsequent hours. The expression patterns of genes coding for 50S or 30S ribosomal components proteins however did not cluster together. Additionally several t-RNA charging molecules show the same pattern (Phatr51120 & Phatr2394).

Nitrogenous compounds breakdown and recycling

Cells under nitrogen starvation increased their capacity to absorb and assimilate nitrogen. Several nitrate transporters are upregulated (Phatr26029 & Phatr54101), together with the nitrate reductase (Phatr54983) and nitrite reductases (Phatr8155 & Phatr13154). These processes show an expression maximum after 20 hours of starvation and are represented in cluster 5. Transporters in general are common among the most highly expressed genes. It should be emphasized that this increased expression does not necessarily correspond to increased protein levels. In the diatom *Cylindrotheca fusiformis* nitrate reductase expression is maintained during starvation but translation only occurs upon addition of nitrate (Poulsen and Kroger 2005).

One way of coping with the reduced availability of N is to increase turnover of nitrogen containing metabolites and polymers. Proteins account for most N in the cell and correspondingly many proteases such as the papaine-like protease (Phatr50321) and the cysteine protease (Phatr25433) have increased transcripts, but this upregulation is more gradual and less distinct. While some amino acid degradation enzymes are present in cluster #5, such as tyrosine aminotransferase (Phatr51609) and proline dehydrogenase (Phatr13232), most of the amidases and aminotransferase are present in cluster #6. These genes respond immediately and peak at 20 hours after medium change. Nucleic acid breakdown also occurs as several genes in purine and pyrimidine degradation or salvaging are upregulated (Phatr15968 & Phatr49782). Other cellular components containing nitrogen are also being recycled. Six proteins containing an amidase domain, which breaks carbon-nitrogen bonds, are present in cluster 6. These are likely involved in the recycling of nitrogen containing proteins. One of the highest induced genes during

N starvation is Phatr55010, containing a pyridoxal binding domain which is often involved in transaminase reactions(Lim et al. 1998).

Incorporation of the liberated ammonia appears to happen through Glutamine oxoglutarate aminotransferase (GOGAT) action (Phatr51214 & Phatr20342), which converts ammonia and 2-oxoglutarate to glutamate. Glutamine synthase can further aminate glutamate to glutamine (Phatr22357). The expression of the two glutamate synthases increases steadily during the course of the experiment while glutamine synthase declines after an initial peak at four hours. Urea cycle transcripts are not significantly upregulated during nitrogen starvation and it has been suggested that this cycle is more involved in assimilating nitrogen rapidly after starvation(Allen et al. 2011).

Energy metabolism

As less carbon is being fixed through photosynthesis and the chloroplast synthesizes less ATP, a sizeable share of energy and carbon skeletons must come from recycling existing products within the cell. Most transcripts of the central TCA cycle are upregulated and it is likely that it serves as the central carbon reprocessing hub. Several TCA cycle genes are present in cluster 6, such as isocitrate dehydrogenase (Phatr14762), isocitrate lyase (Phatr14401) and aconitate hydratase (Phatr26290). This has been seen before in *T. pseudonana* (Hockin et al. 2012b). Several members of the DHLAP complex, which convert pyruvate into acetyl CoA are also present in this cluster. It has previously been reported that the pool of 2-oxoglutarate/alpha-ketoglutarate increases during N starvation (Guerra et al. 2013). Several 2-oxoglutarate (2-OG) oxygenases are upregulated hinting that the oxidation of 2-OG might be used directly as source of energy.

It has been hypothesized that a major sink of excess energy is the synthesis of lipids. Microscopic analysis with the lipophilic dye Nile red showed that lipid accumulate. However, in general transcripts involved in lipid biosynthesis generally did not increase their expression levels dramatically, and many even appear to decrease. Two notable exceptions are Phatr3262 and Phatr4551, both encoding acyltransferases, which are both also upregulated during phosphate limitation (Yang et al. 2014).

Many of the aforementioned processes take place in either the mitochondria or the chloroplast. Using TargetP we predicted the localisation of the up and downregulated genes during N starvation. Although photosynthesis is repressed in N starved cells, the proportion of chloroplast targeted proteins was roughly equal to that of mitochondrial targeted proteins (7% and 5% resp.). It is possible that these predicted localizations for the chloroplasts are underestimates since the N-terminal portion of the proteins is often truncated in the gene models. The TargetP tool used a green plant training set for the chloroplast targeting signals while differences exist with the targeting signals in diatoms(Lang et al. 1998).

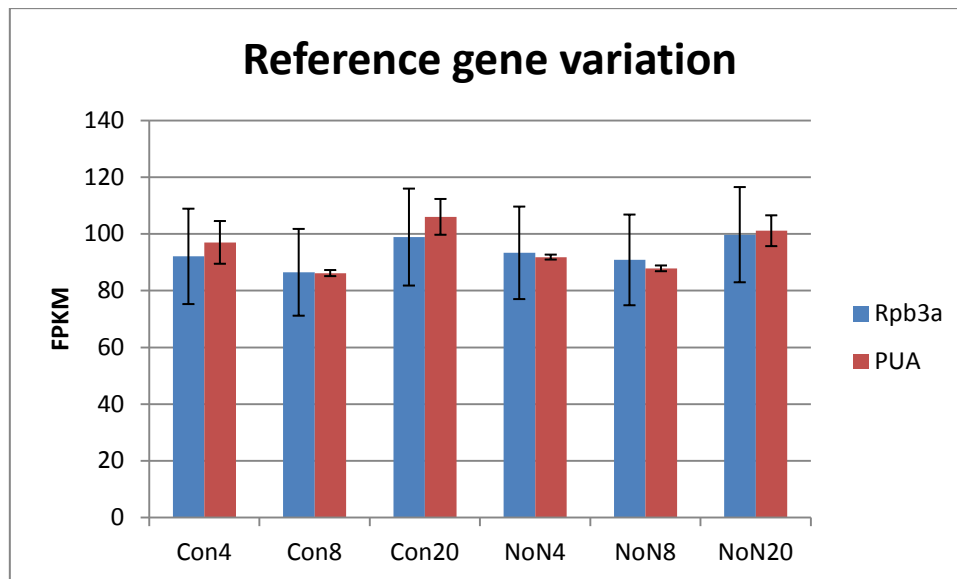


Figure 2: Reference gene variation of PUA and Rpb3a during nitrogen starvation. FPKM values, error bars represent standard deviation

QPCR was used for the validation of the RNA-sequencing results but it was soon apparent that most of the proposed reference genes for *P. tricornutum* strongly fluctuate during nitrogen starvation, with the notable exception of the TATA binding protein (Phatr10199) (Siaut et al. 2007). Expression of Phatr10199 however was below the median of our dataset. We therefore searched for genes with little variation during our assayed condition. Two genes associated with RNA polymerase emerged as suitable candidates: Phatr13566 a rpb3a domain containing protein and Phatr16787 containing a PUA domain (IPR015947), the absolute expression is represented in figure 2. The coefficient variation of the selected genes was on average lower than 20%, outperforming the previously described reference genes by a wide margin during nitrogen starvation e.g. CV 55% for Histon H4.

COMPARISON WITH PREVIOUS STUDIES AND OTHER SPECIES

The available data on *T. pseudonana* shows that out of the 28 transcripts highly upregulated during N starvation, 12 have orthologs in *P. tricornutum* (Mock et al. 2008). Almost all of these genes are also upregulated during N starvation and comprise the previously reported processes of amino acid degradation, oxidative stress and TCA-cycle involvement. These findings support the main conclusions of our study. At the time of writing, there were few transcriptomic datasets available for other species, therefore it was decided to look at which genes were present in all sequenced diatoms. Using a reciprocal best blast hit approach, orthology tables were made for *Thalassiosira pseudonana*, *Fragiliariopsis cylindrus* and the draft genome of *Pseudo-nitzschia multiseriis*. Although there are other more stringent methods for orthologous group determination, this approach was deemed sufficient to outline the conserved processes.

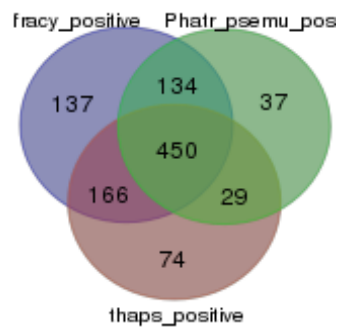


Figure 3: Venn diagram of Orthologous genes shared between *Fragiliariopsis cylindrus* (Fracy), *Pseudonitzschia multiseries* and *Thalassiosira pseudonana* (thaps). Only those genes are shown where the *P. tricornutum* orthology is 2 fold upregulated during N starvation conservation (Total 1481 genes for Phatr), 24% is shared

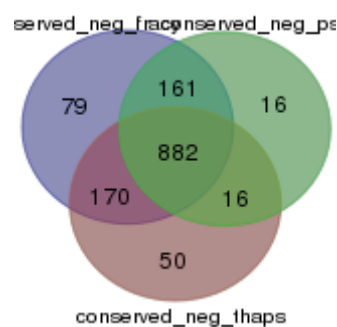


Figure 4: Venn diagram of Orthologous genes shared between *Fragiliariopsis cylindrus* (Fracy), *Pseudonitzschia multiseries* and *Thalassiosira pseudonana* (thaps). Only those genes are shown where the *P. tricornutum* orthology is 2 fold upregulated during N starvation conservation (Total 1481 genes for Phatr), 24% is shared 2fold downregulated during N starvation (Phatr 1600 genes), 55% is shared

Of the genes upregulated at least two fold during nitrogen starvation only a quarter had orthologs in all three investigated species(fig. 3). Looking at downregulated genes however we saw that over half was shared among all species(fig. 4). In total 43% of genes are shared among the three diatoms based on orthologous groups(Vandepoele et al. 2013). Unsurprisingly the shared core of nitrogen responsive genes was mainly involved in primary metabolism and nitrogen uptake. In the *P. tricornutum* specific genes the number of genes with unassigned functions is higher, but otherwise no clear enrichment for any process could be seen in this group. The pennate diatoms *P. multiseries* and *F. cylindrus* are more closely related to each other than to *P. tricornutum*, there is little overlap in orthologous pairs(Lundholm et al. 2002).

While no data is available for nitrogen starvation, comparative studies have taken place for iron and silicon starvation. Changes in gene expression caused by iron starvation showed little overlap in *T. pseudonana* and *P. tricornutum* (Allen et al. 2008).

TWO HIGH INFORMATION CONTENT MOTIFS ARE OVERREPRESENTED IN NITROGEN STARVATION RESPONSIVE PROMOTERS

As 86 transcription factors were upregulated during nitrogen starvation, it was unfeasible to screen all by gain-of-function experiments in transformed diatoms. We decided to find overrepresented motifs in genes responsive to N deficiency and subsequently screen for transcription factors that can bind these motifs. From our RNA-sequencing dataset, 861 genes were selected with the highest fold induction during nitrogen starvation and a median expression higher than 20 FPKM. The intergenic sequences upstream of the coding sequences were selected and used as input for the MEME program to identify motifs in these promoters. The putative promoters were selected as being 500bp upstream from the start codon. The genome of *P. tricornutum* is gene dense and the selected interval often overlapped with the promoter or UTR of the previous gene. These short intergenic spaces make statistical analysis and motif discovery difficult as two genes with different expression profiles were often assigned overlapping intergenic sequences. It has been reported that genes from the same pathway or related process are spaced closely together (Bowler et al. 2008). Despite these genomic and functional overlaps, the expression of adjacent genes is only rarely correlated in our dataset with the notable exception of a cluster of histon H2 genes that have a Pearson score of almost 0.9.

Using a chi-squared based approach, only the motifs that were significantly enriched in the test set were kept (table 1). Motifs that showed a high repetition were also excluded. Using this approach, only two motifs were retained: Motif6 and Motif17. Initially motif6 was deemed the most promising as it contained the TTC core typical of Heat Shock Factor (HSF) type transcription factors (Akerfelt et al. 2010).






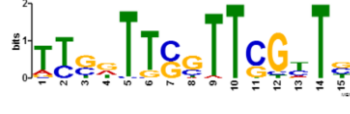
Overrepresented motifs in N up regulated genes			
Name	Motif	p value	# sites in test set
Motif 17		6.07 e-04	26
Motif 6		6.32 e-03	27
Overrepresented motifs in N down regulated genes			
Name	Motif	p value	# sites in test set
dMotif 4		6.96 e-08	100
dMotif 3		3.83 e-06	49
dMotif 17		4.37 e-04	23
dMotif15		4.65 e-03	72

Table 1: Overrepresented motifs in up and downregulated genes

Transcription factors in heterokonts have previously been characterized on the basis of sequence homology (Rayko et al. 2010). On this basis it is estimated that there are 210 TF's in *P. tricornutum*, many of which are Heat Shock Factors, a relatively small group in most other species. Of these predicted transcription factors, 86 were at some point expressed at levels two fold higher than the relevant controls. As this list was still too extensive we decided to focus only on those with the highest fold expression, a summary of which can be seen in figure 5.

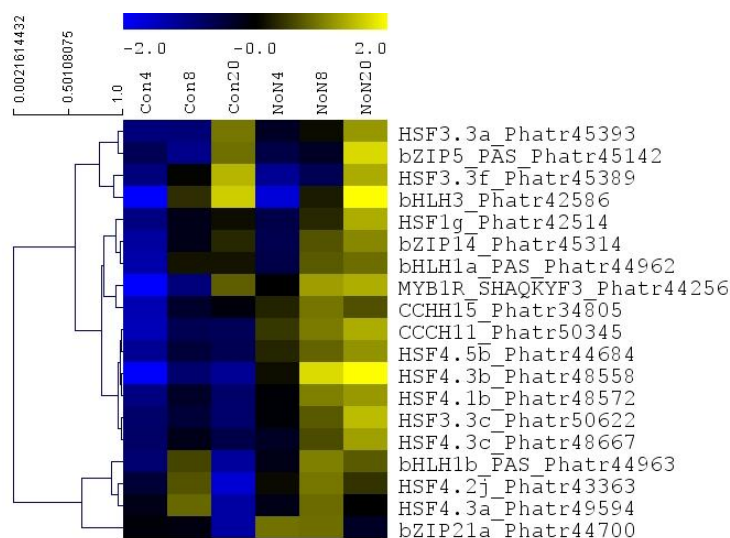


Figure 5: Normalized expression patterns of the 20 most highly induced transcription factors during N starvation

TRANSIENT PROTOPLAST ACTIVATION ASSAYS

In order to find the transcriptional factors that bind these motifs Transient Expression Assays (TEA) were performed for these motifs using 20 transcription factors that were strongly upregulated during nitrogen starvation. Motif 17 and Motif6 were placed separately in front of the firefly luciferase coding sequence and combined with the different transcription factors. No effects were seen in the assay. This could be due to poor translation of diatom transcription factors since diatom sequences were not codon optimized for plant cells. Furthermore the diatom promoters were less competent in driving transcription than typical plant promoters, and background levels in promoter only controls were thus far lower than typically observed for plant promoters. Finally, the evolutionary distance between plants and diatoms is extensive and diatom transcription factors may lack the interactors required for efficient transcription. Because the failure of the assay could be on many different levels, it was decided to abandon this approach and to use other techniques for transcription factor identification.

YEAST ONE HYBRID LIBRARY SCREENING IDENTIFIES A PROTEIN CAPABLE OF BINDING MOTIF 17

Because using the TEA assay was unfeasible for screening all reported transcription factors, the decision was made to attempt a DNA binding assay with the Yeast One Hybrid method (Y1H). An existing cDNA library was present in the department for *P. tricornutum*. This library was made from RNA isolated at different stages of growth in during a day-night light regime. Nutrient limiting conditions were not present

in the library. Open reading frames were fused to the GAL4-AD, the activating domain of the yeast GAL4 transcription factor.

Screening of motif6 was problematic as most clones showed very high autoactivation in yeast. Nevertheless we screened a colony with relatively low autoactivation using the cDNA library. Unfortunately positive colonies did not yield any recurring hits.

Screening of motif17 was more successful. After library transformation thirteen colonies were positive. The vector within ten of these colonies contained the open reading frame of Phatr50304, which we have subsequently renamed Nitrogen Motif Binding 1 (NMB1). This protein was not picked up with any previously screened motif or promoter fragment in our department. The interaction was also seen when directly transforming the motif17 bait strain with the NMB1-GAL4-AD plasmid. Curiously the coding sequence of NMB1 does not contain any known DNA binding domains. A schematic representation is given in figure 6. Interproscan searches indicated a RING type zinc finger in the N-terminal part. RING domains are usually found in E3-ligases, which determine the substrate specificity of ubiquitin conjugating enzyme complexes (Deshaies and Joazeiro 2009). Furthermore a domain called bHLH-myc (pfam14215) was identified. This domain is associated with several plant transcription factors such as MYC2, but it does not contain the DNA binding part of these proteins. This bHLH-myc type part of the protein appears to contain a GAF like domain. GAF domains are commonly associated with a sensor type function and can sense several types of signals such as light, redox status or a variety of small molecules (Unden et al. 2013; Auldrige and Forest 2011; Kelley and Sternberg 2009). Whatever signal is sensed by NMB1 is thus likely to be detected by this part of the protein. The bHLH-myc and RING domains were also found when utilizing the Phyre2 program in combination with the Backphyre option.



Figure 6: Domain organisation of the NMB1 protein, NLS: nuclear localization signal, Q: glutamine rich stretch

The NMB1 protein is predicted to have an importin- α nuclear localization signal which would fit with a transcription factor (Kosugi et al. 2009). It also contains a C-terminal stretch of glutamine residues which is commonly associated with transcriptional activation. To test the transcriptional activation the protein was fused to the GAL4 DNA binding Domain (GAL4-DBD). This fusion protein was able to drive transcription of a UAS::HIS3 selectable marker in *S. cerevisiae* (fig. 7).

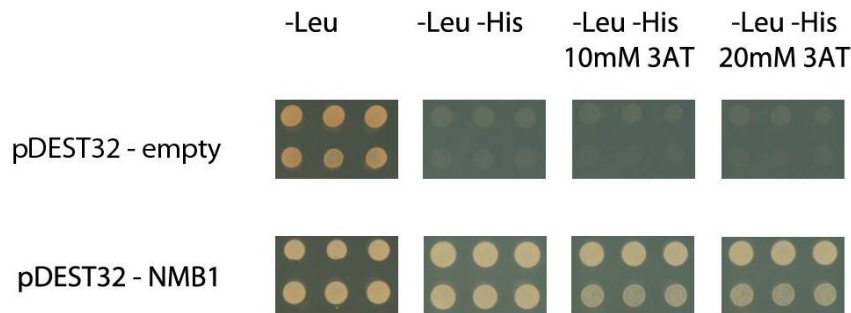


Figure 4: Autoactivation assay of NMB1 fused to the DNA binding domain of GAL4. NMB1 was able to drive the expression of the HIS3 gene placed behind the UAS sequence

In order to pinpoint the activation activity, the N-terminal part of the protein (amino acid 1 to 378) and the C-terminal half (amino acid 379 to 742) were fused to GAL4-BD. Both were able to activate the selectable marker. While puzzling at first, this does not appear to be unusual for transcription factors and has been described before (Eklund et al. 2010). DNA binding was not seen when either of these halves was fused to the GAL4-AD. The exact localization of this functionality is therefore not known.

THE RING LIKE DOMAIN OF NMB1 IS CRUCIAL FOR DNA BINDING

By aligning three full length NMB1 orthologs the most conserved cysteine and histidine residues were identified in the putative RING domain (fig. 8). RING type proteins are usually classified according to the spacing of these residues but NMB1 did not fit any of the known plant E3 ligase classes (Stone et al. 2005).

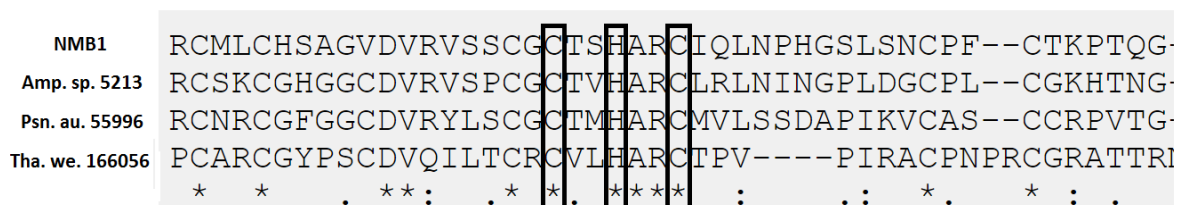


Figure 8: Alignment of the *P. tricornutum* NMB1 with three orthologs. Residues mutated to alanine for the RING-dead version are boxed. Numbers represent MMETSP CDS sequences. *Amphora species* (Amp. Sp), *Pseudo-nitzschia australis* (Psn. au.), *Thalassiosira weissflogii* (Tha. we)

To test if the RING type domain of NMB1 was functioning as a E3 ligase, an autoubiquitination assay was performed, since E3 ligases often ubiquitinate themselves. As a negative control, a version of the protein was made where the conserved cysteine and histidine residues in the RING domain were mutated into alanine (RING-dead), since this has been shown to abolish activity (Plans et al. 2006). No ubiquitination could be seen in the assay for either the native or the RING-DEAD protein.

Although RING domain-containing proteins are typically E3-ligases, they have also been previously shown to bind DNA or RNA. While the majority of reported nucleotide binding activity was non-specific,

there are a number of examples where these proteins bind specific sequences. The RING containing protein Mel18 is a transcriptional repressor, which binds a specific pattern(Kanno et al. 1995). Another example is the STYLISH1 protein of *Arabidopsis thaliana* which has been characterized as a transcription factor(Eklund et al. 2010).

In order to confirm the DNA binding of NMB1, the coding sequence was cloned in the MBP-HIS vector for expression in *Escherichia coli*. The recombinant protein was used for a protein binding microarray using 12 random nucleotides(Franco-Zorrilla and Solano 2014). The length of the probes was therefore shorter than the expected motif as the reasoning was that the final portion of the motif would likely be the DNA binding domain. No nucleotide combination was proven to be overrepresented in the results.

A second test to validate the DNA binding was a Electrophoretic Mobility Shift Assay (EMSA). The same motif of the yeast one hybrid screening was used as probe. Probe retardation was only seen after adding zinc to the reaction buffer(fig. 9). This is consistent with RING as the DNA binding as it consists out of Zinc fingers which coordinate a single Zn^{2+} ion. Standard EMSA buffers contain large amount of EDTA that can chelate this ion. Unfortunately this result could not be repeated in subsequent retrials due to unknown reasons.

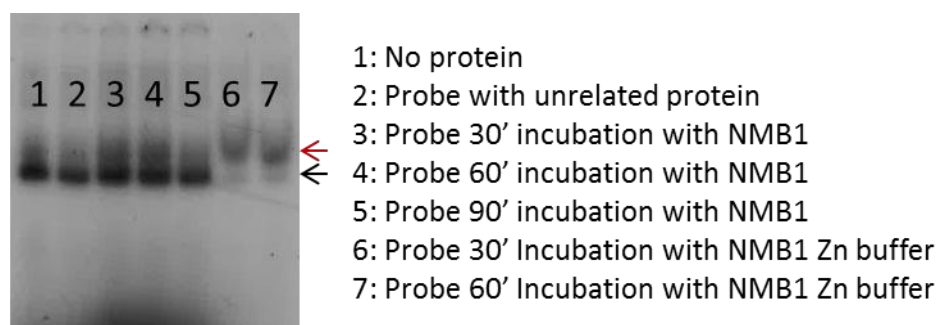


Figure 9: EMSA result of M17 with NMB1

In parallel, the RING dead construct was fused to GAL4-AD and transformed into the original motif17 Y1H strain. No DNA binding could be detected (fig. 10). These results taken together with the unusual spacing of cysteine and histidine residues suggests that the RING like domain is crucial for DNA binding.

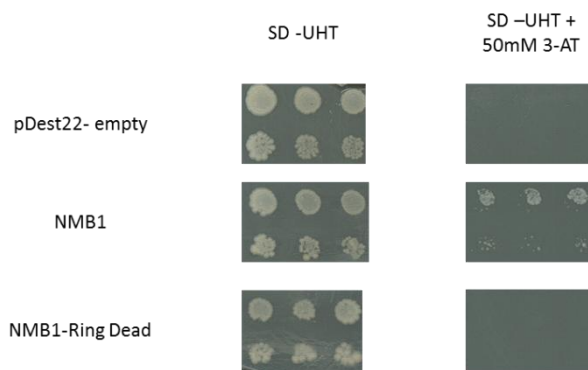


Figure 10: Growth of yeast cells containing the M17 motif upstream of the HIS3 selectable marker only show growth on 3-AT when the NMB1 RING domain is functional

NMB1 BELONGS TO A PROTEIN FAMILY CONSERVED IN DIATOMS

A BLASTp homology search showed that there are two other proteins with a high sequence homology present in the *P. tricornutum* genome: Phatr44641 (NMB2) and Phatr50305 (NMB3) which have 31% and 41% identity resp. All three proteins have a similar domain organization. Interestingly, NMB1 and NMB3 are adjactant on the genome, pointing towards each other. This peculiar organization is also present in *Fragilariopsis cylindrus*, (Fragi206149 & Fragi234602) and in *Thalassiosira pseudonana* (Thaps25161 & 10465), hinting that this duplication event is not recent. The expression pattern however is very different, NMB3 is hardly expressed in any of the conditions we tested. This tandem arrangement and poorly correlated expression was also seen for other genes in *P. tricornutum* e.g. Phatr42555 & Phatr42556.

NMB1 and NMB2-like sequences are present in each sequenced diatom investigated, and the evolutionary relationships between these orthologs is shown in figure 11. Moreover, a BLAST search showed that orthologous proteins containing the combination of the RING type domain with the bHLH-myc domain are present in all sequenced heterokonts. Besides diatoms, the more distantly related micro-algae *Nannochloropsis gaditana* and the brown seaweed *Ectocarpus siliculosus* (CBN79689) also contain this protein. Even various saprophytic water moulds of the *Phytophthora* genus contain similar proteins (E value cutoff of 10^{-3}). The glutamine rich stretch is present in most of the orthologs, but the other regions between the domains show conservation only within the diatoms. Interestingly the bHLH-myc like domain shows homology with two histidine kinase proteins in cyanobacteria e.g. YP_007142195 of *Crinalium epipsammum*.

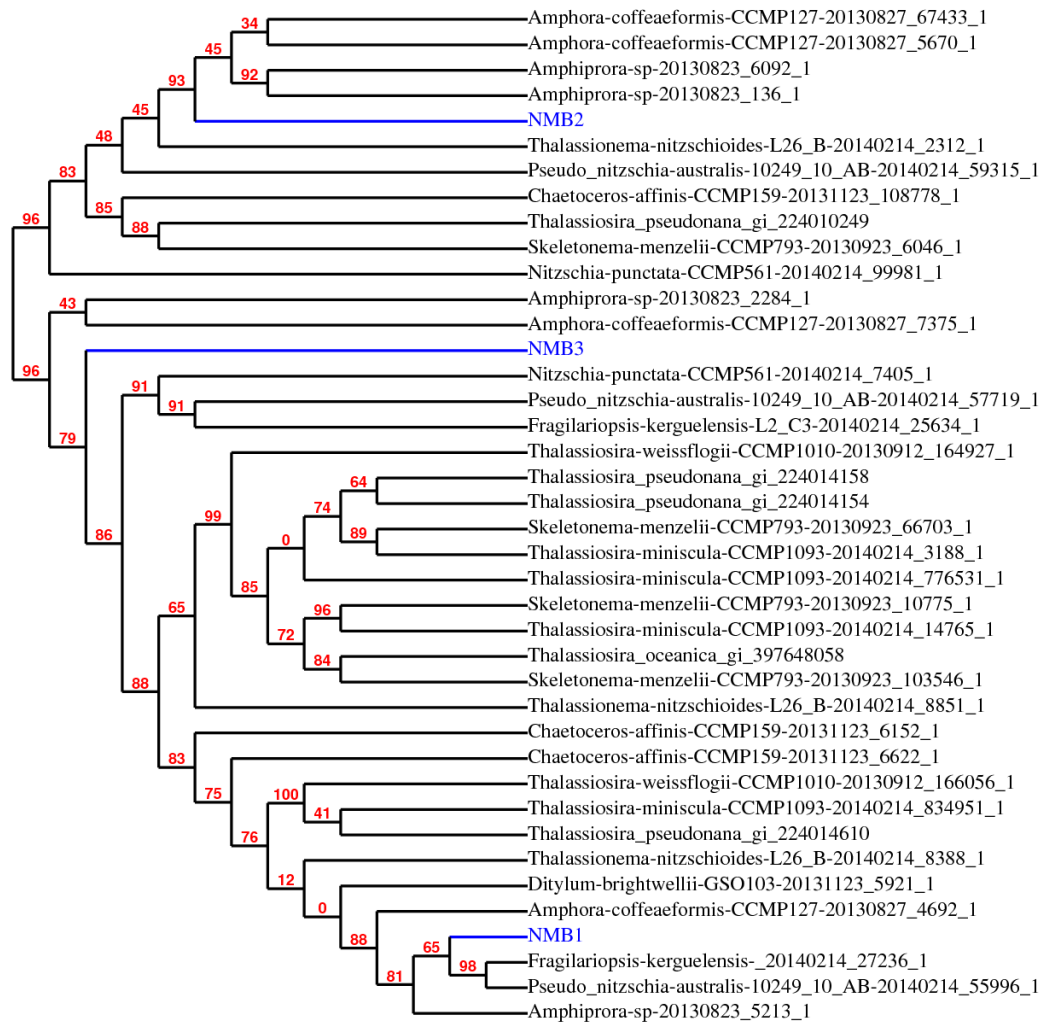


Figure 11: Phylogenetic tree of the NMB type proteins in the genomes of three diatoms with a completed genome and nine species selected from the MMETSP collection, numbers represent branch support values in percentages. Branches with less than 50% support were collapsed.

THE EXPRESSION PATTERN OF NMB1 AND NMB2 POINT TOWARDS A LIGHT MEDIATED ROLE IN NITROGEN METABOLISM

While NMB1 was picked up as binding a motif enriched in nitrogen starvation responsive genes, it was unclear which process was guided by this transcription factor. The RNA-sequencing data showed that both NMB1 and NMB2, but not NMB3, are induced during nitrogen starvation. The expression patterns of NMB1 and NMB2 were confirmed in a QPCR using newly RNA generated under identical conditions as for the RNA-seq RNA (fig. 12). However, there are large differences in the expression pattern of NMB1 and NMB2: while NMB1 is also transcribed in exponentially growing cells, while NMB2 is much more specifically induced upon nitrogen starvation, with 15 fold induction four hours after the removal of nitrogen. Furthermore, when looking at an additional RNA-sequencing library it was seen that upon

reillumination after a period of darkness NMB1 was highly induced while NMB2 levels remained low (fig. 13).

The publicly available EST data showed that NMB3 had higher expression during urea feeding (Maheswari et al. 2009). We checked gene levels upon urea feeding but only saw mild induction (data not shown). No further attention was paid to NMB3 as even though it lies directly upstream of NMB1, it is only very slightly expressed during the assayed conditions and associated controls.

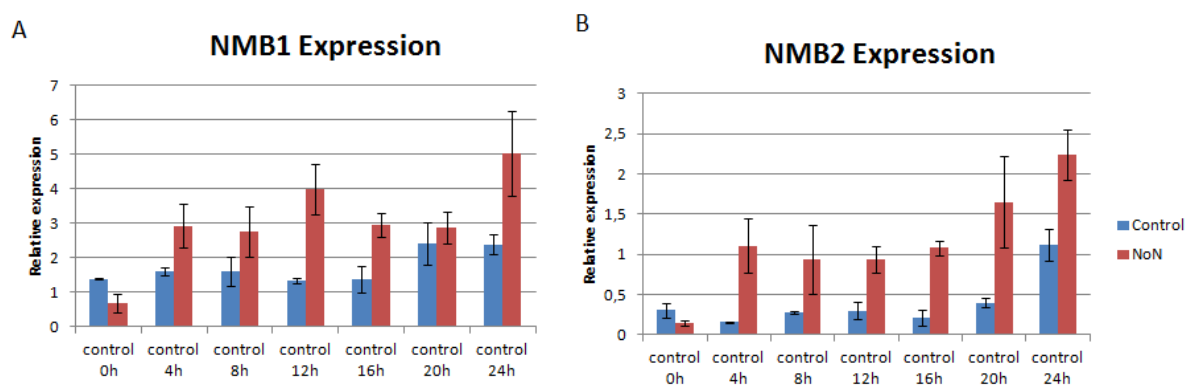


Figure 12: Independent validation of NMB1 and NMB2 expression patterns during N starvation

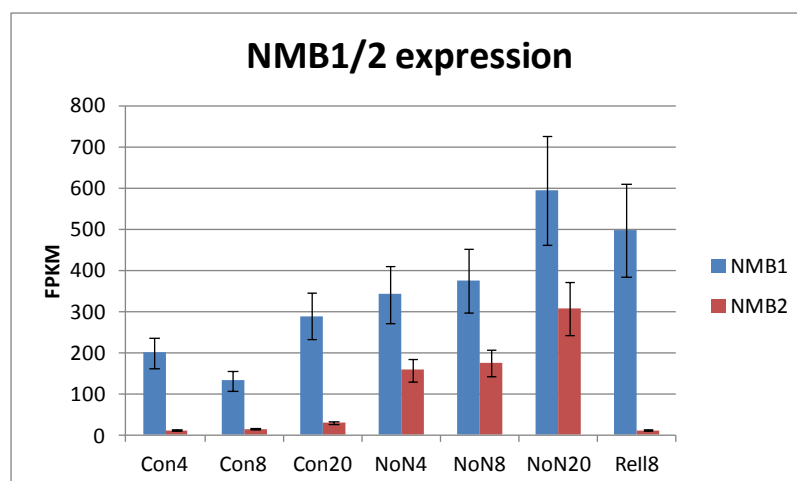


Figure 13: Expression pattern of NMB1 and 2 as seen in the RNA-seq dataset

A useful tool for the functional elucidation of unknown genes is co-expression, as genes involved in the same process often show similar expression patterns (Vandepoele et al. 2009). The twenty genes with the highest correlation to NMB1 and NMB2 respectively are listed in figure 14 & 15. Although both clusters are distinct, both contain genes involved in the assimilation of nitrogen and to a lesser degree genes involved in redox reactions. Interestingly, the NMB1 cluster contains two genes related to the biosynthesis of molybdopterin (Phatr15652 & Phatr35473), a cofactor required for nitrate reductase function.

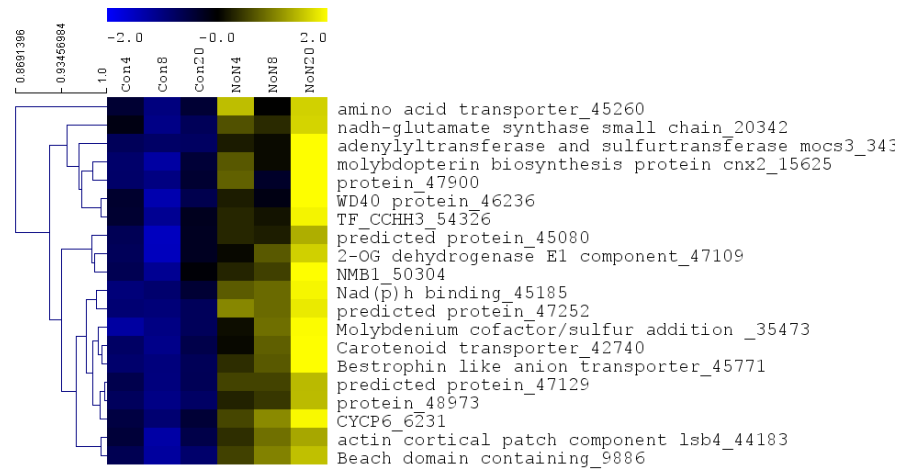


Figure14: Co-expression neighborhood of NMB1 generated using the CummrBund Package and visualized using TMEV4

In the co-expression neighborhood of NMB2, several genes are present related to synthesis of amino acids (Phatr15217, Phatr16499, Phatr51214) and the transport of nitrogen containing compounds (Phatr17344, Phatr54560, Phatr52619, Phatr27877). While this is only indirect proof, it points towards a role in nitrogen metabolism distinct from that NMB1.

The presence of motif 17, which is bound by NMB1, was also checked in the intergenic sequences of the distantly related centric diatom *T. pseudonana*. Filtering the positive hits by only retaining those genes that are present in both diatoms resulted in twenty intergenic sequences. Interestingly, among these genes there was both a glutamate synthase (Phatr51214/Thaps269900) and a nitrate reductase (Phatr54983/Thaps25299). The ortholog of *T. pseudonana* glutamate synthase is present in the co-expression cluster of NMB2(fig. 15).

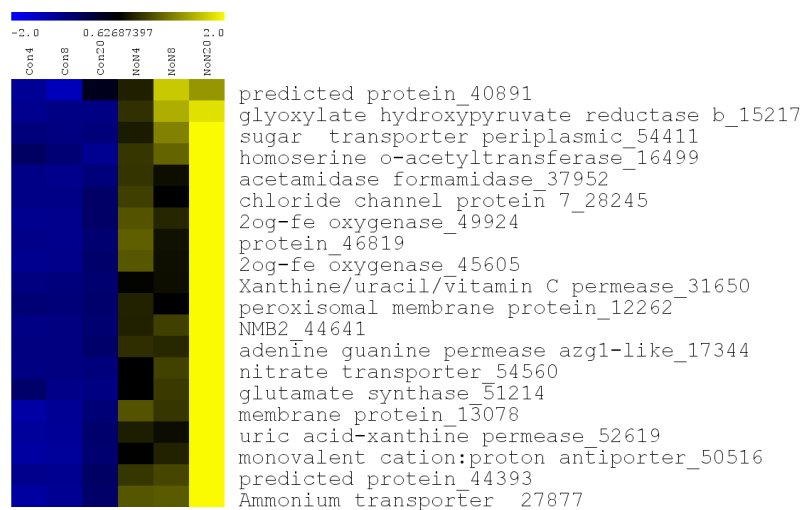


Figure 15: Co-expression neighborhood of NMB1 generated using the CummrBund Package and visualized using TMEV4, color scale represents normalized FPKM values

Materials and methods are presented in chapter 6.

DISCUSSION

In summary we have performed a comprehensive transcriptomic study of the *P. tricornutum* during the early stages of nitrogen deprivation. As nitrogen is key to both nucleotide and protein synthesis the changes in the transcriptome are massive compared to exponential growth. This study combined with those done before by Valenzuela and Yang, have provided the algal community with a set of markers for nitrogen starvation in *P. tricornutum*.

Despite the importance of transcriptional regulation, few transcription factors have been functionally characterized in diatoms. The amount of transcription factors in *P. tricornutum* was estimated at 212 by Rayko *et al.*, which is more than the 147 predicted for the green algae *Chlamydomonas reinhardtii* but in the same range as the 216 predicted TF's in the unicellular yeast *S. cerevisiae* (Riano-Pachon *et al.* 2008; Gordan *et al.* 2011). Interestingly, the centricate diatom *T. pseudonana* contains over thirty more transcription factors while having a similar gene number. Of this extensive set of TFs, in *P. tricornutum* only the blue light responsive Aureochrome 1a and the CO₂ responsive bZIP11 has been described (Huysman, 2012, Ohno 2010). Our study has shown the feasibility of identifying transcription factors in diatoms based solely on expression data and the genome. The result of this approach was the identification of a novel family of heterokont specific transcription factors that are likely involved in the assimilation of nitrogen. Our results have important implications, showing that the predicted number of transcription factors in diatoms is likely to be an underestimate. Poorly sampled branches of the evolutionary tree often contain previously unknown families of DNA binding proteins. In the alveolate *Entamoeba histolytica*, for example, several novel transcription factors with no recognizable DNA binding domain were discovered (Gilchrist *et al.* 2001; Pearson *et al.* 2013).

The widely reported increases in lipid synthesis during nitrogen starvation are likely to be post-transcriptionally regulated. This could happen either through extra synthesis of these proteins or by increasing the activity of the existing enzymes. The latter seems more likely as it has been reported that the activity of key enzymes such as Acetyl CoA Carboxylase is tightly regulated in mammalian cells. This protein complex is inhibited by phosphorylation of the energy sensing complex AMPK and allosterically activated by citrate (Brownsey *et al.* 2006). Similar regulation can be expected because all AMPK components are present in *P. tricornutum* and there are increased levels of TCA intermediates in nitrogen starved cells (Guerra *et al.* 2013).

The expression patterns of NMB1 and NMB2 clearly point towards a role in regulating nitrogen assimilation, but unfortunately we were unable to generate overexpression lines for either of these proteins. It could be that overexpression is lethal for the cell, but it is more likely that the transformations failed because of technical reasons. While it is possible that the overexpression of these transcription factors would have resulted in the increased transcription of target genes, there is a distinct possibility that NMB1 and NMB2 are some kind of sensors and might require additional signals for transcriptional

activation. Now that knock out lines can be generated in transformable diatoms it will be possible to ascertain whether the nitrogen starvation response of the cell can still function after the inactivation of NMB1. This seems to be a more solid approach than using either overexpression or RNAi.

The coupling of *de novo* cis-motif discovery with Yeast One Hybrid screening has been used before for transcription factor discovery. In principle this method can be used for any condition, but there is certainly potential for a more extensive look at nitrogen specific motifs in *P. tricornutum*. A major improvement would likely come from the use of a cDNA library, specifically made from N deficient cells. This approach would also benefit from updated gene models as many of those published are clearly truncated at both 3' and 5' prime end. A more accurate delimitation of the intergenic regions would reduce the number of false positives. The rising number of RNAseq studies, such as ours, should make it possible to greatly improve these gene models.

In comparison to other studies, this approach showed that the nitrogen deprivation response is not static and many responses are transient. It is clear that more in depth analysis need to be done with more timepoints and different diatom species. Particularly intriguing is the manner in which the diatom cell would initially sense the lack of nitrogen.

Unfortunately we were unable to confirm DNA binding with either EMSA or a protein binding nucleotide array. While this might be caused by a myriad of technical reasons, it has been reported that the GAF domain – present in NMB1 – requires the attachment of a cofactor, a process which is unlikely to be completed when expressing the gene in an *E. coli* or *S. cerevisiae* background. While there are many types of co-factors that can be attached to the GAF domain, some of the most common are light sensing molecules termed billins in cyanobacteria or phytochromobilins in plants (Fischer et al. 2005; Yoshihara et al. 2006).

Because the expression patterns of NMB1 and NMB2 are similar but not identical, it is possible that they both regulate a similar process in different phases. The most striking difference between the two is the peak in NMB1 expression in a sample incubated for eight hours in the dark and briefly exposed to the light. One of the possible points where light signaling and nitrogen biosynthesis intersect is the regulation of the two isoforms of both nitrite reductase (NiR) and glutamate synthase (GOGAT). Diatoms contain plastidial enzymes of these enzymes which require ferredoxin (Fd) as an electron donor and cytosolic enzymes that uses NAD(P)H. The ferredoxin and NAD(P)H are distinct in their expression patterns during day/night cycles with the plastidial Fd-GOGAT and Fd-NiR showing similar trends. NMB1 and NMB2 might be the regulators for these light/dark versions. This would require that the NMB proteins have some way of sensing light and this could either be due to bound pigments mentioned previously or to the integration of a redox sensor such as a heme group (Auldrige and Forest 2011). Interestingly this domain shows homology to a region of the plant transcription factor *Myc2*, which has been proven to be

under diurnal control. Although this control is apparently affected by another protein related to the circadian rhythm(Shin et al. 2012).

In conclusion this study has demonstrated the existence of a previously unknown protein family that is likely to be transcription factors. This family is conserved in stramenopiles and is involved in nitrogen metabolism.

REFERENCES

- Akerfelt M, Morimoto RI, Sistonen L (2010) Heat shock factors: integrators of cell stress, development and lifespan. *Nature reviews Molecular cell biology* 11 (8):545-555. doi:10.1038/nrm2938
- Allen AE, Dupont CL, Oborník M, Horák A, Nunes-Nesi A, McCrow JP, Zheng H, Johnson DA, Hu H, Fernie AR (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* 473 (7346):203-207
- Allen AE, LaRoche J, Maheswari U, Lommer M, Schauer N, Lopez PJ, Finazzi G, Fernie AR, Bowler C (2008) Whole-cell response of the pennate diatom *Phaeodactylum tricornutum* to iron starvation. *Proceedings of the National Academy of Sciences of the United States of America* 105 (30):10438-10443
- Armbrust EV (2009) The life of diatoms in the world's oceans. *Nature* 459 (7244):185-192. doi:10.1038/nature08057
- Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou SG, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WWY, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Oborník M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamtrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science* 306 (5693):79-86
- Auldridge ME, Forest KT (2011) Bacterial phytochromes: more than meets the light. *Critical reviews in biochemistry and molecular biology* 46 (1):67-88. doi:10.3109/10409238.2010.546389
- Bender SJ, Parker MS, Armbrust E (2012) Coupled Effects of Light and Nitrogen Source on the Urea Cycle and Nitrogen Metabolism over a Diel Cycle in the Marine Diatom *Thalassiosira pseudonana*. *Protist* 163 (2):232-251
- Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otiilar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret JP, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, Fournet J, Haruta M, Huysman MJJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kroger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin-Jezequel V, Lopez PJ, Lucas S, Mangogna M, McGinnis K, Medlin LK, Montsant A, Oudot-Le Secq MP, Napoli C, Oborník M, Parker MS, Petit JL, Porcel BM, Poulsen N, Robison M, Rychlewski L, Rynearson TA, Schmutz J, Shapiro H, Siaut M, Stanley M, Sussman MR, Taylor AR, Vardi A, von Dassow P, Vyverman W, Willis A, Wyrwicz LS, Rokhsar DS, Weissenbach J, Armbrust EV, Green BR, Van De Peer Y, Grigoriev IV (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456 (7219):239-244. doi:10.1038/nature07410
- Brownsey RW, Boone AN, Elliott JE, Kulpa JE, Lee WM (2006) Regulation of acetyl-CoA carboxylase. *Biochemical Society transactions* 34 (Pt 2):223-227. doi:10.1042/bst20060223
- Deshaies RJ, Joazeiro CA (2009) RING domain E3 ubiquitin ligases. *Annual review of biochemistry* 78:399-434
- Eklund DM, Staldal V, Valsecchi I, Cierlik I, Eriksson C, Hiratsu K, Ohme-Takagi M, Sundstrom JF, Thelander M, Ezcurra I, Sundberg E (2010) The *Arabidopsis thaliana* STYLISH1 protein acts as a transcriptional activator regulating auxin biosynthesis. *The Plant cell* 22 (2):349-363. doi:10.1105/tpc.108.064816

- Elser JJ, Bracken ME, Cleland EE, Gruner DS, Harpole WS, Hillebrand H, Ngai JT, Seabloom EW, Shurin JB, Smith JE (2007) Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecology letters* 10 (12):1135-1142. doi:10.1111/j.1461-0248.2007.01113.x
- Fabris M, Matthijs M, Rombauts S, Vyverman W, Goossens A, Baart GJ (2012a) The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *Plant J* 70 (6):1004-1014. doi:10.1111/j.1365-313X.2012.04941.x
- Fabris M, Matthijs M, Rombauts S, Vyverman W, Goossens A, Baart GJE (2012b) The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *Plant J* 70 (6):1004-1014. doi:10.1111/j.1365-313X.2012.04941.x
- Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, Taylor FJ (2004) The evolution of modern eukaryotic phytoplankton. *Science* 305 (5682):354-360. doi:10.1126/science.1095964
- Fischer AJ, Rockwell NC, Jang AY, Ernst LA, Waggoner AS, Duan Y, Lei H, Lagarias JC (2005) Multiple roles of a conserved GAF domain tyrosine residue in cyanobacterial and plant phytochromes. *Biochemistry* 44 (46):15203-15215. doi:10.1021/bi051633z
- Franco-Zorrilla JM, Solano R (2014) High-throughput analysis of protein-DNA binding affinity. *Methods in molecular biology* (Clifton, NJ) 1062:697-709. doi:10.1007/978-1-62703-580-4_36
- Gilchrist CA, Holm CF, Hughes MA, Schaenman JM, Mann BJ, Petri WA (2001) Identification and Characterization of an *Entamoeba histolytica* Upstream Regulatory Element 3 Sequence-specific DNA-binding Protein Containing EF-hand Motifs. *Journal of Biological Chemistry* 276 (15):11838-11843
- Giordano M, Beardall J, Raven JA (2005) CO₂ concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. *Annual review of plant biology* 56:99-131. doi:10.1146/annurev.arplant.56.032604.144052
- Gordan R, Murphy KF, McCord RP, Zhu C, Vedenko A, Bulyk ML (2011) Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. *Genome biology* 12 (12):R125. doi:10.1186/gb-2011-12-12-r125
- Guerra LT, Levitan O, Frada MJ, Sun JS, Falkowski PG, Dismukes GC (2013) Regulatory branch points affecting protein and lipid biosynthesis in the diatom *Phaeodactylum tricornutum*. *Biomass and Bioenergy* 59:306-315
- Hockin NL, Mock T, Mulholland F, Kopriva S, Malin G (2012a) The response of diatom central carbon metabolism to nitrogen starvation is different from that of green algae and higher plants. *Plant Physiol* 158 (1):299-312. doi:10.1104/pp.111.184333
- Hockin NL, Mock T, Mulholland F, Kopriva S, Malin G (2012b) The response of diatom central carbon metabolism to nitrogen starvation is different from that of green algae and higher plants. *Plant physiology* 158 (1):299-312
- Ingall ED, Diaz JM, Longo AF, Oakes M, Finney L, Vogt S, Lai B, Yager PL, Twining BS, Brandes JA (2013) Role of biogenic silica in the removal of iron from the Antarctic seas. *Nature communications* 4:1981. doi:10.1038/ncomms2981
- Kanno M, Hasegawa M, Ishida A, Isono K, Taniguchi M (1995) mel-18, a Polycomb group-related mammalian gene, encodes a transcriptional negative regulator with tumor suppressive activity. *The EMBO journal* 14 (22):5672-5678
- Kelley LA, Sternberg MJ (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* 4 (3):363-371. doi:10.1038/nprot.2009.2

- Kolber Z, Zehr J, Falkowski P (1988) Effects of Growth Irradiance and Nitrogen Limitation on Photosynthetic Energy Conversion in Photosystem II. *Plant Physiol* 88 (3):923-929
- Kosugi S, Hasebe M, Tomita M, Yanagawa H (2009) Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. *Proc Natl Acad Sci U S A* 106 (25):10171-10176. doi:10.1073/pnas.0900604106
- Kroth PG, Chiovitti A, Gruber A, Martin-Jezequel V, Mock T, Parker MS, Stanley MS, Kaplan A, Caron L, Weber T (2008) A model for carbohydrate metabolism in the diatom *Phaeodactylum tricornutum* deduced from comparative whole genome analysis. *PloS one* 3 (1):e1426
- Lang M, Apt KE, Kroth PG (1998) Protein transport into “complex” diatom plastids utilizes two different targeting signals. *Journal of Biological Chemistry* 273 (47):30973-30978
- Lim YH, Yoshimura T, Kurokawa Y, Esaki N, Soda K (1998) Nonstereospecific transamination catalyzed by pyridoxal phosphate-dependent amino acid racemases of broad substrate specificity. *The Journal of biological chemistry* 273 (7):4001-4005
- Lommer M, Specht M, Roy AS, Kraemer L, Andreson R, Gutowska MA, Wolf J, Bergner SV, Schilhabel MB, Klostermeier UC, Beiko RG, Rosenstiel P, Hippler M, Laroche J (2012) Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome biology* 13 (7):R66. doi:10.1186/gb-2012-13-7-r66
- Lundholm N, Daugbjerg N, Moestrup Ø (2002) Phylogeny of the Bacillariaceae with emphasis on the genus *Pseudo-nitzschia* (Bacillariophyceae) based on partial LSU rDNA. *European Journal of Phycology* 37 (01):115-134
- Maheswari U, Mock T, Armbrust EV, Bowler C (2009) Update of the Diatom EST Database: a new tool for digital transcriptomics. *Nucleic Acids Res* 37 (Database issue):D1001-1005. doi:10.1093/nar/gkn905
- Marchetti A, Schrueth DM, Durkin CA, Parker MS, Kodner RB, Berthiaume CT, Morales R, Allen AE, Armbrust EV (2012) Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences of the United States of America* 109 (6):E317-325. doi:10.1073/pnas.1118408109
- Martin-Jézéquel V, Hildebrand M, Brzezinski MA (2000) Silicon metabolism in diatoms: implications for growth. *Journal of Phycology* 36 (5):821-840
- Mock T, Samanta MP, Iverson V, Berthiaume C, Robison M, Holtermann K, Durkin C, Bondurant SS, Richmond K, Rodesch M, Kallas T, Huttlin EL, Cerrina F, Sussman MR, Armbrust EV (2008) Whole-genome expression profiling of the marine diatom *Thalassiosira pseudonana* identifies genes involved in silicon bioprocesses. *Proceedings of the National Academy of Sciences of the United States of America* 105 (5):1579-1584. doi:10.1073/pnas.0707946105
- Mühlroth A, Li K, Røkke G, Winge P, Olsen Y, Hohmann-Marriott MF, Vadstein O, Bones AM (2013) Pathways of lipid metabolism in marine algae, co-expression network, bottlenecks and candidate genes for enhanced production of EPA and DHA in species of Chromista. *Marine drugs* 11 (11):4662-4697
- Pearson RJ, Morf L, Singh U (2013) Regulation of H₂O₂ stress-responsive genes through a novel transcription factor in the protozoan pathogen *Entamoeba histolytica*. *Journal of Biological Chemistry* 288 (6):4462-4474
- Plans V, Scheper J, Soler M, Loukili N, Okano Y, Thomson TM (2006) The RING finger protein RNF8 recruits UBC13 for lysine 63-based self polyubiquitylation. *Journal of cellular biochemistry* 97 (3):572-582. doi:10.1002/jcb.20587

- Poulsen N, Kroger N (2005) A new molecular tool for transgenic diatoms: control of mRNA and protein biosynthesis by an inducible promoter-terminator cassette. *The FEBS journal* 272 (13):3413-3423. doi:10.1111/j.1742-4658.2005.04760.x
- Rayko E, Maumus F, Maheswari U, Jabbari K, Bowler C (2010) Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. *New Phytol* 188 (1):52-66. doi:10.1111/j.1469-8137.2010.03371.x
- Riano-Pachon DM, Correa LG, Trejos-Espinosa R, Mueller-Roeber B (2008) Green transcription factors: a chlamydomonas overview. *Genetics* 179 (1):31-39. doi:10.1534/genetics.107.086090
- Shin J, Heidrich K, Sanchez-Villarreal A, Parker JE, Davis SJ (2012) TIME FOR COFFEE represses accumulation of the MYC2 transcription factor to provide time-of-day regulation of jasmonate signaling in Arabidopsis. *The Plant cell* 24 (6):2470-2482. doi:10.1105/tpc.111.095430
- Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falcatore A, Bowler C (2007) Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. *Gene* 406 (1-2):23-35. doi:10.1016/j.gene.2007.05.022
- Stone SL, Hauksdottir H, Troy A, Herschleb J, Kraft E, Callis J (2005) Functional analysis of the RING-type ubiquitin ligase family of Arabidopsis. *Plant Physiol* 137 (1):13-30. doi:10.1104/pp.104.052423
- Tozzi S, Schofield O, Falkowski P (2004) Historical climate change and ocean turbulence as selective agents for two key phytoplankton functional groups. *Marine Ecology Progress Series* 274:123-132
- Uden G, Nilkens S, Singenstreu M (2013) Bacterial sensor kinases using Fe-S cluster binding PAS or GAF domains for O₂ sensing. *Dalton transactions (Cambridge, England : 2003)* 42 (9):3082-3087. doi:10.1039/c2dt32089d
- Valenzuela J, Mazurie A, Carlson RP, Gerlach R, Cooksey KE, Peyton BM, Fields MW (2012) Potential role of multiple carbon fixation pathways during lipid accumulation in *Phaeodactylum tricornutum*. *Biotechnol Biofuels* 5 (1):40
- Vandepoele K, Quimbaya M, Casneuf T, De Veylder L, Van de Peer Y (2009) Unraveling Transcriptional Control in Arabidopsis Using cis-Regulatory Elements and Coexpression Networks. *Plant Physiology* 150 (2):535-546. doi:10.1104/pp.109.136028
- Vandepoele K, Van Bel M, Richard G, Van Landeghem S, Verhelst B, Moreau H, Van de Peer Y, Grimsley N, Piganeau G (2013) pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environmental microbiology* 15 (8):2147-2153. doi:10.1111/1462-2920.12174
- Yang Z-K, Niu Y-F, Ma Y-H, Xue J, Zhang M-H, Yang W-D, Liu J-S, Lu S-H, Guan Y, Li H-Y (2013) Molecular and cellular mechanisms of neutral lipid accumulation in diatom following nitrogen deprivation. *Biotechnol Biofuels* 6 (67):1-67
- Yang ZK, Zheng JW, Niu YF, Yang WD, Liu JS, Li HY (2014) Systems-level analysis of the metabolic responses of the diatom *Phaeodactylum tricornutum* to phosphorus stress. *Environmental microbiology* 16 (6):1793-1807
- Yongmanitchai W, Ward OP (1991) GROWTH OF AND OMEGA-3-FATTY-ACID PRODUCTION BY PHAEODACTYLUM-TRICORNUTUM UNDER DIFFERENT CULTURE CONDITIONS. *Applied and Environmental Microbiology* 57 (2):419-425
- Yoshihara S, Shimada T, Matsuoka D, Zikihara K, Kohchi T, Tokutomi S (2006) Reconstitution of blue-green reversible photoconversion of a cyanobacterial photoreceptor, PixJ1, in phycocyanobilin-producing *Escherichia coli*. *Biochemistry* 45 (11):3775-3784. doi:10.1021/bi051983l

Zheng Y, Quinn AH, Sriram G (2013) Experimental evidence and isotopomer analysis of mixotrophic glucose metabolism in the marine diatom *Phaeodactylum tricornutum*. *Microbial cell factories* 12:109. doi:10.1186/1475-2859-12-109

Chapter 5:

Coordinated transcript expression of the TCA cycle enzymes during nitrogen starvation is guided by the transcription factor bZIP14

Manuscript in preparation for publication

Matthijs M, Fabris M, Obata T, Carbonelle S, Vanden Bossche R, Foubert I, Fernie A, Vyverman W, Goossens A

Author contributions:

MM wrote the manuscript, analyzed the data, designed the experiments and performed the majority

ABSTRACT

Diatoms are non-green algae that are among the most abundant micro-algae in aquatic environments. Their growth and photosynthesis is often limited by the lack of macro-nutrients such as nitrogen. In this study the pennate diatom *Phaeodactylum tricornutum* was used to understand the changes on the transcriptomic and metabolomic level during the switch from exponential growth to the halt of cell division as triggered by the lack of nutrients or light. While some metabolic shifts were already recognized in previous studies, the genes controlling this switch in diatoms were completely unknown. In order to find these controlling genes, RNA sequencing was undertaken on cells in four different conditions: nocodazole treatment, phosphate starvation, nitrogen starvation and reillumination after a period of darkness. RNA-sequencing was used to profile the transcriptome of ten timepoints with a focus on nitrogen starvation. Particularly striking in our dataset was the co-ordinated upregulation of the Krebs or Tricarboxylic Acid Cycle (TCA cycle). Using functional studies in *P. tricornutum*, we were able to show that the transcription factor bZIP14 influences the TCA transcriptional behaviour.

INTRODUCTION

One fifth of all oxygen produced is the result of diatom photosynthesis. They are present in fresh water, oceans and even colonize arctic ice. Despite their abundance and beauty, molecular research has only started in the last decade. Diatoms belong to the kingdom of the heterokontae together with multicellular kelps and water moulds such as the Irish potato blight. Even though diatoms are eukaryotic photosynthetic organisms, they have an evolutionary history that is distinct from land plants and green algae. It has been widely accepted that they originated from a secondary endosymbiosis event in which a bikont heterotroph engulfed a red algae. The genome was further enriched by horizontal gene transfer resulting in the uptake of many bacterial and 'green' genes. It has been hypothesized that these genes represent the footprint of an earlier endosymbiotic relationship with a green algae although this theory has come under doubt (Deschamps and Moreira 2012). As a result the diatom genome is a mix of several features that were thought to be exclusive to other kingdoms such as a complete urea-cycle (Allen et al. 2011).

Diatoms live in a highly variable environment and rapidly need to adapt to changes in nutrients. Two of the most important nutrients are nitrogen and phosphate, and their low availability in most environments limits diatom growth. As nutrient starvation stress is frequently encountered by diatoms, they must have evolved coping mechanisms. The physiological response to nitrogen or phosphate starvation has been studied for decades, but genetic information was scarce until the arrival of genome projects, proteomics and RNA-sequencing. These tools have only recently become available to the community but already several studies have been devoted to nitrogen or phosphate deficiency in diatoms but never with this many replicates or timepoints (Valenzuela et al. 2012; Yang et al. 2014).

Understanding how an organism coordinates its response on the molecular level to survive the lack of an essential nutrient is key to understand the metabolic capabilities and spatio-temporal distribution of diatoms. In addition to their ecological significance, nutrient responses have been investigated for their ability to induce lipid production in diatoms and other alga (Yongmanitchai and Ward 1991). Besides lipids diatoms are also able to store carbon in the form of the polysaccharide chrysolaminarin. Apprehension of the control mechanisms that alter metabolic flux to storage product synthesis will allow the rational improvement of lipid biosynthesis to attain commercially relevant levels.

A catalog of all predicted transcription factors in *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* has been published (Rayko et al. 2010). Identifying the function of a transcription factor using orthology alone is nearly impossible in heterokonts because of the evolutionary distance with more well established model organisms. Since this list was based on orthology with other organisms, it cannot be excluded that there are novel types of TF's present in heterokont genomes. Transcriptional control mechanisms are therefore poorly understood in diatoms. Although recent publications have identified the function of three diatom transcription factors involved in blue light signaling and CO₂ assimilation (Huysman et al. 2013; Ohno et al. 2012), over 90% of transcription factors still have no function assigned to them.

Previous studies have mainly focused on diatoms adapted to long-term nitrogen starvation. In this study we focus on the first twenty hours after the removal of nitrogen and contrast our finding with three other stresses: phosphate starvation, nocodazole treatment and cells placed in darkness. The transcriptomic changes during these stresses were investigated through RNA-seq. This data coupled to a profiling of primary metabolite changes resulted in the identification of the pathways under transcriptional control. This knowledge was then applied to discover transcription factors controlling these changes.

RESULTS

SELECTION OF CONDITIONS

The main goal was to identify the transcription factors steering metabolic changes stress. To this end, four conditions that negatively impact diatom cell growth and affect lipid biosynthesis were selected. It was decided to focus on the transition point from growth to the halt of cell division, rather than on cells adapted to the conditions. Therefore samples for or transcriptome analysis were taken during the first 36 hours after medium replacement.

As nitrogen starvation is one of the best studied triggers for lipid biosynthesis, it was chosen as the main component of our dataset with three out of ten RNA-seq samples. Phosphorous starvation was also included because, like nitrogen, it is also a macronutrient. However it should be noted that its removal halts cell division much slower than nitrogen starvation (Yang et al. 2014). To contrast with these two nutrient starvation conditions, cells placed in the dark were included since the lack of light also results in a G1/S arrest but does not lead to lipid production. This condition has allowed us to remove genes from our candidate set that are not directly involved in metabolic changes during nutrient starvation but are the result of a halting cell division. Although these cells were briefly exposed to the light while harvesting, which triggered exponential growth, neither dark adapted cells nor reilluminated cells are expected to accumulate lipids, and this sample still serves as a useful counterpoint to the nutrient starvation points.

As most stresses above halt the cell cycle at G1/S, we expected that a sizeable fraction of affected genes would be involved in the arrest of cell division rather than metabolic adaptation. Therefore we included the microtubule polymerization inhibitor nocodazole. During a normal mitotic phase the chromosomes are aligned in the middle of the cell and subsequently divided over the two cell halves by microtubule action. Nocodazole treatment prevents this process and halts cell division during M-phase. The disadvantage of this drug is that other microtubule dependent processes are also affected and cells do not survive long term treatment (more than 48 hours). The ability of nocodazole to induce G2/M cell cycle arrest in *P. tricornutum* had been previously shown (Huysman et al. 2010).

In total ten samples of the four different stresses were assayed together with the relevant control sample. These provide a detailed insight in the transcriptional reprogramming of the cell under cell division halting conditions.

The generation time of *P. tricornutum* is approximately 20 hours which means that cells have sufficient time to complete one round of cell division during the timecourse from what was observed (fig 1) (Mann & Myers, 1968). The cells placed in medium without nitrogen still undergo approximately one more division. The cells illuminated after dark show no increase in OD while the flow cytograms (discussed further below) still show progression of the G2/M phase cells to G1/S arrest. This apparent contradiction is the result of measuring cells by OD only since it has been shown that cells dividing in the dark have a smaller volume (Chauton et al. 2013). The toxic effect of nocodazole is increasing visible from hour 20 onwards. This corresponds to the flow histograms as almost the entire population is in G2/M arrest.

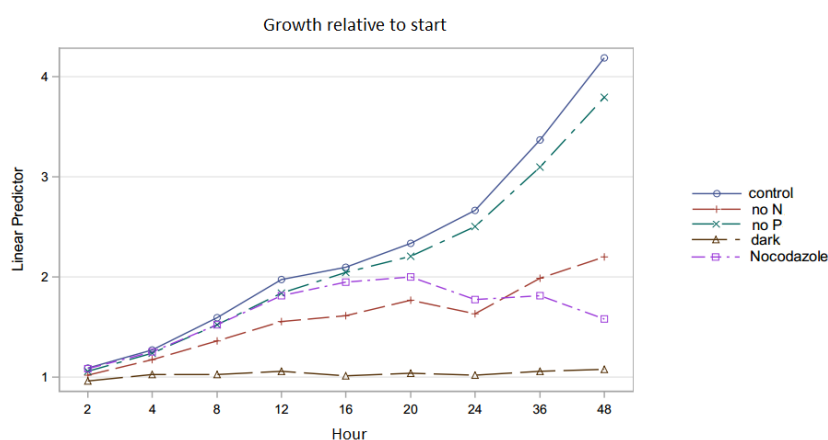


Figure1: Relative growth profiles of the conditions, error bars omitted for clarity, growth was measure by OD405nm and rescaled to timepoint 0 to show the general trend of growth

The effects of phosphate starvation are very mild compared to the other stress responses. This was expected as the amount of phosphate required for diatom growth is 1/16 of that of nitrogen as determined by the classic Redfield ratio and cells often contain substantial intracellular stores of this nutrient (Falkowski et al. 2004). All of these observations fit with what had been observed previously.

All four investigated stresses halt the cell cycle in order in order to which degree these are correlated to the transcriptome changes we decided to perform flow cytometry on DAPI stained nuclei (fig. 2). The intensity of staining correlates with the DNA content of the cell and shows whether the cell has only one full set of chromosomes (G1 or G1/S arrest), is in the process of duplicating its DNA (S phase) or is awaiting the completion of cell division (G2 or G2/M arrest). The transition through the different phases of the cell cycle are controlled by the activity of Cyclin Dependent Kinases (CDK's), and their associated cyclins. While the mode of action is exceptionally well preserved in eukaryotes, diatoms have an expanded family of these cyclins. It has been suggested that some of these are able to integrate nutrient availability. This would make sense as cells need to know beforehand if they have enough resources, both in energy and cellular components, to successfully complete a round of cell division (Huysman et al. 2013). In *S. cerevisiae* there is a metabolic checkpoint in the cell cycle that halts cells in G1/S when these requirements,

e.g. the lack of amino acids, are not met. Nitrogen starvation, phosphate starvation and darkness are well known to halt the cell at the G1/S transition phase. Nocodazole on the other hand prevents microtubule polymerization and therefore traps cells in the anaphase as it is unable to separate the chromosomes.

After the transfer to new medium at timepoint zero, all conditions show the same image. Most are in G1 phase with a gradual decrease to the right of cells that are duplicating their genome. Only a small proportion is in the G2/M phase.

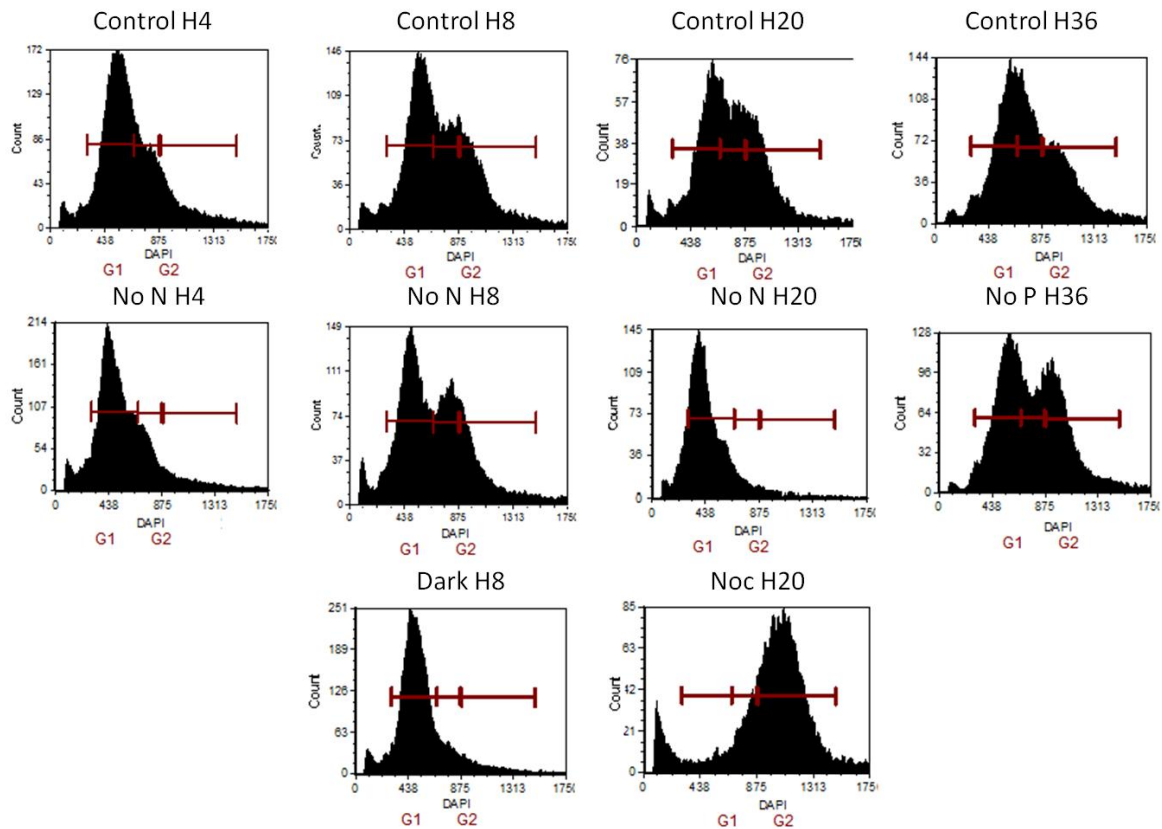


Figure 2: Representative flow histograms of the sequenced timepoints. 2C and 4C DNA content are indicated by G1 and G2. The gradual procession from exponential growth towards cell cycle arrest is visible for all treated samples except Phosphate starved

There is little difference after four hours of incubation in the different media but a gap is appearing between the cells containing 2C and 4C amounts of DNA's. In other words: no new cells are transitioning in S phases. After this timepoint the 4C peak almost completely disappears and the 2C peak becomes sharper and higher, the dark adapted cells are arresting in the G1/S phase. After 8 hours all conditions except the aforementioned dark conditioned the 4C peak gradually increases. This synchrony was unexpected as cells were kept in constant light and in exponential growth. It is likely that the transfer to new medium or the preceding centrifugation step have still induced some measure of synchrony.

From twelve hours onwards the effect of the different conditions has become clear in all samples with the exception of phosphate limitation. While control cells look similar to those at timepoint 8, the 2C cells in the nocodazole treated flask have almost disappeared. The 4C peak is clearly dominant and much broader

than in any other condition, indicating that G2/M arrest has taken hold. Nitrogen starved cells are also beginning to halt their cell cycle at the G1/S point as illustrated by a small peak at C4 and an increasingly sharp peak at 2C. The residual 4C peak has disappeared by 16 hours onward and while this peak still increases in height and sharpness during the next 12 hours it appears that most of the cells are already entering G1/S arrest.

In conclusion we can state that with the exception of phosphate starvation, all investigated conditions halt the cell cycle within the timeframe of the experiment and that the timepoints at which RNA-sequencing was performed, coincide with these arrests.

CARBON STORES INCREASE DURING NITROGEN STARVATION

While the accumulation of lipids during N starvation is well established, it was unknown whether lipid levels increased during the short timespan of the RNA-seq timecourse. To that end lipids were quantified using gas chromatography after methyl esterification, trends are plotted in figure 3. Lipid levels increases in nitrogen deprived cells were already visible two hours after medium change. This rapid onset of lipid accumulation was unexpected. Cells placed in darkness consumed the limited amount of lipids they had stored during exponential phase. The phosphate starvation and nocodazole treatment did not significantly alter compared to the control. Control cells also started to accumulate more lipids at 16 hours, it could be that some of the nutrients were becoming limiting even though cell division continued for another hour as diatoms tend to accumulate more lipids than they immediately use for growth.

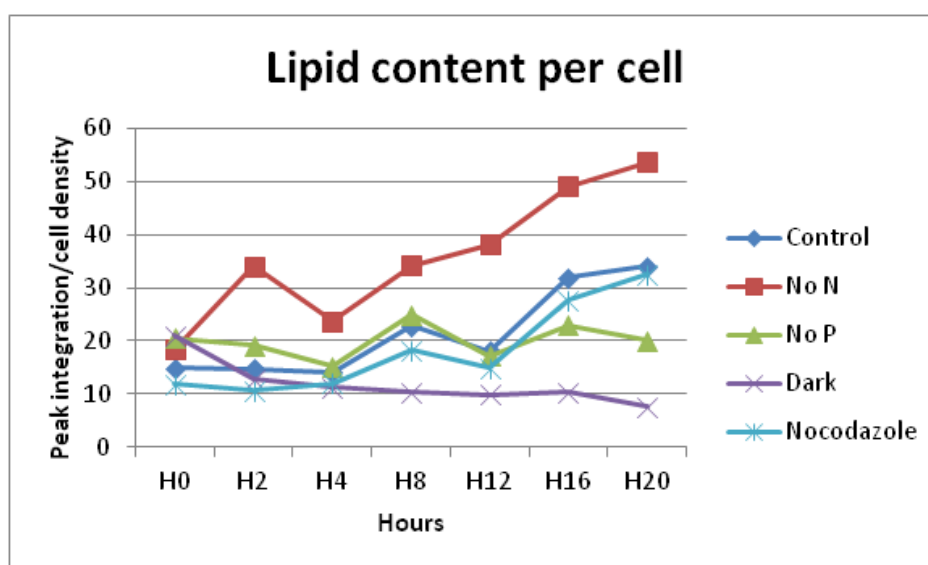


Figure 3: Lipid content per cell, error bars omitted for clarity of trend, numbers represent fatty acid area under peak divided by the OD at that timepoint to account for differences in cell density.

Soluble carbohydrate levels were measured in parallel to the lipid measurements. Chrysolaminarin was extracted from cells using a mild acid solution and quantified according to the method of Granum and Myklestad(Granum and Myklestad 2002). The trends observed mirror those seen for lipid accumulation

and it appears storage of lipids and carbohydrates occurs concurrently. Again nocodazole and phosphate treatment behave similarly to the control, and dark treated cells rapidly consume their available carbohydrate reserves. In nitrogen starved cells, extractable carbohydrates increase simultaneously with the accumulation of lipids but peaks after eight hours in nitrogen starved cells (figure 4). This is in contradiction to previous studies where the reverse was seen (Valenzuela, 2012). One possible explanation is that the method used in this study differs from the one used in most other studies and might extract more chrysolaminarin or extract additional types of sugars (Chaplin and Kennedy 1994). A significant carbon flux towards carbohydrates under nitrogen limiting conditions is supported by the recent finding that knocking out the committing step of polysaccharide biosynthesis increases lipid accumulation (Daboussi et al. 2014). This increase is also visible under nitrogen deprivation; it therefore seems likely that cells store a significant amount of carbon as carbohydrates during nitrogen starvation. An alternative explanation is that the carbohydrates measured are not purely chrysolaminarin but some other form of carbohydrate. A paper by the same authors found that a substantial amount of photosynthetic production was excreted as sugars from the cell during stationary phase (Myklestad et al. 1989). The higher carbohydrate measurements made during nitrogen starvation could be if these sugars are extracted together with the internally stored carbohydrates.

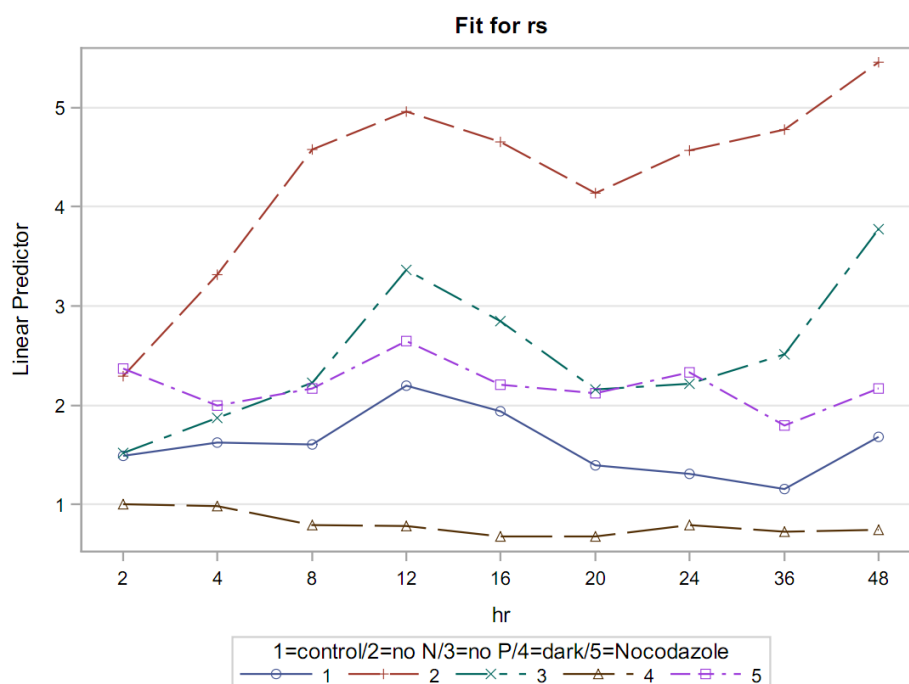


Figure 4: Relative soluble carbohydrate content compared to the control

RNA-SEQUENCING RESULTS

DARKNESS RESULTS IN EXTENSIVE TRANSCRIPTOME REPROGRAMMING OF A MYRIAD OF PROCESSES

It was our intention to stop lipid accumulation of diatom cells through stopping photosynthesis by depriving them of light. Cells were placed in complete darkness for eight hours but filtering, although performed in under 5 minutes, exposed cells to light for a short period of time. The effect of the inadvertent light exposure is likely enough to warrant the labelling of these samples as re-illuminated rather than pure dark samples. One of the best indicators for this is the upregulation of dsCyclin2 which has been reported to rapidly respond to light and this transcript is highly induced in this timepoint(Huysman et al. 2013).

Placing cells for eight hours in the dark after they have been kept in continuous light results in a massive transcriptome shift. The number of genes upregulated at least twofold was 1415, while 2098 were at least twofold downregulated. This is the most disruptive treatment in our dataset. In total 120 transcription factors, over half of all transcription factors, have their expression maximum at this timepoint. The most dramatic change is for Hsf2.2c which is over a hundredfold induced. Hsf3.2e, bHLH2 and Hox1 are more than tenfold induced.

Several transcripts related to photosynthetic transcripts such as the light harvesting protein LHCR7 or steps in the chlorophyll biosynthesis such as geranyl diphosphate synthase (Phatr15180) are highly upregulated but in general photosynthesis is not conspicuously present in the list of positively responding genes. Genes related to chloroplast control however are prominently present in the upregulated set. Sigma factors are important regulators of chloroplast RNA transcription and sigma70.5 is especially responsive to light. Several aureochrome transcription factors that have been shown to be light responsive have been upregulated, namely Aureo1, Aureo3 and Aureo4(Huysman et al. 2013). The transcription factor bHLH2 is also strongly upregulated (Phatr54435) and contains a PAS domain, these domains have been reported to sense light(Brudler et al. 2006).

NOCODAZOLE, A MICROTUBULE POLYMERIZATION INHIBITOR HALTS THE CELL CYCLE AND INDUCES A PROTEIN STRESS RESPONSE

After twenty hours of incubation with nocodazole 965 genes were more than twofold upregulated while 398 genes were more than twofold downregulated. The effect of treatment on lipid biosynthesis was less than expected and the main value of the generated data in this timepoint was as a contrast to other conditions in co-expression analysis (fig. 3). Nocodazole destabilizes the microtubule based cytoskeleton which disturbs intracellular trafficking and the mitotic spindle(Jordan and Wilson 1998). The effect on intracellular trafficking can be seen by the upregulation of ras-gtpases (Phatr43251&Phatr22713). Several

chaperones are upregulated during nocodazole treatment (Phatr36981, Phatr55230, Phatr41417) hinting that the unfolded protein response is activated. The link with nocodazole action is unclear although it has been reported that there are tight links between microtubules and the ER and a similar overrepresentation of HSP's is seen in yeast (Terasaki et al. 1986; Hillenmeyer et al. 2008).

The silicon transporters (Phatr55090 & Phatr23423) are among the highest fold induced genes but in general nutrient uptake appears to be strongly repressed. Genes such as an alkaline phosphatase (Phatr47869), a low iron induced protein (Phatr55031) and nitrite reductase (Phatr12902) feature prominently in the list of downregulated genes.

The most obvious effect of nocodazole treatment is a G2/M cell cycle arrest. The most G2/M specific transcription factors are CCHH9 and Myb1R2, expression peaks in the nocodazole treated cells and is low in the G1/S arrested cells of darkness. Flow cytograms of nocodazole treated cells at later timepoints begin to show increasing amounts of debris, this indicates that cells are close to death. Upregulation is seen of strikingly one of the most upregulated genes is a gag-pol protein (Phatr39576), usually associated with retrotransposon multiplication. This behavior has been seen previously as transposons often become more active during stress conditions (Maumus et al. 2009).

THE EFFECTS OF PHOSPHATE STARVATION ARE FEW AFTER 36 HOURS

Phosphate deprivation had the mildest effects of all stresses, 36 hours after medium change only 681 genes show twofold differential expression when compared to the control. Cells kept in phosphate free medium after 72 hours did stop their growth while controls continued growing (Data not shown). Likely the cell still has a large internal pool of stored phosphate and large scale changes probably occur only later as phosphate reserves are depleted. As with other nutrient stresses, one of the main effects is an increase in the uptake capacity for the limiting nutrient and the use of alternative phosphorous sources. The phosphate transporter Phatr39515, two sodium/phosphate antiporters (Phatr47239 & Phatr47666) increases in expression and do several alkaline phosphatases (Phatr39432, Phatr49678 and Phatr45757). As nucleotides form a large pool of phosphate within the cell it is not surprising that several nucleosides phosphatases are upregulated (Phatr49679 & Phatr43694). The genes mentioned above form a co-expression cluster that contains no annotated transcription factors. The only possible control element in this cluster is a bacterial type histidine sensor kinase (Phatr46628). Two transmembrane proteins (Phatr47434 & Phatr19586) contain a SPX domain which is prevalent among proteins involved in phosphate homeostasis (Secco et al. 2012).

Strongly downregulated are several genes involved in the creatine cycle, such as creatine kinase (Phatr11733) and two creatine transporters (Phatr47653 & Phatr47656). This cycle is able to rapidly regenerate ATP from ADP transferring a phosphate group from phosphocreatine. Interestingly several genes that are severely downregulated during phosphate starvation are highly upregulated during N

starvation including Phatr55010, which is one of the most responsive genes to N starvation and the aforementioned creatine transporters.

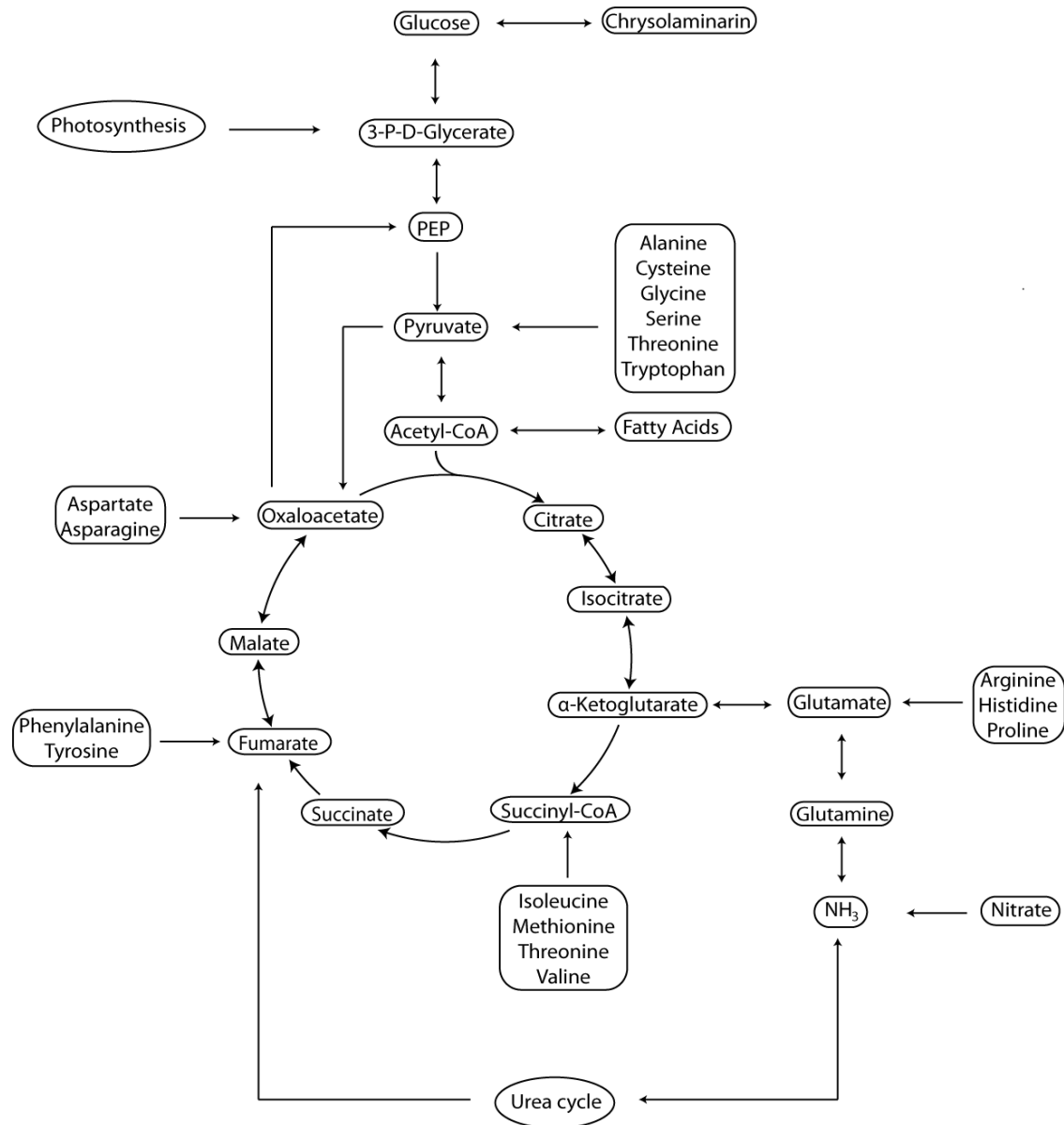


Figure 1: Overview of amino acid degradation and central carbon metabolism. Adapted from Obata et al. 2012 and Hockin et al. 2010. Not all reaction products are shown and some arrows represent more than one reaction

NITROGEN STARVATION

The most downregulated process in nitrogen starved cells compared to the control condition is photosynthesis. With a few exceptions, almost all Fucoxanthin Binding Protein genes are downregulated as are a host of genes involved in pigment biosynthesis. Carbon metabolism is strongly affected, some of the relevant pathways are represented in figure 5. Gluconeogenesis is strongly downregulated, this is most

apparent when looking at the metabolism of three carbon compounds and oxaloacetate. Pyruvate carboxylase (Phatr49339) which also produces alpha-ketoglutarate is strongly downregulated together with other genes that are part of gluconeogenesis such as phosphoenolpyruvate carboxykinase (Phatr27976). One isoform of fructose biphosphate aldolase clusters together with these genes, which implies this isoform is involved in gluconeogenesis. A reduction in sugar biosynthesis is expected as there is no excess photosynthate that needs to be stored. At the same time transcripts of glycolytic enzymes are also downregulated e.g. pyruvate kinase (Phatr45997 & Phatr46001) as are genes that are part of both pathways such as phosphoglycerate mutase (Phatr41063) and triosephosphate isomerase (Phatr32747). In this timepoint the citric acid cycle is strongly downregulated e.g. the transcripts for fumarase (Phatr19708) and Isocitrate dehydrogenase (Phatr14762).

Assimilation of nutrients on the other hand is fully active as both nitrate and ferric reductase (Phatr54983 resp. Phatr54982) are highly expressed, the latter is over thirty fold induced compared to the control. Tentatively the short burst of light makes the cell gear up its metabolism to assimilate nutrients so it can to resume cell division. However the short period of light to which the cells were exposed, was not able to restore normal cell function completely. Ribosomal biosynthesis is still strongly downregulated compared to the controls (e.g. Phatr10196 & Phatr11032). Proteins synthesis apparently needs more time to resume. DNA replication has also not recovered in this timespan as polymerase epsilon (Phatr52678) and several condensins (Phatr30352, Phatr25506 & Phatr50280) are expressed at very low levels.

Surprisingly, in our RNAseq dataset, lipid metabolism was not grouped in a co-expression cluster, despite a clear increase in fatty acid content. In fact several lipid biosynthesis genes were over twofold downregulated, such as a diacylglycerol transferase (Phatr49462), an elongase (Phatr20508), a desaturase (Phatr46275) and a monoacyl transferase. It is possible that the transcriptomic shift to lipid biosynthesis had already peaked before four hours, and this would explain why no major changes in enzyme transcripts involved in the process was seen. Similar findings were made for the brown unicellular microalgae *Nannochloropsis gaditana* (Carpinelli et al. 2014). Other studies have reported increases in transcript levels of individual lipid biosynthesis genes. This contrast sharply with observations in green algae where clear clusters of lipid anabolic genes were reported (Guarnieri et al. 2011; Rismani-Yazdi et al. 2012).

THE TCA CYCLE PLAYS A CENTRAL ROLE IN CARBON REALLOCATION DURING N STARVATION

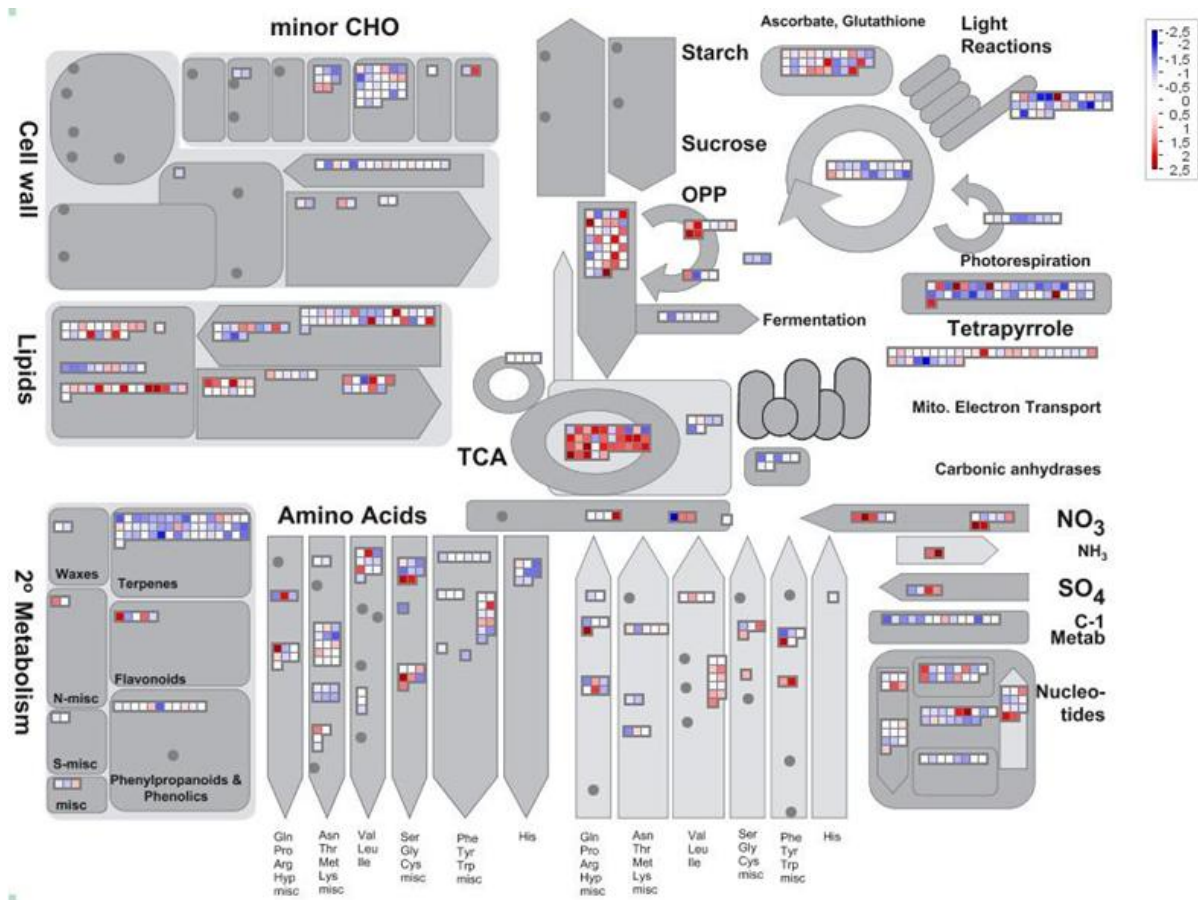


Figure 6: Overview of primary metabolism as visualized by the Mapman program. Points represent the log fold change of Nitrogen starved cells 20 hours after medium change versus control cells at the same timepoint. Adapted from Obata et al. 2012 and Hockin et al.

Because lipid biosynthesis had no apparent transcriptional regulation, we decided to investigate the general primary metabolism of *P. tricornutum* under nitrogen starvation. Using the metabolic pathway visualization tool Mapman it became apparent the tricarboxylic acid cycle (TCA cycle) was the pathway that showed the strongest differential regulation at the RNA level (fig. 6). Furthermore this upregulation appeared to be coordinated (fig. 7).

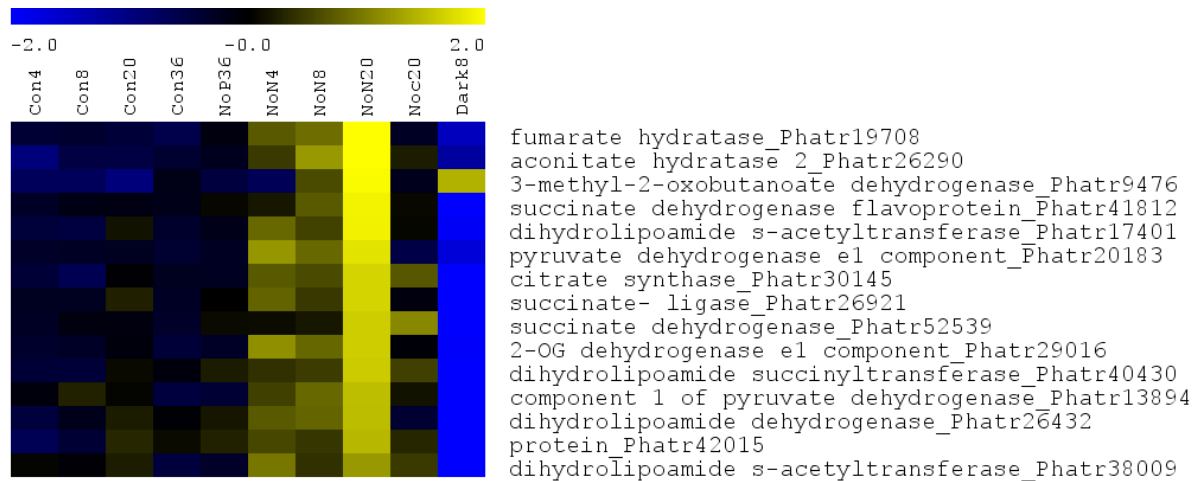


Figure 7: Expression profile of TCA cycle genes upregulated during nitrogen starvation, normalized Fragments Per Kilobase Of Exon Per Million Fragments Mapped (FPKM) values

The upregulation under nitrogen starvation however does not appear to occur in each photosynthetic organism since green algae do not show this pattern (Schmollinger et al. 2014). Nonetheless, It does appear widespread in diatoms since the upregulation of the TCA cycle was also observed in *Thalassiosira pseudonana* on the protein level during nitrogen starvation (Hockin et al. 2012). In cyanobacteria many TCA cycle enzymes are upregulated during nitrogen deprivation, but it is thought that it has a purely anabolic role in capturing recycled ammonia (Steinhauser et al. 2012).

Because of its central role in metabolism, the TCA cycle can impact many processes (fig. 5). Several reactions are reversible and it has both anabolic as catabolic roles. Many intermediates form the carbon skeletons for amino acid synthesis. The most likely reason for the upregulation would be the recycling of amino acids in order to get nitrogen to those processes where they are most needed. Degradation of amino acids, like the TCA cycle, mainly occurs in the mitochondria and ultimately all amino acid carbon skeletons can be used by the TCA cycle. Several amino acids directly form TCA intermediates after deamination (e.g. glutamate and aspartate), while deamination of other amino acids such as alanine result in pyruvate which can be converted to acetyl-CoA, or in the case of some branched chain amino acids and methionine to propionyl-CoA.

Besides the enzymes of the TCA cycle itself, the acetyl-CoA production from pyruvate by the pyruvate dehydrogenase complex (DHLP) was also upregulated. Part of this might come from the degradation of sugars or the degradation of some amino acids such as serine and alanine. Through the action of the methylmalonyl pathway propionyl-CoA is converted to succinyl-CoA, which can be fed into the TCA cycle. Of the four enzymes of this pathway three gradually increase in expression when cell growth is halting (Phatr51830, Phatr51245 & 45886), indicating that at least some of the Acetyl-CoA is not used for lipid biosynthesis but enters the TCA cycle. Transcripts of anaplerotic reactions of the TCA cycle such as PEPC2 and glutamate dehydrogenase (Phatr13951) generally increased.

Some prokaryotes use the TCA cycle as a carbon fixation mechanism, running the reaction in the other way around (Schauder et al. 1987). While this is unlikely to happen in diatoms, there are links between the TCA cycle and anabolism through enzymes such as citrate lyase which converts citrate into oxaloacetate and acetyl-CoA, which feed into gluconeogenesis and fatty acid synthesis respectively. However the enzyme catalyzing this function (Phatr54477) was downregulated upon nitrogen starvation.

METABOLITE PROFILING

In order to get a better understanding of the metabolism during nitrogen starvation, a metabolome analysis was performed in the lab of prof. Alisdair Fernie. Cells were incubated with $^{13}\text{CO}_2$ twenty hours after the removal of nitrogen and in exponentially growing cells as control. Large differences in incorporation were seen in many of the 38 detected metabolites.

The levels of most amino acids were lower in nitrogen starved cells compared to the control. However, the branched chain amino acids showed a less pronounced decline in relative abundance, even though the degradation of these metabolites has recently been linked to lipid synthesis (Ge et al. 2014). This observation has also been made for *T. pseudonana* (Hockin et al. 2012). Other notable exceptions are proline and arginine, which are linked to the urea cycle. Metabolite changes in the urea cycle are correlated with changes in tricarboxylic acid (TCA) cycle metabolites. This suggests that the two cycles are tightly coupled by the action of argininosuccinate lyase which produces fumarate and arginine. However no coordination of these pathways is seen on the transcriptional level. Another exception was tryptophan where levels were actually higher in the N depleted cells and novel synthesis appeared similar as in control cells. Incorporation of novel carbon was decreased in all other amino acids but this decline was less pronounced in arginine and proline, two amino acids which can be linked to the urea cycle. Urea production and levels were not significantly different in both conditions and ornithine was almost undetectable (fig. 8).

Another exception in amino acid metabolism is the level of tryptophan. The synthesis of this amino acid is distinct from all other metabolites. A possible explanation is the use of this compound as a precursor of signaling compounds as it has been shown that diatom growth increases upon addition of indole ring containing compounds such as 3-Indole acetic acid (Amin et al. 2012).

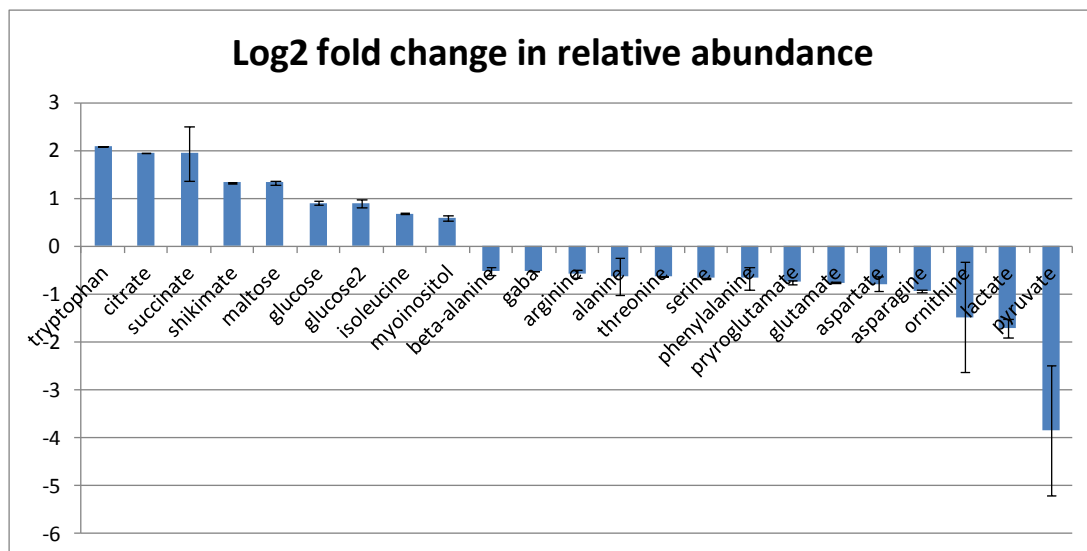


Figure 8: The log2 fold change in metabolite abundance for nitrogen starved vs. control cells. As measured 20 minutes after incubation with ^{13}C labeled CO_2 . Out of the 38 detectable metabolites only those are represented for which there was a significant change.

Abundance and ^{13}C enrichment of carbohydrates was practically unchanged in nitrogen starved cells despite the decrease in photosynthetic efficiency. This is not unexpected as the largest increase in carbohydrates happens during the earlier phases of nitrogen starvation as shown in figure 4.

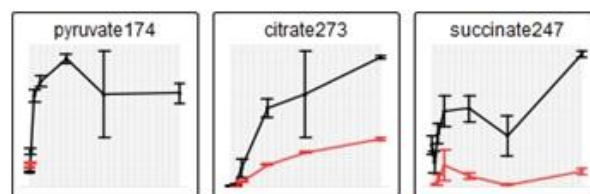


Figure 9: ^{13}C incorporation in three selected metabolites, the red line indicates nitrogen starved cells, the black line control cells. the outermost timepoint is 120 minutes after incubation with $^{13}\text{CO}_2$. These graphs show that relatively little new carbon is fixed into these TCA cycle intermediates compared to the control

The pool of the two measured TCA intermediates citrate and succinate was higher in nitrogen starved cells but incorporation of novel carbon into this metabolites is slower (fig. 9). This was unexpected and hints that the TCA cycle is mainly involved in the degradation or repurposing of existing carbon reserves in the cell.

THE SEARCH FOR TRANSCRIPTIONAL REGULATORS OF THE TCA CYCLE

Since the TCA cycle was one of the most upregulated and coordinated processes in the generated dataset, we attempted to identify potential transcriptional regulators of this pathway. An initial screening with 8 transcription factors upregulated during nitrogen starvation was performed using a tobacco transcriptional activation protoplast test (TEA). The intergenic regions of two of the most induced genes of the TCA

cycle were chosen as baits: Malate dehydrogenase and Citrate Synthase. Unfortunately no transcriptional activation was seen in the assay but the reasons for this failure might be purely technical as explained in the previous chapter.

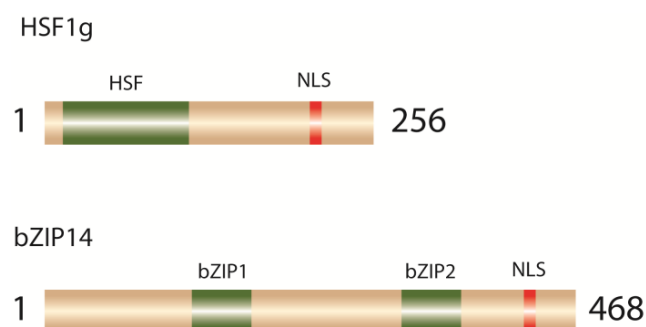


Figure 10: Schematic representation of Hsf1g and bZIP14, NLS: Nuclear localization signal, HSF: Heat shock factor domain, bZIP: Basic leucine zipper domain

Nonetheless we used the previously generated list (Chapter 4) to select two possible candidates. Of the twenty transcription factors with a maximum during nitrogen starvation, Hsf1g and bZIP14 have the highest absolute expression during nitrogen starvation in our dataset and are also upregulated in the dataset of Valenzuela et al. Additionally, they showed specific expression during nitrogen starvation, had a higher than average expression in all conditions and, like the TCA cycle enzymes, gradually increased in expression under nitrogen limitation. A schematic representation of protein domain architecture of the two TF's is shown in figure 10. Their expression pattern during nitrogen starvation was validated independently from the transcriptome dataset by QPCR on samples taken from a new, independent experiment(fig. 11). The co-expression clusters of bZIP14 and Hsf1g are shown in figure 1 and 13 respectively. No clear function seems apparent for either protein when looking at either cluster, nor are there TCA cycle enzymes in the immediate neighborhood.

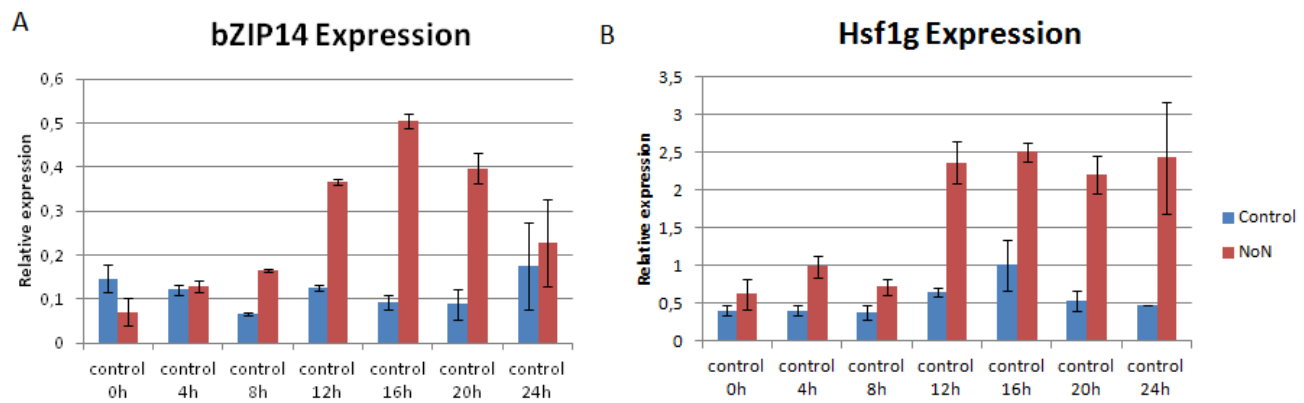


Figure 11: Expression patterns of Hsf1g and bZIP14 during nitrogen starvation in an independent repeat of the RNA-seq timecourse. Bar heights are relative to the two reference genes used. Error bars are the standard error of two biological repeats

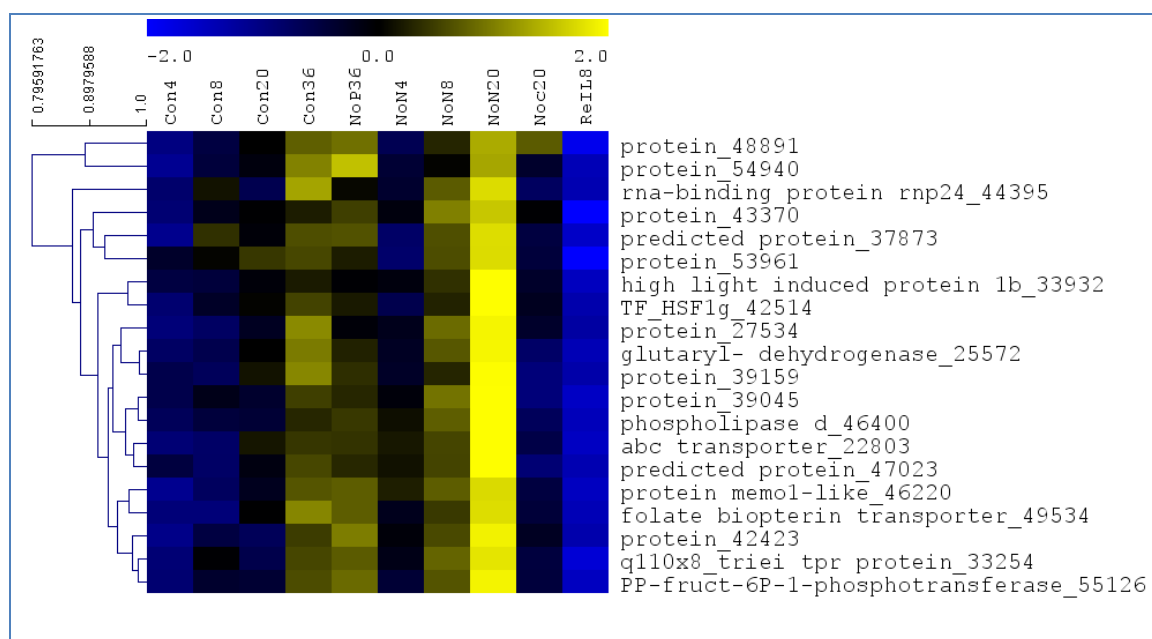


Figure 12: Hsf1g Co-expression cluster, FPKM values after normalisation, CummRbund generated

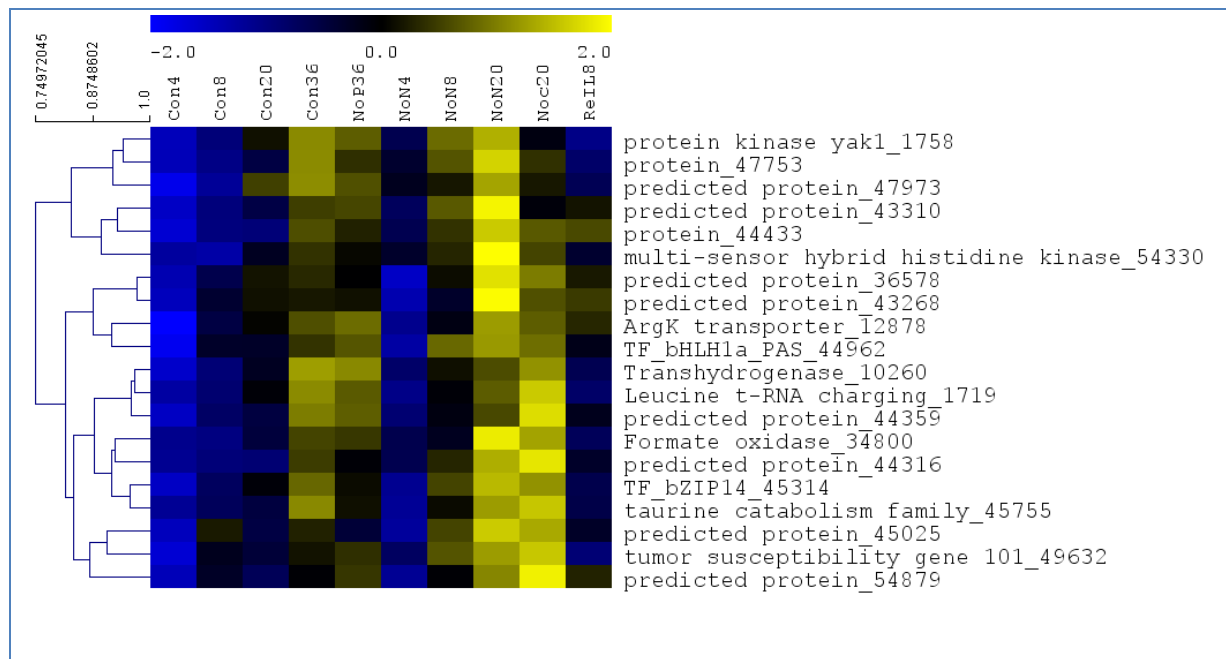


Figure 13: bZIP14 Co-expression cluster, FPKM values after normalisation, CummeRbund generated

BZIP14 IS CONSERVED IN HETEROKONT'S

It had been previously observed that bZIP14 transcription factor (Phatr45314), contains two distinct bZIP DNA binding domains (PfamPF07716 & Pfam PF00170) (Rayko et al. 2010). This unusual domain organization makes it easy to find orthologs of bZIP14 in other species and it was found that the protein is evolutionary conserved in heterokonts. Every sequenced diatoms has a single copy and it is also present in sequenced *Phytophthora sp.* and *Ectocarpus siliculosus*.

In order to find a potential biological function we used the structure based program Backphyre to find structural homologs of bZIP14. Surprisingly, homology was found in *Saccharomyces cerevisiae* with the GCN4 protein of yeast, one of the main transcription factors during nitrogen starvation in yeast.

Clustering bZIP14 with all other transcription factors revealed that two other TF's show a similar expression pattern: HLH1a-PAS and HLH3. HLH1a was present in the list of transcription factors most highly induced under nitrogen starvation, and although variations in HLH3 expression were smaller, they conformed to the same overall pattern. These two transcription factors are present in all sequenced diatoms and both contain a myc-type HLH DNA binding domain (IPR011598) and a PAS type sensor domain. The PAS domain of HLH3 was seen in *F. cylindrus* by interproscan5 but not detected in *P. tricornutum*. However when aligning the two orthologs the region with the predicted domain showed 80% identity, indicating that this function is present in both proteins.

Using the Backphyre program with these two additional proteins indicated structural homology to two yeast transcription factors: RDS2 and RTG3. BLAST searches showed hits with the same protein sequences but a statistically negligible E-value. The aligned portion of RTG3/HLH3 is rather small but

HLH1a spans the entire RDS2 protein. GCN4, RDS2 and RTG3 are all involved in the amino acid deficiency response in yeast and have surprisingly also been reported to alter expression of the TCA cycle (Soontorngun et al. 2007; Liu and Butow 1999). However, since the cohesion of expressions pattern could not be confirmed in a independent QPCR, these genes were dropped for further analysis.

PROTEIN BINDING MICRO-ARRAY

Identifying transcription factors on the basis of homology is straightforward, and for several transcription factors families the core nucleotide motif bound by them is known. However, dedicated experiments are still required to find the exact binding sites. In order to do this for bZIP14 and HSF1g the protein binding micro-array developed in the lab of Roberto Solano was successfully used. Both proteins were expressed in *E. coli* and tested for interaction with all possible combinations of 11 nucleotides. The top three positional weight matrixes of recurrently bound nucleotides are shown in figure 14.

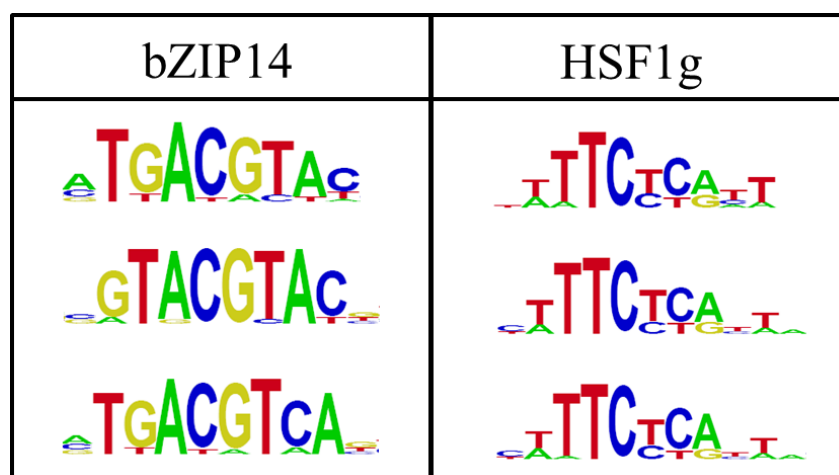


Figure 14: TF binding sites as predicted by the protein binding array represented as positional weight matrixes. Top three motifs shown, obtained by generating the consensus motif from all oligonucleotides bound by the recombinant protein.

The success of this approach was surprising for HSF1g since the array consists only out of combinations of eleven nucleotides which is less than the expected fourteen basepair long consensus sequence for the heat shock factor family. In fact single repeats of the consensus motif were bound more strongly than appropriately space double repeats of the core TTC motif, further strengthening the hypothesis that HSF1g binds as a monomer. Narrowing down potential Hsf1g binding sites in the genome is difficult as only six basepairs of sequence show a high level of conservation. Accordingly, over 60% of all intergenic regions contained the consensus sequence. The bZIP14 protein binds the core ACGT sequence flanked by GT/TG on the 5' end and CA/AC on the 3' prime end. It does not appear to bind a G-box which is typical for plant bZIP transcription factors. In total 635 genes or 6% of all *P. tricornutum* genes contained this motif in the 500bp upstream from their start codon. This included bZIP14 itself and two enzymes related to the TCA cycle: DLDH1(Phatr26432), a subunit of the pyruvate dehydrogenase complex which turns pyruvate into acetyl CoA and Malate Dehydrogenase(Phatr54834). Citrate synthase was also picked

up when allowing one nucleotide divergence from the consensus. Because of a potential link with the TCA cycle it was decided to focus solely on this transcription factor.

The protein binding array was also used for HLH1a and HLH3 but no significant results were obtained. Unfortunately the technique does not seem to work for all transcription factors.

OVEREXPRESSION BZIP14 RESULTS IN THE UPREGULATION OF TCA CYCLE TRANSCRIPTS AND METABOLIC CHANGES

Lines overexpressing HSF1g were generated but although they had a slightly elongated phenotype, A direct link between bZIP14 and the TCA cycle was obtained upon overexpression of bZIP14. Two lines showed robust ectopic expression patterns (fig. 13). Using QPCR it was shown that five transcripts of the TCA cycle are significantly higher expressed compared to control lines. This was also shown to be the case for FBA3, a class II cytosolic enzyme that has previously been reported as upregulated during iron and nitrogen starvation, possibly to mobilize stored carbohydrate reserves(Shrestha et al. 2012). As a negative control we tested the expression of the urease gene which showed robust induction upon nitrogen starvation. Levels of urease transcript were not higher than those seen in empty vector lines (data not shown).

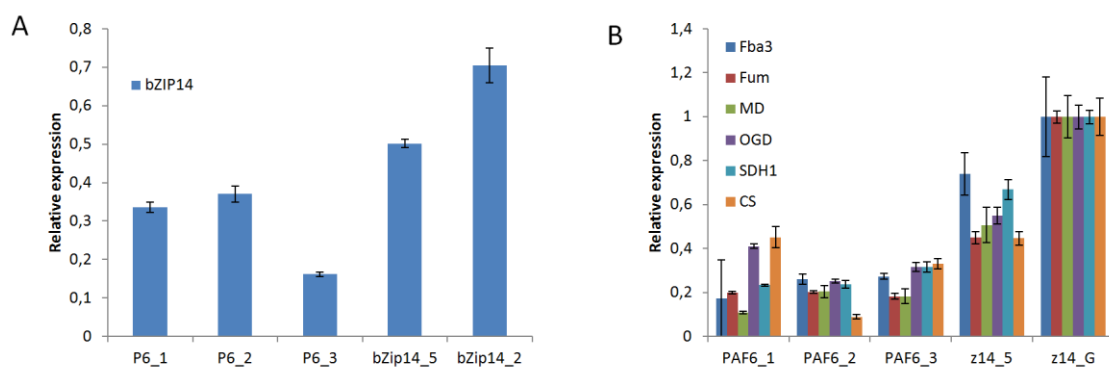


Figure 15: Expression values of TCA cycle enzymes in bZip14 OE lines. FBA3: Phatr29014, MD: Malate Dehydrogenase Phatr42398, FUM: Fumarase Phatr 36139, OGD: 2-OG dehydrogenase complex component 1 Phatr 29016, CS: Citrate Synthase Phatr 30145, SDH1: Succinate Dehydrogenase1 Phatr41812. Paf6 lines are empty vector transformed

Primary metabolites abundance were measured in the overexpression lines (fig. 16). For these analysis HSF1g overexpression lines were also analysed but no clear pattern could be detected. Amino acids in general tended to have higher levels in the bZIP14 overexpression lines. While only slightly higher levels of succinate and citrate were seen compared to the controls, a clear difference was observed with γ -Butyric Acid (GABA) and glutamate. These metabolites are close to the TCA cycle as glutamate can be converted to alpha-ketoglutarate by a single deamination. GABA can be converted into succinate after deamination by 4-amino butyrate transaminase transfers of its amino group to alpha-ketoglutarate making

glutamate in the process. These conversions might be a reaction from the cell to rid itself of excess TCA cycle intermediates. These results continue to hint that bZIP14 is involved in the TCA cycle and nitrogen metabolism.

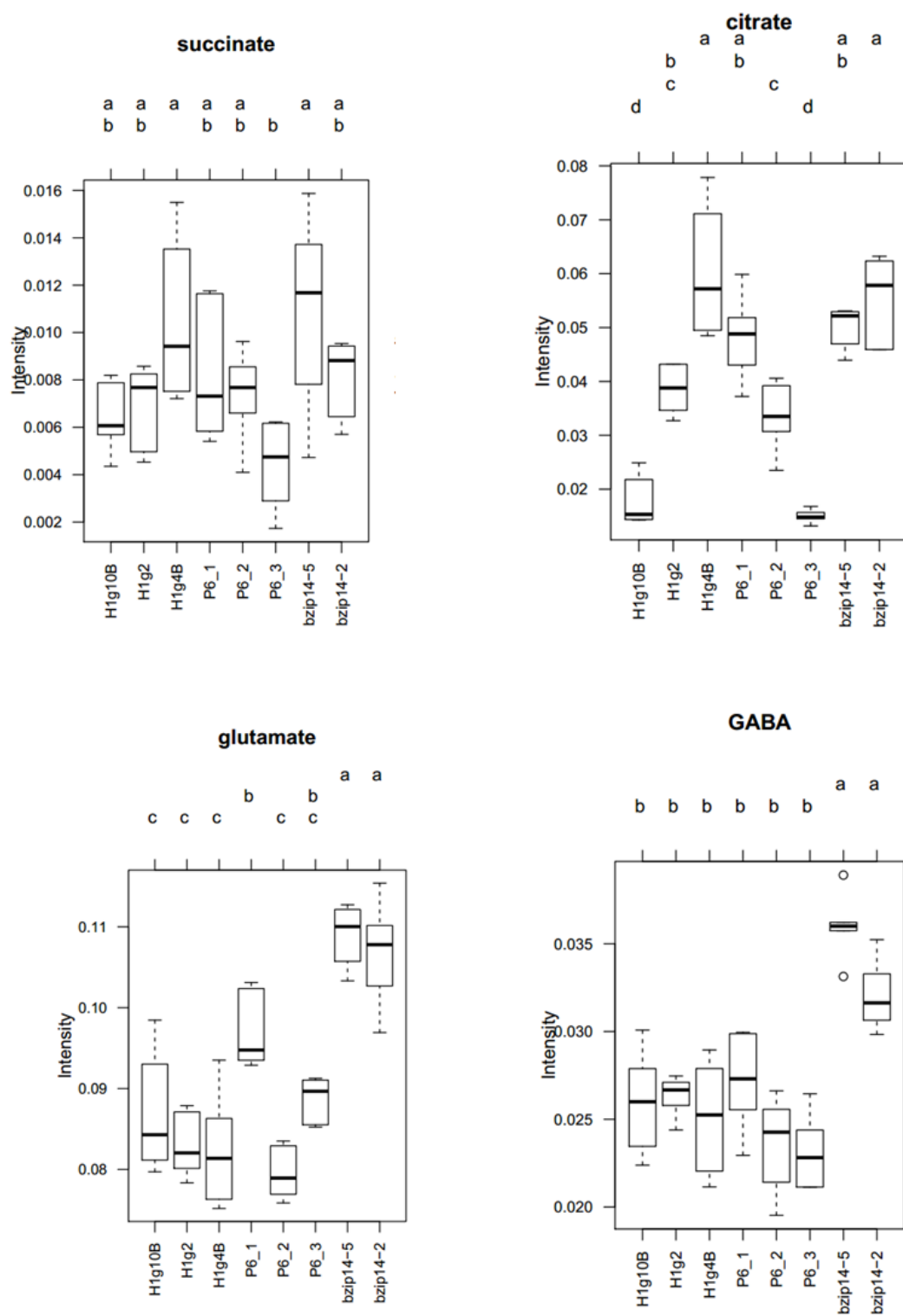


Figure 16: Metabolite levels of two TCA cycle intermediates, glutamate and gamma-butyric acid. Graphs were the result of a one way ANOVA on metabolite intensities

Materials and methods are presented in chapter 6.

DISCUSSION

In this study a detailed analysis was performed the transcriptome and the primary metabolites of *Phaeodactylum tricornutum* during N starvation and three additional stresses. To the best of our knowledge this represents the most extensive RNA-seq dataset on stress conditions generated under identical growing conditions. This resulted in the identification of pathways affected by nitrogen starvation and a list of transcriptional regulators affected by the various stresses. All stresses showed distinctive transcriptome changes. One of the most striking features of this dataset was that the genes encoding the enzymes of the TCA cycle are highly upregulated during nitrogen starvation. While some studies had noted the upregulation of the TCA before on the transcriptional or protein level, the tight transcriptional co-ordination between the different enzymes had not been reported (Hockin et al. 2012; Valenzuela et al. 2012). Clustering of gene expression patterns was greatly aided by the other stress conditions, since co-regulated genes also behave in a coherent manner in other conditions besides nitrogen deprivation. This was a major advantage compared to other studies.

As the TCA cycle is a central pathway in the primary metabolism, it is unclear which processes are most affected, it can oxidize acetyl CoA derived from a variety of sources, or supply carbon skeletons for the biosynthesis of many amino acids. From our dataset we expect that the TCA cycle provides energy while potentially serving as a redistribution hub for the deaminated carbon molecules of amino acids. Labeling experiments showed that while the levels of the metabolites increased, only small amounts of newly fixed carbon ended up in these measured compounds. It is therefore likely that the TCA cycle serves to break down carbon locked in amino acids and other carbon containing molecules for both energy generation and biosynthesis of storage compounds. The photosynthetic apparatus of diatoms is strongly impaired in nitrogen starved cells and it is likely that the cells can satisfy their energy needs for some time by oxidizing stored carbon. Paradoxically carbon fixation must continue as the carbon to nitrogen ratio in the cell strongly increases during nitrogen starvation as lipids are accumulated (Mühlroth et al. 2013). A possible follow up experiment would be to check whether TCA transcript levels remain high after several days of nitrogen starvation. An unresolved question is why intermediates such as citrate and succinate are abundant when enzyme levels are expected to be high while the inverse is expected.

Most pathways of the primary metabolism are regulated on the posttranscriptional level as fluxes can be by altered by fine-tuning the activity of existing enzyme pools, e.g. through allosteric regulation or phosphorylation. It was therefore surprising that many of the enzyme transcripts were upregulated in a coordinated fashion. This has likely to do with the size of the TCA cycle as it requires a larger number of enzymes than most pathways (Fendt et al. 2010). Diatoms do not appear to be the only organisms that control the TCA cycle on the transcriptional level. Several transcription factors in budding yeast are able to influence transcript levels, e.g. the aforementioned RTG3, RDS2 and GCN4. Studies in this organism have also shown that the flux through the TCA cycle is the sole metabolic cycle that can be affected by transcription factors knock outs (Fendt et al. 2010).

In order to find the transcription factors responsible for these transcriptomic changes, two possible candidates regulating nitrogen starvation processes were chosen: HSF1g and bZIP14. The upregulation of both transcription factors during nitrogen starvation was seen by previous work and confirmed in this study by QPCR on independently generated RNA (Valenzuela et al. 2012).

HSF1g (Phatr42514) belongs to a family of transcription factors first identified as a single gene in *Drosophila melanogaster* where it regulates the stress response upon heat treatment (Akerfelt et al. 2010). In stramenopiles such as diatoms and *Phytophthora sp.* there has been an expansion of Heat Shock Factor transcription factors and they are likely to be involved in many processes besides heat tolerance (Rayko et al. 2010). A nuclear localization signal (PS50079) and a canonical HSF DNA binding domain (SM00415) are present in the protein. The transcription factor appears to be present in all sequenced diatom genomes but it is difficult to say whether it is conserved in more distantly related species because of the prevalence of HSF's in stramenopiles genomes. Although lines were created that ectopically expressed HSF1g, no further leads were available to resolve the function of this transcription factor and no clear link was found with metabolism. The consensus motif bound by Hsf1g was identified but, in contrast to bZIP14, the processes in which Hsf1g is involved remain elusive, due to the prevalence of the motif in the genome. The ubiquity of the motif in *P. tricornutum* promoters is hard to interpret. Transcription factor binding is a complex process that often depends on more than sequence alone and usually requires co-factors and a favorable local chromatin structure. This makes it unlikely that all putative sites are bound.

More successes were obtained with the TF bZIP14. Co-expression analysis showed that the transcription profile resembled that of the TCA cycle. Unlike HSF1g, which in this dataset only increases expression under nitrogen limiting conditions, bZIP14 also shows higher expression 20 hours after nocodazole treatment, the significance of this is unknown but it corresponds to the expression pattern of citrate synthase. Furthermore the motif identified by the protein binding array was found in several TCA cycle enzymes. The core motif bound by bZIP14 is also present in the promoter of *T. pseudonana* malate dehydrogenase (Thaps20726) (TGACGTTA), strengthening the link to the TCA cycle. The increasing availability of closely related genomes will make the identification of conserved cis-regulatory motifs in diatoms much easier as it has done in mammalian genomes (King et al. 2005). The most conclusive proof of the involvement of bZIP14 in TCA cycle regulation was found when QPCR data showed that bZIP14 overexpression lines had increased abundance of five TCA cycle transcripts. An open question remains if bZIP14 directly influences expression levels of all enzymes involved or if it only controls a few and steers the others indirectly e.g. by substrate availability. Experiments are ongoing to prove direct interaction between the promoters of these genes and bZIP14.

Originally it was hoped that one of the examined transcription factors would increase lipid levels but no changes were seen in any of the HSF1g or bZIP14 overexpression lines and determined by microscopic examination after staining with Nile Red.

It is unlikely that bZIP14 is the only regulator of the TCA cycle, since several transcription factor knockouts of *S. cerevisiae* were found to be impaired in TCA cycle function. Moreover, transcription factors are by no means the only type of protein able to influence metabolic pathways, but they are easy to recognize because most contain readily identifiable domains. The amount of tools available for TF DNA binding identification is also well established in contrast to other protein classes such as kinases. The unlocking of diatom metabolism will go hand in hand with the identification of its control mechanisms.

REFERENCES

- Akerfelt M, Morimoto RI, Sistonen L (2010) Heat shock factors: integrators of cell stress, development and lifespan. *Nature reviews Molecular cell biology* 11 (8):545-555. doi:10.1038/nrm2938
- Allen AE, Dupont CL, Oborník M, Horák A, Nunes-Nesi A, McCrow JP, Zheng H, Johnson DA, Hu H, Fernie AR (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* 473 (7346):203-207
- Amin SA, Parker MS, Armbrust EV (2012) Interactions between diatoms and bacteria. *Microbiology and Molecular Biology Reviews* 76 (3):667-684
- Brudler R, Gessner CR, Li S, Tyndall S, Getzoff ED, Woods Jr VL (2006) PAS Domain Allostery and Light-induced Conformational Changes in Photoactive Yellow Protein upon I2 Intermediate Formation, Probed with Enhanced Hydrogen/Deuterium Exchange Mass Spectrometry. *Journal of Molecular Biology* 363 (1):148-160. doi:10.1016/j.jmb.2006.07.078
- Carpinelli EC, Telatin A, Vitulo N, Forcato C, D'Angelo M, Schiavon R, Vezzi A, Giacometti GM, Morosinotto T, Valle G (2014) Chromosome scale genome assembly and transcriptome profiling of *Nannochloropsis gaditana* in nitrogen depletion. *Molecular plant* 7 (2):323-335
- Chaplin MF, Kennedy JF (1994) Carbohydrate analysis: a practical approach. vol Ed. 2. IRL Press Ltd,
- Chauton MS, Winge P, Brembu T, Vadstein O, Bones AM (2013) Gene regulation of carbon fixation, storage, and utilization in the diatom *Phaeodactylum tricornutum* acclimated to light/dark cycles. *Plant Physiol* 161 (2):1034-1048. doi:10.1104/pp.112.206177
- Daboussi F, Leduc S, Marechal A, Dubois G, Guyot V, Perez-Michaut C, Amato A, Falcatore A, Juillerat A, Beurdeley M, Voytas DF, Cavarec L, Duchateau P (2014) Genome engineering empowers the diatom *Phaeodactylum tricornutum* for biotechnology. *Nature communications* 5:3831. doi:10.1038/ncomms4831
- Deschamps P, Moreira D (2012) Reevaluating the green contribution to diatom genomes. *Genome biology and evolution* 4 (7):795-800
- Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, Taylor FJ (2004) The evolution of modern eukaryotic phytoplankton. *Science* 305 (5682):354-360. doi:10.1126/science.1095964
- Fendt SM, Oliveira AP, Christen S, Picotti P, Dechant RC, Sauer U (2010) Unraveling condition-dependent networks of transcription factors that control metabolic pathway activity in yeast. *Molecular systems biology* 6:432. doi:10.1038/msb.2010.91
- Ge F, Huang W, Chen Z, Zhang C, Xiong Q, Bowler C, Yang J, Xu J, Hu H (2014) Methylcrotonyl-CoA Carboxylase Regulates Triacylglycerol Accumulation in the Model Diatom *Phaeodactylum tricornutum*. *The Plant cell* 26 (4):1681-1697. doi:10.1105/tpc.114.124982

- Granum E, Mykkestad SM (2002) A simple combined method for determination of beta-1,3-glucan and cell wall polysaccharides in diatoms. *Hydrobiologia* 477 (1-3):155-161
- Guarnieri MT, Nag A, Smolinski SL, Darzins A, Seibert M, Pienkos PT (2011) Examination of triacylglycerol biosynthetic pathways via de novo transcriptomic and proteomic analyses in an unsequenced microalga. *PLoS One* 6 (10):e25851
- Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D, Altman RB, Davis RW, Nislow C, Giaever G (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science (New York, NY)* 320 (5874):362-365. doi:10.1126/science.1150021
- Hockin NL, Mock T, Mulholland F, Kopriva S, 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* 158 (1):299-312. doi:10.1104/pp.111.184333
- Huysman MJJ, Fortunato AE, Matthijs M, Schellenberger Costa B, Vanderhaeghen R, Van den Daele H, Sachse M, Inzé D, Bowler C, Kroth PG, Wilhelm C, Falcatore A, Vyverman W, De Veylder L (2013) AUREOCHROME1a-mediated induction of the diatom-specific cyclin *dsCYC2* controls the onset of cell division in diatoms (*Phaeodactylum tricornutum*). *The Plant cell* 25 (1):215-228. doi:10.1105/tpc.112.106377
- Huysman MJJ, Martens C, Vandepoele K, Gillard J, Rayko E, Heijde M, Bowler C, Inzé D, Van de Peer Y, De Veylder L, Vyverman W (2010) Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome biology* 11 (2). doi:10.1186/gb-2010-11-2-r17
- Jordan MA, Wilson L (1998) Microtubules and actin filaments: dynamic targets for cancer chemotherapy. *Current opinion in cell biology* 10 (1):123-130
- King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome research* 15 (8):1051-1060
- Liu Z, Butow RA (1999) A transcriptional switch in the expression of yeast tricarboxylic acid cycle genes in response to a reduction or loss of respiratory function. *Molecular and cellular biology* 19 (10):6720-6728
- Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, Vardi A, Grandbastien MA, Bowler C (2009) Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics* 10:624. doi:10.1186/1471-2164-10-624
- Mühlroth A, Li K, Røkke G, Winge P, Olsen Y, Hohmann-Marriott MF, Vadstein O, Bones AM (2013) Pathways of lipid metabolism in marine algae, co-expression network, bottlenecks and candidate genes for enhanced production of EPA and DHA in species of Chromista. *Marine drugs* 11 (11):4662-4697
- Mykkestad S, Holm-Hansen O, Vårum KM, Volcani BE (1989) Rate of release of extracellular amino acids and carbohydrates from the marine diatom *Chaetoceros affinis*. *Journal of Plankton Research* 11 (4):763-773
- Ohno N, Inoue T, Yamashiki R, Nakajima K, Kitahara Y, Ishibashi M, Matsuda Y (2012) CO₂-cAMP-responsive cis-elements targeted by a transcription factor with CREB/ATF-like basic zipper domain in the marine diatom *Phaeodactylum tricornutum*. *Plant physiology* 158 (1):499-513
- Rayko E, Maumus F, Maheswari U, Jabbari K, Bowler C (2010) Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. *New Phytologist* 188 (1):52-66

- Rismani-Yazdi H, Haznedaroglu BZ, Hsin C, Peccia J (2012) Transcriptomic analysis of the oleaginous microalga *Neochloris oleoabundans* reveals metabolic insights into triacylglyceride accumulation. *Biotechnol Biofuels* 5 (1):74. doi:10.1186/1754-6834-5-74
- Schauder R, Widdel F, Fuchs G (1987) Carbon assimilation pathways in sulfate-reducing bacteria II. Enzymes of a reductive citric acid cycle in the autotrophic *Desulfobacter hydrogenophilus*. *Archives of microbiology* 148 (3):218-225
- Schmollinger S, Muhlhaus T, Boyle NR, Blaby IK, Casero D, Mettler T, Moseley JL, Kropat J, Sommer F, Strenkert D, Hemme D, Pellegrini M, Grossman AR, Stitt M, Schroda M, Merchant SS (2014) Nitrogen-Sparing Mechanisms in *Chlamydomonas* Affect the Transcriptome, the Proteome, and Photosynthetic Metabolism. *The Plant cell* 26 (4):1410-1435. doi:10.1105/tpc.113.122523
- Secco D, Wang C, Arpat BA, Wang Z, Poirier Y, Tyerman SD, Wu P, Shou H, Whelan J (2012) The emerging importance of the SPX domain-containing proteins in phosphate homeostasis. *The New phytologist* 193 (4):842-851
- Shrestha RP, Tesson B, Norden-Krichmar T, Federowicz S, Hildebrand M, Allen AE (2012) Whole transcriptome analysis of the silicon response of the diatom *Thalassiosira pseudonana*. *BMC Genomics* 13:499. doi:10.1186/1471-2164-13-499
- Soontorngun N, Larochelle M, Drouin S, Robert F, Turcotte B (2007) Regulation of gluconeogenesis in *Saccharomyces cerevisiae* is mediated by activator and repressor functions of Rds2. *Molecular and cellular biology* 27 (22):7895-7905. doi:10.1128/mcb.01055-07
- Steinhauser D, Fernie AR, Araujo WL (2012) Unusual cyanobacterial TCA cycles: not broken just different. *Trends Plant Sci* 17 (9):503-509. doi:10.1016/j.tplants.2012.05.005
- Terasaki M, Chen LB, Fujiwara K (1986) Microtubules and the endoplasmic reticulum are highly interdependent structures. *The Journal of cell biology* 103 (4):1557-1568
- Valenzuela J, Mazurie A, Carlson RP, Gerlach R, Cooksey KE, Peyton BM, Fields MW (2012) Potential role of multiple carbon fixation pathways during lipid accumulation in *Phaeodactylum tricornutum*. *Biotechnol Biofuels* 5 (1):40
- Yang ZK, Zheng JW, Niu YF, Yang WD, Liu JS, Li HY (2014) Systems-level analysis of the metabolic responses of the diatom *Phaeodactylum tricornutum* to phosphorus stress. *Environmental microbiology* 16 (6):1793-1807
- Yongmanitchai W, Ward O (1991) Growth of and omega-3 fatty acid production by *Phaeodactylum tricornutum* under different culture conditions. *Applied and environmental microbiology* 57 (2):419-425

Chapter 6:

Materials and methods

Author contributions:

MM wrote this chapter

MATERIALS AND METHODS

GROWTH

All experiments used the *Phaeodactylum tricornutum* (Pt1) Bohlin Strain 8.6 obtained from the diatom culture collection of the PAE-Ugent. Cells were grown in 500ml erlenmeyers with ESAW medium containing 7.5mg sodium nitrate per liter, other nutrients were added as originally listed (Berges et al. 2001). For RNA-sequencing the pre-culture cells were diluted 1/2 daily to keep growth exponential. For the timecourse experiment cells were harvested by 30 minutes of 6000g centrifugation and washed with nitrogen and phosphate free ESAW. This starter culture was split into equal parts and used to inoculate ESAW with and without added nitrogen. Growth was monitored by OD measurements at 450nm.

RNA EXTRACTION

For RNA-sequencing cells were captured on a 3µm pore size PVDF membrane (Millipore) by filtering 20 ml of culture as rapidly as possible using a mild vacuum. For all other experiments 50 ml of culture was centrifuged at 2600g for 10 minutes. Filters or algal pellets were transferred to 2 ml eppendorfs and flash frozen in liquid nitrogen. For RNA extraction the pellets were resuspended or cells were washed of the filter with 1.5 ml of Tri Reagent (Molecular research) and processed according to the manufacturer's instructions up until the RNA precipitation step. The aqueous phase was mixed with an equal volume of 70% ethanol and transferred to a RNA-Easy mini spin column (Qiagen). On column DNA'se digest was performed according to the manufacturer's instructions using RQ1 DNase (Promega). RNA was eluted in RNase free water and quality control was performed on Nanodrop and 1% agarose gel. Ten TruSeq RNA samples were sequenced paired-end 100bp with a Illumina HiSeq2000 was performed at UZ Leuven Genomics core.

RNA-SEQ ANALYSIS

Using a cutoff value of 20, FASTQ files were quality filtered and adaptor sequences were removed using Cutadapt (Martin 2011). Only the sequences that were still paired were retained for further analysis. Mapping and counting the reads was done with Tophat/Cufflinks/Cuffdiff using the GFF filtered models and genome files available on the JGI website (Trapnell et al. 2013; Bowler et al. 2008). The expression of significantly expressed genes was log2 transformed and visualized using the TM4 MeV package (Saeed et al. 2006). Additional visualization was performed using the Mapman program (Usadel et al. 2009). Initial mapping was performed with the Mercator webserver and manually refined using Diatomcyc data (Lohse et al. 2014). The JGI functional annotation was supplemented with BLAST2GO, Pico-Plaza and Diatomcyc data (Fabris et al. 2012) (Conesa et al. 2005).

MOLECULAR CLONING

The GATEWAY procedure was used for subcloning and the generation of expression clones according to the manufacturer's instructions (Life Technologies). Models were manual checked for completeness and correspondence with RNA-seq coverage using the IGV browser. The destination vectors used were generated previously (Siaut et al. 2007).

A list of primers is available in appendix 2. Picking up genes was done with Primestart DNA polymerase (TaKaRa Biosciences), diagnostic PCR with GoTaq (Promega).

FLOW CYTOMETRIC ANALYSIS

Two milliliter aliquots were taken from the cell culture in triplicate and subsequently centrifuged at 6000g for two minutes. The supernatants were decanted and the cells were resuspended in 70% ethanol and stored at 4°C until analysis. Cells were stained by pelleting the cells as above and washing them twice with PBS, DNA staining was performed with 4',6-diamidino-2-phenylindole (DAPI) for 15 minutes at a final concentration of 1ng/ml. Cytometric analysis was performed on a Partec CyFlow ML using the Flomax software tool (Partec). A minimum of 10⁴ cells were processed for each replicate.

DETERMINATION OF INTERGENIC REGIONS

Using the gene models as available on the JGI website, 500bp were taken upstream from the first exon for each gene. Intergenic regions are often small in *P. tricornutum* and a large percentage of putative promoters overlap with promoters of adjacent genes or UTR's.

DETERMINATION OF OVERREPRESENTED MOTIFS

Starting from the RNA-sequencing data generated, 681 genes were chosen with the highest induction during nitrogen starvation. The MEME program (version 4.6.1) was run with standard settings with a minimum motif size of 6bp and a maximum size of 20 (Bailey et al. 2006). Motifs were then screened for overrepresentation using a set of 680 random intergenic sequences. +

Y1H

A yeast one hybrid screening was performed using a cDNA library synthesized by Invitrogen cloned in pDest22 as reported in a previous study (Huysman et al. 2013; Deplancke et al. 2006). The vectors and method were taken from (Deplancke et al. 2006), yeast strain YM4271 was used as described. Baits of Motif6 and Motif17 were generated by synthetically synthesizing a oligo with four occurrences of each respective motif with ten nucleotides of flanking sequence. Primer design was done with TmPrime (Bode et al. 2009). The sequence for motif 17 was derived from the upstream regions of Phatr13154, 45851/2,

37436,29711. Several independent repeats were tested for LacZ and His auto-activation. Library screening was performed on 60mM 3-AT for the M17 construct and 80mM for M6. Transformation efficiencies were above 10⁶. Colonies were picked after four days and re-streaked on new selective medium. The pdest22 insert of positive colonies was amplified using the primers recommended by Deplancke et al.

SITE DIRECT MUTAGENESIS

Conserved cysteine and histidine residues were identified by aligning the NMB1 proteins in *P. tricornutum* and *Thalassiosira pseudonana*. The sequence CTSHARC was mutated to ATSAARA using an adapted version of the QuickChange method (Zheng et al. 2004), primers are listed in appendix as NMB1_RD.

RECOMBINANT PROTEIN PRODUCTION

Proteins were expressed in *Escherichia coli* BL21-AI (Life Technologies). The CDS of NMB1 and the NMB1-RD were cloned into pdest MBP-HIS which resulted in N-terminal fusion product with the His tag and the Maltose Binding Protein. Cells were grown to a OD600 of 0.4-0.8 and induced with 0.4 mM IPTG and 0.2% Arabinose. After induction cells were transferred to 21°C and incubated for four hours. Protein was liberated by four minutes of sonication on ice with 10 seconds on/off time. The protein was captured on agarose amylose resin (New England Biolabs). Protein purity was checked by SDS PAGE on Coomassie stained 4-15% Mini Protean TGX precast SDS-PAGE gels (Biorad).

TRANSFORMATION OF PHAEODACTYLUM TRICORNUTUM

Transformants of *P. tricornutum* were generated using microparticle bombardment according to the protocol of Kroth (Kroth 2007). Zeocin resistance was introduced on the paf6 plasmid while genes of interest were cloned into pDEST-FCP as described previously (Siaut et al. 2007). Resistant colonies were restreaked on selective medium and afterwards brought into liquid culture.

IN VITRO UBIQUITINATION ASSAY

Human E2 ligase was combined with recombinant NMB1 as described in (Liu and Stone 2010). Briefly the protein was produced in *E. coli* as a MBP fusion protein and purified as described above. The reaction was performed in 30 µl. 50 ng of human E1 ligase, 250 ng of human E2 ligase and 500 ng of NMB1 or NMB1-Ring Dead were combined together with 0,5 µL 10 mg/mL Ubiquitin (5 µg) and 1 µL Creatine Kinase (0,1 U). 3 µL of 10x Ub-buffer which resulted in a final concentration of containing 50 mM Tris-HCl, pH 7.5, 10 mM MgCl₂, 0.05 mM ZnCl₂, 1 mM ATP, 0.2 mM DTT, 10 mM phosphocreatine, was added and H₂O to 30 µl. No substrate was added as it was suspected that protein was able to auto-ubiquitinate.

EMSA

EMSA was performed according to (Memelink 2013). Briefly complementary oligonucleotides containing the M17 repeat used for screening the Y1H library were annealed and labeled with ATP- γ -P³² using T4 polynucleotide kinase (Promega). Excess label was removed using Sephadex G25 columns (GE-Healthcare). Binding buffer consisted out of 10% glycerol, 0.1 M KCl, 0.1 M HEPES pH 7.4 and 10mg/ml BSA. ZnCl₂ was added to a final concentration of 10 mM. Recombinant protein was incubated in 1X binding buffer with 5mg of protein and loaded onto 10% acrylamide/bis-acrylamide 0.5X TBE buffer. Gels were run for 1 hour at a constant voltage of 120V. Autoradiograms were exposed overnight.

TRANSIENT PROTOPLAST ASSAY

Transient protoplast assays were performed as described earlier (Vanden Bossche et al. 2013). Briefly four repeats of the NMB1 recognition motif were cloned in a vector upstream of firefly Luciferase (fLuc). The transcription factor NMB1 was expressed by cloning the coding sequence in a vector containing the 35S promoter. Renilla luciferase (rLUC) placed under control of the 35S promoter was used as a transfection efficiency control. Bright Yellow-2 tobacco cell cultures were co-transfected with the above constructs using 2 μ g of each plasmid. To correct for auto activation cells containing only the putative binding site (motif only) and rLUC were generated and quantified. Transfected protoplasts were incubated overnight at room temperature in an orbital shaker. After lysis fLUC activity was measured and corrected for autoactivation (motif only) and transfection efficiency (rLUC) using the Dual-Luciferase reporter assay system (Promega). All assays were performed with six technical replicates and analysed using T-tests.

QPCR

cDNA was generated with iScript kit (Bio-Rad) from RNA extracted as described above. qRT-PCR was carried out with a Lightcycler 480 (Roche) and SYBR Green QPCR Master Mix (Stratagene). PUA, VTC4 and RP3A were used as reference genes with primers listed in appendix. Analysis was performed using the $\Delta\Delta$ Ct method as described in (Schmittgen and Livak 2008).

SELECTION OF PUTATIVE INTERGENIC REGIONS

Using the gene models as available on the JGI website, 500bp were taken upstream from the first exon for each gene. Intergenic regions are often small in *P. tricornutum* and a large percentage of putative promoters overlap with promoters of adjacent genes or UTR's.

CONSTRUCTION OF PHYLOGENETIC TREES

The consensus coding sequences of nine diatoms were downloaded from the CAMERA MMETSP website and used to construct a local BLAST database. The tBLASTn program was used to find

homologous sequences, taking only the highest scoring hit for each sequence. For tree construction the default method was chosen on phylogeny.fr (Dereeper et al. 2008). Briefly, alignment was performed using MUSCLE 3.7 set at highest accuracy. Poorly aligned regions and gaps were removed using Gblocks v0.91b with removing all nonconserved positions longer than 8 amino acids with, a minimum remaining block length of 10, not allowing any gaps in the final alignment and at least 85% of the sequences present in any flanking regions. The maximum likelihood method was as used as implemented in the PhyML program v3.0 aRLT using the WAG substitution model was used assuming an estimated proportion of invariant sites and four gamma distributed rate categories with an estimated parameter of 1.293. Branch reliability was tested using the aLRT test. Graphics were generated with TreeDyn v1.9.3.

RECOMBINANT PROTEIN PRODUCTION

Proteins were expressed in *Escherichia coli* BL21-AI (Life Technologies). The CDS of bZIP14 and Hsf1g were cloned into pdest MBP-HIS which resulted in N-terminal fusion product with the His tag and the Maltose Binding Protein. Cells were grown to a OD₆₀₀ of 0.4-0.8 and induced with 0.4 mM IPTG and 0.2% Arabinose. After induction cells were transferred to 21°C and incubated for four hours. Protein was liberated by four minutes of sonication on ice with 10 seconds on/off time. The protein was captured on agarose amylose resin (New England Biolabs). Protein purity was checked on Coomassie stained 4-15% Mini Protean TGX precast SDS-PAGE gels (Biorad).

REFERENCES

- Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* 34 (suppl 2):W369-W373
- Berges JA, Franklin DJ, Harrison PJ (2001) Evolution of an artificial seawater medium: improvements in enriched seawater, artificial water over the last two decades. *J Phycol* 37 (6):1138-1145
- Bode M, Khor S, Ye H, Li MH, Ying JY (2009) TmPrime: fast, flexible oligonucleotide design software for gene synthesis. *Nucleic Acids Res* 37 (Web Server issue):W214-221. doi:10.1093/nar/gkp461
- Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret JP, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, Fournet J, Haruta M, Huysman MJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kroger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin-Jezequel V, Lopez PJ, Lucas S, Mangogna M, McGinnis K, Medlin LK, Montsant A, Oudot-Le Secq MP, Napoli C, Obornik M, Parker MS, Petit JL, Porcel BM, Poulsen N, Robison M, Rychlewski L, Ryneerson TA, Schmutz J, Shapiro H, Saut M, Stanley M, Sussman MR, Taylor AR, Vardi A, von Dassow P, Vyverman W, Willis A, Wyrwicz LS, Rokhsar DS, Weissenbach J, Armbrust EV, Green BR, Van de Peer Y, Grigoriev IV (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456 (7219):239-244. doi:10.1038/nature07410
- Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21 (18):3674-3676. doi:10.1093/bioinformatics/bti610

- Deplancke B, Vermeirssen V, Arda HE, Martinez NJ, Walhout AJ (2006) Gateway-compatible yeast one-hybrid screens. *CSH protocols* 2006 (5). doi:10.1101/pdb.prot4590
- Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids research* 36 (Web Server issue):W465-469. doi:10.1093/nar/gkn180
- Fabris M, Matthijs M, Rombauts S, Vyverman W, Goossens A, Baart GJ (2012) The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *Plant J* 70 (6):1004-1014. doi:10.1111/j.1365-313X.2012.04941.x
- Huysman MJ, Fortunato AE, Matthijs M, Costa BS, Vanderhaeghen R, Van den Daele H, Sachse M, Inze D, Bowler C, Kroth PG, Wilhelm C, Falciatore A, Vyverman W, De Veylder L (2013) AUREOCHROME1a-mediated induction of the diatom-specific cyclin dsCYC2 controls the onset of cell division in diatoms (*Phaeodactylum tricornutum*). *The Plant cell* 25 (1):215-228. doi:10.1105/tpc.112.106377
- Kroth PG (2007) Genetic transformation: a tool to study protein targeting in diatoms. *Methods in molecular biology* (Clifton, NJ) 390:257-267
- Liu H, Stone SL (2010) Absciscic acid increases Arabidopsis ABI5 transcription factor levels by promoting KEG E3 ligase self-ubiquitination and proteasomal degradation. *Plant Cell* 22 (8):2630-2641. doi:10.1105/tpc.110.076075
- Lohse M, Nagel A, Herter T, May P, Schroda M, Zrenner R, Tohge T, Fernie AR, Stitt M, Usadel B (2014) Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environ* 37 (5):1250-1258. doi:10.1111/pce.12231
- Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17 (1):10--12
- Memelink J (2013) Electrophoretic mobility shift assay for the analysis of interactions of jasmonic acid-responsive transcription factors with DNA. *Methods Mol Biol* 1011:209-225. doi:10.1007/978-1-62703-414-2_17
- Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J (2006) [9] TM4 Microarray Software Suite. *Methods Enzymol* 411:134-193
- Schmittgen TD, Livak KJ (2008) Analyzing real-time PCR data by the comparative CT method. *Nat Protoc* 3 (6):1101-1108
- Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falciatore A, Bowler C (2007) Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. *Gene* 406 (1-2):23-35. doi:10.1016/j.gene.2007.05.022
- Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 31 (1):46-53. doi:10.1038/nbt.2450
- Usadel B, Poree F, Nagel A, Lohse M, CZEDIK-EYSENBERG A, Stitt M (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. *Plant, cell & environment* 32 (9):1211-1229
- Vanden Bossche R, Demedts B, Vanderhaeghen R, Goossens A (2013) Transient expression assays in tobacco protoplasts. *Methods in molecular biology* (Clifton, NJ) 1011:227-239. doi:10.1007/978-1-62703-414-2_18
- Zheng L, Baumann U, Reymond JL (2004) An efficient one-step site-directed and site-saturation mutagenesis protocol. *Nucleic Acids Res* 32 (14):e115. doi:10.1093/nar/gnh110

Chapter 7:

Conclusion and perspectives

Author contributions:

MM wrote this chapter

The original aim of this work was to use RNA-sequencing in nutrient deprived cultures to find transcription factors regulating lipid biosynthesis in *P. tricornutum*. Surprisingly, while fatty acid levels increased and a handful of lipid biosynthesis genes were upregulated, it does not appear that lipid biosynthesis is controlled on the transcriptional level. This observation has also been made for *Nannochloropsis gaditana*, another unicellular heterokont algae (Corteggiani Carpinelli et al. 2014). Less studies have been performed on the proteome level, but the emerging picture appears to be similar. No large scale changes in fatty acid biosynthetic enzymes were visible in the proteomes of either the diatom *Thalassiosira pseudonana* or *N. gaditana* during nitrogen deprivation (Dong et al. 2013; Hockin et al. 2012). It may well be that the increase of the substrate pool available for lipid biosynthesis is controlled on the transcriptional level, but the metabolic models for *P. tricornutum* are not complete enough to make accurate predictions.

One of the major hurdles in determining lipid levels lies in the comparing of growth conditions between unmodified strains and transgenic strains. Slight changes in culture conditions or effects that impair growth might alter lipid accumulation and complicate the estimation of the effect of a single gene. Nevertheless, many transgenic diatoms have been generated with the aim of increasing the production of a set of metabolites present in the cell, e.g. lipids. Because of the limited amount of tools available for diatoms, these attempts have focused on altering the expression of a single gene. Enzymes however rarely operate alone, most often they belong to a complex pathway where fluxes depend on a multitude of enzymes. These activities are regulated at the transcriptional, translational and post-translational level in a species-dependent regulatory context. This means that even when the amount or activity of an enzyme is increased, it does not necessarily result in more product. While there is a single report of increased lipid levels after overexpression of a native enzyme, there does not appear much success in this approach (Niu et al. 2013). Predicting increases or decreases of a metabolite by modelling the effect of ectopic expression of a single enzyme is difficult in even the most characterized model system and this is almost impossible in the underdetermined metabolic models that have been generated for diatoms to date (Tran et al. 2008). In order to bypass the regulation of an enzyme it is possible to introduce an enzyme that performs the same function as the rate limiting enzyme from a distantly related organism in order to bypass any regulatory signals, as has been done with an acyltransferase in *S. cerevisiae* (Kalscheuer et al. 2004). To the best of our knowledge this has not been attempted in diatom engineering although the author is not the first to suggest this.

Biosynthetic pathways of secondary metabolites are often under stringent transcriptional control and altering the expression levels of a single transcription factor can increase target metabolite levels dramatically (Memelink et al. 2001). Primary metabolism is regulated more on the post-transcriptional level. Nonetheless transcription factors exist which control lipid biosynthesis fluxes in other organisms such as the human SREBP and ChREBP (Horton et al. 1998; Yamashita et al. 2001). If there is

transcriptional regulation on lipid biosynthesis in diatoms, it may not have been discovered, because it could be radically different compared to what is seen in other photosynthetic organisms. At the time of writing, the genes controlling lipid biosynthesis in diatoms are completely unknown. With a better understanding of the regulation, major opportunities will arise to increase lipid levels through a rational approach.

Blocking competing reactions on the other hand is feasible with the current resources and several positive results have been achieved using RNAi or knock out technology. The targets chosen were chrysolaminaran synthesis and lipid degradation (Daboussi et al. 2014a; Trentacoste et al. 2013). Since gene editing using TAL nucleases has been demonstrated by the former paper, these approaches are bound to be used more often in the near future.

The current genes that have been proven to have a deleterious effect on lipid accumulation after knockdown are not likely to show an effect upon overexpression. The branched chain amino acid degradation enzyme for example cannot degrade more amino acids than are present in the cell (Ge et al. 2014). Blocking competing reactions on the other hand is much more predictable and several good results have been achieved by using RNAi or knockout technology, either by blocking chrysolaminaran synthesis or lipid degradation (Daboussi et al. 2014b; Trentacoste et al. 2013).

Changing the activity of native pathways usually requires changing the expression of multiple genes and often it is unknown which of the genes involved in a process should be targeted. An alternative to this approach is ectopic expression of a regulatory protein controlling the pathway (Courchesne et al. 2009). The most easily identifiable regulators are kinases and transcription factors with the latter having a large advantage due to the available tools for screening DNA binding proteins. In this work we attempted to identify the transcription factors steering metabolic pathways in *P. tricornutum*.

THE NITROGEN STARVATION RESPONSE AND THE TRANSCRIPTION FACTORS GUIDING IT

Despite the unexpected setback in identifying lipid related transcription factors, the RNA-seq profiling has allowed us to get a better understanding of the nitrogen starvation response in this diatom. The transcriptome profile showed that the shifts in other primary metabolic processes involves coordinated transcriptional regulation, which was most apparent for the TCA cycle. This finding was corroborated on the protein level in *T. pseudonana* (Hockin et al. 2012). At the moment, it is still unclear how this fits in with the metabolite profiles observed in cells under nitrogen starvation, because it would be expected that the TCA cycle and lipid biosynthesis are in competition for the same pool of Acetyl-CoA. A possible explanation is that the TCA cycle is fed mainly by carbon skeletons obtained from the deamination of nucleotides and amino acids. The observation in *N. gaditana* that most of the carbon ending up in lipids is derived from photosynthesis, which makes it unlikely that the breakdown of stored carbohydrates or amino acids contributes significantly to lipid biosynthesis. The TCA cycle could provide acetyl CoA

through the action of citrate lyase, but there was no evidence of this on the transcriptional level. A more plausible hypothesis is that by catabolizing organic molecules, the cycle could provide the reducing equivalents and energy required for lipid biosynthesis, even when photosynthesis is impaired.

The generation of overexpression lines was the most extensively used tool in our lab for *P. tricornutum* gene characterization. The aim of our work shifted away from lipids and towards the identification of transcription factors that control the TCA cycle and nitrogen metabolism in general. Two different approaches were employed to study three different transcription factors (TF's) which were all significantly induced during nitrogen starvation. HSF1g and bZIP14 were selected for further study on the basis of their expression profile while the novel TF family of NMB1 was identified by a yeast one hybrid screening after *de novo* motif discovery. The latter result could not be obtained on the basis of expression analysis alone and this underlines the need for complementary techniques in TF discovery. To the best of our knowledge these are the first transcription factors unambiguously linked to nutrient stress adaptation in *P. tricornutum*.

While previous studies used RNA-sequencing to profile nitrogen starved diatoms, our study was unique because it included three other conditions that also hamper cell division in *P. tricornutum*. Using cluster analysis this allowed us to select genes that were specifically expressed during nitrogen starvation and not during other conditions. These contrasting conditions greatly aided gene clustering.

SHINING A LIGHT ON THE BLACK BOX OF METABOLISM IN THE 'OMICS AGE

Diatoms have not one but two commonly used model organisms: *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. Both have a sequenced genome and can be genetically transformed but they only share 60% of their genes (Bowler et al. 2008). Because of the availability of 'omic techniques the lines are blurring between model and non-model species. One hurdle that remains in place is the ease of transformation which will likely keep the current models in the focal point of most studies. Nevertheless it has become clear that inter-species differences between diatoms are large and that it will not suffice to base our concepts of diatom metabolism on two rather atypical diatoms.

Sequencing projects can be turned into database of metabolism as has been illustrated by the Diatomcyc website. This database was made with pathway-tools, a semi-automated way of grouping genes into pathways. While this process still requires substantial manual refining, orthology allows the easy classification of reactions. This work has greatly benefited from the availability of this database. Undoubtedly diatom metabolism will continue to surprise with novel combinations of enzymes obtained through their unique evolutionary background or horizontal gene transfer. The number of genes with no allocated function made it impossible to assign a function to many co-expression clusters as they chiefly existed out of hypothetical genes.

Both model diatoms have diploid genomes and no known sexual cycle. This means that mutagenesis based techniques for phenotype improvement or study is non-trivial. The only exception being activation tagging(Weigel et al. 2000). However transformation efficiencies are not particularly high, on the order of 10-100 transformants on 10^8 cells, necessitating screening for easily identifiable phenotypes(Apt et al. 1996; Weigel et al. 2000). Therefore strain improvement is most likely to succeed when based on previous knowledge and rational engineering. This makes the understanding of diatom metabolism a necessity.

The lack of a gene inactivation tool was a handicap during the project and only overexpression was ultimately used. While the lack of sexual reproduction makes *P. tricornutum* easy to maintain, it also precludes it from the generation of homozygous knockouts by mutagenesis, as it is a diploid organism. Combined with the low rate of homologous recombination, this limited molecular tools to overexpression, RNAi and expression of fusion constructs at the time of this work. While RNAi is a useful tool for gene inactivation, it is difficult to assess the impact of RNAi lines. It is impossible to predict what the true impact is when a transcript is reduced by half. Furthermore it has been shown that RNAi works on the posttranscriptional level, making it impossible to quantify the severity of knock down without an antibody specific for the target protein.

Because of this last concern we decided to focus solely on overexpression in our screening. In retrospect different promoters should have been selected as the RNA-sequencing data showed that FCP-B gene is not transcribed at low levels in either darkness or nitrogen deprivation. Besides FCP-B and A only one other promoter is routinely used, namely that of nitrate reductase which was not suitable for this project as we were examining the regulation of nitrogen assimilation. There is an urgent need for more validated promoters, one of the candidates would be the histon H4. Besides, promoter choice, the other main problem, was the stability of transgenic lines. Substantial variation was seen when lines were kept in continuous culture, requiring frequent testing. This was especially problematic with RNAi lines as differences in expression are often small. With the arrival of CRISPR based knock-out technology in *P. tricornutum*, these limitations are bound to be overcome soon and gene or promoter replacement will also become possible.

THE IDENTIFICATION OF NMB1 BY DE NOVO MOTIF DISCOVERY AND THE CHARACTERISATION BZIP14

Using the list previously compiled by Rayko *et al.*, we found that many transcription factors had an expression maximum during nitrogen starvation. The number was too large to screen with overexpression, so it was decided to use a bioinformatics approach to find relevant transcription factors and their targets. The problem was reversed and instead of looking at the transcription factors, the genes with the highest fold induction during nitrogen starvation were selected and their putative promoters were used as input for motif discovery. Our goal was to find overrepresented motifs in these genes and subsequently identify transcription factors capable of binding them. The same method was used to find motifs present in genes

strongly downregulated during nitrogen starvation. While many of these motifs were statistically robust however it was decided to focus on the positive motifs because of the limitations listed above. Two motifs were synthesised and used in a Yeast One Hybrid Screening assay. Unexpectedly the transcription factor binding one of the motifs was not annotated as a transcription factor and did not even contain a clear DNA binding domain. The characterisation of this factor is listed in chapter 5. Because cis-regulatory elements are under purifying selection compared to the surrounding sequence, it is possible to validate a found motif by looking at orthologous promoters in related species(Korkuc et al. 2014). It is essential that the species used in the comparison are closely related, which means this approach is only suited for those species which have several sequenced close relatives (Martino et al. 2007). Because sequencing costs are dropping, the genomes of suitable species will soon come within reach.

A more labour intensive endeavour would be the determination of the binding sites all *P. tricornutum* transcription factors. Chromatin immune precipitation remains the golden standard but antibodies would need to be generated against each of the estimated two hundred transcription factors a average diatom. An additional complication is the apparent promiscuity of transcription factors as many binding large proportions of the genome, making it difficult to draw conclusions without prior knowledge. The protein binding microarray method employed in this work, while in a heterologous systems, gave excellent data but only worked for half of the tested proteins. A systematic screen would require protein production for all TF's, although it has been done for most of the *Saccharomyces cerevisiae* transcription factors(Gordan et al. 2011). Because the aforementioned techniques are labor intensive, they are better suited for validation of binding. Bacterial one hybrid systems on the other hand allow rapid screening of all potential binding sites for a TF as has been done for 35 transcription factor in *Drosophila melanogaster*(Noyes et al. 2008). The only way to get meaningful biological data will require the overlap between different techniques which, while requiring a significant investment of time, is by no means impossible in the 'omics age. As the examples listed in this thesis illustrate, it is possible to start from an expression atlas and a genome to find transcriptional regulators of a process. The method used in this study could be expanded towards other expression clusters of interest.

FUTURE PERSPECTIVES

Despite their problems diatoms are tractable organisms for genetic modification and several research groups have altered the metabolism of *P. tricornutum* by heterologous gene expression. Examples are the production of the omega-3 fatty acid docosahexaenoic acid, which is absent in wild type cells, after introduction of a heterologous desaturase(Hamilton et al. 2014). Another commercially relevant product was the production of biodegradable plastic feedstocks by introducing the biosynthetic genes from bacteria(Hempel et al. 2011). Researchers appear to be shifting away from bulk products an many groups are currently trying to manufacture high value compounds such as antibodies.

In terms of lipid engineering it seems that transcriptomics is not the technique of choice to investigate changes in lipid metabolism. It is possible that the regulation takes place but to work this out we need far better metabolic models. Many unexpected pathways are present in diatom genomes and localizations are unsure for many enzyme isoforms. Furthermore there appear to be substantial differences in between the main diatom model systems (McGinn and Morel 2008). This results in problematic modelling with uncertain conclusions (Zheng et al. 2013). It appears that sequencing has got us this far that there is an urgent need for enzyme localization data, metabolic flux modelling and measuring enzyme activities more directly.

While both yeast one hybrid and motif discovery are prone to false positives, it is possible to compare transcription factors and motifs in a number of species. Conserved elements are much more likely to be important. With an ever expanding range of species and increased RNA-sequencing to validate expression patterns the risks involved in screening can be minimized.

A major aid for high-throughput screening would be the availability of a transcription factor library. Admittedly, such a library would have precluded us from identifying the NMB class of transcription factors but transcription factors are often conditionally expressed and it thus would require the generation of a cDNA library for each investigated condition.

Another hurdle which will be overcome is the annotation state of the *P. tricornutum* genome. When visualizing the mapped RNA-seq reads, it quickly became apparent that most gene models do not correspond to reality and many genes are simply not present in the gene models. Although attempts were made to remedy this problem by generating new models from the RNA seq data, this attempt was eventually abandoned as the genome is dense and transcriptional units often overlap at 5' or 3' ends. Still most transcriptionally active genes are at least partially represented in the gene models and expression can be reliably estimated even when only a fragment of the gene is present, this interferes with the putative promoter predictions. Localization predictions are also hampered as localization signals are often present in the N-terminal part of the protein, precisely the part that is missing from many models. Finally the orthology based predictions are also affected as homology searches with partial proteins yields less reliable hits. However this situation can easily be remedied. The advent of strand specific RNA library construction allows the resolution of most overlapping transcriptional units. RNA-seq data was not available at the time of the genome completion. Re-annotation on the basis of stranded data should solve many of these problems.

Most of this work was performed at the dawn of affordable genome and transcriptome sequencing. The genomic revolution has provided us with the largest, ever expanding, catalogue of genes ever created. At the same time gene editing will give us the tools to find their function. This work is far from over as the ocean is the host to an enormous genetic diversity in eukaryotes and the limited overlap between diatom

genomes. Evolutionary relationships will become increasingly clear with additional sequencing data and they will allow the field to move away from the green centric paradigm of eukaryotic photoautotrophy.

Tackling the problems of pattern identification and indexing this knowledge will take far longer than it will to generate the data. The arrival of low cost gene synthesis and targeted knockouts mean that we have all the tools available to make this list of parts into a construction manual to improve diatoms.

REFERENCES

- Apt KE, Kroth-Pancic PG, Grossman AR (1996) Stable nuclear transformation of the diatom *Phaeodactylum tricornutum*. *Molecular & general genetics* : MGG 252 (5):572-579
- Corteggiani Carpinelli E, Telatin A, Vitulo N, Forcato C, D'Angelo M, Schiavon R, Vezzi A, Giacometti GM, Morosinotto T, Valle G (2014) Chromosome scale genome assembly and transcriptome profiling of *Nannochloropsis gaditana* in nitrogen depletion. *Mol Plant* 7 (2):323-335. doi:10.1093/mp/sst120
- Courchesne NMD, Parisien A, Wang B, Lan CQ (2009) Enhancement of lipid production using biochemical, genetic and transcription factor engineering approaches. *Journal of Biotechnology* 141 (1):31-41
- Daboussi F, Leduc S, Maréchal A, Dubois G, Guyot V, Perez-Michaut C, Amato A, Falciatore A, Juillerat A, Beurdeley M (2014a) Genome engineering empowers the diatom *Phaeodactylum tricornutum* for biotechnology. *Nature communications* 5
- Daboussi F, Leduc S, Marechal A, Dubois G, Guyot V, Perez-Michaut C, Amato A, Falciatore A, Juillerat A, Beurdeley M, Voytas DF, Cavarec L, Duchateau P (2014b) Genome engineering empowers the diatom *Phaeodactylum tricornutum* for biotechnology. *Nature communications* 5:3831. doi:10.1038/ncomms4831
- Dong HP, Williams E, Wang DZ, Xie ZX, Hsia RC, Jenck A, Halden R, Li J, Chen F, Place AR (2013) Responses of *Nannochloropsis oceanica* IMET1 to Long-Term Nitrogen Starvation and Recovery. *Plant Physiol* 162 (2):1110-1126. doi:10.1104/pp.113.214320
- Ge F, Huang W, Chen Z, Zhang C, Xiong Q, Bowler C, Yang J, Xu J, Hu H (2014) Methylcrotonyl-CoA Carboxylase Regulates Triacylglycerol Accumulation in the Model Diatom *Phaeodactylum tricornutum*. *The Plant cell* 26 (4):1681-1697. doi:10.1105/tpc.114.124982
- Gordan R, Murphy KF, McCord RP, Zhu C, Vedenko A, Bulyk ML (2011) Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. *Genome biology* 12 (12):R125. doi:10.1186/gb-2011-12-12-r125
- Hamilton ML, Haslam RP, Napier JA, Sayanova O (2014) Metabolic engineering of *Phaeodactylum tricornutum* for the enhanced accumulation of omega-3 long chain polyunsaturated fatty acids. *Metabolic engineering* 22:3-9. doi:10.1016/j.ymben.2013.12.003
- Hempel F, Bozarth AS, Lindenkamp N, Klingl A, Zauner S, Linne U, Steinbuchel A, Maier UG (2011) Microalgae as bioreactors for bioplastic production. *Microbial cell factories* 10:81. doi:10.1186/1475-2859-10-81

- Hockin NL, Mock T, Mulholland F, Kopriva S, 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* 158 (1):299-312. doi:10.1104/pp.111.184333
- Horton JD, Shimomura I, Brown MS, Hammer RE, Goldstein JL, Shimano H (1998) Activation of cholesterol synthesis in preference to fatty acid synthesis in liver and adipose tissue of transgenic mice overproducing sterol regulatory element-binding protein-2. *The Journal of clinical investigation* 101 (11):2331-2339. doi:10.1172/jci2961
- Kalscheuer R, Luftmann H, Steinbüchel A (2004) Synthesis of novel lipids in *Saccharomyces cerevisiae* by heterologous expression of an unspecific bacterial acyltransferase. *Applied and environmental microbiology* 70 (12):7119-7125
- Korkuc P, Schippers JH, Walther D (2014) Characterization and identification of cis-regulatory elements in arabidopsis based on single-nucleotide polymorphism information. *Plant Physiol* 164 (1):181-200. doi:10.1104/pp.113.229716
- Martino AD, Meichenin A, Shi J, Pan K, Bowler C (2007) Genetic and phenotypic characterization of *Phaeodactylum tricornutum* (Bacillariophyceae) accessions1. *Journal of Phycology* 43 (5):992-1009
- McGinn PJ, Morel FM (2008) Expression and inhibition of the carboxylating and decarboxylating enzymes in the photosynthetic C4 pathway of marine diatoms. *Plant Physiol* 146 (1):300-309. doi:10.1104/pp.107.110569
- Memelink J, Verpoorte R, Kijne JW (2001) ORCANization of jasmonate-responsive gene expression in alkaloid metabolism. *Trends in plant science* 6 (5):212-219
- Niu YF, Zhang MH, Li DW, Yang WD, Liu JS, Bai WB, Li HY (2013) Improvement of neutral lipid and polyunsaturated fatty acid biosynthesis by overexpressing a type 2 diacylglycerol acyltransferase in marine diatom *Phaeodactylum tricornutum*. *Mar Drugs* 11 (11):4558-4569. doi:10.3390/md11114558
- Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA (2008) A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic acids research* 36 (8):2547-2560. doi:10.1093/nar/gkn048
- Tran LM, Rizk ML, Liao JC (2008) Ensemble modeling of metabolic networks. *Biophysical journal* 95 (12):5606-5617
- Trentacoste EM, Shrestha RP, Smith SR, Gle C, Hartmann AC, Hildebrand M, Gerwick WH (2013) Metabolic engineering of lipid catabolism increases microalgal lipid accumulation without compromising growth. *Proceedings of the National Academy of Sciences of the United States of America* 110 (49):19748-19753. doi:10.1073/pnas.1309299110
- Weigel D, Ahn JH, Blazquez MA, Borevitz JO, Christensen SK, Fankhauser C, Ferrandiz C, Kardailsky I, Malanchruvil EJ, Neff MM, Nguyen JT, Sato S, Wang ZY, Xia Y, Dixon RA, Harrison MJ, Lamb CJ, Yanofsky MF, Chory J (2000) Activation tagging in *Arabidopsis*. *Plant Physiol* 122 (4):1003-1013
- Yamashita H, Takenoshita M, Sakurai M, Bruick RK, Henzel WJ, Shillinglaw W, Arnot D, Uyeda K (2001) A glucose-responsive transcription factor that regulates carbohydrate metabolism in the liver. *Proceedings of the National Academy of Sciences of the United States of America* 98 (16):9116-9121. doi:10.1073/pnas.161284298

Zheng Y, Quinn AH, Sriram G (2013) Experimental evidence and isotopomer analysis of mixotrophic glucose metabolism in the marine diatom *Phaeodactylum tricornutum*. *Microbial cell factories* 12:109. doi:10.1186/1475-2859-12-109

Chapter 8:

Acknowledgements

A PhD is not the work of one person, many hands helped me carry the load or dragged me along. I have to thank the following people for completing this work and apologize for any that I have forgotten (in no particular order).

The decision to work on diatoms was made when Prof. Wim Vyverman gave a guest lecture on algae. My enthusiasm after class was not shared by my fellow students. I am very grateful to him for introducing me to this wonderful world of diatoms, in the end it even got me a job after my PhD.

The other great force present during my PhD was Prof. Alain Goossens. Not being a diatom biologist he was standing on the side while I was peddling furiously but rudderless throughout most of my PhD. In the last years I managed to steer closer to familiar shores and I am grateful for the guidance provided by him and the ongoing international opportunities he provided. Without his expertise and the great group of people he built around him I would not have gotten half as far.

I would like to thank all members of the reading committee and the chairwoman: Prof. Dr. Sofie Goormachtig, Prof. Dr. Bart Devreese, Dr. Marc Heijde, Prof. Dr. Peter Kroth, Prof. Dr. Filip Rolland and Prof. Dr. Bartel Vanholme. Without your invaluable criticism, this work would be near to unreadable.

My fellow diatom traveler through most of my years was Michele Fabris. Thanks to him I learned a lot of nice Italian words which I cannot repeat here. Also thanks to him I am co-author on two great papers. We had a great time during our banishment in the far corner of the lab. The bond formed over all those empty gels, westerns and transformation plates will last a lifetime. I will remain ever grateful for the company during the all night sampling sessions and I know I will never beat you in FIFA, Michele.

I also express my sincerest gratitude to Sophie Carbonelle. Sophie, you provided invaluable technical assistance during your stay in the lab. Thanks for keeping things going practically while I was pouring over the RNA-seq data. We hit many roadblocks together but in the end you helped pave the way to results, it was a pity you were not there to see its conclusion. You will be delighted to know that at the lab I currently work, the miniprep yields are only a third of what you got. I hope you and Bruno build and inhabit that house soon and pray the electrical work will hold.

Marie Huysman, thanks for breaking the *Phaeodactylum* ground in the VIB. Your protocols made the first years a lot more streamlined than they otherwise would have been. I am happy to have collaborated on the Aureochrome paper and I am sure there is much more to follow. I always enjoyed our diatom conversations and wish you and Lasse many happy moments together.

Kenneth Goossens, we shared a great year together. It was strange not to be the most pessimistic person, we certainly learned a lot during this year and I hope you can apply some of this knowledge in your PhD. It was great to get to know you.

The 4:00 coffee club also has to be mentioned: Marlies, Jasper, Stefan, Hannes. I am glad to have met you guys during my studies and the extended period that followed. Since we were absolute nerds, we mainly talked science. The many conversations we have furthered my understanding and often solved problems. Our friendship extended beyond the workplace and I fondly remember many nights together. Extra thanks to Stefan for the many perl scripts and the fruitful collaboration. Special thanks to Hannes for helping me with the manuscript preparation, it was incredibly helpful. I would also like to ask Pieter where the other half of my Orval is.

When Michele left, the void was filled by Alex Antwerp, Karel, and Jean Martin. Respectively an irreverent parrot and an awesome scientist, a unique fin and finally the most friendly Walloon I have ever met. It was nice sharing the same corner and have all those (non) science discussions. A special thank you to Alex for reminding me twice daily that I was colorblind and therefore utterly handicapped. Gino Baart was the first diatomist to leave our group, I will remain grateful for the support during the IWT writing process. That's the beauty of science!

Jacob, the other gorgeous blonde in my life from 'het Meetjesland' and Tessa, the wizard of the East, were/are the accessible gurus of our lab. Thanks to you Jacob I cloned my first gene (full of point mutations) and thanks to you Tessa I did my first yeast transformation (where I was ultimately scooped by another group). I learned the ropes of the lab thanks to you guys, all the tips and tricks you taught me will be useful for the rest of my lab life. It makes me look like I know my stuff. Not to mention all the discussions we had about politics and the meaning of life.

The lab manager (or mayor?) Robin, also has my sincerest gratitude for helping me with the protoplast assays and keeping the lab in order. Too bad this part of the work was not as successful as hoped, but you helped me a lot along the way. Also a lot of thanks to Rebecca for completing the last bits of yeast work and being a great person in general. In alphabetical order I pay my respect to the Latin ladies of the lab: Amparro, Andres, Astrid, Patricia and Sabrina. I learned a lot of Spanish by diffusion and truly enjoyed the warm atmosphere you all brought to the lab. Although the young ones of the lab: Marie-Laure and Jonas were equally warm and awesome. Mein liebster Freund Phillip, unsere Gespräche, obwohl politisch nicht korrekt, waren auf jeden Fall einige der besseren die ich im der zweiten Hälfte des PhD hatte. Ihr Wissen über Hunde und Panther ist nur durch Ihre Liebe von ihnen erreicht. Janine, it is too bad you will not be able to read this in person but thank you for all the help (the awesome AFLP manual!) and being so friendly. Nathan, it was always good to realize that there is more in science than academia.

Laurens, I hope the Antwerp mobility problems get sorted out soon, genetically modified plants get accepted in the EU and/or high blood pressure medication makes rapid advances. Keep fighting the good fight. In similar light I would like to thank the local eco warriors Ruben and Christa, I truly enjoyed our conversations. I felt your disapproving glare after every wasted tip or briefly unbuttoned lab coat. In the same hallway another fearsome warrior took up residence in an office bigger and vastly more expensive

than that of Dirk and Jo Bury combined: Kris Morreel, thank you for shining a light on my metabolic ignorance and being an all out helpful and friendly guy. Dirk V.A, I am glad to have met you and I fully appreciate all those times you returned a dead centrifuge to life. I hope you find your path outside the VIB.

I did manage to leave the PSB from time to time and also there were a lot of friendly scientists. Chronologically I went first to KULAK under the no nonsense supervision of Prof. Imogen Foubert, although my PhD eventually went in a different direction. I am very grateful for the initial lipid analysis performed by Eline Ryckbosch and 'logistical' help from party crasher extraordinaire Dries Vandamme. I have a feeling we will share many drinks at the most unexpected congresses. I would also like to thank the entire group at the Protistology and Aquatic Ecology. Although I do not speak the ecology language, I often found essential tools and knowledge there.

I have to thank Prof. Alisdair Fernie and Dr. Toshihiro Obata for the warm welcome in snowy Golm. Thank you for the metabolite data, it is puzzling but very intriguing.

Langs deze weg wil ik ook graag mijn ouders bedanken voor hun steun over al die jaren. Omdat ze zoveel voor mij gedaan hebben, heb ik er moeite mee om een gepaste omschrijving te vinden. Tussen het ontdekken van de wetenschapsboeken in de kast van mijn vader, het in brand steken van de handdoeken van mijn moeder na een vetextractie met ether en dit doctoraat behalen, heb ik onnoemelijk veel op jullie gesteund. Bedankt.

Ze zijn pas veel later in mijn leven gekomen, maar ook een welgemeende merci aan Filip en Christine. Zonder jullie hulp had de verbouwing vast nog aangesleept en had de verhuis nog meer moeite gekost. Geen zorgen, we keren nog terug naar België.

Last but not least I have to thank the anchor in the storm and love of my life: Liesbeth. You were there from day one to support me and helped me to keep going the entire time. Thanks for putting up with me during my PhD, I promise never to do one again.

Chapter 9:

Appendices

Appendix 1: Curriculum Vitae

CONTACT INFORMATION

Name: Michiel Matthijs
Address: 134 Honey Hill Road, Bedford MK40 4PD, United Kingdom
Telephone: +32 (0)9 331 38 56
Email : mimat@psb.ugent.be

PERSONAL INFORMATION

Date of Birth: 29/08/1985
Place of Birth: Ghent
Citizenship: Belgian
Sex: Male

EDUCATION

University: Ghent University 2004-2009
Diploma: Master in Biochemistry/Biotechnology, major plant biotechnology
Thesis: 'The identification of novel oncogenes and tumor suppressors through a comparative analysis of cell cycle correlated genes in *Arabidopsis thaliana* and mammals' under supervision of Prof. dr. Lieven De Veylder
Grade: Distinction

LANGUAGE PROFICIENCIES

Dutch: Native speaker
English: Fluent both written and orally, used on a daily basis
French: Fluent orally, passive written skills
German: Basic, 4 years of lessons in secondary school and one after school course
Arabic: Beginner

CURRENT EMPLOYMENT

Research scientist at Algeniuty (July 2014 - Current)

EMPLOYMENT HISTORY

Academic Positions: PhD student VIB/Ghent University (August 2009-June 2014).
Supervisors: Prof. dr. Alain Goossens (VIB Ghent)/Prof dr. Wim Vyverman (UGent)
PhD Thesis title: *A first look at the genetic control mechanisms shaping metabolic changes during nitrogen starvation in Phaeodactylum tricornutum*
Grant: Personal grant provided by the Flanders Institute for Research and Innovation (IWT)

RESEARCH EXPERIENCE

- Algal Culturing, Photobioreactor operation
- PCR, cloning, vector construction and protein expression in *S. cerevisiae*, *E. coli* and *Phaeodactylum tricornutum*
- Yeast complementation assays and interaction screening (Yeast one/two Hybrid)
- Flow cytometry for cell cycle analysis
- GC/FID and GC/MS for metabolite analysis, mainly FAME's
- RNA-seq analysis
- Working knowledge of R, Linux, Python and Perl

SUPERVISION

Supervised one master thesis, one professional bachelor, two master students and 7 bachelorstudents. Supervising one lab technician.

ADDITIONAL TRAINING

- 2009: Perl Training Course (BITS VIB)

- 2010: Advanced Statistics (Doctoral Schools)
- 2011: ANOVA (Doctoral Schools)
- 2012: Basic R course (Doctoral Schools)

CONFERENCES

Speaker at 'International Diatom Symposium', 2012, Ghent, Belgium

Speaker at TEDxGhent: 'Algal lipid engineering', 2012, Ghent, Belgium

Speaker at 'Young Algaeneers Symposium' Wageningen, 2012, Netherlands

Speaker at 'The molecular life of diatoms EMBO conference', 2013, Paris, France

Speaker at '4th International Bielefeld-CeBiTec Research Conference: Prospects and challenges for the development of algal biotechnology', 2014, Bielefeld, Germany

PUBLICATIONS

1. Identification of putative cancer genes through data integration and comparative genomics between plants and humans.

Quimbaya M, Vandepoele K, Raspé E, Matthijs M, Dhondt S, Beemster GT, Berx G, De Veylder L.

Cell Mol Life Sci. 2012 Jan 5

2. The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff glycolytic pathway

Michele Fabris, Michiel Matthijs, Stephane Rombauts, Wim Vyverman, Alain Goossens and Gino J.E. Baart

The Plant Journal Nov 2012

3. AUREOCHROME1a-Mediated Induction of the Diatom-Specific Cyclin dsCYC2 Controls the Onset of Cell Division in Diatoms (*Phaeodactylum tricornutum*)

Huysman MJ, Fortunato AE, Matthijs M, Costa BS, Vanderhaeghen R, Van den Daele H, Sachse M, Inzé D, Bowler C, Kroth PG, Wilhelm C, Falciatore A, Vyverman W, De Veylder L

Plant Cell. Jan 2013

4. Tracking the sterol biosynthesis pathway of the diatom *Phaeodactylum tricornutum*.

Fabris M, Matthijs M, Carbonelle S, Moses T, Pollier J, Dasseville R, Baart GJ, Vyverman W, Goossens A.

New Phytol. 2014 July

Appendix 2: Primers Used

Cloning primers for NMB	
FW_NMB1	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGGCGACTATGAACGGCTTC
RV_NMB1	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMGTTTGATACAATCGACATC
RV_NMB1_1stpart	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMTAAATCTTCGCCATCTC
FW_NMB1_2ndpart	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGGATAGCAGACTTGTCG
FW_NMB2_B1	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGGATCGTCAAAGCAATAATG
RV_NMB2_44641	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMACCAAACCTCGTCCACAATGCTG
FW_NMB1_RD	TCCGAGTCAGCTCCTGTGGAGCTACAAGCGCTGCCCCGTGCCATCCAACTCAACCCTCACGG
RV_NMB1_RD	CCGTGAGGGTTGAGTTGGATGGCACGGGCAGCGCTTGTAGCTCCACAGGAGCTGACTCGGA
QPCR primers for NMB	
FW_Q_NMB1	TGCGTCTCCTTCTTCTCGT
RV_Q_NMB1	CAACCCTGTGTTTGTCTGTTG
FW_Q_NMB2	TCCCGATCAAAGCAAATCTC
RV_Q_NMB2	CAAACCTCGTCCACAATGCTG
Cloning primers for selected transcription factors	
FW_AP2-EREBP2_45659	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGCAAGTTGGAGGCAAACAATTTC
RV_AP2-EREBP2_45659	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMTCTTTGTGATCAGGACAGATA
FW_bHLH1a_PAS_44962	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGAATAAGCCAGGACAGCGGGGAA
RV_bHLH1a_PAS_44962	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMCAGCTTCGCTGCATCGTCTGAT
FW_bHLH3_42586	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGGACGATAACGACGATATCGACT
RV_bHLH3_42586	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMTCTTCTCACATCTAAGGACTTG
FW_bZIP12_42560	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGTCGCCGACGGCCTACGGAACTA
RV_bZIP12_42560	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMATTTCTATCTCTATTGGAAATT
FW_bZIP14_45314	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGCGATTTCCAGCCGCTTCAGTGG
RV_bZIP14_45314	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMCCAACCCACAGGGGCTGCAAGA
FW_bZIP19_48701	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGAACCAAGCTGCGTCTTCGACAA
RV_bZIP19_48701	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMAGGTACCCGGGCCTTTCTGTTCA
FW_CBF_NF5_46740	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGGCAACCGTTAGGTCGTCGAACC
RV_CBF_NF5_46740	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMCTCCGGTTGCGGGACCGCCGCA
FW_CCCH11_50345	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGTCTTCTGGTTCTTCTTTCCC

RV_CCCH11_50345	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMGGGTATAAAGAGACCGATATCA
FW_CCCH8_44042	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGCCAACAGCGAAAAAGTACAAGC
RV_CCCH8_44042	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMGTCCGACGTCGTCAAAGCACGA
FW_CCHH11_38018	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGGCAACGCCAGAGCCGTACTTTG
RV_CCHH11_38018	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMTCTCCGCTTTTGCTTGAATGGT
FW_CCHH5_5508	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGCCCCAAGCGGAGAAGGGTTTCGC
RV_CCHH5_5508	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMCGCAATTTTAGAGAAATCTTCG
FW_fungal_TRF_50045	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGAACGAAGACAGTACCGAAGACC
RV_fungal_TRF_50045	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMACTTGAAATGAGATTGCTAATC
FW_HSF1c_AP255108	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGCACAAGTTACCCTACAGGGACT
RV_HSF1c_AP255108	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMGAGCAGAGATTCACCTTCAGCG
FW_HSF1g_42514	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGACGACTACTTCATCGAAGCGTA
RV_HSF1g_42514	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMCTCCAAGACCTTCGGGCCAAG
FW_HSF4_5b_44684	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGGTAATCGAGGGAAACGCAACTG
RV_HSF4_5b_44684	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMCTTGTCGTGGACGCGAACGGG
FW_Myb1R_46535	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGTTTCACGAAAATCTGAACGATT
RV_Myb1R_46535	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMAGATTCAGACTCGAGAAAGAGA
FW_HMG_39045	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGGTAAAGCCTCTCACC
RV_HMG_39045	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMTTCCTTGCCAAGTTCGTC
FW_TFIIB_37556	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGTCCCCTGGGAGCAT
RV_TFIIB_37556	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMCAATTTAGAAGCTTTGAGCAAGG
FW_WHTH_43731	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGAAAGATTTTACAGTAAATCC
RV_WHTH_43731	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMTGCTTTTGAGAAACTAGAATG
FW_HSF4,3b	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCatggacaacaggcctgcacc
RV_HSF4,3b	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMgatcttaccagatacggtcgagttg
FW_HSf_4.1b	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCatgaacattagccaacttaacgaag
RV_HSf_4.1b	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMatcctcggtttcatctccttgggatc
FW_MybB_37257	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCatgaatgatttagttcggaagttctcttc
RV_MybB_37257	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMtcttcaaacgagtctacctcttgatc
Cloning Primers for TEA promoter fragments	
FW_prom_MD_42398	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCGTACAACAAAGTAAGTCGTTTC
RV_prom_MD_42398	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMGAAGGATGAGGGAAAAGG

FW_prom_CS_30145	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCCATGCTAAGTGTAAGTTGACTTG
RV_prom_CS_30145	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMGGTTAATGATGATAACAGTGACTG
QPCR primers for transcription factors transformed into <i>P. tricornutum</i>	
RV_Q_HSF4_1b	CAACAACACGCTCAGCTTCT
FW_Q_HSF4_1b	TTCCACCAAATGCACTTCGG
FW_Q_WHTH	CCATTGCCAACGGTATAGCC
RV_Q_WHTH	GCCCAAATCAACAACCGACT
Q_MYBb_FW	TGGCCCTGTTTCGATAAAGTC
Q_MYBb_RV	TGCGACCATTGTAGCTTGAC
Q_TRF_FW	GCCGGTGTACCTGAATAATG
Q_TRF_RV	GAAATGGATTTCGGAAGGAAT
RV_Q_HSF4_3	ATTTCTGGCCTTCACCACAC
FW_Q_HSF4_3	TGGAAGGGGTTGTATTCGAG
Q_FW_HMG	CTTCGCAAATCACCAAAGCA
Q_RV_HMG	GCATCAGCAAACACCGCATT
FW_Q_HSF1g	GCTGGAGAACGACGACGTAA
RV_Q_HSF1g	TTTGACGACCCTTGCCAACT
FW_Q_Bzip14	agaaatggccgaaatgtgtc
RV_Q_Bzip14	gaagttggctcggactcaag
QPCR primers for TCA cycle genes	
FW_26290_Q_AH	ACTGAAAGAACCCGTGTTGG
RV_26290_Q_AH	CGTCCATCTTGGTAGGAGGA
FW_20934_Q_IsoCitDehy	TGGATTCGACATGAACCTGA
RV_20934_Q_IsoCitDehy	TCTTTGCGCATTTTCAACAG
FW_29016_Q_OGD	CACATTTGGAATGCGTCAAC
RV_29016_Q_OGD	GTTGGTGGTGAAGCCAATCT
FW_42015_Q_SCSa	ATTAAACCCGGTGAATGCAA
RV_42015_Q_SCSa	GGTGAAACGTTCCAAGCAAT
FW_26921_Q_SCSa2	AAGCCATCCTCGTCAACATC
RV_26921_Q_SCSa2	GCATCTTCCAGATCCTCTGC
FW_52539_Q_SDH2	CGCTCGCTGTCTATCTAC
RV_52539_Q_SDH2	ATACATTCCGTCGAGCTTGG

FW_18516_Q_SDHcb	CGCTGTCGCCAGTATTACAA
RV_18516_Q_SDHcb	TCAAACCTCCCCATAATGA
FW_41812_Q_SDH1	CCAACCAGGACAACGAGTTT
RV_41812_Q_SDH1	TGAAACGACGTCTCTCGATG
FW_36139_Q_FUM	ACGAGCGAACTCTACCAGGA
RV_36139_Q_FUM	CCCAAGGCTAAACGGTACAA
FW_19708_Q_FUM1	GGTTACCGGACTACGCAAAA
RV_19708_Q_FUM1	CTTTTGAAGCATTCGTCA
FW_Q_Urease	GACAGGATTCGGTTGCTGAT
RV_Q_Urease	GCCGAGACAGGGTTGTGTAT
FW_Q_CS_30145	TCCTGAAATGGATCCAGGAG
RV_Q_CS_30145	AAAATGTCGGGCATGACTTC
FW_Q_MD_42398	GGTAGCTGCGGATCTCAGTC
RV_Q_MD_42398	CAGCGTTGGTGTGAAGAGA
QPCR primers for reference genes	
Q_M_FW_RP3A	AAAGAGCATGCCAAGTGGTG
Q_M_RV_RP3A	TCTACAGCTCGAATGTCCCC
Q_M_FW_PUA	CTGGATTGACAAACCCTGGC
Q_M_RV_PUA	TCCCCCTCACTCCAAACTG
Q_L_FW_VTC4	GGCACATTTGCGCTACGATT
Q_L_RV_VTC4	TATCCTCGTGTACGTGTCCG
Primers for synthesis of motif17	
NoN17_P1_F1	GTTTCGCGTTGCTTTGCTGGAGTGggaatttccggcagtcctgatgt
NoN17_P1_R1	ACACAAAATTCCCAAACCTGCACGATTcacatcaggcactgccggaattcc
NoN17_P1_F2	GAATCGTGCAGGTTTGGGAATTTTGTGTcgtcgtcgcaacaactcttaggtaa
NoN17_P1_R2	AAGAGGAAGTATCCTACAAAGCAAGTTCCGttacctacaagagtgtgtgcgacgacg
NoN17_P1_F3	CGGAACTTGCTTTGTAGGATACTTCCTCTTgtgtcgggtgcggaagtgttagattctt
NoN17_B1	GGGGACAAGTTTGTACAAAAAGCAGGCTCCGTTTCGCGTTGCTTTGCTGGAG
NoN17_B2	GGGGACCACTTTGTACAAGAAAGCTGGGTCAAGAATCTACAACTCCGACACCGAC
NoN17_B4	GGGGACAACCTTTGTATAGAAAAGTTGAAGTTCGCGTTGCTTTGCTGGAG
NoN17_B1r	GGGGACTGCTTTTTGTACAAACTTGAAGAATCTACAACTCCGACACCGAC

Primers for synthesis of motif6	
NoNdown4_P1_F1	CTTTCAACCACACCATGAAGTTTTCCCTCACgctgggtcagctctcatcatgaagttttcc
NoNdown4_P1_R1	TTCATGGTGAGGAGGGACTTGTTGCAATggaaaacttcatgatgagagctgaaccagc
NoNdown4_P1_F2	ATTGCAACAAGTCCCTCCTCACCATGAAggtcgtaccacgctaaccct
NoNdown4_P1_R2	AGCGAGAATGGCAAACCTTCATCGTGAAAGaggggtagcgtggtagcgacc
NoNdown4_B1	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCCTTTCAACCACACCATGAAGTTTTCCC
NoNdown4_B2	GGGGACCACTTTGTACAAGAAAGCTGGGTCAGCGAGAATGGCAAACCTTCATCGT
NoNdown4_B4	GGGGACAACCTTTGTATAGAAAAGTTGAACTTTCAACCACACCATGAAGTTTTCCC
NoNdown4_B1r	GGGGACTGCTTTTTTTGTACAAACTTGGAGCGAGAATGGCAAACCTTCATCGT
Primers for synthesis of motif4	
NoN6_P1_F1	GGTCGGCCATTGGAATTCTGGTACGgcctgcctccaggttggtgaa
NoN6_P1_R1	ACGGGGGTAACGCCCCAGAAAttccacaacctggaggcaggc
NoN6_P1_F2	TTCTGGGGCGTTACCCCCGTggaacgtgaaattctggtgtttgtaattggt
NoN6_P1_R2	ATCGGTTGTTCCAGAATTCCGAATCGAaccaattacaaacaccagaatttcacgttcc
NoN6_B1	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCGGTCGGCCATTGGAATTCTGG
NoN6_B2	GGGGACCACTTTGTACAAGAAAGCTGGGTCATCGGTTGTTCCAGAATTCCGAATC
NoN6_B4	GGGGACCACTTTGTACAAGAAAGCTGGGTCATCGGTTGTTCCAGAATTCCGAATC
NoN6_B1r	GGGGACTGCTTTTTTTGTACAAACTTGGATCGGTTGTTCCAGAATTCCGAATC

Appendix 3: other papers by the author

Author contributions

1. Tracking the sterol biosynthesis pathway of the diatom *Phaeodactylum tricornutum*

MM helped in the identification of fusion enzymes and performed several experiments

2. AUREOCHROME1a-mediated induction of the diatom-specific cyclin dsCYC2 controls the onset of cell division in diatoms (*Phaeodactylum tricornutum*)

MM conceived the cycloheximide treatment control and helped in performing experiments

3. The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff glycolytic pathway

MM provided bioinformatic input and performed experiments

4. Identification of putative cancer genes through data integration and comparative genomics between plants and humans

MM provided practical assistance for several experiments

Tracking the sterol biosynthesis pathway of the diatom *Phaeodactylum tricornutum*

Michele Fabris^{1,2,3}, Michiel Matthijs^{1,2,3}, Sophie Carbonelle^{1,2}, Tessa Moses^{1,2}, Jacob Pollier^{1,2}, Renaat Dasseville³, Gino J. E. Baart^{1,2,3}, Wim Vyverman³ and Alain Goossens^{1,2}

¹Department of Plant Systems Biology, VIB Technologiepark 927, B-9052 Gent, Belgium; ²Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, B-9052, Gent, Belgium; ³Department of Biology, Laboratory of Protistology and Aquatic Ecology, Ghent University, Krijgslaan 281 (S8), B-9000, Gent, Belgium

Author for correspondence:
Alain Goossens
Tel: +32 9 3313851
Email: alain.goossens@psb.vib-ugent.be

Received: 10 December 2013
Accepted: 2 June 2014

New Phytologist (2014)
doi: 10.1111/nph.12917

Key words: chimeric pathway, diatom, fusion enzymes, isopentenyl diphosphate isomerase, oxidosqualene cyclase, *Phaeodactylum tricornutum*, squalene epoxidase, sterol biosynthesis.

Summary

- Diatoms are unicellular photosynthetic microalgae that play a major role in global primary production and aquatic biogeochemical cycling. Endosymbiotic events and recurrent gene transfers uniquely shaped the genome of diatoms, which contains features from several domains of life. The biosynthesis pathways of sterols, essential compounds in all eukaryotic cells, and many of the enzymes involved are evolutionarily conserved in eukaryotes. Although well characterized in most eukaryotes, the pathway leading to sterol biosynthesis in diatoms has remained hitherto unidentified.
- Through the DiatomCyc database we reconstructed the mevalonate and sterol biosynthetic pathways of the model diatom *Phaeodactylum tricornutum* *in silico*. We experimentally verified the predicted pathways using enzyme inhibitor, gene silencing and heterologous gene expression approaches.
- Our analysis revealed a peculiar, chimeric organization of the diatom sterol biosynthesis pathway, which possesses features of both plant and fungal pathways. Strikingly, it lacks a conventional squalene epoxidase and utilizes an extended oxidosqualene cyclase and a multifunctional isopentenyl diphosphate isomerase/squalene synthase enzyme.
- The reconstruction of the *P. tricornutum* sterol pathway underscores the metabolic plasticity of diatoms and offers important insights for the engineering of diatoms for sustainable production of biofuels and high-value chemicals.

Introduction

Sterols are fundamental terpenoids in eukaryotes. They are important components of the plasma membrane, play relevant roles in cellular defence and signalling, and are precursors of several hormones and bioactive secondary metabolites (Adolph *et al.*, 2004; Benveniste, 2004; Dufourc, 2008; Vinci *et al.*, 2008; Galea & Brown, 2009; Tomazic *et al.*, 2011). The ability to synthesize sterols is a common feature of eukaryotes, with rare exceptions represented by some insects, nematodes and oomycete plant pathogens, such as *Phytophthora* spp. (Desmond & Gribaldo, 2009; Gaulin *et al.*, 2010). A few examples of bacteria with a minimal sterol pathway have been reported, although most prokaryotes synthesize hopanoids, structurally similar compounds that do not incorporate oxygen in position C-3 (Pearson *et al.*, 2003; Lamb *et al.*, 2007). Therefore, being deeply rooted in the early history of eukaryotic life, the sterol biosynthesis pathway can be considered as a prime example of metabolic evolutionary conservation. It is believed that the Last Eukaryotic Common Ancestor (LECA) already possessed some

of the metabolic enzymes of the present sterol pathway (Desmond & Gribaldo, 2009). The presumed presence of a primitive form of this pathway in the LECA is reflected by the conservation of many aspects of the pathway in all sterol-producing organisms.

Originally classified into three main variants yielding cholesterol in animals, ergosterol in fungi and diverse phytosterols in land plants (Fig. 1), it is now becoming evident that this subdivision oversimplifies the actual organization of the sterol biosynthesis pathways. The recent increase in the number of sequenced nonmodel organism genomes, and technical advances in genomics and metabolomics triggered a re-evaluation of the organization of sterol biochemistry in fungi (Weete *et al.*, 2010), green algae (Massé *et al.*, 2004; Miller *et al.*, 2012), choanozoa (Kodner & Summons, 2008), kinetoplastids (Nes *et al.*, 2012), dinoflagellates (Leblond & Lasiter, 2012) and even land plants, in which an alternative branch of the pathway has been discovered in *Arabidopsis thaliana* (Ohya *et al.*, 2009).

Despite the diversity in the end products of the sterol biosynthesis pathway, many upstream reactions and intermediates are

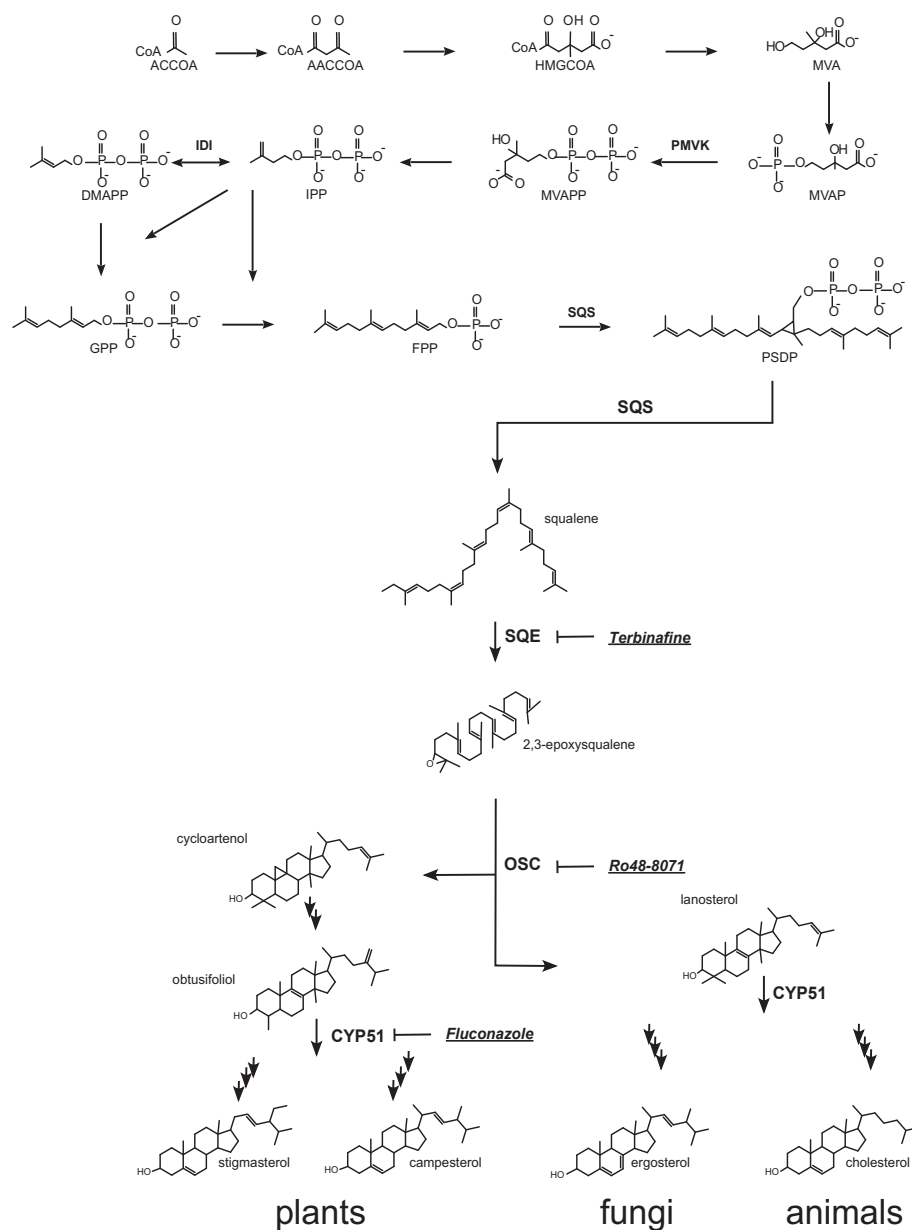


Fig. 1 Conserved reactions of the mevalonate (MVA) and sterol biosynthesis pathways. Enzymes described in the text are highlighted in bold. Chemical enzyme inhibitors are indicated in *italics*, bold and underlined font. IDI, isopentenyl diphosphate isomerase; PMVK, phosphomevalonate kinase; SQS, squalene synthase; SQE, squalene epoxidase; OSC, oxidosqualene cyclase; CYP51, cytochrome P450 sterol-14-demethylase; ACCoA, acetyl-CoA; AACCoA, aceto-acetyl-CoA; HMGCoA, 3-hydroxy-3-methylglutaryl-coenzyme A; MVA, mevalonate; MVAP, mevalonate phosphate; MVAPP, mevalonate diphosphate; IPP, isopentenyl diphosphate; DMAPP, dimethylallyl diphosphate; GPP, geranyl diphosphate; FPP, farnesyl diphosphate; PSDP, presqualene diphosphate.

ubiquitously conserved in the different taxonomical groups (Fig. 1). Land plants and several photosynthetic organisms use two distinct parallel pathways for the synthesis of terpenoid precursors. Generally, the sterol precursors are produced through the cytosolic mevalonate (MVA) pathway, whereas precursors of, for example, carotenoids are synthesized through the plastidic methylerythritol phosphate (MEP) pathway. In green algae and some members of red algae, however, the sterol building blocks are provided by the MEP pathway because the MVA pathway was lost in these organisms (Massé *et al.*, 2004; Lohr *et al.*, 2012). Isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) are isomeric intermediates of both pathways. In the MVA pathway, IPP is formed from mevalonate diphosphate (MVAPP) and isomerized to DMAPP by isopentenyl diphosphate isomerase (**IDI**). The condensation of IPP and DMAPP produces geranyl diphosphate (GPP), which forms farnesyl

diphosphate (FPP) by condensing with another molecule of IPP. FPP is the substrate of squalene synthase, which initiates sterol biosynthesis. The squalene in prokaryotes is cyclized to hopanoids by squalene-hopanoid cyclases (SHCs), whereas in eukaryotes it is first epoxidized by squalene epoxidase (**SQE**) before cyclization by an oxidosqualene cyclase (**OSC**). The cyclization of 2,3-epoxysqualene to lanosterol in animals and fungi, and cycloartenol in plants and green algae is catalysed by distinct **OSCs**: lanosterol synthase (LAS) and cycloartenol synthase (CAS), respectively. The next conserved reaction is catalysed by a cytochrome P450 (P450), sterol-14-demethylase (**CYP51**), which removes a methyl group from lanosterol in animals and fungi, and from obtusifoliol, a product of cycloartenol conversions, in the green lineage (i.e. the Viridiplantae that include land plants and green algae). After this reaction, the synthesis of sterols becomes more specific and varies depending on the phylogeny.

Contrary to the well understood biochemistry of sterols in animals, plants and fungi, our knowledge of other organisms, and in particular the many groups of unicellular eukaryotes, remains fragmentary. Although efforts to chemically characterize the sterol composition of several diatom species have revealed a marked diversity in products (Volkman, 2003; Rampen *et al.*, 2010; Giner & Wikfors, 2011), the diatom sterol biosynthesis pathway remains unknown. Some diatom sterols seem to derive from cycloartenol, like phytosterols, whereas others originate from lanosterol, making the collocation of the pathway in the tripartite subdivision difficult (Rampen *et al.*, 2010). Diatoms belong to the same kingdom as the oomycetes, one of the rare eukaryotic groups that have lost the ability to synthesize sterols during their evolution (Gaulin *et al.*, 2010). Therefore, the enzymatic and genetic study of sterol biosynthesis in diatoms is fascinating, from both evolutionary and ecological perspectives.

Here, we report the reconstruction of the MVA and sterol biosynthesis pathways of the model pennate diatom *Phaeodactylum tricornutum*. Using DiatomCyc (www.diatomcyc.org; Fabris *et al.*, 2012) to identify the set of genes putatively involved in diatom sterol synthesis, we mapped and experimentally validated the main biochemical steps of the pathway. The proposed biochemical route is a hybrid pathway that shares elements with fungal and plant pathways. We report that diatoms, and many other groups of marine organisms, surprisingly lack the ubiquitously conserved *SQE* and harbour uncommon enzymes, including an extended *OSC* and an *IDI-SQS* multifunctional fusion enzyme.

Materials and Methods

Chemicals

Squalene, 2,3-epoxysqualene, cycloartenol, lanosterol, ergosterol, fenpropimorph, fluconazole, imidazole and terbinafine were purchased from Sigma-Aldrich, and Ro 48-8071 from Cayman Chemicals (Ann Arbor, MI, USA).

Generation of plasmid vectors

Full-length (FL) genes were amplified from a cDNA library of *Phaeodactylum tricornutum* (Huysman *et al.*, 2013) with PrimeSTAR[®] HS DNA Polymerase (Takara Bio, Ōtsu, Japan), Gateway[™] cloned into pDONR221 (Invitrogen) and sequence verified. PCR primers (Table S1) were designed with Vector NTI (Invitrogen). Destination vectors were generated with the Gateway[™] cloning technology, unless specified otherwise.

For heterologous expression in *Escherichia coli*, pDEST17 (Invitrogen) was used. *PrOSC* was fused to Maltose Binding Protein (MBP) at its N-terminus by cloning into an in-house modified pDEST17 (pDEST17-MBP). For the co-expression of *PrIDISQS* and *SQE* candidate genes, the *PrIDISQS* cassette flanked by the T7 promoter and terminator was amplified with *SmaI* and *SpeI* containing primers and cloned into the backbone of pDONR223, *de novo* amplified by PCR to introduce compatible restriction sites and remove the Gateway cassette.

For heterologous expression in *Saccharomyces cerevisiae* W303, *PHATRDRAFT_51757* and *PHATRDRAFT_30461* were cloned into pAG426GAL-ccdB and pAG423GAL-ccdB (Alberti *et al.*, 2007), respectively. For heterologous expression in *S. cerevisiae* strain TM5 (Moses *et al.*, 2014), *PrOSC* was cloned into pAG423GAL-ccdB (Alberti *et al.*, 2007).

Bacterial and yeast culturing

E. coli BL21 (DE3) (Invitrogen) transformants were grown in Luria-Bertani (LB) broth supplemented with 50 mg l⁻¹ chloramphenicol and 100 mg l⁻¹ carbenicillin at 37°C until A_{600nm} of 0.6. Gene expression was induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), followed by a 48-h incubation at 20°C in darkness. Cells were harvested and lyophilized for HPLC analysis.

S. cerevisiae transformants were selected on SD medium (Clontech) containing appropriate drop out supplements (Clontech Laboratories, Mountain View, CA, USA). W303 derived strains were grown overnight in liquid medium at 30°C with shaking. Gene expression was induced by inoculating washed precultures in SD GAL/RAF medium (Clontech) containing appropriate drop out supplements and incubated for 2 d at 30°C. TM5 derived strains were cultured as described (Moses *et al.*, 2014).

Protein extraction and immunoblot analysis

Cells from a 200-ml *E. coli* culture were resuspended in phosphate buffered saline and lysed by sonication for 1 min with 10-s pulses using a Heat Systems Ultrasonics sonicator (Heat Systems Inc., Newtown, CT, USA). Cell lysates were directly used for *in vitro* enzymatic assays or separated from cell debris by centrifugation at 10 000 g for 20 min. Protein purification was performed using Ni-NTA Superflow resin (Qiagen). Immunoblot analysis was carried out with Mini-PROTEAN[®] Precast Gels and related equipment (Bio-Rad), anti-His Antibody Selector Kit (Qiagen), HRP-linked anti-mouse antibody (GE Healthcare, Dienen, Belgium) and the Clarity[™] Western ECL Substrate (Bio-Rad).

Treatments with chemical inhibitors

Three-day-old *P. tricornutum* CCAP 1055/1 cultures grown in ESAW medium (Berges *et al.*, 2001) at 21°C in continuous light (average intensity 75 μmol photons m⁻² s⁻¹), were treated in triplicate with terbinafine, fluconazole, Ro 48-8071, fenpropimorph or imidazole for 48 h. Samples were harvested 1, 2, 4, 6, 8, 12, 24 and/or 48 h after treatment. Mock treatments were performed with the corresponding solvents: dimethyl sulfoxide for terbinafine and fenpropimorph, methyl acetate for Ro 48-8071, 10% ethanol for fluconazole, and water for imidazole.

Sterol extraction and gas chromatography – mass spectrometry (GC-MS) analysis

Cells from 50 ml of *P. tricornutum*, 2 ml of *E. coli*, 1 ml of *S. cerevisiae* W303 or 12 ml of *S. cerevisiae* TM5 cultures were collected and snap frozen. Cells were lysed by incubation for

10 min at 95°C in equal volumes (250 µl) of 40% KOH and 50% ethanol. Extraction was achieved by adding 900 µl hexane to the lysate and collecting the organic phase. This procedure was repeated two times. The pooled organic fractions were evaporated and derivatised with 20 µl pyridine (Sigma-Aldrich) and 100 µl *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (Sigma-Aldrich). Authentic standards were dissolved in hexane, evaporated and derivatised similarly.

GC-MS analysis was performed with the GC model 6890 and MS model 5973 (Agilent, Santa Clara, CA, USA). One microlitre of sample was injected in splitless mode into a VF-5ms capillary column (Varian CP9013; Agilent). Helium carrier gas was set at a constant flow of 1 ml min⁻¹. The injector temperature was set to 280°C and the oven temperature was programmed as follows: 80°C for 1 min post injection; ramped to 280°C at 20°C min⁻¹, held at 280°C for 45 min, ramped to 320°C at 20°C min⁻¹, held at 320°C for 1 min and cooled to 80°C at 50°C min⁻¹ at the end of the run. The MS transfer line was set to 250°C, the MS ion source to 230°C, and the quadrupole to 150°C, throughout. Full EI-MS spectra were generated by scanning the *m/z* range of 60–800 with a solvent delay of 7.8 min.

Extraction and quantification of lycopene

Aliquots of lyophilized *E. coli* samples were weighed (0.1 mg accuracy) and extracted with acetone : water, 90 : 10 (v/v) and sonicated with a tip sonicator at 40 W for 30 s, 2-s pulses. Extracts were filtered through a 0.2-µm Alltech[®] nylon syringe filter (Thermo Fisher Scientific, Waltham, MA, USA) to remove cell debris and injected into an Agilent 1100 series HPLC system equipped with a Grace[®] reverse phase Eclipse XDB C₁₈ column (150 × 4.6 mm; 3.5 µm). Lycopene was analysed as described (Van Heukelem & Thomas, 2001) using two solvents: solvent A, 70 : 30 (v/v) methanol, 28 mM aqueous *N*-tert-butylacrylamide (TBAA), pH 6.5; solvent B, methanol. Lycopene was identified by comparing retention times and absorption spectra, and quantified by calculating response factors, using pure lycopene standards (DHI, Hørsholm, Denmark).

Gene silencing in *P. tricornutum*

The *PtOSC* RNA interference (RNAi) construct was prepared as described elsewhere (De Riso *et al.*, 2009) using the primers listed in Supporting Information Table S1. *P. tricornutum* cells were transformed with *PtOSC*-RNAi or empty pAF6 vectors by biolistic transformation as reported (Falcioro *et al.*, 1999) with a PDS-1000/HeTM System (Bio-Rad). Transformants were selected and cultivated on ESAW medium containing 100 µg ml⁻¹ phleomycin.

Nile Red staining and fluorescence microscopy

One millilitre of *P. tricornutum* culture was stained with 5 µl Nile Red (9-(diethylamino)-5H benzo [α] phenoxazin-5-one) (from a stock solution of 2 mg ml⁻¹ in acetone) and

incubated for 15 min in the dark. Fluorescence microscopy images were acquired with a Zeiss Axio Imager.M2m microscope (Carl Zeiss, Germany) as described previously (Green-span *et al.*, 1985).

Quantitative real-time PCR (qRT-PCR) analysis

cDNA was generated with iScript (Bio-Rad) from RNA extracted with RNeasy (Qiagen) from *P. tricornutum* cultures. qRT-PCR was carried out with a Lightcycler 480 (Roche) and SYBR Green QPCR Master Mix (Stratagene, La Jolla, CA, USA). *Histone H4* (*H4*) and *Tubulin β chain* (*TubB*) were used as the reference genes for normalisation (Siaut *et al.*, 2007). Primers for amplification of *PtOSC* (Table S1) were designed with Beacon Designer (Premier Biosoft; www.premierbiosoft.com).

Results

P. tricornutum cultures accumulate the C-28 sterols brassicasterol and campesterol

In order to profile the pool of sterols produced by *P. tricornutum*, we analysed 3-d-old cultures with GC-MS. *P. tricornutum* only accumulated C-28 sterols, mainly brassicasterol (24-methylcholest-5,22-dien-3β-ol) and campesterol (24-methylcholest-5-en-3β-ol) (Fig. 2), in agreement with earlier reports (Rampen *et al.*, 2010).

In silico reconstruction of the *P. tricornutum* sterol pathway

We mined DiatomCyc (www.diatomcyc.org; Fabris *et al.*, 2012) to retrieve the set of genes putatively encoding the biosynthetic enzymes involved in MVA and sterol biosynthesis in *P. tricornutum*. The identified genes, listed in Table 1, generally show homology with orthologues from plants, yeast, algae and animals, with high orthology scores (Table S2). Most genes lacked either a correct start or stop codon or both, thus their gene models were manually corrected with the aid of in-house available RNA-Seq data and uploaded in DiatomCyc v2.0 (Fabris *et al.*, 2012).

We identified the whole set of genes encoding the enzymes involved in the MVA pathway (Fig. 3). Many of these genes show pronounced similarity with plant orthologues (Tables 1, S2), in accordance with the fact that this pathway also provides the sterol precursors in plants, but not in green algae, in which it is lost. Initially, a phosphomevalonate kinase (PMVK) could not be identified in *P. tricornutum* by the orthology prediction method (Fabris *et al.*, 2012), or by BLAST searches on the NCBI and JGI databases (<http://genome.jgi-psf.org/Phatr2/Phatr2.home.html>), whereas we could find corresponding orthologues in all other sequenced diatom genomes. By performing BLAST searches with Arabidopsis *PMVK* (*At1g31910*) as the query sequence and public and in-house generated *P. tricornutum* transcript sequences, we identified a 1395-nt transcript encoding the missing PMVK. The novel *PtPMVK* transcript could be mapped, in reverse

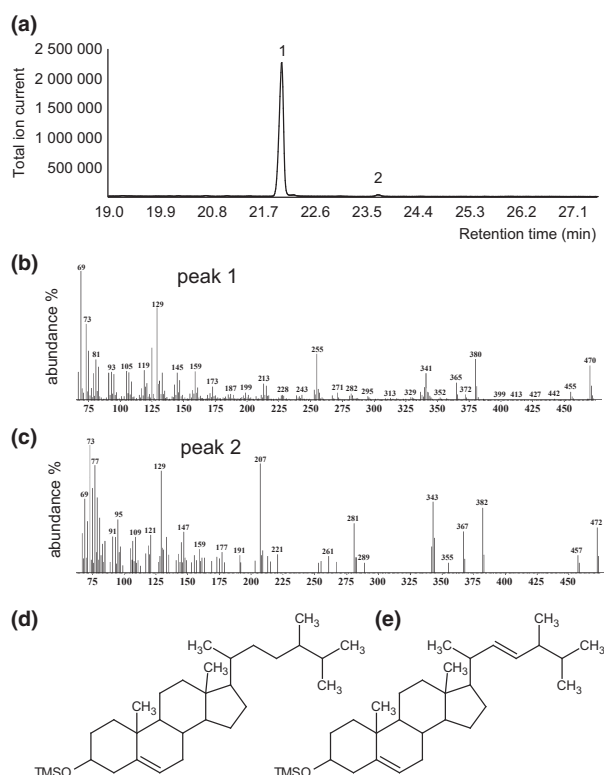


Fig. 2 *Phaeodactylum tricornutum* sterol composition. (a) GC-MS chromatogram of a TMS-derivatised extract from a 3-d-old *P. tricornutum* culture. (b, c) EI-MS spectra of TMS-derivatised brassicasterol (b) and campesterol (c) standards. (d, e) Structures of TMS-campesterol (d) and TMS-brassicasterol (e).

orientation, onto a specific region on chromosome 4 (chr_4:691145–692539), as part of the locus *PHATRDRRAFT_44478*, erroneously predicted as an intron (Fig. S1).

P. tricornutum possesses two isoforms of IDI, encoded by *PHATRDRRAFT_12533* and *PHATRDRRAFT_12972* (Fig. 3), and localized on chromosome 8 and 10, respectively. Although the first exists as an independent ORF, the latter is part of an elongated gene model, not predicted by the original genome annotation (Bowler *et al.*, 2008). Remarkably, the manual refinement of the gene model revealed an in-frame fusion with the only locus encoding the SQS enzyme (*PHATRDRRAFT_13160*). GPP is generated by the GPP synthases *PHATRDRRAFT_47271* and *PHATRDRRAFT_49325*. Likely, the latter also catalyses the subsequent conversion to FPP, as it shows high similarity with ERG20, which in *S. cerevisiae* catalyses both reactions (Tables 1, S2).

Beyond FPP, the *P. tricornutum* sterol biosynthesis pathway is predicted to involve 11 genes (Table 1) that encode enzymes necessary for the conversion of squalene to the sterols brassicasterol and campesterol. These enzymes allow different hypothetical pathway reconstructions; both of the synthetic routes used by fungi and plants are theoretically possible (Figs 1, 3). Among the identified genes we found a single OSC, encoded by a noticeably extended ORF. Another fact that emerged from the *in silico* analysis of the pathway is the lack of a conventional SQE. The

two latter observations, the occurrence of the *IDI-SQS* fusion gene, and the overall pathway structure were investigated further.

Identification of the *IDI-SQS* fusion gene in *P. tricornutum*

The gene model of the ORF that putatively encodes a fusion enzyme with predicted IDI and SQS activities was manually adjusted, resulting in a 2335-nt long ORF. Originally predicted as independent adjacent ORFs, *PHATRDRRAFT_12972* (*PtIDI*) and *PHATRDRRAFT_13160* (*PtSQS*) are actually spaced by a short coding sequence that does not include stop codons. The encoded protein consists of 763 amino acids (AAs) subdivided in a NUDIX hydroxylase/isopentenyl diphosphate delta-isomerase domain at the N-terminus and a squalene/phytoene synthase domain at the C-terminus (Fig. 4a), both sharing high similarity with the IDI and SQS enzymes of brown and green algae, respectively (Table S2). The same *IDI-SQS* fusion gene is present in all sequenced diatoms, as well as in *Aureococcus anophagefferens* and *Ectocarpus siliculosus* (Fig. S2), suggesting that it is conserved among Stramenopiles. A notable exception is the parasitic heterokont oomycete *Phytophthora*, which possesses the *IDI* gene, but lacks the *SQS* gene, as well as the whole sterol biosynthesis pathway (Gaulin *et al.*, 2010).

The *in silico* predicted *IDI-SQS* gene structure was validated at the transcript level by mining public and in-house generated *P. tricornutum* transcript sequences and by performing RT-PCR (Fig. 4b), and at the protein level by producing recombinant *PtIDISQS* proteins in *E. coli* (Fig. 4c).

Functional characterization of *PtIDISQS*

The isomerization of IPP to DMAPP is a key step in the biosynthesis of triterpenoids through the MVA pathway. Also for IPP synthesized through the MEP pathway, the equilibrium of the conversion favours the formation of DMAPP, which is considered a rate-limiting step for the synthesis of other isoprenoids such as the carotenoids (Sun *et al.*, 1998). Therefore, we evaluated the functionality of the IDI domain of *PtIDISQS* in *E. coli* transformed with the pAC-LYC plasmid that carries a set of *Erwinia herbicola* genes that enable the synthesis of the carotenoid lycopene (Cunningham & Gantt, 2007). *E. coli* transformed with pAC-LYC yield pink colonies, due to the accumulation of lycopene made from DMAPP and IPP precursors provided by the endogenous MEP pathway of the bacteria. Enhancing IDI activity results in increased lycopene accumulation, producing a visible change in the colour of *E. coli* cultures (Cunningham & Gantt, 2007). Accordingly, expression of *PtIDISQS* produced darker pink coloured cells resulting from an increased accumulation of lycopene (Fig. 5a), demonstrating the functionality of the IDI domain in the fusion protein.

In order to determine whether *PtIDISQS* also encodes a functional SQS, its ability to convert FPP to squalene was tested in *E. coli* as well. Wild-type *E. coli* does not synthesize squalene or sterols, but naturally produces FPP. Plasmids carrying *PtIDISQS* were transformed into *E. coli* and 48 h after induction of

Table 1 List of genes putatively involved in the mevalonate (MVA) and sterol biosynthesis pathway of *Phaeodactylum tricornutum*

Gene ID	EC #	Predicted function	Closest orthologue in orthology-prediction (DiatomCyc) ¹ (score) <i>GeneID</i> (organism)	
			1st InParanoid hit	2nd InParanoid hit
PHATRDRRAFT_23913	2.3.1.9	Acetyl-coa c-acetyltransferase	(423) AT5G47720 (<i>A. thaliana</i>)	(399) YPL028W (<i>S. cerevisiae</i>)
PHATRDRRAFT_16649	2.3.3.10	Hydroxymethylglutaryl-CoA (HMG-CoA) synthase	(385) AT4G11820 (<i>A. thaliana</i>)	(355) 3157 (<i>H. sapiens</i>)
PHATRDRRAFT_16870	1.1.1.34	Hydroxymethylglutaryl-CoA (HMG-CoA) reductase	(410) AT1G76490 (<i>A. thaliana</i>)	(270) MJ0705 (<i>M. jannaschii</i>)
PHATRDRRAFT_53929	2.7.1.36	Mevalonate kinase	(164) YMR208W (<i>S. cerevisiae</i>)	(140) 4598 (<i>H. sapiens</i>)
PHATRDRRAFT_44478	2.7.4.8	Phosphomevalonate kinase	–	–
PHATRDRRAFT_BD1325	4.1.1.33	Diphosphomevalonate decarboxylase	(318) AT2G38700 (<i>A. thaliana</i>)	(315) 4597 (<i>H. sapiens</i>)
PHATRDRRAFT_12972	5.3.3.2	Isopentenyl-diphosphate δ -isomerase	(265) CHLREDRAFT_24471 (<i>C. reinhardtii</i>)	(237) OSTLU_13493 (<i>O. lucimarinus</i>)
PHATRDRRAFT_12533	5.3.3.2	Isopentenyl-diphosphate δ -isomerase	(265) CHLREDRAFT_24471 (<i>C. reinhardtii</i>)	(156) 3422 (<i>H. sapiens</i>)
PHATRDRRAFT_13160 ²	2.5.1.21	Squalene synthase	(279) OSTLU_31144 (<i>O. lucimarinus</i>)	(274) 2222 (<i>H. sapiens</i>)
PHATRDRRAFT_47271	2.5.1.1	Geranyl-diphosphate synthase	(309) 9453 (<i>H. sapiens</i>)	(202) YPL069C (<i>S. cerevisiae</i>)
PHATRDRRAFT_49325	2.5.1.1	Geranyl-diphosphate synthase	(343) Ot03g02400 (<i>O. tauri</i>)	(340) YJL167W (<i>S. cerevisiae</i>)
PHATRDRRAFT_49325	2.5.1.10	Farnesyl-diphosphate synthase	(343) Ot03g02400 (<i>O. tauri</i>)	(340) YJL167W (<i>S. cerevisiae</i>)
not found	1.14.99.7	Squalene epoxidase	–	–
PHATR_645 ²	5.4.99.8	Oxidosqualene cyclase	(602) Ot03g00850 (<i>O. tauri</i>)	(596) AT2G07050 (<i>A. thaliana</i>)
PHATRDRRAFT_10824	2.1.1.41	24-methylenesterol C-methyltransferase	(301) AT5G13710 (<i>A. thaliana</i>)	(296) OSTLU_30710 (<i>O. lucimarinus</i>)
PHATRDRRAFT_49447	5.5.1.9	Cycloeucalenol cycloisomerase	(250) Ot04g04910 (<i>O. tauri</i>)	(231) OSTLU_6401 (<i>O. lucimarinus</i>)
PHATRDRRAFT_31339	1.14.13.70	14- α -demethylase	(435) CMS319C (<i>C. merolae</i>)	(427) OSTLU_43938 (<i>O. lucimarinus</i>)
PHATRDRRAFT_10852	1.14.13.72	Methylsterol monooxygenase	(82) AT1G07420 (<i>A. thaliana</i>)	–
PHATRDRRAFT_48864	1.1.1.170	3- β -hydroxysteroid-4- α -carboxylate-3-dehydrogenase	(318) Ot04g04390 (<i>O. tauri</i>)	(305) OSTLU_87094 (<i>O. lucimarinus</i>)
PHATRDRRAFT_5780	1.1.1.270	putative SDR oxidoreductase	(103) AT5G65205 (<i>A. thaliana</i>)	(97) 3248 (<i>H. sapiens</i>)
PHATR_36801	5.3.3.5	3-Beta-hydroxysteroid-delta(8), delta(7)-isomerase	–	–
PHATRDRRAFT_14208	1.14.21.6	δ -7-sterol δ -5-dehydrogenase	(193) AT3G02580 (<i>A. thaliana</i>)	(178) CHLREDRAFT_59933 (<i>C. reinhardtii</i>)
PHATRDRRAFT_30461	1.3.1.21	Δ 7-sterol reductase	(426) AT1G50430 (<i>A. thaliana</i>)	–
PHATRDRRAFT_51757	1.3.1.-.	sterol C-22 desaturase	(230) CHLREDRAFT_196874 (<i>C. reinhardtii</i>)	(219) CMJ284C (<i>C. merolae</i>)

Genes retrieved from DiatomCyc are listed with their predicted function and first and second InParanoid hits determined in (Fabris *et al.*, 2012) and listed in Supporting Information Table S2.

¹Stramenopiles are excluded.

²Reconstructed gene models are discussed in the text.

expression, the nonsaponifiable lipid fraction was extracted and analysed with GC-MS. The resulting chromatograms showed a recurrent small peak corresponding to squalene in strains expressing *PtDISQS*, but not in control strains (Fig. 5b,c), confirming the predicted SQS activity of the fusion enzyme, as well as supporting the occurrence of squalene as a sterol pathway intermediate in *P. tricornutum*.

Many Stramenopiles lack a conventional *SQE* gene

Because we could not identify *SQE* orthologues in the *P. tricornutum* genome in the dataset generated for the reconstruction of DiatomCyc, we performed an extensive BLASTP search in eukaryotes, excluding the group of fungi, animals and land plants, and using the *SQE* sequences from *S. cerevisiae*, *Homo sapiens* and *Arabidopsis* as queries. Surprisingly, *SQE*

seemed to be widely absent in the analysed groups, as already noticed previously for some of the species therein (Desmond & Gribaldo, 2009), particularly in most of the analysed Stramenopiles, with the exception of *E. siliculosus* and the oomycete pathogens *Aphanomyces euteiches* and *Saprolegnia diclina*. No hits were obtained in any Alveolata and Choanoflagellida either, but *SQE* seemed to be conserved in the red and green algal lineages, with the exception of *Chlamydomonas reinhardtii* (Fig. 6).

Chemical inhibitor treatments indicate that 2,3-epoxysqualene is a sterol pathway intermediate in *P. tricornutum*

The absence of a conventional *SQE* raised the question of whether diatoms would use 2,3-epoxysqualene as the precursor for the cyclization step. To determine this, we treated diatom

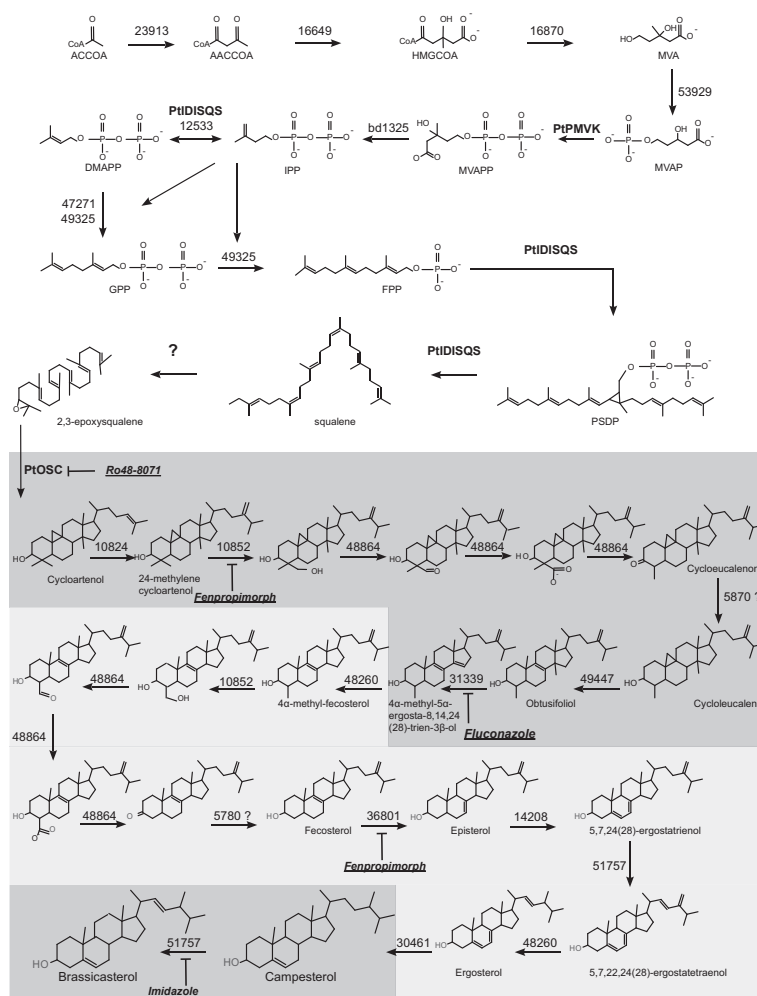


Fig. 3 *In silico* reconstruction of the mevalonate (MVA) and sterol biosynthesis pathways of *Phaeodactylum tricornutum*. The numbers above the arrows indicate the PHATRDRAFT accession number of the corresponding enzyme. Dark and light grey shading indicate similarities with the sterol biosynthesis of plants and fungi, respectively. Chemical enzyme inhibitors are indicated in italics, bold and underlined font. See Fig. 1 for explanation of abbreviations.

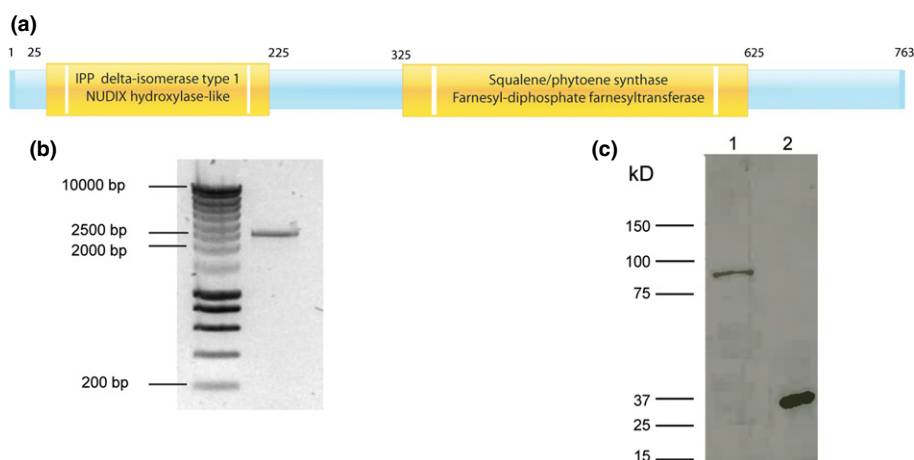


Fig. 4 The *Phaeodactylum tricornutum* *PtDISQS* fusion gene. (a) Schematic organization of the IDI and SQS domains in *PtDISQS* according to InterProScan predictions (Hunter *et al.*, 2009). Numbers refer to the approximate amino acid (AA) residue position. (b) RT-PCR amplification of *PtDISQS* from cDNA of *P. tricornutum*. (c) Immunoblot analysis with anti-His antibodies of *Escherichia coli* BL21 (DE3) samples expressing 6xHis-*PtDISQS* (lane 1) or 6xHis-*PtEDA* (Fabris *et al.*, 2012) as a control (lane 2). According to the *in silico* predictions, the 6xHis-*PtDISQS* protein should have a mass of 87.7 kDa, which is in agreement with the size of the product found.

cultures with terbinafine and Ro 48-8071, specific inhibitors of conventional SQE and OSC enzymes, respectively. Pilot experiments revealed that *P. tricornutum* cultures could be treated with 40 μ M terbinafine without major growth impairments, but that

Ro 48-8071 at a concentration of 30 μ M killed them within 48 h. Therefore, time course experiments were initiated, both with terbinafine and Ro 48-8071, to allow timely detection of sterol pathway intermediates.

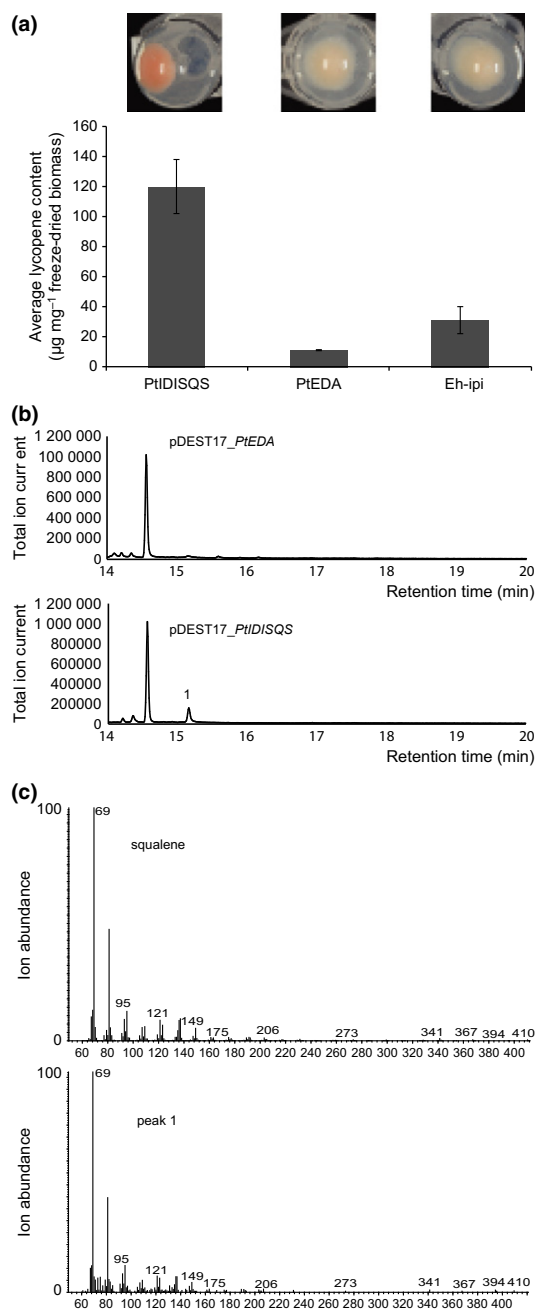


Fig. 5 Enzymatic activity of PtIDISQS. (a) Screening of lycopene accumulation in *Escherichia coli* cells transformed with the pAC-LYC plasmid. Quantification by HPLC of the lycopene content of *E. coli* colonies transformed with the pAC-LYC and *IDI* expression plasmids. Bacterial pellets showing a darker pink colour indicate an increased *IDI* activity caused by PtIDISQS overexpression (see insets on top). As a positive control, *E. coli* cells co-transformed with the *IDI* gene of *E. herbicola* (Eh-ipt) (Cunningham & Gantt, 2007) were used. As a negative control, *E. coli* cells transformed with the *PtEDA* gene were used. The latter strain indicates the basal amounts of lycopene accumulation caused by the endogenous *idi* gene of *E. coli*. Error bars, \pm SE of the mean ($n=3$). (b) GC-MS chromatogram of *E. coli* BL21 (DE3) cells expressing PtIDISQS and showing accumulation of squalene (peak1), compared to *E. coli* cultures expressing PtEDA as negative control. (c) Comparison of the MS of an authentic squalene standard with that of peak 1 from panel (b).

In accordance with the predicted absence of a conventional SQE, treatments with 40 μ M terbinafine had no effect on growth, appearance or squalene accumulation in *P. tricornutum* cultures, not even at higher concentrations (up to 340 mM). Terbinafine treatment triggered some accumulation of ergosterol (Fig. 7b), perhaps because of interference with another oxidoreductase in the pathway, such as PHATRDRRAFT_30461 or PHATRDRRAFT_51757.

By contrast, *P. tricornutum* cells treated with 10 μ M Ro 48-8071 displayed pronounced growth effects 8–12 h after treatment, and accumulated both squalene and 2,3-epoxysqualene (Fig. 7a). Furthermore, Ro 48-8071-treated cells accumulated intracellular lipid droplets, as became apparent after staining with Nile Red (Fig. 7e). Overall, these findings confirm that diatoms produce sterols through the cyclization of the conventional precursor 2,3-epoxysqualene by a conventional OSC but using a nonconventional, terbinafine-insensitive SQE enzyme to generate the 2,3-epoxysqualene precursor.

Screening for the *P. tricornutum* SQE

We mined the *P. tricornutum* genome for genes possibly encoding enzymes catalysing epoxidation of squalene to 2,3-epoxysqualene. This led to a list of candidate SQEs that included P450s, FAD-dependent monooxygenases, carotenoid epoxidases and hydroxylases (Table S3). Given its extended structure (see further below), the *P. tricornutum* OSC was included in this list as well, as a possible multifunctional SQE-OSC enzyme.

First, we attempted to detect SQE activity in an *in vitro* assay, using lysates of *E. coli* cells transformed with pDEST17 plasmids carrying the candidate genes. Established protocols were followed for *in vitro* SQE activity determination (Nagumo *et al.*, 1995; Laden *et al.*, 2000; Germann *et al.*, 2005) using nonlabelled squalene, with or without the Arabidopsis Cytochrome P450 reductase, an enzyme required for functional reconstitution of most P450 and FAD monooxygenases. However, no 2,3-epoxysqualene could be detected by GC-MS in any of the samples. Deletion of the putative targeting peptide of PtOSC (see next paragraph) or the transmembrane domains of PHATRDRRAFT_46438, PHATRDRRAFT_26422 and PHATRDRRAFT_45845 in an attempt to optimize did not lead to any detectable SQE activity, either. Second, we exploited the ability of *E. coli* cells expressing PtIDISQS to accumulate low amounts of squalene (Fig. 5b) to detect potential SQE activity of the candidate genes *in vivo*. This assay also did not reveal any accumulation of 2,3-epoxysqualene. Because immunoblot analysis of tagged versions of the candidate SQEs indicated that most of them were not or hardly expressed in the transformed *E. coli* cells (data not shown), we did not pursue SQE discovery any further.

Diatoms possess an extended OSC gene

We identified the locus PHATRDRRAFT_645 as part of an ORF putatively encoding an OSC. The gene model provided by the

Stramenopiles	Bacillariophyta	<i>Phaeodactylum tricornutum</i>		Viridiplantae	Chlorophyta	<i>Ostreococcus tauri</i>	
	Bacillariophyta	<i>Thalassiosira pseudonana</i>			Chlorophyta	<i>Ostreococcus lucimarinus</i>	
	Bacillariophyta	<i>Thalassiosira oceanica</i>			Chlorophyta	<i>Micromonas pusilla</i>	
	Bacillariophyta	<i>Fragilariopsis cylindrus</i>			Chlorophyta	<i>Chlamydomonas reinhardtii</i>	
	Bacillariophyta	<i>Pseudo-nitzschia multistriata</i>			Chlorophyta	<i>Chlorella variabilis</i>	
	Pelagophyceae	<i>Aureococcus anophagefferens</i>			Chlorophyta	<i>Volvox carteri</i>	
	Oomycetes	<i>Albugo laibachii</i>			Chlorophyta	<i>Coccomyxa subellipsoidea</i>	
	Oomycetes	<i>Phytophthora infestans</i>			Chlorophyta	<i>Bathycoccus prasinos</i>	
	Oomycetes	<i>Aphanomyces euteiches</i>			Chlorophyta	<i>Chlorella variabilis</i>	
	Oomycetes	<i>Saprolegnia diclina</i>		Euglenozoa	Kinetoplastida	<i>Trypanosoma cruzi</i>	
Alveolata	Apicomplexa	<i>Toxoplasma gondii</i>			Kinetoplastida	<i>Leishmania braziliensis</i>	
	Apicomplexa	<i>Neospora caninum</i>			Kinetoplastida	<i>Strigomonas culicis</i>	
	Perkinsea	<i>Perkinsus marinus</i>			Kinetoplastida	<i>Angomonas deanei</i>	
Cryptophyta	Pyrenomonadales	<i>Guillardia theta</i>		Rhodophyta	Bangiophyceae	<i>Cyanidioschyzon merolae</i>	
Opisthokonta	Choanoflagellida	<i>Salpingoeca sp.</i>			Bangiophyceae	<i>Galdieria sulphuraria</i>	
	Choanoflagellida	<i>Monosiga brevicollis</i>		Heterolobosea	Schizopyrenida	<i>Naegleria gruberi</i>	
Rhizaria	Cercozoa	<i>Paulinella chromatophora</i>					

Fig. 6 Conservation (grey) and loss (black) of the *SQE* gene in representative organisms belonging to the groups of Stramenopiles, green algae, Alveolata, Choanoflagellida, Kinetoplastids, Cryptophyta, Heterolobosea and Rhizaria, according to BLASTP searches.

JGI assembly (Bowler *et al.*, 2008) was incomplete and wrongly annotated as acetyl coenzyme A synthase. Therefore, it was manually adjusted, resulting in an ORF of 2931 nt, denominated *PtOSC* and spanning the coding regions of both the preceding and the succeeding gene on chromosome 11, *PHATRDRRAFT_46724* and *PHATRDRRAFT_46726*. The reconstructed *PtOSC* model was confirmed at the transcript level by mining of public and in-house generated *P. tricornutum* transcript sequences and by RT-PCR analysis, producing a single band of the predicted size (Fig. S3a). Notably, as for the *PtDISQS* gene, a similarly extended *OSC* locus could be identified in all annotated diatom genomes (Fig. S4). In the predicted *PtOSC* protein, the sequence corresponding to the first 46 N-terminal AAs is possibly part of an ER-targeting peptide (Emanuelsson *et al.*, 2007), the following 103 AAs potentially correspond to three transmembrane domain repeats (Emanuelsson *et al.*, 2007), and the C-terminus shows similarity to common CAS and LAS enzymes. Although all diatom *OSC* sequences share an N-terminal extension, its sequence is apparently species-specific and not conserved among diatoms (Fig. S4). The existence of the FL protein could not be verified yet as tagged versions could not be expressed in *E. coli* or *S. cerevisiae*, perhaps due to differences in codon usage, the considerable size of the polypeptide and/or the presence of potential membrane targeting domains at the N-terminus. Therefore, a truncated version of *PtOSC*, in which the N-terminal 46 AAs were removed, was fused to MBP and expressed in *E. coli*. Immunoblot analysis revealed the existence of a large protein of *c.* 104 kDa (Fig. S3b), matching the *in silico* predicted protein model.

PtOSC encodes a cycloartenol synthase (CAS)

Being a highly conserved enzyme, the specificity of an *OSC* can be inferred from the AA residues at positions 381, 449 and 453

(numbering relative to the *OSC* of *H. sapiens*) that are specific for the formation of cycloartenol (Y381,H449,I453) or lanosterol (T381,C/Q449,V453) (Fig. S4) (Summons *et al.*, 2006), which would simultaneously also allow location of this portion of the diatom pathway within the photosynthetic or the nonphotosynthetic lineage, respectively. Analysis of these active site residues in *PtOSC* suggested that *P. tricornutum*, as well as other diatoms, cyclizes 2,3-epoxysqualene to cycloartenol (Fig. S4), like plants and in agreement with previous postulations for the *Thalassiosira pseudonana* *OSC* (Desmond & Gribaldo, 2009).

We were not able to detect any enzymatic activity with the truncated, recombinant *PtOSC* protein produced in *E. coli* (Fig. S3b). Considering the inefficient expression of tagged *PtOSC*, either in *E. coli* or *S. cerevisiae*, we assessed *PtOSC* activity in a yeast strain (TM5) that we have engineered for high, inducible, accumulation of 2,3-epoxysqualene (Moses *et al.*, 2014). Expression of full-length, nontagged *PtOSC* in this strain led to the accumulation of cycloartenol, indicating that *PtOSC* encodes a CAS (Fig. 8a).

Additionally, we employed a pharmacological approach and determined the accumulation of sterol pathway intermediates in *P. tricornutum* cultures treated with different concentrations of fluconazole, a specific inhibitor of CYP51 (*PHATRDRRAFT_31339*), another conserved enzyme in the sterol pathway acting downstream of the *OSC* (Fig. 3). The strongest effect of fluconazole on *P. tricornutum* cultures was observed to occur 48 h after treatment at a concentration of 653 mM (200 mg ml⁻¹), but lower concentrations (tested down to 30 µM) produced similar results (data not shown). No visible effects on growth were apparent, even at the highest concentration tested, but GC-MS analysis demonstrated the presence of two distinct peaks, as compared to the mock-treated samples (Fig. 7c). The first corresponds to obtusifoliol (Figs 7c, S5), a known derivative of cycloartenol in plants, thus confirming the

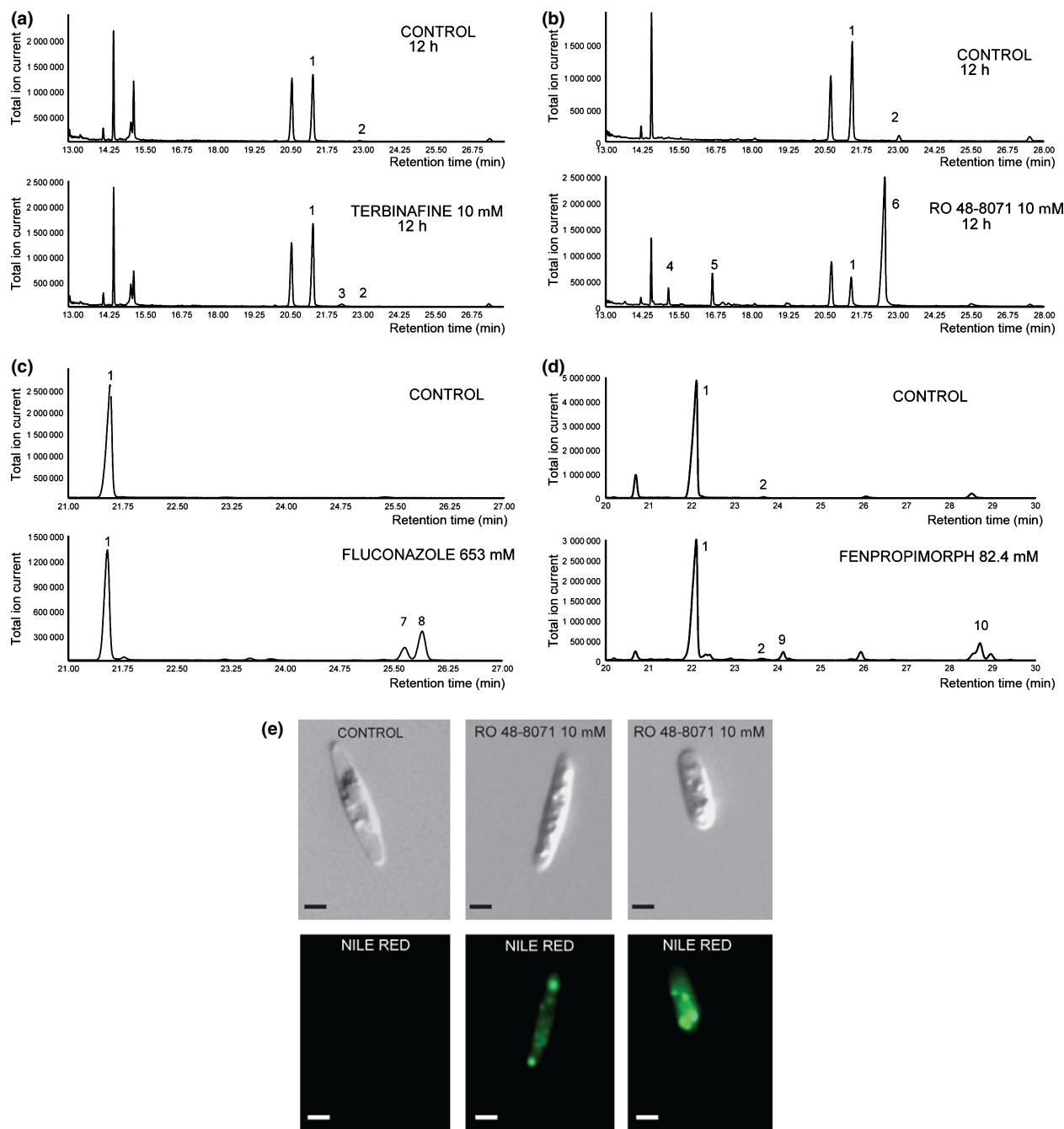


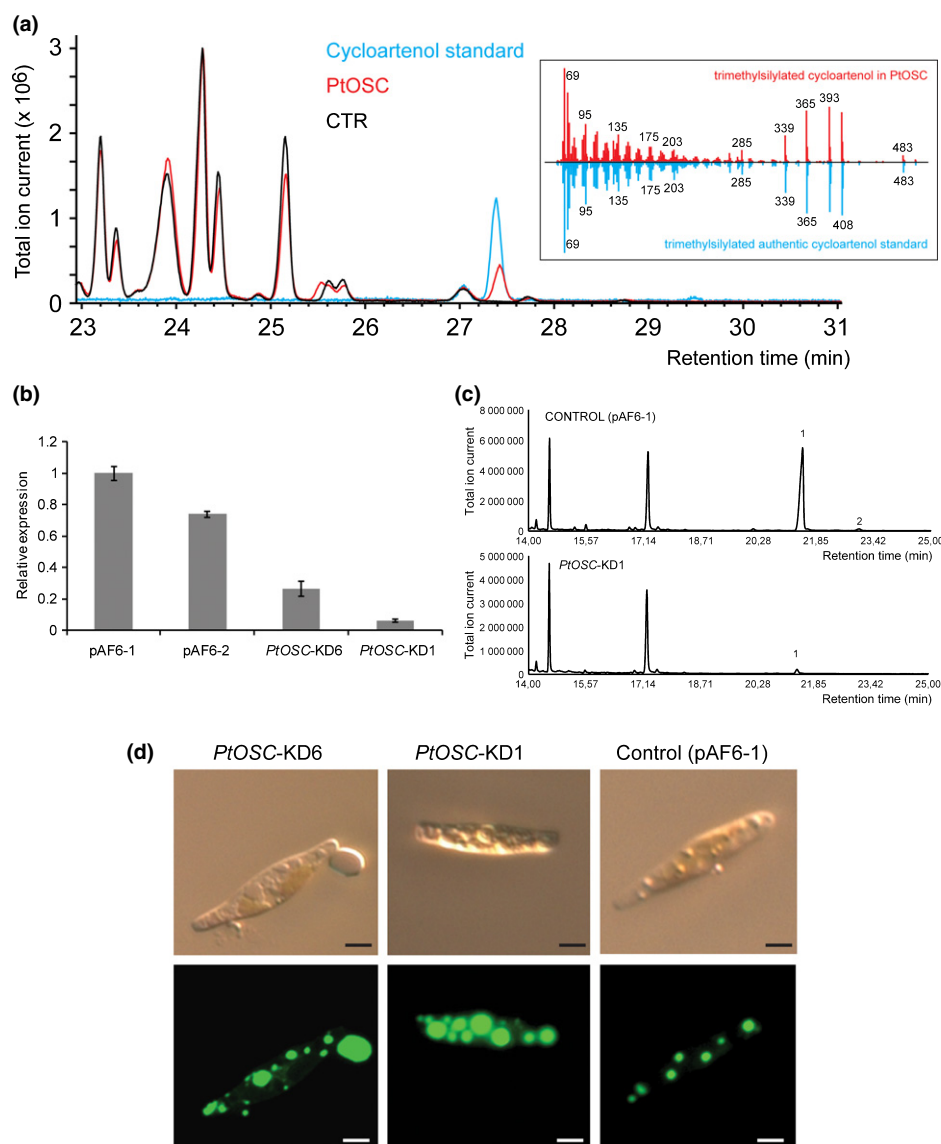
Fig. 7 Chemical perturbation of *Phaeodactylum tricornutum* sterol synthesis. (a–d) GC-MS chromatograms of TMS-derivatised nonsaponifiable lipid extracts of *P. tricornutum* cells treated for 12 h with 10 μ M Ro 48-8071 (a), 40 μ M terbinafine (b), 653 mM fluconazole (c) and 82.4 mM fenpropimorph (d) and compared to the respective mock treatments. Peak numbering: 1, TMS-brassicasterol; 2, TMS-campesterol; 3, TMS-ergosterol; 4, squalene; 5, 2, 3-epoxysqualene; 6, Ro 48-8071; 7, TMS-obtusifoliol; 8, TMS derivative of unknown steroid compound of FW 500; 9, putative TMS-fecosterol; 10, TMS-cycloartenol. EI-MS spectra of these peaks are given in Supporting Information Fig. S5. (e) Effects of Ro 48-8071 on lipid accumulation in *P. tricornutum* cells. Upper and lower panels show differential interference contrast (DIC) images of *P. tricornutum* cells and epifluorescence images of the same cells upon staining with Nile Red, to visualise intracellular lipids. Bars, 3 μ m.

specificity of PtOSC. The second peak is an unknown steroid compound with the formula weight 428, which possibly corresponds to a reduced form of obtusifoliol (Fig. S5).

In order to further support the role of PtOSC in diatom sterol biosynthesis, *P. tricornutum* cells were transformed with RNAi constructs targeting this gene within the locus *PHATRDR4*

.FT_645. No detectable amounts of OSC precursors accumulated in the *PrOSC* RNAi lines but they exhibited a striking phenotype characterized by dramatically impaired growth, reduced sterol content, and a significant increase in lipid accumulation (Fig. 8b–d). These phenotypes are typical for unicellular organisms with impaired sterol biosynthesis (Wentzinger *et al.*, 2002;

Fig. 8 PtOSC is a cycloartenol synthase. (a) Cycloartenol synthase activity of PtOSC in transformed yeast cells. Overlay of GC chromatograms showing accumulation of cycloartenol in cells transformed with *PtOSC* but not in control cells (CTR). Blue, Cycloartenol standard; red, *PtOSC*; black, CTR. The inset shows the EI-MS spectrum of trimethylsilylated cycloartenol. The GC retention time and EI-MS spectra of cycloartenol produced in yeast match those of the authentic standard. (b–d) Effects of *PtOSC* silencing in transformed *Phaeodactylum tricornutum* cells. (b) qRT-PCR analysis of *PtOSC* transcript levels in *PtOSC* RNAi lines (KD) relative to control lines transformed with the pAF6 vector. The expression ratio was normalized to the pAF6_1 control line. Error bars, \pm SE of the mean ($n = 3$). (c) GC-MS chromatograms of the TMS-derivatised nonsaponifiable lipid fraction of *PtOSC-KD1*, showing a dramatic decrease in the accumulation of the main steroid compound compared to the pAF6 control line. Peak numbering: 1, TMS-brassicasterol; 2, TMS-campesterol. (d) Effects of *PtOSC* silencing on lipid accumulation in cells of 3-d-old *Phaeodactylum tricornutum* cultures. Upper and lower panels show differential interference contrast (DIC) images of *P. tricornutum* cells and epifluorescence images of intracellular lipid accumulation, in cells stained with Nile Red. Bars, 3 μ m.



Ta *et al.*, 2012). Also in plant and yeast cells, the perturbation of the sterol metabolism through chemical inhibitors causes the appearance of lipid droplets, whereas in human cells, the blockage of HMGCR, catalysing the rate-limiting step of the MVA pathway, triggers accumulation of polyunsaturated fatty acids and upregulation of fatty acid biosynthesis genes (Wentzinger *et al.*, 2002; Plée-Gautier *et al.*, 2012; Ta *et al.*, 2012). Hence, these findings confirm the involvement of PtOSC in the sterol pathway and support the link between sterol biosynthesis and lipid accumulation, already observed after treatment with Ro 48-8071, the specific inhibitor of conventional OSC enzymes (Fig. 7e).

P. tricornutum employs a chimeric sterol synthesis route

The chemical fenpropimorph has strong effects and possibly multiple targets in sterol and other synthesis pathways in eukaryotes. Although drug concentrations between 50 and

100 μ g ml⁻¹ rapidly killed *P. tricornutum* cells, treatments at concentrations of 12.5 and 25 μ g ml⁻¹ caused an altered sterol profile. GC-MS analysis of sterols accumulated by fenpropimorph-treated diatom cultures revealed the presence of cycloartenol and another peak, presumably fecosterol (Figs 7d, S5), a typical fungal sterol. The accumulation of cycloartenol, which further confirms the predicted activity of PtOSC, is likely caused by the inhibition of the methylsterol monooxygenase encoded by *PHATRDRRAFT_10852*, as this enzyme type is also a known target of fenpropimorph (Burden *et al.*, 1989). The presence of fecosterol may be supported by the fact that the enzyme Δ 7-sterol isomerase, likely encoded by *PHATRDRRAFT_36801*, that converts this compound to episterol, is a known target of fenpropimorph (Campagnac *et al.*, 2009). Based on the outcomes of the inhibitor treatments, the full *P. tricornutum* sterol pathway could be tentatively reconstructed and is proposed to be a hybrid of the plant and fungal sterol synthesis routes (Fig. 3).

The enzymes $\Delta 7$ -sterol reductase and sterol C-22 desaturase catalyse the final steps in the synthesis of the main *P. tricornutum* sterols

The genes *PHATRDRRAFT_30461* and *PHATRDRRAFT_51757* putatively encode a sterol $\Delta 7$ -reductase and a sterol C-22 desaturase, respectively (Table 1). Both show high similarity to reported Arabidopsis sterol enzymes (Table S2). *PHATRDRRAFT_30461* is the putative orthologue of *DWARF5* (*At1g50430*) that encodes a sterol $\Delta 7$ -reductase, whereas *PHATRDRRAFT_51757* is the putative orthologue of *At2g34490*, which encodes a P450 of the CYP710 subfamily with sterol 22-desaturase activity (Morikawa *et al.*, 2006). In our proposed pathway reconstruction (Fig. 3), these enzymes are associated with the last two reactions, theoretically converting ergosterol to campesterol (*PHATRDRRAFT_30461*) and subsequently to brassicasterol (*PHATRDRRAFT_51757*), the two main sterols found in *P. tricornutum* (Fig. 2).

In order to confirm their predicted activity, both genes were cloned and expressed in *S. cerevisiae*, which naturally accumulates ergosterol as the main sterol compound. GC-MS analysis of yeast expressing *PHATRDRRAFT_30461* confirmed its $\Delta 7$ -reductase activity and its capacity to convert ergosterol to campesterol (Fig. 9a). Additionally, in the same samples, brassicasterol was also detected, presumably as a result of a desaturation of campesterol in position C-22 by ERG5, the endogenous C-22 sterol reductase. Both sterols were also detected in yeast cells expressing both diatom genes (Fig. 9b). In support of the possible activity of *PHATRDRRAFT_51757* in converting campesterol to

brassicasterol, we observed that in the presence of increasing concentrations of imidazole, which inhibits P450s of the CYP710 subfamily with minor specificity, *P. tricornutum* cultures increasingly accumulate campesterol in a concentration-dependent manner (Fig. 9c).

Discussion

Reconstruction of the *Phaeodactylum tricornutum* sterol synthesis pathway

Through computational analysis, we reconstructed the sterol biosynthesis pathway of *P. tricornutum* *in silico*. Several elements of the predicted pathway were experimentally supported with data from pharmacological assays, metabolite profiling of wild-type and transgenic *P. tricornutum* lines, and activity assays using recombinant *P. tricornutum* enzymes produced in bacteria and yeast. Together these data suggest that *P. tricornutum* utilizes a chimeric pathway that leads to the main sterols brassicasterol and campesterol, and has the initial and final parts in common with the pathway from plants, whereas the central part appears to be similar to that of fungi (Fig. 3). This supports the postulation, based on phylogenomic analysis only, that another model diatom, *T. pseudonana*, also has a sterol pathway that displays a mixture of features from fungi, animals and land plants (Desmond & Gribaldo, 2009).

Based on our analysis, the following reaction scheme is postulated. Despite the absence of a conventional SQE, *P. tricornutum* also generates 2,3-epoxysqualene as the precursor for the

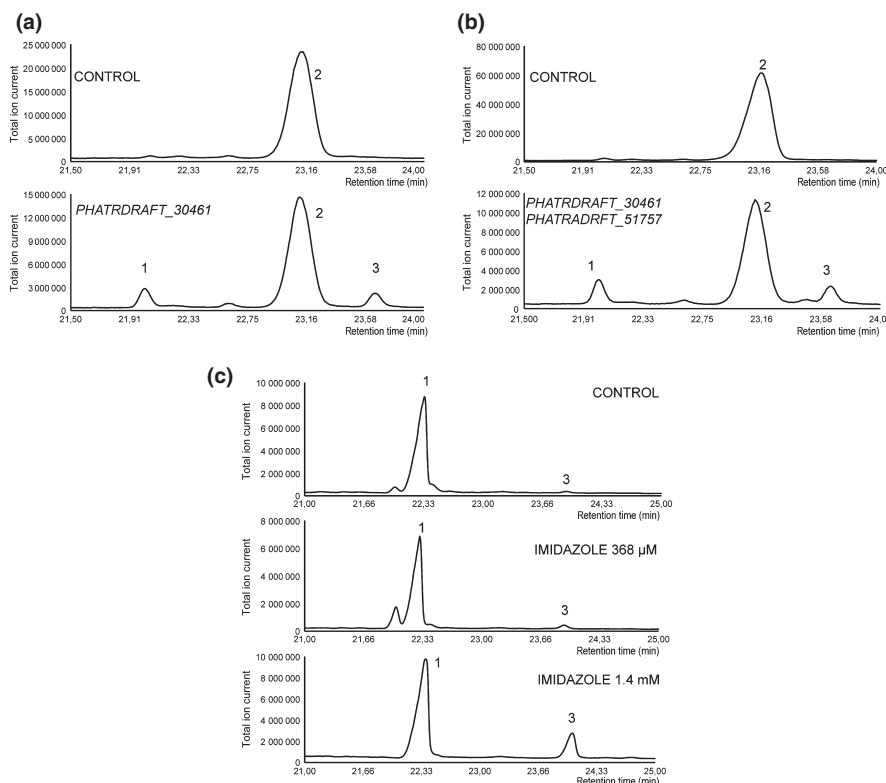


Fig. 9 Functional characterization of *PHATRDRRAFT_30461* and *PHATRDRRAFT_51757*. (a, b) GC-MS chromatograms of TMS-derivatised nonsaponifiable lipid extracts of *Saccharomyces cerevisiae* cells expressing *PHATRDRRAFT_30461* alone (a) or in combination with *PHATRDRRAFT_51757* (b) compared to controls transformed with corresponding empty vector(s). (c) GC-MS chromatograms showing the perturbed sterol composition of *Phaeodactylum tricornutum* cultures treated for 48 h with 368 µM or 1.4 mM imidazole, and compared to mock-treated cultures. Peak numbering: 1, TMS-brassicasterol; 2, TMS-ergosterol; 3, TMS-campesterol.

cyclization to cycloartenol, like plants and green algae. Cycloartenol is then methylated at C-24 by *PHATRDRRAFT_10824*, a sterol methyltransferase yielding 24-methylene-cycloartenol. One of the two methyl groups at C-4 is then removed by the subsequent actions of three enzymes. The first is the methylsterol monooxygenase encoded by *PHATRDRRAFT_10852*, which shows significant similarity with plant orthologues. The resulting hydroxysterol is decarboxylated on the C-4 hydroxy-methyl group and at the same time dehydrogenated on the hydroxyl group at the C-3 carbon by an NADPH-dependent reaction, catalysed by a 3 β -hydroxysteroid-4 α -carboxylate-3-dehydrogenase encoded by *PHATRDRRAFT_48864*, with highest orthology scores to the *Ostreococcus* enzymes postulated previously as the homologues of the yeast C-3 dehydrogenase/C-4 decarboxylase ERG26 (Desmond & Gribaldo, 2009). The resulting keto-sterol requires a reduction of the oxygen at C-3 in order to be further converted. This step requires a 3-keto-steroid reductase, which is currently unidentified in plants and algae. We postulate that a reductase involved in other reactions might have acquired specificity for keto-sterols. For example, using Pathologic (Karp *et al.*, 2002), we identified *PHATRDRRAFT_5870*, encoding a short-chain dehydrogenase/reductase as the best candidate for this reaction, potentially yielding the molecule cycloleucalenol, the typical precursor of obtusifolol in plants. Similar to land plants, a cycloleucalenol cycloisomerase breaks the cyclopropane ring present between C-19 and C-9 with a consequent desaturation at position C-9/C-8. This enzyme is possibly encoded by *PHATRDRRAFT_49447*, which shares highest orthology with the corresponding enzymes in the Viridiplantae lineage. The resulting product, obtusifolol, a common intermediate in the sterol biosynthesis of plants and green algae, is the substrate of the P450 sterol 14 α -demethylase (PtCYP51), likely encoded by *PHATRDRRAFT_31339*. This enzyme catalyses the removal of the methyl group at C-14 and the formation of the double bond between C-14 and C-15, yielding (4 α)-methyl-(5 α)-ergosta-8,14,24(28)-trien-3 β -ol. The presence of a sterol C14-24 reductase, corresponding to the locus *PHATRDRRAFT_48260* (PtSC14-24R), allows the conversion of this compound to methyl-fecosterol, which in turn, through the re-iteration of the oxidative demethylation at C-4 catalysed by the same enzymes as described above, is converted to fecosterol, a typical intermediate of ergosterol biosynthesis in yeast and fungi. *PHATRDRRAFT_36801* encodes a $\Delta 8$ - $\Delta 7$ sterol isomerase that can possibly catalyse the shift of the double bond from C-8/C-9 to C-8/C-7. The resulting episterol is first desaturated by a Δ (7)-sterol 5-desaturase, possibly encoded by *PHATRDRRAFT_14208*, subsequently subjected to a second desaturation reaction at position C-22 by the P450 C-22 sterol desaturase (PtCYP710) encoded by *PHATRDRRAFT_51757*, and lastly reduced at position C-24 by PtSC14-24R (*PHATRDRRAFT_48260*), to yield the typical fungal sterol ergosterol. Despite being the main sterol of fungi, ergosterol is often found in algae as well (Miller *et al.*, 2012). Finally, in two steps the ergosterol is converted to the phytosterols

campesterol and brassicasterol, the final products of sterol biosynthesis in *P. tricornutum*, by the $\Delta 7$ -sterol reductase encoded by *PHATRDRRAFT_30461* and the C-22 desaturase encoded by *PHATRDRRAFT_51757*, respectively. Interestingly, orthologues of *PHATRDRRAFT_51757* have not been found in *T. pseudonana* (Table S2), indicating a possible difference in the pathways of the two diatoms and in agreement with the fact that the sterols produced by *T. pseudonana* are not desaturated at position C-22 (Rampen *et al.*, 2010).

Peculiar enzymes in the diatom sterol pathway

Besides its chimeric nature, the sterol biosynthesis pathway of *P. tricornutum* harbours several other peculiarities that make it unique among the known variants of this pathway across the kingdoms of life. Particularly, the pathway is characterized by: the fusion of the IDI and SQS activities in a single multifunctional enzyme; the lack of a conventional SQE; and the presence of an exotic CAS that is composed of a conserved C-terminal domain and a large, less conserved N-terminal region. Remarkably, these features recur in all sequenced diatoms.

Joining different enzymatic activities in a single fusion protein is a frequent event in protein evolution and often implicates enzymes subjected to co-regulation or involved in the same pathway (Hwang *et al.*, 2011). In diatoms, the presence of fusion enzymes that catalyse subsequent reactions appears relatively frequent. Examples are found in carbohydrate metabolism, where a triosephosphate-isomerase/glyceraldehyde-3-phosphate dehydrogenase (*PHATRDRRAFT_25308*), a UDP-glucose-pyrophosphorylase/phosphoglucomutase (*PHATRDRRAFT_50444*) and a glucose-6-phosphate-dehydrogenase/6-phosphogluconate-dehydrogenase (*PHATRDRRAFT_54663*) were predicted to exist as fusion proteins in *P. tricornutum* (Kroth *et al.*, 2008). In contrast to the former fusion enzymes, IDI and SQS activities do not occupy a consecutive position within the sterol pathway. DMAPP, the product of IDI, is converted to FPP before being converted by SQS to squalene. This intermediate conversion occurs through two additional reactions encoded by the loci *PHATRDRRAFT_49325* and *PHATRDRRAFT_47271*, located on different chromosomes than *PtIDISQS*. The IDI-SQS fusion gene is conserved in diatoms, and presumably even in the Stramenopiles, suggesting that its origin might be considerably ancient and confer a selective advantage. Although the occurrence of protein–protein interactions with other enzymes of the pathway cannot be excluded, it is plausible that a potential selective advantage of the IDI-SQS protein resides in a more efficient co-regulation, rather than in metabolic channelling.

The epoxidation of squalene by SQE, requiring FAD as cofactor and molecular oxygen, is believed to be an ubiquitous reaction, occurring in aerobic conditions in every sterol-producing organism through an identical mechanism. As for most of the other proteins working upstream of the OSC, the degree of conservation of SQE is so high that no sterol-producing organisms with alternative SQE enzymes have been reported yet. Our and previous (Desmond & Gribaldo, 2009) surveys clearly indicate that the SQE gene seems to be lost in several groups. The existence of alternative

biochemical mechanisms for the epoxidation of squalene thus seems plausible and diatoms might use one of them.

For example, diatoms might have evolved a particular P450 that acquired a novel activity or specificity. In rat tumour cells it has been demonstrated that a P450 17- α hydroxylase-17,20 lyase (CYP17) has a secondary SQE activity (Liu *et al.*, 2005). Alternatively, among the substantial number of genes encoding proteins with unknown function (Maheswari *et al.*, 2010; Fabris *et al.*, 2012) diatoms might harbour an enzyme with unprecedented SQE activity. Theoretically, squalene epoxidation could occur anaerobically using water as source of oxygen instead of O₂ (Raymond & Blankenship, 2004), although the thermodynamics of such a reaction would be significantly less favourable (Summons *et al.*, 2006). Because the synthesis of sterols presumably evolved in anaerobic eukaryotes facing an increasingly oxidative environment, it is possible that the primordial SQE was anaerobic (Raymond & Blankenship, 2004) and substantially different. Before being replaced by the 'modern' SQE, this hypothetical primitive enzyme might have persisted in some groups. Based on some of these assumptions, we compiled a list of possible alternative enzymes that might have replaced SQE in diatoms (Table S3) and we tested their activity in a series of preliminary assays in *E. coli*; these have been unsuccessful thus far.

The utility of DiatomCyc

The reconstruction of the complex diatom sterol pathway starting from the DiatomCyc database highlights the value and reliability of this tool for research on diatom metabolism. The proposed sterol biosynthesis pathway follows a chimeric fungal/plant route and introduces novel multifunctional enzymes and novel, yet unknown, enzymatic alternatives to a highly conserved biochemical event. This further underscores the prominent metabolic plasticity of diatoms, suggests that sterol biosynthesis in the marine environment might have evolved differently, and requires a general reconsideration of the sterol biosynthetic pathway, currently considered as a highly conserved pathway and subdivided into rigid phylogenetic variants. Further efforts will be required to identify the alternative SQE and confirm the role of some of the sterol enzymes postulated in this study. Similarly, the existence of a conserved link between the regulation of sterol biosynthesis and lipid accumulation, suggested by the increased lipid accumulation in diatoms with impaired sterol biosynthesis, might warrant further investigation at a physiological, metabolic and enzymatic level. Such multi-level analysis might benefit from the information made available through DiatomCyc and other, analogous resources.

Acknowledgements

We thank Dr Hubert Schaller (Institut de Biologie Moléculaire des Plantes University of Strasbourg, France) for the obtusifolioside standard, Professor Francis X. Cunningham (University of Maryland, USA) for the pAC LYC and pAC LYCipi plasmids, Wilson Ardiles-Diaz for sequencing and Annick Bleys for help in preparing the manuscript. This work was supported by the Agency for Innovation by

Science and Technology in Flanders ('Strategisch Basisonderzoek' grant no. 80031 and by a predoctoral fellowship to M.M.). J.P. is a Postdoctoral Fellow of the Research Foundation Flanders.

References

- Adolph S, Bach S, Blondel M, Cueff A, Moreau M, Pohnert G, Poulet SA, Wichard T, Zuccaro A. 2004. Cytotoxicity of diatom-derived oxylipins in organisms belonging to different phyla. *The Journal of Experimental Biology* 207: 2935–2946.
- Alberti S, Gitler AD, Lindquist S. 2007. A suite of Gateway cloning vectors for high-throughput genetic analysis in *Saccharomyces cerevisiae*. *Yeast* 24: 913–919.
- Benveniste P. 2004. Biosynthesis and accumulation of sterols. *Annual Review of Plant Biology* 55: 429–457.
- Berges JA, Franklin DJ, Harrison PJ. 2001. Evolution of an artificial seawater medium: improvements in enriched seawater, artificial water over the last two decades. *Journal of Phycology* 37: 1138–1145.
- 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.
- Burden RS, Cooke DT, Carter GA. 1989. Inhibitors of sterol biosynthesis and growth in plants and fungi. *Phytochemistry* 28: 1791–1804.
- Campagnac E, Fontaine J, Lounès-Hadj Sahraoui A, Laruelle F, Durand R, Grandmougin-Ferjani A. 2009. Fenpropimorph slows down the sterol pathway and the development of the arbuscular mycorrhizal fungus *Glomus intraradices*. *Mycorrhiza* 19: 365–374.
- Cunningham FX, Gantt E. 2007. A portfolio of plasmids for identification and analysis of carotenoid pathway enzymes: *Adonis aestivalis* as a case study. *Photosynthesis Research* 92: 245–259.
- De Riso V, Raniello R, Maumus F, Rogato A, Bowler C, Falcitatore A. 2009. Gene silencing in the marine diatom *Phaeodactylum tricornutum*. *Nucleic Acids Research* 37: e96.
- Desmond E, Gribaldo S. 2009. Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature. *Genome Biology and Evolution* 1: 364–381.
- Dufour E. 2008. Sterols and membrane dynamics. *Journal of Chemical Biology* 1: 63–77.
- Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* 2: 953–971.
- 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. *Plant Journal* 70: 1004–1014.
- Falcitatore A, Casotti R, Leblanc C, Abrescia C, Bowler C. 1999. Transformation of nonselectable reporter genes in marine diatoms. *Marine Biotechnology* 1: 239–251.
- Galea AM, Brown AJ. 2009. Special relationship between sterols and oxygen: were sterols an adaptation to aerobic life? *Free Radical Biology & Medicine* 47: 880–889.
- Gaulin E, Bottin A, Dumas B. 2010. Sterol biosynthesis in oomycete pathogens. *Plant Signaling & Behavior* 5: 258–260.
- Germann M, Gallo C, Donahue T, Shirzadi R, Stukey J, Lang S, Ruckenstein C, Oliaro-Bosso S, McDonough V, Turnowsky F *et al.* 2005. Characterizing sterol defect suppressors uncovers a novel transcriptional signaling pathway regulating zymosterol biosynthesis. *The Journal of Biological Chemistry* 280: 35 904–35 913.
- Giner J-L, Wikfors GH. 2011. "Dinoflagellate Sterols" in marine diatoms. *Phytochemistry* 72: 1896–1901.
- Greenspan P, Mayer EP, Fowler SD. 1985. Nile red: a selective fluorescent stain for intracellular lipid droplets. *The Journal of Cell Biology* 100: 965–973.
- Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L *et al.* 2009. InterPro: the integrative protein signature database. *Nucleic Acids Research* 37: D211–D215.

- Huysman MJJ, Fortunato AE, Matthijs M, Costa BS, Vanderhaeghen R, Van den Daele H, Sachse M, Inzé D, Bowler C, Kroth PG *et al.* 2013. AUREOCHROME1a-mediated induction of the diatom-specific cyclin *dsCYC2* controls the onset of cell division in diatoms (*Phaeodactylum tricornutum*). *The Plant Cell* 25: 215–228.
- Hwang S, Rhee SY, Marcotte EM, Lee I. 2011. Systematic prediction of gene function in *Arabidopsis thaliana* using a probabilistic functional gene network. *Nature Protocols* 6: 1429–1442.
- Karp PD, Paley S, Romero P. 2002. The Pathway Tools software. *Bioinformatics* 18: S225–S232.
- Kodner R, Summons R. 2008. Sterols in a unicellular relative of the metazoans. *Proceedings of the National Academy of Sciences, USA* 105: 9897–9902.
- 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.
- Laden BP, Tang Y, Porter TD. 2000. Cloning, heterologous expression, and enzymological characterization of human squalene monooxygenase. *Archives of Biochemistry and Biophysics* 374: 381–388.
- Lamb DC, Jackson CJ, Warrilow AGS, Manning NJ, Kelly DE, Kelly SL. 2007. Lanosterol biosynthesis in the prokaryote *Methylococcus capsulatus*: insight into the evolution of sterol biosynthesis. *Molecular Biology and Evolution* 24: 1714–1721.
- Leblond JD, Lasiter AD. 2012. Sterols of the green-pigmented, aberrant plastid dinoflagellate, *Lepidodinium chlorophorum* (Dinophyceae). *Protist* 163: 38–46.
- Liu Y, Yao Z-X, Papadopoulos V. 2005. Cytochrome P450 17 α hydroxylase/17,20 lyase (CYP17) function in cholesterol biosynthesis: identification of squalene monooxygenase (epoxidase) activity associated with CYP17 in Leydig cells. *Molecular Endocrinology* 19: 1918–1931.
- Lohr M, Schwender J, Polle JE. 2012. Isoprenoid biosynthesis in eukaryotic phototrophs: a spotlight on algae. *Plant Science* 185–186: 9–22.
- Maheswari U, Jabbari K, Petit J-L, Porcel BM, Allen AE, Cadoret J-P, De Martino A, Heijde M, Kaas R, La Roche J *et al.* 2010. Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome Biology* 11: R85.
- Massé 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.
- Miller MB, Haubrich B, Wang Q, Snell WJ, Nes WD. 2012. Evolutionarily conserved $\Delta^{25(27)}$ -olefin ergosterol biosynthesis pathway in the alga *Chlamydomonas reinhardtii*. *Journal of Lipid Research* 53: 1636–1645.
- Morikawa T, Mizutani M, Aoki N, Watanabe B, Saga H, Saito S, Oikawa A, Suzuki H, Sakurai N, Shibata D *et al.* 2006. Cytochrome P450 CYP710A encodes the sterol C-22 desaturase in *Arabidopsis* and tomato. *The Plant Cell* 18: 1008–1022.
- Moses T, Pollier J, Almagro L, Buyst D, Van Montagu M, Pedreño MA, Martins JC, Thevelein JM, Goossens A. 2014. Combinatorial biosynthesis of sapogenins and saponins in *Saccharomyces cerevisiae* using a C-16 α hydroxylase from *Bupleurum falcatum*. *Proceedings of the National Academy of Sciences, USA* 111: 1634–1639.
- Nagumo A, Kamei T, Sakakibara J, Ono T. 1995. Purification and characterization of recombinant squalene epoxidase. *Journal of Lipid Research* 36: 1489–1497.
- Nes CR, Singha UK, Liu J, Ganapathy K, Villalta F, Waterman MR, Lepesheva GI, Chaudhuri M, Nes WD. 2012. Novel sterol metabolic network of *Trypanosoma brucei* procyclic and bloodstream forms. *The Biochemical Journal* 443: 267–277.
- Ohyama K, Suzuki M, Kikuchi J, Saito K, Muranaka T. 2009. Dual biosynthetic pathways to phytosterol via cycloartenol and lanosterol in *Arabidopsis*. *Proceedings of the National Academy of Sciences, USA* 106: 725–730.
- Pearson A, Budin M, Brocks JJ. 2003. Phylogenetic and biochemical evidence for sterol synthesis in the bacterium *Gemmata obscuriglobus*. *Proceedings of the National Academy of Sciences, USA* 100: 15 352–15 357.
- Plée-Gautier E, Antoun J, Goulitquer S, Le Jossic-Corcos C, Simon B, Amet Y, Salauin J-P, Corcos L. 2012. Statins increase cytochrome P450 4F3-mediated eicosanoids production in human liver cells: a PXR dependent mechanism. *Biochemical Pharmacology* 84: 571–579.
- Rampen SW, Abbas BA, Schouten S, Sinnighe Damsté J. 2010. A comprehensive study of sterols in marine diatoms (Bacillariophyta): implications for their use as tracers for diatom productivity. *Limnology and Oceanography* 55: 91–105.
- Raymond J, Blankenship RRE. 2004. Biosynthetic pathways, gene replacement and the antiquity of life. *Geobiology* 2: 199–203.
- Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falcatore A, Bowler C. 2007. Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. *Gene* 406: 23–35.
- Summons RE, Bradley AS, Jahnke LL, Waldbauer JR. 2006. Steroids, triterpenoids and molecular oxygen. *Philosophical transactions of the Royal Society B* 361: 951–968.
- Sun Z, Cunningham FX, Gantt E. 1998. Differential expression of two isopentenyl pyrophosphate isomerases and enhanced carotenoid accumulation in a unicellular chlorophyte. *Proceedings of the National Academy of Sciences, USA* 95: 11 482–11 488.
- Ta MT, Kapterian TS, Fei W, Du X, Brown AJ, Dawes IW, Yang H. 2012. Accumulation of squalene is associated with the clustering of lipid droplets. *The FEBS Journal* 279: 4231–4244.
- Tomazic ML, Najle SR, Nusblat AD, Uttaro AD, Nudel CB. 2011. A novel sterol desaturase-like protein promoting dealkylation of phytosterols in *Tetrahymena thermophila*. *Eukaryotic Cell* 10: 423–434.
- Van Heukelem L, Thomas CS. 2001. Computer-assisted high-performance liquid chromatography method development with applications to the isolation and analysis of phytoplankton pigments. *Journal of Chromatography* 910: 31–49.
- Vinci G, Xia X, Veitia RA. 2008. Preservation of genes involved in sterol metabolism in cholesterol auxotrophs: facts and hypotheses. *PLoS ONE* 3: e2883.
- Volkman JK. 2003. Sterols in microorganisms. *Applied Microbiology and Biotechnology* 60: 495–506.
- Weete J, Abril M, Blackwell M. 2010. Phylogenetic distribution of fungal sterols. *PLoS ONE* 5: e10899.
- Wentzinger LF, Bach TJ, Hartmann M-A. 2002. Inhibition of squalene synthase and squalene epoxidase in tobacco cells triggers an up-regulation of 3-hydroxy-3-methylglutaryl coenzyme A reductase. *Plant Physiology* 130: 334–346.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Screenshot of the JGI genome browser of *P. tricornutum* (<http://genome.jgi-psf.org/Phatr2/Phatr2.home.html>) corresponding to the erroneously predicted locus *PHATRDRR FT_44478* (in dark blue) on chromosome 4.

Fig. S2 Alignment of the AA sequences of the reconstructed IDI-SQS fusion enzyme of diatoms with the corresponding orthologues of *Aureococcus anophagefferens* and *Ectocarpus siliculosus*.

Fig. S3 The *PrOSC* gene.

Fig. S4 Alignment of the AA sequences of the reconstructed *OSC* gene models of diatoms with those of plants (*A. thaliana*), animals (*H. sapiens*) and fungi (*S. cerevisiae*).

Fig. S5 EI-MS spectra of the main peaks of the chromatograms shown in Fig. 7 and that of the corresponding authentic standards.

Table S1 Primers used

Table S3 List of candidate genes screened for SQE activity

Table S2 Orthology scores of *P. tricornutum* MVA and sterol pathway enzymes

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.

AUREOCHROME1a-Mediated Induction of the Diatom-Specific Cyclin *dsCYC2* Controls the Onset of Cell Division in Diatoms (*Phaeodactylum tricornutum*)^W

Marie J.J. Huysman,^{a,b,c,d,e} Antonio E. Fortunato,^d Michiel Matthijs,^{a,b,c} Benjamin Schellenberger Costa,^f Rudy Vanderhaeghen,^{b,c} Hilde Van den Daele,^{b,c} Matthias Sachse,^g Dirk Inzé,^{b,c} Chris Bowler,^e Peter G. Kroth,^g Christian Wilhelm,^f Angela Falciatore,^d Wim Vyverman,^{a,1} and Lieven De Veylder^{b,c,1,2}

^aProtistology and Aquatic Ecology, Department of Biology, Ghent University, B-9000 Gent, Belgium

^bDepartment of Plant Systems Biology, VIB, B-9052 Gent, Belgium

^cDepartment of Plant Biotechnology and Bioinformatics, Ghent University, B-9052 Gent, Belgium

^dLaboratoire de Génomique des Microorganismes, Université Pierre et Marie Curie, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 7238, 75006 Paris, France

^eEnvironmental and Evolutionary Genomics Section, Institut de Biologie de l'Ecole Normale Supérieure, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 8186, Institut National de la Santé et de la Recherche Médicale U1024, Ecole Normale Supérieure, 75230 Paris cedex 05, France

^fDepartment of Plant Physiology, Institute of Biology, University of Leipzig, 04103 Leipzig, Germany

^gFachbereich Biologie, Universität Konstanz, Konstanz 78457, Germany

Cell division in photosynthetic organisms is tightly regulated by light. Although the light dependency of the onset of the cell cycle has been well characterized in various phototrophs, little is known about the cellular signaling cascades connecting light perception to cell cycle activation and progression. Here, we demonstrate that *diatom-specific cyclin 2 (dsCYC2)* in *Phaeodactylum tricornutum* displays a transcriptional peak within 15 min after light exposure, long before the onset of cell division. The product of *dsCYC2* binds to the cyclin-dependent kinase CDKA1 and can complement G1 cyclin-deficient yeast. Consistent with the role of *dsCYC2* in controlling a G1-to-S light-dependent cell cycle checkpoint, *dsCYC2* silencing decreases the rate of cell division in diatoms exposed to light-dark cycles but not to constant light. Transcriptional induction of *dsCYC2* is triggered by blue light in a fluence rate-dependent manner. Consistent with this, *dsCYC2* is a transcriptional target of the blue light sensor AUREOCHROME1a, which functions synergistically with the basic leucine zipper (bZIP) transcription factor bZIP10 to induce *dsCYC2* transcription. The functional characterization of a cyclin whose transcription is controlled by light and whose activity connects light signaling to cell cycle progression contributes significantly to our understanding of the molecular mechanisms underlying light-dependent cell cycle onset in diatoms.

INTRODUCTION

In eukaryotes, the presence of various cell cycle checkpoints ensures that the genetic information in a cell is inherited correctly by inhibiting the replication and distribution of incomplete or damaged chromosomes to the daughter cells. The major cell cycle checkpoints occur during the onset of DNA replication (G1-to-S transition) and mitosis (G2-to-M transition). During the mid-to-late G1 phase, most organisms exhibit a commitment point, before which a number of intra- and extracellular conditions must be fulfilled (Hartwell et al., 1974; Pardee, 1974; Spudich and Sager, 1980; Moulager et al., 2010). Beyond this commitment point, cells complete their cell cycle and become independent of mitogenic stimuli, such as growth

factors or nutrients and, in the case of phototrophs, light. In *Chlamydomonas reinhardtii*, the commitment point has been shown to be preceded by a primary arrest point in G1 at which cell cycle progression becomes light dependent (Spudich and Sager, 1980).

Despite the fact that light plays a key role in the growth of photoautotrophic organisms, as demonstrated by the light-driven expression of various cell cycle genes (Bisova et al., 2005; Moulager et al., 2007, 2010; López-Juez et al., 2008; Huysman et al., 2010; Moriyama et al., 2010), little is known about the cellular signaling mechanisms that connect light perception with the activation of the cell cycle machinery in the nucleus, which includes cyclin-dependent kinases (CDKs) and their interaction partners, the cyclins (CYCs) (Morgan, 1997; Inzé and De Veylder, 2006). In the green alga *Ostreococcus tauri*, cyclin A plays an important role during S phase entry. This gene, the first cell cycle gene to be transcribed in the organism after dawn, is translated in a cyclic adenosine monophosphate (cAMP)-dependent manner only when cells have acquired adequate levels of light energy, thereby reflecting the metabolic state of the cells (Moulager et al., 2010). In the red alga *Cyanidioschyzon merolae*, the inhibition of cyclin 1 degradation through

¹ These authors contributed equally to this work.

² Address correspondence to lieven.deveyllder@psb.vib-ugent.be.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Lieven De Veylder (lieven.deveyllder@psb.vib-ugent.be).

^W Online version contains Web-only data.

www.plantcell.org/cgi/doi/10.1105/tpc.112.106377

a tetrapyrrole-mediated signaling pathway during the shift from dark to light was shown to be crucial for connecting organellar and nuclear DNA replication (Kobayashi et al., 2011).

Here, we studied the molecular regulation of the light-dependent checkpoint in diatoms. Diatoms are unicellular algae that dominate primary production in many aquatic ecosystems and are responsible for about 20% of global photosynthetic carbon fixation (Van den Hoek et al., 1995; Field et al., 1998; Mann, 1999). Diatoms can grow and photosynthesize over a wide range of different light intensities and wavelengths (Holdsworth, 1985; Mercado et al., 2004), and they possess specific light sensing and acclimation strategies (Nymark et al., 2009; Bailleul et al., 2010; Park et al., 2010; Zhu and Green, 2010; Lepetit et al., 2012). Light quality and intensity not only determine the photosynthetic capacity of diatoms, but it also affects different cellular processes, including motility, sexual reproduction, and cell division (Brzezinski et al., 1990; Chen et al., 2004; Cohn et al., 2004; McLachlan et al., 2009; Mouget et al., 2009). Analogous to *C. reinhardtii*, the cell cycle of diatoms consists of light-dependent and -independent segments (Vaulot et al., 1986). The two major diatom groups, the centrics and the pennates (Kooistra et al., 2003; Sims et al., 2006), appear to have evolved different light-sensitive phases during their mitotic cell cycles. Flow cytometric analyses of dark-adapted cells have shown that in centric species, two light-sensitive stages are present during their cell cycle, namely, the G1 and G2/M phases (Olson et al., 1986; Vaulot et al., 1986; Brzezinski et al., 1990). Some pennate species have been reported to show a similar G1 and G2/M arrest, as reported for *Cylindrotheca fusiformis* (Brzezinski et al., 1990), while others display only a G1 arrest, as in *Phaeodactylum tricornutum* (Huysman et al., 2010) and *Seminavis robusta* (Gillard et al., 2008). For those species with only a light-dependent segment at the G1 phase, the immediate release of dark-arrested cells has proven to be a useful characteristic to synchronize and study the cell division process (Gillard et al., 2008; Huysman et al., 2010).

Although the light dependency of the diatom cell cycle was demonstrated more than 20 years ago (Olson et al., 1986; Vaulot et al., 1986; Brzezinski et al., 1990), to date, nothing is known about the molecular regulators that control the light-dependent cell cycle checkpoints in diatoms. In a previous study, we identified many members of the cyclin gene family in the pennate diatom *P. tricornutum* and the centric *Thalassiosira pseudonana* and described a class of diatom-specific cyclins involved in environmental signaling (Huysman et al., 2010). One of the most strongly and earliest expressed genes during the switch from dark to light in synchronized cells is the *diatom-specific cyclin2* (*dsCYC2*), hinting at a role for this cyclin in cell cycle activation after dark arrest. To address this hypothesis, we studied the role of *dsCYC2* at the light-dependent G1 checkpoint and investigated the light-dependent transcriptional regulation of this gene in *P. tricornutum*.

RESULTS

Light-Dependent Transcriptional and Translational Control of *dsCYC2*

As previously shown, the transcript level of *dsCYC2* changes abruptly upon exposure of dark-grown *P. tricornutum* cells to light

(Huysman et al., 2010). To document the kinetics of *dsCYC2* transcript and protein abundance upon illumination, we generated a transgenic marker line that expressed the full-length *dsCYC2* open reading frame (ORF) C-terminally fused to a hemagglutinin (HA) tag under the control of the *dsCYC2* promoter (p_{dsCYC2}), which we will refer to as the HA marker line (Figure 1A). To determine the kinetics of *dsCYC2* transcript levels after light exposure, we conducted a finely resolved sampling experiment during the first hour after illumination of dark-arrested cells. To this end, cells were grown exponentially under a 12-h-light/12-h-dark (12L/12D) regime and then transferred to the dark for a prolonged period (24 h) that, due to a light-dependent segment within the G1 phase, enriches cultures for G1 phase cells (Brzezinski et al., 1990; Huysman et al., 2010). When returned to light, cells progress synchronously through the cell cycle starting

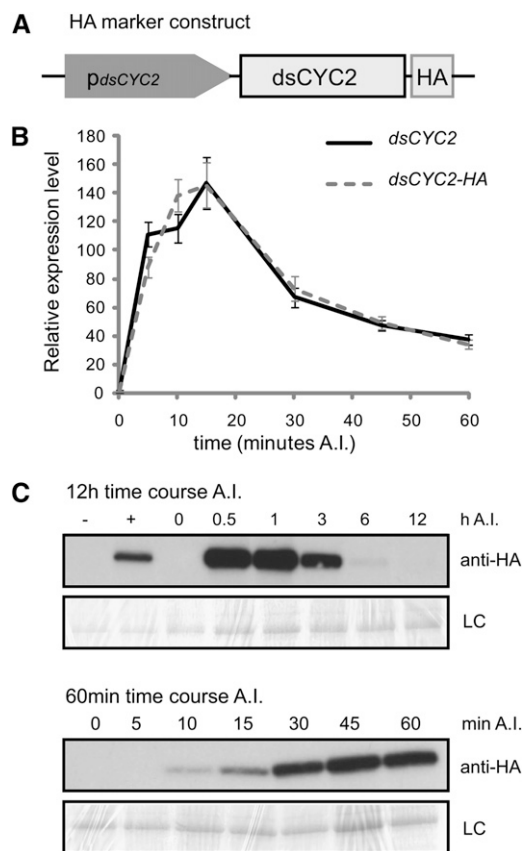


Figure 1. Light-Dependent Transcription and Translation of *dsCYC2*.

(A) Schematic representation of the HA marker construct.

(B) Transcript levels of *dsCYC2* (solid line) and HA-tagged *dsCYC2* (dashed line) during a 60-min time course after illumination (A.I.) of 24-h dark-adapted HA marker cells. Values were normalized against those obtained for *histone H4* and then rescaled to the gene expression levels at 0 min after illumination (=1). Error bars represent SE of three technical replicates.

(C) *dsCYC2*-HA protein levels during a 12-h (top panel) and 60-min (bottom panel) time course after illumination of 24-h dark-adapted HA marker cells. —, Negative control (wild-type 4 h light); +, positive control (HA 4 h light). LC, loading control by Coomassie blue staining.

from the G1 phase (Huysman et al., 2010). After illumination, samples were taken at 0, 5, 10, 15, 30, 45, and 60 min for real-time quantitative PCR to monitor *dsCYC2* transcript levels. An initial increase in transcript levels was observed after only 5 min of illumination, reaching a peak at 15 min, followed by a rapid decrease of the *dsCYC2* mRNA levels (Figure 1B). Protein gel blot analysis over a 12-h time course showed that *dsCYC2*-HA protein was undetectable immediately after illumination but reached high levels at 30 to 60 min, decreasing gradually thereafter to become undetectable by 12 h (Figure 1C). Protein analysis during the first hour after illumination showed increasing levels of *dsCYC2*-HA starting from 10 min until 60 min after light exposure (Figure 1C), although the transcript levels were markedly lower at the later time point (Figure 1B). These data show that upon illumination, *dsCYC2* transcript levels instantly reach a peak within 10 to 15 min, followed by a translational peak 30 to 60 min after light exposure.

***dsCYC2* Interacts with CDKA1 but Not CDKA2**

To explore the role of *dsCYC2* during the cell cycle, we tested whether *dsCYC2* can bind to the most conserved CDKs of *P. tricornutum*. To this end, we performed a yeast two-hybrid (Y2H) interaction assay in which *P. tricornutum* CDKA1, a G1/S-regulated cyclin-dependent kinase (CDK) containing the amino acid PSTAIRE motif, and CDKA2, a mitotically expressed CDK containing a PSTALRE motif (Huysman et al., 2010), were used as bait and *dsCYC2* as prey. Growth on selective His-lacking medium was observed for the combination of *dsCYC2* with CDKA1, but not with CDKA2 or any of the controls (Figure 2A). Complex formation between *dsCYC2* and CDKA1 is supported by their coexpression at the G1-to-S transition in synchronized cells (Huysman et al., 2010).

Complementation of a Conditional G1 Cyclin-Deficient Yeast Mutant by Expression of *dsCYC2*

The interaction of *dsCYC2* with CDKA1, and its peak in abundance during the early cell cycle, suggested that *dsCYC2* might

encode a G1-specific cyclin controlling the G1/S transition. Therefore, we examined whether *dsCYC2* is able to functionally substitute for yeast G1 cyclins using a complementation assay in the yeast strain BF305-15d-21. BF305-15d-21 cells contain mutations in the endogenous *cyclin1* (*CLN1*) and *cyclin2* (*CLN2*) genes and express *cyclin3* (*CLN3*) from a Gal-inducible promoter (Xiong et al., 1991). Hence, these cells are able to divide only in the presence of Gal. On Glc-containing medium, *CLN3* expression is repressed and cells arrest at a regulatory transition point in the G1 phase. BF305-15d-21 cells were transformed with the *pTH-dsCYC2* vector, containing the *dsCYC2* ORF under control of a doxycycline-repressible promoter. Cells containing *pTH-dsCYC2* were able to resume division in the presence of Glc (Figure 2B). When *dsCYC2* expression was repressed by doxycycline, complementation did not occur (Figure 2B), confirming that the complementation was linked to *dsCYC2* expression. These results demonstrate that *dsCYC2* encodes a functional cyclin that is able to complement a G1 cyclin-deficient yeast strain.

Silencing *dsCYC2* Slows Cell Cycle Progression by Prolonging the Light-Dependent G1 Phase

The early light-dependent transcription of *dsCYC2* suggests that its gene product plays a role in the reactivation of cell division upon illumination. To test this hypothesis, *P. tricornutum dsCYC2* knockdown lines were generated by introducing a hairpin construct under control of the constitutive *histone H4* promoter (De Riso et al., 2009) targeting the N-terminal region of *dsCYC2* (Figure 3A). Silencing was evaluated at 15 min after illumination by comparing *dsCYC2* transcript levels in wild-type cells and six independent transgenic lines harboring the RNA interference constructs. Two lines (*dscyc2*-2.4 and *dscyc2*-2.8) showed no silencing of *dsCYC2*, while four other lines (*dscyc2*-2.6, *dscyc2*-2.9, *dscyc2*-3.4, and *dscyc2*-3.5) showed a 40 to 75% reduction in transcript level compared with wild-type cells (Figure 3B). To test whether *dsCYC2* silencing had an effect on cell cycle progression, growth rate analysis was performed on wild-type

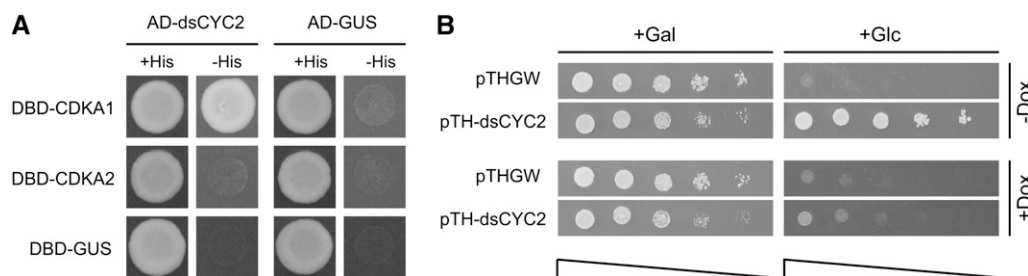


Figure 2. *dsCYC2* Functions as a G1-Cyclin.

(A) Interaction of *dsCYC2* with CDKA1. Yeast PJ694- α cells were cotransformed with bait and prey plasmid as indicated. Cotransformation was analyzed on medium lacking Leu and Trp (+His). Cotransformants were tested for their ability to activate the His marker gene by assessing yeast growth on medium lacking Leu, Trp, and His (-His). Constructs containing β -glucuronidase (GUS) were used as negative controls. For each combination, three independent colonies were screened, one of which is shown.

(B) Complementation of G1 cyclin-deficient yeast by *dsCYC2*. BF305-15d-21 cells were transformed with pTHGW (vector control) or pTH-*dsCYC2*. Yeast cells were serially diluted and spotted onto SD-Ura plates containing Gal (+Gal) or Glc (+Glc). When Glc was the sole carbon source, control cells were not able to grow because of the lack of G1 cyclin expression, while cells that expressed *dsCYC2* overcame this phenotype. When *dsCYC2* expression was repressed by the addition of doxycycline (+Dox), the complementation was lost.

and *dsCYC2* knockdown lines grown under a 12L/12D regime. No effect was observed for the nonsilenced internal control lines (*dscyc2-2.4* and *dscyc2-2.8*) (Figure 3C). By contrast, all knockdown lines (*dscyc2-2.6*, *dscyc2-2.9*, *dscyc2-3.4*, and *dscyc2-3.5*) showed a significant increase in generation time compared with wild-type cells (Figure 3C), indicating that *dsCYC2* is crucial for proper cell cycle progression.

Expression analysis of different cell cycle marker genes during the light-dependent cell cycle reentry of 24-h dark-arrested wild-type and *dscyc2-2.9* cells indicated that silencing of *dsCYC2* results in the attenuation of G1 progression upon light exposure. Expression of the early cell cycle marker genes *cyclin H1* (*CYCH1*) and the transcription factor *E2F1* was extended in the silenced versus wild-type cells (Figure 3D), indicating that cells with lower *dsCYC2* expression levels spend more time in the G1 phase and are delayed in the onset of the cell cycle upon illumination. Also, the timing of expression of the G2/M markers *CYCB1* and *MAD3* (Figure 3E) was clearly delayed in the *dscyc2-2.9* cells compared with wild-type cells. Thus, in addition to having an effect on cell cycle initiation at the G1 checkpoint after dark arrest, the

absence of *dsCYC2* expression also affects the timing of all downstream cell cycle transitions. The possibility that *dsCYC2* silencing affected transcription in general or produced a general stress response could be excluded, as several miscellaneous genes that were examined showed no differential expression in wild-type versus silenced lines (see Supplemental Figure 1 online).

If *dsCYC2* acts primarily at the light-dependent G1 checkpoint, no growth defects would be expected in *dscyc2* cells that do not experience a dark arrest. Therefore, we monitored the growth rates of wild-type and *dscyc2-2.9* cells grown under constant light conditions. Because cells grow faster and reach the stationary phase earlier in constant light compared with 12L/12D cycles, this experiment was performed at lower light intensities (50 μ E) to enable the detection of the exponential phase in the growth curves. While a clear reduction of cell growth rate was observed in *dscyc2-2.9* compared with wild-type cells grown in 12L/12D (Figure 3C), no significant difference in growth rate was observed when cells were grown in constant light (Table 1). However, when cells grown under continuous

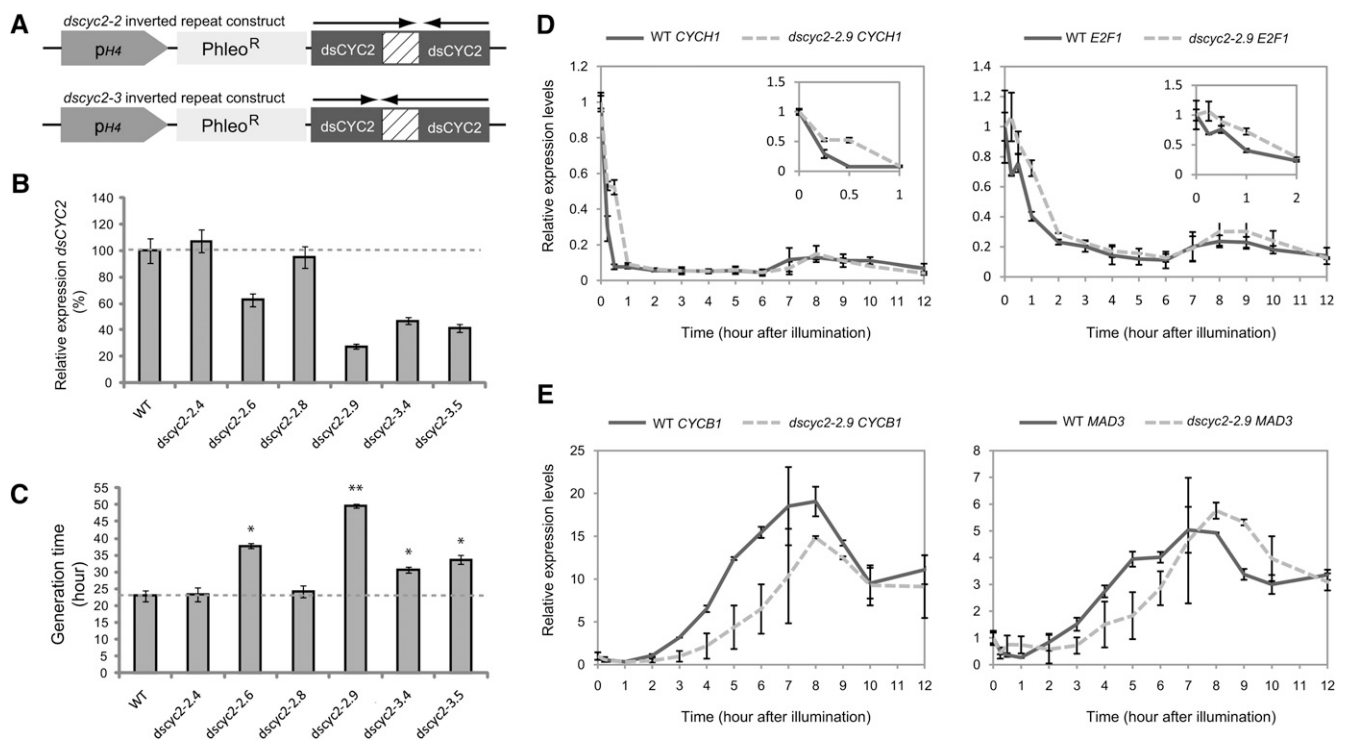


Figure 3. Effect of *dsCYC2* Silencing on Cell Cycle Progression.

(A) Schematic representation of the *dsCYC2* inverted repeat constructs used for silencing analysis. In the *dscyc2-2* construct, the large fragment is positioned first, followed by the small fragment. In the *dscyc2-3* construct, the small fragment is followed by the large fragment (arrows).

(B) Real-time quantitative PCR analysis of *dsCYC2* transcript levels in wild-type (WT) and silenced lines. Cells were dark adapted for 24 h, and transcript levels were measured 15 min after light exposure. Transcript levels of wild-type cells were set at 100%.

(C) Generation times of wild-type and *dsCYC2* silenced lines grown at 100 μ E 12L/12D cycles. Error bars (in **(B)** and **(C)**) represent sd of the mean of three independent experiments. * $P < 0.005$; ** $P < 0.001$ (two-tailed Student's t test).

(D) and **(E)** Transcript expression profiles of G1 marker genes **(D)** and mitotic markers **(E)** during a synchronized time course in wild-type and *dscyc2-2.9* knockdown cells. Error bars represent se of two biological replicates.

light were moved to 12L/12D conditions, the cells regained the growth phenotype within 3 d (Table 1). Since *P. tricornutum* cells are only light dependent at the G1 phase, these data suggest a primary role for *dsCYC2* at the G1 phase.

Wavelength and Fluence Rate Dependency of *dsCYC2* Transcription

To determine whether the light-regulated induction of *dsCYC2* transcripts is photoreceptor mediated, transcript levels were examined under different light conditions, including blue and red light at different fluence rates. As observed for white light, dark-adapted cells that were shifted to blue light showed an increase of *dsCYC2* transcript levels 10 min after light exposure, with a stronger effect at lower light intensities (Figure 4A). Although *dsCYC2* induction was lower under 90 μ E blue light, the transcript levels remained high for a longer period of time compared with lower light intensities. In contrast with blue light, exposing dark-adapted cells to red light did not result in major changes in *dsCYC2* transcript levels at either low or higher light intensities (Figure 4A).

To determine whether *dsCYC2* induction is solely photoreceptor mediated or, to some extent, also controlled by photosynthesis-mediated metabolic changes, the effect of the addition of DCMU during the light period was tested (Figure 4B). DCMU is a specific inhibitor of noncyclic photosynthetic electron transport (PET) and blocks the transfer of electrons from photosystem II to the plastoquinone pool. The addition of DCMU prior to blue light exposure had no effect on the induction of *dsCYC2* at 10 or 30 min after illumination (Figure 4B), suggesting that *dsCYC2* induction is photoreceptor mediated and not dependent on PET.

To assess the effects of light color and intensity on diatom growth and the role of *dsCYC2* under these conditions, the growth rates of wild-type and *dsCYC2* knockdown cells exposed to a 12L/12D photoperiod of white, blue, or red light adjusted to equal values of photosynthetically absorbed radiation (QPhar) were determined. Because *dsCYC2* is only induced when blue light is present, no difference in growth rate was expected to occur under red light conditions. Indeed, while *dsCYC2* knockdown cells grew more slowly than wild-type cells under white and blue light, no difference was observed between the control and transgenic cultures under red light (Figure 4C). These results support the role of blue light-induced *dsCYC2* expression during the cell cycle in *P. tricornutum* cells. In general, cells grown in red light showed a lower growth rate than those grown in blue or white light, highlighting the importance of blue light for diatom growth.

Table 1. Generation Times (h) of Wild-Type (WT) and *dsCYC2-2.9* Cells Grown in Constant White Light at 50 μ E or Shifted to a 12L/12D Cycle

Strain	LL	LD (Early, Days 1 and 2)	LD (Late, Days 3 to 5)
WT	29.5 \pm 3.4	29.2 \pm 0.0	42.3 \pm 3.9
<i>dsCYC2-2.9</i>	30.1 \pm 1.5	31.2 \pm 0.3	59.1 \pm 1.0

LD, light-dark cycle; LL, constant white light.

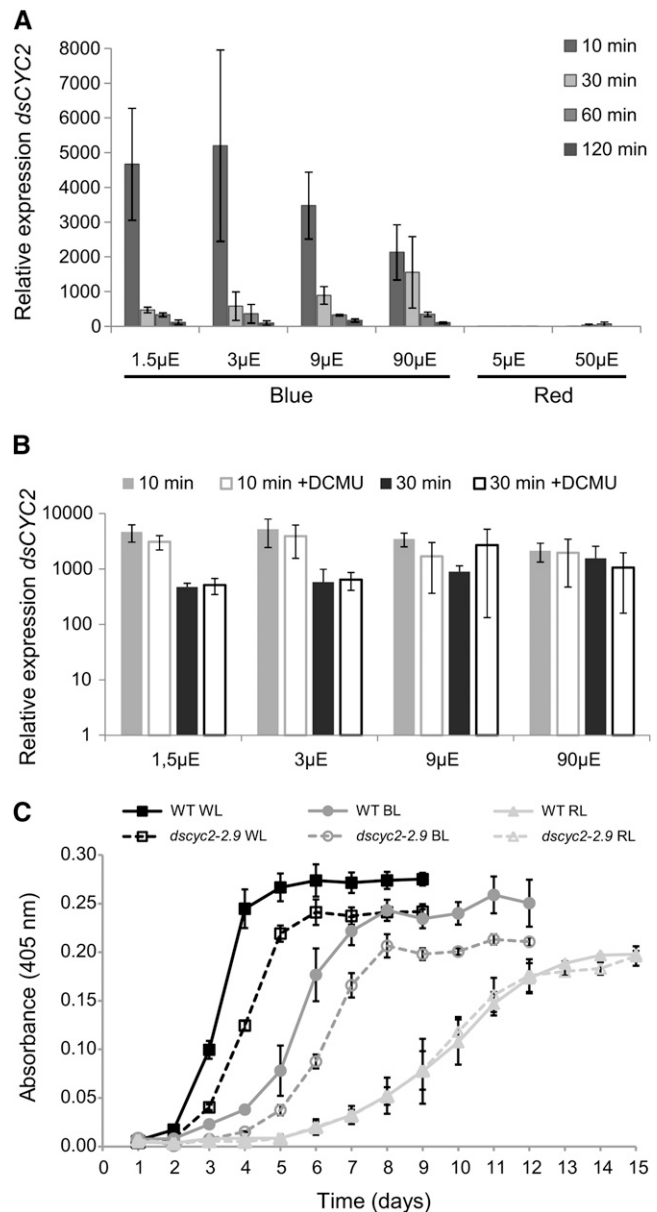


Figure 4. Blue Light Photoreceptor-Mediated Control of *dsCYC2* Induction.

(A) Wavelength and fluence rate dependency of *dsCYC2* induction. Wild-type cultures were dark incubated for 60 h and switched to blue or red light at different light intensities, as indicated. Relative mRNA levels of *dsCYC2* at 10, 30, 60, and 120 min after light exposure are shown. Relative levels were normalized to *histone H4* levels and rescaled to the expression level in dark-incubated cells (=1).

(B) Effect of DCMU on *dsCYC2* induction. Log scale representation of the relative mRNA levels of *dsCYC2* at 10 and 30 min after blue light exposure at different light intensities in the absence or presence of DCMU. In (A) and (B), error bars represent SE of two biological replicates.

(C) Growth curves of wild-type (WT) and *dsCYC2-2.9* cells grown in white (WL), blue (BL), and red (RL) light adjusted to equal values of photosynthetically absorbed radiation. Error bars represent SD of three biological replicates.

Regulation of the *dsCYC2* Promoter by Light

The observation that *dsCYC2* transcript levels were markedly affected by light suggests that either light has a direct effect on *dsCYC2* transcript stability or there is a yet undefined light-dependent signaling pathway that targets the *dsCYC2* promoter sequence. In order to distinguish between these possibilities, we analyzed the short-term expression kinetics of a reporter gene placed under control of the *dsCYC2* promoter during the light period following dark incubation. To construct this reporter fusion, we combined the 1018-bp region upstream of the translational start of *dsCYC2* (p_{dsCYC2}) and the coding region of enhanced yellow fluorescent protein (*eYFP*) (Figure 5A). Similar to *dsCYC2* transcript levels, *eYFP* transcript levels were induced shortly after light exposure and dropped again to basal levels after longer periods of illumination (Figure 5B). The slight delay in the decrease of *eYFP* transcript compared with the kinetics of the *dsCYC2* transcript (Figure 5B) is likely due to the higher intrinsic stability of *eYFP* versus *dsCYC2* mRNA. Nevertheless, the overall parallel kinetics of the endogenous *dsCYC2* transcript and the *eYFP* reporter transcript over time suggest that changes in *dsCYC2* mRNA are primarily a consequence of changes in promoter activity rather than transcript stability.

AUREOCHROME1a and bZIP10 Associate with the *dsCYC2* Promoter

To identify transcription factors that can bind to and regulate the *dsCYC2* promoter, a genome-wide yeast one-hybrid (Y1H) cDNA library screen was conducted. To this end, a p_{dsCYC2} reporter yeast strain was generated harboring the *dsCYC2* promoter upstream of the *HIS3* and the *LacZ* reporter genes, and this strain was transformed with a yeast-compatible *P. tri-comutum* cDNA library. The screen yielded two predicted basic leucine zipper (bZIP) transcription factors, AUREOCHROME1a (AUREO1a) and bZIP10. AUREO1a is a putative blue light photoreceptor that contains an N-terminal bZIP domain responsible for DNA binding and dimer formation and a C-terminal LOV (for light, oxygen, voltage) domain responsible for light sensing (Takahashi et al., 2007; Depauw et al., 2012). bZIP10 is a classical bZIP transcription factor (Rayko et al., 2010). Retransformation of AUREO1a and bZIP10 in the Y1H reporter strain confirmed their binding to the *dsCYC2* promoter, as indicated by auxotrophic growth on selective medium and the expression of the *LacZ* gene (Figure 5C).

Posttranslational Control of *dsCYC2* Induction

Light-dependent transcriptional induction of *dsCYC2* through the activation of the LOV domain of AUREO1a would be expected to occur without the need for de novo protein synthesis. To test this hypothesis, *dsCYC2* transcription was measured in wild-type cells treated with cycloheximide (CHX), an inhibitor of eukaryotic translation, just before illumination. As predicted, CHX treatment did not impair the light-dependent induction of *dsCYC2* (Figure 6A). Surprisingly, in contrast with the control cultures that showed a decrease of *dsCYC2* transcript levels following the initial transcriptional peak during the first hours

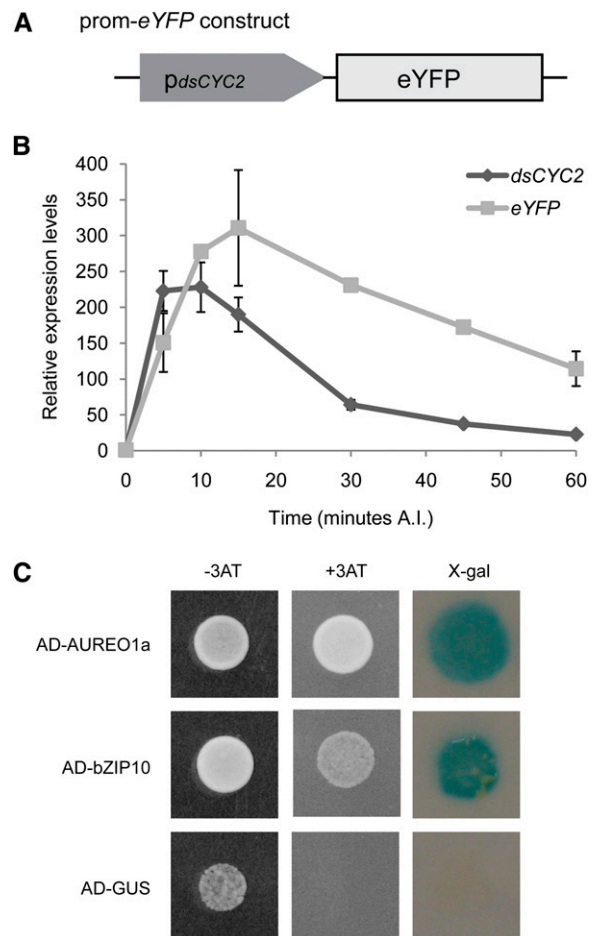


Figure 5. Regulation of the *dsCYC2* Promoter by Light.

(A) Schematic representation of the prom-*eYFP* marker (right) constructs.

(B) Transcript levels of *dsCYC2* and *eYFP* during a 60-min time course after illumination (A.I.) of 24-h dark-adapted p_{dsCYC2} -*eYFP* cells. Values were normalized against *H4* expression levels and rescaled to the levels at 0 min after illumination (= 1). Error bars represent SE of two biological replicates.

(C) Y1H protein-DNA interaction assay. Interactions are positive when *HIS3* (growth on 3-aminotriazole-containing medium (+3AT)) and *LacZ* (X-Gal turns blue) expression is induced. Constructs containing GUS were used as negative controls. For each combination, three independent colonies were screened, one of which is shown.

after light exposure, transcripts accumulated to high levels in the CHX-treated cultures (Figure 6A), suggesting that upon illumination, a repressor is produced de novo that specifically targets the promoter activity of *dsCYC2* to repress its expression. The addition of CHX during the dark period did not alter *dsCYC2* transcript levels (see Supplemental Figure 2 online), indicating that the effect of CHX was specific to light exposure. Together, these data corroborate the hypothesis of induction of *dsCYC2* transcription through activation of a photoreceptor, such as AUREO1a. Moreover, protein gel blot analysis demonstrated constitutive levels of AUREO1a during the switch from dark to

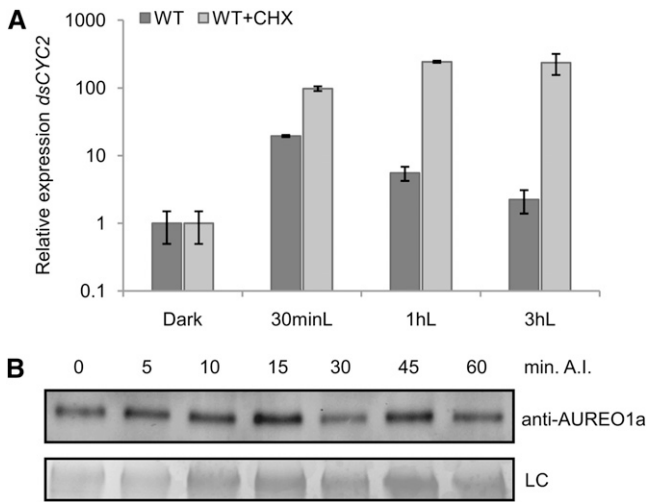


Figure 6. Posttranslational Regulation of *dsCYC2* Induction upon Light Exposure.

(A) Wild-type (WT) cultures were synchronized by 24-h dark treatment (Dark) and then exposed to light for 0.5 (30 minL), 1 (1 hL), or 3 h (3 hL) in the absence (dark gray) or presence (light gray) of 2 μ g/mL CHX. Relative expression levels of *dsCYC2* are shown. Values were normalized against *H4* expression levels and then rescaled to the gene expression levels of the dark sample (=1). Error bars represent SE of two independent experiments.

(B) AUREO1a protein levels during a 60-min time course after illumination (A.I.) of 24-h dark-adapted HA marker cells. LC, loading control by Coomassie blue staining.

light (Figure 6B), which suggests posttranslational activation of AUREO1a upon light exposure.

Activation of the *dsCYC2* Promoter by AUREO1a and bZIP10

Because bZIP proteins are known to function as homo- or heterodimers (Schütze et al., 2008), a Y2H interaction assay was performed to test whether AUREO1a and bZIP10 can interact with themselves or with each other. In this test, when AUREO1a was used as a bait, there were high levels of self-activation, precluding any conclusions about interactions. However, when bZIP10 was used as a bait, an interaction was found to occur with both bZIP10 and AUREO1a, as indicated by auxotrophic growth on His-lacking medium (Figure 7A).

To assess the effect of AUREO1a and bZIP10 on *dsCYC2* promoter activity, a transient activity assay was performed. AUREO1a and bZIP10 effector plasmids were transiently transformed either alone or together, along with a *p_{dsCYC2}:fLUC* reporter construct, into tobacco (*Nicotiana tabacum*) Bright Yellow-2 (BY-2) protoplast cells. When provided alone, both AUREO1a and bZIP10 slightly activated the *dsCYC2* promoter (Figure 7B). However, the activation effect was significantly increased when both effector plasmids were coexpressed (Figure 7B). These data suggest that AUREO1a and bZIP10 function in a synergistic manner to activate the *dsCYC2* promoter in response to light.

DISCUSSION

dsCYC2 Functions at the Light-Dependent G1 Checkpoint

For any photosynthetic organism, including diatoms, light is an extremely important factor that influences growth. Because diatoms can grow over a wide range of light intensities and wavelengths, these organisms are believed to have developed specific photoacclimation and photoadaptation mechanisms (Huisman et al., 2004; Lavaud et al., 2004; Lavaud et al., 2007). As with most other phytoplankton species, the timing of diatom cell division can be entrained by alternating periods of light and dark, implying that the cell cycle consists of light-dependent and -independent segments (Vaulot et al., 1986). Accordingly, both by light limitation and deprivation experiments, light-controlled restriction points have been identified in several diatom species, either during the G1 phase or during both the G1 and G2/M phases of the cell cycle (Olson et al., 1986; Vaulot et al., 1986; Brzezinski et al., 1990; Gillard et al., 2008; Huysman et al., 2010).

Previous work highlighted the role of *dsCYCs* in linking diverse environmental conditions to the cell cycle in diatoms

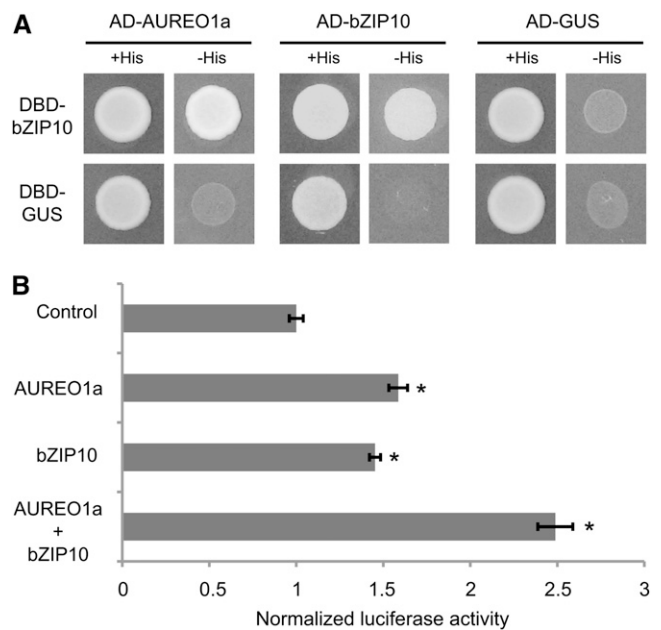


Figure 7. Activation of the *dsCYC2* Promoter by AUREO1a and bZIP10.

(A) Y2H protein-protein interaction assay. Yeast cells were cotransformed with bait and prey plasmid as indicated. Cotransformation was analyzed on medium lacking Leu and Trp (+His). Cotransformants were tested for their ability to activate the His marker gene by assessing yeast growth on medium lacking Leu, Trp, and His (-His). Constructs containing GUS were used as negative controls. For each combination, three independent colonies were screened, one of which is shown.

(B) Protoplast transactivation assay using *p_{dsCYC2}:fLUC* as reporter, *p35S::LUC* as normalization, and *p35S::AUREO1a* and *p35S::bZIP10* as effector constructs. Luciferase activity of the control was arbitrarily set to 1. Error bars represent SE of three biological replicates (* $P \leq 0.05$, two-sided *t* test).

(Bowler et al., 2008; Huysman et al., 2010). Here, we functionally characterized *dsCYC2* as a crucial regulator of cell cycle onset after a period of darkness in *P. tricornutum*. Upon light exposure, *dsCYC2* mRNA levels increase within minutes, followed by the induction of *dsCYC2* protein. The specific expression of *dsCYC2* during the G1 phase and its ability to complement G1 cyclin-deficient yeast cells suggest that *dsCYC2* operates early in the cell cycle. A role for *dsCYC2* in cell cycle entry is supported by the observation that lower *dsCYC2* levels following light exposure prolong the G1-to-S phase transition, as shown by the delayed and altered transcript levels of the G1 markers. Also G2/M marker genes displayed a delayed expression of about 1 to 2 h compared with wild-type cells. Thus, although *dsCYC2* likely acts primarily at the point of cell cycle onset, it appears that its effects go well beyond the early time points. From growth rate analysis, we determined that *dscyc2-2.9* cells have a generation time almost double that of wild-type cells; thus, we would have expected to observe a longer mitotic delay. Most likely this difference can be explained by the observation that ~10 to 15% of the *dscyc2-2.9* cells are not cycling but appear to be arrested at the S phase, as observed from DNA abundance measurements (see Supplemental Figure 3 online). Therefore, the observed longer generation time most likely results from the cumulative effect of a slower cell cycle progression at the G1 phase of the cycling cells and an S phase arrest of a subset of the cells. The increase in S phase cells suggests that a prime action of *dsCYC2* is to activate the CDK/cyclin complexes that are required for DNA replication, a function similar to that of the G1-specific *CLN1/2* and *cyclin E* genes in budding yeast and mammalian somatic cells, respectively (Morgan, 2007). Although *dsCYC2* protein can be detected during the S phase, it is unlikely that *dsCYC2* controls DNA replication itself, as no difference in generation time was observed between control and *dsCYC2* silenced lines under constant light conditions. Together with the observation that *dsCYC2* levels only peak at the G1/S transition under dark/light cycles, these data suggest that *dsCYC2* functions to relieve light-dependent G1 arrest, rather than regulating DNA replication.

In yeast, appropriate cell growth and metabolic status trigger G1 progression by activating *CLN3*, followed by *CLN1* and *CLN2*, in association with *CDC28* to finally activate the G1/S transcription factor *SBF/MBF* and the transcription of S phase genes (Tyers et al., 1993; reviewed in Mendenhall and Hodge, 1998). In animals and plants, D-type cyclins are stimulated by serum growth factors and hormones or Suc, respectively. D-type cyclins associate with CDKs and phosphorylate retinoblastoma (Rb) protein, leading to the release and activation of E2F transcription factors and the G1-to-S phase transition (reviewed in Oakenfull et al., 2002). Overexpression or silencing of these G1 cyclins has been reported to exhibit various effects on the G1 phase duration and overall cell cycle length, depending on the type of cyclin and cells (Quelle et al., 1993; Resnitzky et al., 1994; Sherr, 1995; Menges et al., 2006). Both *CLN1-3* and D-type cyclins are characterized by PEST sequences that render the proteins unstable and confer rapid turnover (Rechsteiner and Rogers, 1996; Renaudin et al., 1996; Mendenhall and Hodge, 1998). Furthermore, plant

and animal D-type cyclins, as well as the *Ostreococcus* cyclin A, possess an LxCxE amino acid motif at their N-terminal region that is responsible for their interaction with the Rb protein (Dowdy et al., 1993; Renaudin et al., 1996; Moulager et al., 2010). None of these motifs can be recognized in the *dsCYC2* sequence, suggesting that *dsCYC2* turnover is regulated by alternative mechanisms and that the protein probably does not interact directly with the *P. tricornutum* Rb orthologous protein. Alternatively, it is possible that *dsCYC2* expression results in the transcription or activation of other G1 cyclins that regulate the Rb protein in *P. tricornutum*. Because diatom cell cycle progression depends not only on light, but also on other environmental factors, such as nutrient availability, it is to be expected that multiple cyclins are involved in G1 control, representing a complex integrative fine-tuning network of different signaling pathways. The presence of a critical molecule, such as *dsCYC2*, that rapidly coordinates the activation of the cell cycle machinery upon changing light conditions is thus of major importance for diatoms living in highly variable environments and potentially allows them to pace their cell division rate to the prevailing light conditions.

Blue Light-Dependent Induction of *dsCYC2*

Promoter-reporter analysis suggests that transcriptional regulation of *dsCYC2* occurs through its promoter sequence. Screening for interactors of the *dsCYC2* promoter yielded two transcription factors belonging to the bZIP transcription factor family: AUREO1a and bZIP10. Of particular interest is AUREO1a, which belongs to the AUREOCHROME family of blue light photoreceptors in photosynthetic stramenopiles (Takahashi et al., 2007; Ishikawa et al., 2009). AUREOCHROMES typically possess an N-terminal bZIP domain expected to be involved in dimerization and DNA binding and a C-terminal LOV domain thought to act as a photosensor (Takahashi et al., 2007; Toyooka et al., 2011). Absorption of blue light by flavin mononucleotide (FMN) attached to the LOV domain induces covalent adduct formation between FMN and a conserved Cys residue in the LOV domain. *P. tricornutum* encodes four AUREOCHROME-like proteins (Rayko et al., 2010; Depauw et al., 2012), but only AUREO1a seems to be involved in *dsCYC2* regulation. In *Vaucheria frigida*, the bZIP domain of AUREO1 was found to recognize the bZIP binding site TGACGT (Jakoby et al., 2002; Takahashi et al., 2007). Interestingly, the promoter of *dsCYC2* contains three of these sites (see Supplemental Figure 4 online), rendering them putative regulatory *cis*-acting elements. The role of AUREO1a in *dsCYC2* induction is further supported by the specific transcription of *dsCYC2* by blue light, but not red light. Treatment of cells with CHX or the redox inhibitor DCMU had no effect on *dsCYC2* induction, indicating that no de novo protein synthesis or PET is required, reinforcing the hypothesis of direct photoreceptor-mediated regulation of *dsCYC2* induction by AUREO1a. The specific response of *dsCYC2* to low fluence rate blue light through AUREO1a signaling could be of particular significance to diatoms as blue light (350 to 500 nm) is the most prevalent color of light below the surface of oceanic waters (MacIntyre et al., 2000); hence, efficient blue light sensing and

signaling mechanisms are expected to play a crucial role in the control of diatom growth.

Various LOV domain–signaling mechanisms have been described for different plant, algal, and bacterial proteins, such as light-induced unfolding, rotation, dimerization, and/or DNA binding of the effector domain (reviewed in Herrou and Crosson, 2011). Here, we have shown that AUREO1a and bZIP10 can form heterodimers and that bZIP10 is able to form homodimers. Either protein could activate the *dsCYC2* promoter in the BY-2 protoplast system, but activation was enhanced when both proteins were coexpressed. Based on these findings, different models of *dsCYC2* regulation can be envisioned. First, upon blue light exposure, AUREO1a and bZIP10 might form heterodimers and as such bind and activate the regulatory sites present in the *dsCYC2* promoter. However, previous reports have suggested that the *V. frigida* AUREO1 LOV domain has a dimeric nature (Mitra et al., 2012) and that two LOV domains would be needed to activate AUREO1 (Toyooka et al., 2011). Therefore, it seems plausible that upon illumination, homodimers of AUREO1a and bZIP10 would occupy different regulatory sites within the *dsCYC2* promoter and act synergistically to activate it (Figure 8). How bZIP10 activates transcription of *dsCYC2* remains unknown, but possible mechanisms include nuclear translocation or post-translational modifications upon light exposure that result in the modulation of the DNA binding activity or activation potential, as described for other bZIP proteins (Jakoby et al., 2002). Further investigations are needed to uncover the precise mechanism of *dsCYC2* activation by AUREO1a and bZIP10.

Interestingly, inhibition of protein synthesis at the dark-to-light transition delays the decrease of *dsCYC2* transcript levels in the light and results in the accumulation of higher transcript levels after illumination, suggesting that upon light exposure, a repressor of *dsCYC2* transcription is generated. Such a repressor might interfere with DNA binding of the activators, either directly through, for example, occupation and repression of the regulatory sites, or indirectly by interfering with the dimerization or DNA binding properties of the activators through, for example, post-translational modifications (Schütze et al., 2008). Future work will focus on identifying the repressor(s) and their mode of regulation.

In conclusion, we identified two bZIP transcription factors that are likely to be involved in the blue light–dependent transcription of a cyclin gene that regulates the onset of the cell cycle in diatoms after a period of darkness. The involvement of aureochromes in blue light–mediated branching and sex organ development have previously been described (Takahashi et al., 2007). This study identifies a possible role for AUREO1a and its target gene *dsCYC2* during the cell cycle. The *dsCYC2* gene appears to be conserved in other pennate diatom species, including *Fragilariopsis cylindrus* (<http://genome.jgi-psf.org/Fracy1/Fracy1.home.html>, Fracy1_253344) and *Pseudo-Nitzschia multiseries* (<http://genome.jgi.doe.gov/Psemu1/Psemu1.home.html>, Psemu1_301178), but no clear homolog was found in the centric *T. pseudonana* (Huysman et al., 2010). However, because of the high number of *dsCYCs* in diatom species (Huysman et al., 2010) and the presence of AUREO1a in both pennates and centrics (Depauw et al., 2012), the mechanism of light-dependent cell cycle activation through AUREO1a-mediated induction of a cyclin gene is most likely conserved in diatoms.

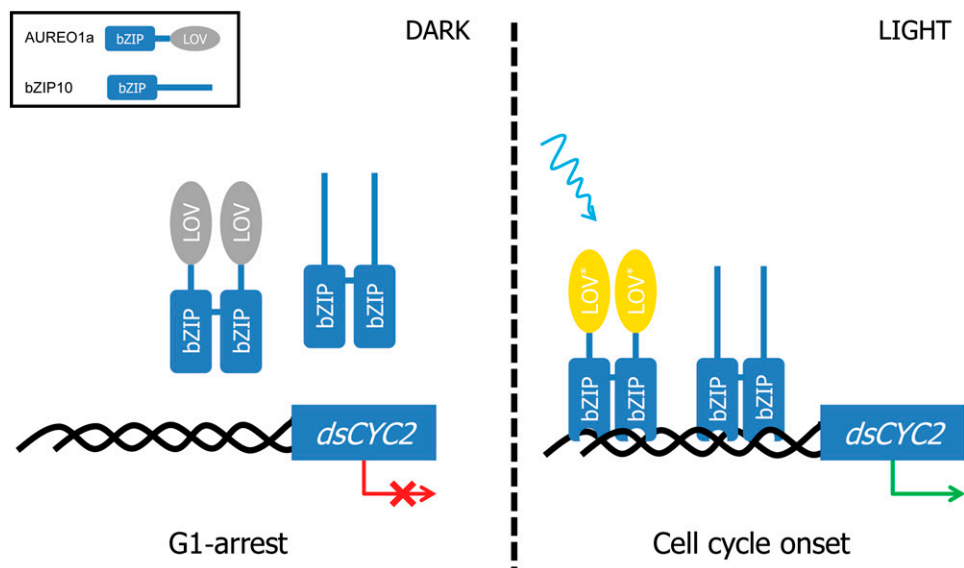


Figure 8. Hypothetical Model of the Light-Dependent Regulation of *dsCYC2* and Cell Cycle Onset in *P. tricornutum*.

Upon light exposure, the LOV domain of AUREO1a is changed from the dark state (gray) into the light state (yellow) through cysteinyl-FMN adduct formation. This induces a conformational change in the homodimer protein complex, resulting in the binding of the bZIP domains to the promoter of *dsCYC2*. Binding of both AUREO1a and bZIP10 homodimers to different regulatory elements in the *dsCYC2* promoter results in the synergistic activation of *dsCYC2* and leads to the onset of the cell cycle.

METHODS

Diatom Culture Conditions

Phaeodactylum tricornutum (Pt1 8.6; accession numbers CCAP 1055/1 and CCMP2561) cells were grown in f/2 medium without silica (f/2-Si) (Guillard, 1975) made from filtered and autoclaved sea water collected from the North Sea (Belgium) or artificial sea water medium (Vartanian et al., 2009). Cultures were cultivated at 18 to 20°C under a 12L/12D regime using 70 to 100 $\mu\text{mol photons m}^{-2} \text{ s}^{-1}$ white light (Radium NL 36W/840 Spectrolux plus, cool white). Liquid cultures were shaken at 100 rpm. For expression studies under blue and red light conditions, blue light (380 to 450 nm) was generated using neon lamps (Osram L36W/67, Lumilux Bleu), while red light (620 to 720 nm with a peak intensity of 670 nm) was generated using an LED source (Flight II DC Red + Black; Quantum Devices), and different light intensities were obtained using neutral density filters.

Vector Cloning and Biolistic Transformation

The 1018-bp promoter sequence alone, the promoter and full-length gene sequence, or the gene sequence alone of *dsCYC2* of *P. tricornutum* was amplified with gene-specific primers (see Supplemental Table 1 online), cloned in the pDONR221 or pENTR-D-TOPO vector (Invitrogen), and subsequently recombined in a *P. tricornutum* destination vector (pDEST) by attL \times attR recombination (Invitrogen) (Siaut et al., 2007). The *dsCYC2* promoter sequence was recombined in pDEST-C-eYFP for C-terminal fusion to construct the prom-eYFP reporter line. The promoter and gene constructs were recombined in pDEST-C-HA to construct the HA marker line. Both plasmids were subsequently digested with *SacII* and *NotI* (Promega) to remove the *fcpB* promoter sequence. The digested product was treated with T4 DNA polymerase in the presence of 10 mM deoxynucleotide triphosphate to produce blunt ends and then ligated using T4 DNA ligase according to the manufacturer's instructions (Promega). For the creation of *dsCYC2* inverted repeat silencing constructs, a 167-bp fragment (corresponding to the *dsCYC2* gene sequence from 13 to 179 bp) and a 301-bp fragment (corresponding to the gene sequence from 13 to 313 bp) were amplified from the *dsCYC2* cDNA with the primers *dsCYC2f1_Fw* (containing a *EcoRI* site) and *dsCYC2f1_Rv* (containing a *XbaI* site), and *dsCYC2f1_Fw* and *dsCYC2f2_Rv* (containing a *XbaI* site) (see Supplemental Table 1 online), respectively. The fragments were digested with *EcoRI* and *XbaI* (Promega) and ligated in sense and antisense orientations to the *EcoRI* site of the linearized hir-PtGUS vector (De Riso et al., 2009).

Constructs were introduced into *P. tricornutum* by microparticle bombardment as previously described (Falcatore et al., 1999). The prom-eYFP reporter plasmid and the HA marker plasmid were each co-transformed with the pAF6 plasmid to confer resistance to phleomycin (Falcatore et al., 1999). Individual phleomycin-resistant colonies were both restreaked on f/2-Si agar plates and grown in liquid f/2-Si medium without antibiotics for further analysis.

Real-Time Quantitative PCR

For RNA extraction, 5×10^7 cells were collected by fast filtration, and filters with cell pellets were fast frozen in liquid nitrogen and stored at -70°C . Cell lysis and RNA extraction were performed using TriReagent (Molecular Research Center) according to the manufacturer's instructions. Contaminating genomic DNA was removed by DNaseI treatment (GE Healthcare), and RNA was purified by ammonium acetate precipitation. RNA concentration and purity were assessed by spectrophotometry. Total RNA was reverse transcribed using iScript reverse transcriptase

(Bio-Rad) or a Quantitect reverse transcription kit (Qiagen) according to the manufacturer's instructions. Finally, an equivalent of 5 or 10 ng of reverse-transcribed RNA (cDNA) was used as template in each quantitative PCR reaction.

Samples in triplicate were amplified on the Lightcycler 480 platform (Roche) or the CFX96 Real-Time PCR detection system (Bio-Rad) with Lightcycler 480 SYBR Green I Master mix (Roche Applied Science) in the presence of 0.5 μM gene-specific primers (*dsCYC2_Fw*, 5'-CTATCA-TGCGACTCGTCATCAAC-3', and *dsCYC2_Rv*, 5'-TGTCCACCAAAGC-CTCCAAAC-3'; *dsCYC2-HA_Fw*, 5'-TCGCTCCTCTGGTGGAA-3', and *dsCYC2-HA_Rv*, 5'-GTCGTAGGGGTAGGCGTAGT-3'; for other primer sequences, see Siaut et al., 2007 and Huysman et al., 2010). The cycling conditions were 10 min polymerase activation at 95°C and 45 cycles at 95°C for 10 s, 58°C for 15 s, and 72°C for 15 s. Amplicon dissociation curves were recorded after cycle 45 by heating from 65 to 95°C . Data were analyzed using the ΔC_t (cycle threshold) relative quantification method using qBase (Hellemans et al., 2007), with the stably expressed *histone H4* used as a normalization gene (Siaut et al., 2007).

Protein Gel Blot Analysis

For protein extraction, 5×10^7 cells were collected by fast filtration, and filters with cell pellets were fast frozen in liquid nitrogen and stored at -70°C . Proteins were extracted by adding 200 μL Laemmli buffer containing Complete Protease Inhibitor Cocktail (Roche) to the frozen cells and vortexing at high speed until the cells were lysed. Cell lysates were incubated for 15 min on ice and centrifuged at 13,000 rpm for 15 min at 4°C to remove insoluble material. Protein concentrations were determined by the Bio-Rad Protein Assay (Bio-Rad) based on the method of Bradford (1976). Equal amounts of protein extracts were resolved on 12% SDS-PAGE gels and transferred to nitrocellulose membranes (Millipore) using the wet-blot method. The *dsCYC2*-HA fusion protein was detected by incubating proteins transferred to nitrocellulose membranes for 1 h with a 1:500 dilution of anti-HA primary antibody (Roche) at room temperature, followed by 1 h incubation in a 1:10,000 dilution of horseradish peroxidase anti-rat secondary antibody (Abcam) at room temperature. AUREO1a protein was detected by incubating proteins transferred to nitrocellulose membranes for 1 h with a 1:1000 dilution of a specific anti-AUREO1a primary antibody at room temperature, followed by 1 h incubation in a 1:10,000 dilution of horseradish peroxidase anti-rabbit secondary antibody (GE Healthcare) at room temperature. Signals were visualized using the Western Lightning detection kit (Thermo Scientific Pierce) according to the manufacturer's instructions.

Y2H Analysis

Y2H bait and prey plasmids were generated through recombinational Gateway cloning (Invitrogen). The full-length ORFs of the *P. tricornutum* *dsCYC2*, *CDKA1*, and *CDKA2* genes were amplified from cDNA using gene-specific primers (see Supplemental Table 1 online), cloned in the pENTR-D-TOPO vector (Invitrogen), and subsequently recombined in the pDEST22 and pDEST32 vectors (Invitrogen) by attL \times attR recombination, resulting in translational fusions between the proteins and the GAL4 transcriptional activator and DNA binding domains, respectively. AUREO1a and bZIP10 cDNAs were derived from plasmid extraction (Zymo-prep; Zymo Research) from positive colonies of a Y1H library screen (see below) and recombined in the pDEST22 and pDEST32 vectors by Gateway recombination. Bait and prey plasmids were cotransformed in the yeast strain PJ694- α by the LiAc method (Gietz et al., 1992). Co-transformed yeast cells were selected on synthetic defined (SD) medium plates lacking Leu and Trp. Interaction between the introduced proteins was scored on SD plates lacking Leu, Trp, and His.

Yeast Complementation

The full-length *dsCYC2* cDNA was cloned into the yeast tetracycline-repressible vector pTHGW (Peres et al., 2007) by LR cloning. The resulting plasmid, pTH-*dsCYC2* (or pTHGW as a control), was transformed into the G1-deficient yeast strain BF305-15d-21 (MATa *leu2-3*, 112his3-11, 15ura3-52 *trp1 ade1 met14*; *arg5,6 GAL1-CLN3 HIS3::cln 1 TRP1::cln2*) by the LiAc method (Gietz et al., 1992). Complementation was assayed on Gal-containing SD medium in the presence or absence of 20 $\mu\text{g/mL}$ doxycycline (Sigma-Aldrich).

Y1H Analysis

For the Y1H library screen, the *dsCYC2* promoter sequence (1018 bp upstream of ATG) was cloned in the pMW#2 and pMW#3 destination vectors (Deplancke et al., 2006), yielding *HIS3* and *LacZ* reporter constructs, respectively. The Y1H bait strain was generated as previously described (Deplancke et al., 2004, 2006). Subsequently, the Y1H bait strain was transformed with 50 μg of prey plasmids derived from a custom-made *P. tricornutum* Y2H cDNA library (Invitrogen) according to the Yeast Protocol Handbook (Clontech), and yeast cells that hosted a successful interaction were selected on selective SD medium lacking His, Ura, and Trp containing 25 mM of 3-amino-1,2,4-triazole and retested using a direct Y1H test.

Growth Analyses

To monitor growth rates of wild-type and *dscyc2* cells, cells were grown at 12L/12D (100 μE of white light) in a 24-well plate (Falcon), in a total volume of 1 mL, over a time period of 11 d. Absorbances of the cultures were measured at 405 nm using the VICTOR³ Multilabel Plate Reader (Perkin-Elmer) each day in the morning. The growth curves of triplicate cultures were LN(2) transformed, and mean generation times were calculated by determination of the derivative of the values between the points of maximal slope (exponential growth phase).

To determine the growth rates of wild-type and *dscyc2*-2.9 cells under constant light conditions, batch cultures were grown under continuous illumination of white light for at least 2 weeks. At the beginning of the experiment, cells were diluted with fresh F/2 medium without silica to the same absorbance at 405 nm (0.025), and cells were either placed under constant light conditions or transferred to 12L/12D conditions at the same light intensity (50 μE). Generation times were calculated as described above.

To monitor growth curves under different light quality conditions, wild-type and *dscyc2*-2.9 batch cultures were cultivated at 12L/12D at 20°C in air-lifted 100-mL test tubes. Flora light-emitting diode panels (CLF Plant Climatics) were used for illumination with monochromatic blue light and red light at wavelengths of 469 nm \pm 10 nm and 659 nm \pm 11 nm, respectively. White fluorescence tubes (18W/865; Osram) provided illumination with white light. The spectral composition of the light sources was recorded with a spectroradiometer (Tristan). The relative absorption of incident light varied for the different light sources. Therefore, the incident light intensity was adjusted to either 72 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ blue light, 120 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ white light, or 123 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ red light. This resulted in similar values of 30 $\mu\text{mol absorbed photons m}^{-2} \text{s}^{-1}$ photosynthetically absorbed radiation (Q_{PAR}) at a culture density of 2 $\mu\text{g chlorophyll a mL}^{-1}$ as calculated according to Gilbert et al. (2000). All cultures were inoculated with 50,000 cells mL^{-1} . The in vivo absorption at 405 nm was recorded with a spectrophotometer (Specord M500; Zeiss).

nCounter Analysis

RNA levels were measured using the Nanostring nCounter analysis system (Nanostring Technologies) by the VIB Nucleomics Core Facility as

previously described (Geiss et al., 2008). An overview of the nCounter probe pairs used in this study is shown in Supplemental Table 2 online. All probes were screened against the *P. tricornutum* annotated transcript database from the Department of Energy Joint Genome Initiative for potential cross-hybridization. Total RNA extract (100 ng) from two biological replicates for both wild-type and *dscyc2*-2.9 cells was used for hybridization, and all genes were measured simultaneously using multiplexed reactions. After a first normalization against the internal spike-in controls, genes were normalized against the four reference genes *EF1a*, *histone H4*, *RPS*, and *UBI-4*. Fold induction calculations for wild-type and *dscyc2*-2.9 cells values were divided by the value at the 0-h time point.

Inhibitor Studies

To determine the effect of PET inhibition, DCMU (Sigma-Aldrich) was dissolved in ethanol to a stock concentration of 100 mM and delivered to the cells 10 min before the onset of light treatment at a final concentration of 20 μM . Identical volumes of ethanol were added to the controls and had no effect on transcript expression.

To determine the effect of inhibition of protein translation on *dsCYC2* transcription, cells were treated with or without CHX (Duchefa Biochemie) at a final concentration of 2 $\mu\text{g/mL}$, 5 min before the onset of light treatment.

Transient Reporter Assays

The *dsCYC2* promoter sequence was cloned simultaneously with the fLUC sequence in the pm42GW7,3 destination vector (Karimi et al., 2007) by multisite Gateway cloning (Invitrogen). To generate the effector constructs, the cDNA clones of *AUREO1a* and *bZIP10* were recombined in the p2GW7 destination vector by Gateway cloning, containing the cauliflower mosaic virus 35S promoter. Both reporter and effector plasmids were used to transfect tobacco (*Nicotiana tabacum*) BY-2 protoplasts using the polyethylene glycol/ Ca^{2+} method as described by De Sutter et al. (2005). Luciferase measurements were performed using the Dual-luciferase Reporter 1000 Assay System (Promega) according to the manufacturer's instructions and as previously described (De Sutter et al., 2005).

Accession Numbers

Sequence data from this article can be found in the *P. tricornutum* genome sequence database through the Joint Genome Initiative portal (<http://genome.jgi-psf.org/Phatr2/Phatr2.home.html>) under the following accession numbers: *dsCYC2*, Phatr2_34956; *CDKA1*, Phatr2_20262; *CDKA2*, Phatr2_51279; *AUREO1a*, Phatr2_49116; and *bZIP10*, Phatr2_43744. Sequence data from other genes discussed in this article can be found in the EMBL/GenBank data libraries under the accessions numbers listed in Supplemental Table 2 online.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Silencing of *dsCYC2* Does Not Result in a General Stress Response.

Supplemental Figure 2. Effect of CHX on *dsCYC2* Transcript Expression in Dark and Light.

Supplemental Figure 3. S-Phase Distribution in Wild-Type (Solid Line) versus *dscyc2*-2.9 (Dashed Line) Cells during a Synchronized Time Course.

Supplemental Figure 4. Mapping of TGACGT Sites in the *dsCYC2* Promoter Sequence.

Supplemental Table 1. Overview of the Cloning Primers.

Supplemental Table 2. Overview of the nCounter Code Set Probe Pairs.

ACKNOWLEDGMENTS

We thank Frederik Coppens, Joke Allemeersch, and Rudy Van Eijsden for technical advice and assistance, Jonas Van Hove, Leila Tirichine, and Frauke Depauw for practical assistance, and Martine De Cock and Annick Bleys for help in preparing the article. We thank Bruce Futcher and Jim Murray for providing the BF305-15d-21 yeast strain. M.J.J.H. and M.M. thank the Agency for Innovation by Science and Technology in Flanders (IWT-Vlaanderen) for a predoctoral fellowship. This work was supported by a grant of the Research Foundation Flanders (G.0288.13). M.J.J.H. acknowledges the Federation of European Biochemical Societies (FEBS) organization for a short-term fellowship to visit A.F.'s lab at Université Pierre et Marie Curie and the European Molecular Biology Organization (EMBO) organization for a short-term fellowship (ASTF 93-2011) to visit C.B.'s lab at Institut de Biologie de l'Ecole Normale Supérieure (IBENS) Paris (France). C.B. acknowledges support from the Agence Nationale de Recherche. P.G.K. is grateful for financial support by the Deutsche Forschungsgemeinschaft (research group 1261, project 8) and the Universität Konstanz. C.W. and B.S. acknowledge the support from the Deutsche Forschungsgemeinschaft Grant FOR 1261 (Wi 764/19). A.F., A.E.F., and M.J.J.H. acknowledge support by the Human Frontier Science Program Young Investigator Grant (RGY0082/2010) and the Action Thématique et Incitative sur Programme award (2009) from Centre National de la Recherche Scientifique.

AUTHOR CONTRIBUTIONS

M.J.J.H., A.E.F., B.S.C., D.I., C.B., P.G.K., C.W., A.F., W.V., and L.D.V. conceived and designed the research. M.J.J.H., A.E.F., M.M., B.S.C., R.V., H.V.d.D., and M.S. performed the experiments. M.J.J.H., W.V., and L.D.V. analyzed the data and wrote the article. All authors read, revised, and approved the article.

Received October 19, 2012; revised December 5, 2012; accepted December 18, 2012; published January 4, 2013.

REFERENCES

- Bailleul, B., Rogato, A., de Martino, A., Coesel, S., Cardol, P., Bowler, C., Falcatore, A., and Finazzi, G. (2010). An atypical member of the light-harvesting complex stress-related protein family modulates diatom responses to light. *Proc. Natl. Acad. Sci. USA* **107**: 18214–18219.
- Bisova, K., Krylov, D.M., and Umen, J.G. (2005). Genome-wide annotation and expression profiling of cell cycle regulatory genes in *Chlamydomonas reinhardtii*. *Plant Physiol.* **137**: 475–491.
- Bowler, C., et al. (2008). The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**: 239–244.
- Bradford, M.M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**: 248–254.
- Brzezinski, M.A., Olson, R.J., and Chisholm, S.W. (1990). Silicon availability and cell-cycle progression in marine diatoms. *Mar. Ecol. Prog. Ser.* **67**: 83–96.
- Chen, M., Chory, J., and Fankhauser, C. (2004). Light signal transduction in higher plants. *Annu. Rev. Genet.* **38**: 87–117.
- Cohn, S.A., Bahena, M., Davis, J.T., Ragland, R.L., Rauschenberg, C.D., and Smith, B.J. (2004). Characterisation of the diatom photophobic response to high irradiance. *Diatom Res.* **19**: 167–179.
- Depauw, F.A., Rogato, A., Ribera d'Alcalá, M., and Falcatore, A. (2012). Exploring the molecular basis of responses to light in marine diatoms. *J. Exp. Bot.* **63**: 1575–1591.
- Deplancke, B., Dupuy, D., Vidal, M., and Walhout, A.J. (2004). A Gateway-compatible yeast one-hybrid system. *Genome Res.* **14**: 2093–2101.
- Deplancke, B., Vermeirssen, V., Arda, H.E., Martinez, N.J., and Walhout, A.J. (2006). Gateway-compatible yeast one-hybrid screens. *CSH Protoc.* **2006**: 5.
- De Riso, V., Raniello, R., Maumus, F., Rogato, A., Bowler, C., and Falcatore, A. (2009). Gene silencing in the marine diatom *Phaeodactylum tricornutum*. *Nucleic Acids Res.* **37**: e96.
- De Sutter, V., Vanderhaeghen, R., Tillemans, S., Lammertyn, F., Vanhoutte, I., Karimi, M., Inzé, D., Goossens, A., and Hilson, P. (2005). Exploration of jasmonate signalling via automated and standardized transient expression assays in tobacco cells. *Plant J.* **44**: 1065–1076.
- Dowdy, S.F., Hinds, P.W., Louie, K., Reed, S.I., Arnold, A., and Weinberg, R.A. (1993). Physical interaction of the retinoblastoma protein with human D cyclins. *Cell* **73**: 499–511.
- Falcatore, A., Casotti, R., Leblanc, C., Abrescia, C., and Bowler, C. (1999). Transformation of nonselectable reporter genes in marine diatoms. *Mar. Biotechnol. (NY)* **1**: 239–251.
- Field, C.B., Behrenfeld, M.J., Randerson, J.T., and Falkowski, P. (1998). Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**: 237–240.
- Geiss, G.K., et al. (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat. Biotechnol.* **26**: 317–325.
- Gietz, D., St Jean, A., Woods, R.A., and Schiestl, R.H. (1992). Improved method for high efficiency transformation of intact yeast cells. *Nucleic Acids Res.* **20**: 1425.
- Gilbert, M., Wilhelm, C., and Richter, M. (2000). Bio-optical modeling of oxygen evolution using in vivo fluorescence: Comparison of measured and calculated photosynthesis/irradiance (P-I) curves in four representative phytoplankton species. *J. Plant Physiol.* **157**: 307–314.
- Gillard, J., et al. (2008). Physiological and transcriptomic evidence for a close coupling between chloroplast ontogeny and cell cycle progression in the pennate diatom *Seminavis robusta*. *Plant Physiol.* **148**: 1394–1411.
- Guillard, R.R.L. (1975). Culture of phytoplankton for feeding marine invertebrates. In *Culture of Marine Invertebrate Animals*, W.L. Smith and M.H. Canley, eds (New York: Plenum Press), pp. 29–60.
- Hartwell, L.H., Culotti, J., Pringle, J.R., and Reid, B.J. (1974). Genetic control of the cell division cycle in yeast. *Science* **183**: 46–51.
- Hellemans, J., Mortier, G., De Paepe, A., Speleman, F., and Vandesompele, J. (2007). qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol.* **8**: R19.
- Herrou, J., and Crosson, S. (2011). Function, structure and mechanism of bacterial photosensory LOV proteins. *Nat. Rev. Microbiol.* **9**: 713–723.
- Holdsworth, E.S. (1985). Effect of growth factors and light quality on the growth, pigmentation and photosynthesis of two diatoms, *Thalassiosira gravida* and *Phaeodactylum tricornutum*. *Mar. Biol.* **86**: 253–262.
- Huisman, J., Sharples, J., Stroom, J.M., Visser, P.M., Kardinaal, W.E.A., Verspagen, J.M.H., and Sommeijer, B. (2004). Changes

- in turbulent mixing shift competition for light between phytoplankton species. *Ecology* **85**: 2960–2970.
- Huysman, M.J.J., Martens, C., Vandepoele, K., Gillard, J., Rayko, E., Heijde, M., Bowler, C., Inzé, D., Van de Peer, Y., De Veylder, L., and Vyverman, W. (2010). Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome Biol.* **11**: R17.
- Inzé, D., and De Veylder, L. (2006). Cell cycle regulation in plant development. *Annu. Rev. Genet.* **40**: 77–105.
- Ishikawa, M., Takahashi, F., Nozaki, H., Nagasato, C., Motomura, T., and Kataoka, H. (2009). Distribution and phylogeny of the blue light receptors aureochromes in eukaryotes. *Planta* **230**: 543–552.
- Jakoby, M., Weisshaar, B., Droge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J., Kroj, T., and Parcy, F.; bZIP Research Group (2002). bZIP transcription factors in *Arabidopsis*. *Trends Plant Sci.* **7**: 106–111.
- Karimi, M., Depicker, A., and Hilson, P. (2007). Recombinational cloning with plant Gateway vectors. *Plant Physiol.* **145**: 1144–1154.
- Kobayashi, Y., Imamura, S., Hanaoka, M., and Tanaka, K. (2011). A tetrapyrrole-regulated ubiquitin ligase controls algal nuclear DNA replication. *Nat. Cell Biol.* **13**: 483–487.
- Kooistra, W.H., De Stefano, M., Mann, D.G., and Medlin, L.K. (2003). The phylogeny of the diatoms. *Prog. Mol. Subcell. Biol.* **33**: 59–97.
- Lavaud, J., Rousseau, B., and Etienne, A.-L. (2004). General features of photoprotection by energy dissipation in planktonic diatoms (Bacillariophyceae). *J. Phycol.* **40**: 130–137.
- Lavaud, J., Strzepek, R.F., and Kroth, P.G. (2007). Photoprotection capacity differs among diatoms: Possible consequences on the spatial distribution of diatoms related to fluctuations in the underwater light climate. *Limnol. Oceanogr.* **52**: 1188–1194.
- Lepetit, B., Goss, R., Jakob, T., and Wilhelm, C. (2012). Molecular dynamics of the diatom thylakoid membrane under different light conditions. *Photosynth. Res.* **111**: 245–257.
- López-Juez, E., Dillon, E., Magyar, Z., Khan, S., Hazeldine, S., de Jager, S.M., Murray, J.A.H., Beemster, G.T.S., Bögre, L., and Shanahan, H. (2008). Distinct light-initiated gene expression and cell cycle programs in the shoot apex and cotyledons of *Arabidopsis*. *Plant Cell* **20**: 947–968.
- MacIntyre, H.L., Kana, T.M., and Geider, R.J. (2000). The effect of water motion on short-term rates of photosynthesis by marine phytoplankton. *Trends Plant Sci.* **5**: 12–17.
- Mann, D.G. (1999). The species concept in diatoms. *Phycologia* **38**: 437–495.
- McLachlan, D.H., Brownlee, C., Taylor, A.R., Geider, R.J., and Underwood, G.J.C. (2009). Light-induced motile responses of the estuarine benthic diatoms *Navicula perminuta* and *Cylindrotheca closterium* (Bacillariophyceae). *J. Phycol.* **45**: 592–599.
- Mendenhall, M.D., and Hodge, A.E. (1998). Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **62**: 1191–1243.
- Menges, M., Samland, A.K., Planchais, S., and Murray, J.A. (2006). The D-type cyclin CYCD3;1 is limiting for the G1-to-S-phase transition in *Arabidopsis*. *Plant Cell Biol.* **18**: 893–906.
- Mercado, J.M., Sánchez-Saavedra, M.P., Correa-Reyes, G., Lubián, L., Montero, O., and Figueroa, F.L. (2004). Blue light effect on growth, light absorption characteristics and photosynthesis of five benthic diatom strains. *Aquat. Bot.* **78**: 265–277.
- Mitra, D., Yang, X., and Moffat, K. (2012). Crystal structures of Aureochrome1 LOV suggest new design strategies for optogenetics. *Structure* **20**: 698–706.
- Morgan, D.O. (1997). Cyclin-dependent kinases: Engines, clocks, and microprocessors. *Annu. Rev. Cell Dev. Biol.* **13**: 261–291.
- Morgan, D.O. (2007). *The Cell Cycle: Principles of Control*. (London: New Science Press).
- Moriyama, T., Terasawa, K., Sekine, K., Toyoshima, M., Koike, M., Fujiwara, M., and Sato, N. (2010). Characterization of cell-cycle-driven and light-driven gene expression in a synchronous culture system in the unicellular rhodophyte *Cyanidioschyzon merolae*. *Microbiology* **156**: 1730–1737.
- Mouget, J.-L., Gastineau, R., Davidovich, O., Gaudin, P., and Davidovich, N.A. (2009). Light is a key factor in triggering sexual reproduction in the pennate diatom *Haslea ostrearia*. *FEMS Microbiol. Ecol.* **69**: 194–201.
- Moulager, M., Corellou, F., Vergé, V., Escande, M.-L., and Bouget, F.-Y. (2010). Integration of light signals by the retinoblastoma pathway in the control of S phase entry in the picophytoplanktonic cell *Ostreococcus*. *PLoS Genet.* **6**: e1000957.
- Moulager, M., Monnier, A., Jesson, B., Bouvet, R., Mosser, J., Schwartz, C., Garnier, L., Corellou, F., and Bouget, F.-Y. (2007). Light-dependent regulation of cell division in *Ostreococcus*: Evidence for a major transcriptional input. *Plant Physiol.* **144**: 1360–1369.
- Nymark, M., Valle, K.C., Brembu, T., Hancke, K., Winge, P., Andresen, K., Johnsen, G., and Bones, A.M. (2009). An integrated analysis of molecular acclimation to high light in the marine diatom *Phaeodactylum tricornutum*. *PLoS ONE* **4**: e7743.
- Oakenfull, E.A., Riou-Khamlichi, C., and Murray, J.A.H. (2002). Plant D-type cyclins and the control of G1 progression. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **357**: 749–760.
- Olson, R.J., Vulot, D., and Chisholm, S.W. (1986). Effects of environmental stresses on the cell cycle of two marine phytoplankton species. *Plant Physiol.* **80**: 918–925.
- Pardee, A.B. (1974). A restriction point for control of normal animal cell proliferation. *Proc. Natl. Acad. Sci. USA* **71**: 1286–1290.
- Park, S., Jung, G., Hwang, Y.S., and Jin, E. (2010). Dynamic response of the transcriptome of a psychrophilic diatom, *Chaetoceros neogracile*, to high irradiance. *Planta* **231**: 349–360.
- Peres, A., et al. (2007). Novel plant-specific cyclin-dependent kinase inhibitors induced by biotic and abiotic stresses. *J. Biol. Chem.* **282**: 25588–25596.
- Quelle, D.E., Ashmun, R.A., Shurtleff, S.A., Kato, J.Y., Bar-Sagi, D., Roussel, M.F., and Sherr, C.J. (1993). Overexpression of mouse D-type cyclins accelerates G₁ phase in rodent fibroblasts. *Genes Dev.* **7**: 1559–1571.
- Rayko, E., Maumus, F., Maheswari, U., Jabbari, K., and Bowler, C. (2010). Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. *New Phytol.* **188**: 52–66.
- Rechsteiner, M., and Rogers, S.W. (1996). PEST sequences and regulation by proteolysis. *Trends Biochem. Sci.* **21**: 267–271.
- Renaudin, J.-P., et al. (1996). Plant cyclins: A unified nomenclature for plant A-, B- and D-type cyclins based on sequence organization. *Plant Mol. Biol.* **32**: 1003–1018.
- Resnitzky, D., Gossen, M., Bujard, H., and Reed, S.I. (1994). Acceleration of the G1/S phase transition by expression of cyclins D1 and E with an inducible system. *Mol. Cell. Biol.* **14**: 1669–1679.
- Schütze, K., Harter, K., and Chaban, C. (2008). Post-translational regulation of plant bZIP factors. *Trends Plant Sci.* **13**: 247–255.
- Sherr, C.J. (1995). D-type cyclins. *Trends Biochem. Sci.* **20**: 187–190.
- Siaut, M., Heijde, M., Mangogna, M., Montsant, A., Coesel, S., Allen, A., Manfredonia, A., Falciatore, A., and Bowler, C. (2007). Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. *Gene* **406**: 23–35.
- Sims, P.A., Mann, D.G., and Medlin, L.K. (2006). Evolution of the diatoms: Insights from fossil, biological and molecular data. *Phycologia* **45**: 361–402.

- Spudich, J.L., and Sager, R.** (1980). Regulation of the *Chlamydomonas* cell cycle by light and dark. *J. Cell Biol.* **85**: 136–145.
- Takahashi, F., Yamagata, D., Ishikawa, M., Fukamatsu, Y., Ogura, Y., Kasahara, M., Kiyosue, T., Kikuyama, M., Wada, M., and Kataoka, H.** (2007). AUREOCHROME, a photoreceptor required for photomorphogenesis in stramenopiles. *Proc. Natl. Acad. Sci. USA* **104**: 19625–19630.
- Toyooka, T., Hisatomi, O., Takahashi, F., Kataoka, H., and Terazima, M.** (2011). Photoreactions of aureochrome-1. *Biophys. J.* **100**: 2801–2809.
- Tyers, M., Tokiwa, G., and Futcher, B.** (1993). Comparison of the *Saccharomyces cerevisiae* G₁ cyclins: Cln3 may be an upstream activator of Cln1, Cln2 and other cyclins. *EMBO J.* **12**: 1955–1968.
- Van den Hoek, C., Mann, D.G., and Jahns, H.M.** (1995). *Algae: An Introduction to Phycology*. (Cambridge, UK: Cambridge University Press).
- Vartanian, M., Desclés, J., Quinet, M., Douady, S., and Lopez, P.J.** (2009). Plasticity and robustness of pattern formation in the model diatom *Phaeodactylum tricornutum*. *New Phytol.* **182**: 429–442.
- Vaulot, D., Olson, R.J., and Chisholm, S.W.** (1986). Light and dark control of the cell cycle in two marine phytoplankton species. *Exp. Cell Res.* **167**: 38–52.
- Xiong, Y., Connolly, T., Futcher, B., and Beach, D.** (1991). Human D-type cyclin. *Cell* **65**: 691–699.
- Zhu, S.-H., and Green, B.R.** (2010). Photoprotection in the diatom *Thalassiosira pseudonana*: Role of LI818-like proteins in response to high light stress. *Biochim. Biophys. Acta* **1797**: 1449–1457.

AUREOCHROME1a-Mediated Induction of the Diatom-Specific Cyclin *dsCYC2* Controls the Onset of Cell Division in Diatoms (*Phaeodactylum tricornutum*)

Marie J.J. Huysman, Antonio E. Fortunato, Michiel Matthijs, Benjamin Schellenberger Costa, Rudy Vanderhaeghen, Hilde Van den Daele, Matthias Sachse, Dirk Inzé, Chris Bowler, Peter G. Kroth, Christian Wilhelm, Angela Falciatore, Wim Vyverman and Lieven De Veylder
Plant Cell 2013;25;215-228; originally published online January 4, 2013;
DOI 10.1105/tpc.112.106377

This information is current as of July 9, 2014

Supplemental Data	http://www.plantcell.org/content/suppl/2013/01/04/tpc.112.106377.DC1.html http://www.plantcell.org/content/suppl/2013/02/15/tpc.112.106377.DC2.html
References	This article cites 68 articles, 23 of which can be accessed free at: http://www.plantcell.org/content/25/1/215.full.html#ref-list-1
Permissions	https://www.copyright.com/ccc/openurl.do?sid=pd_hw1532298X&issn=1532298X&WT.mc_id=pd_hw1532298X
eTOCs	Sign up for eTOCs at: http://www.plantcell.org/cgi/alerts/ctmain
CiteTrack Alerts	Sign up for CiteTrack Alerts at: http://www.plantcell.org/cgi/alerts/ctmain
Subscription Information	Subscription Information for <i>The Plant Cell</i> and <i>Plant Physiology</i> is available at: http://www.aspb.org/publications/subscriptions.cfm

The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner–Doudoroff glycolytic pathway

Michele Fabris^{1,2,3}, Michiel Matthijs^{1,2,3}, Stephane Rombauts^{1,2}, Wim Vyverman³, Alain Goossens^{1,2} and Gino J.E. Baart^{1,2,3,*}

¹Department of Plant Systems Biology, VIB, B-9052 Gent, Belgium,

²Department of Plant Biotechnology and Bioinformatics, Ghent University, B-9052 Gent, Belgium, and

³Department of Biology, Laboratory of Protistology and Aquatic Ecology, Ghent University, B-9000 Gent, Belgium

Received 10 November 2011; revised 3 February 2012; accepted 10 February 2012; published online 1 April 2012.

*For correspondence (e-mail gino.baart@ugent.be).

SUMMARY

Diatoms are one of the most successful groups of unicellular eukaryotic algae. Successive endosymbiotic events contributed to their flexible metabolism, making them competitive in variable aquatic habitats. Although the recently sequenced genomes of the model diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* have provided the first insights into their metabolic organization, the current knowledge on diatom biochemistry remains fragmentary. By means of a genome-wide approach, we developed DiatomCyc, a detailed pathway/genome database of *P. tricornutum*. DiatomCyc contains 286 pathways with 1719 metabolic reactions and 1613 assigned enzymes, spanning both the central and parts of the secondary metabolism of *P. tricornutum*. Central metabolic pathways, such as those of carbohydrates, amino acids and fatty acids, were covered. Furthermore, our understanding of the carbohydrate model in *P. tricornutum* was extended. In particular we highlight the discovery of a functional Entner–Doudoroff pathway, an ancient alternative for the glycolytic Embden–Meyerhof–Parnas pathway, and a putative phosphoketolase pathway, both uncommon in eukaryotes. DiatomCyc is accessible online (<http://www.diatomcyc.org>), and offers a range of software tools for the visualization and analysis of metabolic networks and ‘omics’ data. We anticipate that DiatomCyc will be key to gaining further understanding of diatom metabolism and, ultimately, will feed metabolic engineering strategies for the industrial valorization of diatoms.

Keywords: diatoms, *Phaeodactylum tricornutum*, pathway/genome database, DiatomCyc, metabolism, Entner–Doudoroff pathway.

INTRODUCTION

Almost all of the world’s known petroleum reserves originated from biomass produced by eukaryotic phytoplankton (Falkowski *et al.*, 2005), among which diatoms are one of the most common groups. Diatoms account for approximately 20% of the global photosynthetic productivity (Field *et al.*, 1998), thereby driving oceanic carbon and silica cycles, and fueling aquatic food chains (Falkowski *et al.*, 1998). They possess the most efficient ribulose-1,5-bisphosphate carboxylase oxygenase (Rubisco) among autotrophs (Giordano *et al.*, 2005), and accumulate high-value compounds, such as pigments and oils, rich in unsaturated fatty acids (Chisti, 2007). Genome-wide analysis revealed that a remarkable number of the genes of the model diatom *Phaeodactylum tricornutum* have diverse origins (Whitaker *et al.*, 2009; Tirichine and Bowler, 2011). As a consequence, it possesses metabolic pathways from all domains of life, as illustrated by the recently described metazoan-like ornithine–urea cycle

(Allen *et al.*, 2011) that is typically absent in plants and green algae. Although several aspects of diatom metabolism have been explored already, a comprehensive overview is still lacking.

Genome-scale databases of metabolism have been constructed for several organisms, including *Escherichia coli* (Keseler *et al.*, 2011), *Homo sapiens* (Romero *et al.*, 2004), *Arabidopsis thaliana* (Radrach *et al.*, 2010) and *Chlamydomonas reinhardtii* (May *et al.*, 2009), and can be used to better understand cellular metabolism (Baart *et al.*, 2008; Chang *et al.*, 2011) in order to develop metabolic engineering strategies (Fong *et al.*, 2005; Smid *et al.*, 2005; Hua *et al.*, 2006), design culture media and processes (Teusink *et al.*, 2005; Baart *et al.*, 2007), and even to develop online process control (Provost and Bastin, 2004). Hence, the aim of this study was to generate a pathway/genome database (PGDB) of *P. tricornutum* based on its genome sequence and the

biochemical literature. Here, we present the features of PGDB, designated DiatomCyc, and discuss some specific metabolic pathways that have been discovered. Mining the *P. tricornutum* genome information led to the identification of unusual metabolic pathways, such as the typically prokaryotic Entner–Doudoroff pathway and a phosphoketolase pathway that, to date, has only been found in a few fungi.

RESULTS AND DISCUSSION

Creation of DiatomCyc

The available translated genomic sequence of *P. tricornutum* (Bowler *et al.*, 2008) was taken as the starting point for the reconstruction of the first diatom PGDB: DiatomCyc. To complete and improve the current genome annotation, we applied an orthology-based methodology (Notabaart *et al.*, 2006). Twenty-three genomes of organisms (i.e. reference organisms, see Experimental Procedures) were used for high-resolution predictions of translated gene sequence equivalency (Remm *et al.*, 2001; O'Brien *et al.*, 2005). Eleven out of the 23 organisms have a published genome-scale metabolic network and curated annotation. Gene functions of the reference organisms were conveyed to the corresponding *P. tricornutum* orthologs, for which we employed the 'PHATRDRAFT' terminology, which refers to the common gene identifiers used in databases such as NCBI and KEGG. Gene-to-function and function-to-reaction associations were transferred semi-automatically with the KEGG (Kanehisa and Goto, 2000) and MetaCyc (Caspi *et al.*, 2010) databases as input. The results were imported into PATHWAY TOOLS 15.0 (Karp *et al.*, 2002), and were subsequently refined. As for the other members of the MetaCyc family (Karp *et al.*, 2010), including the highly curated databases of *Saccharomyces cerevisiae*, *E. coli*, *H. sapiens* and *A. thaliana*, DiatomCyc is accessible online (<http://www.diatomcyc.org>), and offers a range of software tools for the visualization and querying of metabolic networks and the analysis of 'omics' data. DiatomCyc provides different levels of information, from the cellular metabolic overview to single gene, protein, reaction and metabolite information, and includes literature references. Through the web interface,

users can query the PGDB through keywords, the implemented BLAST utility and the genome browser. In addition, comparative analysis between different PGDBs can be carried out and the 'Omics Viewer' utility allows the graphical visualization of transcriptomics, proteomics and metabolomics data. Table 1 shows the types of data and the number of entries for each data type included in DiatomCyc, and compares these with data held in other species-specific MetaCyc-based databases, including EcoCyc (Keseler *et al.*, 2011), YeastCyc (Ball *et al.*, 2001), ChlamyCyc (May *et al.*, 2009), AraCyc (Zhang *et al.*, 2005) and HumanCyc (Romero *et al.*, 2004). Currently, DiatomCyc comprises 286 pathways with 1719 metabolic reactions and 1613 enzymes, which is comparable with other eukaryotic PGDBs, spanning both the central and parts of the secondary metabolism of *P. tricornutum* (Figure 1a; Table 1).

Pathway visualization and manual curation

Given the broad interest in diatoms as potential polyunsaturated oil and biodiesel producers, the metabolic pathways related to lipid biosynthesis (i.e. saturated and polyunsaturated fatty acids, triacylglycerols and phospholipid biosynthesis) are completely charted in DiatomCyc. As described below, the available MetaCyc templates were adjusted or refined manually when required. To illustrate the content of DiatomCyc, the fatty acid biosynthetic pathway was taken as an example. In *P. tricornutum* the most abundant saturated fatty acid is palmitic acid (C16:0) (Siron *et al.*, 1989; Zhukova and Aizdaicher, 1995; Domergue *et al.*, 2003; Patil *et al.*, 2007). Unlike mammals that use a single gene product to carry out the entire reaction set necessary to produce fatty acids, *P. tricornutum*, analogously to plants and bacteria, uses discrete proteins encoded by different genes involved in separate steps (<http://www.diatomcyc.org/DIATOM/NEW-IMAGE?type=PATHWAY&object=PWY-5156>). The pathway can be selected from the cellular overview of DiatomCyc (Figure 1a), and can be visualized in detail (Figure 1b). All reactions starting from acetyl-CoA to palmitate and stearate are associated with the corresponding enzymes and genes. By clicking on a specific reaction on the pathway (Figure 1c), the enzyme(s) and the associated gene(s) are shown, as well

Table 1 Main contents of DiatomCyc and other pathway/genome databases (PGDBs; with version between brackets)

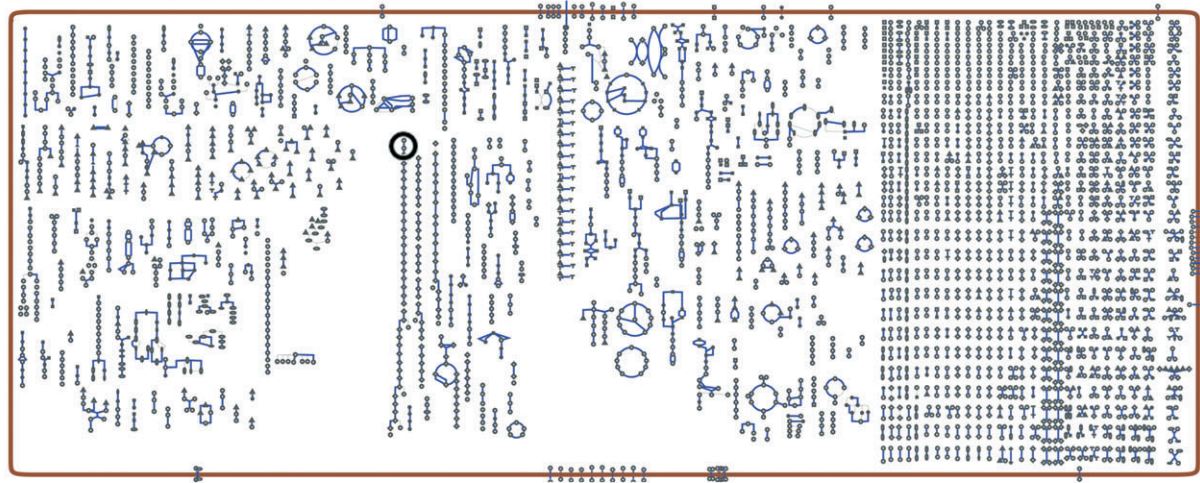
Content	MetaCyc	EcoCyc (15.1)	YeastCyc (15.0)	ChlamyCyc (2010-03-03)	AraCyc (8.0)	HumanCyc (15.1)	DiatomCyc (1.0)
Total genes	–	5305	8069	15 025	33 602	21 086	10 647
Protein coding	–	5115	6607	14 339	27 416	17 566	10 565
Pathways	1745	281	152	263	393	263	286
Enzymatic reactions	9458	1492	981	1419	2627	1866	1719
Enzymes	7424	1467	708	2851	3203	3623	1613
Compounds	9187	2161	688	1066	2825	1196	1173

(a) **DIATOMCYC**
A comprehensive database of diatom metabolism

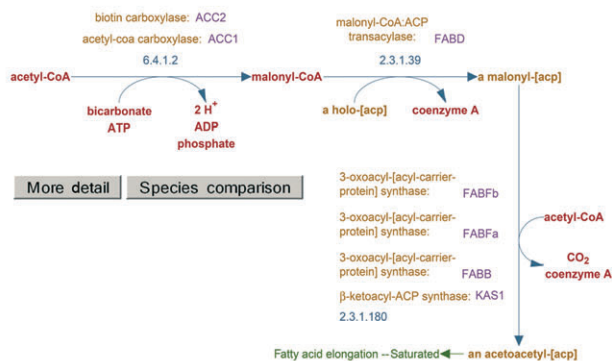
Quick search | Gene search
Searching *Phaeodactylum tricornutum* | change organism database

Home | Search | Tools | Help | Cellular Overview

Cellular overview of *Phaeodactylum tricornutum*



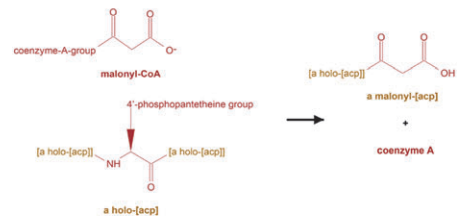
(b) *Phaeodactylum tricornutum* pathway: Fatty acid biosynthesis



(c) Reaction: 2.3.1.39

Enzymes and genes:

malonyl-CoA:ACP transacylase: **FABD**



Enzyme commission primary name for this reaction: [Acyl-carrier protein] S-malonyltransferase

Gene-reaction schematic:

Unification links: **BRENDA:2.3.1.39, ENZYME:2.3.1.39**

(d) Gene: **FABD** Accession number: PHATRDRRAFT_37652 (diatomCyc)

Protein sequence | Nucleotide sequence | Nucleotide sequence, advanced

Enzymatic reaction of: malonyl-CoA:ACP transacylase



The reaction direction shown, that is, $A + B \rightleftharpoons C + D$ versus $C + D \rightleftharpoons A + B$, is in accordance with the Enzyme Commission system.
The reaction is favored in the direction shown.

In Pathways: **fatty acid biosynthesis**

Created by: mifab on 7-Apr-2010

Gene Local Context (not to scale):

(e)

```
>PHATRDRRAFT_37652 FABD
atggggatgc tcctactgac tgttggctgt ctggatggg caacagcttt tggcgctgcc
ttccaatgga gcattgctac ggccagcgga gctgaaacct catctgcca agccagtgat
tctgactttg atgacttttc ttccagagag gcattcatgt tcccgggcca gggcgacaaa
tttggcgcca tgggtgggga gctggttaag gatgaccca aagcccaagc gctcttgat
caagccagcg agattctggg ctacgacttg ctgcaagtgt gggttgaagg acccaaggaa
aagctggatt caacgtagt ttccagcgca gctatttttg ttgacagcat gggcggggtg
gaaaagctcc gacaggaaca gggcgagcac gccatcaagc cgccacactg ggcagggga
ttgagtttgg gogaatactc gggcgtttgc ttgcoagac ccatctcctt tgcagacgtg
gtcaaaatca ccaaggctcg gggagaagcc atgcaagcgc cgctgatgac gctgatgatg
ggcatggttt cgttatttgg actagacag gaaaagtggt cgagctctg tggcatggcc
agtgaagaaa gtggcgaaag gatcagattt gctaaacttc ttgacagcgg aaactacgcc
gtgagtgcca gttccaagcc gtgtgagccc gtcaatgaac tgccaaacc cgagttcaaa
gcacgatga cgtgaaact tgcgtggccc gggcgctttc atacgattt catgaaagcc
gctgttgctt ccttgaaaaa ggtcctggcc gatgtgaga ttcaaaacc acgaatcccc
gtcattagca acgtggatgc caagccacat tccgatcccg aaacattcaa aaagctcttg
ggacccaag tcaggtcacc gctcttggg gaaaatacta tggatcttat gctttcagat
ggcttgaaaa aggccttga actagggcct ggtaaagtta cgccggggat tctcaagcga
ttgataaaa aagcgaatg tgaaaagctt gagggtgaa
```

Figure 1. Screenshots from DiatomCyc describing the pathway for the biosynthesis of fatty acids in *Phaeodactylum tricornutum*, and illustrating the different levels of information.

- (a) Cellular overview of the complete metabolism of *P. tricornutum*. Part of the pathway, shown in more detail on the information page (b) is encircled.
- (b) Details of genes and proteins associated with corresponding reactions. A general description of the pathway and links to the literature are provided (not shown in the figure). Comparative analysis can be carried out by the occurrence of the same pathway in different organisms (Species comparison).
- (c) Selection and visualization of a single reaction.
- (d) Gene information page that, in turn, includes links to external databases and literature references, and information relative to gene length and protein size. The genomic localization of the protein is graphically represented and the genomic coordinates are indicated.
- (e) Genome browser, protein and nucleotide sequences are accessible by clicking the gray boxes in (d).

as a detailed graphical representation of the reaction. When a gene is selected (Figure 1d), different types of information become available. A short description of the gene, its genomic coordinates, its length and the molecular weight of the protein it encodes are visible in the main window. In the genome browser, the gene sequence (Figure 1e) and its translation are directly accessible by clicking the corresponding boxes. Furthermore, links to the relevant literature and external resources [such as links from each gene to UniProt, the JGI genome sequence, the expressed sequence tag (EST) database, the KEGG database and NCBI] are provided.

The semi-automatic reconstruction of PGDBs is based on known metabolic pathways, often preventing the detection of variations within a given biochemical route. The biosynthesis of eicosapentaenoic acid (EPA) present in DiatomCyc is an example of the literature-based manual curation that had been applied to construct the diatom PGDB. EPA is an omega-3 polyunsaturated fatty acid (PUFA), and one of the valuable compounds synthesized by diatoms. In *P. tricornutum*, the accumulation of EPA is variable in time and can reach remarkable portions of the total fatty-acid fraction (Siron *et al.*, 1989; Zhukova and Aizdaicher, 1995; Domergue *et al.*, 2003; Patil *et al.*, 2007). Fatty-acid profiling showed that *P. tricornutum* accumulates large quantities of EPA, whereas intermediates of the pathway are present only in traces (Siron *et al.*, 1989; Zhukova and Aizdaicher, 1995; Domergue *et al.*, 2003), suggesting that the pathway has been optimized for the specific accumulation of EPA. The pathway template provided by MetaCyc had been adapted according to the outcome of other studies (Domergue *et al.*, 2002, 2003; Wen and Chen, 2003) to reconstruct the putative pathway that leads to the EPA synthesis (Figure 2). The reconstructed pathway is directly connected to the previously described fatty acid biosynthesis pathways, from which stearyl-ACP (C18:0) is desaturated to oleoyl-ACP by a Δ^9 desaturase, encoded by *PHATRDRRAFT_9316*. The set of genes involved in the subsequent desaturation and elongation steps consists of two Δ^6 elongases (encoded by *PHATRDRRAFT_22274* and *PHATRDRRAFT_20508*), two Δ^5 desaturases (*PHATRDRRAFT_22459* and *PHATR_46830*), a Δ^{12} desaturase (*PHATR_25769*), a Δ^6 desaturase (*PHATRDRRAFT_29488*) and an omega-3 desaturase (*PHATRDRRAFT_41570*) (Figure 2).

The 'Omics Viewer'

The web interface of DiatomCyc allows the user to import and picture several types of data, including microarray expression data, proteomics data, metabolomics data, reaction flux data or data from any other high-throughput 'omics' experiment, related to genes, protein reactions or compounds. With the 'Omics Viewer' tool implemented in DiatomCyc, data sets can be visualized on the cellular overview, and reactions and genes are marked with different color scales. As examples, the regulation of two isoprenoid precursor pathways will be illustrated: those of methyl-erythrol 4-phosphate (MEP) and the mevalonate (MVA). Both pathways were mapped in DiatomCyc: the MEPP (<http://www.diatomcyc.org/DIATOM/NEW-IMAGE?type=PATHWAY&object=NONMEVIPP-PWY>) and MVAP (<http://www.diatomcyc.org/DIATOM/NEW-IMAGE?type=PATHWAY&object=PWY-922>). Isoprenoids are important precursors for the biosynthesis of carotenoids and other pigments, such as β -carotene and the diatom-specific fucoxanthin (Bertrand, 2010), which have several applications in the food and pharmaceutical industries (Pangestuti and Kim, 2011). Previously, *P. tricornutum* had been found to use these two different pathways to synthesize isoprenoid precursors, just like plants (Cvejić and Rohmer, 2000). It is commonly believed that the MEPP operates in the chloroplasts, providing precursors for carotenoids, and that the MVAP contributes to the biosynthesis of sterols in the cytosol (Massé *et al.*, 2004). Carbon-labeling experiments (Cvejić and Rohmer, 2000; Massé *et al.*, 2004) demonstrated that the two pathways are triggered by different precursors. The main carbon source for the MEPP is CO₂, whereas the MVAP uses acetate-derived carbon.

Currently, the Diatom EST Database (Maheswari *et al.*, 2009) represents the only publicly available large-scale collection of gene expression data. Although this data set is still relatively small and does not allow the analysis of statistically significant changes in gene expression, it was used nonetheless to illustrate the utility of the DiatomCyc 'Omics Viewer' tools. As an example, the transcriptional reprogramming of the two isoprenoid precursor pathways under different culture conditions will be used. With the 'Omics Viewer' utility, the imported EST data are mapped on the cellular pathway overview, and the reactions and genes involved are marked with different color scales

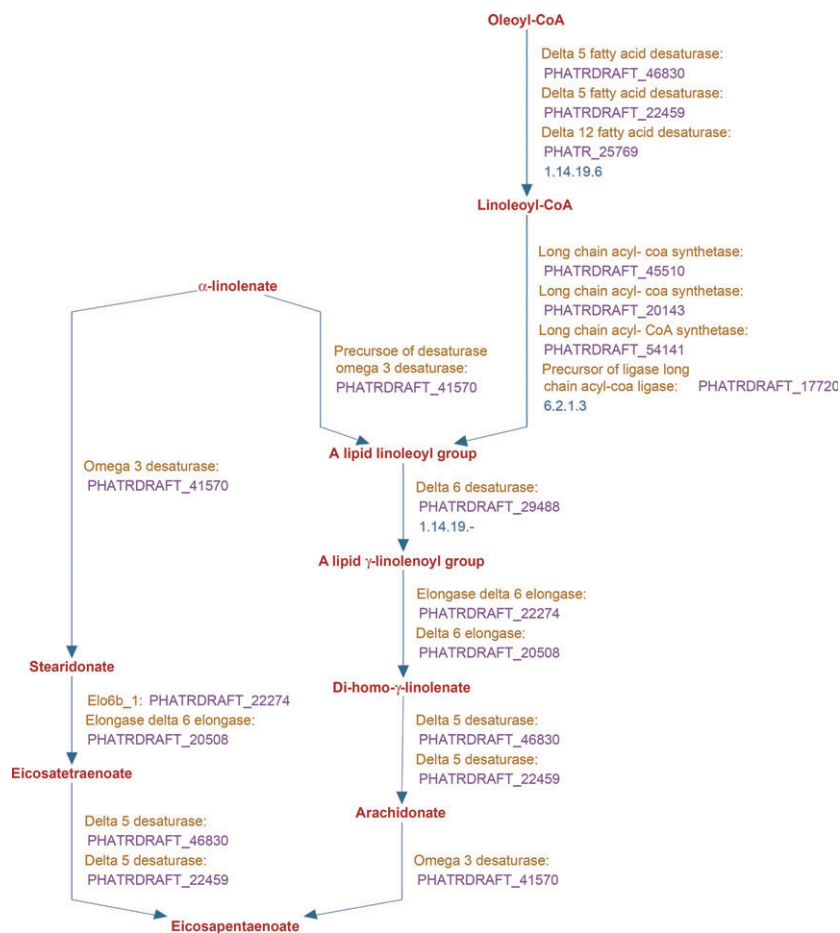


Figure 2. DiatomCyc screenshot of eicosapentaenoic acid biosynthesis in *Phaeodactylum tricornutum*.

(Figure 3). Specifically, as an example of the potential utility, we imported the EST data (Maheswari *et al.*, 2009) of the three culture conditions that affected the EST levels of MVAP and MEPP most markedly when compared with those in control pseudo steady-state cultures, namely the presence of a high concentration of CO₂, the addition of a toxic aldehyde and iron limitation. This analysis indicated that contigs corresponding to genes involved in the MEPP were induced in cells exposed to a high concentration of CO₂, whereas genes involved in the MVAP pathway were not (Figure 3a), according to their main carbon sources (Cvejić and Rohmer, 2000). In contrast, in the presence of the toxic aldehyde 2E,4E-decadienal (DD), the MVAP genes were induced, but not those of the MEPP (Figure 3b). This observation is plausible, because an *S. cerevisiae* mutant lacking the *ERG6* gene [encoding the $\delta(24)$ sterol C-methyltransferase involved in ergosterol biosynthesis] has been demonstrated to be significantly more sensitive to DD, which points to a direct link between the stress induced by the aldehyde and the synthesis of sterols (Adolph *et al.*, 2004). Like many other toxic oxylipins, DD increases membrane permeability and programmed cell death in various organisms (Adolph *et al.*, 2004; Ribalet

et al., 2007). Sterols, fed by the MVAP, might represent an important cellular defense barrier against such toxic compounds. Finally, the lack of iron in culture media has been reported to trigger a rearrangement of the photosynthetic components, and to stimulate the expression of genes involved in the biosynthesis of chlorophyll, carotenoids and their precursors (Allen *et al.*, 2008). Accordingly, the EST data indicate that the MEPP genes, as well as those of the MVAP, are induced under iron restriction (Figure 3c). Although these data need to be interpreted with care because of the limited number of ESTs in the database, the above examples illustrate the usefulness of DiatomCyc as an interactive laboratory tool in the study of diatom metabolism, and its reprogramming under different culture or stress conditions.

Completion of gaps in the current carbohydrate model

A genome-based overview of diatom metabolism provided the first insights into the acquisition of dissolved inorganic carbon, photorespiration and a detailed synthesis of carbohydrate metabolism (Kroth *et al.*, 2008). Our results confirmed the proposed carbohydrate model and allowed us to fill in several of the remaining gaps. For example, acetyl-CoA

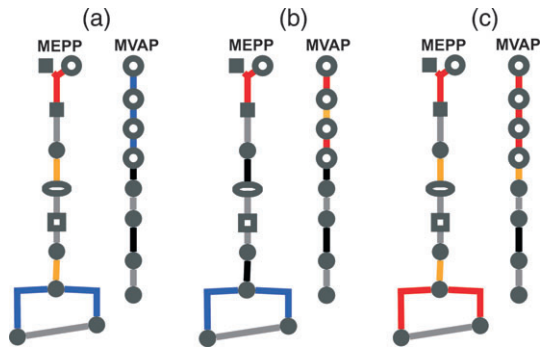


Figure 3. 'Omics Viewer' in DiatomCyc.

Examples of visualization of expressed sequence tag (EST) data on the reconstructed methyl-erythrol-4-phosphate pathway (MEPP) and mevalonate pathway (MVAP). Each icon represents a single metabolite: squares, carbohydrates; ellipses, pyrimidines; circles, other compounds; and filled shapes indicate phosphorylated compounds. Different inductions of MEPP and MVAP under different culture conditions are shown, i.e. pseudo steady-state growth of *Phaeodactylum tricornutum* in the presence of: a high concentration of CO₂ (a); 2E,4E-decadienal (b); and under iron limitation (c) (Maheswari *et al.*, 2009). Before importing EST data in DiatomCyc, we normalized the data by dividing the absolute number of ESTs for every gene in a given condition by the total number of ESTs for that condition. Subsequently, these normalized values were divided by the value relative to the condition used as a comparison (the standard condition). Colors indicate repression of a gene in a given condition (blue), no difference in expression relative to the standard condition (black) or induction of a gene in a given condition (red). Genes for which there are ESTs in the 'stress' condition but not in the standard condition are indicated in orange, and those for which no unambiguous EST data are available in any condition are indicated in light gray.

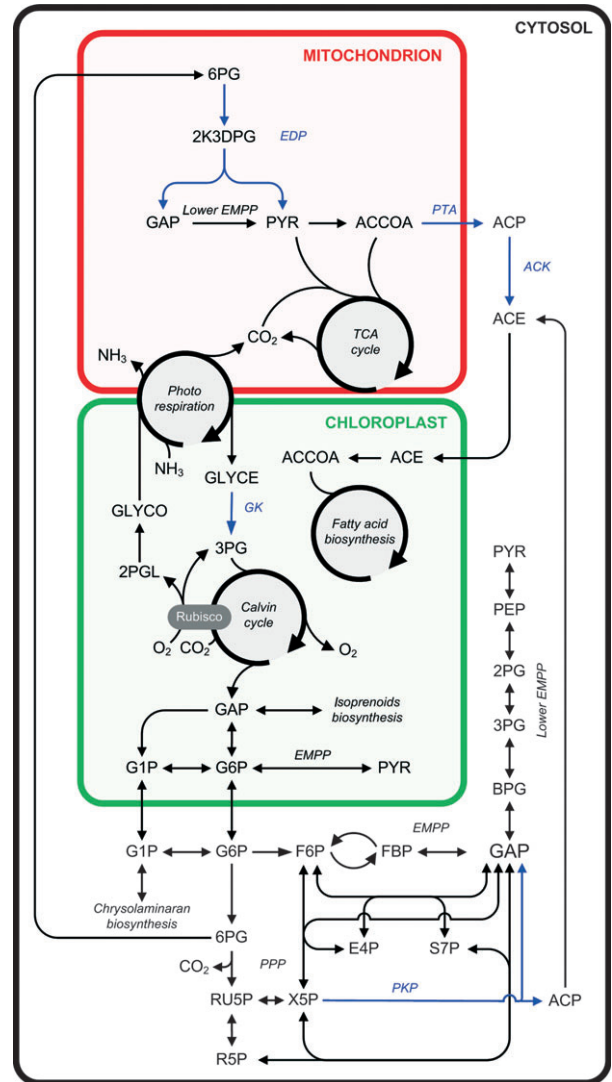


Figure 4. Simplified overview of carbohydrate metabolism and photorespiration in *Phaeodactylum tricornutum*. The Entner–Doudoroff pathway (EDP), phosphoketolase pathway (PKP) and completed gaps are indicated in blue. 2K3DPG, 2-keto-3-deoxyphosphogluconate; 2PGL, 2-phosphoglycolate; 3PG, 3-phosphoglycerate; 6PG, 6-phospho gluconate; ACCoA, acetyl-CoA; ACE, acetate; ACK, acetate kinase; ACP, acetylphosphate; BPG, 1,3-bisphosphoglycerate; CO₂, carbon dioxide; EMPP, Embden–Meyerhof–Parnas pathway; E4P, erythrose-4-phosphate; FBP, fructose-1,6-bisphosphate; F6P, fructose-6-phosphate; GAP, glyceraldehyde-3-phosphate; GK, glyceralate kinase; GLYCE, glyceralate; GLYCO, glycolate; G1P, glucose-1-phosphate; G6P, glucose-6-phosphate; NH₃, ammonium; O₂, oxygen; PEP, phosphoenolpyruvate; 2PG, 2-phosphoglycerate; PPP, pentose phosphate pathway; PTA, phosphate acetyltransferase; PYR, pyruvate; R5P, ribulose-5-phosphate; RU5P, ribulose-5-phosphate; S7P, sedoheptulose-7-phosphate; X5P, xylulose-5-phosphate.

In silico prediction of a phosphoketolase pathway

Besides the Embden–Meyerhof–Parnas pathway (EMPP; Figure 4), which is the most widely distributed glycolytic pathway in nature, we predicted the presence of a phosphoketolase pathway (PKP; Figure 4). The PKP is a catabolic

variant of the pentose phosphate pathway (PPP), and uses phosphoketolase (XFP) as the key enzyme. To date, two types of XFP activities have been described: fructose-6-phosphate phosphoketolase (EC 4.1.2.22) and xylulose-5-phosphate phosphoketolase (EC 4.1.2.9) (Sánchez *et al.*, 2010).

Organisms with both XFP activities are able to exploit a unique metabolic sugar pathway, termed the bifid shunt, that is generally considered a distinctive taxonomic mark for *Bifidobacteria*. Most XFPs have dual substrate specificity (Sánchez *et al.*, 2010). In *P. tricornutum*, a putative xylulose-5-phosphate/fructose-6-phosphate phosphoketolase (PtXPK), encoded by the gene *PHATRDRAFT_36257*, was identified. PtXPK might catalyze the cleavage of xylulose-5-phosphate to acetyl-phosphate and glyceraldehyde-3-phosphate (GAP) that can be further converted in pyruvate by entering the lower part of the glycolysis. As mentioned above, acetyl phosphate can be converted to acetate by PtACK. Similar to the enzymes involved in the oxidative PPP (Kroth *et al.*, 2008), PtXPK and PtACK are putatively cytosolic enzymes. Among the reference organisms used in the orthology prediction, PtXPK shared orthologs with Cyanobacteria (*Acaryochloris marina*, *Synechococcus* sp. and *Trichodesmium erythraeum*) and Lactobacillaceae (*Lactobacillus plantarum* and *Lactococcus lactis*) (Table S1). Interestingly, no orthologs were found in *Thalassiosira pseudonana*. In the genome of *P. tricornutum*, PtXFP is localized on chromosome 9, adjacent to the PtACK-encoding locus that catalyzes the next reaction in the pathway (Figure 5). The spatial association between phosphoketolase and acetate kinase open reading frames has been reported to occur relatively frequently in Proteobacteria and Cyanobacteria (Sánchez *et al.*, 2010), and might hint at a bacterial origin of the PKP in *P. tricornutum*, as indicated previously by computational analysis (Bowler *et al.*, 2008). To date, eukaryotic phosphoketolases have been indicated only in a few fungal species (Sánchez *et al.*, 2010).

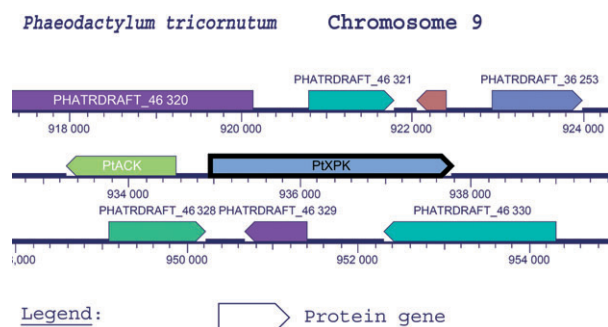


Figure 5. Genome browser of DiatomCyc.

A particular region of chromosome 9 of *Phaeodactylum tricornutum*, where the genes encoding acetate kinase (PtACK) and phosphoketolase (PtXPK), key enzymes of the PKP, are spatially associated, is shown. Numbers indicate genomic positions.

In silico prediction and experimental confirmation of the Entner–Doudoroff pathway

A striking finding that emerged from the orthology-based approach was the identification of the Entner–Doudoroff pathway (EDP) in *P. tricornutum*. The EDP (Figure 4) is considered to be the ancient glycolytic pathway (Romano and Conway, 1996), and is predominantly restricted to prokaryotic lineages. The involved genes encode 6-phosphogluconate dehydratase (EDD, EC 4.2.1.12) and 2-keto-3-deoxyphosphogluconate aldolase (EDA, EC 4.2.1.14). The EDP degrades glucose after its conversion to 6-phosphogluconate (6PG). EDD catalyzes the first reaction in which the dehydration of 6PG produces one molecule of 2-keto-3-deoxyphosphogluconate (KDPG). The second and final step of the pathway, catalyzed by EDA, cleaves KDPG to pyruvate and GAP. The lower glycolysis can then further degrade GAP to pyruvate. Both EDP enzymes in *P. tricornutum* have a predicted mitochondrial signal peptide and might form a complete mitochondrial EDP when combined with the lower glycolysis that had been identified previously (Kroth *et al.*, 2008). Whereas the EDD enzyme is widely conserved in nature, as a multifunctional protein involved in several cellular metabolic processes (Peekhaus and Conway, 1998), EDA is considered the key and distinctive enzyme of the EDP. For instance, the *PtEDD* gene (*PHATRDRAFT_20547*) has orthologs in 17 out of 21 reference organisms (Table S1). In contrast, the hits for *PtEDA* (*PHATRDRAFT_34120*) are primarily restricted to prokaryotes, although orthologs have been identified in some eukaryotes as well, such as *Cyanidioschyzon merolae*, *Ostreococcus tauri* and *Ostreococcus lucimarinus*. Notably, the two EDP genes of *P. tricornutum* are not part of the list of genes previously predicted to be of bacterial origin (Bowler *et al.*, 2008). The two genes of the EDP have also been identified in *T. pseudonana*, suggesting that the EDP might be conserved in diatoms, in contrast to the PKP.

To confirm the function of the predicted EDP genes, we performed a genetic complementation experiment in the triple knock-out mutant strain $\Delta edd\Delta eda\Delta gnd$ of *E. coli*, and assayed growth on minimal medium containing gluconate as the sole carbon source (Figure 6). In this mutant, the native ED genes *eda* (*b4477*) and *edd* (*b3771*) were inactivated by knock out. Additionally, to avoid gluconate flux being channeled through the PPP, the gene coding for 6-phosphogluconate dehydrogenase (*GND*, *b2029*) was also deleted. The resulting triple knock-out strain was unable to grow on the gluconate-containing minimal medium (Figure 6). Transformation of the *E. coli* mutant with a polycistronic plasmid carrying the *PtEDA* and *PtEDD* genes from *P. tricornutum* restored growth on the minimal gluconate medium (Figure 6), thus confirming the functionality of the *PtEDA* and *PtEDD* proteins. The restored EDP allowed the metabolization of gluconate in the bacterial cell, after its

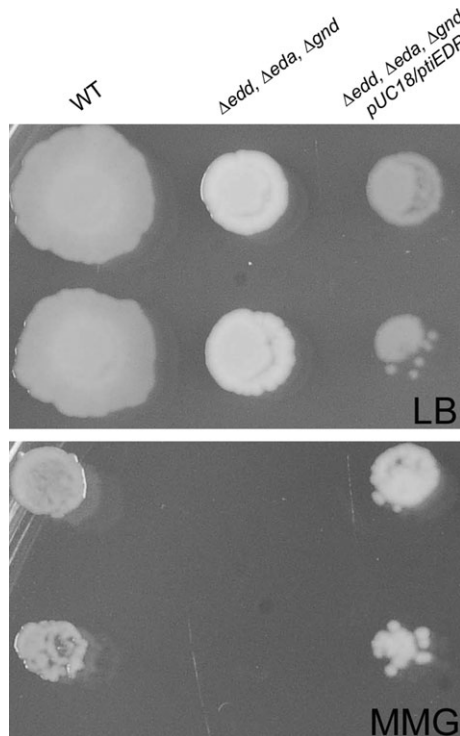


Figure 6. Genetic complementation of the Entner–Doudoroff pathway in *Escherichia coli*. Transformed wild-type and mutant *E. coli* strains ($\Delta\text{edd}\Delta\text{eda}\Delta\text{gnd}$) were grown for 48 h at 37°C on LB medium or minimal medium supplemented with gluconate as the sole carbon source (MMG).

phosphorylation to 6-phosphogluconate by endogenous gluconate kinase.

Two sets of experiments confirmed the functionality of the EDP in diatom cells. First, the occurrence of the EDP was supported by the detection of *PtEDA* and *PtEDD* expression in cultured *P. tricornutum* cells. Moreover, differential transcript accumulation in cells grown under different light regimes was observed (Figure S1). In particular, the transcript levels corresponding to *PtEDA*, the key enzyme of the EDP, showed a pronounced light-modulated pattern, with a gradual decrease under continuous light and a sharp increase following the switch to a dark phase. Second, the activity of the EDP enzymes was tested with an enzymatic assay with soluble protein extracts from *P. tricornutum* cells grown under different light regimes and the different *E. coli* strains that we generated. 6PG was used as substrate and its conversion to pyruvate was measured with a fluorometric assay (Table S2). Pyruvate was detected in cells of *P. tricornutum*, wild-type *E. coli* and in mutant *E. coli* $\Delta\text{edd}\Delta\text{eda}\Delta\text{gnd}$ complemented with pUC18-*PtiEDP*, but not in cells of the uncomplemented $\Delta\text{edd}\Delta\text{eda}\Delta\text{gnd}$ mutant that lacks an EDP. These observations provide direct evidence for the activity of the *PtEDP* enzymes in both endogenous and heterologous cells. Interestingly, the measured pyruvate concentration was higher in samples of *P. tricornutum* grown in a

prolonged dark phase than in the samples obtained from diatoms grown in continuous light, which correlates with the *PtEDA* transcript levels (Figure S1). These preliminary results support the hypothesis that EDP regulation might be linked to changes in the energy status of the cell.

The presence of multiple glycolytic pathways in marine eukaryotes, such as diatoms, remains unclear. We hypothesize that the coordination of multiple central carbon pathways in *P. tricornutum* might be the consequence of cellular ‘economical strategies’. Although the EDP produces less energy per molecule of glucose, it requires fewer resources to synthesize the enzymes than the EMPP (Carlson, 2007), and possibly enables a fast shunt to match the required production and demands for ATP/NADPH (Kramer and Evans, 2011). The shift to energetically inefficient metabolism originates from the trade-off between low investment cost in enzyme synthesis and a high operation cost for alternative catabolic pathways (Molenaar *et al.*, 2009). In *E. coli*, a low operation cost/high investment cost strategy (i.e. maximizing energy yield), has been recognized as a competitive strategy during nutrient-limited chemostat growth (Schuetz *et al.*, 2007). The combined use of different pathways for glucose use might reflect a great metabolic flexibility: ‘expensive’ pathways in terms of investment costs, such as the EMPP, are more energetically efficient and finely tuned by transcriptional regulation (Wessely *et al.*, 2011). In contrast, ‘cheaper’ pathways, such as the EDP, might be less tightly regulated and predominant in an organism subjected to frequent environmental changes, resulting in faster metabolic responses that provide an immediate selective advantage (Wessely *et al.*, 2011).

CONCLUSIONS

The PGDBs and genome-scale metabolic networks represent relevant resources for the study of cellular metabolism. As shown by well-curated databases, such as EcoCyc (Keseler *et al.*, 2011), the level of information can have an impressive resolution and an important impact on the scientific community by becoming a common laboratory tool, in particular for comparative analysis and visualization of high-throughput experiments. DiatomCyc is a first step toward the generation of a comprehensive overview of diatom metabolism, and provides a user-friendly, interactive platform for current and future diatom research. The comparison between DiatomCyc and other PGDBs (Table 1) reflects the current status of knowledge of *P. tricornutum* metabolism, which still contains gaps in some metabolic pathways. Such gaps can sometimes be ascribed to missing or incomplete EC numbers in MetaCyc. Furthermore, isolated reactions without a gene association might involve diatom-specific genes. Proteins with obscure functions (POFs) that cannot yet be linked *in silico* to any cellular process, generally represent a significant portion in every eukaryotic genome, ranging from 18 to 38% (Gollery *et al.*, 2006). In *P. tricornu-*

tum, 44% of the genes code for such POFs and lack detectable functional domains (Maheswari *et al.*, 2010). Functional annotation and assembly of the *P. tricornutum* genome is an ongoing project; hence, continuous curation will be necessary to address new annotations and to complete the metabolic pathways. Therefore, DiatomCyc will be updated accordingly on a regular basis.

The orthology-based approach that was used to construct DiatomCyc led to the characterization of the central metabolism of the model diatom *P. tricornutum*, with particular interest in pathways with a biotechnological relevance, such as carbohydrate, isoprenoid and fatty acid biosynthesis. Moreover, it allows us to propose hypothetical routes for unknown pathways, such as of PUFA biosynthesis. The utility and power of the methods used are illustrated by the identification of uncommon eukaryotic pathways, such as the EDP, of which the functionality has been confirmed experimentally. The implication that *P. tricornutum* has multiple pathways for glucose metabolism emphasizes its marked metabolic versatility. DiatomCyc will become a powerful resource, both for fundamental research and the development of metabolic engineering strategies aiming at the industrial exploitation of diatoms. Diatoms produce a variety of compounds, such as polyunsaturated fatty acids, for the production of health foods and marine animal feed, and saturated fatty acids and monoenic fatty acids for the production of biodiesel (Bozarth *et al.*, 2009). Although bioenergy processes from diatoms (and other microalgae) are not yet economically competitive, their theoretical energy potential (9–154 kW ha⁻¹, depending on oil content) is significantly higher than that of *Saccharum* sp. (sugar cane) and *Elaeis guineensis* (palm) (Chisti, 2007, 2008; Demirbas, 2009), making diatoms a promising renewable and carbon-neutral alternative to petroleum fuels for the future.

EXPERIMENTAL PROCEDURES

Orthology prediction and database construction

The annotated genomes of *P. tricornutum* (Bowler *et al.*, 2008), 10 organisms with a published genome-scale metabolic model and curated annotation (*Arabidopsis thaliana*, *Homo sapiens*, *Saccharomyces cerevisiae*, *Escherichia coli*, *Helicobacter pilory*, *Neisseria meningitidis*, *Methanococcus jannaschii*, *Lactococcus lactis*, *Lactobacillus plantarum* and *Bacillus subtilis*) and 12 additional species (*Acaryochloris marina*, *Ectocarpus siliculosus*, *Prochlorococcus marinus*, *Ostreococcus tauri*, *Ostreococcus lucimarinus*, *Cyanidioschyzon merolae*, *Synechococcus* sp. JA-3-3Ab, *Synechococcus* sp. JA-2-3B'a, *Chlamydomonas reinhardtii*, *Trichodesmium erythraeum*, *Thalassiosira pseudonana* and *Entamoeba histolytica*) were downloaded from the KEGG database (<http://www.genome.jp/kegg>, February 2010; Kanehisa and Goto, 2000). Additionally the genome of *Aureococcus anophagefferens* was downloaded from the JGI website (<http://genome.jgi.doe.gov/Auran1/Auran1.home.html>; Gobler *et al.*, 2011).

Orthology prediction was carried out with INPARANOID 3.0 with a default cut-off value of 50 bits (Remm *et al.*, 2001; O'Brien *et al.*, 2005). Gene-to-function and function-to-reaction associations were

transferred semi-automatically by means of the KEGG (Kanehisa and Goto, 2000) and MetaCyc (<http://metacyc.org>; Caspi *et al.*, 2010) databases as input. In addition, FASTA headers of the reference genomes were screened and mined to improve the functional annotation of the *P. tricornutum* genome. Functions of the translated gene sequences with the highest score among the reference organisms were transferred to the corresponding *P. tricornutum* orthologs, yielding the primary genome-scale metabolic network. The genome-scale network was converted into a PathoLogic-specific data set according to the PATHWAY TOOLS documentation (Karp *et al.*, 2010), imported in PATHWAY TOOLS 15.0 (Karp *et al.*, 2002), and subsequently refined and manually curated with literature references and bioinformatic tools, such as InterProScan (Hunter *et al.*, 2009), TargetP (Emanuelsson *et al.*, 2007) and TransportDB (Ren *et al.*, 2007). Metabolic pathways absent in the MetaCyc framework were added manually.

Expression of *P. tricornutum* genes in *E. coli*

Wild-type *E. coli* strain K-12 MG1655 and the $\Delta edd\Delta eda$ mutant were kindly provided by Prof. Daniël Charlier (Vrije Universiteit Brussel, Belgium). A third gene deletion (*gnd*, 6-phosphogluconate dehydrogenase) was introduced into the latter strain to block the flux through the PPP by replacing the target gene by a kanamycin-resistant gene (Datsenko and Wanner, 2000). Luria broth (LB) medium, enriched with 50 μ M kanamycin (Duchefa Biochemie, <http://www.duchefa.com>) was used to select the triple knock-out *E. coli* mutants. The final mutant ($\Delta edd\Delta eda\Delta gnd$) was checked with PCR.

Axenic cultures of the *P. tricornutum* CCAP 1055/1 were grown in f/2 medium without silica (Guillard and Ryther, 1962) at 21°C in a 12-h light/12-h dark regime (average 75 μ mol photons m⁻² s⁻¹) and shaken at 100 rpm. Total RNA was extracted with Tri-Reagent (Molecular Research Center, <http://www.mrcgene.com>) according to the manufacturer's protocol. DNase treatment was performed with RQ1 RNase-Free DNase (Promega, <http://www.promega.com>) and cDNA was synthesized with the iScript cDNA synthesis kit (Bio-Rad, <http://www.bio-rad.com>). The predicted open reading frames of *EDD* and *EDA* (*PHATRDRRAFT_34120* and *PHATRDRRAFT_20547*) were amplified with PrimeSTAR[®] HS DNA Polymerase (Takara Bio, <http://www.takara-bio.com>). For primer design, the JGI gene model of *PHATRDRRAFT_20547* was manually adjusted. The full-length open reading frames of *EDD* and *EDA* were sequentially cloned into the pUC18 vector as a polycistron (Ye *et al.*, 2010) to yield the pUC18-PtiEDP construct.

The *E. coli* mutants were transformed with the pUC18-PtiEDP vector by heat shock, selected on LB medium enriched with 100 μ M ampicillin and verified by PCR. The minimal medium for the complementation experiments contained 15 g L⁻¹ agar, 6.75 g L⁻¹ NH₄Cl, 1.25 g L⁻¹ (NH₄)₂SO₄, 1.15 g L⁻¹ KH₂PO₄, 0.5 g L⁻¹ NaCl, 0.5 g L⁻¹ MgSO₄·7H₂O, 1 ml L⁻¹ vitamin solution and 100 μ l L⁻¹ molybdc acid solution. The vitamin solution consisted of 4.89 g L⁻¹ FeCl₃·6H₂O, 5 g L⁻¹ CaCl₂·2H₂O, 1.3 g L⁻¹ MnCl₂·2H₂O, 0.5 g L⁻¹ CoCl₂·6H₂O, 0.94 g L⁻¹ ZnCl₂, 0.0311 g L⁻¹ H₃BO₄, 0.4 g L⁻¹ Na-EDTA·2H₂O and 1.01 g L⁻¹ thiamine-HCl. The molybdate solution contained 0.76 g L⁻¹ molybdc acid. In the complementation assays, D-gluconic acid sodium salt (16.5 g L⁻¹) was used as the sole carbon source. Isopropyl β -D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 500 μ M to induce transgene expression.

Analysis of PtiEDP gene expression

Complementary DNA (cDNA) obtained from RNA from axenic cultures of *P. tricornutum* CCAP 1055/1 grown in f/2 medium without silica (Guillard and Ryther, 1962) at 21°C either in continuous light (average 75 μ mol photons m⁻² s⁻¹) for 22 h or switched to the dark

after 12 h in continuous light, was prepared and analyzed by quantitative real-time PCR, as described previously (Huysman *et al.*, 2010). *Histone H4 (H4)* and *Tubulin-β chain (TubB)* were used as the internal control genes (Siaut *et al.*, 2007) for the normalization of the relative expression ratio of *PtEDA* and *PtEDD* transcripts. Primers for the amplification of *PtEDA* (forward, 5'-CGCTACTTCGGATGATTGC-3'; reverse, 5'-GGAGTCGTGAGGGTGAAC-3') and *PtEDD* (forward, 5'-AGAAGCGAAGAACAAGATGG-3'; reverse, 5'-GGAGCGGCAAT CACAATC-3') were designed with Beacon Designer (Premier Bio-soft, <http://www.premierbiosoft.com>).

Analysis of PtEDP enzymatic activity

Axenic cultures of *P. tricornutum* CCAP 1055/1 were grown in ESAW (Enriched Artificial Seawater) medium (Harrison *et al.*, 1980) at 21°C in a 16-h light/8-h dark regime (average 75 μmol photons m⁻² s⁻¹) and harvested either after 10 h of cultivation in the light or after 10 h of a prolonged dark phase. *E. coli* K-12 MG1655, *E. coli* K-12 MG1655 *ΔeddΔedaΔgnd* and *E. coli* K-12 MG1655 *ΔeddΔedaΔgnd* pUC18-PtEDP were grown on LB medium, LB medium enriched with 50 μM kanamycin and 100 μM ampicillin, respectively, for 12 h at 37°C on an orbital shaker. Cells were harvested by centrifugation, washed in PBS and resuspended in 0.5 ml of lysis buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100; Complete Protease Inhibitor) according to the manufacturer's protocol (Roche, <http://www.roche.com>). Resuspended samples were sonicated on ice for 5 min with 10-s pulses with a Heat Systems Ultrasonics sonicator (Heat Systems Incorporated, <http://www.misonix.com>). Cell debris and non-solubilized material were removed by centrifugation (30 min at 14 000 g), and the supernatant was subsequently centrifuged (2 h at 40 000 g) in order to obtain the soluble enzyme fraction. The protein concentration was determined by the Bradford assay (Bio-Rad, <http://www.bio-rad.com>) according to the manufacturer's protocol, and using bovine serum albumin (BSA) as a standard.

The soluble protein concentration of *P. tricornutum* and *E. coli* lysates was equalized between samples of the same organism. The EDP *in vitro* reactions were carried out in 40 μl of TRIS-HCl, pH 7.5, 20 μl sodium arsenite 0.2 M, by adding 40 μl of soluble enzyme fraction and 40 μl of 0.2 M 6PG as the substrate. The reaction mix was incubated for 90 min at 37 and 21°C for *E. coli* and *P. tricornutum* samples, respectively. Pyruvate concentrations were determined fluorometrically with a Pyruvate Assay Kit (BioVision, <http://www.biovision.com>) according to the manufacturer's protocol.

ACKNOWLEDGEMENTS

We would like to thank Dr Marie Huysman for providing cDNA samples and helpful discussions. This work was supported by the Agency for Innovation by Science and Technology in Flanders ('Strategisch Basisonderzoek' grant no. 80031 and by a predoctoral fellowship to MM).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Expression of PtEDP genes in different light regimes.

Table S1. Orthology prediction results.

Table S2. Enzymatic activity of the PtEDP proteins.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

REFERENCES

- Adolph, S., Bach, S., Blondel, M., Cuff, A., Moreau, M., Pohnert, G., Poulet, S.A., Wichard, T. and Zuccaro, A. (2004) Cytotoxicity of diatom-derived oxylipins in organisms belonging to different phyla. *J. Exp. Biol.* **207**, 2935–2946.
- Allen, A.E., LaRoche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P.J., Finazzi, G., Fernie, A.R. and Bowler, C. (2008) Whole-cell response of the pennate diatom *Phaeodactylum tricornutum* to iron starvation. *Proc. Natl. Acad. Sci. USA*, **105**, 10438–10443.
- Allen, A.E., Dupont, C.L., Obornik, M. *et al.* (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature*, **473**, 203–207.
- Baart, G.J.E., Zomer, B., de Haan, A., van der Pol, L.A., Beuvery, E.C., Tramper, J. and Martens, D.E. (2007) Modeling *Neisseria meningitidis* metabolism: from genome to metabolic fluxes. *Genome Biol.* **8**, R136.
- Baart, G.J.E., Willemsen, M., Khatami, E., de Haan, A., Zomer, B., Beuvery, E.C., Tramper, J. and Martens, D.E. (2008) Modeling *Neisseria meningitidis* B metabolism at different specific growth rates. *Biotechnol. Bioeng.* **101**, 1022–1035.
- Ball, C.A., Jin, H., Sherlock, G. *et al.* (2001) *Saccharomyces* Genome Database provides tools to survey gene expression and functional analysis data. *Nucleic Acids Res.* **29**, 80–81.
- Bertrand, M. (2010) Carotenoid biosynthesis in diatoms. *Photosynth. Res.* **106**, 89–102.
- Bowler, C., Allen, A.E., Badger, J.H. *et al.* (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature*, **456**, 239–244.
- Bozarth, A., Maier, U.-G. and Zauner, S. (2009) Diatoms in biotechnology: modern tools and applications. *Appl. Microbiol. Biotechnol.* **82**, 195–201.
- Carlson, R.P. (2007) Metabolic systems cost-benefit analysis for interpreting network structure and regulation. *Bioinformatics*, **23**, 1258–1264.
- Caspi, R., Altman, T., Dale, J.M. *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **38**, D473–D479.
- Chang, R.L., Ghamsari, L., Manichaikul, A. *et al.* (2011) Metabolic network reconstruction of *Chlamydomonas* offers insight into light-driven algal metabolism. *Mol. Syst. Biol.* **7**, 518.
- Chisti, Y. (2007) Biodiesel from microalgae. *Biotechnol. Adv.* **25**, 294–306.
- Chisti, Y. (2008) Biodiesel from microalgae beats bioethanol. *Trends Biotechnol.* **26**, 126–131.
- 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*, **53**, 21–28.
- Datsenko, K.A. and Wanner, B.L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. USA*, **97**, 6640–6645.
- Demirbas, A. (2009) Progress and recent trends in biodiesel fuels. *Energy Convers. Manage.* **50**, 14–34.
- Domergue, F., Lerchl, J., Zähringer, U. and Heinz, E. (2002) Cloning and functional characterization of *Phaeodactylum tricornutum* front-end desaturases involved in eicosapentaenoic acid biosynthesis. *Eur. J. Biochem.* **269**, 4105–4113.
- Domergue, F., Spiekermann, P., Lerchl, J., Beckmann, C., Kilian, O., Kroth, P.G., Boland, W., Zähringer, U. and Heinz, E. (2003) New insight into *Phaeodactylum tricornutum* fatty acid metabolism. Cloning and functional characterization of plastidial and microsomal Δ¹²-Fatty acid desaturases. *Plant Physiol.* **131**, 1648–1660.
- Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2**, 953–971.
- Falkowski, P.G., Barber, R.T. and Smetacek, V. (1998) Biogeochemical controls and feedbacks on ocean primary production. *Science*, **281**, 200–206.
- Falkowski, P.G., Katz, M.E., Milligan, A.J., Fennel, K., Cramer, B.S., Aubry, M.P., Berner, R.A., Novacek, M.J. and Zapol, W.M. (2005) The rise of oxygen over the past 205 million years and the evolution of large placental mammals. *Science*, **309**, 2202–2204.
- Field, C.B., Behrenfeld, M.J., Randerson, J.T. and Falkowski, P. (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, **281**, 237–240.

- Fong, S.S., Burgard, A.P., Herring, C.D., Knight, E.M., Blattner, F.R., Maranas, C.D. and Palsson, B.O. (2005) In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol. Bioeng.* **91**, 643–648.
- Giordano, M., Beardall, J. and Raven, J.A. (2005) CO₂ concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. *Annu. Rev. Plant Biol.* **56**, 99–131.
- Gobler, C.J., Berry, D.L., Dyhrman, S.T. et al. (2011) Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics. *PNAS*, **108**, 4352–4357.
- Gollery, M., Harper, J., Cushman, J., Mittler, T., Girke, T., Zhu, J.-K., Bailey-Serres, J. and Mittler, R. (2006) What makes species unique? The contribution of proteins with obscure features. *Genome Biol.* **7**, R57.1–R57.8.
- Guillard, R.R. and Ryther, J.H. (1962) Studies of marine planktonic diatoms. I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (Cleve) Gran. *Can. J. Microbiol.* **8**, 229–239.
- Harrison, P.J., Waters, R.E. and Taylor, F.J.R. (1980) A broad spectrum artificial seawater medium for coastal and open ocean phytoplankton. *J. Phycol.* **16**, 28–35.
- Hua, Q., Joyce, A.R., Fong, S.S. and Palsson, B.O. (2006) Metabolic analysis of adaptive evolution for in silico-designed lactate-producing strains. *Biotechnol. Bioeng.* **85**, 992–1002.
- Hunter, S., Apweiler, R., Attwood, T.K. et al. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215.
- Huysman, M.J., Martens, C., Vandepoele, K. et al. (2010) Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome Biol.* **11**, R17.
- Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30.
- Karp, P.D., Paley, S. and Romero, P. (2002) The Pathway Tools software. *Bioinformatics*, **18**(Suppl. 1), S225–S232.
- Karp, P.D., Paley, S.M., Krummenacker, M. et al. (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.* **11**, 40–79.
- Keseler, I.M., Collado-Vides, J., Santos-Zavaleta, A. et al. (2011) EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.* **39**, D583–D590.
- Kramer, D.M. and Evans, J.R. (2011) The importance of energy balance in improving photosynthetic productivity. *Plant Physiol.* **155**, 70–78.
- Kroth, P.G., Chiovitti, A., Gruber, A. et al. (2008) A model for carbohydrate metabolism in the diatom *Phaeodactylum tricornutum* deduced from comparative whole genome analysis. *PLoS ONE*, **3**, e1426.
- Maheswari, U., Mock, T., Armbrust, E.V. and Bowler, C. (2009) Update of the Diatom EST Database: a new tool for digital transcriptomics. *Nucleic Acids Res.* **37**, D1001–D1005.
- Maheswari, U., Jabbari, K., Petit, J.-L. et al. (2010) Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome Biol.* **11**, R85.
- Massé, G., Belt, S.T., Rowland, S.J. and Rohmer, M. (2004) Isoprenoid biosynthesis in the diatoms *Rhizosolenia setigera* (Brightwell) and *Haslea ostrearia* (Simonsen). *Proc. Natl. Acad. Sci. USA*, **101**, 4413–4418.
- May, P., Christian, J.-O., Kempa, S. and Walther, D. (2009) ChlamyCyc: an integrative systems biology database and web-portal for *Chlamydomonas reinhardtii*. *BMC Genomics*, **10**, 209.
- Molenaar, D., van Berlo, R., de Ridder, D. and Teusink, B. (2009) Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol. Syst. Biol.* **5**, 323.
- Notebaart, R.A., van Enkevort, F.H.J., Francke, C., Siezen, R.J. and Teusink, B. (2006) Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics*, **7**, 296.
- O'Brien, K.P., Remm, M. and Sonnhammer, E.L.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**, D476–D480.
- Pangestuti, R. and Kim, S.-K. (2011) Biological activities and health benefit effects of natural pigments derived from marine algae. *J. Funct. Foods*, **3**, 255–266.
- Patil, V., Källqvist, T., Olsen, E., Vogt, G. and Gislerød, H.R. (2007) Fatty acid composition of 12 microalgae for possible use in aquaculture feed. *Aquacult. Int.* **15**, 1–9.
- Peekhaus, N. and Conway, T. (1998) What's for dinner?: Entner-Doudoroff metabolism in *Escherichia coli*. *J. Bacteriol.* **180**, 3495–3502.
- Provost, A. and Bastin, G. (2004) Dynamic metabolic modelling under the balanced growth condition. *J. Process Control*, **14**, 717–728.
- Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A., Swainston, N., Baart, G. and Schwartz, J.-M. (2010) Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Syst. Biol.* **4**, 114.
- Remm, M., Storm, C.E.V. and Sonnhammer, E.L.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **14**, 1041–1052.
- Ren, Q., Chen, K. and Paulsen, I.T. (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res.* **35**, D274–D279.
- Ribaut, F., Wichard, T., Pohnert, G., Ianora, A., Miralto, A. and Casotti, R. (2007) Age and nutrient limitation enhance polyunsaturated aldehyde production in marine diatoms. *Phytochemistry*, **68**, 2059–2067.
- Romano, A.H. and Conway, T. (1996) Evolution of carbohydrate metabolic pathways. *Res. Microbiol.* **147**, 448–455.
- Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M. and Karp, P.D. (2004) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* **6**, R2.
- Sánchez, B., Zúñiga, M., González-Candelas, F., de los Reyes-Gavilán, C.G. and Margolles, A. (2010) Bacterial and eukaryotic phosphoketolases: phylogeny, distribution and evolution. *J. Mol. Microbiol. Biotechnol.* **18**, 37–51.
- Schuetz, R., Kuepfer, L. and Sauer, U. (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.* **3**, 119.
- Siaut, M., Heijde, M., Mangogna, M., Montsant, A., Coesel, S., Allen, A., Manfredonia, A., Falcatore, A. and Bowler, C. (2007) Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. *Gene*, **406**, 23–35.
- Siron, R., Giusti, G. and Berland, B. (1989) Changes in the fatty acid composition of *Phaeodactylum tricornutum* and *Dunaliella tertiolecta* during growth and under phosphorus deficiency. *Mar. Ecol.-Prog. Ser.* **55**, 95–100.
- Smid, E.J., Molenaar, D., Hugenholtz, J., de Vos, W.M. and Teusink, B. (2005) Functional ingredient production: application of global metabolic models. *Curr. Opin. Biotechnol.* **16**, 190–197.
- Teusink, B., van Enkevort, F.H.J., Francke, C., Wiersma, A., Wegkamp, A., Smid, E.J. and Siezen, R.J. (2005) In silico reconstruction of the metabolic pathways of *Lactobacillus plantarum*: comparing predictions of nutrient requirements with those from growth experiments. *Appl. Environ. Microbiol.* **71**, 7253–7262.
- Tirichine, L. and Bowler, C. (2011) Decoding algal genomes: tracing back the history of photosynthetic life on Earth. *Plant J.* **66**, 45–57.
- Wen, Z.-Y. and Chen, F. (2003) Heterotrophic production of eicosapentaenoic acid by microalgae. *Biotechnol. Adv.* **21**, 273–294.
- Wessely, F., Bartl, M., Guthke, R., Li, P., Schuster, S. and Kaleta, C. (2011) Optimal regulatory strategies for metabolic pathways in *Escherichia coli* depending on protein costs. *Mol. Syst. Biol.* **7**, 515.
- Whitaker, J.W., McConkey, G.A. and Westhead, D.R. (2009) Prediction of horizontal gene transfers in eukaryotes: approaches and challenges. *Biochem. Soc. Trans.* **37**, 792–795.
- Ye, Q., Cao, H., Yan, M. et al. (2010) Construction and co-expression of a polycistronic plasmid encoding carbonyl reductase and glucose dehydrogenase for production of ethyl (S)-4-chloro-3-hydroxybutanoate. *Bioreour. Technol.* **101**, 6761–6767.
- Zhang, P., Foerster, H., Tissier, C.P., Mueller, L., Paley, S., Karp, P.D. and Rhee, S.Y. (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol.* **138**, 27–37.
- Zhukova, N.V. and Aizdaicher, N.A. (1995) Fatty acid composition of 15 species of marine microalgae. *Phytochemistry*, **39**, 351–356.

Identification of putative cancer genes through data integration and comparative genomics between plants and humans

Mauricio Quimbaya · Klaas Vandepoele · Eric Raspé ·
Michiel Matthijs · Stijn Dhondt · Gerrit T. S. Beemster ·
Geert Berx · Lieven De Veylder

Received: 22 August 2011 / Revised: 11 December 2011 / Accepted: 13 December 2011 / Published online: 5 January 2012
© Springer Basel AG 2012

Abstract Coordination of cell division with growth and development is essential for the survival of organisms. Mistakes made during replication of genetic material can result in cell death, growth defects, or cancer. Because of the essential role of the molecular machinery that controls DNA replication and mitosis during development, its high degree of conservation among organisms is not surprising. Mammalian cell cycle genes have orthologues in plants, and vice versa. However, besides the many known and characterized proliferation genes, still undiscovered regulatory genes are expected to exist with conserved functions

Electronic supplementary material The online version of this article (doi:[10.1007/s00018-011-0909-x](https://doi.org/10.1007/s00018-011-0909-x)) contains supplementary material, which is available to authorized users.

M. Quimbaya · K. Vandepoele · M. Matthijs · S. Dhondt ·
G. T. S. Beemster · L. De Veylder (✉)
Department of Plant Systems Biology, VIB,
Technologiepark 927, 9052 Gent, Belgium
e-mail: lieven.deveylder@psb.vib-ugent.be

M. Quimbaya · K. Vandepoele · M. Matthijs · S. Dhondt ·
G. T. S. Beemster · L. De Veylder
Department of Plant Biotechnology and Bioinformatics,
Ghent University, Technologiepark 927, 9052 Gent, Belgium

M. Quimbaya · E. Raspé · G. Berx
Molecular and Cellular Oncology Unit,
Department for Molecular Biomedical Research, VIB,
Technologiepark 927, 9052 Gent, Belgium

M. Quimbaya · E. Raspé · G. Berx
Department of Biomedical Molecular Biology,
Ghent University, Technologiepark 927,
9052 Gent, Belgium

G. T. S. Beemster
Department of Biology, University of Antwerp,
Groenenborgerlaan 171, 2020 Antwerpen, Belgium

in plants and humans. Starting from genome-wide *Arabidopsis thaliana* microarray data, an integrative strategy based on coexpression, functional enrichment analysis, and *cis*-regulatory element annotation was combined with a comparative genomics approach between plants and humans to detect conserved cell cycle genes involved in DNA replication and/or DNA repair. With this systemic strategy, a set of 339 genes was identified as potentially conserved proliferation genes. Experimental analysis confirmed that 20 out of 40 selected genes had an impact on plant cell proliferation; likewise, an evolutionarily conserved role in cell division was corroborated for two human orthologues. Moreover, association analysis integrating *Homo sapiens* gene expression data with clinical information revealed that, for 45 genes, altered transcript levels and relapse risk clearly correlated. Our results illustrate how a systematic exploration of the *A. thaliana* genome can contribute to the experimental identification of new cell cycle regulators that might represent novel oncogenes or/and tumor suppressors.

Keywords *Arabidopsis thaliana* · MCF7 · Cell cycle · Cancer genomics · Comparative genomics

Abbreviations

CDK	Cyclin-dependent kinase
EI	Endoreduplication index
fRMA	Frozen Robust Multiarray Analysis
GO	Gene ontology
HU	Hydroxyurea
PCC	Pearson correlation coefficient
PWM	Positional Weight Matrix
QPCR	Quantitative polymerase chain reaction
siRNA	Small interfering RNA

Introduction

The cell cycle represents a precisely programmed series of events that enables a cell to duplicate its content and to generate two daughter cells. In all eukaryotes studied to date, the cell division process is controlled by cyclin-dependent kinases (CDKs) [1, 2]. The numerous components controlling the activity of these kinases form a complex molecular network that has not been fully dissected even 30 years after their initial discovery. All physiological signals and signaling pathways affecting cell proliferation are in some way connected to the cell cycle regulators. Therefore, it is not surprising that mutations in key steps within these signaling pathways provoke dramatic changes in DNA replication, DNA repair efficiency, and cell proliferation rate. In mammals, a deregulated cell cycle is directly linked with malignant transformation processes that lead to tumorigenesis and cancer.

A wide spectrum of strategies has been used to identify new oncogenes or cell malignancy modulators, from proteomics studies [3] and cytogenetics [4] to cancer epigenetics [5]. With the technological progress in gene expression techniques, methods such as digital differential display [6, 7] and serial analysis of gene expression (SAGE) [8] have been used as tools to discover new oncogenes and tumor suppressors. Microarrays have also been employed as a highly preferred technology to characterize cancer-specific expression patterns (cancer fingerprints) and cancer-deregulated pathways [9–13]. Additionally, recent technological advances have provided platforms that allow hundreds of thousands of single nucleotide polymorphisms (SNPs) to be analyzed in genome-wide association studies (GWAS), providing a basis for the identification of moderate-risk alleles that contribute to cancer progression [14–16]. Nevertheless, in spite of the invaluable information obtained with these tools, cancer persists as one of the major killing diseases in the world [17]. Therefore, it is desirable to develop additional approaches that allow us to get better and more systemic, insight into the origin, progression, and outcome of cancer.

Comparative genomics represents a complementary tool for cancer research [18–20]. Although 1.6 billion years ago the mammalian and plant clades had diverged, commonly shared pathways and signaling cascades inherited from their last common ancestor still persist. Correspondingly, *Arabidopsis thaliana* not only has had a great impact on the understanding of the plant kingdom itself but has also contributed extensively to the dissection of specific mechanisms that have been evolutionarily conserved. Innate immunity [21], circadian clock [22], DNA methylation [23], RNAi processing mechanisms [24], and G protein signaling [25] are some of the traits firstly studied in *Arabidopsis*. Similarly, the *Arabidopsis* and *Homo*

sapiens genomes contain a highly comparable repertoire of “disease genes”. Almost 70% of the genes implicated in cancer have *Arabidopsis* homologues, which is comparable to the percentage found in *Drosophila melanogaster* (67%), *Caenorhabditis elegans* (72%), and *Saccharomyces cerevisiae* (41%) [26].

Regarding cancer, nowadays an old paradigm has been reinforced, namely that, underlying the variability among different tumors, only a relatively small number of critical events lie at the origin of their development. In most instances, deregulated cell proliferation provides the fundamental platform for neoplastic transformation [27, 28]. Through microarray expression analysis of different types of cancers, it has been possible to detect the cancer core mechanisms, represented by an early deregulation of the mitotic cell cycle, DNA replication, DNA repair, and chromatin assembly. Interestingly, all these processes are largely controlled by the RB-E2F pathway [29], in agreement with the common alteration of this pathway in cancer [30, 31].

The RB-E2F pathway is one of the most conserved pathways between plants and mammals, as illustrated by the large amount of E2F target genes that are shared by both organisms [32, 33]. Therefore, given that at its early stages abnormal cell proliferation is a cancer hallmark, new cell cycle regulators with a specific role in carcinogenesis might be identified by a systematic study of the cell replication machinery in *Arabidopsis*. Here, we applied a combination of functional prediction and comparative genomics strategies to identify evolutionarily conserved cell cycle genes. A subset of the computational identified genes was tested experimentally, both in plant and human cell cultures, to validate their role in cell cycle progression. A Cox survival analysis revealed a strong enrichment for genes that upon misexpression might result in cancer relapse, demonstrating that the designed integrative strategy had been successful in detecting novel cell division genes that were conserved between humans and plants.

Materials and methods

Arabidopsis microarray expression data analysis and clustering

Microarray data were retrieved from the NASC transcriptomics service [34]. Based on the Affymetrix ATH1 array, 20,777 *A. thaliana* genes were analyzed using 213 microarray CEL files covering different tissues and under different experimental conditions (Supplementary Table 1). To detect coexpressed genes, all 20,777 *Arabidopsis* genes were used as seed to detect coexpression neighborhoods using the complete expression compendium. The Pearson correlation coefficient (PCC) was calculated for each pair of

genes within the dataset, generating a $20,777 \times 20,777$ data matrix. For all the pair-wise comparisons, a significance value of coexpression between the compared genes was established [35].

Gene ontology associations

Gene ontology (GO) associations for *Arabidopsis* proteins were retrieved from TAIR [36] and for human proteins from AmiGO [37]. The assignments of genes to the original GO categories were extended to include parental terms (i.e., a gene assigned to a given category was automatically also assigned to all the parent categories). Enrichment values for the GO terms DNA repair (GO:0006281) and DNA replication (GO:0006260) for both *Arabidopsis* and *H. sapiens* were calculated as the ratio of the relative occurrence in a set of genes (coexpression neighborhood) to the relative occurrence in the genome. The statistical significance of the functional enrichment within sets of genes was evaluated with the hypergeometric distribution adjusted by the Bonferroni correction for multiple hypothesis testing. Corrected *P* values smaller than 0.05 were considered as significant. GO enrichment analysis for validating the different filtering steps was performed using ATCOECIS (<http://bioinformatics.psb.ugent.be/ATCOECIS/>).

Cis-regulatory elements detection

One-kb promoter regions of the set of genes significantly enriched for the terms DNA repair and/or DNA replication were scanned for the presence of an E2F binding-site by means of a positional weight matrix (PWM), with TTTsCGC as consensus sequence (based on a set of E2F-upregulated genes; [38]). E2F motif instances were identified with MotifLocator and using a threshold of 0.95 [39].

Detection of orthologous genes

Orthologous genes between *Arabidopsis* and *H. sapiens* were identified with OrthoMCLDB [40], a comparative genomics resource hosting orthologous families based on protein clustering. Starting from the selected *Arabidopsis* genes, the corresponding orthologous gene families were retrieved and evaluated by phylogenetic inference. For each family, protein sequences were aligned using MUSCLE [41] and a neighbor-joining phylogenetic tree was constructed using TREECON [42], with the Poisson correction for evolutionary distance calculation. Highly supported nodes (bootstrap support >90%), indicating the speciation between plants and mammals, were used to identify orthologous genes and copy numbers.

Human microarray data analysis

The human microarray data analysis comprised CEL files of studies performed on Affymetrix array platforms compatible with the mRNA expression data (HG133A or HG133plus2), involving at least 50 breast tumor samples (Supplementary Table 2) published before September 2009 in the GEO or Array Express databases. Data were extracted, background-subtracted, normalized, and summarized (median polish option) using frozen (f)RMA, the new summarization Bioconductor package [43]. Data from the nine selected studies were merged in a pooled dataset. To avoid over-fitting, data corresponding to the same patient analyzed in different studies were included only once in the pooled dataset containing 1,400 patients. Statistical processing and Cox survival analysis were performed as given in the Supplemental Methods file.

Plant growth conditions and phenotypic analysis

Arabidopsis thaliana (L.) Heyhn. accession Columbia-0 and the mutant plants were grown under long-day conditions (16 h/8 h light/darkness) at 22°C on half-strength Murashige and Skoog (MS) agar plates. All the insertion T-DNA lines were obtained from the European *Arabidopsis* Stock Centre (NASC). To screen for homozygous insertion alleles, primers were designed following the instructions of the Salk Institute genomic analysis laboratory (<http://signal.salk.edu/tdnaprimers.2.html>). The complete list of the used primers for the selection of homozygous lines is detailed in Supplementary Table 3. For characterization of embryo-lethal mutants, independent seedpods (>10) from different plants were harvested and dissected. Pictures were taken with a Leica MZ16 stereoscope using a $\times 5$ magnification factor. The number of aborted seeds was correlated with the proportion of expected homozygous seeds; the significance of this correlation was tested with the χ^2 statistical test. For DNA ploidy analysis, the first developed leaf (harvested 3 weeks after sowing) was chopped with a razor blade in 200 μ l of nucleus extraction solution, supplemented with 800 μ l of staining solution (<http://www.partec.com>). The homogenate was filtered through a 30- μ m mesh. The nuclei were analyzed using a CyFlow cytometer and FloMax software (<http://www.partec.com>). The EI was calculated as the fraction of nuclei of each represented ploidy level multiplied by the number of endoreduplication cycles necessary to reach the corresponding ploidy level. Leaf cell number and cell size measurements and root growth analysis were performed as given in the Supplemental Methods file.

MCF7 cell culture and transfection

MCF7 cell cultures were grown in complete medium (Dulbecco's modified MEM Eagle medium with 5% fetal calf serum, supplemented with L-Gln, NaPy, NEAA, and 6 ng/ml bovine insulin) at 37°C and 5% CO₂. The following small interfering (si)RNA sequences (DharmaFECT; Thermo Fisher Scientific, Waltham, MA, USA), were used for the specific transfections: human HEATR6 (SMARTpool; J-015921-09, J-015921-10, J-015921-11, J-015921-12), human STAT1P1 (SMARTpool; J-021064-05, J-021064-06, J-021064-07, J-021064-08), human C14ORF21 (SMARTpool; J-017798-09, J-017798-10, J-017798-11, J-017798-12) and control (SMARTpool non-targeting pool). Growth and ploidy content were measured as given in the Supplemental Methods file.

QPCR analysis of *Homo sapiens* siRNAs

MCF7 cells (250,000 cells approximately) were seeded in 5 ml of MCF7 medium without antibiotics in a 6-well plate and grown under the previously described conditions. STAT1P1, C14ORF21, HEATR6, and control siRNAs were transfected into the cells according to the manufacturer's instructions (DharmaFECT; Thermo Fisher Scientific). The final concentration of each siRNA was 30 nM. Cells were collected 48 h after transfection with a rubber policeman. RNA was extracted with an RNeasy animal Mini Kit (Qiagen) and cDNA was prepared with the cDNA synthesis system according to the manufacturer's instructions (Roche Diagnostics, Indianapolis, USA). For quantitative PCR, a Light-Cycler 480 SYBR Green I Master (Roche Diagnostics) was used with 100 nM primers and 0.1 mg of reverse transcription reaction product. Reactions were run and analyzed on the LightCycler 480 Real Time PCR System according to the manufacturer's instructions (Roche Diagnostics). All quantifications were normalized to the TATA binding protein (TBP) and Ubiquitin C (UBC) expression levels. Quantitative reactions were done in triplicate and averaged. Primers used for QPCR analysis are given in Supplementary Table 4.

Results

Selection of target genes using data integration and comparative genomics

To identify new genes playing a putative role in the regulation of the cell cycle in plants and humans, we applied an integrative genomics strategy (Fig. 1). Starting from >200 microarray experiments (Supplementary Table 1), the expression levels for 20,777 *Arabidopsis* genes were

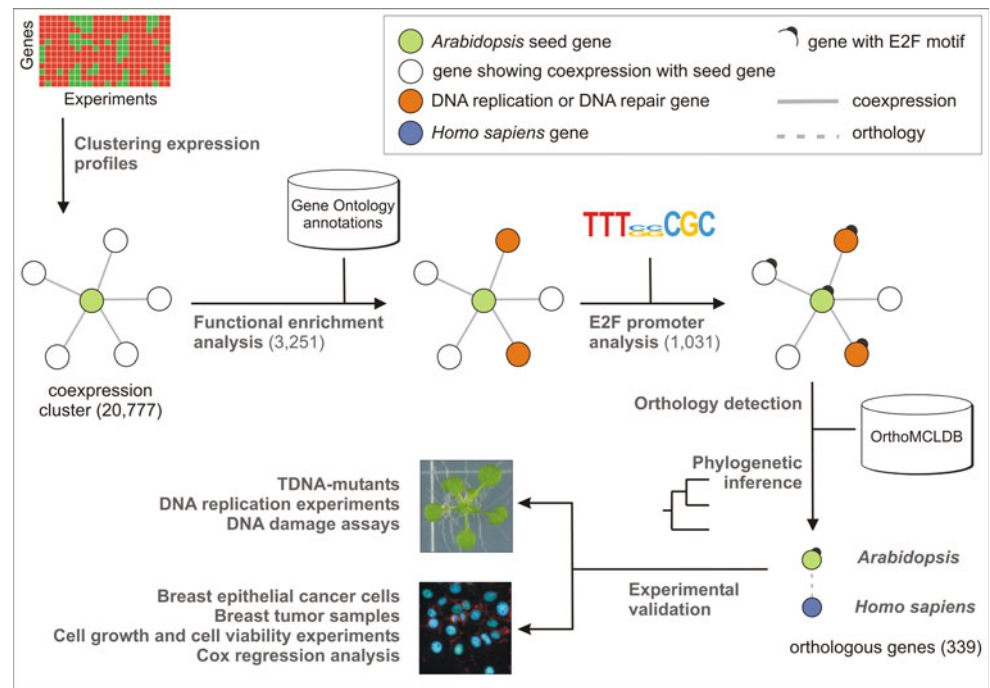
used to identify gene coexpression neighborhoods based on the Pearson correlation coefficient (PCC) (see “Materials and methods”). Depending on the seed gene, neighborhood clusters of coexpressed genes contained between 10 and 450 genes. Subsequently, each gene cluster was tested for functional enrichment with GO. The terms “DNA replication” (GO:0006260) and “DNA repair” (GO:0006281) were scanned within the annotations of the coexpressed neighbors of all seed genes. In total, 3,251 genes were significantly enriched ($P < 0.05$) for one or both terms (Supplementary Table 5). To identify within this list the genes with a putative role in DNA replication or DNA repair, the 1-kb promoter regions of the 3,251 genes were scanned for the presence of E2F *cis*-regulatory elements by means of a PWM with a consensus sequence TTTssCGC (see “Materials and methods”). A total of 1,031 *Arabidopsis* genes were found, harboring one or more predicted E2F-binding sites within their promoter region (Supplementary Table 6). Subsequently, to select only those genes with a putatively conserved role across species, plant genes with a mammalian orthologue were identified with the OrthoMCL database (<http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi>). The sets of orthologues were verified by means of phylogenetic inference (see “Materials and methods”), and for 515 genes at least one human orthologue was identified. As functional redundancy might obscure downstream functional analysis upon gene knock-out, only those genes that were part of a low copy number family in both *Arabidopsis* and human were retained. A total of 339 genes fitted this criterion (Supplementary Table 7).

A GO enrichment analysis was performed to validate the effectiveness of the used filters (see “Materials and methods”). This analysis demonstrated a progressive enrichment for both GO terms after each filter applied (Supplementary Table 8), illustrating that the application of the E2F and the *Arabidopsis*–*H. sapiens* orthology filters effectively resulted in an enrichment of the candidate genes with a putative role in DNA replication and/or DNA repair.

Validation of the putative cell cycle regulators using the plant model

To experimentally validate a subset of the above-identified genes as novel plant cell cycle genes, we screened for potential *Arabidopsis* knock-out lines in the available T-DNA insertion collections (<http://signal.salk.edu/cgi-bin/tdnaexpress>). Forty genes were randomly selected that harbored a T-DNA insertion inbetween the translational start and stop codons, either in an intron or in an exon (Table 1). No homozygous T-DNA insertion lines could be identified for three genes (*AT1G06590*, *AT4G07410*, and

Fig. 1 Schematic representation of the applied methodology for the selection of the target genes using data integration and comparative genomics. Starting from genome-wide *Arabidopsis thaliana* microarray data, an integrative strategy based on coexpression, functional enrichment analysis, and *cis*-regulatory element annotation was combined with a comparative genomics approach between plants and humans to detect conserved cell cycle genes involved in DNA replication and DNA repair processes. Numbers in parentheses report the number of genes that were retained after each step



AT5G22370), indicating that their deficiency was embryonically lethal. Indeed, when the seedpods of the hemizygous lines were analyzed in detail, 25% of the embryos were aborted, indicative of an embryo lethal phenotype ($P < 0.01$ according to the statistical χ^2 test), and suggesting that the proteins encoded by these three genes are essential for embryogenesis (Supplementary Fig. 1).

For the available homozygous insertion lines, effects on overall cell division and DNA replication activity were determined for the first developed leaf pair harvested at maturity (3 weeks after sowing). As demonstrated previously, the first leaf of *Arabidopsis* is an excellent model system to study cell division and DNA replication parameters [44–47]. As the leaf grows, its cells progressively shift from a dividing mode to a phase during which they exit their cell cycle program and start to expand. Mutations that affect cell division affect the total number of cells formed at leaf maturity. Furthermore, the cell expansion phase is correlated with the onset of endoreduplication, an alternative cell cycle during which cells continue to replicate their DNA without cell division. Mutations that affect the endoreduplication index (EI; the mean number of endoreduplication cycles) of the leaf are indicative of a change in the cell differentiation timing, with a decreased or increased EI reflecting a delayed or premature cell cycle exit, respectively.

EI measurements revealed a shift in DNA ploidy distribution for 15 of the 40 knockout lines (37 homozygous knockouts and the 3 hemizygous mutants) (Fig. 2a;

Supplementary Fig. 2). In contrast, among 11 randomly selected insertion lines, only 1 (*AT5G46160*) displayed a replication phenotype (Supplementary Fig. 3), illustrating a strong enrichment for replication mutants in the selected set of mutants. In five mutant lines, the EI was lower than that in wild-type plants, whereas for ten knockout lines it was higher (Table 2). Although the mutant line for *AT1G72320* (*APUM23*) had an EI almost identical to that of the control plants, it displayed a totally different DNA ploidy distribution (Supplementary Fig. 2), which implies that proliferation in this line was both stimulated and inhibited, probably in a tissue-specific manner.

Changes in the DNA content due to an altered cell differentiation timing should affect the total leaf cell number and cell size distribution, in which a delayed or premature onset of cell differentiation often correlates with smaller or bigger cells, respectively [48]. Therefore, cell number and cell size distribution analyses of the leaf epidermal cells were performed. When the average cell numbers and cell sizes were plotted, two main subgroups of mutants could be recognized: one characterized by more but smaller cells, and one with few but larger cells, than those of the wild-type plants (Fig. 2b). According to the flow cytometric measurements, a subgroup of mutant lines in the first group had a reduced EI (green dots), showing that the differences at the DNA ploidy level originated from enhanced cell proliferation or delayed cell differentiation. Conversely, the other subgroup of mutants comprised plants displaying an increased DNA ploidy content (red dots), indicative of premature cell cycle exit. The data were substantiated by

Table 1 Genes selected for downstream experimental validation

Arabidopsis line	TAIR Annotation	T-DNA accession	HUGO
AT1G01940	F22M8.7	061120.53.75.X-Intronic	PPIL3
AT1G03110	TRM82	025857.27.50.X-Exonic	WDR4
AT1G03530	ATNAF1	013589.53.50.X-Exonic	NAF1
AT1G04020	ATBARD1	031862.53.75.X-Exonic	BARD1
AT1G06590	F12K11.7	024997.29.40.X-Intronic	ANAPC5
AT1G08410	T27G7.9	119395.38.15.X-Exonic	LSG1
AT1G10490	T10O24.10	070262.56.00.X-Intronic	NAT10
AT1G13330	AHP2	136002.41.85.X-Exonic	PSMC3IP
AT1G49540	ATELP2	106485.50.75.X-Intronic	ELP2-STATIP1
AT1G72320	APUM23	052992.53.50.X-Intronic	C14ORF21
AT1G74150	F9E11.8	088010.26.55.X-Exonic	KHLDC3
AT1G76260	DWA2	143341.50.65.X-Exonic	TSSC1
AT2G15790	CYCLOPHILIN 40	033511.51.20.X-Intronic	PPID
AT2G19430	ATTHO6	051022.41.15.X-Exonic	THOC6
AT2G28450	T1B3.3	039998.52.40.X-Exonic	TRMT2A
AT2G40550	ETG1	145460.18.05.X-Exonic	MCMBP
AT2G34260	F13P17.10	063054.55.75.X-Intronic	WDR55
AT3G02220	F14P3.13	028532.34.35.X-Exonic	C9ORF85
AT3G07050	F17A9.21	099852.47.75.X-Exonic	GNL3
AT3G26410	ATTRM11	122158.32.05.X-Exonic	TRMT11
AT3G42660	T12K4.110	052512.12.95.X-Exonic	WDHD1
AT3G49990	F3A4.70	090801.18.60.X-Exonic	LTV1
AT3G55160	T26I12.40	006621.56.00.X-Exonic	THADA
AT3G56990	EDA7	098429.45.45.X-Exonic	NOL10
AT3G60660	T4C21.70	041743.49.40.X-Exonic	C18ORF24-SKA1
AT4G00850	GIF3	052744.30.10.X-Exonic	SS18
AT4G01270	F2N1.19	056467.55.00.X-Exonic	TRAIP
AT4G07410	F28D6.14	022607.45.25.X-Exonic	CIRH1A
AT4G15890	DL3985 W	094776.23.50.X-Intronic	NCAPD3
AT4G20350	F9F13.6	138864.18.85.X-Exonic	ALKBH6
AT4G22970	AESP	037016.52.60.X-Intronic	ESPL1
AT4G35910	T19K4.40	030197.20.30.X-Intronic	CTU2
AT4G38120	F20D10.240	066582.56.00.X-Exonic	HEATR6
AT5G05660	ATNFXL2	017558.18.75.X-Exonic	NFXL1
AT5G11240	F2I11.130	052897.39.70.X-Exonic	WDR43
AT5G14600	T15N1.90	024680.34.10.X-Exonic	TRMT61B
AT5G22370	EMB1705	059852.56.00.X-Intronic	GPN2
AT5G40530	MNF13.4	102154.30.95.X-Intronic	RRP8
AT5G49110	K20J1.8	055483.52.00.X-Exonic	FANCI
AT5G61770	PAN-LIKE	088929.56.00.X-Exonic	PPAN

HUGO Gene nomenclature,
Homo sapiens official symbol

cell size distribution analysis, with those mutants showing a decreased EI exhibiting an increased subpopulation of small cells, in comparison with control plants. Conversely, the mutants that displayed an increased EI were enriched in enlarged cells (Supplementary Fig. 4). In the mutant line for *AT1G72320*, the population of both small and large cells had increased, hinting again at a dual effect of this gene on cell proliferation.

DNA damage assays

As the screening method involved a selection of genes displaying a significant enrichment of genes involved in DNA repair among coexpressed neighbors, the knock-out lines were tested for hypersensitivity toward DNA replication inhibiting stress treatments, including UV-B (UV) radiation and hydroxyurea (HU) treatment. UV-B radiation

Fig. 2 Experimental association of *Arabidopsis thaliana* candidate genes with cell replication. The *Arabidopsis* first leaf was used to measure cell division and DNA replication parameters. **a** The mean number of endoreduplication cycles denoted as Endoreduplication Index (EI) of the T-DNA insertion lines [*statistically different from the control (*Col-0*) plants, according to the *t* test $P < 0.05$ ($n = 10$); \pm represents hemizygous mutants]. **b** Scatter plot of the analyzed mutants. Mutants were plotted according to their respective number of cells and cell size. Mutant lines are color-coded according to their DNA ploidy content phenotype. *Green* and *red* dots represent mutants with a reduced and increased EI, respectively

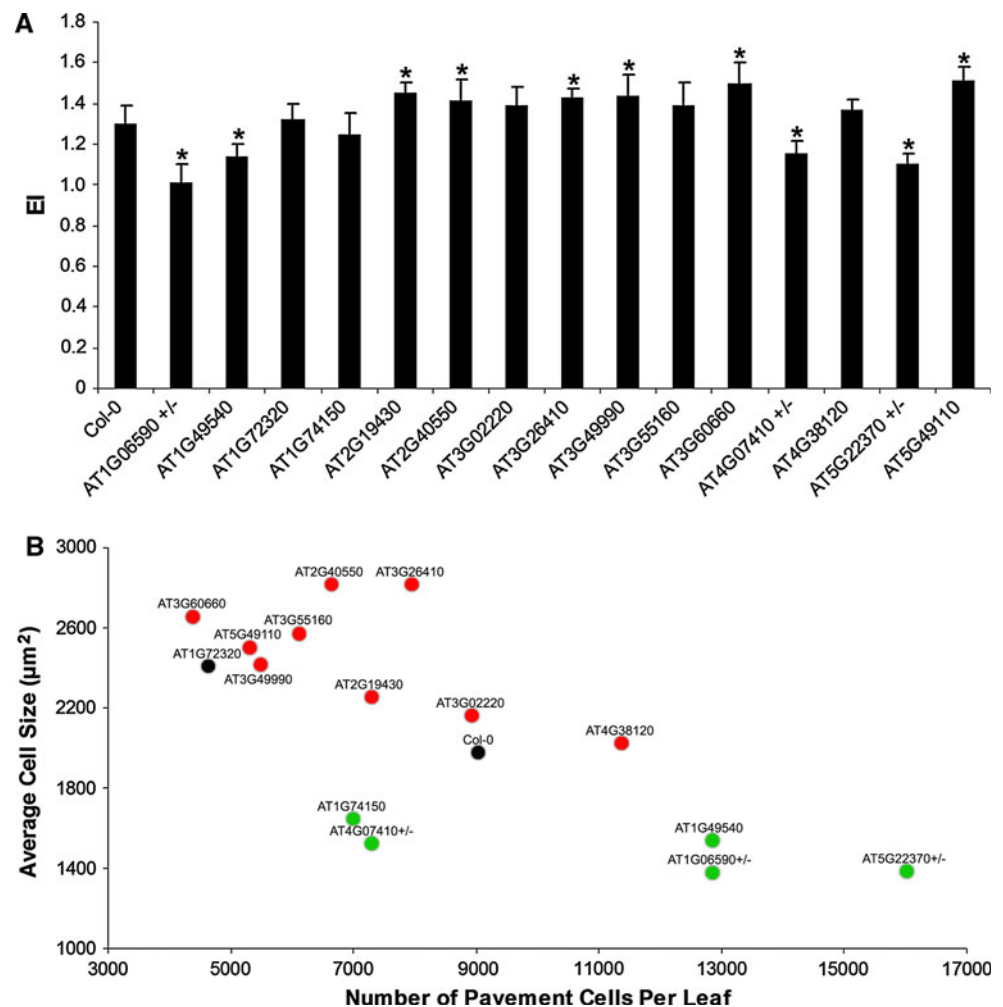


Table 2 Analysis of endoreduplication index (EI), pavement cell size, and cell number in the first developed leaf pair of the studied T-DNA insertion mutants

Line	EI	Average leaf area (mm ²)	Average cell size (μm ²)	Pavement cells per mm ² (Cell density)	Pavement cells per leaf
Col-0	1.29	28.2	1,976	320	9,040
AT1G06590 +/-	1.01	23.0	1,380	559	12,864
AT1G49540	1.13	38.7	1,537	332	12,850
AT1G72320	1.32	17.9	2,409	258	4,626
AT1G74150	1.24	20.1	1,643	347	6,987
AT2G19430	1.45	27.0	2,252	270	7,289
AT2G40550	1.42	21.5	2,819	309	6,651
AT3G02220	1.39	40.0	2,165	223	8,937
AT3G26410	1.43	33.0	2,816	241	7,939
AT3G49990	1.44	26.4	2,419	208	5,499
AT3G55160	1.39	29.9	2,569	204	6,110
AT3G60660	1.50	26.6	2,651	165	4,376
AT4G07410 +/-	1.16	18.3	1,524	398	7,286
AT4G38120	1.37	30.1	2,020	377	11,359
AT5G22370 +/-	1.10	29.1	1,387	550	16,016
AT5G49110	1.51	26.2	2,498	203	5,320

+/- Hemizygous lines

dimerizes adjacent pyrimidine bases, and inhibits replication and transcription, eventually causing a growth delay. Similarly, HU treatment causes a collapse of the replication fork, with inhibition of growth as a consequence. DNA damage was measured by comparing root growth under control and DNA-damaging growth conditions (see “Materials and methods”). Without any DNA stress treatment, the mutants for *ATG06590* (hemizygous mutant), *AT1G49540* (*ATELP2*), *AT1G72320* (*APUM23*), *AT2G40550* (*ETG1*), *AT3G55160*, and *AT3G60660*, showed a significant root growth reduction ($P < 0.01$ according to Student's *t* test), when compared to wild-type Col-0 plants, displaying at 7 days after germination 35, 46, 67, 23, 33, and 44% of growth reduction, respectively. Conversely, the hemizygous mutants for *AT4G07410* and *AT5G22370* showed a significant increase in root growth (Fig. 3a). Wild-type plants were not hypersensitive towards UV-B (1.9 W/m^2). In contrast, the lines mutant for *AT1G01940* and *AT1G04020* showed a clear growth inhibition 72 h after the treatment (Fig. 3b). Similarly, these two mutants displayed a root growth inhibition stronger than that observed for the wild-type plants when treated with 1 mM HU for 6 days (Fig. 3c).

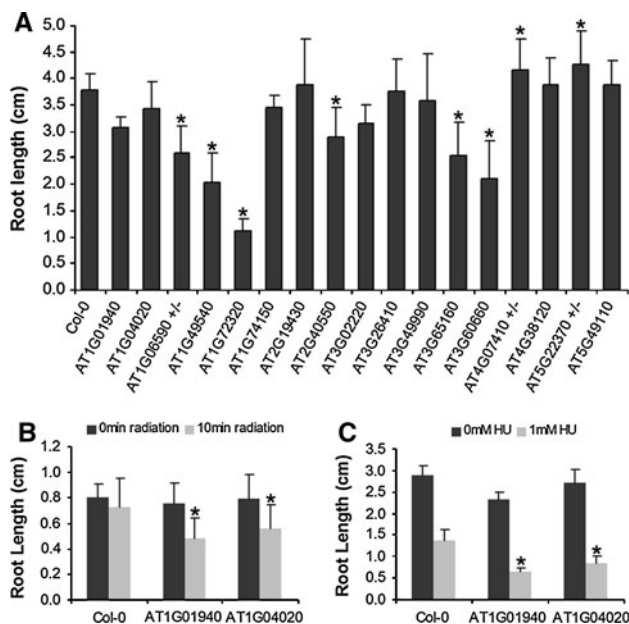


Fig. 3 Hypersensitivity of selected T-DNA insertion lines towards DNA replication-inhibiting treatments. **a** Root length under standard growth conditions for the analyzed T-DNA insertion lines. Roots were measured after 7 days of growth on vertical MS plates. **b**, **c** Mutants displaying a differential root growth response upon UV-B irradiation or in the presence of 1 mM HU, respectively [*Statistically different from the control (*Col-0*) plants according to the *t* test $P < 0.05$ ($n = 30$); \pm represents hemizygous mutants]

Validation of putative cell cycle regulators in MCF7 cells

To test whether the obtained gene list had a predictive power for detecting cell cycle-related genes in the mammalian model, three human genes that, to our knowledge, had not been implicated in cancer origin or progression, were silenced in breast epithelial cancer cell cultures (MCF7 cells), including the orthologues of *AT1G49540* (*STAT1P1*), *AT4G38120* (*HEATR6*), and *AT1G72320* (*C14ORF21*). In *Arabidopsis*, knock-out of the *AT1G49540* gene resulted into an enhanced cell division phenotype, and the knock-out of the *AT4G38120* gene caused an early induction of the differentiation processes, whereas the knockout of the *AT1G72320* gene was responsible for a dual phenotype. Similarly to its plant counterpart, the co-expression neighborhood of *HEATR6* was enriched for the GO term “DNA repair” ($P < 0.01$ according to the hypergeometric distribution) (Table 3). This was not the case for *STAT1P1* and *C14ORF21*.

After transient knock-down of *STAT1P1*, *C14ORF21*, and *HEATR6* through specific siRNA pools, cell culture growth was monitored by the colorimetric MTT assay [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide] [49]. In comparison with controls [untransfected cells (WT) and cells transfected with si-control (NT)], knock-down of *C14ORF21* and *HEATR6* clearly affected growth (Fig. 4a). The reduced number of cells might be caused by a cell cycle arrest. To corroborate this possibility, flow cytometric experiments revealed a larger number of G2/M cells in the knock-down cultures of the *C14ORF21* and *HEATR6* genes than that in the controls (Fig. 4b), indicative of a transient G2 arrest. In agreement with these results, the transcripts of the G2/M marker genes *CDK1*, *CyclinB1*, and *CyclinB2* were up-regulated upon knock-down of *C14ORF21* and *HEATR6* (Fig. 4c).

Associations with cancer relapse probability

To assess the potential correlation between the phenotypes of the plant genes selected by means of the designed integrative approach with those of their corresponding human orthologues, we created a database of transcriptional profiles of 1,400 non-redundant breast cancer samples linked to well-annotated clinical information; including relapse events and relapse time (see “Materials and methods”). The significance of a particular association between gene expression and a relapse event was assessed by Cox regression analysis. To ensure that the increased statistical power of the analysis due to the great number of patients in the database did not lead to irrelevant association with relapse risk, we iteratively and randomly subdivided the initial patient set into two complementary

Table 3 Gene ontology (GO) enrichment conservation

	<i>Arabidopsis</i> fold enrichments			<i>Homo sapiens</i> fold enrichments		
	AGI code	DNA replication	DNA repair	HUGO	DNA replication	DNA repair
GO enrichment for the terms <i>DNA repair</i> and <i>DNA replication</i> was calculated for <i>Arabidopsis</i> and <i>Homo sapiens</i> and is given for each GO class. Statistical significance according to the hypergeometric distribution: * $P < 0.01$, ** $P < 0.05$, - no significant enrichment	AT1G01940	3.86*	2.92*	PPIL3	4.26*	2.59**
	AT1G04020	6.31*	3.07**	BARD1	22.23*	12.59*
	AT1G06590	5.07*	3.89*	APC5	1.89-	1.85-
	AT1G49540	3.73*	2.59*	STATIP1	0.00-	1.11-
	AT1G72320	5.04*	2.98*	C14ORF21	0.95-	1.85-
	AT1G74150	4.98*	3.68*	KLHDC3	0.95-	1.11-
	AT2G19430	4.38*	3.02*	THOC6	7.09*	4.07*
	AT2G40550	5.18*	3.03*	MCMBP	3.31*	5.93*
	AT3G02220	3.95*	0.00-	C9ORF85	0.00-	0.00-
	AT3G26410	4.31*	2.91*	TRMT11	3.31*	3.33*
	AT3G49990	3.95*	0.00-	LTV1	3.31*	1.85-
	AT3G55160	3.58*	2.72*	THADA	0.47-	0.00-
	AT3G60660	3.88*	2.69*	C18ORF24	23.65*	13.7*
	AT4G07410	4.06*	2.57*	CIRH1A	6.15*	2.22**
	AT4G38120	5.46*	2.76*	HEATR6	1.42-	3.33*
	AT5G22370	4.04*	2.69*	GPN2	0.95-	1.85-
	AT5G49110	3.94*	2.81*	FANCI	23.18*	14.07*

subsets of 100 training sets of 75% of the samples ($n = 1,050$) and 100 validation sets of the corresponding remaining samples ($n = 350$). The Cox survival analysis was performed independently in parallel with both the training and validation sets. We considered for further analysis only the probe sets with significant association (at the 0.01 level) with increased or decreased risk in at least 95% of the corresponding training sets and validation sets. After stability evaluation, 182 out of the 9,976 available reliable probe sets (see Supplemental Methods file) were associated with decreased risk of relapse, while 995 probe sets were associated with an increased relapse risk. Among these, genes known to be associated with good disease outcome, such as the *ESR1* estrogen and *PGR* progesterone receptors, were associated with a decreased risk of relapse. Conversely, genes known to be associated with poor disease outcome such as *ERBB2* or *TOP2A* were correlated with a significantly increased relapse risk, proving the validity of our database (Supplemental Fig. 5).

For the list of candidate genes resulting from the comparative analysis between plant and human, 211 reliable probe sets (corresponding to 169 human orthologues; Supplementary Table 9) were available, for which 162 were not significantly associated with relapse risk or their association with it was not stable upon cross-validation. Only one was stably associated with decreased risk of relapse. In contrast, 48 probe sets (corresponding to 45 genes) were stably associated with an increased relapse risk (Supplementary Table 9). Thus, compared to the 9,976 probe sets present in the whole database, the 221 probes

were significantly enriched in probe sets associated with an increased risk of relapse ($P = 1.14 \times 10^{-7}$ according to the hypergeometric distribution). Interestingly, among the 15 analyzed *Arabidopsis* mutant lines that displayed a leaf growth phenotype upon mutation, 6 were associated with an increased relapse risk. For these genes, comprising four uncharacterized genes and the well-characterized replication genes *BARD1* and *FANCI*, Cox survival curves showed a clear association between altered expression levels and a diminished probability of survival, indicating that they can be considered as good markers to predict disease outcome in human breast cancer (Fig. 5).

Discussion

The field of comparative genomics has been growing and evolving rapidly thanks to the massive amount of genomic data generated over the last decade. Here, we have integrated coexpression analysis with comparative genomics to identify putative new cell cycle genes. Previously, we had demonstrated that in *Arabidopsis* coexpression alone performs poorly to infer known biological gene functions [35]. To improve the predictive power of coexpression networks, we have combined different functional prediction elements (GO enrichment analysis and *cis*-regulatory element scoring) to create a reliable platform for the detection of novel conserved cell cycle regulators. Interestingly, recently available ChIP-Seq data [50] revealed that there is a highly significant overlap ($P = 2.75 \times 10^{-14}$ according to the

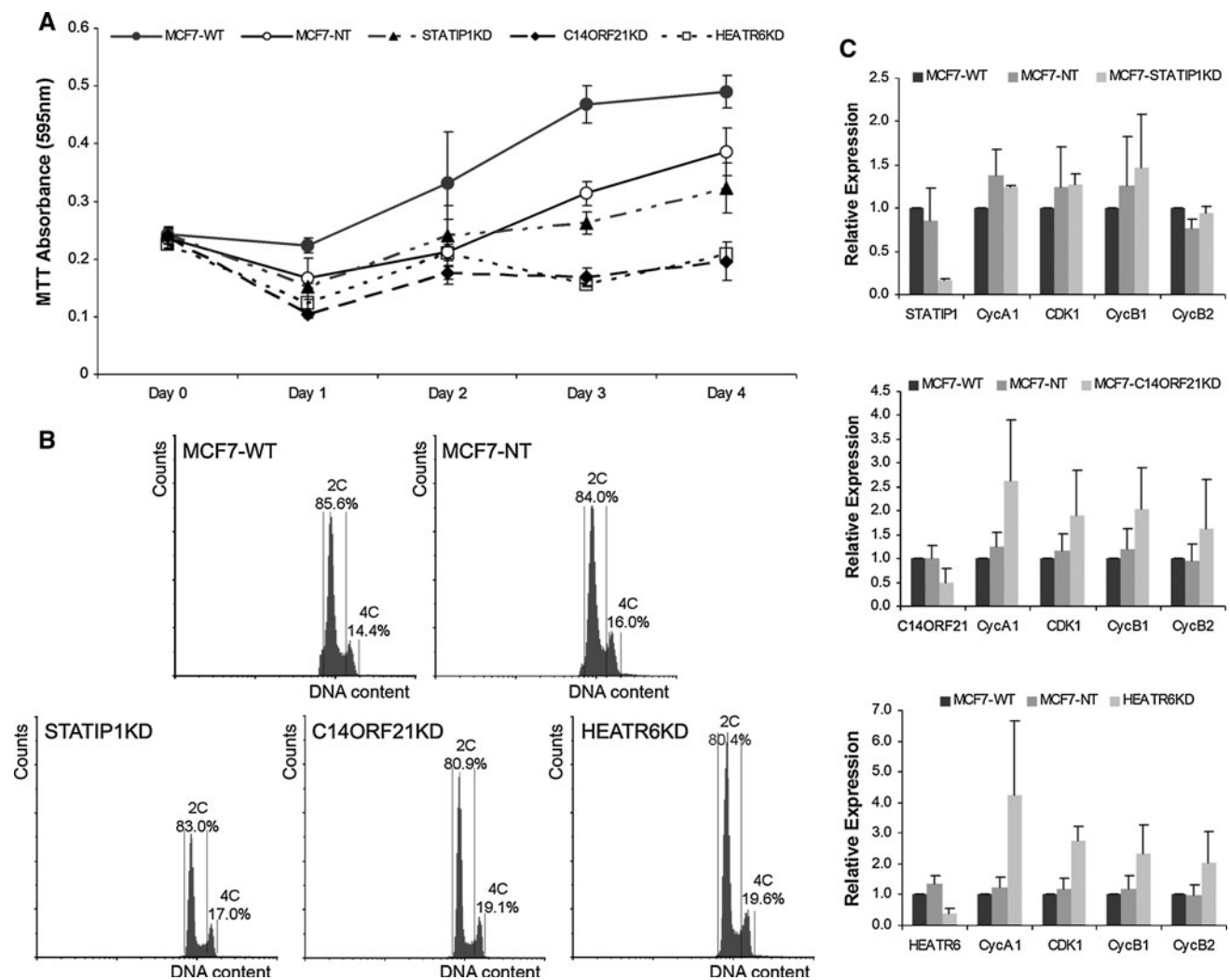


Fig. 4 Experimental association of human candidate genes with cell division in MCF7 cell cultures. Genes were silenced in breast epithelial cancer cell cultures (MCF7 cells) using small interfering (si) RNA sequences. **a** Growth curves of the siRNA knocked-down MCF7 cultures, illustrating growth inhibition by knock-down of *C14ORF21* and *HEATR6*. **b** Ploidy distributions of the *STAT1P1*, *C14ORF21*, and *HEATR6* knocked-down cultures in comparison with

controls assessed by flow cytometry, illustrating a significant increased number of G2/M cells in *C14ORF21* and *HEATR6* knock-down cultures ($P < 0.05$ ($n = 9$) according to a t test). **c** Expression levels of cell cycle phase makers measured by Q-PCR, illustrating transcriptional upregulation of the G2/M marker genes *CDK1*, *CyclinB1*, and *CyclinB2* in *C14ORF21* and *HEATR6* knock-down cell cultures, indicative for a transient G2 arrest

hypergeometric distribution), between the E2F target genes detected by our strategy and the genes that are predicted to be direct E2F targets on basis of the ChIP analysis.

The success rate of the integrative approach was illustrated by the observation that among 11 randomly selected T-DNA insertion lines only 1 displayed a DNA ploidy distribution profile different from wild-type plants, generating an identification rate of mutants possibly involved in replication events of 9%. In contrast, out of 40 plant candidate genes selected for downstream functional analysis, 15 were experimentally proven to affect cell proliferation, representing a success rate fivefold higher than that of the random approach. Moreover, two *Arabidopsis* mutant lines could be related with DNA stress responses and two human selected

orthologues clearly affected cell proliferation when knocked-down in breast epithelial cancer cells, emphasizing the highly significant predictive value of our integrative approach.

The importance of including *Arabidopsis* data in our search for novel cancer genes is illustrated by the observation that our final list of 339 genes retains 79 human genes that, according to the gene ontology classification (based on AMIGO), do not have a defined category (genes with unknown function). Similarly, there are 82 human genes that according to the GO classification are involved in functions totally unrelated to DNA replication and repair (Supplementary Table 7). This total of 161 genes represents half of the final list, illustrating the importance of the *Arabidopsis*-*H. sapiens* orthology relationship in order to

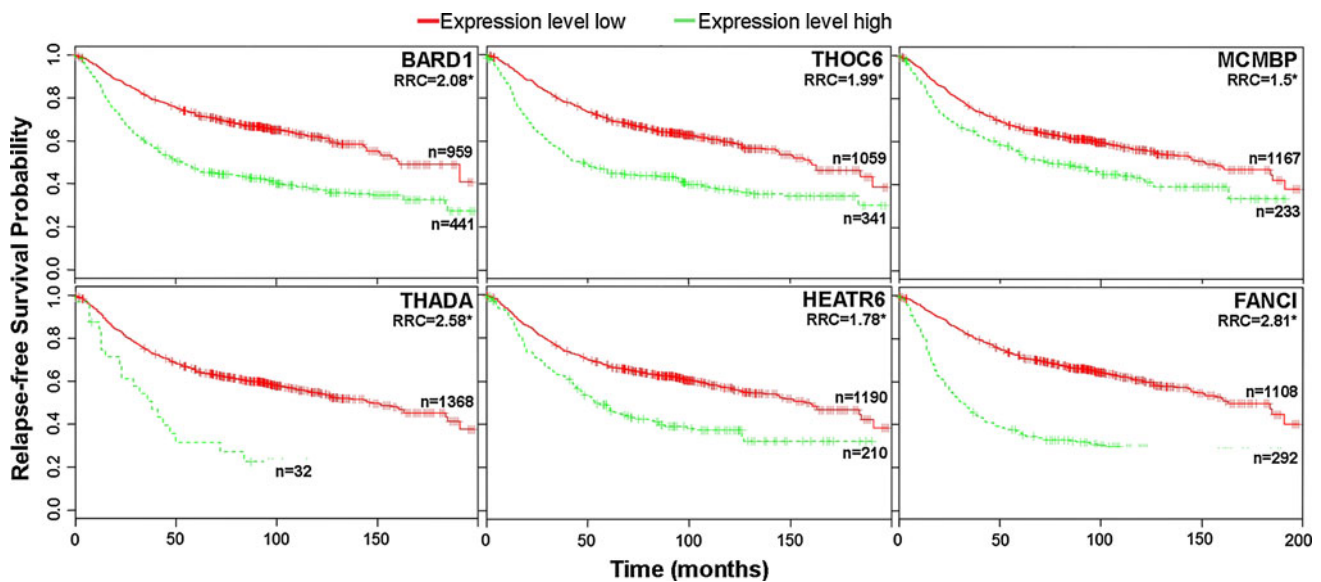


Fig. 5 Association of human orthologues of *Arabidopsis* genes involved in cell replication with specific cancer outcomes (relapse risk). Cox survival plots for the human orthologues of *Arabidopsis* genes with a direct influence on cell proliferation were constructed. A

clear association between increased gene expression levels and a diminished probability of relapse-free survival is shown. *RRC* relative risk coefficient, *statistically significant differences in the survival probability, $P < 0.01$

give or detect new gene functions even in highly distant organisms. A good example of the importance of the *Arabidopsis* filtering process is that the genes *THADA*, *HEATR6*, and *MCMBP*, which, according to the analyses presented, might represent important predictors of breast carcinomas, were exclusively retained in the final list of candidate genes due to the fact that their respective *Arabidopsis* orthologues were strongly associated with DNA replication processes.

Known proliferation genes populate the list of 339 candidate genes, that encode cell division control proteins (CDC6, CDC7, and CDC27), the retinoblastoma protein RB, replication proteins (MCM1, MCM2, MCM3, MCM4, MCM, MCM8, ORC1L, ORC2L, ORC3L, ORC5L, ORC6L, and PCNA), repair proteins (WEE1, PARP1, RAD50, RAD51, DDB1, and MRE11A), and previously characterized oncogenes (BARD1, BRIP1, API5, and ESPL1). These genes can be considered as positive controls. It suggests that the new genes found with this approach might be new cell cycle regulators. Indeed, we showed that 48 of the candidate genes have a significant prognostic value, at least for breast cancer, being associated with specific clinical outcomes when deregulated. In other words, 30%, of the retained genes are putative cancer predictors and represent highly significant cancer associations ($P < 0.01$ according to the hypergeometric distribution). Interestingly, comparing the data of a cancer gene census study [51], the candidate list of 339 genes showed a largely similar set of GO categories, although at a slightly different relative abundance (Supplementary Fig. 6).

Different facts argue in favor of the list of new cell cycle regulators to hold important elements in the mammalian cell cycle. First, two of the orthologous genes that are embryo lethal in *Arabidopsis* have an important role in the origin and progression of different diseases, including cancer. APC5, the human orthologue of the mutant line *AT1G06590*, is part of the gene set that is commonly misregulated during the onset and progression of breast and colorectal cancers [52]. CIRH1A, the human orthologue of the *Arabidopsis* embryo-lethal line *AT4G07410*, is the cause of the North American Indian Childhood Cirrhosis (NAIC/CIRH1A), a severe autosomal recessive intrahepatic cholestasis. All NAIC patients have a homozygous mutation in the CIRH1A protein, of which the function is still unknown [53]. Nevertheless, CIRH1A can upregulate a canonical NF- κ B element and might participate in the regulation of other genes containing NF- κ B responsive elements [54]. Because the activities of genes regulated through NF- κ B responsive elements are especially important during development, this interaction might explain not only the appearance of NAIC but it also suggests that *CIRH1A* misregulation is a new important element in the NF- κ B pathway, alterations of which have been extensively proved to lie at the basis of cancer origin and progression [55, 56].

Secondly, three of the genes in the final list have been linked recently with cell proliferation or DNA repair in plants and/or mammals. *SKA1*, orthologue of the *Arabidopsis* *AT3G60660* gene, plays a critical role in coupling chromosome movement to microtubule dynamics at the

outer kinetochore [57]. The plant orthologue of the well-studied mammalian breast cancer associated RING domain protein 1 gene (BARD1), involved in DNA repair, also controls DNA repair in plants [58]. Whereas this gene had been established to be essential for responding to the DNA cross-linking agent mitomycin, our results reveal that BARD1 knocked-down plants are sensitive toward UV irradiation and HU. Another example is the E2F TARGET GENE 1 (ETG1) protein that had been identified recently as a novel evolutionarily conserved replisome factor. ETG1 is associated with the minichromosome maintenance complex, being crucial for efficient DNA replication [59]. Additionally, depletion of ETG1 or its human orthologue MCM-BP, results in a stringent late G2 cell cycle arrest that correlates with a partial loss of sister chromatids cohesion [60, 61], hinting at an equally important developmental role for this molecule in plants and mammals.

Here, we found that the knock-down of the genes *C14ORF21* and *HEATR6*, which are orthologues of the *Arabidopsis* *AT1G72320* and *AT4G38120* genes, respectively, have an inhibitory effect on cell proliferation. We showed that depletion of *C14ORF21* and *HEATR6* resulted in an increase in the population of cells with a 4C DNA content, which is supported by an upregulation of G2/M cell cycle marker genes. Interestingly, *HEATR6* is present on one of the most commonly amplified fragments in breast cancer [62] and, accordingly, its transcript is significantly overexpressed in gastric, brain, and breast carcinomas. Similarly, the *C14ORF21* transcript is upregulated in colorectal, gastric, and prostate cancers (Supplementary Fig. 7).

Some of the genes found in the present study might at first sight not fit the classical picture of tumor suppressors or oncogenes, like those related to ribosomes and ribogenesis (such as *AT2G28450*, *AT3G02220*, and *AT3G49990* genes, orthologues of the human genes *TRMT2A*, *C9ORF85*, and *LTV1*, respectively). Ribosomal proteins are ubiquitous, abundantly present, and mostly regarded as constants in the cells. Approximately 80 proteins have been reported to be part of the ribosomes, and many more are involved in their biogenesis and assembly. However, recent data showed that some of these proteins appear to have extra-ribosomal functions [63], and some are even linked to cancer [64, 65]. The imbalance of ribosomal subunits leads to p53 activation and apoptosis [66]. Additionally, in recent years, drugs that disrupt ribosome production, such as *rapamycin*, have been applied successfully to cancer treatments. As cell division requires the synthesis of a large amount of proteins, deregulation of ribosome biogenesis emerges as a novel strategy to control abnormal cell proliferation, given that without a protein synthesis machinery that can cope with an altered DNA replication process, no division can occur. The best example of this is that inactivation of *SSF1* (orthologue

of *AT5G61770*), involved in ribosome synthesis, leads to loss of contact inhibition [67].

The data presented argue in favor of an applied integrative approach as a powerful strategy to discover new conserved cell cycle regulators. Nevertheless, this strategy suffers from restrictions, especially because it is based on gene coexpression, and thus cannot provide a full perspective of molecular interactions, such as protein–protein interactions, as exemplified by the *Arabidopsis* gene *AT1G49540* and its human orthologue *STAT1P1*. Although the knockdown of the *Arabidopsis* gene triggered cell proliferation, the knock-down of *STAT1P1* did not. In contrast to the plant gene, the coexpression neighborhood of *STAT1P1* is not enriched for DNA replication or DNA repair (Table 3), indicating that despite their orthology relationship both molecules may have diverged functionally during evolution. The contrasting phenotypic effects between these two orthologous genes illustrate that not only the components belonging to a specific network are important but also their wiring.

Conclusions

To understand the origin and progression of the carcinogenic process, and to shed light onto the complex mechanisms that lead to tumorigenesis and cancer, different model organisms have been used. Some of them, like *Mus musculus*, are relatively closely related with humans, and several mouse models are currently used in cancer research [68–70], whereas some others, like *Drosophila melanogaster* or *Saccharomyces cerevisiae*, are distantly related. Nevertheless, they have also contributed extensively to the understanding of the disease [71–74]. With the data presented in this study, we demonstrated that through the use of comparative genomics the plant model species *A. thaliana*, but likely any model organism for which large expression datasets and genome data are available, can aid in the discovery of putative cancer genes.

Acknowledgments We thank all members of the cell cycle and oncology groups for fruitful discussions and suggestions, the *Arabidopsis* Biological Research Center for providing the T-DNA insertion lines, and Martine De Cock and Lorena López for help in preparing the manuscript. This work was supported by grants from the Interuniversity Poles of Attraction Programme (IUAP VI/33), initiated by the Belgian State, Science Policy Office, the Research Foundation-Flanders (grant no. G008306), Ghent University (“Geconcerteerde Onderzoeksaacties” no.01G013B7), the Stichting tegen Kanker (no. 189-2008), the Association for International Cancer Research (Scotland), the EU-FP6 framework program BRECOSM LSHC-CT-2004-503224, and the EU-FP7 framework program TuMIC 2008-201662. M.Q. is indebted with the VIB international PhD program. K.V. acknowledges the support by Ghent University (Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks”) and the Interuniversity Attraction Poles Programme (IUAP P6/25), initiated by the Belgian

State, Science Policy Office (BioMaGNet). S.D. is indebted to the Agency for Innovation through Science and Technology for a predoctoral fellowship.

References

- Morgan DO (1997) Cyclin-dependent kinases: engines, clocks, and microprocessors. *Annu Rev Cell Dev Biol* 13:261–291
- Inze D, De Veylder L (2006) Cell cycle regulation in plant development. *Annu Rev Genet* 40:77–105
- Srinivas PR, Verma M, Zhao Y, Srivastava S (2002) Proteomics for cancer biomarker discovery. *Clin Chem* 48:1160–1169
- Pekarsky Y, Zanasi N, Palamarchuk A, Huebner K, Croce CM (2002) *FHIT*: from gene discovery to cancer treatment and prevention. *Lancet Oncol* 3:748–754
- Jones PA, Laird PW (1999) Cancer epigenetics comes of age. *Nat Genet* 21:163–167
- Marone M, Scambia G, Giannitelli C, Ferrandina G, Masciullo V, Bellacosa A, Benedetti-Panici P, Mancuso S (1998) Analysis of cyclin E and cdk2 in ovarian cancer: gene amplification and RNA overexpression. *Int J Cancer* 75:34–39
- Scheurle D, DeYoung MP, Binninger DM, Page H, Jahanzeb M, Narayanan R (2000) Cancer gene discovery using digital differential display. *Cancer Res* 60:4037–4043
- Argani P, Rosty C, Reiter RE, Wilentz RE, Murugesan SR, Leach SD, Ryu B, Skinner HG, Goggins M, Jaffee EM, Yeo CJ, Cameron JL, Kern SE, Hruban RH (2001) Discovery of new markers of cancer through serial analysis of gene expression: prostate stem cell antigen is overexpressed in pancreatic adenocarcinoma. *Cancer Res* 61:4320–4324
- Alizadeh AA, Ross DT, Perou CM, van de Rijn M (2001) Towards a novel classification of human malignancies based on gene expression patterns. *J Pathol* 195:41–52
- Korkola JE, DeVries S, Fridlyand J, Hwang ES, Estep ALH, Chen Y-Y, Chew KL, Dairkee SH, Jensen RM, Waldman FM (2003) Differentiation of lobular versus ductal breast carcinomas by expression microarray analysis. *Cancer Res* 63:7167–7175
- Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci USA* 101:9309–9314
- Miller LD, Liu ET (2007) Expression genomics in breast cancer research: microarrays at the crossroads of biology and medicine. *Breast Cancer Res* 9:206
- Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe J-P, Tong F, Speed T, Spellman PT, DeVries S, Lapuk A, Wang NJ, Kuo W-L, Stilwell JL, Pinkel D, Albertson DG, Waldman FM, McCormick F, Dickson RB, Johnson MD, Lippman M, Ethier S, Gazdar A, Gray JW (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10:515–527
- Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z, Welch R, Staats BJ, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Gelmann EP, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover R, Hunter DJ, Chanock SJ, Thomas G (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 39:645–649
- Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, Struwing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, SEARCH Collaborators, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen C-Y, Wu P-E, Wang H-C, Eccles D, Evans DG, Peto J, Fletcher O, Johnson N, Seal S, Stratton MR, Rahman N, Chenevix-Trench G, Bojesen SE, Nordestgaard BG, Axelsson CK, Garcia-Closas M, Brinton L, Chanock S, Lissowska J, Peplonska B, Nevanlinna H, Fagerholm R, Eerola H, Kang D, Yoo K-Y, Noh D-Y, Ahn S-H, Hunter DJ, Hankinson SE, Cox DG, Hall P, Wedren S, Liu J, Low Y-L, Bogdanova N, Schürmann P, Dörk T, Tollenaar RAEM, Jacobi CE, Devilee P, Klijn JGM, Sigurdson AJ, Doody MM, Alexander BH, Zhang J, Cox A, Brock IW, MacPherson G, Reed MWR, Couch FJ, Goode EL, Olson JE, Meijers-Heijboer H, van den Ouweland A, Uitterlinden A, Rivadeneira F, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Hopper JL, McCredie M, Southey M, Giles GG, Schroen C, Justenhoven C, Brauch H, Hamann U, Ko Y-D, Spurdle AB, Beesley J, Chen X, kConFab, AOCs Management Group, Mannermaa A, Kosma V-M, Kataja V, Hartikainen J, Day NE, Cox DR, Ponder BAJ (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447:1087–1093
- Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, Sullivan K, Matakidou A, Wang Y, Mills G, Doheny K, Tsai Y-Y, Chen WV, Shete S, Spitz MR, Houlston RS (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* 40:616–622
- Jemal A, Siegel R, Xu J, Ward E (2010) Cancer statistics, 2010. *CA Cancer J Clin* 60:277–300
- Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302:249–255
- Ala U, Piro RM, Grassi E, Damasco C, Silengo L, Oti M, Provero P, Di Cunto F (2008) Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS Comput Biol* 4:e1000043
- McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM (2010) Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci USA* 107:6544–6549
- Jones JDG, Dangl JL (2006) The plant immune system. *Nature* 444:323–329
- Thresher RJ, Vitaterna MH, Miyamoto Y, Kazantsev A, Hsu DS, Petit C, Selby CP, Dawut L, Smithies O, Takahashi JS, Sancar A (1998) Role of mouse cryptochrome blue-light photoreceptor in circadian photoresponses. *Science* 282:1490–1494
- Chan SW-L, Henderson IR, Jacobsen SE (2005) Gardening the genome: DNA methylation in *Arabidopsis thaliana*. *Nat Rev Genet* 6, 351–360 (Err. *Nat Rev Genet* 6, 590)
- Matzke MA, Matzke AJM, Pruss GJ, Vance VB (2001) RNA-based silencing strategies in plants. *Curr Opin Genet Dev* 11:221–227
- Ma H (1994) GTP-binding proteins in plants: new members of an old family. *Plant Mol Biol* 26:1611–1636
- Jones AM, Chory J, Dangl JL, Estelle M, Jacobsen SE, Meyerowitz EM, Nordborg M, Weigel D (2008) The impact of *Arabidopsis* on human health: diversifying our portfolio. *Cell* 133:939–943
- Evan GI, Vousden KH (2001) Proliferation, cell cycle and apoptosis in cancer. *Nature* 411:342–348
- Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* 144:646–674
- Goodarzi H, Elemento O, Tavazoie S (2009) Revealing global regulatory perturbations across human cancers. *Mol Cell* 36:900–911

30. Sherr CJ, McCormick F (2002) The RB and p53 pathways in cancer. *Cancer Cell* 2:103–112
31. Nevins JR (2001) The Rb/E2F pathway and cancer. *Hum Mol Genet* 10:699–703
32. Chen H-Z, Tsai S-Y, Leone G (2009) Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat Rev Cancer* 9:785–797
33. Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* 443:594–597
34. Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res* 32:D575–D577
35. Vandepoele K, Quimbaya M, Casneuf T, De Veylder L, Van de Peer Y (2009) Unraveling transcriptional control in *Arabidopsis* using cis-regulatory elements and coexpression networks. *Plant Physiol* 150:535–546
36. Poole RL (2007) The TAIR database. *Methods Mol Biol* 406:179–212
37. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, Hub AmiGO, Group Web Presence Working (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics* 25:288–289
38. Vandepoele K, Vlieghe K, Florquin K, Hennig L, Beemster GTS, Gruijssem W, Van de Peer Y, Inzé D, De Veylder L (2005) Genome-wide identification of potential plant E2F target genes. *Plant Physiol* 139:316–328
39. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouzé P, Moreau Y (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol* 9:447–464
40. Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 34:D363–D368
41. Edgar RC (2004) MUSCLE: a multiple sequence alignment with reduced time and space complexity. *BMC Bioinformatics* 5:113
42. Van de Peer Y, De Wachter R (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput Appl Biosci* 10:569–570
43. McCall MN, Bolstad BM, Irizarry RA (2010) Frozen robust multivariate analysis (fRMA). *Biostatistics* 11:242–253
44. De Veylder L, Beeckman T, Beemster GTS, de Almeida Engler J, Ormenese S, Maes S, Naudts M, Van Der Schueren E, Jacqmard A, Engler G, Inzé D (2002) Control of proliferation, endoreduplication and differentiation by the *Arabidopsis* E2Fa-DPa transcription factor. *EMBO J* 21:1360–1368
45. Boudolf V, Vlieghe K, Beemster GTS, Magyar Z, Torres Acosta JA, Maes S, Van Der Schueren E, Inzé D, De Veylder L (2004) The plant-specific cyclin-dependent kinase CDKB1;1 and transcription factor E2Fa-DPa control the balance of mitotically dividing and endoreduplicating cells in *Arabidopsis*. *Plant Cell* 16:2683–2692
46. Vlieghe K, Boudolf V, Beemster GTS, Maes S, Magyar Z, Atanassova A, de Almeida Engler J, De Groodt R, Inzé D, De Veylder L (2005) The DP-E2F-like *DELI* gene controls the endocycle in *Arabidopsis thaliana*. *Curr Biol* 15:59–63
47. Beemster GTS, De Veylder L, Vercruysse S, West G, Rombaut D, Van Hummelen P, Galichet A, Gruijssem W, Inzé D, Vuylsteke M (2005) Genome-wide analysis of gene expression profiles associated with cell cycle transitions in growing organs of *Arabidopsis*. *Plant Physiol* 138:734–743
48. Tsukaya H, Beemster GTS (2006) Genetics, cell cycle and cell expansion in organogenesis in plants. *J Plant Res* 119:1–4
49. Cory AH, Owen TC, Bartrop JA, Cory JG (1991) Use of an aqueous soluble tetrazolium/formazan assay for cell growth assays in culture. *Cancer Commun* 3:207–212
50. Cao AR, Rabinovich R, Xu M, Xu X, Jin VX, Farnham PJ (2011) Genome-wide analysis of transcription factor E2F1 mutant proteins reveals that N- and C-terminal protein interaction domains do not participate in targeting E2F1 to the human genome. *J Biol Chem* 286:11985–11996
51. Puente XS, Velasco G, Gutierrez-Fernandez A, Bertranpetit J, King MC, Lopez-Otin C (2006) Comparative analysis of cancer genes in the human and chimpanzee genomes. *BMC Genomics* 7:15
52. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JKV, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* 314:268–274
53. Chagnon P, Michaud J, Mitchell G, Mercier J, Marion J-F, Drouin E, Rasquin-Weber A, Hudson TJ, Richter A (2002) A missense mutation (R565 W) in *Cirrhin* (FLJ14728) in North American Indian childhood cirrhosis. *Am J Hum Genet* 71:1443–1449
54. Yu B, Mitchell GA, Richter A (2009) Cirrhin up-regulates a canonical NF- κ B element through strong interaction with CiriP/HIVEP1. *Exp Cell Res* 315:3086–3098
55. Pikarsky E, Porat RM, Stein I, Abramovitch R, Amit S, Kasem S, Gutkovich-Pyest E, Urieli-Shoval S, Galun E, Ben-Neriah Y (2004) NF-KappaB functions as a tumour promoter in inflammation-associated cancer. *Nature* 431:461–466
56. Huber MA, Azoitei N, Baumann B, Grünert S, Sommer A, Pehamberger H, Kraut N, Beug H, Wirth T (2004) NF- κ B is essential for epithelial-mesenchymal transition and metastasis in a model of breast cancer progression. *J Clin Invest* 114:569–581
57. Welburn JPI, Grishchuk EL, Backer CB, Wilson-Kubalek EM, Yates JR III, Cheeseman IM (2009) The human kinetochore Skl complex facilitates microtubule depolymerization-coupled motility. *Dev Cell* 16:374–385
58. Reidt W, Wurz R, Wanick K, Chu HH, Puchta H (2006) A homologue of the breast cancer-associated gene BARD1 is involved in DNA repair in plants. *EMBO J* 25:4326–4337
59. Takahashi N, Lammens T, Boudolf V, Maes S, Yoshizumi T, De Jaeger G, Witters E, Inzé D, De Veylder L (2008) The DNA replication checkpoint aids survival of plants deficient in the novel replisome factor ETG1. *EMBO J* 27:1840–1851
60. Takahashi N, Quimbaya M, Schubert V, Lammens T, Vandepoele K, Schubert I, Matsui M, Inzé D, Bex G, De Veylder L (2010) The MCM-binding protein ETG1 aids sister chromatid cohesion required for postreplicative homologous recombination repair. *PLoS Genet* 6:e1000817
61. Nishiyama A, Frappier L, Méchali M (2011) MCM-BP regulates unloading of the MCM2–7 helicase in late S phase. *Genes Dev* 25:165–175
62. Wu G-j, Sinclair C, Hinson S, Ingle JN, Roche PC, Couch FJ (2001) Structural analysis of the 17q22–23 amplicon identifies several independent targets of amplification in breast cancer cell lines and tumors. *Cancer Res* 61:4951–4955
63. Lai M-D, Xu J (2007) Ribosomal proteins and colorectal cancer. *Curr Genomics* 8:43–49
64. Macias E, Jin A, Deisenroth C, Bhat K, Mao H, Lindström MS, Zhang Y (2010) An ARF-independent c-MYC-activated tumor suppression pathway mediated by ribosomal protein-Mdm2 interaction. *Cancer Cell* 18:231–243
65. Leontieva OV, Ionov Y (2009) RNA-binding motif protein 35A is a novel tumor suppressor for colorectal cancer. *Cell Cycle* 8:490–497

66. Warner JR, McIntosh KB (2009) How common are extraribosomal functions of ribosomal proteins? *Mol Cell* 34:3–11
67. Welch PM, Gabal M, Betts DM, Whelan NC, Studer ME (2000) In vitro analysis of antiangiogenic activity of fungi isolated from clinical cases of equine keratomycosis. *Vet Ophthalmol* 3:145–151
68. White DE, Kurpios NA, Zuo D, Hassell JA, Blaess S, Mueller U, Muller WJ (2004) Targeted disruption of $\beta 1$ -integrin in a transgenic mouse model of human breast cancer reveals an essential role in mammary tumor induction. *Cancer Cell* 6:159–170
69. Pearson T, Greiner DL, Shultz LD (2008) Humanized SCID mouse models for biomedical research. *Curr Top Microbiol Immunol* 324:25–51
70. Vucur M, Roderburg C, Bettermann K, Tacke F, Heikenwalder M, Trautwein C, Luedde T (2010) Mouse models of hepatocarcinogenesis: What can we learn for the prevention of human hepatocellular carcinoma? *Oncotarget* 1:373–378
71. Hartwell LH (1992) Role of yeast in cancer research. *Cancer* 69:2615–2621
72. Rosengard AM, Krutzsch HC, Shearn A, Biggs JR, Barker E, Margulies IMK, King CR, Liotta LA, Steeg PS (1989) Reduced Nm23/Awd protein in tumour metastasis and aberrant *Drosophila* development. *Nature* 342:177–180
73. Moberg KH, Bell DW, Wahrer DCR, Haber DA, Hariharan IK (2001) Archipelago regulates Cyclin E levels in *Drosophila* and is mutated in human cancer cell lines. *Nature* 413:311–316
74. Caussinus E, Gonzalez C (2005) Induction of tumor growth by altered stem-cell asymmetric division in *Drosophila melanogaster*. *Nat Genet* 37:1125–1129