Expanding the molecular toolbox in diatoms: developing a transformation system, CRISPR-Cas and Inverse Yeast-1-hybrid

Amanda Hopes

A thesis submitted for the degree of Doctor of Philosophy

University of East Anglia, Norwich, UK

School of Environmental Sciences

July 2017

© This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived there from must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

Abstract

Diatoms are single celled microalgae with intricately patterned silica cell walls. This cosmopolitan group is a dominant primary producer with many species playing key roles in marine, estuarine and freshwater habitats. Furthermore, due to their silica frustule, lipid production and a range of other chemical and physiological adaptations, diatoms have high potential for biotechnology. Despite their diversity and ecological relevance, molecular tools for diatoms are often underrepresented and limited to a small number of species. This PhD expands the molecular toolbox for two key species: *Thalassiosira pseudonana*, a model, centric, temperate diatom with a heavily silicified frustule and *Fragilariopsis cylindrus*, a key, pennate diatom in marine psychrophilic waters and sea-ice.

A transformation system has been developed in *F. cylindrus* leading to the expression of egfp and shble transgenes under the control of an endogenous FCP promoter. This method has been applied to understanding the role of the SITMyb gene, a potential transcription factor with links to silica metabolism, by overexpression. In-silico and in-vitro modelling of the SITMyb gene has been performed and preliminary development of an inverse yeast-1-hybrid system, to elucidate potential transcription factor binding sites, has been carried out. *F. cylindrus* is the first genetically tractable polar microalgae and appears to be the first psychrophilic eukaryote to be transformed.

CRISPR-Cas is a targeted genome editing tool, fast becoming an essential method in any molecular toolbox. This thesis demonstrates development in *T. pseudonana* by successfully editing the urease gene through a programmed deletion using two sgRNAs. As a model diatom, several molecular tools are already available for *T. pseudonana*, however this is the first time a targeted knock-out has been achieved in this species. In addition Golden-Gate cloning has been used to produce the construct, giving this method a large degree of flexibility and future potential for multiplexing.

Contents

List of tables

Tables are listed numerically. When tables are present in a publication this is stated and published tables retain their original numbering system.

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of figures

Figures are listed numerically. When figures are present in a publication this is stated and published figures retain their original numbering system.

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Preface

This statement confirms that the work contained in this thesis was conceived, planned, conducted, interpreted and written by Amanda Hopes. Prof. Thomas Mock, my primary supervisor, was involved throughout all stages of this PhD, including reviewing the five chapters contained within this thesis. Involvement of other members of the Mock lab and collaborators is outlined below.

Chapter 1 introduces the topic of model diatoms and diatoms in a polar systems. It explains the need for molecular tools within the diatom community. Three publications, explained below, are included which give a broad overview of diatom biology, adaptations, molecular tools in diatoms and diatoms within a psychrophilic environment. In all three I have first authorship.

Diatoms: glass dwelling dynamos was published in 2014 in Microbiology Today. I wrote the article and created the second figure. Thomas Mock edited the article and provided the remaining figures.

Evolution of Microalgae and Their Adaptations in Different Marine Ecosystems was published in 2015 in ELS. I wrote the article which was later edited by Thomas Mock. I produced figure three. Figure four was jointly produced by Thomas Mock and myself. A few corrections have been made since publication for this thesis.

Polar Microalgae: Functional Genomics, Physiology and the Environment is a book chapter published in Psychrophiles: From Biodiversity to Biotechnology in 2017. The section 'adaptation of microalgae at high latitudes' has been included in the introduction of this thesis. The book chapter was originally written by Thomas Mock and David. N. Thomas in 2008. I updated the included section for the 2017 edition. A few corrections have been made since publication for this thesis.

Chapter 2 describes the development of a transformation system in *Fragilariopsis cylindrus*. RNA-sequencing data used to establish genes with high expression levels in *F. cylindrus* was provided by Jan Strauss during his post-doc in Thomas Mock's lab. Jan also conducted preliminary tests to establish the concentration of zeocin needed to inhibit growth of *F. cylindrus* in liquid media.

Chapter 3 details the development of a gene editing system using CRISPR Cas in *Thalassiosira pseudonana*. Vladimir Nekrasov gave advice on Golden-Gate cloning and the band shift-assay method. He also provided the domesticated L0 Cas9:YFP module. Golden-gate vector backbones were provided by Vladimir Nekrasov and Oleg Raitskin from the repository at the Sainsbury Laboratory.

Lewis Dunham, a Masters student in Thomas Mock's lab, performed the bench work for TSO RACE to elucidate the end of the U6 promoter in *T. pseudonana* under my supervision. Gene specific primers were designed by me, as was the experiment, using the method by Pinto and Lindblad (2010). I also carried out the analysis.

Chapter 4 examines the function of the SITMyb gene in *F. cylindrus*. Nigel Belshaw carried out the gel electrophoresis, membrane transfer and antibody labelling steps for western blots performed on HA-tagged SITMyb proteins. For the SITMyb CRISPR-Cas construct, Irina Grouneva made the construct using my FCP:shble module and my FCP promoter/terminator sequences, developed during my transformation chapter (chapter 2). I also designed the sgRNAs for this construct. I elucidated the U6 promoter in *F. cylindrus* in-silico and Nigel Belshaw confirmed this empirically using the TSO RACE method, used in chapter 3. Both Nigel and Irina work in the Mock lab.

Acknowledgements

I would like to thank my primary supervisor Thomas Mock for his constant support and advice. Thank you for the pushing me and encouraging me to explore new opportunities. I wouldn't have achieved this much without your enthusiasm. I would also like to thank my secondary supervisor Richard Bowater for his support and encouragement.

Thanks to Vladimir Nekrasov and Sophien Kamoun at the Sainsbury lab for collaborating on the CRISPR-Cas project. I'd particularly like to thank Vlad for the opportunity to help co-organise a workshop on CRISPR-Cas in plants and diatoms. This was an invaluable experience and gave me my first chance to speak at a conference.

The Natural Environment Research Council UK (NERC) provided the funding which made this PhD possible. NERC also funded two courses to help improve my bioinformatics and statistical skills.

Thanks to the Microbiology Society, the British Phycological Society (BPS) and the Molecular Life of Diatoms (MLD) conference organisers in Kobe for providing me with travel funds which allowed me to attend several excellent conferences.

My life would have been a lot more difficult these last three and half years without the support of the technical lab staff at UEA. In particular Rob Utting was an absolute star and provided a friendly face as well as first rate technical help.

I would also like to express my gratitude to Shawn Burgess for providing the pYOH1 and pYOH366 plasmids for yeast-1-hybrid.

I'd also like to thank the other members of the Mock lab, both past and present. Thanks to Nigel for his technical support and advice, particularly when setting up yeast culturing. There's only so many times you can ask someone if they think a particular yeast culture is white or slightly pink before they start to lose patience with you. I'd like to thank Jan Strauss, Amy Kirkham, Krisztina Sarkozi and Katrin Schmidt for welcoming me into the lab and helping me settle in. Thanks for all the interesting lab conversations and providing a friendly ear when necessary!

Special thanks to my friends and family - this would have been a lot harder without your emotional support. Especially my parents Elaine and Graham who are always there on the end of the phone when I need someone to talk to.

Last but certainly not least, a huge thank you to my wonderful partner Tony. Thank you for the love and mental support, the encouragement and for putting up with me throughout - particularly whilst writing my thesis. You've kept me housed, clothed, fed and sane these last few months.

Chapter 1: Introduction

Introduction to *Thalassiosira pseudonana* and *Fragilariopsis cylindrus*

The published reviews within this introduction give a broad overview of diatom biology, adaptations and molecular advances in diatoms. This thesis focuses on two key, ecologically important diatoms: *Thalassiosira pseudonana* and *Fragilariopsis cylindrus*. The first is a cosmopolitan, centric diatom found in temperate, freshwater, brackish and marine coastal environments (Guiry & Guiry 2017), whilst *F. cylindrus* is psychrophilic, raphid, pennate diatom and a dominant photoautotroph in marine polar waters and sea ice (Cefarelli et al. 2010).

As higher plants are rarely found in the polar environment, the associated ecosystems are reliant on primary production from algae and cyanobacteria (Lizotte 2001). Diatoms are often dominant in these systems, especially species from the *Fragilariopsis* genus, including *F. cylindrus* (Leventer 1998; Lizotte 2001; Cefarelli et al. 2010; Kang & Fryxell 1992). As a result they are responsible for large amounts of carbon fixation and play a pivotal role in the Arctic and Antarctic food webs (Mock & Thomas 2008).

Both diatoms are model organisms, and have been subject to prior molecular analysis. *T. pseudonana* in particular, was the first diatom with a sequenced genome (Armbrust et al. 2004) and a transformation for this species has been available since 2006 (Poulsen et al. 2006). Molecular analysis such as genome sequencing and expression analysis including microarrays, expressed sequence tags (ESTs) and RNA sequencing (Hook & Osborn 2012; Ashworth et al. 2016; Smith et al. 2016; Mock et al. 2017; Bowler et al. 2008), provide a wealth of a data for environmental, evolutionary and physiological understanding. Molecular tools such as overexpression (Yao et al. 2014; Matthijs et al. 2017; Cook & Hildebrand 2015), RNA-silencing (De Riso et al. 2009; Kirkham et al. 2017) and gene knock-out (Weyman et al. 2015; Daboussi et al. 2014; Nymark et al. 2016) can then be used to answer specific biological questions through reverse genetics.

Due to its intricate and heavily silicified silica frustule, as well as availability of molecular tools, *T. pseudonana* has been one of the main species used to study formation and molecular basis of silicification in diatoms (Poulsen et al. 2013; Kröger et al. 1999; Scheffel et al. 2011; Tesson & Hildebrand 2010; Shrestha & Hildebrand 2015). It has also been the focus of several biotechnology applications (Delalat et al. 2015; Sheppard et al. 2012; Cook & Hildebrand 2015). Although RNA silencing is available for this species (Kirkham et al. 2017), gene knockout is an important tool for reverse genetics and can entirely eliminate protein production of a gene for downstream phenotyping, functional analysis and physiological modification. To this end, CRISPR-Cas has been developed for this species (chapter 3), to provide a cheap, quick, adaptable method for genome editing and knock-out.

In comparison to *T. pseudonana* and especially the model pennate diatom *Pheodactylum tricornutum*, molecular tools and analysis in *F. cylindrus* are underrepresented, despite its large ecological significance. Recently the genome has been published for *F. cylindrus* along with RNA sequencing data under various different conditions linked to the polar environment (Mock et al. 2017). In additional several expressed sequence tag (EST) libraries under cold and salt shock (Mock et al. 2005; Krell 2006) have been produced to give insight into physiological adaptations in this environment. Details of molecular analyses in polar diatoms and other psychrophilic microalgae can be found in the 'adaptation of microalgae at high latitudes' section of the book chapter 'Polar Microalgae: Functional Genomics, Physiology and the Environment', found within this introduction.

A transformation system is a key tool for reverse genetics and has been previously developed in several diatom species, detailed in the transformation chapter (chapter 2), to employ methods such as overexpression (Cook & Hildebrand 2015; Matthijs et al. 2017), protein tagging (Apt et al. 2002; Joshi-Deo et al. 2010), gene silencing (De Riso et al. 2009; Kirkham et al. 2017) and protein localisation (Siaut et al. 2007). The first data chapter of this thesis details the development of a transformation system for *F. cylindrus*, which is later used to overexpress a transcript for a potential endogenous transcription factor with possible links to silica metabolism (chapter 4: SITMyb). This chapter also details in-silico and in-vitro analysis to characterise the SITMyb gene as well as preliminary development of yeast-1-hybrid for *F. cylindrus* to elucidate transcription factor binding sites.

Development of methods such as CRISPR-Cas, transformation and yeast-1-hybrid provide a strong contribution to the growing but, still underrepresented diatom molecular toolbox. This will hopefully help to further illuminate our molecular, physiological, environmental and evolutionary understanding of diatoms. These techniques also have a place in biotechnology, with the potential for genome editing to be used to enhance production of oils (Daboussi et al. 2014), for medical purposes (Delalat et al. 2015; Hempel et al. 2011) and to modify the silica frustule for nanotechnology applications (Dolatabadi & de la Guardia 2011; Wang et al. 2013; Jeffryes et al. 2011). Furthermore, the development of a transformation system in *F. cylindrus* appears to be the first example of genetic transformation in a polar eukaryotic species. Although several psychrophilic bacteria can be transformed, eukaryotic proteins are not always correctly processed in a prokaryotic host (Demain & Vaishnav 2009), therefore transformation in *F. cylindrus* has the additional potential for producing recombinant proteins, particularly given that growth at lower temperatures can lead to higher yields, correct folding and higher solubility in some proteins (Vasina & Baneyx 1996; San-Miguel et al. 2013).

References

Apt, K.E. et al., 2002. In vivo characterization of diatom multipartite plastid targeting signals. *Journal of cell science*, 115(Pt 21), pp.4061–4069.

Armbrust, E. V. et al., 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science*, 306(5693), pp.79–86.

Ashworth, J. et al., 2016. Pan-transcriptomic analysis identifies coordinated and orthologous functional modules in the diatoms Thalassiosira pseudonana and Phaeodactylum tricornutum. *Marine Genomics*, 26, pp.21–28. Available at: http://dx.doi.org/10.1016/j.margen.2015.10.011.

Bowler, C. et al., 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, 456(7219), pp.239–244.

Cefarelli, A.O. et al., 2010. Diversity of the diatom genus Fragilariopsis in the Argentine Sea and Antarctic waters: Morphology, distribution and abundance. *Polar Biology*, 33(11), pp.1463–1484.

Cook, O. & Hildebrand, M., 2015. Enhancing LC-PUFA production in Thalassiosira pseudonana by overexpressing the endogenous fatty acid elongase genes. *Journal of Applied Phycology*, 28(2), pp.897–905. Available at: http://link.springer.com/10.1007/s10811-015-0617-2.

Daboussi, F. et al., 2014. Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology. *Nature communications*, 5, p.3831. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24871200.

Delalat, B. et al., 2015. Targeted drug delivery using genetically engineered diatom biosilica. *Nature Communications*, 6, p.8791. Available at: http://www.nature.com/doifinder/10.1038/ncomms9791.

Demain, A.L. & Vaishnav, P., 2009. Production of recombinant proteins by microbes and higher organisms. *Biotechnology Advances*, 27(3), pp.297–306.

Dolatabadi, J.E.N. & de la Guardia, M., 2011. Applications of diatoms and silica nanotechnology in biosensing, drug and gene delivery, and formation of complex metal nanostructures. *TrAC - Trends in Analytical Chemistry*, 30(9), pp.1538–1548. Available at: http://dx.doi.org/10.1016/j.trac.2011.04.015.

Guiry, M.D. & Guiry, G.., 2017. *AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. http://www.algaebase.org; searched November 2016.*

Hempel, F. et al., 2011. Algae as Protein Factories: Expression of a Human Antibody and the Respective Antigen in the Diatom *Phaeodactylum tricornutum*. *PLoS ONE*. 6(12), e28424.

Hook, S.E. & Osborn, H.L., 2012. Comparison of toxicity and transcriptomic profiles in a diatom exposed to oil, dispersants, dispersed oil. *Aquatic Toxicology*, 124-125, pp.139–151. Available at: http://dx.doi.org/10.1016/j.aquatox.2012.08.005.

Jeffryes, C. et al., 2011. The potential of diatom nanobiotechnology for applications in solar cells, batteries, and electroluminescent devices. *Energy & Environmental Science*, 4(10), p.3930. Available at: http://xlink.rsc.org/?DOI=c0ee00306a.

Joshi-Deo, J. et al., 2010. Characterization of a trimeric light-harvesting complex in the diatom Phaeodactylum tricornutum built of FcpA and FcpE proteins. *Journal of experimental botany*, 61(11), pp.3079–3087.

Kang, S.H. & Fryxell, G. a., 1992. Fragilariopsis cylindrus (Grunow) Krieger: The most abundant diatom in water column assemblages of Antarctic marginal ice-edge zones. *Polar Biology*, 12(6-7), pp.609–627.

Kirkham, A. et al., 2017. A role for the cell-wall protein silacidin in cell size of the diatom Thalassiosira pseudonana. *ISME*.

Krell, A., 2006. *Salt stress tolerance in the psychrophilic diatom Fragilariopsis cylindrus.* University of Bremen, Germany.

Kröger, N., Deutzmann, R. & Sumper, M., 1999. Polycationic Peptides from Diatom Biosilica That Direct Silica Nanosphere Formation. *Science*, 286(5442), pp.1129–1132. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.286.5442.1129.

Leventer, A., 1998. The fate of Antarctic "Sea ice diatoms" and their use as paleoenvironmental indicators. *Antarctic Research Series*, 73, pp.121–137.

Lizotte, M.P., 2001. The Contributions of Sea Ice Algae to Antarctic Marine Primary Production. *American Zoologist*, 41(1), pp.57–73. Available at: http://az.oxfordjournals.org/content/amzoo/41/1/57.full.pdf.

Di Martino Rigano, V. et al., 2006. Temperature dependence of nitrate reductase in the psychrophilic unicellular alga Koliella antarctica and the mesophilic alga Chlorella sorokiniana. *Plant, Cell and Environment*, 29(7), pp.1400–1409.

Matthijs, M. et al., 2017. The transcription factor bZIP14 regulates the TCA cycle in the diatom Phaeodactylum tricornutum. *EMBO Journal*, 36(11), pp.1559–1576.

Mock, T. et al., 2005. Analysis of Expressed Sequence Tags (ESTs) from the Polar Diatom Fragilariopsis cylindrus. *Journal of Phycology*, 42, pp.78–85.

Mock, T. et al., 2017. Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. *Nature*, 541(7638), pp.536–540.

Mock, T. & Thomas, D.N., 2008. Microalgae in Polar regions: Linking functional genomics and physiology with environmental conditions. , pp.285–312. Available at: http://hdl.handle.net/10242/37672.

Nymark, M. et al., 2016. A CRISPR/Cas9 system adapted for gene editing in marine algae. *Scientific reports*, 6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27108533.

Poulsen, N. et al., 2013. Pentalysine clusters mediate silica targeting of silaffins in Thalassiosira pseudonana. *Journal of Biological Chemistry*, 288(28), pp.20100–20109.

Poulsen, N., Chesley, P.M. & Kröger, N., 2006. Molecular genetic manipulation of the diatom Thalassiosira pseudonana (Bacillariophyceae). *Journal of Phycology*, 42(5), pp.1059–1065.

De Riso, V. et al., 2009. Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic Acids Research*, 37(14).

San-Miguel, T., Perez-Bermudez, P. & Gavidia, I., 2013. Production of soluble eukaryotic recombinant proteins in E. coli is favoured in early log-phase cultures induced at low

temperature. *Springerplus*, 2(1), p.89. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23525091.

Scheffel, A. et al., 2011. Nanopatterned protein microrings from a diatom that direct silica morphogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 108(8), pp.3175–80. Available at: http://www.pnas.org/content/108/8/3175.short.

Sheppard, V.C. et al., 2012. Live diatom silica immobilization of multimeric and redox-active enzymes. *Applied and Environmental Microbiology*, 78(1), pp.211–218.

Shrestha, R.P. & Hildebrand, M., 2015. Evidence for a regulatory role of diatom silicon transporters in cellular silicon responses. *Eukaryotic Cell*, 14(1), p.29.

Siaut, M. et al., 2007. Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene*, 406(1-2), pp.23–35.

Smith, S.R. et al., 2016. Transcript level coordination of carbon pathways during silicon starvation-induced lipid accumulation in the diatom Thalassiosira pseudonana. *New Phytologist*, 210(3), pp.890–904.

Tesson, B. & Hildebrand, M., 2010. Extensive and intimate association of the cytoskeleton with forming silica in diatoms: Control over patterning on the meso- and micro-scale. *PLoS ONE*, 5(12).

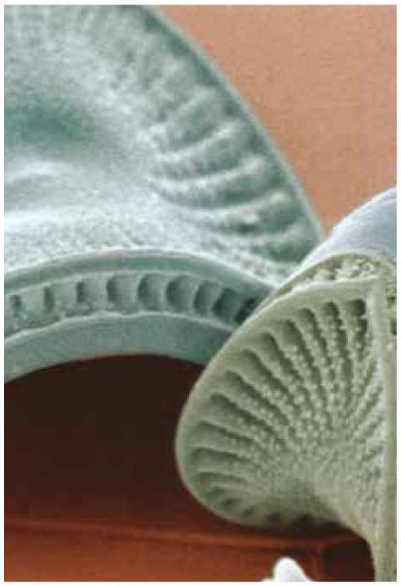
Vasina, J. a & Baneyx, F., 1996. Recombinant protein expression at low temperatures under the transcriptional control of the major Escherichia coli cold shock promoter cspA. *Applied and environmental microbiology*, 62(4), pp.1444–1447.

Wang, Y. et al., 2013. Preparation of biosilica structures from frustules of diatoms and their applications: Current state and perspectives. *Applied Microbiology and Biotechnology*, 97(2), pp.453–460.

Weyman, P.D. et al., 2015. Inactivation of Phaeodactylum tricornutum urease gene using transcription activator-like effector nuclease-based targeted mutagenesis. *Plant Biotechnology Journal*, 13(4), pp.460–470.

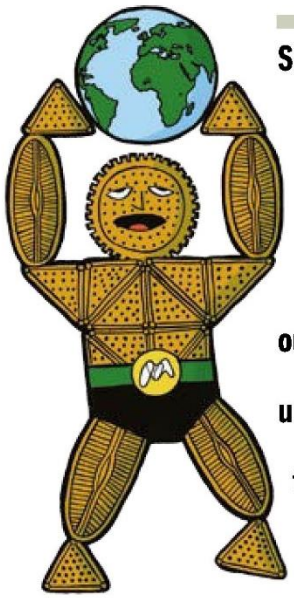
Yao, Y. et al., 2014. Glycerol and neutral lipid production in the oleaginous marine diatom Phaeodactylum tricornutum promoted by overexpression of glycerol-3-phosphate dehydrogenase. *Biotechnology for Biofuels*, 7(1), p.110. Available at: http://www.biotechnologyforbiofuels.com/content/7/1/110.

Diatoms: glass-dwelling dynamos



Diatoms glass-dwelling dynamos

Amanda Hopes & Thomas Mock



Superheroes have a reputation for being larger than life, but it is the unseen micro-organisms that can have a substantial impact on our lives that for most will go unnoticed. One such group are the unicellular algae known as diatoms.

Members of the heterokontophyta, diatoms have both plant- and animal-like characteristics. Most are photosynthetic, and use chlorophylls *a* and *c* to store energy from the sun as lipids or polysaccharides. However, some are obligate or facultative heterotrophs and can live on an external food source either permanently or during extended periods of little or no light. Diatoms are abundant and diverse with an estimated 200,000 extant species spread across almost all aquatic habitats.

One of the most outstanding features of the diatoms is their ability to produce complex, beautiful, silica

Coloured SEM of the diatom *Campylodiscus hibernicus*. Power and Syred/Science Photo Library

frustules that are effectively intricate glass shells. The form and shape of the frustule is species-specific, and with so many diatom species there is a vast array of morphologies with many different shapes, sizes and projections, including spines, ridges and protuberances. The basic form, however, consists of two overlapping valves known as theca that contain pores and are bound together with girdle bands.

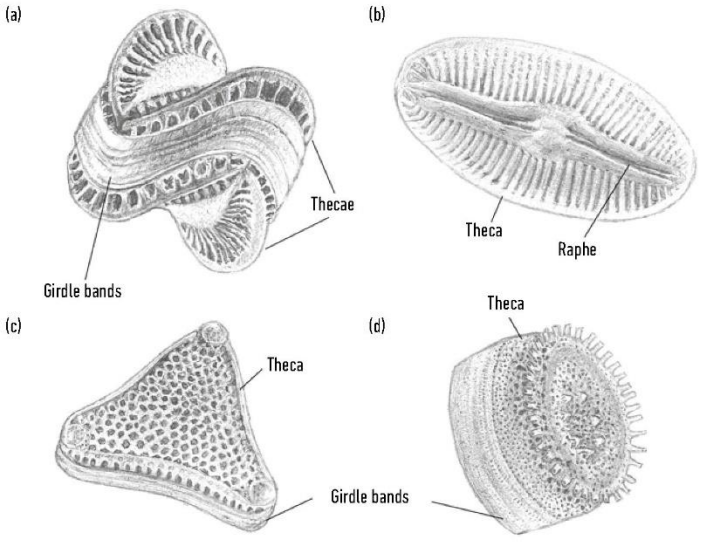
Many aspects of the mechanisms by which diatoms form their frustules remain to be discovered. However, much has been learnt in the last few decades. Diatoms use silicic acid to create their frustules. This soluble form of silica is taken up by silicon transporters into the silica deposition vesicle where it is precipitated. Several molecules have been implicated in the precipitation and structure on a nanoscale: silaffins, silacidins, cingulins and long-chain polyamines. Evidence suggests that structuring of assemblages and the final frustule shape is influenced by actin microfilaments and microtubules of the cytoskeleton.

Historically, it is the shapes formed by these processes that have influenced diatom taxonomy. However, phylogenetics following molecular sequencing has determined that diatoms can be split into two clades. The first contains the centric diatoms that have radial valve symmetry and tend to be circular in shape. The second clade is split into two further groups: the bi-/multipolar centrics and the pennate diatoms. Multipolar centrics can be a variety of shapes, whereas pennate diatoms are elongated with bilateral symmetry. Pennate diatoms can be further broken down into araphid and raphid pennates, the latter of which can move through sediments or over surfaces by passing secretions through a slit (raphe) present in one or both of the valves.

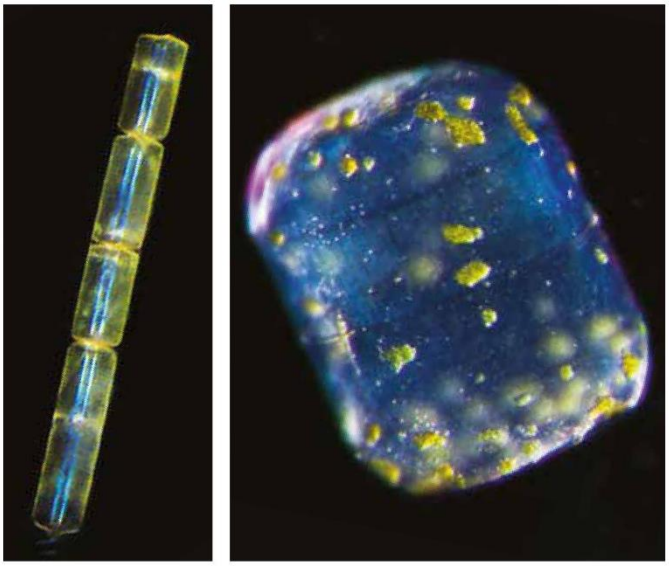
Global importance

Diatoms have a tremendous impact on many global events, which is influenced and connected by different aspects of their physiology. Photosynthesis, biogenic silica formation, environmental diversity and a propensity to dominate phytoplankton communities has led to the major involvement of diatoms in primary production, nutrient cycling and support of organisms further up the food chain.

It is estimated that diatoms contribute 40–45% of oceanic primary productivity, which amounts to 20% of global carbon fixation and oxygen production. Unsurprisingly, given the amount of carbon they fix, diatoms



Diatom structure. (a) Centric diatom, *Campylodiscus* sp.; (b) raphid pennate diatom, *Diploneis* sp.; (c) multipolar centric diatom, *Triceratum* sp.; (d) centric diatom, *Cyclotella* sp. A. Hopes



Melosira sp. (left) and *Lauderia annulata* (right) collected during the Tara Oceans expedition 2009–2012.
Christian Sardet, Jennifer Gillette & Chris Bowler

are also heavily involved in the ocean carbon cycle: this involvement is further substantiated by the presence of the silica frustule. Sinking of organic matter below the photic zone to the ocean floor provides both essential nutrients for organisms living in the ocean depths and exports carbon to the ocean interior. The density of the silica frustules expedites sinking of diatom cells, which combined with their abundance makes them a key player in the biological pump as well as the silica cycle. Carbon dioxide (CO_2) present in the photic zone is fixed by diatoms and sinks to the ocean depths, leading to a loss from surface waters. This in turn is replaced by atmospheric CO_2, thereby maintaining the balance of both CO_2 and global temperature. This contribution towards the carbon cycle is especially evident upon the formation of large diatom blooms. Diatom growth is limited by factors such as nutrient availability; however, when there are large nutrient influxes or seasonal changes, diatoms can form large blooms up to several kilometres long. As nutrients, particularly nitrate and silicate, begin to run out the bloom dies back and aggregates of dense silicified cells sink towards the ocean floor, depositing large amounts of organic matter. Deposits of frustules which can be hundreds of metres thick can result from sinking diatoms and lead to silica-rich sediments known as diatomaceous earth. These deposits have been widely used within industry due to their chemical composition and large structural surface area, properties which have led to perhaps the most famous use of diatomaceous earth: the stabilising component in Alfred Nobel's dynamite.

As primary producers diatoms provide a food source and support for higher levels of the food chain. For some regions, such as the Southern Ocean surrounding the Antarctic continent, they are particularly important as they are able to photosynthesise where many other phytoplankton are not. As a result, diatoms are responsible for feeding the entire Antarctic food web including krill, penguins and whales.

It is evident that this successful group provides several beneficial superhero services to the planet. The question, therefore, is how have these single-celled organisms developed and diversified to fill multiple niches, outcompeting other phytoplankton to take a key position in driving global biological and biogeochemical processes?

Secret of success

The diversity of diatoms in terms of morphology and habitat show that they are highly adaptable and have been able to take advantage of different environments in order to evolve and spread since their origin about 240 million years ago. With genome sequencing of four diatom species, insights into their success have been revealed.

Diatoms have what has been referred to as a 'mix-and-match genome' due to diverse evolutionary origins. These superheroes are the product of secondary endosymbiosis in which a eukaryotic heterotroph (exosymbiont) engulfed a red alga (endosymbiont). Although the resultant plastid lost the majority of its genes over time, several have been incorporated into the nucleus of the host cell evidenced by red algal

It seems that the ability to combine plant, animal and bacterial abilities has led to a highly adaptable group of species with several advantages over other phytoplankton, leading to a dominant primary producer in a beautiful yet practical glass shell.

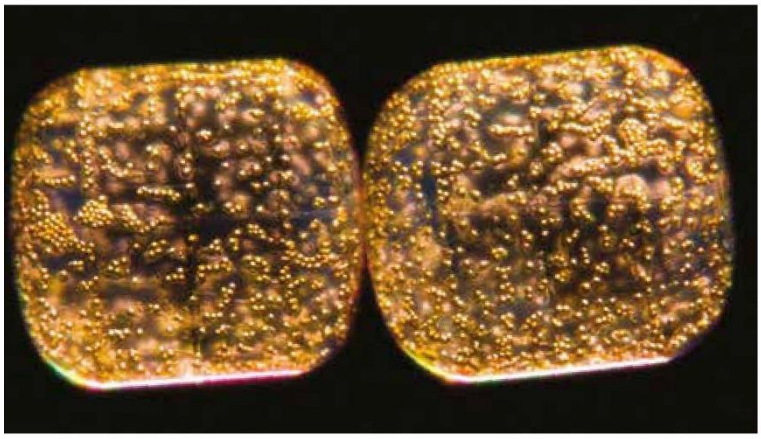
genes observed in the sequenced genomes. Interestingly, green algae and bacterial genes are also present. It has been speculated that secondary endosymbiosis involving a green alga may have also occurred with gene transfer to the host nucleus followed by subsequent plastid loss. Presence of bacterial genes is thought to be due to horizontal gene transfer, probably aided by mutually beneficial relationships documented with bacteria.

Deriving genes from several sources means that diatoms have a potentially advantageous range of abilities that would not normally be found in a single organism. The silica frustule, thought to be inherited from the exosymbiont, may increase fitness through a range of different morphologies; for example, protection against predation, pathogens and desiccation, focusing light into the cell, and nutrient acquisition and storage. The evolution of a more refined frustule is thought to have allowed diatoms to colonise the pelagic oceans and it has been calculated that a silica-based wall is less costly than an organic one.

The spread of diatoms into several niches may be explained by abilities originating from bacteria. For example, some diatoms express proton-pump-like rhodopsins that may be advantageous in areas with low iron availability, while others produce ice-binding proteins, allowing them to live in sub-zero temperatures.

These processes and many others derived from this unique evolutionary background have ensured diatom success. It seems that the ability to combine plant, animal and bacterial abilities has led to a highly adaptable group of species with several advantages over other phytoplankton, leading to a dominant primary producer encased in a beautiful yet practical glass shell.

This practicality, however, may reach beyond the immediate benefit to the diatom itself. The structural and physical properties of the frustule are the focus of several research areas into nanotechnology applications. These include drug delivery, solar technology, microfluidics, catalyst production and bio-sensing. Lipid production in diatoms is also drawing interest as a source of renewable oil. Although the global contributions made by diatoms are already significant, these technologies may be the key to drawing the public eye to the importance of these microscopic algae and the roles they play within our lives.



***Coscinodiscus* sp. collected during the Tara Oceans expedition 2009–2012.** Christian Sardet & Chris Bowler

Amanda Hopes & Thomas Mock
School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK
a.hopes@uea.ac.uk, **t.mock@uea.ac.uk**

Further reading

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

Bowler, C., Vardi, A. & Allen, A. E. (2010). Oceanographic and biogeochemical insights from diatom genomes. *Ann Rev Mar Sci* **2**, 333–365.

Falkowski, P. G. & Raven, J. A. (2007). *Aquatic Photosynthesis: An Introduction to Photosynthesis in Aquatic Systems*. Princeton: Princeton University Press.

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

Mock, T. & Medlin, L. K. (2012). Genomics and genetics of diatoms. *Adv Bot Res* **64**, 245–284.

Smetacek, V. (1999). Diatoms and the ocean carbon cycle. *Protist* **150**, 25–32.

Tesson, B. & Hildebrand, M. (2010). Extensive and intimate association of the cytoskeleton with forming silica in diatoms: control over patterning on the meso- and micro-scale. *PLoS ONE* **5**, 1–13.

Evolution of Microalgae and Their Adaptations in Different Marine Ecosystems

Advanced article

Article Contents

Online posting date: 15th October 2015

Amanda Hopes, *University of East Anglia, Norwich, UK*

Thomas Mock, *University of East Anglia, Norwich, UK*

Microalgae are unicellular eukaryotic organisms that are predominantly photosynthetic. They are found in a wide range of habitats, particularly marine ecosystems, and are responsible for a significant portion of the oceans biogeochemical cycling. Microalgae have a varied evolutionary history with genes derived from photosynthetic organisms, heterotrophic eukaryotes and bacteria. This has led to a wide range of adaptations, allowing them to thrive in a variety of conditions. Microalgae in coastal regions are adapted to turbulence, high nutrients and low light whilst open ocean microalgae have to contend with high irradiance and low nutrient concentrations. Polar microalgae are adapted to freezing temperatures, high nutrients and long periods of light and darkness. Microalgae genomes and transcriptomes uncover and interpret these adaptations, providing information on how microalgae have become a dominant force within marine ecosystems.

Introduction

The term microalgae is used to describe small unicellular eukaryotic algae, which live individually or associate in chains or colonies. Microalgae can be up to 1000 µm in size with the smallest at 0.95 µm (Courties *et al*., 1994) and can be motile, either swimming with a flagella or gliding across surfaces or through sediments via a raphe. They occupy a diverse range of freshwater, marine and terrestrial habitats including extreme environments, such as snow, sea ice (psychrophiles), hot springs (thermophiles) and salt lakes (halophiles) (Rothschild and Mancinelli, 2001). Microalgae can be found in most habitats where water and sufficient light are available, however, the majority are found in the oceans (**Figure 1**) as part of the phytoplankton community, which includes both the eukaryotic microalgae and cyanobacteria. Together they are responsible for just over 45% of global primary production (Field *et al*., 1998) substantially impacting the ocean's carbon fixation, oxygen production, nutrient cycling and food web. The majority of microalgae are photosynthetic, though they can be heterotrophic or mixotrophic, as a result, the majority can be found within the photic zone. A significant proportion of net production comes from temperate and polar frontal zones and in areas where nutrient rich water comes to the surface by upwelling (Field *et al*., 1998).

Microalgae play a large role in nutrient cycling, as they provide nutrients to other organisms through consumption and by sinking of organic matter. Sinking also transports carbon, fixed from atmospheric CO_2 to the oceans interior, and therefore microalgae are an important factor in CO_2 levels and climate change. These processes are especially transparent during algal blooms (**Figure 2**), where an excess of normally limiting nutrients such as nitrate and iron allows rapid and dense growth of often a single or a few species of microalgae (Boyd *et al*., 2000). These blooms have been known to grow to an area the size of the United Kingdom, with some species responsible for these elevated densities producing toxins which are damaging to other organisms. Once the nutrients are depleted, the algal community sinks, cycling organic and inorganic products. Sinking of algal blooms highlights the geological contributions made by diatoms and coccolithophores, which produce intricate shells of silica and calcium carbonate, respectively. Blooms formed from these species result in diatomite and chalk deposits. Oil reserves are also linked to microalgal lipid deposition. **See also**: **Algal Calcification and Silification**

This review paper focuses on the adaptations of microalgae in different important marine ecosystems. Understanding this allows us to determine how these globally important organisms are able to thrive in their current environments and to predict the effect of global change on their biogeography. **See also**: **Biogeography of Marine Algae**

eLS subject area: Plant Science

How to cite:
Hopes, Amanda and Mock, Thomas (October 2015) Evolution of Microalgae and Their Adaptations in Different Marine Ecosystems. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0023744

eLS © 2015, John Wiley & Sons, Ltd. www.els.net

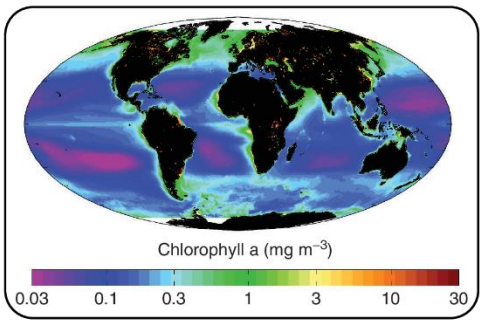


Figure 1 Average sea surface chlorophyll *a* concentration from 1998 to 2006. Chlorophyll *a* is used as an indicator of phytoplankton biomass.

Figure 2 *Emiliania huxleyi* bloom off the coast of South West England.

The Evolution of Microalgae

There are approximately 30 000 described species of phytoplankton and around 90% of these are eukaryotic (**Table 1**). In terms of species, diatoms account for the majority (>50%) of microalgae (Guiry and Guiry, 2015), however, numbers are estimated to be much higher than currently described (Guiry, 2012). The key to microalgae success is their evolution, which has enabled them to adapt to a myriad of different environments. In addition to the four major forces of evolution (mutation, selection, genetic drift and gene flow), evolution has been shaped by three additional evolutionary processes: endosymbiosis (**Figure 3**), vertical gene transfer and horizontal gene transfer (HGT). These three processes have significantly contributed to mosaic genomes characterised by a mix and match of genes from different organisms (Armbrust, 2009). Chlorophytes, glaucophytes and rhodophytes are the products of separate endosymbiotic events in which heterotrophic, single-celled, eukaryotic hosts engulfed a cyanobacteria, eventually becoming membrane-bound organelles known as plastids (Falkowski *et al*., 2004). Secondary endosymbiosis took place in new heterotrophic eukaryotes with green and red algae becoming plastids with additional membranes (Cavalier-Smith, 1999). This led to the evolution of two groups with green plastids: the euglenoids and chlorarachniophyte and four groups with red plastids: the cryptophytes, dinoflagellates, haptophytes (including coccolithophores) and heterokontophytes (including diatoms). There is also evidence of prior secondary endosymbiosis in the chromalveolates with a green alga, followed by gene transfer to the nucleus and plastid loss (Moustafa *et al*., 2009). Some dinoflagellate members have undergone a tertiary endosymbiosis, whereby heterotrophic dinoflagellates have engulfed haptophytes or diatoms (Falkowski *et al*., 2004). Algal genomes are made up of host, plastid and bacterial genes likely acquired by HGT (Bowler *et al*., 2008; Raymond and Kim, 2012). The evolutionary origins of microalgae have most likely enabled them to adapt to a wide range of environmental conditions and habitats. **See also**: **Algae: Phylogeny and Evolution**

Arguably the most important aspect in microalgal evolution is endosymbiosis leading to gene transfer from an endosymbiont's nucleus, plastid and mitochondria to the host nucleus. Genes gained through endosymbiotic and HGT have shaped the metabolic pathways of microalgae that include haem (Oborník and Green, 2005), lipid (Chan *et al*., 2013) and amino acid (Jiroutová *et al*., 2007) biosynthesis. Fatty acid biosynthesis in diatoms is controlled by genes from green (prasinophytes) and red algal lineages and from the heterotrophic host (Chan *et al*., 2013). The tryptophan biosynthesis pathway in diatoms also contains genes from the red algal symbiont and heterotrophic host as well as genes from bacterial origins. In *Phaeodactylum tricornutum*, this includes a gene fused to a sequence coding for a hypothetical protein of cyanobacterial origin (Jiroutová *et al*., 2007). Genes from the primary endosymbiont and eukaryotic host are also found for haem biosynthesis in *Thalassiosira pseudonana* and *Cyanidioschyzon merolae* as well as a gene with possible mitochondrial origin (Oborník and Green, 2005).

Adaptations in Different Marine Ecosystems

Coastal and estuarine regions

Coastal and estuarine regions are characterised by high and fluctuating nutrient levels, turbulence and light conditions which can be both low and variable (**Figure 4**). Turbulent conditions result in sediment suspension maintaining high nutrient concentrations in the water column (Smetacek, 1985). Diatoms and prasinophytes dominate these regions (Rodrigues *et al*., 2014) and are adapted to use a 'velocity strategy' for nitrate (Litchman *et al*., 2007) and phosphate (Kwon *et al*., 2013) acquisition, characterised by high maximum uptake rates and high growth rates (Litchman *et al*., 2007). High uptake and lower growth rates allow some diatoms to store nutrients in large vacuoles during high nutrient pulses (Litchman *et al*., 2007), allowing

eLS © 2015, John Wiley & Sons, Ltd. www.els.net

Table 1 Number of described microalgae species

Marine groups of microalgae	Level	Number of species
Bacillariophyta	Phylum	13 426
Chlorarachniophyceae	Class	14
Chlorophyta	Phylum	4500[a]
Cryptophyta	Phylum	194
Dinophyta	Phylum	3308
Euglenozoa	Phylum	1323
Glaucophyta	Phylum	20
Haptophyta	Phylum	626
Rhodophyta	Phylum	600[a]
Heterokonts besides Bacillariophyta	Phylum	1670

[a]Approximate numbers are used for chlorophyta and rhodophyta owing to large numbers of macroalgae within these groups. Based on data from AlgaeBase (http://www.algaebase.org).

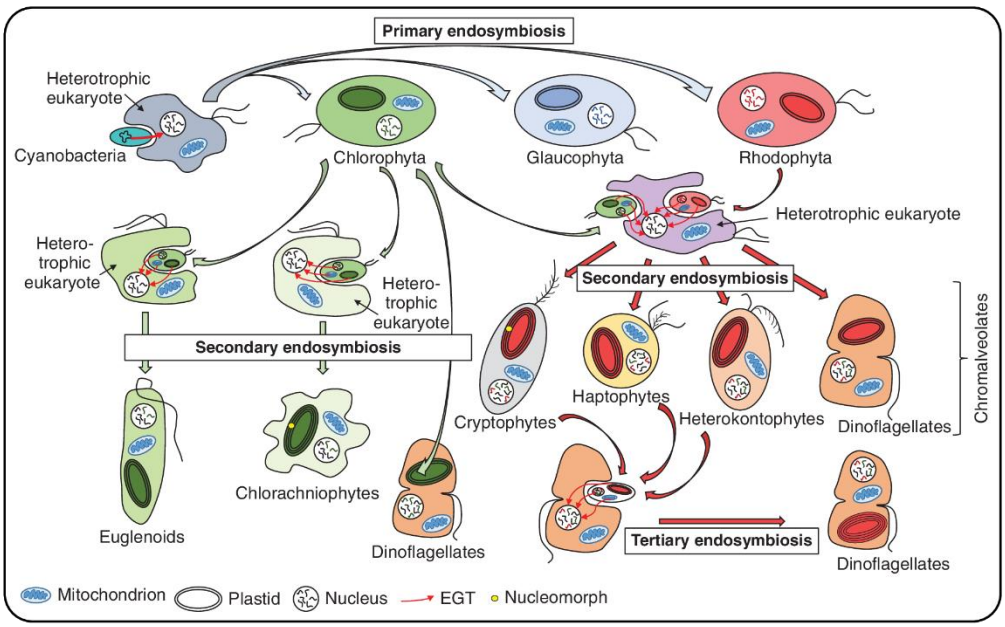


Figure 3 Evolution of algae according to primary, secondary and tertiary endosymbiotic events. EGT, endosymbiotic gene transfer.

them to thrive at fluctuating nutrient levels. Nitrate transporters are highly expressed in the diatom *Skeletonema costatum* under nitrate-available and nitrogen-limiting conditions (Hildebrand and Dahlin, 2000), suggesting that diatoms take up nitrate when it is both plentiful and scarce. Coastal diatoms, *S. costatum* and *Cylindrotheca fusiformis*, are adapted to maximise uptake of ammonium over nitrate as it is less costly to assimilate. This is accomplished by downregulating nitrate transporter genes in the presence of ammonium and increasing ammonium transporters in nitrogen-limiting conditions (Hildebrand and Dahlin, 2000; Hildebrand, 2005; Liu *et al*., 2013). Coastal and estuarine microalgae can uptake and use a variety of essential nutrients from their surroundings, including nitrate, ammonium, phosphate, sulfate, silicate and iron. Genes for transport of organic nutrients such as amino acids and purines (Armbrust *et al*., 2004; Bowler *et al*., 2008; Worden *et al*., 2009) may help provide nitrogen during oscillating nutrient levels, whilst genes for a complete urea cycle, thought to originate from the host of secondary endosymbiosis, may allow rapid recovery after nitrogen deprivation (Allen *et al*., 2011). A major limiting factor for

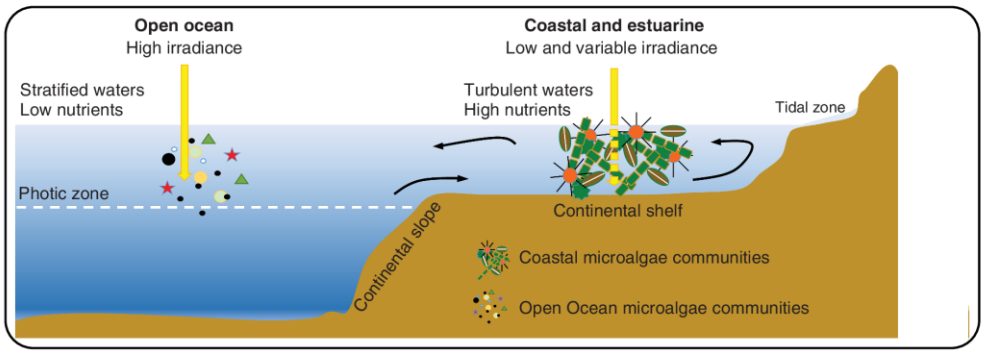


Figure 4 Major differences between nutrient, light and turbulance in marine coastal and open-ocean ecosystems.

growth in some coastal regions is iron concentration (Hutchins and Bruland, 1998). Microalgae have adapted mechanisms for uptake and storage of iron at low levels, as well as reducing iron requirements. Allen *et al.* (2008) looked at transcript levels in the pennate diatom *P. tricornutum* under low iron conditions. A marked decrease in mechanisms that rely on iron-dependent enzymes such as photosynthesis and respiration was shown. In response to these decreases *P. tricornutum* reduced carbon flux to respiratory processes and increased polysaccharide breakdown to release sugars. Iron-requiring and -independent antioxidants were down- and upregulated, respectively allowing reactive oxygen species (ROS) formed by the Fe-limited photosynthetic components to be scavenged. To reduce the production of ROS, excess electrons were channelled to mitochondrial alternative oxidase and dissipated by upregulation of genes to increase non-photochemical quenching (NPQ) capacity. Downregulation in nitrate assimilation genes was compensated by recycling proteins. Bacterial-like ferrichrome-binding proteins and a suite of iron-sensitive genes have been found in *P. tricornutum* but not in *T. pseudonana,* which requires much higher iron concentrations to survive (Allen *et al.*, 2008). Although some genes such as ferric reductases (Allen *et al.*, 2008) are shared, *T. pseudonana* appears to display different iron-associated genes under iron-limitation compared to *P. tricornutum*, including a ferroxidase, metal transporters and two iron permeases (Kustka *et al.*, 2007). This highlights the role that different genes play in adaptation to different conditions.

Suspension of sediment, organic matter and cells by turbulence along with tidal changes and scattering of light can create low and variable light conditions (Depauw *et al.*, 2012). Microalgae are able to acclimate to low light conditions by increasing the concentration of their photosynthetic pigments (Rodrigues *et al.*, 2014; Ezequiel *et al.*, 2015). Whilst this increases photosynthetic efficiency, it also increases excess light energy under high irradiance, leading to photoinhibition and production of ROS which can damage the cell (Goss and Lepetit, 2015). It is therefore important to have an effective mechanism to dispel excess photons. This role is normally carried out by NPQ. The role of NPQ has been recently reviewed for several plant and algal groups by Goss and Lepetit (2015). Diatoms use an NPQ system that converts diadinoxanthin (Dd) to diatoxanthin (Dt) in the presence of a proton gradient induced by high light. The Dd–Dt xanthophyll cycle is beneficial in coastal regions as it responds faster to high light than the VAZ (violaxanthin, antheraxanthin and zeaxanthin) cycle found in algae such as the chlorophytes. Components of the Dd–Dt cycle are more soluble within the thylakoid membrane, meaning they can occur at higher concentrations. Several genes for LHCX proteins, linked to modulation of NPQ under high and fluctuating light conditions (Goss and Lepetit, 2015), have been found in diatoms (Armbrust *et al.*, 2004; Bowler *et al.*, 2008). Some microalgae are also able to use motility to avoid photodamage and maximise light harvesting (Ezequiel *et al.*, 2015). Genes with homology to photoreceptors have been found in diatoms (Armbrust *et al.*, 2004), which may mean that some microalgae are able to detect their position in the water column relative to light levels.

The silica frustule may also play a role in the adaptation and dominance of diatoms in coastal systems. Drag on frustules spines may help cells to spin in the water, increasing nutrient uptake by microturbulence (Smetacek, 1985). For chain-forming microalgae (**Figure 5**) in turbulent conditions, stiffer chains were modelled to have higher nutrient uptake per cell, indicating a potential advantage for silicified diatom chains (Musielak *et al.*, 2009). Adaptations such as the formation of gaps between cells reduce the impact of lower nutrient diffusion per cell seen for chains, whilst taking advantage of higher nutrient transport seen in chains during turbulent conditions (Pahlow *et al.*, 1997). The shape of the frustule may also give some diatoms an advantage, with elongated cells, such as pennate diatoms, modelled to have a greater diffusive nutrient supply (Pahlow *et al.*, 1997). The silica frustule may also act to focus light into the cell (De Stefano *et al.*, 2007), benefiting diatoms in low light conditions. **See also**: **Diatoms**

Open oceans and oligotrophic waters

The open ocean is characterised by stratified waters, low nutrients and high irradiance (**Figure 4**). It homes a range of microalgae

eLS © 2015, John Wiley & Sons, Ltd. www.els.net

Figure 5 Melosira sp. chain illustrating the advantages of a silica frustule: rigidity in microalgae chains can increase nutrient uptake in turbulent conditions. Images of diatoms collected during the Tara Oceans expedition 2009–2012. Courtesy of Christian Sardet and Chris Bowler.

including coccolithophores, prasinophytes, dinoflagellates and pelagophytes (Litchman *et al*., 2007; Rodrigues *et al*., 2014). Affinity strategists such as coccolithophores have low half saturation constants, which allow them to compete for nutrients at low concentrations (Litchman *et al*., 2007). The important pelagic coccolithophore *Emiliania huxleyi* has a high affinity for phosphate but not nitrate (Riegman *et al*., 2000), giving it an advantage in phosphate-limiting conditions. When comparing genes between prasinophytes, an oligotrophic Micromonas sp. was found to encode a higher number of transporters from a larger range of protein families than a high-nutrient sourced Ostreococcus. This may reflect the need to optimise nutrient uptake in low nutrient conditions (Worden *et al*., 2009). The number of phosphate transporters also varied between *E. huxleyi* strains, suggesting differences in uptake capability within the same species (Read *et al*., 2013). Low uptake rates for nitrate in *E. huxleyi* may be assuaged by the ability to uptake a variety of nitrogen forms including nitrate, nitrite, urea and ammonium (Read *et al*., 2013). The importance of ammonium as a nitrogen source can be seen in the high number of ammonium transporter genes found in this species (Read *et al*., 2013).

Adaptations such as mixotrophy may supplement nutrient uptake or allow species with less capacity to compete for nutrients such as dinoflagellates (Litchman *et al*., 2007) to survive. Interactions with bacteria may also provide vitamins (see Biological Interactions).

Many microalgae in open oceans need to cope with low iron concentrations. Production of several enzymes, which use zinc, copper, manganese and nickel instead of iron, helps to reduce iron demands, whilst iron transporters and binding proteins optimise acquisition (Read *et al*., 2013; Worden *et al*., 2009). A reduction in photosystem I complexes reduces the demand for iron in the oceanic diatom *Thalassiosira oceanica*. This has led to the loss of rapid NPQ under variable light conditions, but as this species thrives under continuous light, it seems that iron requirement exerts higher selective pressure (Strzepek and Harrison, 2004). Acclimation to high light levels in open waters involves a reduction in photosynthetic pigments such as fucoxanthin chlorophyll *a*–*c*-binding protein (FCP) seen in *E. huxleyi* (McKew *et al*., 2013). Removal of excess energy by NPQ through an increase in xanthophyll, Dd (McKew *et al*., 2013) and photoprotection-linked LHC Lhcz proteins (McKew *et al*., 2013; Read *et al*., 2013) is also important. Adaptations to protect cells from ROS produced under high light include production of antioxidants that can be utilised under iron-limiting conditions such as Cu or Zn superoxide dismutase (Worden *et al*., 2009; Read *et al*., 2013) and production of proteins for repair of photosystem II. **See also**: **Genomics of Algae**

Marine polar systems

Microalgae in marine polar systems are subjected to a variety of extreme conditions: cold temperatures, high irradiance during summer and total darkness during winter. Additionally, freezing of surface waters leads to high salt and oxygen concentrations and limited gas exchange in sea ice (Mock and Thomas, 2008). In the Southern Ocean, nutrient levels are high owing to upwelling and strong vertical mixing. This leads to high nitrate, phosphate and silica concentrations although iron tends to be limiting. Iron limitation is less pronounced in the Arctic Ocean owing to river run-off (Mock and Thomas, 2008). Pennate diatoms dominate the polar oceans (e.g. *Fragilariopsis*, *Navicula),* haptophytes and dinoflagellates also occur (Mock and Thomas, 2008), suggesting that chromalveolates are particularly well adapted to environmental conditions of polar marine ecosystems.

It is important to maintain photosynthetic efficiency, whilst minimising damage during long periods of irradiance. As seen with microalgae in warmer conditions, *Fragilariopsis cylindrus* responds to high irradiance by increasing NPQ and reducing proteins involved in light harvesting. This is accomplished by upregulating genes linked with energy dissipation such as high-light FCPs along with downregulating other light-harvesting genes (Mock and Hoch, 2005). LHCx proteins are a component of the FCP complex involved in light harvesting (Goss and Lepetit, 2015) and are also capable of binding xanthophyll pigments associated with NPQ (Lyon and Mock, 2014). An increase in LHCx proteins during high light conditions, and a greater number of LHCx genes in *F. cylindrus* compared to temperate diatoms, suggests that LHCx proteins play a role in adaptation to both high light and lower temperatures. Another adaptation to high light levels including ultraviolet irradiance is the presence of antioxidants that are able to scavenge ROS. An increase in antioxidants can be seen in response to high light in *F. cylindrus* (Mock and Hoch, 2005). The ability to repair proteins damaged by light-related processes is also important in these conditions, which may explain the presence of multiple peptidase genes in *F. cylindrus* (Mock and Valentin, 2004). Adaptations to low light levels under freezing conditions include upregulation of FCP genes (Mock and Valentin, 2004), leading to an increase in photosynthetic pigments and light-harvesting ability (McKew *et al*.,

2013). The combined presence of both high light and extreme cold imposes more stress on polar microalgae, which explains upregulation of stress response genes under these conditions (Mock and Valentin, 2004). Polar microalgae also experience long periods of darkness. Adaptations such as the ability to accumulate and use storage molecules including lipids, glucan and starch and enter hibernating states can aid survival without photosynthesis (Lyon and Mock, 2014).

Membranes surrounding the cell and organelles in microalgae need to be kept fluid during cold temperatures. Fluidity can be maintained by increasing concentration of polyunsaturated fatty acids (PUFA) (Lyon and Mock, 2014), seen by an upregulation of genes involved in fatty acid production under freezing conditions (Mock and Thomas, 2008).

Genes for ice-binding proteins (IBPs) have been found in diatoms, haptophytes and prasinophytes adapted to polar environments (Raymond and Kim, 2012). *F. cylindrus* responds to freezing temperatures by upregulating IBPs genes (Mock and Thomas, 2008). Presence of IBPs leads to ice-pitting activity and formation of smaller ice pockets (Raymond and Kim, 2012). This keeps the immediate environment surrounding microalgae as a liquid state, allowing import and export of molecules between the cell and surrounding water. Interestingly, IBP genes in the Raymond and Kim (2012) study showed homology to bacterial genes, suggesting that they were acquired by HGT. The acquisition of genes such as IBP may be a key component in the colonisation of sea-ice by microalgae.

It is important for microalgae to have sufficient access to nutrients and carbon whether they are limited by entrapment in sea-ice or freely available. It has been shown that enzymes involved in uptake of nitrate, ammonium and carbon in ice diatoms have evolved to be active at very low temperatures, whilst others such as nitrate reductase have been found to work best at slightly higher temperatures but are less sensitive to temperature changes (Priscu *et al*., 1989). Enzymes such as ribulose-1,6-bisphosphate carboxylase (RUBISCO) involved in carbon fixation (Devos *et al*., 1998) or ribosomal proteins for translation (Toseland *et al*., 2013) make up for a decrease in activity in cold environment by increasing concentration.

Polar microalgae produce a range of osmoprotectants that protect cells by reducing the intracellular freezing point, an adaptation essential for sea-ice microalgae, which have to contend with high salinity and freezing temperatures. These include betaine, DMSP, sugars, polyols and amino acids (Lyon and Mock, 2014). Genes for proline synthesis, an important osmolyte, have been found in *F. cylindrus* salt shock EST libraries (Krell, 2006). In the same study, several abundant ion-transporters including antiporters were found, allowing *F. cylindrus* to maintain cellular homeostasis in high salt concentrations by passing ions both in and out of the cell (Krell, 2006).

Whilst some genes involved in adaptation can be linked to specific mechanisms and conditions, others may be harder to immediately identify. Many of the genes upregulated in transcription libraries or found after genome sequencing have no homology to known proteins (Mock *et al*., 2006; Hwang *et al*., 2008). Some genes such as putative transcription factors or those involved in protein editing show differential regulation (Hwang *et al*., 2008), but further work is needed to determine their targets. As mentioned earlier, several proteins display an increase or decrease in concentration in relation to conditions such as the cold, salinity and irradiance. Transcriptional and translational regulations underline these changes, and as such adaptations in gene regulatory networks are likely to be important in polar microalgae.

Biological Interactions

Associations between microalgae and other organisms occur across marine ecosystems and shape microalgae evolution through both direct and indirect interactions.

Several reports detailing interactions between microalgae and bacteria show an exchange of products (reviewed by Natrah *et al*., 2014), including organic compounds (e.g. sugars), inorganic nutrients (e.g. nitrogen and iron) and vitamins (e.g. B12 and biotin). This may be enhanced by attachment of bacteria to algal cells (Amin *et al*., 2012). Bacterial products have the potential to impact microalgae evolution by selective pressure. A good example of this is dependence on cobalamin (vitamin B12) produced exclusively by bacteria. Microalgae contain two forms of methionine synthase: METH, which is B12 dependent and METE which is B12 independent, but thought to produce methionine at lower efficiencies (Helliwell *et al*., 2011). Loss of METE has occurred independently on several occasions in a range of microalgae, likely due to downregulation of METE in the presence of B12 leading to a possible mechanism for loss in B12-rich environments (Helliwell *et al*., 2011). The B12-dependent coccolithophore *E. huxleyi* was able to grow in media without B12, possibly due to close interactions seen with bacteria (Helliwell *et al*., 2011). Interactions with bacteria provide the opportunity for HGT, which likely explains the large number of bacterial genes found in several microalgae (Armbrust *et al*., 2004; Bowler *et al*., 2008). Chances for HGT may be increased by algae themselves, by stimulating the release of transformable plasmid DNA in bacteria (Matsui *et al*., 2003).

Dinoflagellates of the Symbiodinium genus are well known for their symbiosis with a range of cnidarians, in particular corals. Leggat *et al.* (2007) identified genes for ammonium and sulfate transporters, possibly associated with the transfer of inorganic nutrients from coral to dinoflagellate. A sugar permease was also identified which may provide the host with sugars, a common role within the coral-*Symbiodinium* symbiosis. A high number of antioxidants genes have been found within symbiotic *Symbiodinium* sp. which may be an adaptation to reduce the impact of ROS species, produced by photosynthesis, on the host (Bayer *et al*., 2012). Amongst the antioxidants found was a bacterial nickel-type SOD containing an ubiquitin domain, suggesting that the gene may be a fusion from eukaryotic and prokaryotic lineages.

The presence of signal transduction pathways linked to cell surface receptors in a cultured *Symbiodinium* sp. may enable them to sense and explore the environment (Voolstra *et al*. 2009), which could be useful in host colonisation. An important consideration in microalgae within a symbiotic relationship is the coevolution between microalgae and host. Voolstra *et al.* (2009) found three species-specific genes (out of 115) in a colonising *Symbiodinium* not associated with other dinoflagellates. The highest dN/dS ratio,

eLS © 2015, John Wiley & Sons, Ltd. www.els.net

an indicator of selective pressure, was found in one of these genes, indicating positive selection. It has been speculated that this may be necessary to coevolve with the host or to compete with other symbionts (Voolstra *et al*., 2009). **See also**: **Algal Symbioses**

Several groups of algae are able to source nutrients from bacteria by phagotrophy. This includes haptophytes, cryptophytes, chlorophytes and dinoflagellates (Unrein *et al*., 2014). The haptophyte *Prymnesium parvum* preys on bacteria and ciliates to complement its photosynthetic capabilities (Liu *et al*., 2015). Genes to process fatty acids were upregulated in the presence of both types of prey, suggesting that *P. parvum* was able to use organic carbon acquired from both. Nitrogen, however, was primarily sourced from ciliates and iron from bacteria. In ciliate-associated cultures, genes for inorganic nitrogen uptake and processing were downregulated, whilst glutamine synthetase was downregulated and glutamate dehydrogenase was upregulated. This suggests that amino acids from ciliate prey are being used as a nitrogen source. In the presence of bacteria, genes for iron uptake were downregulated.

Whilst some interactions are beneficial, others can be harmful, leading to the evolution of defence mechanisms. Polyunsaturated aldehydes (PUA) are produced by several diatoms to reduce fitness in copepod grazers when mechanically disrupted (Pohnert *et al*., 2002). Genes involved in PUA production include phospholipase and lipoxygenase (Amin *et al*., 2012). The ability to use PUAs to reduce grazing depends not just on the species but the strain as well. One strain of *Thalassiosira rotula* was able to reduce the fitness of copepods, whilst the other was not (Pohnert *et al*., 2002), demonstrating that different adaptations can be found within the same species. Dimethylsulfoniopropionate (DMSP), an organosulfur compound, is produced by several algal groups and may act to deter grazers. DMSP has other functions, so selective pressure may be influenced by factors other than predation. Mechanical defences such as the ability to produce a silica frustule in diatoms deter grazers (Hamm *et al*., 2003) and antimicrobial compounds protect against pathogenic bacteria (Amin *et al*., 2012).

Being able to sense either synergistic, pathogenic, competitive or predatory organisms may be highly advantageous. Very little is known regarding the ability of algae to sense other life forms, however, putative GAF domains have been found in *T. psuedonana* and *P. tricornutum*, which may bind acyl homoserine lactones associated with bacterial sensing (Amin *et al*., 2012).

Concluding Remarks

There are many adaptations found within marine microalgae, which enable them to survive in a variety of marine habitats. Differences in essential processes such as photosynthesis, NPQ and nutrient acquisition as well as acquisition of specialist genes allow microalgae to cope with a range of conditions. Their diversity and abundance has a far-reaching impact on ocean ecosystems including carbon cycling, primary production, geochemical deposits and as a food source. Microalgae have acquired their mosaic gene pool from photosynthetic organisms, heterotrophic eukaryotes and bacteria, giving them a unique range of abilities.

References

Allen AE, LaRoche J, Maheswari U, Lommer M, *et al.* (2008) Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation. *Proceedings of the National Academy of Sciences* **105** (30): 10438–10443.

Allen AE, Dupont CL, Oborník M, *et al.* (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473** (7346): 203–207.

Amin SA, Parker MS and Armbrust EV (2012) Interactions between diatoms and bacteria. *Microbiology and Molecular Biology Reviews* **76** (3): 667–684.

Armbrust EV, Berges JA, Bowler C, *et al.* (2004) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* **306** (5693): 79–86.

Armbrust EV (2009) The life of diatoms in the world's oceans. *Nature* **459** (7244): 185–192.

Bayer T, Aranda M, Sunagawa S, *et al.* (2012) Symbiodinium transcriptomes: genome insights into the dinoflagellate symbionts of reef-building corals. *PLoS One* **7** (4): e35269.

Bowler C, Allen AE, Badger JH, *et al.* (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456** (7219): 239–244.

Boyd PW, Watson AJ, Law CS, *et al.* (2000) A mesoscale phytoplankton bloom in the polar Southern Ocean stimulated by iron fertilization. *Nature* **407** (6805): 695–702.

Cavalier-Smith T (1999) Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree1, 2. *Journal of Eukaryotic Microbiology* **46** (4): 347–366.

Chan CX, Baglivi FL, Jenkins CE and Bhattacharya D (2013) Foreign gene recruitment to the fatty acid biosynthesis pathway in diatoms. *Mobile Genetic Elements* **3** (5): e27313.

Courties C, Vaquer A, Troussellier M, *et al.* (1994) Smallest eukaryotic organism. *Nature* **370**: 255.

De Stefano L, Rea I, Rendina I, De Stefano M and Moretti L (2007) Lensless light focusing with the centric marine diatom Coscinodiscus walesii. *Optics Express* **15** (26): 18082–18088.

Depauw FA, Rogato A, d'Alcalá MR and Falciatore A (2012) Exploring the molecular basis of responses to light in marine diatoms. *Journal of Experimental Botany* **63** (4): 1575–1591.

Devos N, Ingouff M, Loppes R and Matagne RF (1998) RUBISCO adaptation to low temperatures: a comparative study in psychrophilic and mesophilic unicellular algae. *Journal of Phycology* **34** (4): 655–660.

Ezequiel J, Laviale M, Frankenbach S, Cartaxana P and Serôdio J (2015) Photoacclimation state determines the photobehaviour of motile microalgae: the case of a benthic diatom. *Journal of Experimental Marine Biology and Ecology* **468**: 11–20.

Falkowski PG, Katz ME, Knoll AH, *et al.* (2004) The evolution of modern eukaryotic phytoplankton. *Science* **305** (5682): 354–360.

Field CB, Behrenfeld MJ, Randerson JT and Falkowski P (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281** (5374): 237–240.

Goss R and Lepetit B (2015) Biodiversity of NPQ. *Journal of Plant Physiology* **172**: 13–32.

Guiry MD (2012) How many species of algae are there? *Journal of Phycology* **48** (5): 1057–1063.

Guiry, M.D. & Guiry, G.M. 2015 *AlgaeBase*. World-wide electronic publication, National University of Ireland, Galway. http://www.algaebase.org; searched on 15 May 2015.

Hamm CE, Merkel R, Springer O, *et al.* (2003) Architecture and material properties of diatom shells provide effective mechanical protection. *Nature* **421** (6925): 841–843.

Helliwell KE, Wheeler GL, Leptos KC, Goldstein RE and Smith AG (2011) Insights into the evolution of vitamin B12 auxotrophy from sequenced algal genomes. *Molecular Biology and Evolution* **28** (10): 2921–2933.

Hildebrand M and Dahlin K (2000) Nitrate transporter genes from the diatom Cylindrotheca fusiformis (Bacillariophyceae): mRNA levels controlled by nitrogen source and by the cell cycle. *Journal of Phycology* **36** (4): 702–713.

Hildebrand M (2005) Cloning and functional characterization of ammonium transporters from the marine diatom cylindrotheca fusiformis (Bacillariophyceae). *Journal of Phycology* **41** (1): 105–113.

Hutchins DA and Bruland KW (1998) Iron-limited diatom growth and Si: N uptake ratios in a coastal upwelling regime. *Nature* **393** (6685): 561–564.

Hwang YS, Jung G and Jin E (2008) Transcriptome analysis of acclimatory responses to thermal stress in Antarctic algae. *Biochemical and Biophysical Research Communications* **367** (3): 635–641.

Jiroutová K, Horák A, Bowler C and Oborník M (2007) Tryptophan biosynthesis in stramenopiles: eukaryotic winners in the diatom complex chloroplast. *Journal of Molecular Evolution* **65** (5): 496–511.

Krell, A. (2006) *Salt Stress Tolerance in the Psychrophilic diatom Fragilariopsis cylindrus*. Dissertation, University of Bremen, Germany.

Kustka AB, Allen AE and Morel FM (2007) Sequence analysis and transcriptional regulation of iron acquisition genes in two marine diatoms. *Journal of Phycology* **43** (4): 715–729.

Kwon HK, Oh SJ and Yang HS (2013) Growth and uptake kinetics of nitrate and phosphate by benthic microalgae for phytoremediation of eutrophic coastal sediments. *Bioresource Technology* **129**: 387–395.

Leggat W, Hoegh-Guldberg O, Dove S and Yellowlees D (2007) Analysis of an EST library from the dinoflagellate (Symbiodinium sp.) symbiont of reef-building corals1. *Journal of Phycology* **43** (5): 1010–1021.

Litchman E, Klausmeier CA, Schofield OM and Falkowski PG (2007) The role of functional traits and trade-offs in structuring phytoplankton communities: scaling from cellular to ecosystem level. *Ecology Letters* **10** (12): 1170–1181.

Liu Y, Song X, Han X and Yu Z (2013) Influences of external nutrient conditions on the transcript levels of a nitrate transporter gene in Skeletonema costatum. *Acta Oceanologica Sinica* **32** (6): 82–88.

Liu Z, Jones AC, Campbell V, *et al.* (2015) Gene expression in the mixotrophic prymnesiophyte, Prymnesium parvum, responds to prey availability. *Frontiers in Microbiology* **6**.

Lyon BR and Mock T (2014) Polar microalgae: new approaches towards understanding adaptations to an extreme and changing environment. *Biology* **3** (1): 56–80.

Matsui K, Ishii N and Kawabata ZI (2003) Release of extracellular transformable plasmid DNA from Escherichia coli cocultivated with algae. *Applied and Environmental Microbiology* **69** (4): 2399–2404.

McKew BA, Davey P, Finch SJ, *et al.* (2013) The trade-off between the light-harvesting and photoprotective functions of fucoxanthin-chlorophyll proteins dominates light acclimation in Emiliania huxleyi (clone CCMP 1516). *New Phytologist* **200** (1): 74–85.

Mock T and Hoch N (2005) Long-term temperature acclimation of photosynthesis in steady-state cultures of the polar diatom Fragilariopsis cylindrus. *Photosynthesis Research* **85** (3): 307–317.

Mock T and Thomas DN (2008) Microalgae in polar regions: linking functional genomics and physiology with environmental conditions. In: Psychrophiles: From Biodiversity to Biotechnology, pp. 285–312. Berlin Heidelberg: Springer.

Mock T and Valentin K (2004) Photosynthesis and cold acclimation: molecular evidence from a polar diatom. *Journal of Phycology* **40** (4): 732–741.

Mock T, Krell A, Glöckner G, Kolukisaoglu Ü and Valentin K (2006) Analysis of expressed sequence tags (ests) from the polar diatom fragilariopsis cylindrus. *Journal of Phycology* **42** (1): 78–85.

Moustafa A, Beszteri B, Maier UG, *et al.* (2009) Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324** (5935): 1724–1726.

Musielak MM, Karp-Boss L, Jumars PA and Fauci LJ (2009) Nutrient transport and acquisition by diatom chains in a moving fluid. *Journal of Fluid Mechanics* **638**: 401–421.

Natrah FM, Bossier P, Sorgeloos P, Yusoff FM and Defoirdt T (2014) Significance of microalgal–bacterial interactions for aquaculture. *Reviews in Aquaculture* **6** (1): 48–61.

Oborník M and Green BR (2005) Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. *Molecular Biology and Evolution* **22** (12): 2343–2353.

Pahlow M, Riebesell U and Wolf-Gladrow DA (1997) Impact of cell shape and chain formation on nutrient acquisition by marine diatoms. *Limnology and Oceanography* **42** (8): 1660–1672.

Pohnert G, Lumineau O, Cueff A, *et al.* (2002) Are volatile unsaturated aldehydes from diatoms the main line of chemical defence against copepods? *Marine Ecology Progress Series* **245** (1): 33–45.

Priscu JC, Palmisano AC, Priscu LR and Sullivan CW (1989) Temperature dependence of inorganic nitrogen uptake and assimilation in Antarctic sea-ice microalgae. *Polar Biology* **9** (7): 443–446.

Raymond JA and Kim HJ (2012) Possible role of horizontal gene transfer in the colonization of sea ice by algae. *PLoS One* **7** (5): e35968.

Read BA, Kegel J, Klute MJ, *et al.* (2013) Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature* **499** (7457): 209–213.

Riegman R, Stolte W, Noordeloos AA and Slezak D (2000) Nutrient uptake and alkaline phosphatase (EC 3: 1: 3: 1) activity of Emiliania huxleyi (Prymnesiophyceae) during growth under N and P limitation in continuous cultures. *Journal of Phycology* **36** (1): 87–96.

Rodrigues SV, Marinho MM, Jonck CCC, *et al.* (2014) Phytoplankton community structures in shelf and oceanic waters off southeast Brazil (20°–25° S), as determined by pigment signatures. *Deep Sea Research Part I: Oceanographic Research Papers* **88**: 47–62.

Rothschild LJ and Mancinelli RL (2001) Life in extreme environments. *Nature* **409** (6823): 1092–1101.

Smetacek VS (1985) Role of sinking in diatom life-history cycles: ecological, evolutionary and geological significance. *Marine Biology* **84** (3): 239–251.

eLS © 2015, John Wiley & Sons, Ltd. www.els.net

Strzepek RF and Harrison PJ (2004) Photosynthetic architecture differs in coastal and oceanic diatoms. *Nature* **431** (7009): 689–692.

Toseland A, Daines SJ, Clark JR, *et al.* (2013) The impact of temperature on marine phytoplankton resource allocation and metabolism. *Nature Climate Change* **3** (11): 979–984.

Unrein F, Gasol JM, Not F, Forn I and Massana R (2014) Mixotrophic haptophytes are key bacterial grazers in oligotrophic coastal waters. *The ISME Journal* **8** (1): 164–176.

Voolstra CR, Sunagawa S, Schwarz JA, *et al.* (2009) Evolutionary analysis of orthologous cDNA sequences from cultured and symbiotic dinoflagellate symbionts of reef-building corals (Dinophyceae: Symbiodinium). *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics* **4** (2): 67–74.

Worden AZ, Lee JH, Mock T, *et al.* (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. *Science* **324** (5924): 268–272.

Further Reading

Bork P, Bowler C, de Vargas C, *et al.* (2015) Tara Oceans studies plankton at planetary scale. *Science (Special Issue)* **348**: 873.

Caron DA, Countway PD, Jones AC, Kim DY and Schnetzer A (2012) Marine protistan diversity. *Annual Review of Marine Science* **4**: 467–493.

Marine Microbiology (2007) *Nature Reviews Microbiology* **5** (10).

Mock T, Daines SJ, Geider R, *et al.* (2015) Bridging the gap between omics and earth system science to better understand how environmental change impacts marine microbes. *Global Change Biology*. DOI: 10.1111/gcb.12983.

Worden AZ, Follows MJ, Giovannoni SJ, *et al.* (2015) Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* **347** 1257594.

Polar Microalgae: Functional Genomics, Physiology and the Environment

The following book section is part of the chapter titled, 'Polar Microalgae: Functional Genomics, Physiology and the Environment', and was published in 'Psychrophiles: From Biodiversity to Biotechnology' in 2017.

14.3 Adaptation of microalgae at high latitudes

14.3.1 Diatoms (Bacillariophyceae)

Psychrophilic diatoms are one of the most abundant groups of phytoplankton in polar oceans. This is mainly due to the presence of higher silicate concentrations in these waters and to their successful adaptation to strong vertical mixing in polar waters, strong seasonality in solar irradiance, freezing temperatures, and extremes of salinity (Cota 1985; Fiala and Oriol 1990; Boyd 2002; Mock and Valentin 2004; Ryan et al. 2004; Ralph et al. 2005). Due to their importance as primary producers, many physiological studies with polar diatoms were related either to growth and its dependency on nutrients and temperature or to regulation of photosynthesis under typical polar condition. This section aims to provide a comprehensive overview of new data regarding physiological and in particular molecular adaptation for this important group of polar algae.

Maximum growth rates for many polar diatoms are in the range of 0.25 to 0.75 divisions per day, that is 2- to 3-fold slower than growth at temperatures above 10 °C (Sommer 1989). Many of these diatoms are psychrophilic and not able to live at warmer temperatures (above ca. 15 °C) which is indicative of the presence of specific molecular adaptations that enable these diatoms to grow under freezing temperatures.

14.3.1.1 Functional genomics

Approaches to uncover the gene repertoire of a polar diatom have been dominated by the genus Fragilariopsis, in particular *Fragilariopsis cylindrus*, a marine indicator species for cold water, found at both poles (Quillfeldt 2004) and in seasonally cold waters (Hendey 1974; Hällfors 2004).

The first approaches involved constructing and sequencing two expressed sequence tag (EST) libraries, one generated under freezing temperatures (Mock et al. 2005) and the another under increased salinity (Krell 2008). 966 EST were generated from the cold stress library and 1691 from the salt stress library. There are now over 21000 EST from *F. cylindrus* on the EST- databank at NCBI and about 200 gene-specific oligonucleotides (70mers) from the original EST libraries for functional gene-array experiments (Mock and Valentin 2004). An important addition to algal research, particularly in terms of understanding polar adaptation, is the recent publication of the *F. cylindrus* genome and RNA-sequencing data generated under a range of polar conditions (Mock et

al. 2017). This is the sixth diatom genome to be published (Armbrust et al. 2004; Bowler et al. 2008; Galachyants et al. 2015; Lommer et al. 2012; Tanaka et al. 2015), and the first polar diatom. There is only one other polar microalga with a published genome, the psychrotolerant freshwater green alga *Coccomyxa subellipsoidea* (Blanc et al. 2012).

All EST-sequences were compared against the genomes of *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*. In addition, 11 algae and plant databanks were consulted to annotate sequences that were not found in the temperate diatom genomes. Nevertheless, over 50 % of sequences showed no similarity to known sequences in these databanks and to both diatom genomes even when using a comparatively high e-value of $\leq 10^{-4}$ (Mock et al. 2005).

In the cold-stress EST library, the most abundant functional categories were related to translation, post-translational modification of proteins and transport of amino acids and peptides by ABC-transporters. Some of these ABC-transporters displayed homology to bacterial permeases and others appeared to be involved in translational control or post-translational. However, most of them could not be assigned a function.

The presence of six different DNA/RNA helicases in the cold-stress library indicated that DNA and RNA coiling and uncoiling are important under freezing temperatures. Minimizing the likely formation of secondary structures and duplexes of mRNAs under low temperature stress is necessary to initiate translation. However, protein domains of DNA/RNA helicases are also the eighth most abundant protein domain in the genome of *T. pseudonana* (Armbrust et al. 2004) and therefore more evidence is necessary to conclude that these enzymes are essential to cope with freezing temperatures. The most abundant sequences in this library in terms of their redundancy were either sequences that were related to energy generation (e.g. fucoxanthin-chlorophyll a, c binding proteins) or completely unknown sequences (Mock et al. 2005).

In the salt-stress library, the most abundant functional categories of sequences were related to post-translational modification of proteins (e.g. heat-shock proteins; hsps) and ion-transport (Krell 2008). Most of them were hsps and different ionic transporter genes reflecting the requirement to re-establish homeostasis under salt stress. Several sequences of different kinds of V-type H+-ATPases and antiporters for various ions such as sodium, potassium and calcium were found in this library. V-type H+-ATPases are of great importance in establishing an electrochemical proton gradient across the tonoplast to drive sodium sequestration into the vacuole (Shi et al. 2003).

One important organic osmolyte under salt stress in diatoms is the amino acid proline. Many genes involved in proline synthesis were found in the salt-stress-EST library indicating that this pathway was active under experimental conditions (Krell 2008). The gene coding for pyrroline-5-carboxylate reductase (P5CR, catalyzing the final step in proline synthesis) could be identified among the most abundant sequences in the salt-stress library (Krell 2008). Furthermore, seven proteins involved in the proline synthesis pathway increased in abundance in response to high salinity (Lyon et al. 2011). This indicates that proline may be important for salt stress acclimation.

One of the interesting aspects of the *F. cylindrus genome* is the high number of divergent alleles. Approximately 25 % of the diploid genome consists of alleles that are highly divergent, particularly in comparison to the temperate diatom genomes of *T. pseudonana* and *P. tricornutum* (Mock et al. 2017).

Differential expression can be seen between divergent alleles under different conditions, many of which are commonplace in the polar environment, including prolonged darkness, freezing and elevated temperatures, iron starvation, and increased CO_2 concentration (Fig. 14.11b). In addition dN/dS analyses suggests that there may be a positive correlation between allelic differentiation and diversifying selection (Mock et al. 2017).

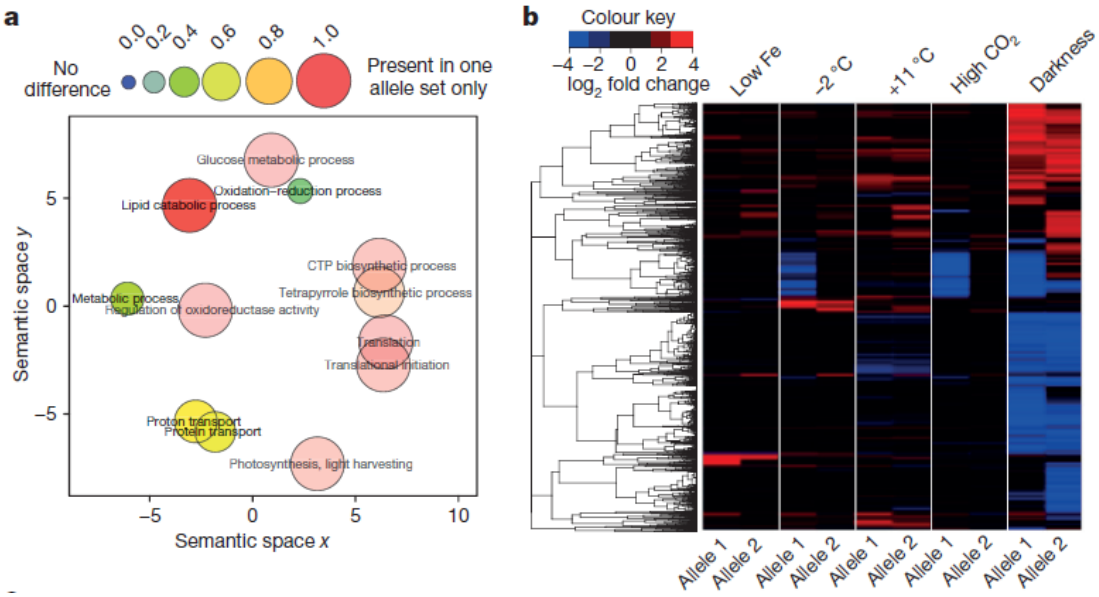


Fig. 14.11 Bi-allelic transcriptome and metatranscriptome profiling. (a) REViGO semantic similarity scatterplot of biological process gene ontology terms for *Fragiariopsis cylindrus*-like sequences (*E*-value ≤1×10−10) in Southern Ocean metatranscriptome samples. Gene ontology terms that are overrepresented in the set of diverged alleles compared to non-diverged alleles are shown in bold. (b) Hierarchical clustering of 4,030 differentially expressed allelic gene pairs in *F. cylindrus* (likelihood ratio test, *P* <0.001; log2 fold change ≤−2 or ≥+2) under low iron, freezing temperature (−2 °C), elevated temperature (+11 °C), elevated carbon dioxide (1,000 ppm CO_2) and prolonged darkness, relative to optimal growth conditions. Each experimental treatment corresponds to two separate columns for both allelic variants and each single-haplotype gene to a single row. Image is taken from Mock et al. (2017)

Copper rather than iron binding proteins are enriched in the *F. cylindrus* genome as are plastocyanin/azurin-like domains. This may facilitate electron transport during photosynthesis whilst reducing iron dependence. In terms of photosynthesis a large number of light harvesting complex (LHC) proteins are also present including Lhcx, which is involved in stress response. There are also a larger number of methionine sulfoxide reductase (MSR) genes in the *F. cylindrus* genome

compared to *T. pseudonana* or *P. tricornutum* that are linked to oxidative stress under cold temperatures (Lyon et al. 2014).

A large number of zinc-binding proteins can be found in this genome compared to the sequenced temperate diatoms. These contain myeloid-Nervy-DEAF-1 domains (MYND) which are associated with protein-protein interactions and regulation.

Enrichment of specific genes groups can be found in within the diverged alleles, these include; catalytic activity, transport, membrane proteins and metabolic processes (Fig.14.11a). Furthermore, divergent alleles were found to be differentially expressed under different conditions, suggesting that they may be involved in adaptation to polar conditions. Given the low sequence identity between promoters of divergent alleles and their differential regulation, it seems likely that individual copies are under different regulatory controls. RNA-seq data focused on changes in expression under prolonged (7 days) darkness as this condition gave rise to the highest number of up and down-regulated genes (Fig.14.11b). Down-regulated genes include those involved in photosynthesis, light harvesting, photoprotection and translation. Genes involved in regulation of gene expression, DNA replication, signal transduction and starch, sucrose or lipid metabolism were up-regulated (Mock et al. 2017). RNA-seq data suggests that during darkness, photosynthetic activity and supporting processes are reduced whilst processes such as chrysolaminarin and fatty acid storage are used instead.

Interestingly, as well as displaying the largest differential expression, growth under prolonged darkness also led to double the number of RNA-seq reads (30 %) that did not map to predicted genes compared to any other condition. Alleles with the largest dN/dS ratios tended to show strong differences in expression between conditions, in addition the majority of these alleles have no known function. As mentioned this suggests a positive correlation between diversifying selection and allelic differentiation. It also highlights the necessity for reverse genetics in polar species to determine the function of these sequences and in turn understand how they are adapted to polar environments.

One of the most interesting discoveries in the *F. cylindrus* EST salt-stress library was a gene involved in antifreeze processes (Krell 2008). The presence of ice-binding protein (IBP) genes in this species was verified following sequencing of the genome (Mock et al. 2017). Shortly after, IBPs were identified and characterised in the polar diatom *Navicula glaciei* (Janech et al. 2006). Since then several papers have been produced which explore the function of IBPs in polar diatoms, this is discussed in more detail in the next section. In diatoms ice-binding proteins have been identified in *F. cylindrus*, *Fragilariopsis curta*, *N. glaciei*, *C. neogracile*, *Attheya sp*., *Amphora sp*. and *Nitzshia stellate* (Janech et al. 2006; Krell 2008; Bayer-Giraldi et al. 2010; Gwak et al. 2010; Raymond and Kim 2012).

The N-terminal sequences of the identified IBPs of *N. glaciei*, *F. cylindrus* and each of the *T. ishikariensis* antifreeze isoforms are most likely signal peptides and have low probabilities of being mitochondrial- or chloroplast-targeting peptides (Janech et al. 2006; Fig. 14.12). N-terminal

sequences were found in *Attheya sp*. but not *Amphora sp*. or *Nitzshia stellate* and therefore may not be secreted (Raymond and Kim 2012).

Many diatom genes show homology to bacterial or fungal genes suggesting origins from horizontal gene transfer (HGT). *N. glaciei* and *F. cylindrus* IBPs show sequence similarity to several antifreeze isoforms of the Basidomycete fungus Typhula ishikaiensis (Fig.s 14.12 and Fig. 14.3), which is known to inhabit sea ice (Janech et al. 2006). Sorhannus (2011), also found homology between IBPs of *F. cylindrus* and *F. curta* to IBPs from basidomycetes, however in contrast to findings from Janech et al. (2006), IBPs from N. glaciei are placed in a separate clade and are suggested to originate from ancestral genes along with IBPs from *C. neogracile*.

Fig. 14.12 ClustalW alignment of ice-binding proteins from *Navicula glaciei* (Acc. no. DQ062566), *Fragilariopsis cylindrus* (CN212299) and *Typhula ishikariensis* (AB109745), and hypothetical proteins from *Cytophaga hutchinsonii* (ZP_00309837) and *Ferroplasma acidarmanus* (ZP_500309837). Predicted signal peptides are underlined. Gaps have been inserted to improve alignment. Conserved residues are shaded. The N-terminal sequence of *Cytophaga* protein and the N-and C-terminal sequences of *Ferroplasma* protein are truncated. Residue numbers are shown at right. Alignment is taken from Janech et al. (2006)

Similarities between *F. cylindrus* and *N. glaciei* IBPs to hypothetical proteins from Gram-negative bacteria such as *Cytophaga hutchonsonii* and *Shewanella denitrificans* (between 43 and 58 % amino acid sequence identity), have been observed. These bacteria have frequently been isolated from Arctic and Antarctic sea ice (Junge et al. 2002) and Cytophaga–Flavobacterium–bacteroides, which include *C. hutchonsonii* are important in well-established sea-ice algal assemblages (Bowman et al. 1997) and the coldest (wintertime) sea ice (Junge et al. 2004). Raymond et al. (2012) found IBPs from *Attheya sp*., *Amphora sp*. and *Nitzshia stellate* to show greatest homology to bacterial IBPs. These diatoms IBPs contain no introns and furthermore, Flavobacterium frigoris, which produces an IBP with 47 % amino acid identity to an IBP in *Nitzschia stellate* was isolated from Antarctic sea ice in the same layer as diatoms.

Expression of IBPs have also been demonstrated in the Antarctic bacterium *Marinomonas primoryensis*, where they aid adherence to ice, allowing *M. primoryensis* to remain near the top of the water column (Jung 2017) and in an Antarctic *Colwellia sp*. where they inhibit ice recrystallization (Raymond et al. 2007). In other organisms, antifreezes appear to have arisen from a variety of proteins with other functions, although some retain the original functions (Cheng 1998). Other genes with homology to bacteria found in the *F. cylindrus* genome include ABC transporters with similarities to bacterial permeases and proton-pumping proteorhodopsins, for trace-metal independent ATP synthesis (Strauss et al. 2013).

14.3.1.2 Molecular physiology

The presence of genes in a genome only indicates the potential for physiological adaptation, but knowledge of the expression and regulation of genes and the irrespective proteins leads to an actual understanding of how these diatoms cope with the extreme polar conditions. Expression analysis can be done by focusing on single genes (e.g. northern blots or quantitative PCR) or multiple genes through gene arrays or RNA sequencing. Arrays can be composed of known genes (gene-specific arrays) or the whole genome sequence (tiling arrays).

One of the most dramatic environmental changes in polar marine sea ice habitats is the freezing of seawater and the melting of the ice. The inclusion of organisms into newly-formed sea ice represents a strong selective pressure. Only those organisms that are capable of acclimation to the relatively fast-changing conditions of temperature, irradiance and salinity can survive.

Several experiments have been conducted to investigate gene expression under polar conditions including freezing temperatures, high salinity, high irradiance and prolonged darkness. Some study multiple genes using macroarrays (Mock and Valentin 2004) or RNA-seq (Mock et al. 2017) whilst others focus on specific genes such as ice binding proteins (Bayer-Giraldi et al. 2010; Bayer-Giraldi et al. 2011).

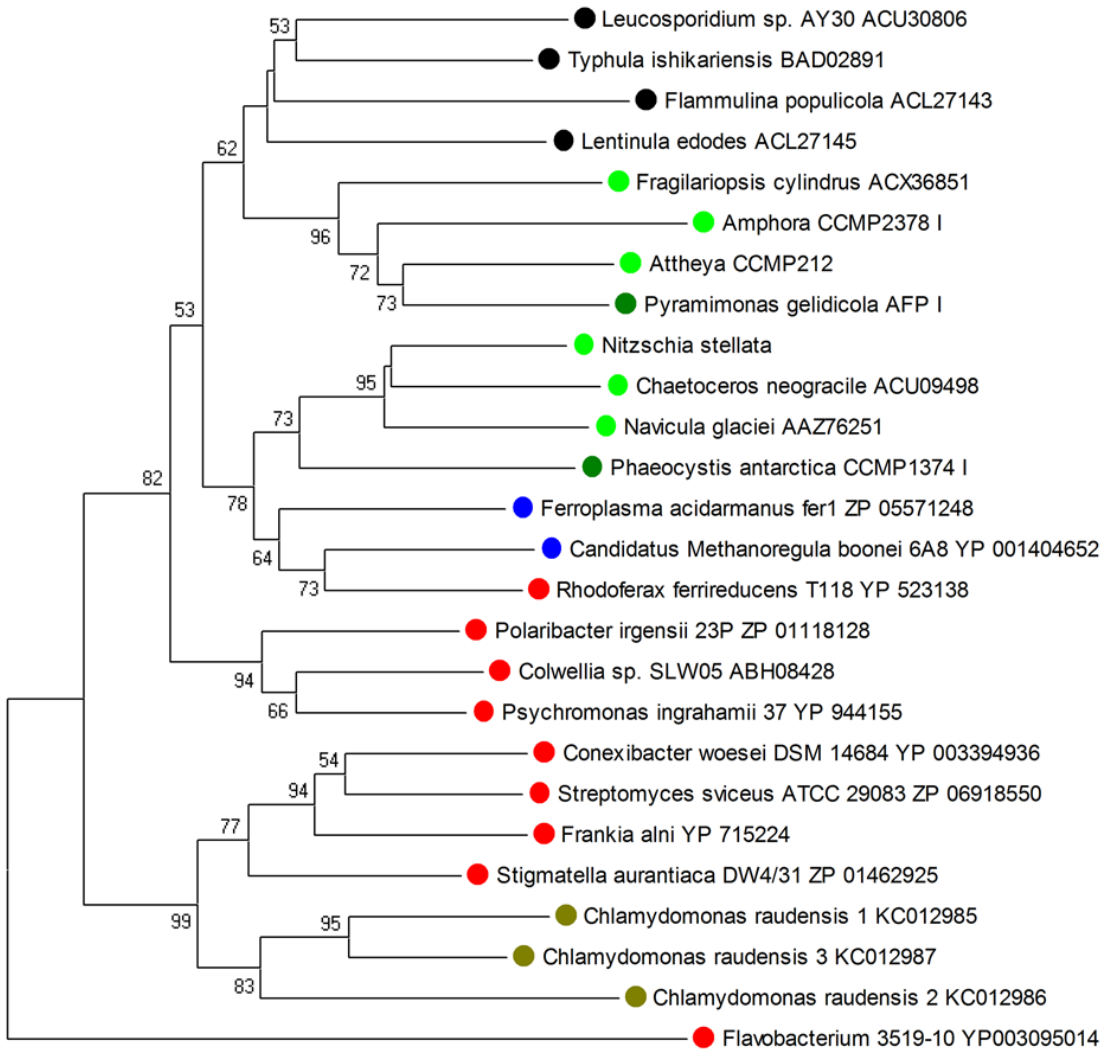


Fig. 14.13 Neighbor-joining tree constructed from amino acid sequences of selected ice-binding proteins (IBP) and IBP-like proteins. The *Chlamydomonas raudensis* IBPs (olive) are closest to IBP-like proteins in several bacteria and relatively distant from other algal IBPs. The tree was rooted with the *Flavobacterium* 3519-10 IBP. Numbers at nodes indicate bootstrap values for 500 replications. Values less than 50 are not shown. Colors: black, fungi; light green, diatoms; dark green prasinophyte and prymnesiophyte; blue, archaea; red, bacteria; olive, *C. raudensis*. (Janech et al. 2006)

Data from EST libraries has been used to produce arrays for two polar diatom species, *F. cylindrus* (Mock and Valentin 2004) and *C. neogracile* (Hwang et al. 2008; Park et al. 2010). About 200 70mer oligonucleotides were compiled into a nylon-membrane-based macro-array to study short-, mid- and long term acclimation to freezing temperatures under high and low irradiance in *F.*

cylindrus. One thousand four hundred *C. neogracile* transcripts were analysed using microarrays to observe expression at 4 and 10 °C (Hwang et al. 2008) as well as under high, moderate, low and changing light intensities (Park et al. 2010).

The short-term response to freezing temperatures, which simulates the incorporation into newly formed sea ice during fall, was characterized by down-regulation of genes encoding proteins for photosystem II (psbA and psbC) and carbon fixation (RUBISCO large subunit, rbcL) regardless of light intensity used (3 and 35 µmol photons m^{-2} s^{-1}). However, under higher irradiance (35 µmol photons m^{-2} s^{-1}) up-regulation of genes encoding chaperons (hsp 70) and genes for plastid protein synthesis and turnover (elongation factor EfTs, ribosomal rpS4 and plastidial ftsH protease) were observed (Mock and Valentin 2004).

In *Chaetoceros neogracile,* increased irradiance led to both up and down regulation of particular LHCx proteins and FCPs (Park et al. 2010). Several genes for cell division, transcription and signalling were up-regulated whilst many genes for photosynthesis (including LHC, FCPs and PSII associated proteins) were down regulated along with some transporter genes including members from the ABC transporter family.

In *Fragilariopsis cylindrus* freezing accompanied with a reduction in irradiance (from 35 to 3 µmol photons m^{-2} s^{-1}) showed a typical response to low-light acclimation by up-regulation of genes encoding specific fucoxanthin-chlorophyll a,c binding proteins (fcps) without signs of a cold stress response. Fcps are a diverse gene family composed of genes involved in light harvesting as well as dissipation of light (see Sect. 15.2; Mock and Valentin 2004). Low irradiance in this species also leads to an increase in chloroplast PUFAs which can maintain electron flow by increasing fluidity of the thylakoid membrane (Mock and Kroon, 2002).

Up-regulation of stress response genes and genes for protein turnover only under higher light intensities and decreasing temperatures, indicates that a decrease in temperature at such light intensities mimics a further increase in light that could be more stressful than the actual decrease in temperature was by itself (Mock and Valentin 2004). This phenomenon is probably part of a cold-shock response that is also known from temperate plants when they get exposed to lower temperatures (Allen and Ort 2001).

Entomoneis kufferathii, a sea-ice diatom, showed high catalase activity, which is linked to protection against oxidative damage, in response to high irradiance and low temperatures (Schriek 2000). Genes for glutathione metabolism, an important antioxidant, were up-regulated soon after exposure to high light in *C. neogracile*, although glutathione S-transferase and superoxide dismutase (SOD), two enzymes involved in scavenging ROS, were down-regulated (Park et al. 2010). A gradual increase in heat-shock proteins was observed under the same conditions over 6 h. Shifts in irradiance from either low to high or high to low light resulted in an increase in SOD in *Chaetoceros brevis* (Janknegt et al. 2008). An increase in temperature in *C. neogracile* from 4 to 10°C resulted in the up-regulation of several antioxidant genes including monoascorbate reductase, glutaredoxin, glutathione

peroxidase, glutathione S-transferase, and alternative oxidase (Hwang et al. 2008). Polar diatoms appear to have tailored and multiple resources for dealing with stress caused by the extreme polar environment.

Psychrophilic plants and diatoms are able to acclimate to higher irradiances under low temperatures (Streb et al. 1998; Mock and Hoch 2005; Ralph et al. 2005; Morgan-Kiss et al. 2006; Park et al. 2010). Long-term acclimation experiments to higher irradiances at freezing temperatures, when compared to the same light intensity but higher temperatures (+5 °C), revealed that cells kept at lower temperatures showed a typical response known from high-light acclimation: higher non-photochemical quenching, up-regulation of the gene psbA and up-regulation of high-light fcps that are involved in energy dissipation (Mock and Valentin 2004; Mock and Hoch 2005). A rapid increase in diatoxanthin (Dtx) in *C. neogracile* under high light also demonstrates energy dissipation through NPQ, along with an increase in expression of specific FCPs (Park et al. 2010).

In *F. cylindrus*, a reduction in expression of other photosynthesis-related genes (such as rbcL) was not observed after several months under freezing conditions indicating that long-term acclimation had been achieved.

Temperature effects that are less dependent on adjustments of the energy flow under freezing temperatures could also be identified by gene expression analysis (Mock and Valentin 2004). In the Mock and Valentin (2004) study, genes were selected that were either abundant in the EST libraries (e.g. ABC transporters), or were important for general acclimation to freezing temperatures (e.g. IBP, fatty-acid desaturase). Three unknown but abundant genes (in EST libraries) were also selected to see whether at least one of them is upregulated under freezing temperatures. Expression of these genes was investigated at+5 °C and 9 days after reducing the temperatures to −1.8 °C.

Up-regulation of a gene encoding a delta5-desaturase under freezing temperatures indicated the necessity for production of polyunsaturated fatty acids (PUFAs) to maintain membrane fluidity at lower temperatures. Delta-5 desaturases produce omega3-fattyacids such as EPA (20:5 n-3), one of the most abundant fatty acid in diatoms and the main fatty acid in the galactolipids MGDG and DGDG. Thus, it can be assumed that more EPA is necessary under freezing temperatures to keep the thylakoid membrane fluid for electron transport or other membrane-bound processes.

Teoh et al. (2013) also found that PUFAs concentration increased in *N. glaciei* with a decrease in temperature. In contrast, a delta-12 desaturase gene also known for producing PUFAs was not up-regulated in temperate cyanobacteria (Nishida and Murata 1996). This indicates a different mechanism of gene regulation for this enzyme in psychrophilic diatoms.

An ABC transporter gene was strongly up-regulated at −1.8 °C in *F. cylindrus*, however, the family of ABC-transporters is composed of genes with very diverse functions so it unclear of its specific function in response to freezing temperatures (Mock and Valentin 2004).

Extracellular polymeric substances (EPS) are linked to adaptation of diatoms in polar environments as both cryoprotectants and through maintenance of the cells microclimate (Underwood et al. 2010). An example of this can be seen in the sea-ice diatom *Melosira arctica*, in which EPS from the sea-ice diatom *Melosira arctica* altered the microstructure of ice-pore morphologies leading to salt retention (Krembs et al. 2011). EPS can include, but are not limited to substances such as polysaccharides (Aslam et al. 2012), uronic acid, peptides, proteins and glycoproteins (Krembs et al. 2011; Underwood et al. 2013).

Cryoprotectants can also help to maintain both the internal and external environment in polar cells and include solutes such as proline, DMSP and betaine (Lyon et al. 2014). Proline synthesis genes were enriched in the *F. cylindrus* cold stress EST library (Mock et al. 2005).

DMSP pathway-linked protein concentrations were also increased in response to high salinity as were two proteins isoforms with homology to bacterial/archaeal glycine betaine methyltransferase (Lyon et al. 2011). DMSP, which has been found in high concentrations in ice-diatom communities has been shown to protect enzymes against denaturation in freezing conditions (DiTullio et al. 1998).

Studies on ice-binding proteins in diatoms show that they have antifreeze properties and are able to inhibit ice recrystallization (Gwak et al. 2010; Bayer-Giraldi et al. 2011; Raymond 2011). As several IBP have similarities to bacterial or fungal sequences it is hypothesised that they have been acquired through HGT (see Sect. 15.3.1.1) and may have allowed diatoms to colonise sea-ice.

An IBP protein in *F. cylindrus* was strongly up-regulated (ca. 50-fold) under freezing temperatures (Mock and Valentin 2004) whilst Bayer-Giraldi et al. (2011) found several isoforms in *F. cylindrus* and *F. curta* to be differentially regulated depending on temperature and salt stress. *F. cylindrus* IBPs in both of these studies have been identified in the recently published genome (Mock et al. 2017). Proteomics studies on *C. neogracile* also showed an increase in concentration of IBPs in response to freezing conditions (Gwak et al. 2010). Furthermore, isolation of IBP transcripts from Arctic and Antarctic sea ice suggests that they are found at similar levels as genes with essential metabolic processes such as photosynthesis (Uhlig et al. 2015). Within the same study it was found that most IBP transcripts originated from diatoms, haptophytes and crustaceans, however many of the IBPs have not been previously characterised (Uhlig et al. 2015). These results support the hypothesis that these proteins are of great importance not only under salt stress but also under freezing temperatures to protect the cells from injury by growing ice crystals.

An important adaptation for polar photosynthetic organisms is the need to survive for periods of prolonged darkness. As discussed in Sect. 14.3.1.1, seven day darkness leads to a decrease in photosynthesis and associated processes. Genes which are involved in starch, sugar and fatty acid metabolism are up-regulated (Mock et al. 2017) suggesting that *F. cylindrus* is able to use existing cellular resources in place of photosynthesis. Diatoms are able to store glucan for use in periods of extended darkness (van Oijen et al. 2003), and are able to uptake molecules such as sugar and starch (Palmisano and Garrison 1993). The urea cycle in diatoms has been suggested as a means to process

inorganic carbon and nitrogen, particularly during low nitrogen availability (Allen et al. 2011) – all genes for the urea cycle can be found in the *F. cylindrus* genome. Proton-pumping proteorhodopsins, for trace-metal independent ATP synthesis (Strauss et al. 2013) were up-regulated under darkness, suggesting a role in energy production. There are also ATP- independent enzymes available to *F. cylindrus* which may save chemical energy such as pyrophosphate-dependent phospho-fructo-kinase which was elevated during salinity acclimation (Lyon et al. 2011).

Information is steadily becoming available for polar diatoms. New insights are being gained into their adaptations and the importance of their roles in polar communities. Although much has been learned, there are vast numbers of genes with unknown or partially characterised functions in many of these studies. For example, many identified transcripts have no homology to existing sequences (Mock et al. 2005; Krell 2008; Mock et al. 2017) and different FCPs and LHC proteins are both up- and down-regulated under the same conditions (Park et al. 2010). Reverse genetics is needed in order establish the function and roles of these genes and their pathways. A transformation system for *F. cylindrus* has been successfully established – as far as we are aware, this is the first transformation system for any polar species (Hopes and Mock, unpublished). Furthermore CRISPR-Cas for gene knock-out and gene silencing in the temperate diatoms *Thalassiosira pseudonana* (Hopes et al. 2016; Kirkham, unpublished) and *Pheodactylum tricornutum* (Nymark et al. 2016; De Riso et al. 2009) have been established. Work on CRISPR-Cas in *F. cylindrus* is also currently ongoing.

With the establishment of additional, elegant molecular tools for diatoms, there is a much greater scope for potential research and therefore our understanding of these psychrophilic and psychrotolerant organisms and their environment.

14.3.2 Green algae (Chlorophyceae)

Most polar green algae live in freshwater ecosystems such as snow, permanently ice-covered lakes or more ephemeral habitats like creeks or melt ponds on top of snow or sea ice. Most species belong either to the genera *Chlamydomonas, Chloromonas* or *Chlorella*, and many of them are very motile due to the presence of flagella.

Ecologically important species that are physiologically and molecularly well characterized are *Chlamydomonas raudensis, Chlamydomonas nivalis* and *Chlamydomomas sp*. ICE-L. *C. raudensis* is an abundant species in permanently ice-covered lakes and the cloneUWO241 have been studied for decades (see review by Morgan-Kiss et al. 2006). *C. nivalis* is a dominant representative of the snow-algae community and also intensively studied (Williams et al. 2003). Therefore, this discussion will mainly focus on *Chlamydomonas* sp. There is less research in this area in terms of functional genomics, however the genome sequencing of Coccomyxa subellipsoidea provides some insight into polar adaptations within the Chlorophyceae as does the cold shock EST library for *Pyramimonas gelidicola*.

14.3.2.1 Functional genomics

Coccomyxa subellipsoidea is a psychrotolerant green alga that has been isolated from dried algal peat in Antarctica, and although it can grow at low temperatures it shows optimal growth at around 20 °C (Blanc et al. 2012). Despite not being a true psychrophile its genome has some pronounced differences to mesophilic chylorphytes and offers several insights into polar adaptation. Although the genomes of several green algae have been sequenced this is the first genome to be published from a polar microalga. An EST library has also been generated under cold shock conditions for the psychrophilic *Pyramimonas gelidicola*, a dominant primary producer from Antarctic sea ice (Jung et al. 2012).

In comparison to other sequenced chlorophytes *C. subellipsoidea* has a large number of mitochondrial and chloroplast sequences integrated into its nuclear genome. GC content of these organelle genomes is also comparatively high. It is important to maintain homeostasis and efficient cellular functions under the extreme conditions found in polar regions. This includes lipid metabolism and membrane fluidity. Four lipid protein families were over-represented in *C. subellipsoidea*; type-I- fatty acid synthases, FA elongases, FA ligases and type 3 lipases. There were also three fatty acid desaturases present that were not found in temperate counter-parts (Blanc et al. 2012). An increase in double bonds in membrane based lipids helps to increase fluidity at cold temperatures (Los and Murata 2004). Within the same species, there were a high number of genes involved in polysaccharide and cell wall metabolism (Blanc et al. 2012). As previously mentioned both glycoproteins and polysaccharides can act as cryopreservants in microalgae. Two genes involved in cryoprotection with homology to late embryogenesis abundant (LEA) proteins have also been found in *C. subellipsoidea* (Liu et al 2011).

Structural parts such as the cytoskeleton of the cell also have to be adapted to low temperatures in order to conduct mitosis, meiosis, secretion and cell motility.

The tublin alpha chain protein domain was the fifth most abundant in the EST library from *P. gelidicola* (Jung et al. 2012). Willem et al. (1999) showed that alpha-tubulin from two *Chloromonas* spp. had five amino acid substitutions compared to the mesophilic *Chlamydomonas reinhardtii*. Two of these substitutions occurred in the region of inter dimer contacts that could therefore positively influence microtubule assembly under low temperatures.

Translation elongation factor-1a was prominent in EST from *P. gelidicola* (Jung et al. 2012). Furthermore, a translation elongation factor-1a was found in the *C. subellipsoidea* genome that is able to functionally replace elongation-factor like EFL found in previously sequenced chlorophytes. Up-regulation of an elongation factor involved in protein synthesis has also been observed in cold shock diatoms (Mock and Valentin 2004).

Given that polar species may be exposed to freezing temperatures and high light, many adaptive strategies include proteins involved in stress response and protection against ROS. DOPA-dioxygenase which provides protection against solubilised oxygen was identified in the *C. subellipsoidea* genome, as were two genes with homologs to phospholipase D and chalcone synthase. The former is involved in stress response, whilst homologs of the latter are involved in metabolites for UV photoprotection and antimicrobial defence in plants (Blanc et al. 2012).

Both heat shock protein 70 (hsp70) and stress related chlorophyll a/b binding protein were enriched in *P. gelidicola* ESTs. Heat shock protein 70 appears to be a key component involved adaptation of several polar microalgae species (Mock and Valentin 2004; Krell 2008; Liu et al. 2010).

When comparing *C. subellipsoidea* to temperate chlorophytes, Blanc et al. (2012), found that as well as enrichment of certain gene families and gene additions there were also several key genes missing. This includes PsaN, which is involved in docking plastocyanin to the PSI complex. This leads to a drop in electron transfer from plastocyanin to PSI which may be beneficial in a polar environment as low temperatures create an excess of electrons through this system which in turn leads to an increase in ROS. As PsaN is not crucial for photosynthesis, loss of this gene may protect the cell from oxidative damage (Blanc et al. 2012). *C. subellipsoidea* also has genes for dioxygenase and FA desaturases that utilize dioxygen and therefore may provide further protection against ROS (Blanc et al. 2012).

One gene loss which could reduce, cellular efficiency in *C. subellipsoidea*, however, is a pyruvate phosphate dikinase (PPDK), which produces ATP through glycolysis. Function of this gene appears to be replaced by three pyruvate kinases, which potentially produce less chemical energy (Blanc et al. 2012).

In terms of nutrient acquisition *C. subellipsoidea* has a large number of genes for amino acid permeases and transporters which may enhance uptake of organic nutrients. It also has cobalamin independent methionine synthase but lacks the cobalamin dependent version of this gene MetH (Blanc et al. 2012), suggesting that this species is not dependent on this often bacteria-associated co-factor (Croft et al. 2005) for synthesis of this important amino acid.

There is still much to discover in establishing the function and origins of many genes specific to polar species. There were a higher number of ESTs with unknown functions under freezing conditions in *P. gelidicola* compared to 4 °C (Jung et al. 2012). Furthermore, there are over 2300 genes in the *C. subellipsoidea* genome with no known homologs in sequenced mesophilic chlorophytes. The majority of these genes show homology to Streptophytes and other Eukaryotes, suggesting origins from a common ancestor to chlorophytes. Interestingly rather than displaying homology to green algae, most of the genes involved in defence, detoxification and carbohydrate metabolism show higher sequence similarity to bacteria, suggesting possible acquisition by HGT.

As discussed in Sect. 14.3.1.1, ice-binding proteins in diatoms appear to have bacterial or fungal origins. Several IBPs have also been identified in psychrophilic or psychrotolerant green algae including *Pyramimonas gelidicola* (Jung et al. 2014), *Chlamydomonas* raudensis (Raymond and Morgan-Kiss 2013), *Chlamydomonas sp.* strain CCMP681 (Raymond et al. 2009) and *Chloromonas brevispina* (Raymond 2014). Raymond and Morgan-Kiss (2013) separate ice-binding proteins into two different groups; IBP I, a group of similar proteins appearing to have fungal or bacterial origins (Raymond and Morgan-Kiss 2013; Sorhannus 2011; Raymond and Kim 2012, Jung et al. 2014; Raymond 2011) and IBP II. So far all studied algal species have type I IBPs with the exception of *Chlamydomonas sp.* strain CCMP681 which has four type II isoforms isolated from ESTs (Raymond et al. 2009; Raymond and Morgan-Kiss 2013). A polyphyletic origin for IBPs has been suggested given their sequential and structural differences, as well as a lack of IBPs in temperate species (Fig. 14.13, Raymond and Morgan-Kiss 2013).

As more genomes and transcriptomes become available for polar chlorophytes, more light can be shed on their intricacies and adaptations to extreme environments. The genome of an important Antarctic sea ice chlorophyte, *Chlamydomonas* ICE-L, has been recently sequenced (personal communication with Naihao Ye, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao 266071, China) and Raymond and Morgan-Kiss (2013) plan to compare the transcriptome of *C. raudensis* to a temperate counterpart.

14.3.2.2 Molecular physiology

Maximum growth rates of polar green algae are comparable to those from polar diatoms. They range from 0.2 to 0.4 day^{-1} (Tang et al. 1997). Temperatures above 18 °C are mostly lethal to these algae. C. *raudensis* has its maximum photosynthetic rates at 8 °C, which declines steadily with increasing temperatures (Morgan-Kiss et al. 2006). This indicates maximal efficiency in converting light into photosynthetic energy at low temperatures. The quality of light also plays an important role and *C. raudensis* is not able to grow under red light (Morgan-Kiss et al. 2005). This is probably a consequence of almost never being exposed to a longer wavelength spectrum in the natural habitat of permanently ice covered lakes where the ice absorbs all longer wavelengths of solar irradiance (Fritsen and Priscu 1999; Morgan-Kiss et al. 2006). However, the majority of this light is reflected on the white surface of ice and scattered while passing through ice. Thus, the environment below the ice is characterized by low intensities enriched in blue-green wavelengths (Lizotte and Priscu 1992).

Many physiological and molecular investigations have been conducted with *C. raudensis* to find the reasons for successful photo adaptation under these extreme conditions. A comparison with the temperate *C. reinhardtii* partly uncovered the mechanisms of photo adaptation in *C. raudensis* (Morgan-Kiss et al. 2005; Morgan-Kiss et al. 2006): In contrast to the temperate *C. reinhardtii*, the psychrophile has lost its ability to live under high light but increased its efficiency of light harvesting under low light in the blue-green spectrum. This adaptation can be seen in structural changes of the

photosynthetic apparatus (Fig. 14.14). For instance, *C. raudensis* has an unusually high ratio of photosystem II to I, and significantly higher levels of light harvesting II complexes than its temperate counterpart *C. reinhardtii*. These changes are probably an adaptive advantage under constant exposure to blue light of low photon flux densities because the light harvesting apparatus of photosystem II (PSII) utilizes chlorophyll b and short-wavelength-absorbing chlorophyll a to absorb light predominantly in the blue region. Interestingly, most marine algae (e.g. red algae, diatoms), which are also living in a blue-green light environment because of optical properties of the seawater, also show a high ratio of PSII to PSI due to chromatic regulation (Fujita 2001). However, most of them, and even psychrophilic diatoms, have the physiological ability to grow under high irradiance levels.

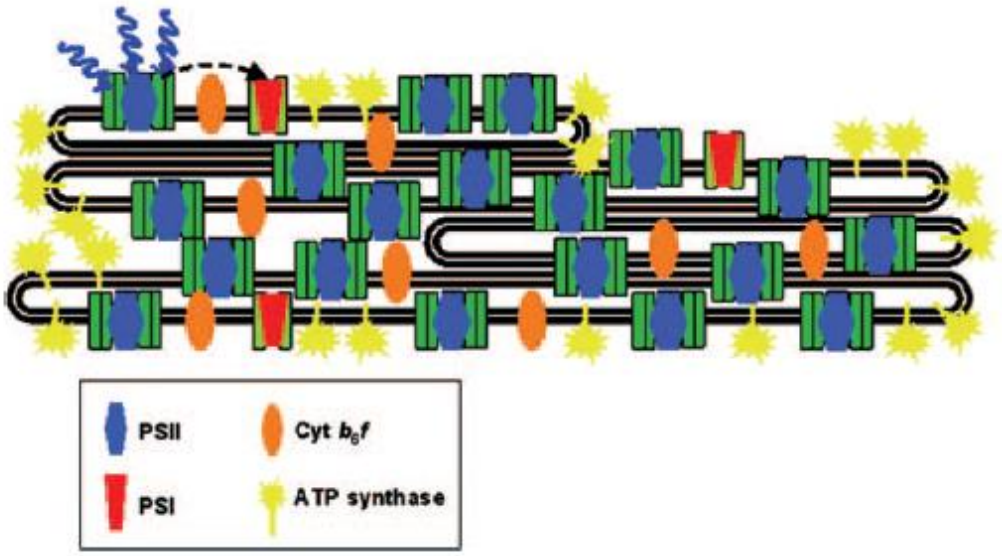


Fig. 14.14 Model for organization of thylakoid pigment-protein complexes of the electron transport chain in the psychrophilic *Chlamydomonas raudensis* UWO 241. In the natural, extremely stable light environment of extreme shade and predominantly blue-green wavelengths (blue lines), the majority of available light would be preferentially absorbed by PSII. Adaptation in *C. raudensis* to this light environment has led to an unusually high PSII/PSI stoichiometry and highly efficient energy transfer from LHCII to PSII. Conversely, PSI and associated light-harvesting complexes are both structurally and functionally downregulated. Given the severe reduction in light-harvesting capacity of PSI, it is proposed that PSI centers are largely excited via a spillover energy transfer mechanism from PSII (dotted line). Photosynthetic membranes may be arranged as loose stacks rather than distinct granal and stromal regions to promote energy spillover between the photosystems. Picture from Morgan-Kiss et al. (2006)

Whilst the ability to dissipate excess energy through NPQ has been reduced in *C. raudensis* (Morgan-Kiss et al. 2006), other polar species in this genera have retained this ability which allows them to photosynthesise under high light conditions. *Chlamydomonas sp*. ICE-L, shows an up-

regulation of light harvesting complex (LHC) genes LhcSR1 and LhcSR2, accompanied by increase in NPQ following high light, UV-B radiation and high salinity. This suggests that these LHC genes play a role in stress response and energy dissipation (Mou et al. 2012). In order to mitigate photoinhibition, a psychrotolerant *Chlorella sp*. isolated from Arctic glacier melt water, decreases the size of its light-harvesting complex (Cao et al. 2016).

Another interesting similarity between diatoms (psychrophilic and temperate) and *C. raudensis* is the biochemistry and architecture of the thylakoid membrane. Diatoms, as well as *C. raudensis*, have high concentrations of poly-unsaturated fatty acids in their thylakoid lipids classes and their thylakoid membranes are not organized in grana and stroma (Mock and Kroon 2002a; Mock and Kroon 2002b; Morgan-Kiss et al. 2006). This possibly means that looser membrane stacks in *C. raudensis* and homogeneously folded membranes in diatom plastids promote energy spill over between photosystems and therefore light energy transfer between photosystems (Morgan-Kiss et al. 2006).

An increase in transcripts for omega-3 fatty acid desaturase (CiFAD3) was measured in a *Chlamydomonas sp*. ICE-L under both high (12 °C) and low temperatures (0 °C) compared a control at 6 °C, as well as at high salinity (Zhang et al. 2011; An et al. 2013). This suggests that PUFAs may also play a role in heat stress and high salinity acclimation. Consumption of PUFAs was also observed in the same species during darkness (Xu et al. 2014), indicating that PUFAs are an important aspect of adaptation to several extreme conditions found in the polar regions. As with diatoms, antioxidants also play an important role in cold-shock adaptation, as seen in an Antarctic *Chlamydomonas sp*. in which an increase in glutathione S-transferase was observed (Kan et al. 2006).

The snow alga *C. nivalis* is exposed to the full spectrum of solar irradiance (UVC to infrared) and must therefore have a completely different photosynthesis performance compared to the low light adapted *C. raudensis* (Remias et al. 2005). The most striking difference between photosynthesis of both psychrophilic green algae is that *C. nivalis* does not seem to be inhibited by high solar irradiances. Even an exposure of cells to photon flux densities of 1,800 µmol photons m^{-2} s^{-1} for 40 min at 1.5 °C did not inhibit net photosynthesis (Remias et al. 2005). This extreme photosynthetic performance is only possible by a change in the life cycle. A combination of factors may trigger the formation of immotile red hypnoblast stages that are most resistant to environmental changes (Muller et al. 1998; Remias et al. 2005).

The transformation into hypnoblasts is characterized by a substantial incorporation of sugars and lipids, and by the formation of esterified extraplastidal secondary carotenoids (Hoham and Duval 2001). The most important carotenoid is astaxanthin which is located in cytoplasmatic lipid globuli (Muller et al. 1998; Remias et al. 2005), and is assumed to be responsible for the high photostability and therefore the absence of photoinhibition under strong solar irradiance on top of snow (Remias et al. 2005). Mature hypnoblasts can contain about 20 times more of this pigment than chlorophyll a, where the astaxanthin is possibly acting as a filter to reduce the irradiance that would otherwise be

damaging to the photosynthetic activity inside the plastids. Exposure to UV-B in *Chlamydomonas sp*. ICE-L led to an increase in expression of heat shock protein 70 (Liu et al. 2010) which suggests a role in protection against high irradiance.

High solar irradiance is not the only harsh condition on top of snow. Drought due to freezing of water is another main stress on the hypnoblast stages of *Chlamydomonas nivalis*. Like cacti in the desert, these stages have very rigid cell walls as the outer boundary to an extreme environment (Muller et al. 1998; Remias et al. 2005). Sometimes cells secrete carbohydrates to produce a visible mucilage sheet around them (Muller et al. 1998). These carbohydrates are not only attractive to bacteria that use them as a substrate but they also trap particles transported into the snow by wind. These particle-covered cells increase the absorption of solar irradiance and therefore the production of heat. This heat might cause melting of surrounding snow crystals and therefore provide liquid water to the cells (Takeuchi 2002). Such small spots of melt events around warm bodies (e.g. rock debris, cells) are called cryoconite holes (Takeuchi 2002). However, these adhering particles may also shade and thus protect *C. nivalis* against high irradiance. This is not universal and hypnoblasts from *C. nivalis*, for example, never show such attached structures.

Chemical reactions are influenced by temperature according to the relationship described by Arrhenius. In general, a 10 °C reduction in growth temperature causes biochemical reaction rates to decline two to three times. However, doubling times of psychrophilic algae can be comparable to mesophilic algae (Sommer 1989) which means that rates of enzyme catalyzed reactions must be optimized to low temperatures in these organisms (Feller and Gerday 2003). Studies with the enzyme nitrate reductase (NR), for instance, showed that these enzymes from psychrophilic algae possess structural modifications that make them more cold adapted, being more catalytically efficient at lower temperatures but at the same time less thermally stable, than NRs from mesophilic species (Di Martino Rigano et al. 2006). It also appears that light and salinity may influence nitrogen metabolism in *Chlamydomonas sp*. ICE-L (Wang et al. 2015).

In contrast to NR, the temperature maximum for carboxylase activity of ribulose-1–5-bisphosphatecarboxylase/oxygenase (RUBISCO), one of the most critical enzymes for inorganic carbon fixation in photoautotrophes, was not altered in some psychrophilic green algae and the specific activity at low temperatures was actually lower in the psychrophilic if compared to the mesophilic Rubisco (Devos et al. 1998). Decreased catalytic efficiency of these RUBISCOs under low temperature seems to be at least partly compensated by an increased cellular concentration of the protein. This is supported by the presence of RUBISCO as the fifth most abundant EST in *P. gelidicola* (Jung et al. 2012). An increase in ribosomal proteins seen at colder temperatures may counteract reduced efficiencies in translation (Toseland et al. 2013), alternatively it may help to cope with up-regulation of proteins due to reduced activity. Cao et al. (2016) found that a strain of arctic *Chlorella* increased both proteins and lipids at lower temperatures.

Expression and secretion of ice-binding proteins in polar Chlorophytes helps to maintain a fluid environment and reduce damage from ice crystals. Studies which look at IBPs through recombinant proteins and culture supernatant have demonstrated functions including changes in ice morphology, ice pitting, recrystallization inhibition and the creation of smaller brine pockets which helps to maintain salinity (Raymond et al. 2009; Raymond and Kim 2012; Raymond and Morgan-Kiss 2013; Jung et al. 2014).

As with diatoms, molecular tools are constantly improving for Chlorophytes, including techniques for gene editing and expression such as TALEs (Gao et al. 2014) and CRISPR-Cas (Shin et al. 2016; Wang et al. 2016). So far only temperate green algae have been selected for targeted gene knock-out, but as molecular tools such as these become available in their polar counter-parts, the potential to discover the function and role of important genes and pathways in polar adaptation drastically increases.

14.4 Conclusions

The application of omics approaches in combination with biochemical and physiological measurements has revealed unique adpatations in polar microalgae. Unsurprisingly, there is evidence that the extreme and highly variable conditions in polar ecosystems were driving those adaptations. While some of these adaptations (e.g. allelic divergence, gene duplications) were the consequence of mutations and subsequent diversification, others were based on biotic interactions that enabled transfer of genes (e.g. ice-binding) between different species and therefore the entire community to thrive under the extreme conditions of polar ecosystems. These mechanisms of adaptive evolution are not unique to polar microalgae but how they are used to produce unique phenotypes required to survive temperatures below freezing, long periods of darkness, strong seasonality and fluctuations in nutrients and salinity is still unknown. Once we have obtained genetically tractable model species such as *Fragilariopsis cylindrus* and *Chlamydomonas sp*. ICE, we'll be able to better understand how genotypes impact phenotypes that matter to thrive under polar conditions. With these model species, we will be able to test, through experimental evolution approaches, how their populations respond to global wamring, which is still largely unknown. Results from these model species can be used to inform studies on natural populations (e.g. barcoding, metatranscriptomes and metagenomes) in terms of identifying their standing pool of genetic variation and evolutionary potential to respond to global warming. Identification of genetic diversity in these organisms not only provides new insights into their evolution and adaptation, but also contributes to extend the pool of marine genetic resources, which so far is dominated by genes and their products from non-polar organisms.

References

Allen, A.E., Dupont, C.L., Oborník, M., Horák, A., Nunes-Nesi, A., McCrow, J.P., Zheng, H., Johnson, D. a, Hu, H., Fernie, A.R., Bowler, C., 2011. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. Nature 473, 203–207. doi:10.1038/nature10074

Allen, D.J., Ort, D.R., 2001. Impacts of chilling temperatures on photosynthesis in warm-climate plants. Trends Plant Sci. 6, 36–42. doi:10.1016/S1360-1385(00)01808-2

An, M., Mou, S., Zhang, X., Zheng, Z., Ye, N., Wang, D., Zhang, W., Miao, J., 2013. Expression of fatty acid desaturase genes and fatty acid accumulation in Chlamydomonas sp. ICE-L under salt stress. Bioresour. Technol. 149, 77–83. doi:10.1016/j.biortech.2013.09.027

Armbrust, E. V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., Brzezinski, M.A., Chaal, B.K., Chiovitti, A., Davis, A.K., Demarest, M.S., Detter, J.C., Glavina, T., Goodstein, D, D., 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science (80-.). 306, 79–86.

Aslam, S.N., Cresswell-Maynard, T., Thomas, D.N., Underwood, G.J.C., 2012. Production and Characterization of the Intra- and Extracellular Carbohydrates and Polymeric Substances (Eps) of Three Sea-Ice Diatom Species, and Evidence for a Cryoprotective Role for Eps. J. Phycol. 48, 1494–1509. doi:10.1111/jpy.12004

Bayer-Giraldi, M., Uhlig, C., John, U., Mock, T., Valentin, K., 2010. Antifreeze proteins in polar sea ice diatoms: Diversity and gene expression in the genus Fragilariopsis. Environ. Microbiol. 12, 1041–1052. doi:10.1111/j.1462-2920.2009.02149.x

Bayer-Giraldi, M., Weikusat, I., Besir, H., Dieckmann, G., 2011. Characterization of an antifreeze protein from the polar diatom Fragilariopsis cylindrus and its relevance in sea ice. Cryobiology 63, 210–219. doi:10.1016/j.cryobiol.2011.08.006

Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Brueggeman, A., Dunigan, D.D., Gurnon, J., Ladunga, I., Lindquist, E., Lucas, S., Pangilinan, J., Pröschold, T., Salamov, A., Schmutz, J., Weeks, D., Yamada, T., Lomsadze, A., Borodovsky, M., Claverie, J.-M., Grigoriev, I. V, Van Etten, J.L., 2012. The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. Genome Biol. 13, R39. doi:10.1186/gb-2012-13-5-r39

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P. and Rayko, E., 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 456, 239-244.

Boyd, P.W., 2002. Review of environmental factors controlling phytoplankton processes in the Southern Ocean 1. J. Phycol. 38, 844–861. doi:10.1046/j.1529-8817.2002.t01-1-01203.x

Cao, K., He, M., Yang, W., Chen, B., Luo, W., Zou, S., Wang, C., 2016. The eurythermal adaptivity and temperature tolerance of a newly isolated psychrotolerant Arctic Chlorella sp. J. Appl. Phycol. 28, 877–888. doi:10.1007/s10811-015-0627-0

Cheng, C.H.C., 1998. Evolution of the diverse antifreeze proteins. Curr. Opin. Genet. Dev. 8, 715–720. doi:10.1016/S0959-437X(98)80042-7

Cota, G.F., 1985. Photoadaptation of high Arctic ice algae. Nature 315, 219–222.

Croft, M.T., Lawrence, A.D., Raux-Deery, E., Warren, M.J., Smith, A.G., 2005. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. Nature 438, 90–3. doi:10.1038/nature04056

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

Devos, N., Ingouff, M., Loppes, R., Matagne, R.F., 1998. Rubisco adaptation to low temperatures: a comparative study in psychrophilic and mesophilic unicellular algae. J. Phycol. 34, 655–660. doi:10.1046/j.1529-8817.1998.340655.x

Ditullio, G.R., Garrison, D.L., Mathot, S.., 1998. Dimethylsulfoniopropionate in sea ice algae from the Ross Sea polynya. Antarct. Sea Ice Biol. Process. Interact. Var. 139–146.

Feller, G., Gerday, C., 2003. Psychrophilic enzymes; hot topics in cold adaptation. Nat. Rev. Microbiol. 1, 200–208.

Fiala, M., Oriol, L., 1990. Light-Temperature Interactions on the Growth of Antarctic Diatoms 629–636.

Fritsen, C.H., Priscu, J.C., 1999. Seasonal change in the optical properties of the permanent ice cover on Lake Bonney, Antarctica: consequences for lake productivity and phytoplankton dynamics. Limnol. Oceanogr. 44, 447–454. doi:10.4319/lo.1999.44.2.0447

Fujita, Y., 2001. Chromatic variation of the abundance of PSII complexes observed with the red alga Prophyridium cruentum. Plant Cell Physiol. 42, 1239–1244. doi:10.1093/pcp/pce164

Galachyants, Y.P., Zakharova, Y.R., Petrova, D.P., Morozov, A.A., Sidorov, I.A., Marchenkov, A.M., Logacheva, M.D., Markelov, M.L., Khabudaev, K.V., Likhoshway, Y.V. and Grachev, M.A., 2015. Sequencing of the complete genome of an araphid pennate diatom Synedra acus subsp. radians from Lake Baikal. In Doklady Biochemistry and Biophysics. 461(1), 84-88).

Gao, H., Wright, D. a., Li, T., Wang, Y., Horken, K., Weeks, D.P., Yang, B., Spalding, M.H., 2014. TALE activation of endogenous genes in Chlamydomonas reinhardtii. Algal Res. 5, 52–60. doi:10.1016/j.algal.2014.05.003

Gwak, I.G., sic Jung, W., Kim, H.J., Kang, S.H., Jin, E., 2010. Antifreeze Protein in Antarctic Marine Diatom, Chaetoceros neogracile. Mar. Biotechnol. 12, 630–639. doi:10.1007/s10126-009-9250-x

Hällfors, G., 2004. Checklist of Baltic Sea phytoplankton species (including some heterotrophic protistan groups). Balt. Sea Environ. Proc. 95, 208.

Hendey, N.I., 1974. A REVISED CHECK-LIST OF BRITISH DIATOMS. J. Mar. Biol. Assoc. United Kingdom 54, 277–300.

Hoham, R.W., Duval, B., 2001. Microbial ecology of snow and freshwater ice with emphasis on snow algae, in: Jones, H.G., Pomeroy, J.W., Walker, D.A., Hoham, R.W. (Eds.), Snow Ecology: An Interdisciplinary Examination of Snow-Covered Ecosystems. Cambridge University Press, Cambridge, pp. 168–228.

Hopes, A., Nekrasov, V., Kamoun, S., Mock, T., 2016. Editing of the urease gene by CRISPR-Cas in the diatom Thalassiosira pseudonana. Plant Methods 12, 49–60. doi:10.1186/s13007-016-0148-0

Hwang, Y.S., Jung, G., Jin, E., 2008. Transcriptome analysis of acclimatory responses to thermal stress in Antarctic algae. Biochem. Biophys. Res. Commun. 367, 635–641. doi:10.1016/j.bbrc.2007.12.176

Janech, M.G., Krell, A., Mock, T., Kang, J.S., Raymond, J. a., 2006. Ice-binding proteins from sea ice diatoms (Bacillariophyceae). J. Phycol. 42, 410–416. doi:10.1111/j.1529-8817.2006.00208.x

Janknegt, P.J., Van De Poll, W.H., Visser, R.J.W., Rijstenbil, J.W., Buma, a. G.J., 2008. Oxidative stress responses in the marine antarctic diatom Chaetoceros brevis (Bacillariophyceae) during photoacclimation. J. Phycol. 44, 957–966. doi:10.1111/j.1529-8817.2008.00553.x

Jung, W., Gwak, Y., Davies, P.L., Kim, H.J., Jin, E.S., 2014. Isolation and Characterization of Antifreeze Proteins from the Antarctic Marine Microalga Pyramimonas gelidicola. Mar. Biotechnol. 16, 502–512. doi:10.1007/s10126-014-9567-y

Jung, W., Lee, S.G., Kang, S.W., Lee, Y.S., Lee, J.H., Kang, S.H., Jin, E.S., Kim, H.J., 2012. Analysis of expressed sequence tags from the Antarctic psychrophilic green algae, Pyramimonas gelidicola. J. Microbiol. Biotechnol. 22, 902–906. doi:10.4014/jmb.1201.01002

Junge, K., Eicken, H., Deming, J.W., 2004. Bacterial Activity at -2 to -20 degrees C in Arctic wintertime sea ice. Appl. Environ. Microbiol. 70, 550–557. doi:10.1128/AEM.70.1.550

Kan, G.F., Miao, J.L., Shi, C.J., Li, G.Y., 2006. Proteomic alterations of antarctic ice microalga Chlamydomonas sp. under low-temperature stress. J. Integr. Plant Biol. 48, 965–970. doi:10.1111/j.1744-7909.2006.00255.x

Krell, A., Beszteri, B., Dieckmann, G., Glöckner, G., Valentin, K., Mock, T., 2008. A new class of ice-binding proteins discovered in a salt-stress-induced cDNA library of the psychrophilic diatom *Fragilariopsis cylindrus* (Bacillariophyceae). Eur. J. Phycol. 43, 423–433. doi:10.1080/09670260802348615

Krembs, C., Eicken, H., Deming, J.W., 2011. Exopolymer alteration of physical properties of sea ice and implications for ice habitability and biogeochemistry in a warmer Arctic. Proc. Natl. Acad. Sci. U. S. A. 108, 3653–3658. doi:10.1073/pnas.1100701108

Liu, S., Zhang, P., Cong, B., Liu, C., Lin, X., Shen, J., Huang, X., 2010. Molecular cloning and expression analysis of a cytosolic Hsp70 gene from Antarctic ice algae Chlamydomonas sp. ICE-L. Extremophiles 14, 329–337. doi:10.1007/s00792-010-0313-8

Liu, X., Wang, Y., Gao, H., Xu, X., 2011. Identification and characterization of genes encoding two novel LEA proteins in Antarctic and temperate strains of Chlorella vulgaris. Gene 482, 51–58. doi:10.1016/j.gene.2011.05.006

Lizotte, M.P., Priscu, J.., 1992. Spectral irradiance and biooptical properties in perennial ice-covered lakes of the dry valleys (McMurdo Sound Antarctica). Antarct. Res. Ser. 57, 1–14.

Lommer, M., Specht, M., Roy, A.S., Kraemer, L., Andreson, R., Gutowska, M.A., Wolf, J., Bergner, S.V., Schilhabel, M.B., Klostermeier, U.C. and Beiko, R.G., 2012. Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome biology, 13(7), R66.

Los, D. a., Murata, N., 2004. Membrane fluidity and its roles in the perception of environmental signals. Biochim. Biophys. Acta - Biomembr. 1666, 142–157. doi:10.1016/j.bbamem.2004.08.002

Lyon, B., Mock, T., 2014. Polar Microalgae: New Approaches towards Understanding Adaptations to an Extreme and Changing Environment. Biology (Basel). 3, 56–80. doi:10.3390/biology3010056

Lyon, B.R., Lee, P. a., Bennett, J.M., DiTullio, G.R., Janech, M.G., 2011. Proteomic analysis of a sea-ice diatom: salinity acclimation provides new insight into the dimethylsulfoniopropionate production pathway. Plant Physiol. 157, 1926–1941. doi:10.1104/pp.111.185025

Mock, T., Hoch, N., 2005. Long-term temperature acclimation of photosynthesis in steady-state cultures of the polar diatom Fragilariopsis cylindrus. Photosynth. Res. 85, 307–317. doi:10.1007/s11120-005-5668-9

Mock, T., Kroon, B.M. a, 2002a. Photosynthetic energy conversion under extreme conditions—I: important role of lipids as structural modulators and energy sink under N -limited growth in Antarctic sea ice diatoms. Phytochemistry 61, 41–51. doi:10.1016/S0031-9422(02)00216-9

Mock, T., Kroon, B.M. a, 2002b. Photosynthetic energy conversion under extreme conditions—II: the significance of lipids under light limited growth in Antarctic sea ice diatoms. Phytochemistry 61, 53–60. doi:10.1016/S0031-9422(02)00216-9

Mock, T., Otillar, R.P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., Ward, B.J., Allen, A.E., Dupont, C.L., Frickenhaus, S., Maumus, F., Veluchamy, A., Wu, T., Barry, K.W., Falciatore, A., Ferrante, M.I., Fortunato, A.E., Glöckner, G., Gruber, A., Hipkin, R., Janech, M.G., Kroth, P.G., Leese, F., Lindquist, E. a., Lyon, B.R., Martin, J., Mayer, C., Parker, M., Quesneville, H., Raymond, J. a., Uhlig, C., Valas, R.E., Valentin, K.U., Worden, A.Z., Armbrust, E.V., Clark, M.D., Bowler, C., Green, B.R., Moulton, V., van Oosterhout, C., Grigoriev, I. V., 2017. Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature 541, 536–540. doi:10.1038/nature20803

Mock, T., Valentin, K., 2004. Photosynthesis and cold acclimation: Molecular evidence from a polar diatom. J. Phycol. 40, 732–741. doi:10.1111/j.1529-8817.2004.03224.x

Morgan-Kiss, R.M., Ivanov, A.G., Pocock, T., Kr??l, M., Gudynaite-Savitch, L., H??ner, N.P. a, 2005. The antarctic psychrophile, Chlamydomonas raudensis Ettl (UWO241) (Chlorophyceae, Chlorophyta), exhibits a limited capacity to photoacclimate to red light. J. Phycol. 41, 791–800. doi:10.1111/j.1529-8817.2005.04174.x

Morgan-Kiss, R.M., Priscu, J.C., Pocock, T., Gudynaite-Savitch, L., Huner, N.P. a, 2006. Adaptation and Acclimation of Photosynthetic Microorganisms to Permanently Cold Environments. Microbiol. Mol. Biol. Rev. 70 , 222–252. doi:10.1128/MMBR.70.1.222-252.2006

Mou, S., Zhang, X., Ye, N., Dong, M., Liang, C., Liang, Q., Miao, J., Xu, D., Zheng, Z., 2012. Cloning and expression analysis of two different LhcSR genes involved in stress adaptation in an Antarctic microalga, Chlamydomonas sp. ICE-L. Extremophiles 16, 193–203. doi:10.1007/s00792-011-0419-7

Müller, T., Bleiß, W., Martin, C.D., Rogaschewski, S., Fuhr, G., 1998. Snow algae from northwest Svalbard: Their identification, distribution, pigment and nutrient content. Polar Biol. 20, 14–32. doi:10.1007/s003000050272

Nishida, I., Murata, N., 1996. Chilling sensitivity in plants and cyanobacteria: the crucial contribution of membrane lipids. Annu. Rev. Plant Physiol. Plant Mol. Biol. 47, 541–568. doi:10.1146/annurev.arplant.47.1.541

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

Palmisano, A.C., Garrison, D.L., 1993. Microorganisms in Antarctic sea ice, in: Friedmann, E. (Ed.), Antarctic Microbiology. Wiley-Liss, New York, pp. 167–218.

Park, S., Jung, G., Hwang, Y.S., Jin, E., 2010. Dynamic response of the transcriptome of a psychrophilic diatom, Chaetoceros neogracile, to high irradiance. Planta 231, 349–360. doi:10.1007/s00425-009-1044-x

Ralph, P.J., McMinn, A., Ryan, K.G., Ashworth, C., 2005. Short-term effect of temperature on the photokinetics of microalgae from the surface layers of Antarctic pack ice. J. Phycol. 41, 763–769. doi:10.1111/j.1529-8817.2005.00106.x

Raymond, J. a, 2011. Algal ice-binding proteins change the structure of sea ice. Proc. Natl. Acad. Sci. 108, E198. doi:10.1073/pnas.1106288108

Raymond, J. a., 2014. The ice-binding proteins of a snow alga, Chloromonas brevispina: probable acquisition by horizontal gene transfer. Extremophiles 18, 987–994. doi:10.1007/s00792-014-0668-3

Raymond, J. a., Fritsen, C., Shen, K., 2007. An ice-binding protein from an Antarctic sea ice bacterium. FEMS Microbiol. Ecol. 61, 214–221. doi:10.1111/j.1574-6941.2007.00345.x

Raymond, J. a., Janech, M.G., Fritsen, C.H., 2009. Novel ice-binding proteins from a psychrophilic antarctic alga (chlamydomonadaceae, chlorophyceae). J. Phycol. 45, 130–136. doi:10.1111/j.1529-8817.2008.00623.x

Raymond, J. a., Kim, H.J., 2012. Possible role of horizontal gene transfer in the colonization of sea ice by Algae. PLoS One 7, 35968. doi:10.1371/journal.pone.0035968

Raymond, J. a., Morgan-Kiss, R., 2013. Separate Origins of Ice-Binding Proteins in Antarctic Chlamydomonas Species. PLoS One 8, e59186. doi:10.1371/journal.pone.0059186

Remias, D., Lutz-Meindl, U., Lutz, C., 2005. Photosynthesis, pigments and ultrastructure of the alpine snow alga Chlamydomonas nivalis. Eur. J. Phycol. 40, 259–268. doi:10.1080/09670260500202148

Ryan, K.G., Ralph, P., McMinn, a., 2004. Acclimation of Antarctic bottom-ice algal communities to lowered salinities during melting. Polar Biol. 27, 679–686. doi:10.1007/s00300-004-0636-y

Schriek, R., 2000. Effects of light and temperature on the enzymatic antioxidative defense systems in the Antarctic ice diatom Entomoneis kufferathii Manguin. Reports Polar Res. 349, 1–130.

Shi, H., Lee, B., Wu, S.-J., Zhu, J.-K., 2003. Overexpression of a plasma membrane Na+/H+ antiporter gene improves salt tolerance in Arabidopsis thaliana. Nat. Biotechnol. 21, 81–85. doi:10.1038/nbt766

Shin, S.-E., Lim, J.-M., Koh, H.G., Kim, E.K., Kang, N.K., Jeon, S., Kwon, S., Shin, W.-S., Lee, B., Hwangbo, K., Kim, J., Ye, S.H., Yun, J.-Y., Seo, H., Oh, H.-M., Kim, K.-J., Kim, J.-S., Jeong, W.-J., Chang, Y.K., Jeong, B., 2016. CRISPR/Cas9-induced knockout and knock-in mutations in Chlamydomonas reinhardtii. Sci. Rep. 6, 27810. doi:10.1038/srep27810

Sommer, U., 1989. Maximum growth rates of Antarctic phytoplankton: only weak dependence on cell size. Limnol. Oceanogr. 34, 1109–1112.

Sorhannus, U., 2011. Evolution of antifreeze protein genes in the diatom genus Fragilariopsis: Evidence for horizontal gene transfer, gene duplication and episodic diversifying selection. Evol. Bioinforma. 2011, 279–289. doi:10.4137/EBO.S8321

Strauss, J., Gao, S., Morrissey, J., Bowler, C., Nagel, G., Mock, T., 2013. A light-driven rhodopsin proton pump from the psychrophilic diatom Fragilariopsis cylindrus, in: In Proceeding of EMBO Workshop: The Molecular Life of Diatoms, Paris, France.

Streb, P., Shang, W., Feierabend, J., Bligny, R., 1998. Divergent strategies of photoprotection in high-mountain plants. Planta 207, 313–324. doi:10.1007/s004250050488

Takeuchi, N., 2002. Optical characteristics of cryoconite (surface dust) on glaciers: the relationship between light absorbency and the property of organic matter contained in the cryoconite. Ann. Glaciol. 34, 409–414.

Tanaka, T., Maeda, Y., Veluchamy, A., Tanaka, M., Abida, H., Maréchal, E., Bowler, C., Muto, M., Sunaga, Y., Tanaka, M. and Yoshino, T., 2015. Oil accumulation by the oleaginous diatom Fistulifera solaris as revealed by the genome and transcriptome. The Plant Cell. 27(1), 162-176.

Tang, E.P.Y., Vincent, W.F., Proulx, D., Lessard, P., Noue, J. d. L., 1997. Polar cyanobacteria versus green algae for tertiary wast-water treatment in cool climates. J. Appl. Phycol. 9, 371–381.

Teoh, M.L., Phang, S.M., Chu, W.L., 2013. Response of Antarctic, temperate, and tropical microalgae to temperature stress. J. Appl. Phycol. 25, 285–297. doi:10.1007/s10811-012-9863-8

Toseland, A., Daines, S.J., Clark, J.R., Kirkham, A., Strauss, J., Uhlig, C., Lenton, T.M., Valentin, K., Pearson, G. a., Moulton, V., Mock, T., 2013. The impact of temperature on marine phytoplankton resource allocation and metabolism. Nat. Clim. Chang. 3, 979–984. doi:10.1038/nclimate1989

Uhlig, C., Kilpert, F., Frickenhaus, S., Kegel, J.U., Krell, A., Mock, T., Valentin, K., Beszteri, B., 2015. In situ expression of eukaryotic ice-binding proteins in microbial communities of Arctic and Antarctic sea ice. ISME J. 9, 2537–2540. doi:10.1038/ismej.2015.43

Underwood, G.J.C., Aslam, S.N., Michel, C., Niemi, A., Norman, L., Meiners, K.M., Laybourn-Parry, J., Paterson, H., Thomas, D.N., 2013. Broad-scale predictability of carbohydrates and exopolymers in Antarctic and Arctic sea ice. Proc. Natl. Acad. Sci. U. S. A. 110, 15734–9. doi:10.1073/pnas.1302870110

Underwood, G.J.C., Fietz, S., Papadimitriou, S., Thomas, D.N., Dieckmann, G.S., 2010. Distribution and composition of dissolved extracellular polymeric substances (EPS) in Antarctic sea ice. Mar. Ecol. Prog. Ser. 404, 1–19. doi:10.3354/meps08557

Van Oijen, T., van Leeuwe, M. a., Gieskes, W.W.C., 2003. Variation of particulate carbohydrate pools over time and depth in a diatom-dominated plankton community at the Antarctic Polar Front. Polar Biol. 26, 195–201. doi:10.1007/s00300-002-0456-x

Von Quillfeldt, C.H., 2004. The diatom Fragilariopsis cylindrus and its potential as an indicator species for cold water rather than for sea ice 54, 137–143.

Wang, D.S., Xu, D., Wang, Y.T., Fan, X., Ye, N.H., Wang, W.Q., Zhang, X.W., Mou, S.L., Guan, Z., 2015. Adaptation involved in nitrogen metabolism in sea ice alga Chlamydomonas sp.

ICE-L to Antarctic extreme environments. J. Appl. Phycol. 27, 787–796. doi:10.1007/s10811-014-0372-9

Wang, Q., Lu, Y., Xin, Y., Wei, L., Huang, S., Xu, J., 2016. Genome editing of model oleaginous microalgae Nannochloropsis spp. by CRISPR/Cas9. Plant J. 1071–1081. doi:10.1111/tpj.13307

Willem, S., Srahna, M., Devos, N., Gerday, C., Loppes, R., Matagne, R.F., 1999. Protein adaptation to low temperatures: A comparative study of α-tubulin sequences in mesophilic and psychrophilic algae. Extremophiles 3, 221–226. doi:10.1007/s007920050119

Williams, W.E., Gorton, H.L., Vogelmann, T.C., 2003. Surface gas-exchange processes of snow algae. Proc. Natl. Acad. Sci. U. S. A. 100, 562–6. doi:10.1073/pnas.0235560100

Xu, D., Wang, Y., Fan, X., Wang, D., Ye, N., Zhang, X., Mou, S., Guan, Z., Zhuang, Z., 2014. Long-Term Experiment on Physiological Responses to Synergetic. Environ. Sci. Technol. 48, 7738–7746.

Zhang, P., Liu, S., Cong, B., Wu, G., Liu, C., Lin, X., Shen, J., Huang, X., 2011. A Novel Omega-3 Fatty Acid Desaturase Involved in Acclimation Processes of Polar Condition from Antarctic Ice Algae Chlamydomonas sp. ICE-L. Mar. Biotechnol. 13, 393–401. doi:10.1007/s10126-010-9309-8

Chapter 2: Transformation of *Fragilariopsis cylindrus*

Introduction

As discussed in the general introduction *Fragilariopsis cylindrus* is an important species in polar ecosystems. Sequencing of the genome (Mock et al., 2017), analysis of transcripts (Mock et al., 2017) and analysis of expressed sequence tags (ESTs; (Krell, 2006; Mock et al., 2005) have already given us insight into some of the adaptations which allow this species to thrive in colder water and sea ice. A transformation system allows a deeper level of analysis through direct manipulation of genes and pathways, as well as the opportunity to utilise cells as hosts for production of recombinant proteins. As such, it is an important tool for understanding both the biology of a species and for biotechnology.

This chapter focuses on the development of a transformation system for *F. cylindrus*.

There are currently 13 transformable diatom species; 12 with stable nuclear transformation and one with transient expression. One species, *P. tricornutum*, also has methods for transforming plastids. An overview of species and methods can be found in table 2.1.

Figure 2.1 shows the geographical distribution of transformable diatom species in terms of recorded location according to Algaebase (Guiry and Guiry, 2017) and strain collection point. The majority of species are marine with only two from freshwater/brackish environments. Up until now all species have been temperate, with the majority of strains used for transformation collected around North-East America, Western Europe and East Asia. In contrast *F. cylindrus* is found in the Arctic, Antarctic and seasonally cold waters, with the strain used in this study isolated in the Southern Ocean near Antarctica. As far as I am aware this is the first transformation system for not only a psychrophilic diatom, but also a psychrophilic eukaryote, although systems do exist for psychrophilic bacteria (Duilio et al. 2004; Vigentini et al. 2006; Miyake et al. 2007).

Species are also clustered in terms of evolutionary background, with all belonging to two out of the four diatom classes (Figure 2.2). Even within classes, several species are clustered in order, particularly for Bacillariophyceae, with 8/9 species belonging to either the Bacillariales or Naviculales, including *F. cylindrus*. This demonstrates that mainly marine, temperate diatoms clustered within a few orders are represented. Considering that diatoms are exceptionally variable in terms of habitat, silicification, size, stress tolerance and evolution (See introduction chapter: Hopes & Mock 2014; Hopes & Mock 2015), developing transformation systems and molecular tools for a greater range of diatoms could help give a better understanding of this broad group of organisms.

Several different methods have been applied to transform diatoms. The most popular is microparticle bombardment which shoots micron sized gold or tungsten particles coated in DNA

into cells under high pressure in a vacuum. This method is often used to penetrate hard cell walls in plants and algae (Kindle et al., 1989; Qin et al., 2005; Taylor and Fauquet, 2002). This also facilitates delivery of DNA into diatoms through the silica frustule. This method was used to transform the first diatoms; *Cyclotella cryptica* and *Navicula saprophila*.

Electroporation is one of the most successful methods in terms of transformation efficiency in diatoms. It works by applying an electrical pulse or several pulses (multi-pulse electroporation) to permeabilise the cell membrane and allow entry of large molecules including DNA, RNA and proteins. However, the cell wall can interfere with delivery via this method (Azencott et al., 2007), which may explain why only lightly silicified diatoms have been transformed in this manner (Ifuku et al., 2015; Miyahara et al., 2013; Zhang and Hu, 2014).

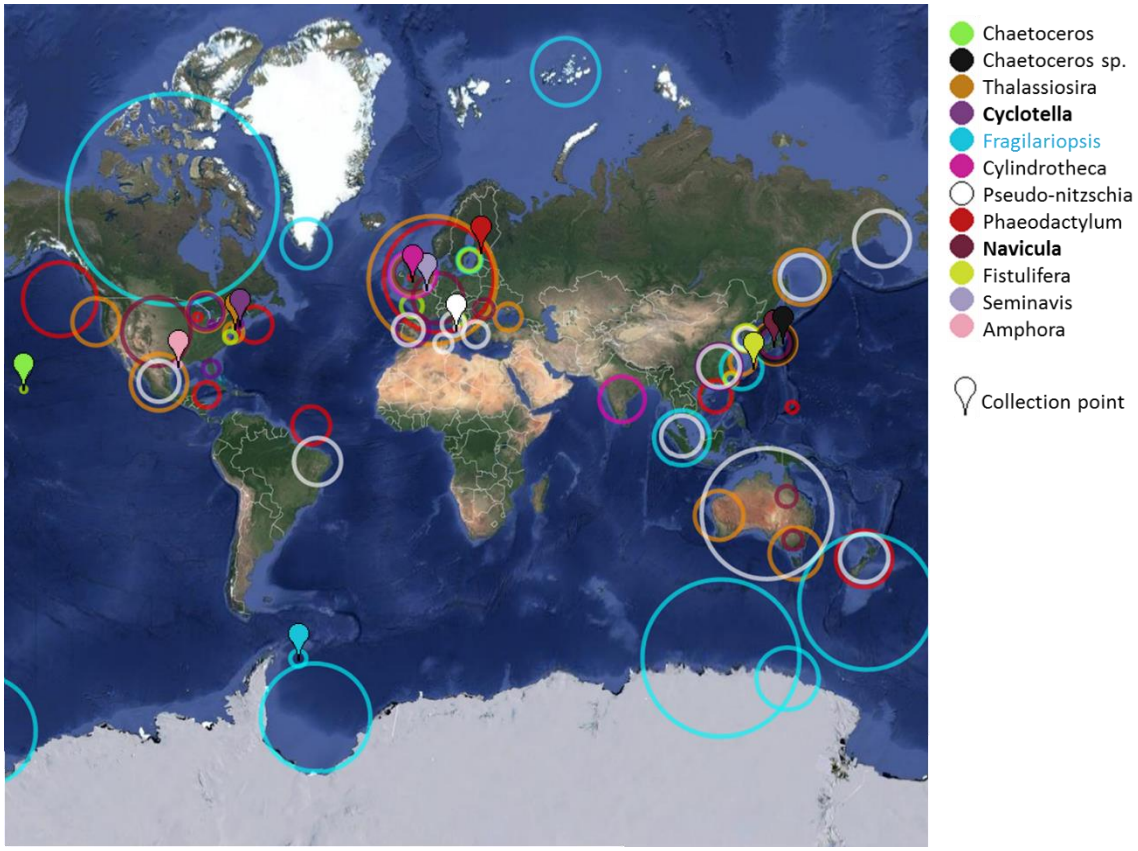


Figure 2.1. Map showing recorded location and collection points for different diatom genera with transformation systems. Circles indicate recorded locations (Guiry and Guiry, 2017). Coloured pins represent collection points for each strain transformed (see table 2.1 for references). Mesophilic species are shown in black, psychrophilic *F. cylindrus* is highlighted in blue and freshwater/brackish species are in bold text. Initial map created with Scribble maps.

Bacterial conjugation has only recently been applied to diatoms (Karas et al., 2015). It involves delivery of the plasmid through a bacterial intermediate. The cargo plasmid carrying an origin of transfer (oriT) and a conjugation plasmid are transformed in to *E. coli* which then delivers the cargo plasmid to the diatom. This method has been applied to two model diatom species: the lightly

silicified *P. tricornutum* and the heavily silicified *T. pseudonana*. Transformation efficiencies from this method compare well with electroporation. Transformation with Polyethylene glycol was also tested (Karas et al., 2015) but gave the lowest transformant yield out of all the methods currently available for diatom transformation.

All bar one species, with the exception of *Chaetoceros gracilis* (Ifuku et al., 2015) and all genera of transformable diatoms have been successfully modified following microparticle bombardment. As a result of the well-characterised protocols and parameters developed across a range of diatom species, this method was chosen to transform *F. cylindrus*.

A toolbox with a variety of promoters, selective markers and reporter genes is available for expression of transgenes in diatoms (table 2.1).

Typically constitutive endogenous promoters from genes such as fucoxanthin chlorophyll a/c binding protein (FCP), acetyl coenzyme A (acetyl-CoA), and histone 4 (H4) have been used to express transgenes, although exogenous promoters from other diatom species (Buhmann et al., 2014; Dunahay et al., 1995; Miyagawa et al., 2009; Miyahara et al., 2013; Sabatino et al., 2015) and viral promoters (Muto et al., 2013; Sakaue et al., 2008) are also common. One inducible promoter from the nitrate reductase gene has been identified for a range of species (Ifuku et al., 2015; Niu et al., 2012; Poulsen and Kröger, 2005; Poulsen et al., 2006) allowing expression to be activated through the addition of nitrate. Expression can also be altered in response to light levels through the use of the FCP promoter (Leblanc et al., 1999; Siaut et al., 2007). To remove this variation the promoter from elongation factor 2 (EF2) was used in *P. tricornutum* for more consistent expression under variable light conditions (Seo et al., 2015).

There is evidence to suggest that endogenous promoters may lead to higher numbers of transformants (Muto et al., 2013) or exogenous promoters to low numbers of colonies (Miyagawa-Yamaguchi et al., 2011). However, higher expression levels have been observed with viral promoters compared to endogenous promoters (Sakaue et al., 2008). In this study an endogenous FCP promoter linked to high expression was chosen based on RNA-seq data and previous success of FCP promoters in the majority of transformable diatoms.

There are several antibiotics that are effective in diatoms, which along with their respective resistance genes can be used to select positive transformants. Species react differently to different antibiotics and their concentrations (Apt et al., 1996; Sabatino et al., 2015) and not all selectable markers are functional or expressed in high enough quantities to counteract antibiotic potency (Poulsen et al., 2006). Furthermore, conditions such as salinity may alter antibiotic activity (Falciatore et al., 1999; Muto et al., 2013) and not all species grow well on plates (Sabatino et al., 2015). This means that antibiotics and markers need to be empirically tested for each species in both liquid media and on plates. Previous work has shown that zeocin, an antibiotic that prevents growth by intercalating to and breaking DNA, is effective against *F. cylindrus* in liquid media at 100µg/ml (Strauss, unpublished).

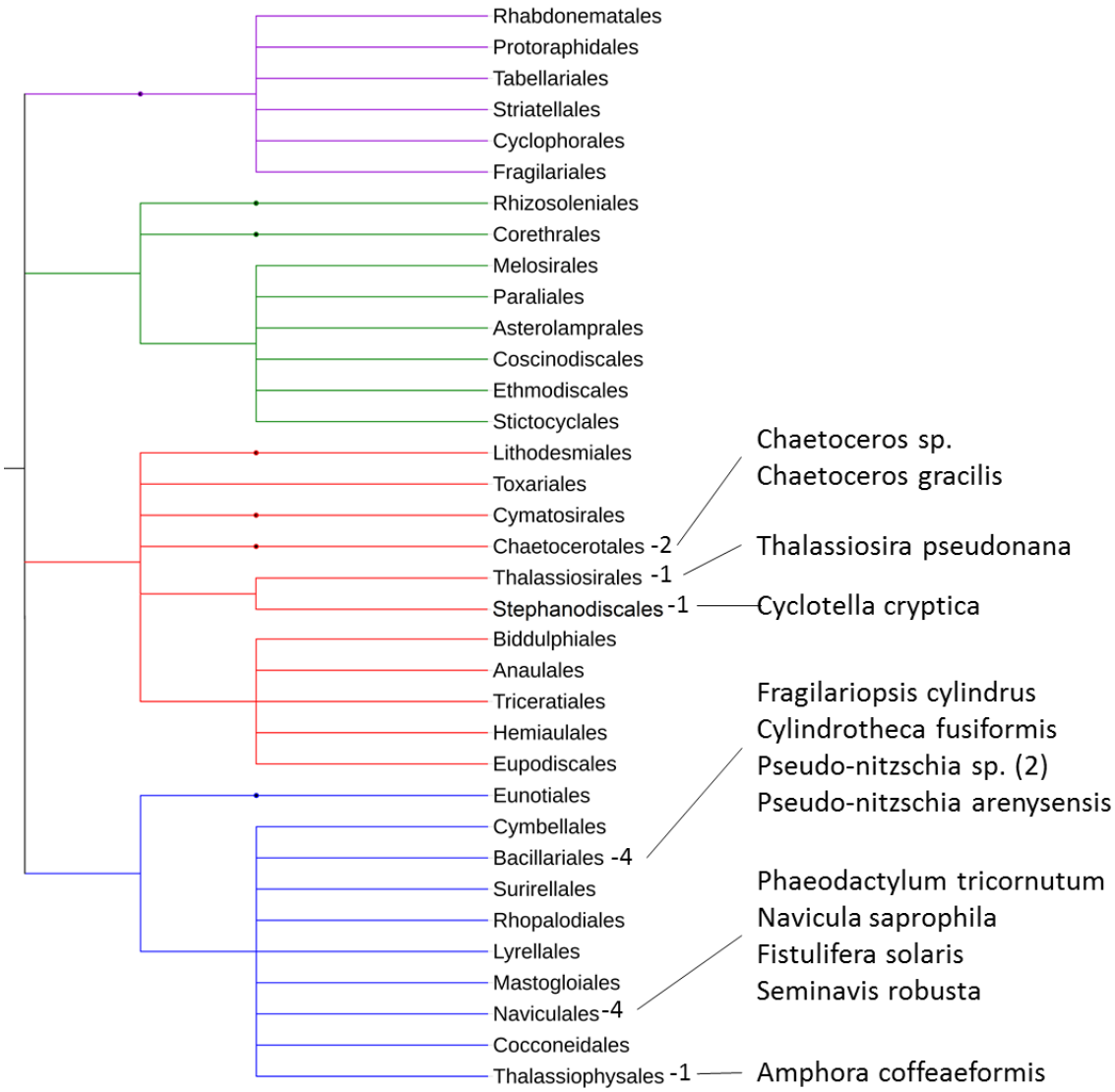


Figure 2.2. Transformable diatoms by phylogeny. Number of transformable diatom species is indicated adjacent to the order. Clades are coloured to represent each class: Fragilariophyceae/ araphid pennate diatoms (Purple), Coscinodiscophyceae/ radial centric diatoms (Green), Mediophyceae/ polar centric diatoms (Red) and Bacillariophyceae/ raphid pennate diatoms (Blue). Diatoms orders were sourced from algaebase (Guiry) and plotted using phyloT and iTOL (http://phylot.biobyte.de/) which links to data from NCBI, orders Chaetocerotales and Thalassiosirales were moved to the class Mediophyceae, according to Medlin and Kaczmarska (2004).

Species	Method	Promoter	Terminator	Reporter	Selective marker	Efficiency/ 10^8 cells	Efficiency/ µg plasmid	Ref
Amphora coffeaeformis	MPB	PTfcpB/ PTfcpA	PTfcpA	eyfp	shble/nat	800	800	Buhmann et al. 2014
Chaetoceros gracilis	MPE	fcp/NR/ acetyl-CoA	fcp	luc, Azami-Green	nat	400	69.6	Ifuku et al. 2015
Chaetoceros sp.	MPB	TPfcp/ TPNR	TPfcp/ TPNR		nat	16.2	8.1	Miyagawa-Yamaguchi et al. 2011
Cyclotella cryptica	MPB	CCacetyl-CoA	CCacetyl-CoA	-	nptII	28	8.5	Dunahay et al. 1995
Cylindrotheca fusiformis	MPB	fruα3	fruα2/ hepA	-	shble	200-500	20-50	Fisher et al. 1999
Cylindrotheca fusiformis	MPB	fcpA/ NR	fcpA/ NR	egfp	shble	360	36	Poulsen & Kröger 2005
Fistulifera solaris	MPB	RSV/CaMV35S/ PTfcpA/H4/ fcpB	PTfcpA	gfp	nptII	23	11.5	Muto et al. 2013
Fragilariopsis cylindrus	MPB	fcp	fcp	egfp	shble	30	15	This thesis
Navicula saprophila (fistifulifera)	MPB	CCacetyl-CoA	CCacetyl-CoA	-	nptII	2.8	8.4	Dunahay et al. 1995
Phaeodactylum tricornutum	MPB	fcpA-fcpE	fcpA	-	shble/ cat	83	103.75	Apt et al. 1996
Phaeodactylum tricornutum	MPB	fcpB/fcpF	fcpA	LUC	shble	238	23.8	Falciatore et al. 1999
Phaeodactylum tricornutum	MPB	fcpA /fcpB	fcpA	GUS/ gfp	nat/ sat-1/ nptII/ shble	10-20	6.25-12.5	Zaslavskaia et al. 2000
Phaeodactylum tricornutum	MPB	fcpA/fcpB	fcpA	egfp	shble	-	-	Zaslavskaia et al. 2001
Phaeodactylum tricornutum	MPB (p)	fcpA/fcpB	fcpA	egfp	shble	-	-	Apt et al. 2002
Phaeodactylum tricornutum	MPB	fcpB	fcpA	GUS	shble	20-40	10-20	Harada et al. 2005
Phaeodactylum tricornutum		fcpB	fcpA	ecfp/eyfp	shble	-	-	Siaut et al. 2007
Phaeodactylum tricornutum	MPB	CMV/LTR/CaMV35S/ fcpB	fcpA	GUS	shble	28-244	14-122	Sakaue et al. 2008
Phaeodactylum tricornutum	MPB	CFfcp/ CFNR	CFfcp/ CFNR	egfp	shble	650	325	Miyagawa et al. 2009
Phaeodactylum tricornutum	E	NR	NR	-	cat	1000	125	Niu et al. 2012
Phaeodactylum tricornutum	MPE	fcpA/fcpB/CaMV35S	fcpA	sgfp/ GUS	shble	4500	~1000	Miyahara et al. 2013
Phaeodactylum tricornutum	E (p)	PrbcL	TrbcS	egfp	cat	-	-	Xie et al. 2014
Phaeodactylum tricornutum	E	fcpA/fcpB	fcpA	egfp/ GUS	shble	2800	1400	Zhang & Hu 2014
Phaeodactylum tricornutum	BC	fcpB/fcpF	fcpA	yfp	shble	~4200	-	Karas et al. 2015
Phaeodactylum tricornutum	PEG	fcp	fcp	-	shble	1	-	Karas et al. 2015
Phaeodactylum tricornutum	MBP	EF2/fcpB/fcpF	fcpA	LUC	shble	-	-	Seo et al. 2015
Pseudo-nitzschia arenysensis	MPB	PM H4	PTfcpA	GUS/ egfp	shble	Liquid selection		Sabatino et al. 2015
Pseudo-nitzschia multistriata	MPB	PM H4	PTfcpA	-	shble	Liquid selection		Sabatino et al. 2015
Seminavis robusta	MPB	AtpBE	AtpEt	-	nat	-	-	Kirupamurthy 2014
Thalassiosira pseudonana	MPB	fcp/ NR	fcp/ NR	egfp	nat	423	423	Poulsen et al. 2006
Thalassiosira pseudonana	MPB	fcp/ NR	fcp/ NR	egfp/ GUS	nat/ GOx (-)	-	-	Sheppard et al. 2012
Thalassiosira pseudonana	MPB	fcp/ NR	fcp/ NR	egfp	nat	-	-	Samakawa et al. 2014
Thalassiosira pseudonana	BC	fcp	fcp	yfp	nat	~1500	-	Karas et al. 2015
Thalassiosira weissflogii	MPB (t)	PTfcpB	PTfcpA	GUS	-	-	-	Falciatore et al. 1999

Table 2.1. Transformable diatoms and overview of methods. Includes studies which have developed or added to methods for diatom transformation. See table 1b below for codes.

Code	Promoter
acetyl-CoA	acetyl coenzyme A
AtpBE	ATPase Beta
CaMV35S	Cauliflower mosaic virus 35S
CMV	cytomegalovirus
EF2	Elongation factor 2
fcp	fucoxanthin chlorophyll a/c binding protein
fruα3	fruα3
H4	Histone 4
LTR	long terminal repeat
NR	nitrate reductase
PrbcL	Rubisco large sub-unit
RSV	Rous sarcoma virus
Code	**Selective marker/ antibiotic**
cat	Chloramphenicol acetyltransferase/ chloramphenicol
nat	N-acetyltransferase/ nourseothricin
nptII	neomycin phosphotransferase/ neomycin & G418
sat-1	streptothricin acetyl transferase/ nourseothricin
shble	Streptoalloteichus hindustanus bleomycin/ zeocin & phleomycin
Code	**Terminator**
AtpET	ATPase Epsilon/Delta
fcp	fucoxanthin chlorophyll a/c binding protein
fruα2	fruα2
hepA	HEP200
TrbcS	Rubisco small sub-unit

Code	Reporter
ecfp	enhanced cyan fluorescent protein
egfp	enhanced green fluorescent protein
gfp	green fluorescent protein
GO (-)	glucose oxidase
GUS	β-Glucuronidase
luc	luciferase
sgfp	superfolder green fluorescent protein
yfp	yellow fluorescent protein
Code	**Transformation method**
BC	Bacterial conjugation
E	Electroporation
MPB	Microparticle bombardment
MPB (t)	Microparticle bombardment (transient expression)
MPB (p)	Microparticle bombardment (plastid transformation)
MPE	Multi-pulse electroporation
PEG	Polyethylene glycol
Code	**Species origin of promoter/ terminator**
CC	*C. cryptica*
CF	*C. fusiformis*
FS	*F. solaris*
PM	*P. multistriata*
PT	*P. tricornutum*
TP	*T. psuedonana*

Table 2.1b. Codes for transformation table. **CaMV35S, CMV and RSV are promoters with viral origins. All other promoters unless stated are endogenous. Exogenous promoters are prefixed with the initial of the species they originate from. (-) following the glucose oxidase reporter indicates a negative selective marker.**

Several reporter genes for fluorescent or colorimetric analysis have been tested in diatoms including enhanced green fluorescent protein (egfp), enhanced yellow fluorescent protein (eyfp), luciferase (luc) and β-glucuronidase (GUS). Enhanced green fluorescent protein (egfp) is the most prevalent and has been successfully expressed in the majority of transformed diatom species. For some diatoms such as *P. tricornutum*, successful expression may be due to similarities in codon usage between the host species and the egfp gene, which has a human codon bias, and therefore coincidently a *P. tricornutum* codon bias. Trials with other variants of gfp designed for expression in other species were not successful (Zaslavskaia et al., 2000). It does appear however, that other factors may contribute to the function of reporter genes, as no expression from egfp can be seen in *Chaetoceros sp*. despite possessing a similar codon bias to *P. tricornutum* (Ifuku et al., 2015; Miyagawa-Yamaguchi et al., 2011). Egfp was chosen for this study due to its functionality in the majority of modified diatoms, following analysis of codon usage in *F. cylindrus* (Figure 2.9).

The ability to stably introduce DNA and express transgenes has been used for a wide range of different applications in diatoms including determining gene function and cellular mechanisms, as well as for biotechnology. Fluorescent marker genes have been fused to genes of interest to investigate potential adhesion proteins in raphid pennate diatoms (Buhmann et al., 2014), localise carbonic anhydrase (Samukawa et al., 2014) and visualise localisation of proteins to different organelles (Apt et al., 2002; Siaut et al., 2007), including determination of localisation signals (Apt et al., 2002). GUS has been fused to several truncated versions of the β-carbonic anhydrase promoter in *P. tricornutum* to determine essential elements for carbon dioxide responsive

transcription (Harada et al., 2005). Overexpression (Yao et al., 2014) and gene knock-out (Daboussi et al., 2014) have been used to increase glycerol and lipid production, whilst gene silencing has been used to study photoreceptors (De Riso et al., 2009), photosynthesis and non-photochemical quenching (Bailleul et al., 2010; Lavaud et al., 2012). As well as RNAi, genes linked to autotrophic growth have also been knocked out using CRISPR-Cas (Nymark et al., 2016). On the opposite end of the spectrum, glucose transporters have been expressed in *P. tricornutum* to allow heterotrophic growth without photosynthesis (Zaslavskaia et al., 2001). Several proteins and active enzymes have been tagged to the cell membrane and stably incorporated into the silica frustule (Fischer et al., 1999; Poulsen et al., 2007; Sheppard et al., 2012). This has powerful applications for biotechnology as demonstrated by the incorporation of antibodies into the frustule for targeted drug delivery (Delalat et al., 2015).

This chapter describes a proof of principle transformation system in *F. cylindrus*. This is the first psychrophilic diatom to be transformed and possibly the first psychrophilic eukaryote. This has implications for not only understanding this ecologically important species and diatoms in general, but also for biotechnology given that solubility, folding, yield and stability of recombinant proteins can be improved at lower temperatures (San-Miguel et al., 2013; Vasina and Baneyx, 1996). These properties can be enhanced in psychrophilic bacterial hosts which are adapted to growth in cold conditions (Giuliani et al., 2014; Miyake et al., 2007; Vigentini et al., 2006). Eukaryotic hosts can be required for expression of large proteins or proteins with specific post-translational modifications (Demain and Vaishnav, 2009), therefore it may advantageous to have a system in which recombinant proteins can be expressed in a psychrophilic, eukaryotic host.

Materials and methods

Strains and growth conditions

Fragilariopsis cylindrus (CCMP 1102) was grown in Aquil synthetic seawater (Price et al., 1989) at 4°C under 24 hour light (100-140µE) conditions. Starter cultures were inoculated with no less than 50,000 cells/ml.

Construct for egfp and shble expression

Choosing promoter and terminator regions

RNA sequencing data produced by Jan Strauss (Mock et al., 2017; Strauss, 2012) from *Fragilariopsis cylindrus* under control, iron limiting, low temperature (2°C), high temperature (8°C), high CO_2 and dark conditions, was examined for genes with the highest expression levels. Sequences of approximately 1000bp, flanking the coding region up and down-stream of the fucoxanthin chlorophyll a/c binding protein (JGI ID 267576) were chosen for the promoter and terminator regions respectively. Primer design determined the final size of the promoter and terminator products. The promoter region spans -986 to -1 bp upstream of the coding region whilst terminator sequences of two different lengths, 1118 and 1099bp starting immediately after the FCP stop codon, have been used in the final construct.

Determining codon bias compared to egfp

F. cylindrus transcripts were downloaded from the Joint Genome Institute (JGI). The 20 highest expressed *F. cylindrus* genes under control conditions were determined from previously generated expression data (Mock et al., 2017; Strauss, 2012). Emboss cusp was used to calculate codon usage for all *F. cylindrus* transcripts, the 20 highest expressed transcripts and egfp. Usage was expressed as a fraction (frequency of one codon/frequency of all codons for a specific amino acid). Human codon bias was downloaded from Genscript (http://www.genscript.com/tools/codon-frequency-table) and *P. tricornutum* codon usage was taken from Scala et al. (2002).

Extraction of F. cylindrus genomic DNA. *F. cylindrus* was grown to exponential phase ($1x10^6$ cells/ml) and gDNA extracted using the Easy-DNA genomic purification kit (ThermoFisher) according to the manufacturer's protocol.

PCR of fragments for Gibson assembly. Two attempts at making the construct via Gibson assembly were made. The first used pBluescript II (SK-) as a backbone. Assembly of all fragments in one reaction as well as assembly in two parts was attempted but was ultimately unsuccessful. The second used puc19 as a backbone and was conducted in two parts. Part one produced FCFCP:shble (figure 2.3) and part 2 produced the final plasmid pucFCFCPshble:FCPegfp (figure 2.4). Phusion DNA polymerase (NEB) was used to amplify fragments for Gibson assembly. Primers include sequences complementary to the adjacent fragment to give either 20bp overhangs (pBluescript assembly, table 2.2) or 40bp overhangs (puc19 assembly, table 2.3). FCP promoter and terminator regions were amplified from genomic DNA whilst egfp was amplified from TPfcpGFP and shble from pPha-T1 (Zaslavskaia et al., 2000). PCR was carried out with final concentrations of 1x HF buffer, 0.2mM dNTPs and 0.3uM of each primer in 20-100µl volumes. Either 1ng of plasmid DNA, 1ng of PCR product or 200-500ng of gDNA was used as a template. Initial denaturation was performed for 2 minutes at 98°C, followed by 35 cycles of denaturation at 98°C for 10 seconds, annealing for 30 seconds (see table 2.2 and table 2.3 for annealing temperatures) and extension at 72°C for 30 seconds per kb. Following 35 cycles, a final extension for 5 minutes at 72°C was performed. Products were run on 0.8% agarose gels in 1x TAE buffer and bands were excised and purified using a GFX PCR DNA and Gel Band Purification Kit (GE).

Preparing pBluescript II by restriction digest. One µg of pBluescript II (SK-) was digested with 5 units of KpnI and NotI in a 20µl reaction with 1x Multicore buffer (Promega) at 37°C for 4 hours. The vector was dephosphorylated by adding 1µl of Antarctic phosphatase (NEB), 2.4µl of 10x Antarctic phosphatase buffer + 0.6µl of water and incubated at 37°C for 15 minutes before heat inactivation at 70°C for 5 minutes. The reaction was run on a 1% agarose gel in 1 x TAE buffer. The band corresponding to the linearized vector was excised and purified using a GFX PCR DNA and Gel Band Purification Kit.

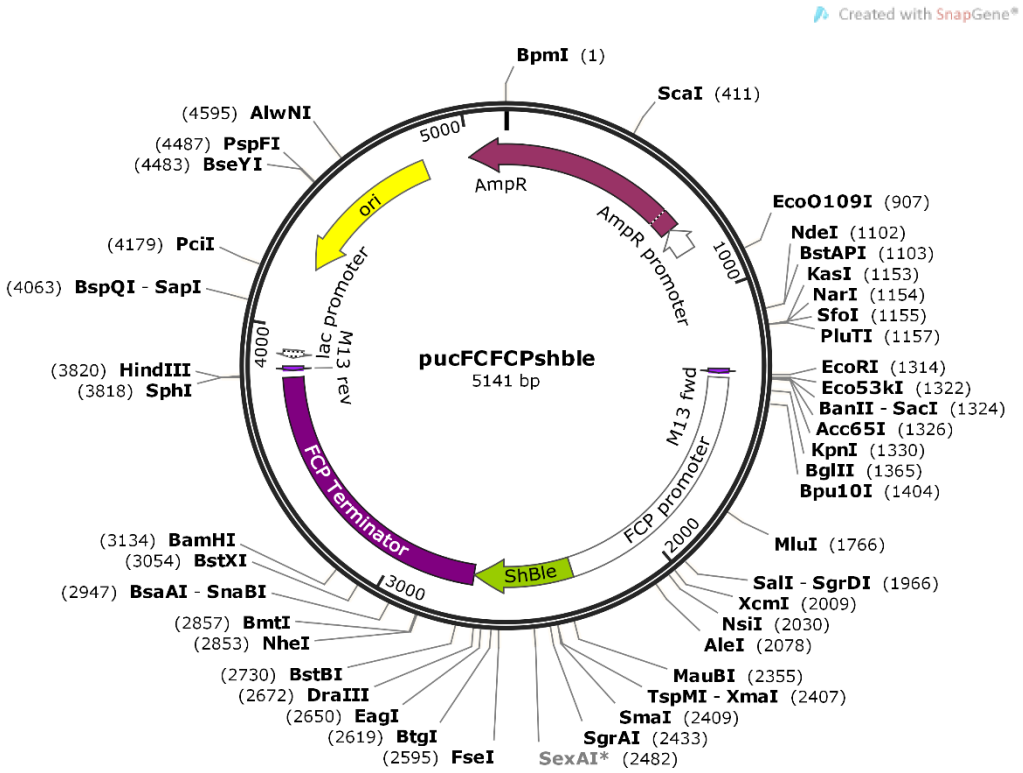


Figure 2.3. Vector map of pucFC_FCPshble. Plasmid constructed during part one of Gibson assembly – FCP:shble cassette in a puc19 backbone.

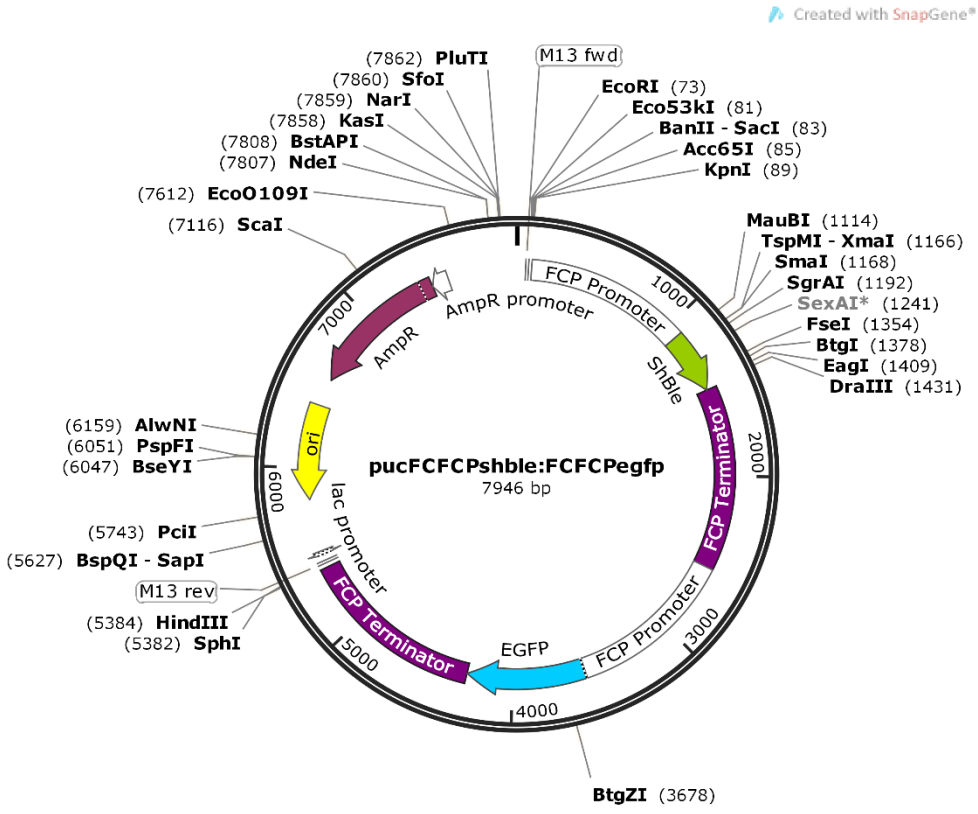


Figure 2.4. Vector map of pucFCFCPshble:FCFCPegfp. Plasmid constructed during part two of Gibson assembly – FCP:shble cassette and FCP:egfp cassette in a puc19 backbone.

Product	Primer	Sequence	Annealling temp (°C)	Product size
pBluescript II	pBlue_amp_F	GCGGCCGCCACCGCGGTG	72	2882
	pBlue_amp_R	GTACCCAATTCGCCCTATAGTGAGTCGTATTACGCGCG		
Prom1	Gib_FC_P1_F	ctatagggcgaattgggtac*catatg* CCCAAAGTAAGGCATAG	53	1024
	Gib_FC_P1_R	ccat*ctcgag* TTTGATATATAAGTTTGTTTTTGG		
sh_ble	Gib_FC_shble_F	tatatatcaaa*ctcgag* ATGGCCAAGTTGACCAGTGC	67	402
	Gib_FC_shble_R	gc*ttaattaa* TCAGTCCTGCTCCTCGGC		
Term1 (1 step)	Gib_FC_T1_F	gcaggactga*ttaattaa* GCATTTTATTAATCCTTATTTGATCG	60	1146
	Gib_FC_T1_R	tt*gcggccgc* AGTCGTTGTTGTTGTGCTG		
Term 1 (2 step)	Gib_FC_T1_F	gcaggactga*ttaattaa* GCATTTTATTAATCCTTATTTGATCG	60	1162
	Gib_FC_T1_R_add	tggagctccaccgcggtg*gcggccgc* AGTCGTTGTTGTTGTGCTG		
Prom2	Gib_FC_P2_F	aacaacgact*gcggccgc* AAAGTAAGGCATAGAAATAATCTG	58	1013
	Gib_FC_P2_R	ccat*gaattc* TTTGATATATAAGTTTGTTTTTGGTAGT		
Egfp	Gib_FC_egfp_F	tatatatcaaa*gaattc* ATGGTGAGCAAGGGC	55	747
	Gib_FC_egfp_R	atgc*gcatgc* TTACTTGTACAGCTCGTCC		
Term2	Gib_FC_T2_F	gtacaagtaa*gcatgc* GCATTTTATTAATCCTTATTTGATCG	60	1154
	Gib_FC_T2_R	tggagctccaccgcggtggcAGTCGTTGTTGTTGTGCTG		

Table 2.2. Primers for amplification of fragments for Gibson assembly using pBluescript II (SK-) as a backbone.

Product	Primer	Sequence	Annealling temp (°C)	Product size
puc19	pucGA vector F	cagcacaacaacaacgactAGGCATGCAAGCTTGGC	65	2700
	pucGA vector R	atttctatgccttactttgGGGTACCGAGCTCGAATTCAC		
Prom 1	pucGA prom1 F	attcgagctcggtacccCAAAGTAAGGCATAGAAATAATC	58	1021
	pucGA prom1 R	tggtcaacttggccatTTTGATATATAAGTTTGTTTTTGGTAG		
Shble	pucGA shble F	aaacaaacttatatatcaaaATGGCCAAGTTGACCAGTGC	67	415
	pucGA shble R	aataaggattaataaaatgcTCAGTCCTGCTCCTCGGC		
Term 1	pucGA term1 F	cgaggagcaggactgaGCATTTTATTAATCCTTATTTGATCG	60	1151
	pucGA term1 R	gccaagcttgcatgcctAGTCGTTGTTGTTGTGCTG		
Puc19:fcp:shble	pucGA vector2 F	actactgttgtcgtctactaAGGCATGCAAGCTTGGC	62	5181
	pucGA vector2 R	tatttctatgccttactttgAGTCGTTGTTGTTGTGCTGTAG		
Prom2	pucGA prom2 F	agcacaacaacaacgactCAAAGTAAGGCATAGAAATAATCTG	61	1021
	pucGA prom2 R	tcgcccttgctcaccatTTTGATATATAAGTTTGTTTTTGGTAG		
Egfp	pucGA egfp F	aaacaaacttatatatcaaaATGGTGAGCAAGGGCGAG	62	760
	pucGA egfp R	aataaggattaataaaatgcTTACTTGTACAGCTCGTCCATG		
Term2	pucGA term2 F	acgagctgtacaagtaaGCATTTTATTAATCCTTATTTGATC	61	1136
	pucGA term2 R	tacgccaagcttgcatgcctTAGTAGACGACAACAGTAGT		
Insert	pucGA Insert F	GCTGCAAGGCGATTAAGTTG	64	Step1: 2653
	pucGA Insert R	GCTCGTATGTTGTGTGGAATTG		Step 2: 5458

Table 2.3. Primers for amplification of fragments for Gibson assembly using puc19 as a backbone.

Gibson assembly. Gibson assembly was carried out with either Gibson Assembly Master Mix (NEB) or with a Master Mix compiled from a recipe on OpenWetWare (Ford, 2013). For the latter, an 'Isothermal Start Mix' stock was made with 1.5g PEG_{8000}, 3ml of 1M Tris-HCl (pH 8.0) and 150µl of 2M $MgCl_2$. The recipe can easily be scaled for smaller volumes. 2x Master Mix: 405µl Isothermal Start Mix, 25µl 1M DTT, 20µl 25mM dNTPs, 50µl NAD+, 31.25µl Phusion polymerase, 250µl Taq Ligase and 467.75µl of nuclease free water. NAD+ and enzymes were all purchased from NEB.

Vector was added at 50ng for the first part of the assembly or the full 6 fragment assembly and 100ng for the second part of the assembly (figure 2.5). All inserts were added at a 3x molar excess compared to the vector. Vector, inserts, and master mix (final 1x concentration) were added together in a 10µl volume. The reaction was incubated at 50°C for 1 hour and stored at -20°C until transformation into *E. coli*.

Bacterial Transformation

One µl of the reaction was used for transformation into high efficiency NEB 5-alpha competent *E. coli* as described in the NEB protocol. One hundred µl of transfomant cells in SOC media were spread onto selective LB agar plates. Plates were made with 100µg/ml ampicillin and spread with 40µl of 20mg/ml x-gal and 40µl of 100mM IPTG an hour prior to plating cells. Plates were incubated at 37°C overnight.

Direct PCR from Gibson assembly

Initial attempts using a 20bp overhang and pBluescript II as a vector failed to produce colonies with the correct inserts. In order to troubleshoot this, PCRs directly from the Gibson assembly reaction were carried out to determine if any of the fragments were being correctly assembled. This involved using primers (table 2.2) spanning across 2-6 fragments, with the lowest annealing temperature of the primer pair being used. Phusion DNA polymerase and 0.5µl of the Gibson assembly reaction in a 50µl volume was used to amplify fragments as described. Products were run on a 1% agarose gel.

Screening constructs

Typically 4 colonies from each assembly were picked and grown in 5ml of LB media with 100µg/ml ampicillin overnight, before harvesting plasmids with a PureYield Plasmid Miniprep kit (Promega). For pBlusecript II assemblies plasmids were digested in single and double digests in buffer D and 1x BSA with XhoI, NotI, EcoRI, NdeI, SphI, KpnI, depending on the fragment being examined. For puc19 assemblies KpnI and HindIII, which flank the insert region, were used in a double digest in 1x Multicore buffer and 1 x BSA. BamHI in Multicore buffer was also used to check the full size of the puc19 assemblies by linearization. Five units of enzyme was used in a 20µl reaction to digest between 250-500ng of plasmid at 37°C for 2 hours. Reactions were run on 1% agarose gels.

Primers to amplify the full insert from puc19 assemblies were designed (Table 2.3). PCR was carried out with Phusion DNA polymerase and 1ng of plasmid as described. Constructs, which screened positive for correct inserts by digest and PCR, were sodium acetate-ethanol precipitated and sent for sequencing. Plasmids with the correct sequence were transformed into NEB 5-alpha competent *E. coli*. Three colonies were picked and grown in 5ml of LB media with 100µg/ml ampicillin overnight. Five hundred µl of each miniprep were then transferred to 100ml of selective LB media and grown overnight before harvesting *plasmids* with a PureYield Plasmid Maxiprep kit (Promega) and eluting in TE buffer. Plasmid Maxipreps were checked by restriction digest using BamHI. Plamids were sodium-acetate ethanol precipitated and re-suspended in nuclease free water to remove salts and concentrate DNA prior to diatom transformation.

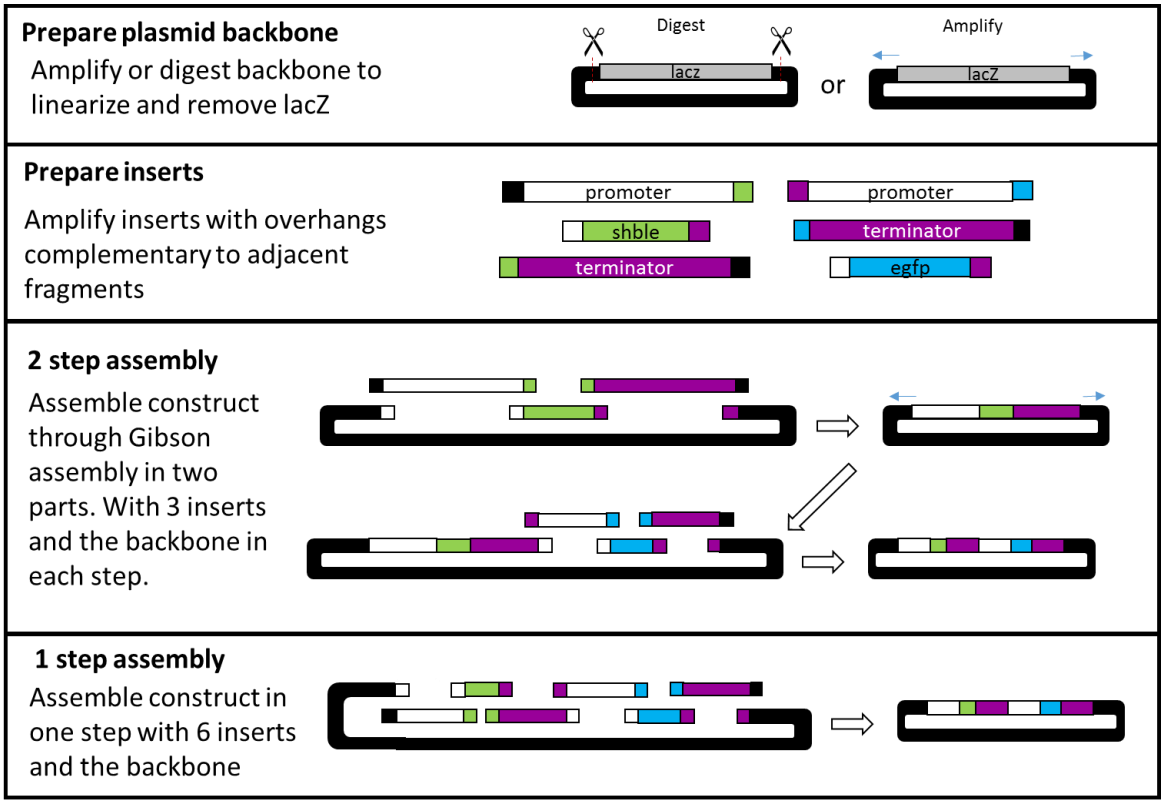


Figure 2.5. Overview of Gibson assembly. Depicts vector preparation, amplification of insert and construction in both 1 and 2 steps.

Transforming *Fragilariopsis cylindrus.*

Testing zeocin concentrations on plates.

Preliminary work by Jan Strauss showed 100µg/ml zeocin in liquid Aquil media prevents growth of *F. cylindrus. F. cylindrus* growth was tested on plates with different zeocin concentrations. Aquil-0.8% Agar plates were made with 0 (control), 50, 100 and 200µg/ml zeocin. As salinity has been shown to affect antibiotic function (Falciatore et al., 1999; Muto et al., 2013) two sets of plates were made, one with full salinity and the other with half salinity. In the latter, all synthetic ocean water (SOW) salts were reduced by half but nutrient concentration stayed the same. In order

to prevent precipitation during autoclaving, plates were made by autoclaving a 2x SOW (or 1x, in the case of half salinity plates) and 2 x agar separately. Once cooled to 50°C the two solutions were added together and nutrients (phosphate, nitrate, silicate, trace metals and vitamins) and zeocin added. Once poured unused plates were kept for up to a month at 4°C. *F. cylindrus* cells were grown to exponential phase (~$1x10^6$ cells/ml), and harvested by spinning at 3000xg for 10 minutes at 4°C in a pre-chilled centrifuge. Supernatant was removed and cells resuspended in fresh media at $5x10^7$ cells/ml. Two hundred µl of culture ($1x10^7$ cells) was spread on the prepared plates using a sterilized Drigalski spatula in triplicate for each condition. Throughout the process cells and plates were kept on ice. Plates were incubated upside down at 4°C in 24 hour light. To avoid condensation plates were not wrapped in parafilm.

Microparticle bombardment.

Non-selective 1.5% agar Aquil plates for shooting and 0.8% agar Aquil plates with 100µg/ml zeocin for selection were made as described. Non-selective 0.8% agar Aquil plates were also made for the positive growth control.

Particles were prepared and coated with plasmid according to Kroth (2007). Briefly, 50µl (3mg) of prepared 0.7µm (M10) tungsten particles were coated in 5µg of plasmid in the presence of $CaCl_2$ and spermidine. This is enough for 5 shots with 600ng of tungsten particles and 1µg of plasmid. A negative control in which water replaced the plasmid was also carried out.

Using vacuum filtration 5 x 10^7 *F. cylindrus* cells in exponential phase ($1x10^6$ cells/ml) were collected onto a 47mm 1.2µm Isopore membrane filter (Millipore) at 4°C. Each filter was placed on a 1.5% Agar plate for shooting. Plates were kept at 4°C until required.

The Bio-Rad PDS-1000/He biolistic microparticle delivery system was used to introduce plasmids into the cells, according to manufacturer's instructions, as described by Kroth (2007). Initially rupture discs at 1100 and 1350 psi were used. A second transformation utilised 1350 and 1550 psi rupture discs. Cells were placed on the 2nd shelf at a 6cm flight distance and shot in a vacuum of 25 Hg. All shots were carried out in triplicate. Following shooting plates were placed on ice and the filter turned upside down so that the cells came into contact with the nutrient-rich agar. Cells were left to incubate at 4°C for 24 hours in the light before being scrapped/ pipetted off into 500µl of fresh non-selective Aquil media. For the first transformation 3 x 100µl aliquots from each bombardment were spread onto 3 selective plates ($1x10^7$ cells/ plate). The remaining 200µl was used to inoculate 20ml of liquid Aquil media containing 100µg/ml zeocin. As well as selective plates 100µl from the negative control was spread onto non-selective plates as a positive growth control. Plates were incubated under standard growth conditions until colonies appeared. Liquid selective cultures were split into two tubes after 1 week. One tube remained at 100µg/ml zeocin and the other increased to 200µg/ml due to lack of bleaching. Liquid cultures were incubated under standard conditions until colour was noticeable. Colonies from plates were picked and resuspended in 500µl of pre-chilled Aquil with 100µg/ml zeocin in 12 well plates and left to grow under standard conditions before transferring to a larger 20ml volume of selective Aquil.

Screening

Once cell density from liquid selection and colony cultures reached 1-2x10^6 cells/ml, gDNA was extracted from 1ml using the Easy DNA gDNA purification kit (ThermoFisher) according to protocol #3 of the product manual. Extracted gDNA was resuspended in 20µl of TE buffer and the concentration measured using a NanoDrop 1000 spectrophotometer.

Phusion PCRs were carried out with the same primers and parameters used to amplify the FCP promoter (positive control), shble and egfp fragments for Gibson assembly. Template consisted of 15-100ng of transformant or wild-type genomic DNA, or plasmid DNA for the positive control.

Flow cytometry

Egfp fluorescence of exponentially growing cells was measured using a Becton Dickinson FACS calibur flow cytometer. Autofluorescence was measured by excitation at 550nm (FL1) and egfp at 435nm (FL3). FL1 (x) and FL3 (y) were plotted to determine if green fluorescence was relatively higher in samples with egfp.

Widefield microscopy

Fluorescence was examined at a cellular level using a widefield Zeiss Axioplan 2ie microscope. Using 63x and 100x objectives, an Alexa 586 filter (Ex=578, Em = 603) and GFP filter (Ex = 488, Em = 509) were used to observe autofluorescence and egfp respectively. Cells were also observed under brightfield.

Testing transformant stability

Transformants were grown in non-selective media under standard conditions for 2 months before transferring back to selective media to test stability of zeocin resistance. Cultures were maintained by passaging to media with 100µg/ml zeocin every two months. PCR of the egfp and shble genes (positive control) was carried out on cultures two years after transformation to determine if non-selective egfp was still present in transformant cultures. Egfp and shble were amplified by GoTaq polymerase (Promega) from lysate of cells at stationary phase (~ 2 x 10^6 cells/ml) using the same primers designed for Gibson assembly (table 2.3). Final concentrations of each reagent can be found in the GoTaq flexi protocol - $MgCl_2$ was added at a final concentration of 1.25mM. Ten µl of three transformant cultures, which originally screened positive for egfp and one that screened negative for egfp were spun down for a minute and the supernatant removed. Cells were resuspended in 20µl of lysis buffer (10% Triton X-100, 20 mM Tris–HCl pH 8, 10 mM EDTA), kept on ice for 15 minutes then incubated at 95 °C for 10 minutes. One µl of lysate was used in a 20µl reaction with the following parameters: Initial denaturation for 5 minutes at 95°C, followed by 35 cycles of denaturation at 95°C for 30 seconds, annealing for 1 minute at 51°C (shble) and 56°C (egfp), and extension at 72°C for 1 minute. Following 35 cycles, a final extension for 10 minutes at 72°C was performed. Products were run on 1% agarose gels.

Results and Discussion

Choosing promoter and terminator sequences

Examination of *F. cylindrus* RNA-seq data across several conditions, showed that two fucoxanthin chlorophyll a/c binding proteins (FCP; JGI ID 267576 and 143190) gave the two highest expression levels in the control condition, and were highly ranked (top 30) under iron limiting, low temperature (2°C), high temperature (8°C) and high CO_2. Expression levels of both proteins were also high under prolonged darkness (in the top 26% of expression levels), especially since FCP in other diatoms species can be heavily down-regulated in the absence of light (Leblanc et al., 1999; Siaut et al., 2007). Endogenous (Falciatore et al., 1999; Ifuku et al., 2015; Poulsen et al., 2006) and exogenous (Buhmann et al., 2014; Miyahara et al., 2013; Muto et al., 2013) diatom FCP promoter have been widely used in diatom transformation systems due to their high expression levels and previous success. FCP terminators are also widely used in conjugation with the FCP promoter as well as other diatom promoters (Falciatore et al., 1999; Ifuku et al., 2015; Poulsen et al., 2006; Seo et al., 2015).

The FCP promoter was chosen for high expression levels, to help ensure resistance against zeocin and clear expression of egfp.

Plasmid construction

Gibson assembly was used to assemble the FCP:Shble and FCP:egfp cassettes in a single pucFCFCPshble:FCFCPegfp construct. This method was chosen as it allows assembly of multiple fragments in a single reaction. Fragments for use in Gibson assembly are amplified to include overlapping sequences to the adjacent fragment (figure 2.5). All fragments and a linearized backbone vector are included in the same reaction, in which a 5’ exonuclease chews back the 5’ end, allowing overhangs of adjacent fragments to anneal. A polymerase then fills in any gaps and a ligase seals the nicks. This should produce a circular plasmid with all inserts in the correct formation. Initial attempts were unsuccessful, however, optimisation and changes to the method led to successful assembly.

Initially pBluescriptII (SK-) was used as backbone vector. This vector contains the LacZ gene for blue/ white colony screening. Assembly was carried out with either a digested and dephosphorylated vector, or an amplified vector, with 20bp overhangs corresponding to the promoter and terminator regions. In both cases this led to white colonies but no insert, as determined by restriction digest and PCR screening. Furthermore vector-only controls also led to white colonies suggesting that the vector can self-seal through Gibson assembly. Colonies from template carryover should be minimised as both digested and amplified vectors are run on a gel and purified prior to assembly. In addition colonies from un-modified, un-cut template should be blue. As a result the pBluescript II (SK-) vector was substituted for puc19. Puc19 for assembly was amplified rather than linearized as this reduces that chance of template being carried over, given that only a small amount is required for PCR.

A 6-fragment GA with all fragments and the pBluescript II vector was attempted. Assembly in two parts using 3 fragments was carried out for with both pBluescript II and puc19. For the two-part assembly, the backbone was first combined with the FCP promoter, shble gene and FCP terminator (figure 2.3). This construct was then used in a second GA as the backbone, with the FCP promoter, egfp gene and FCP terminator (figure 2.4). In this case the second FCP terminator has a different 3' end so that the overhang on the 5' end of the vector matches the second terminator rather than the first, thus avoiding resealing of the first construct. Initially 20bp overhangs were built into fragments for assembly into pBluescript II. Although no plasmids with inserts were observed, PCRs carried out on the Gibson assembly reaction showed that smaller numbers of fragments were being assembled. Each adjacent fragment was confirmed to assemble as were the FCP:Shble and FCP:egfp cassettes. No full 6 fragment assembly was observed – this may be due to either difficulties carrying out PCR of a large sequence (~6000bp) or a reduced Gibson assembly efficiency due to the larger number of fragments (NEB). As a result, it was decided to carry out the puc19 Gibson assembly in two parts as described. Overhangs for the optimised assembly were also increased to ~40bp as NEB suggests increasing overhangs to increase assembly efficiency. Primers for amplification of the fragments for assembly into pBluescript also included restriction sites so that genes and promoters could be changed in future constructs. These were removed from the final optimised assembly due to high levels of primer dimers potentially caused by the palindromic sequences.

In the final construct, successful assembly was carried out by amplifying vectors, promoters, coding regions and terminators with 40bp overhangs. The FCP:shble cassette was initially assembled into puc19 via Gibson assembly (figure 2.3). The resulting construct was then used in a second assembly with fragments for the FCP:egfp cassette, creating a single construct with both cassettes (figure 2.4). The FCP:shble construct has also been used in later CRISPR-Cas applications (see CRISPR chapter).

No difference in number of colonies or assembly was observed when using either the NEB Gibson assembly master mix or the homemade version according to openwetware (Ford, 2013). The homemade master mix was preferentially used due to lower costs.

Testing zeocin concentrations on plates.

After two weeks cells bleached on plates with all zeocin concentrations. No growth was observed after several months. This compares to positive growth control plates without zeocin which showed a lawn of cells after two weeks. There was no clear difference between growth (control), or lack of growth (zeocin) between the two salinities. Plates with zeocin at 100µg/ml and 100% salinity were chosen to select transformants. The zeocin concentration required is consistent with several other diatom species which use shble as a selective marker, such as *P. tricornutum and Pseudo-nitzchia species* (Apt et al., 1996; Falciatore et al., 1999; Sabatino et al., 2015). Salinity and cell density can impact zeocin performance. Falciatore et al. (1999) found that plates with cell densities at 10^6 compared to 10^5, required 100µg/ml zeocin and a reduction in salinity to 50% in order to inhibit

growth. Selection of this species is routinely carried out with 10^7 cells/ml with 75-100µg/ml zeocin (Apt et al., 2002, 1996; Kroth, 2007). In *A. coffeaeformis* 600µg/ml is required (Buhmann et al., 2014) whilst on the opposite end of the spectrum, *T. pseudonana* shows a very high sensitivity to zeocin, leading to difficulties producing transformant colonies (Poulsen et al., 2006). In contrast 50 µg/ml zeocin on 100% salinity plates was able to inhibit growth of 10^7 *F. cylindrus* cells, but allow the growth of shble transformants, suggesting this is a suitable antibiotic for selection in this species.

Microparticle bombardment

After 3-5 weeks transformant colonies appeared on plates. Colonies grew in size over the next couple of weeks and were picked at weeks 5 and 7. Transformation efficiencies for each transformation at different biolistic pressures are shown in table 2.4. The highest average efficiency was seen at 1550 psi with 30 ± 14 colonies/10^8 cells. Transformation efficiency was variable, with large differences between replicates and transformation events using the same rupture discs. There are reports of variable transformation efficiencies seen when using microparticle bombardment in other systems (Sabatino et al., 2015), and between studies (table 2.1), although it can be difficult to discern variability within particular studies as averages or highest efficiencies are often given within the same parameters. The highest efficiencies are seen with bacterial conjugation for *P. tricornutum* and *T. pseudonana*, and with either electroporation or multi-pulse electroporation in *P. tricornutum* (table 2.1). The most common method of transformation in diatoms is microparticle bombardment. This was the first method used to transform a diatom species, providing a means to introduce transgenic DNA into the cell through the silica frustule (Dunahay et al., 1995). Transformation efficiencies with this method vary widely between species and parameter, with the highest seen in *A. coffeaeformis* at 800 colonies/10^8 cells and lowest in *N. saprophila* at 2.8 colonies/10^8 cells. Efficiencies can be calculated either by colonies per number of cells transformed or µg plasmid used, which can drastically alter results. In either case *F. cylindrus* ranks 10^{th} out of 15 (table 2.1) for number of colonies from microparticle bombardment, with similar efficiencies to studies with *P. tricornutum*, *C. cryptica* and *F. solaris* (Dunahay et al., 1995; Harada et al., 2005; Muto et al., 2013; Zaslavskaia et al., 2000). The majority of transformations with this method favour flight distances between 6-7 cm, tungsten particles between 0.7-1.1µm and higher pressure rupture discs (1350 – 2000 psi). Pressure at 1550 psi is used for most species, from those with highly silicified shells such as *T. pseudonana* to those with lightly silicified frustules such as *P. tricornutum*. An exception to this rule is *F. solaris*, a lightly silicified pennate diatom (Matsumoto et al., 2014) which sees the highest number of colonies with rupture discs at 450 and 650 psi (Muto et al., 2013). The highest number of transformants in *F. cylindrus* was seen with the highest pressure at 1550 psi, but due to variability it's difficult to say whether or not this is significant. Optimisation of the system such as exploring pressure and particle size in *F. cylindrus* may yield larger numbers of colonies.

Whilst the *F. cylindrus* transformation system bears many similarities to microparticle bombardment in other diatoms, some key aspects of the method have required alterations to adapt the system to a psychrophilic species. First and foremost it was essential to keep cells at 4°C or on ice throughout all procedures. This also meant that cells could not be dried onto plates for transformation. Instead, cells were gently filtered onto 1.2µm isopore membranes and the filter placed on an agar plate for shooting. Further tests would be required to see if this positively or negatively affects the delivery of transgenes into cells. It is possible that the filter on agar provides a different level of shock absorption during shooting compared to agar alone. Filtration does however provide an even layer of cells in an exact diameter compared to drying which is more irregular and can often lead to clumping. The 24 hour non-selective incubation was carried out by flipping the filter upside down so that cells came into contact with the nutrient rich agar. It may be worth optimising this step to see if recovery in liquid media is preferred. It took 3 to 5 weeks for colonies to appear. In comparison, temperate species, with well-established systems such as *P. tricornutum* and *T. pseudonana*, typically take 1 ½ – 3 weeks (Apt et al., 1996; Falciatore et al., 1999; Zaslavskaia et al., 2000) and 8-10 days (Poulsen et al., 2006) to form colonies, respectively. The growth rate for *P. tricornutum* is around 0.92 with a doubling time of ~18 hours (Mann and Jack, 1968) whilst growth rate for *T. pseudonana* is about 1.2, doubling approximately every 14 hours (CRISPR-Cas chapter). In contrast the growth rate for *F. cylindrus* is about 0.35 with cells doubling every two days. By taking the growth rate into account, growth on plates is comparable to the two temperate species, with the first colonies appearing slightly quicker for *F. cylindrus*. Some diatoms such as *P. arenysensis* and *P. multistriata* have been selected in liquid media due to difficulties growing cells on plates (Sabatino et al., 2015). Selection in liquid media was also carried out for *F. cylindrus*. Cultures took three weeks to bleach with both 100 and 200µg/ml zeocin. It took 7-10 weeks for cultures to reach a high cell density (2 x 10^6 cells/ ml) and only two out of three replicates for each rupture disc (1100 psi and 1350 psi) showed growth in zeocin. In this case growth on plates is a much faster method to select transformants for screening. *F. cylindrus* is known to live in sea-ice and brine pockets (Mock and Thomas, 2008) and is therefore adapted to life in harsh conditions and on solid substrates. This may give it an advantage when growing on plates.

	Colonies/ Transformation			
Replicate	**1100 psi**	**1350 psi**	**1350 psi**	**1550 psi**
1	2	4	16	22
2	1	1	11	9
3	0	0	4	14
Average/ event	1.0	1.7	10.3	15.0
Cells plated	3 x 10^7	3 x 10^7	5 x 10^7	5 x 10^7
Efficiency/ 10^8 cells	3.3	5.6	20.7	30.0

Table 2.4. Numbers of colonies and transformation efficiency following transformation of *F. cylindrus*.

Screening

PCRs on gDNA extracts from the first transformation were carried out to check the presence of transgenes in both colonies and liquid selection cultures. Three colonies were tested from cells transformed with 1100 psi rupture discs and two from 1350psi. All picked colonies (figure 2.6) and cultures from liquid selection screened positive for the shble gene. Faint egfp bands were seen for 1 replicate from each pressure in the liquid selection cultures. Strong egfp bands were observed for three of the five colonies tested: 1100 FC2, 1100 FC3-2 and 1350 FC2 (named by pressure-replicate-colony) meaning that 60% of colonies with shble also screened positive for egfp. Similar results are seen in other diatoms systems. Studies in which two transgenes have been introduced simultaneously via co-transformation show that 37.5 – 70% of transformants which contain the selective marker also contain the reporter gene (Falciatore et al., 1999; Harada et al., 2005; Ifuku et al., 2015). In transformations where two transgenes are introduced on the same construct, as is this case in this study, 50-60% of resistant colonies also contained the second marker (Harada et al., 2005; Zhang and Hu, 2014). In this study presence of transgenes was screened simply by presence/absence through PCR. Whilst this method is also used in other papers (Miyahara et al., 2013), many studies also look at the number of copies integrated via Southern blotting. Typically between 1-6 copies are integrated (Harada et al., 2005; Miyagawa-Yamaguchi et al., 2011; Muto et al., 2013; Poulsen et al., 2006; Seo et al., 2015) although, as many as 10 copies have been found (Falciatore et al., 1999). As this method was conducted as a proof of principle, and given that integration of transgenes appears to be random in both position and number (so is likely to change between transformations), it was decided not to carry out Southern blots, although it may be worth considering in future work, in which the number of copies may affect expression and phenotype.

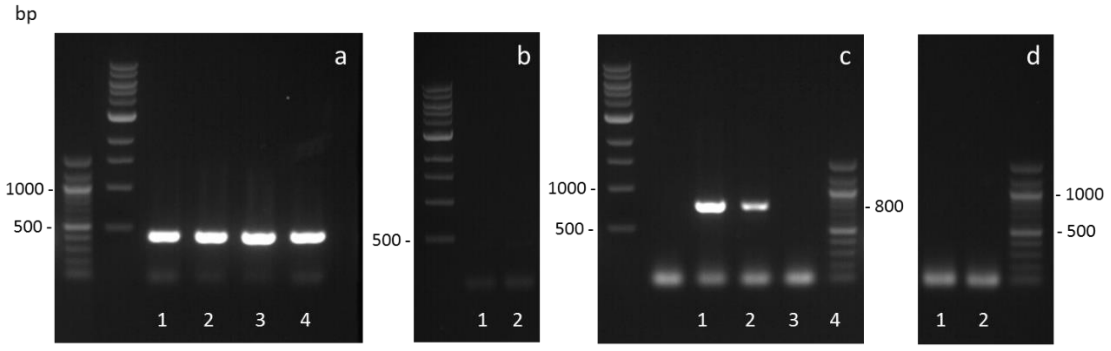


Figure 2.6. PCR of transgenes from gDNA. a. PCR of the shble gene. 1-4) 1100 FC3-1, 1100 FC3-2, 1350 FC2, 1350 FC3, b. PCR of the shble gene. 1) WT, 2) negative control c. PCR of the egfp gene. 1) 1100 FC3-1, 2) 1100 FC3-2, 3) 1350 FC2, 4) 1350 FC3 d. PCR of the egfp gene. 1) WT, 2) negative control.

Flow cytometry and fluorescence microscopy

Egfp expression was measured on a flow cytometer by comparing green fluorescence to red autofluorescence (figure 2.7). Relatively higher fluorescence in the green channel was seen in 1100 FC2, 1100 FC3-2 and 1350 FC2. This is supported by amplification of the egfp gene from these 3 cultures. WT, 1100 FC3-1 and 1350 FC3 showed no evidence of egfp fluorescence which is also supported by a lack of banding during PCR (figure 2.6). Further evidence for egfp expression in these cultures can be seen by widefield fluorescence microscopy (figure 2.8). Strong signal in the egfp channel shows fluorescence in the cytosol, separate to the faint bleed-through from the plastids. In the WT control the only green fluorescence seen is linked directly to the plastids and is substantially fainter. The majority of cells in egfp transformant cultures viewed under the microscope showed clear expression of egfp.

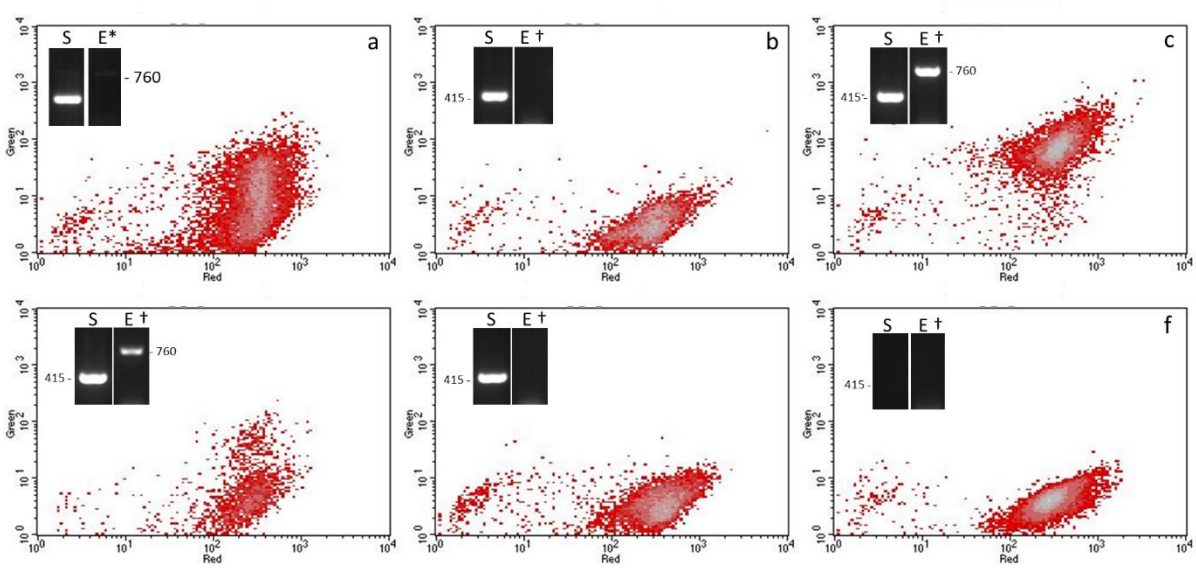


Figure 2.7. Flow cytometry of egfp (green) and autofluorescence (red) in transgenic and WT cell lines with PCR of the shble (S) and egfp (E) genes. a: 1100 FC2, b: 1100 FC3_1, c: 1100 FC3_2, d: 1350 FC2, e:, 1350 FC3 f: WT. † PCR from gDNA: see figure 2.6. * PCR from lysate 2 years after transformation: see figure 2.10.

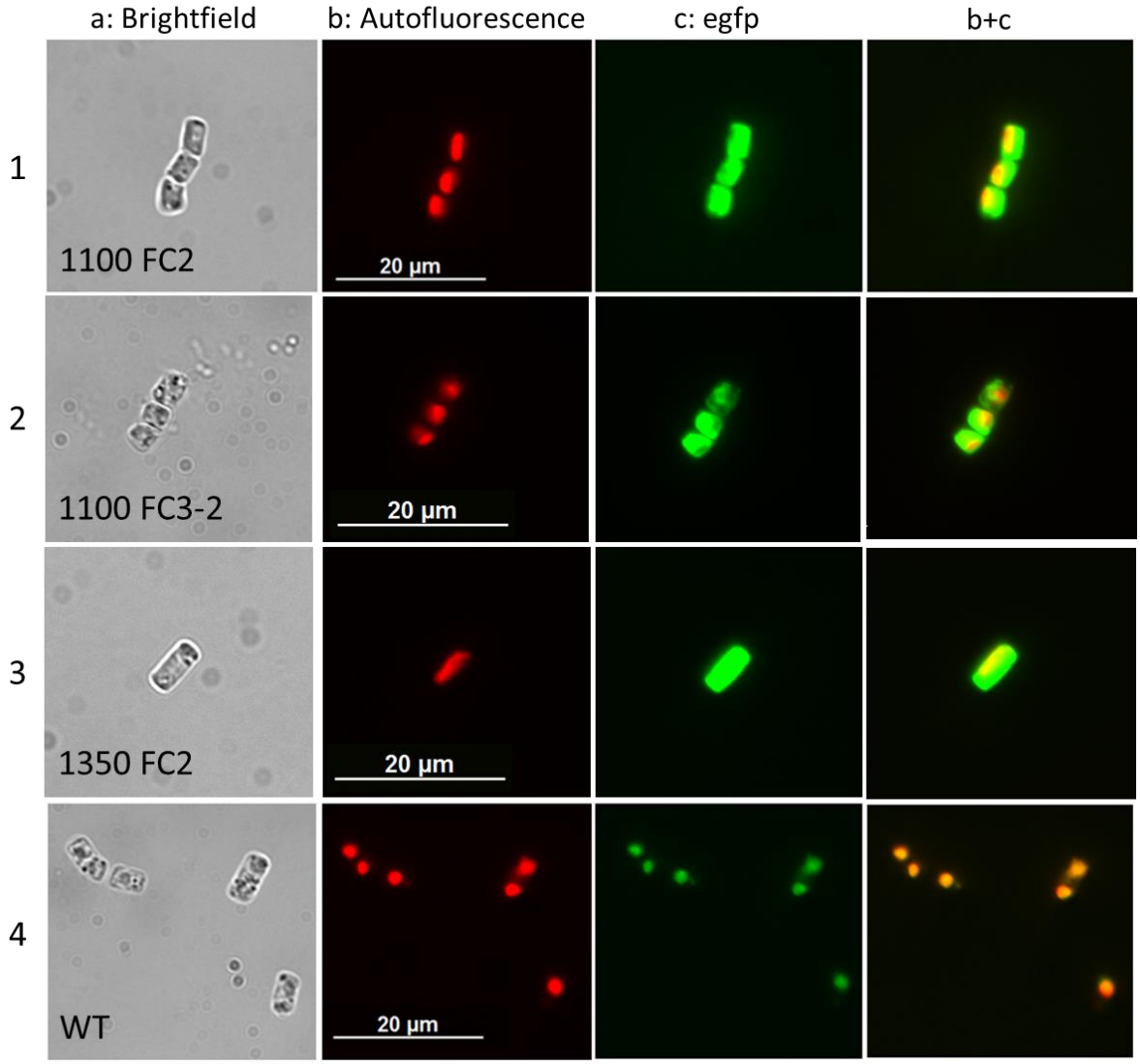


Figure 2.8. Images from widefield fluorescence microsopy. Rows 1-3 show cells from cultures 1100 FC2, 1100 FC3-2 and 1350 FC2 respectively, all of which screened positive for egfp. Row 4 shows the wildtype (WT). Columns 1-4 show brightfield, autofluorescece (red channel), egfp (green channel) and an overlay of the red and green channels respectively.

Egfp has been adapted from the original green fluorescent protein found in the *jellyfish Aequorea victoria* to be 35 times brighter, and provide an optimal codon usage for mammalian cells (Zhang et al., 1996). It has been successfully expressed in several different diatom species, individually (Apt et al., 1996; Miyagawa-Yamaguchi et al., 2011; Poulsen and Kröger, 2005; Poulsen et al., 2006; Sabatino et al., 2015; Zaslavskaia et al., 2000) and as a fusion gene (Poulsen et al., 2007; Samukawa et al., 2014). When trying several different gfp variants in *P. tricornutum* , Zaslavskaia et al. (2000) found that only egfp was functional. This may be due to differences in codon usage, as egfp is designed with a human codon bias which happens to be similar to that of *P. tricornutum* (Zaslavskaia et al., 2000). It is worth mentioning however, that *C. gracilis* also has a similar codon bias but is unable to express either egfp or sgfp (Ifuku et al., 2015), suggesting that other factors may be affecting expression.

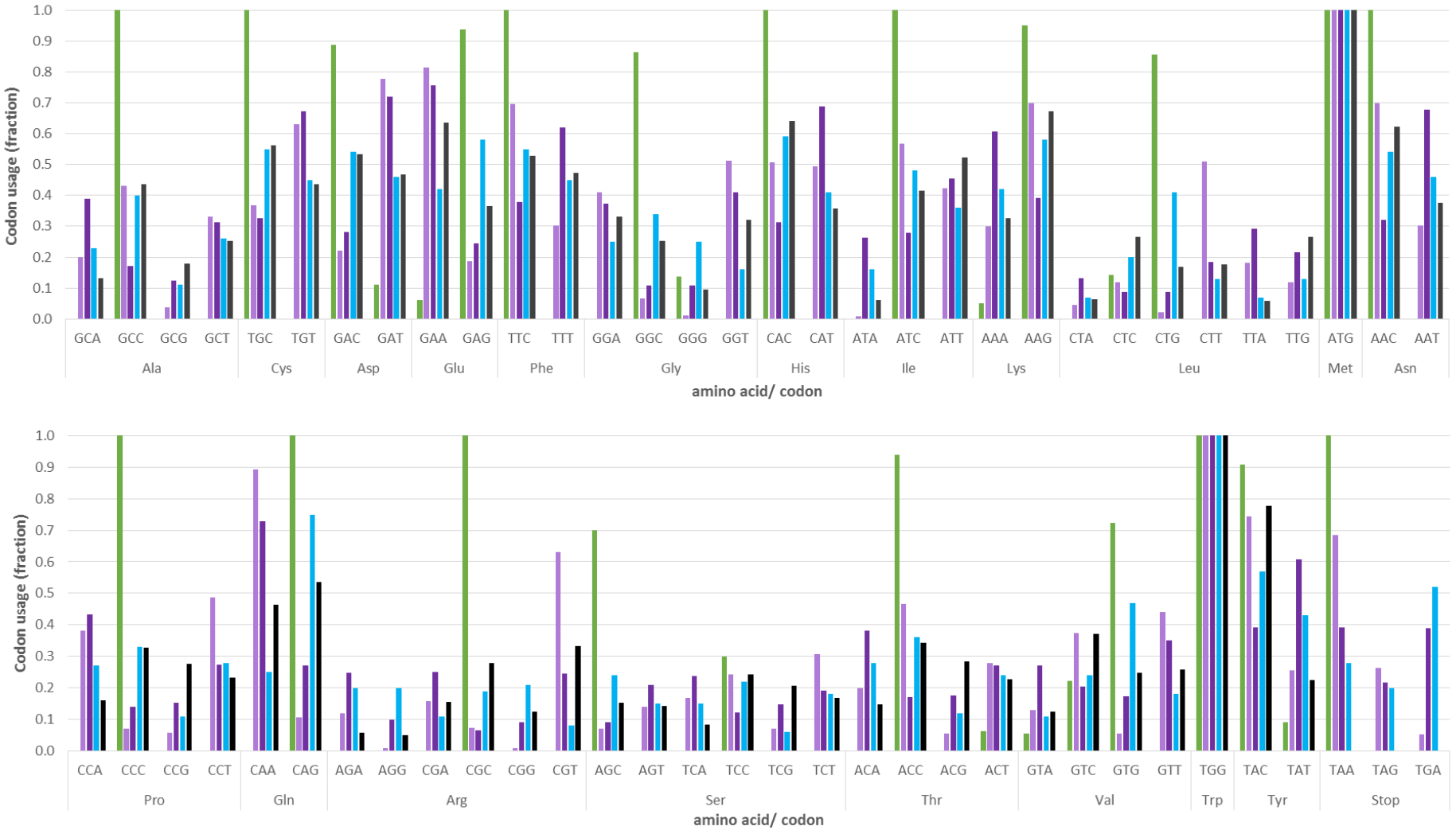


Figure 2.9. Comparison of codon usage. Codon usage as a fraction of triplets used per amino acid, is plotted for the egfp gene (green ■), all *F. cylindrus* transcripts (light purple ■), the top 20 expressed *F. cylindrus* transcripts (dark purple ■), *Homo sapiens* (blue ■) and *P. tricornutum* (black ■)

Codon usage in *F. cylindrus* was examined in comparison to *P. tricornutum*, humans and the egfp gene. Codon usage was calculated for all transcripts in *F. cylindrus*, as well as for the top 20 expressing genes under control conditions (figure 2.9). Codon usage can alter gene expression within a genome, with the potential for highly expressed genes to possess a different codon bias to those with low levels of expression (Gingold and Pilpel, 2011). *F. cylindrus* shows a slightly different codon bias to *P. tricornutum*, favouring AT rich triplets, as described in Mock et al. (2017). However, most of the codons used by egfp are reasonably well represented by the *F. cylindrus* transcriptome, both overall and for the highly expressed genes. It would however be interesting to alter egfp codon usage to favour *F. cylindrus* and monitor expression levels. Currently expression levels with the FCP promoter are high enough to confer resistance to zeocin and give a strong egfp signal.

Stability of transgenes

Shble resistant cultures were grown for two months in media without selection before transferring back to Aquil with 100µg/ml zeocin. Cells continued to grow suggesting that zeocin resistance is stable. Cultures have been maintained in zeocin however, as there are reports of diatoms losing antibiotic resistance following long-term growth in non-selective media (Falciatore et al., 1999). To check the stability of a transgene that is not directly under selective pressure, egfp was amplified from cell lysate two years after transformation (figure 2.10). All three cultures which originally contained egfp (1100 FC2, 1100 FC3-2 and 1350 FC2), screened positive for the same gene 2 years later. This suggests that at least a population of cells still contain the egfp gene. No further work has been carried out on this, however re-plating cells and testing a selection of colonies may give a better indication of egfp stability within the culture. It is possible that egfp may be protected by proximity if integrated at the same loci as shble. Ideally a long term experiment to study growth without zeocin would be required for more substantive evidence if planning to maintain transgenic cultures for long periods of time.

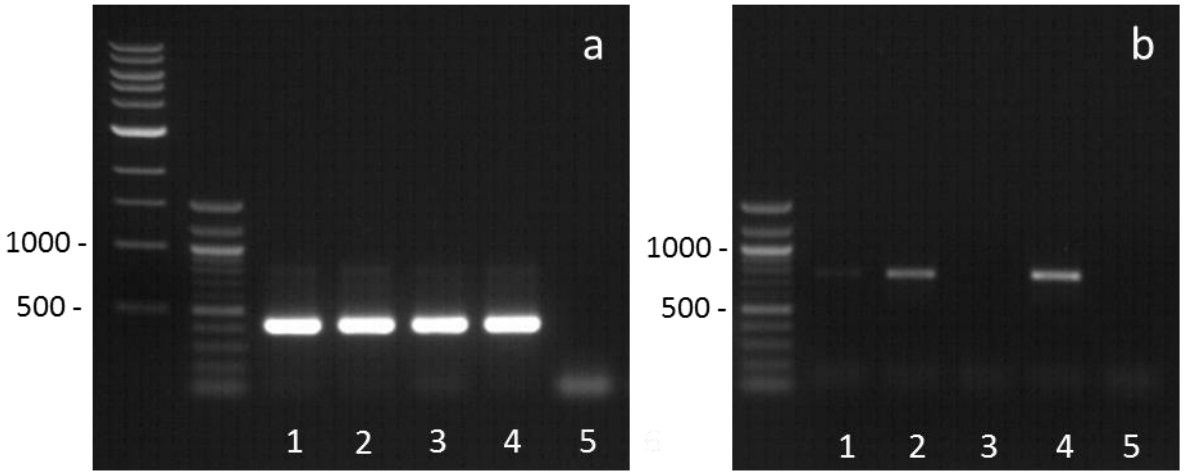


Figure 2.10. PCR from lysate of transgenic lines and WT, 2 years after transformation. a. shble. b. egfp. 1) 1100 FC2, 2) 1350 FC2, 3) 1100 FC3-1, 4) 1100 FC3-2, 5) WT

Future considerations for *F. cylindrus* transformation

This proof of principle study demonstrates that *F. cylindrus* can be transformed with adaptations to existing established diatom transformation methods. The FCP promoter, shble selective marker and egfp reporter gene are all functional, providing zeocin resistance and clear egfp fluorescence. Transformation efficiency is high enough to produce several colonies for screening, however it would be beneficial to increase efficiency for future work, particularly if working with multiple or large cassettes. Microparticle bombardment can lead to chemical and mechanical fragmentation of the plasmids upon delivery (Jacobs et al., 2015; Krysiak et al., 1999) leading to partial integration of the plasmid (Apt et al., 1996) and colonies with a selective marker but no reporter gene (Harada et al., 2005; Zhang and Hu, 2014). Integration of transgenes also appears to be random (Dunahay et al., 1995; Zhang and Hu, 2014). As discussed, approximately 50% of transformants will contain both genes. If plasmids are fragmented and multiple genes randomly integrated, then the chance of larger genes or several genes being integrated may be reduced. Later chapters of this thesis show that introduction of a Cas9 cassette for CRISPR-Cas in *T. pseudonana* occurs in about 11% of transformants (CRISPR-Cas chapter) and a SITMyb overexpression cassette in *F. cylindrus* is present for about 25% of colonies (SITMYb chapter). Both of these cassettes are large at over 5000bp. In these cases screening larger number of colonies is required to find transformants with both cassettes. As multiple copies of genes can be delivered and as they are randomly integrated, it makes sense to have access to several cell-lines when phenotyping.
Optimisation of the microparticle bombardment method or exploration of other transformation methods may increase the number of *F. cylindrus* transformants, or the way in which they are expressed. Development of the transformation toolbox to provide alternative promoters and markers would also improve the utility of this method.

Developing inducible expression with promoters such as nitrate reductase (Miyagawa et al., 2009; Niu et al., 2012; Poulsen and Kröger, 2005; Poulsen et al., 2006) or constitutive promoters besides FCP, which are not affected by light such as EF2 (Seo et al., 2015), could give greater control over gene expression. A further valuable addition to the toolbox would be to determine localisation signals to direct products to specific parts of the cell, or to identify functional tags to label cells in vivo or pull out products. Sequences to localise proteins to the cell wall (Fischer et al., 1999), nucleus (Nymark et al., 2016; Sabatino et al., 2015; Siaut et al., 2007), mitochondria (Siaut et al., 2007), endoplasmic reticulum and plastid (Apt et al., 2002; Siaut et al., 2007) have been utilised in other diatom species, as have tags to pull out and label proteins (Fischer et al., 1999; Xie et al., 2014).

Testing additional antibiotics and selective markers may allow transformant cell lines with resistance to one antibiotic to be further modified. Optimising microparticle bombardment parameters such as flight distance, particle size and pressure have been shown to increase transformation efficiency (Apt et al., 1996; Buhmann et al., 2014; Muto et al., 2013). It may also be worth considering how cells are recovered after transformation.

Although microparticle bombardment is a popular choice for transformation, the highest efficiencies come from either electroporation (Miyahara et al., 2013; Zhang and Hu, 2014) or bacterial conjugation (Karas et al., 2015). Two species have been transformed through either electroporation or multi-pulse electroporation; *P. tricornutum* (Miyahara et al., 2013; Niu et al., 2012; Zhang and Hu, 2014) and *C. gracilis* (Ifuku et al., 2015). Electroporation works by increasing the permeability of the cell membrane through the application of an electrical field. There is evidence that cell walls may limit DNA delivery through electroporation in algae (Azencott et al., 2007) and so far only lightly silicified diatoms have been transformed in this manner. Whether or not this method would be viable for species with more robust frustules remains to be seen. However, *F. cylindrus* is lightly silicified so electroporation could be a consideration. Bacterial conjugation, which has been carried out in *P. tricornutum* and *T. pseudonana* (Karas et al., 2015), involves gene transfer of the plasmid to diatoms through a bacterial intermediate. The cargo plasmid carrying the transgenes needs to contain an origin of transfer (oriT) and is transformed into *E. coli* alongside a conjugative plasmid. *E. coli* is then incubated with the diatom at 30°C for 1 ½ hours to transfer the cargo plasmid to the diatom. Inclusion of a sequence for autonomous replication of the plasmid (CEN-ARS-HIS) allows expression without integration and leads to higher numbers of transgenes (Karas et al., 2015). It may be possible to deliver episomes via electroporation as this method also allows delivery of circular plasmids, though at a lower efficiency compared to linearized plasmids (Miyahara et al., 2013). In its current format, bacterial conjugation would not be possible with *F. cylindrus* due to the high temperature required during conjugation. However, psychrophilic bacteria can also be transformed via conjugation with *E.coli*, but at 18°C (Duilio et al., 2004; Miyake et al., 2007). Therefore, it may also be possible to transform diatoms through bacterial conjugation at lower temperatures. Alternatively there may be a psychrophilic bacteria which is capable of delivering the cargo plasmid. Interestingly one of the transformable species of psychrophilic bacteria belong to the *Shewanella* genus (Miyake et al., 2007) which includes *Shewanella denitrificans*, a species associated with possible horizontal gene transfer of ice-binding proteins to diatoms, including *Navicula glaciei* and *F. cylindrus* (Janech et al., 2006).

This is the first transformation system for a psychrophilic diatom and alga. Literature searches suggest it may also be the first for any psychrophilic eukaryote. This tools provides the opportunity for more complex and targeted studies into molecular mechanisms in polar diatoms and polar eukaryotes in general. It also offers a system to express recombinant proteins in a cold-adapted eukaryote which opens up possibilities for biotechnology applications.

References

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

Apt, K.E., Zaslavkaia, L., Lippmeier, J.C., Lang, M., Kilian, O., Wetherbee, R., Grossman, A.R., Kroth, P.G., 2002. In vivo characterization of diatom multipartite plastid targeting signals. J. Cell Sci. 115, 4061–4069. doi:10.1242/jcs.00092

Azencott, H.R., Peter, G.F., Prausnitz, M.R., 2007. Influence of the Cell Wall on Intracellular Delivery to Algal Cells by Electroporation and Sonication. Ultrasound Med. Biol. 33, 1805–1817. doi:10.1097/OPX.0b013e3182540562.The

Bailleul, B., Rogato, A., de Martino, A., Coesel, S., Cardol, P., Bowler, C., Falciatore, A., Finazzi, G., 2010. An atypical member of the light-harvesting complex stress-related protein family modulates diatom responses to light. Proc. Natl. Acad. Sci. 107, 18214–18219. doi:10.1073/pnas.1007703107

Buhmann, M.T., Poulsen, N., Klemm, J., Kennedy, M.R., Sherrill, C.D., Kröger, N., 2014. A tyrosine-rich cell surface protein in the diatom Amphora coffeaeformis identified through transcriptome analysis and genetic transformation. PLoS One 9. doi:10.1371/journal.pone.0110369

Daboussi, F., Leduc, S., Maréchal, A., Dubois, G., Guyot, V., Perez-Michaut, C., Amato, A., Falciatore, A., Juillerat, A., Beurdeley, M., Voytas, D.F., Cavarec, L., Duchateau, P., 2014. Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology. Nat. Commun. 5, 3831. doi:10.1038/ncomms4831

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

Delalat, B., Sheppard, V.C., Rasi Ghaemi, S., Rao, S., Prestidge, C. a., McPhee, G., Rogers, M.-L., Donoghue, J.F., Pillay, V., Johns, T.G., Kröger, N., Voelcker, N.H., 2015. Targeted drug delivery using genetically engineered diatom biosilica. Nat. Commun. 6, 8791. doi:10.1038/ncomms9791

Demain, A.L., Vaishnav, P., 2009. Production of recombinant proteins by microbes and higher organisms. Biotechnol. Adv. 27, 297–306. doi:10.1016/j.biotechadv.2009.01.008

Duilio, A., Tutino, M., Marino, G., 2004. Recombinant protein production in Antarctic Gram-negative bacteria. Methods Mol. Biol. 267, 225–37.

Dunahay, T.G., Jarvis, E.E., Roessler, P.G., 1995. Genetic Transformation of the Diatoms Cyclotella Cryptica and Navicula Saprophila1. J. Phycol. 31, 1004–1012. doi:10.1111/j.0022-3646.1995.01004.x

Falciatore, A., Casotti, R., Leblanc, C., Abrescia, C., Bowler, C., 1999. Transformation of Nonselectable Reporter Genes in Marine Diatoms. Mar. Biotechnol. (NY). 1, 239–251. doi:10.1007/PL00011773

Fischer, H., Robl, I., Fisher, H., Al, E.T., 1999. Targeting and covalent modification of cell wall and membrane proteins heterologously expressed in the diatom Cylindrotheca fusiformis (Bacillariophyceae) 120, 113–120.

Ford, E., 2013. 2X Gibson Assembly Master Mix [WWW Document]. URL https://ethanomics.files.wordpress.com/2013/06/2x-gibson-assembly-master-mix.pdf

Gingold, H., Pilpel, Y., 2011. Determinants of translation efficiency and accuracy. Mol. Syst. Biol. 7, 481. doi:10.1038/msb.2011.14

Giuliani, M., Parrilli, E., Sannino, F., Apuzzo, G.A., Marino, G., Tutino, M.L., 2014. Recombinant production of a single-chain antibody fragment in Pseudoalteromonas haloplanktis TAC125. Appl. Microbiol. Biotechnol. 98, 4887–4895. doi:10.1007/s00253-014-5582-1

Guiry, M.D., Guiry, G.., 2017. No Title. AlgaeBase. World-wide Electron. Publ. Natl. Univ. Ireland, Galway. http//www.algaebase.org; searched Novemb. 2016.

Harada, H., Nakatsuma, D., Ishida, M., Matsuda, Y., 2005. Regulation of the Expression of Intracellular b-Carbonic Anhydrase in Response to CO 2 and Light in the Marine Diatom Phaeodactylum tricornutum 1. Plant Physiol. 139, 1041–1050. doi:10.1104/pp.105.065185.2004

Hopes, A., Mock, T., 2015. Evolution of Microalgae and Their Adaptations in Different Marine Ecosystems. eLS 1–9. doi:10.1002/9780470015902.a0023744

Hopes, A., Mock, T., 2014. Diatoms: glass-dwelling dynamos. Microbiol. Today 41, 20–23.

Ifuku, K., Yan, D., Miyahara, M., Inoue-Kashino, N., Yamamoto, Y.Y., Kashino, Y., 2015. A stable and efficient nuclear transformation system for the diatom Chaetoceros gracilis. Photosynth. Res. 123, 203–211. doi:10.1007/s11120-014-0048-y

Jacobs, T.B., LaFayette, P.R., Schmitz, R.J., Parrott, W. a, 2015. Targeted genome modifications in soybean with CRISPR/Cas9. BMC Biotechnol 15, 16. doi:10.1186/s12896-015-0131-2

Janech, M.G., Krell, A., Mock, T., Kang, J.S., Raymond, J. a., 2006. Ice-binding proteins from sea ice diatoms (Bacillariophyceae). J. Phycol. 42, 410–416. doi:10.1111/j.1529-8817.2006.00208.x

Karas, B.J., Diner, R.E., Lefebvre, S.C., McQuaid, J., Phillips, A.P.R., Noddings, C.M., Brunson, J.K., Valas, R.E., Deerinck, T.J., Jablanovic, J., Gillard, J.T.F., Beeri, K., Ellisman, M.H., Glass, J.I., Hutchison III, C. a., Smith, H.O., Venter, J.C., Allen, A.E., Dupont, C.L., Weyman, P.D., 2015. Designer diatom episomes delivered by bacterial conjugation. Nat. Commun. 6, 6925. doi:10.1038/ncomms7925

Kindle, K.L., Schnell, R.A., Fernandez, E., Lefebvre, P.A., 1989. Stable Nuclear Transformation of Chlamydomonas Using the Chlamydomonas Gene for Nitrate Reductase. J. Cell Biol. 109, 2589–2601.

Kirupamurthy, D., 2014. Studies of light responses and the development of a transformation system for the benthic diatom Seminavis robusta. Norwegian University of Science and Technology.

Krell, A., 2006. Salt stress tolerance in the psychrophilic diatom Fragilariopsis cylindrus. University of Bremen, Germany.

Kroth, P.G., 2007. Genetic Transformation A Tool to Study Protein Targeting in Diatoms. Methods Mol. Biol. 390, 257–267.

Krysiak, C., Mazus, B., Buchowicz, J., 1999. Relaxation, linearization and fragmentation of supercoiled circular DNA by tungsten microprojectiles. Transgenic Res. 8, 303–306. doi:10.1023/A:1008990712122

Lavaud, J., Materna, A.C., Sturm, S., Vugrinec, S., Kroth, P.G., 2012. Silencing of the violaxanthin de-epoxidase gene in the diatom Phaeodactylum tricornutum reduces diatoxanthin synthesis and non-photochemical quenching. PLoS One 7. doi:10.1371/journal.pone.0036806

Leblanc, C., Falciatore, a, Watanabe, M., Bowler, C., 1999. Semi-quantitative RT-PCR analysis of photoregulated gene expressio in marine diatoms. Plant Moleuclar Biol. 40, 1031–1044.

Mann, J.E., Jack, M., 1968. On pigments, growth, and photosynthesis of Phaeodactylum tricornutum. J. Phycol. 4, 349–355.

Matsumoto, M., Mayama, S., Nemoto, M., Fukuda, Y., Muto, M., Yoshino, T., Matsunaga, T., Tanaka, T., 2014. Morphological and molecular phylogenetic analysis of the high triglyceride-producing marine diatom, Fistulifera solaris sp. nov. (Bacillariophyceae). Phycol. Res. 62, 257–268. doi:10.1111/pre.12066

Medlin, L.K., Kaczmarska, I., 2004. Evolution of the diatoms: V. Morphological and cytological support for the major clades and taxonomic revision. Phycologia 43, 245–270.

Miyagawa, A., Okami, T., Kira, N., Yamaguchi, H., Ohnishi, K., Adachi, M., 2009. Research note: High efficiency transformation of the diatom Phaeodactylum tricornutum with a promoter from the diatom Cylindrotheca fusiformis. Phycol. Res. 57, 142–146. doi:10.1111/j.1440-1835.2009.00531.x

Miyagawa-Yamaguchi, A., Okami, T., Kira, N., Yamaguchi, H., Ohnishi, K., Adachi, M., 2011. Stable nuclear transformation of the diatom Chaetoceros sp. Phycol. Res. 59, 113–119. doi:10.1111/j.1440-1835.2011.00607.x

Miyahara, M., Aoi, M., Inoue-Kashino, N., Kashino, Y., Ifuku, K., 2013. Highly Efficient Transformation of the Diatom Phaeodactylum tricornutum by Multi-Pulse Electroporation. Biosci. Biotechnol. Biochem. 77, 874–876. doi:10.1271/bbb.120936

Miyake, R., Kawamoto, J., Wei, Y.L., Kitagawa, M., Kato, I., Kurihara, T., Esaki, N., 2007. Construction of a low-temperature protein expression system using a cold-adapted bacterium, Shewanella sp. strain Ac10, as the host. Appl. Environ. Microbiol. 73, 4849–4856. doi:10.1128/AEM.00824-07

Mock, T., Krell, A., Glöckner, G., Kolukisaoglu, Ü., Valentin, K., 2005. Analysis of Expressed Sequence Tags (ESTs) from the Polar Diatom Fragilariopsis cylindrus. J. Phycol. 42, 78–85. doi:10.1111/j.1529-8817.2005.00164.x

Mock, T., Otillar, R.P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., Ward, B.J., Allen, A.E., Dupont, C.L., Frickenhaus, S., Maumus, F., Veluchamy, A., Wu, T., Barry, K.W., Falciatore, A., Ferrante, M.I., Fortunato, A.E., Glöckner, G., Gruber, A., Hipkin, R., Janech, M.G., Kroth, P.G., Leese, F., Lindquist, E. a., Lyon, B.R., Martin, J., Mayer, C., Parker, M., Quesneville, H., Raymond, J. a., Uhlig, C., Valas, R.E., Valentin, K.U., Worden, A.Z., Armbrust, E.V., Clark, M.D., Bowler, C., Green, B.R., Moulton, V., van Oosterhout, C., Grigoriev, I. V., 2017. Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature 541, 536–540. doi:10.1038/nature20803

Mock, T., Thomas, D.N., 2008. Microalgae in Polar regions: Linking functional genomics and physiology with environmental conditions. 285–312.

Muto, M., Fukuda, Y., Nemoto, M., Yoshino, T., Matsunaga, T., Tanaka, T., 2013. Establishment of a Genetic Transformation System for the Marine Pennate Diatom Fistulifera sp. Strain JPCC DA0580-A High Triglyceride Producer. Mar. Biotechnol. 15, 48–55. doi:10.1007/s10126-012-9457-0

Niu, Y.F., Yang, Z.K., Zhang, M.H., Zhu, C.C., Yang, W.D., Liu, J.S., Li, H.Y., 2012. Transformation of diatom Phaeodactylum tricornutum by electroporation and establishment of inducible selection marker. Biotechniques 52, 1–3. doi:10.2144/000113881

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

Poulsen, N., Berne, C., Spain, J., Kröger, N., 2007. Silica immobilization of an enzyme through genetic engineering of the diatom Thalassiosira pseudonana. Angew. Chemie - Int. Ed. 46, 1843–1846. doi:10.1002/anie.200603928

Poulsen, N., Chesley, P.M., Kröger, N., 2006. Molecular genetic manipulation of the diatom Thalassiosira pseudonana (Bacillariophyceae). J. Phycol. 42, 1059–1065. doi:10.1111/j.1529-8817.2006.00269.x

Poulsen, N., Kröger, N., 2005. A new molecular tool for transgenic diatoms: Control of mRNA and protein biosynthesis by an inducible promoter-terminator cassette. FEBS J. 272, 3413–3423. doi:10.1111/j.1742-4658.2005.04760.x

Price, N.M., Harrison, G.I., Hering, J.G., Hudson, R.J., Nirel, P.M., Palenik, B., Morel, F.M., 1989. Preparation and Chemistry of the Artificial Algal Culture Medium Aquil. Biol. Oceanogr. 6, 443–461.

Qin, S., Jiang, P., Tseng, C., 2005. Transforming kelp into a marine bioreactor. Trends Biotechnol. 23, 264–268. doi:10.1016/j.tibtech.2005.03.010

Sabatino, V., Russo, M.T., Patil, S., d'Ippolito, G., Fontana, A., Ferrante, M.I., 2015. Establishment of Genetic Transformation in the Sexually Reproducing Diatoms Pseudo-nitzschia multistriata and Pseudo-nitzschia arenysensis and Inheritance of the Transgene. Mar. Biotechnol. 17, 452–462. doi:10.1007/s10126-015-9633-0

Sakaue, K., Harada, H., Matsuda, Y., 2008. Development of gene expression system in a marine diatom using viral promoters of a wide variety of origin. Physiol. Plant. 133, 59–67. doi:10.1111/j.1399-3054.2008.01089.x

Samukawa, M., Shen, C., Hopkinson, B.M., Matsuda, Y., 2014. Localization of putative carbonic anhydrases in the marine diatom, Thalassiosira pseudonana. Photosynth. Res. 121, 235–249. doi:10.1007/s11120-014-9967-x

San-Miguel, T., Perez-Bermudez, P., Gavidia, I., 2013. Production of soluble eukaryotic recombinant proteins in E. coli is favoured in early log-phase cultures induced at low temperature. Springerplus 2, 89. doi:10.1186/2193-1801-2-89

Scala, S., Carels, N., Falciatore, A., Chiusano, M.L., Bowler, C., Plant, M., C, M.E.N., 2002. Genome Properties of the Diatom Phaeodactylum tricornutum. Society 129, 993–1002. doi:10.1104/pp.010713.2

Seo, S., Jeon, H., Hwang, S., Jin, E., Chang, K.S., 2015. Development of a new constitutive expression system for the transformation of the diatom Phaeodactylum tricornutum. Algal Res. 11, 50–54. doi:10.1016/j.algal.2015.05.012

Sheppard, V.C., Scheffel, a., Poulsen, N., Kröger, N., 2012. Live diatom silica immobilization of multimeric and redox-active enzymes. Appl. Environ. Microbiol. 78, 211–218. doi:10.1128/AEM.06698-11

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, 23–35. doi:10.1016/j.gene.2007.05.022

Strauss, J., 2012. A genomic analysis using RNA - Seq to investigate the adaptation of the psychrophilic diatom Fragilariopsis cylindrus to the polar environment. University of East Anglia.

Taylor, N.J., Fauquet, C.M., 2002. Microparticle bombardment as a tool in plant science and agricultural biotechnology. DNA Cell Biol. 21, 963–977. doi:10.1089/104454902762053891

Vasina, J. a, Baneyx, F., 1996. Recombinant protein expression at low temperatures under the transcriptional control of the major Escherichia coli cold shock promoter cspA. Appl. Environ. Microbiol. 62, 1444–1447.

Vigentini, I., Merico, A., Tutino, M.L., Compagno, C., Marino, G., 2006. Optimization of recombinant human nerve growth factor production in the psychrophilic Pseudoalteromonas haloplanktis. J. Biotechnol. 127, 141–150. doi:10.1016/j.jbiotec.2006.05.019

Xie, W.H., Zhu, C.C., Zhang, N.S., Li, D.W., Yang, W.D., Liu, J.S., Sathishkumar, R., Li, H.Y., 2014. Construction of Novel Chloroplast Expression Vector and Development of an Efficient Transformation System for the Diatom Phaeodactylum tricornutum. Mar. Biotechnol. 16, 538–546. doi:10.1007/s10126-014-9570-3

Yao, Y., Lu, Y., Peng, K.-T., Huang, T., Niu, Y.-F., Xie, W.-H., Yang, W.-D., Liu, J.-S., Li, H.-Y., 2014. Glycerol and neutral lipid production in the oleaginous marine diatom Phaeodactylum tricornutum promoted by overexpression of glycerol-3-phosphate dehydrogenase. Biotechnol. Biofuels 7, 110. doi:10.1186/1754-6834-7-110

Zaslavskaia, L. a, Lippmeier, J.C., Kroth, P.G., Grossman, A.R., Apt, K.E., 2000. Transformation Of the Diatom Phaeodactylum Tricornutum With a Variety of Selectable Marker and Reporter Genes. J. Phycol. 386, 379–386. doi:10.1046/j.1529-8817.2000.99164.x

Zaslavskaia, L. a, Lippmeier, J.C., Shih, C., Ehrhardt, D., Grossman, a R., Apt, K.E., 2001. Trophic Conversion of an Obligate Photoautotrophic Organism Through Metabolic Engineering. Science (80-.). 292, 2073–2075. doi:10.1126/science.160015

Zhang, C., Hu, H., 2014. High-efficiency nuclear transformation of the diatom Phaeodactylum tricornutum by electroporation. Mar. Genomics 16, 63–66. doi:10.1016/j.margen.2013.10.003

Zhang, G., Gurtu, V., Kain, S.R., 1996. An enhanced green fluorescent protein allows sensitive detection of gene transfer in mammalian cells. Biochem Biophys Res Commun 227, 707–711. doi:10.1006/bbrc.1996.1573

Chapter 3: Developing CRISPR-Cas in *Thalassiosira pseudonana*

Introduction

CRISPR-Cas is arguably one of the most important methods in molecular biology since PCR. The ability to guide a double strand break (DSB) inducing nuclease to a specific site in the genome, through base complementarity, is a powerful tool with a large range of growing applications. It is an increasingly popular method, with tens of thousands of publications produced since the first papers describing its use as a precise and adaptable gene editing tool (Jinek et al. 2012) and its application to eukaryotic organisms (Cong et al. 2013; Mali et al. 2013). The majority of this project is described in the accompanying paper (Hopes et al. 2016). In these sections, the CRISPR-Cas system described in the paper is placed into a greater context in terms of the history behind the method and the decisions made in order to develop it for the diatom *T. pseudonana*.

An additional CRISPR construct is also described which includes a CEN-ARS-HIS sequence, previously shown to induce plasmid replication in diatoms (Karas et al. 2015), to investigate Cas9 expression from an episome. The discussion includes comparison to other algal genome editing methods and consideration of additional CRISPR-Cas applications that can now be applied to diatoms.

History and adaptation to eukaryotic organisms

CRISPR-Cas as a gene editing tool is adapted from a CRISPR-Cas type II viral defence mechanism found in several species of bacteria and archaea (Lander 2016). In these organisms, arrays of short fragments of viral DNA are found between repeat sequences known as clustered regularly interspersed palindromic repeats (CRISPR). The CRISPR complex responsible for inducing DSBs consists of the Cas9 nuclease, CRISPR RNA (crRNA), containing a viral spacer and repeat, and transactivating CRISPR RNA (tracrRNA). CrRNA is transcribed as a longer precursor and processed into smaller fragments with RNAase III. TracrRNA hybridises to the crRNA through a complementary sequence within the repeat region and forms part of the scaffold which forms a complex with the Cas9 nuclease. The viral spacer ‘target’ sequence then guides the nuclease to the invading viral DNA through base pairing. Cas9 is then able to latch onto a protospacer adjacent motif (PAM) in the viral DNA and cleave both strands using two active domains, RuvC and HNH, each of which is responsible for cutting a specific strand (Lander 2016). This mechanism allows organisms with CRISPR machinery to keep libraries of viral DNA and protect against future infections.

The type II CRISPR-Cas system has since been modified and stream-lined to allow precise gene editing in a wide range of eukaryotic organisms, including higher plants and algae (Nekrasov et al. 2013; Brooks, C. et al. 2014; Nymark et al. 2016; Shin et al. 2016; Wang et al. 2016).

In-vitro work demonstrated that the crRNA and tracrRNA could be combined into a chimeric single guide RNA (sgRNA; Jinek et al. 2012) negating the need for RNase III. However independent transcription of the crRNA, trRNA and Cas9 in human cells, still led to efficient genome editing, even in the absence of a bacterial RNase III, suggesting that the target cells were able to process the RNA duplex (Cong et al. 2013). In order for CRISPR-Cas to function efficiently in eukaryotes a full length tracrRNA is required containing an essential hairpin loop (Cong et al. 2013; Mali et al. 2013). Cas9 from *Streptococcus pyogenes* was codon optimised with a human codon bias for expression in human and mouse cell lines and a nuclear localisation signal (NLS) added to direct the enzyme to the nucleus (Cong et al. 2013). Furthermore, only a 20nt 'target' sequence in the crRNA is required for specific binding of the CRISPR complex to the target (Gasiunas et al. 2012). Expression of the crRNA/tracrRNA or the chimeric sgRNA is often controlled by a promoter which recruits polymerase III for transcription of small non-coding RNAs. The U6 promoter is a popular choice (Cong et al. 2013; Brooks, C. et al. 2014; Nymark et al. 2016; Wang et al. 2016; Nekrasov et al. 2013) and can be easily identified for use in endogenous systems as the U6 gene contains regions of high conservation. The bacterial RHO independent terminator includes a polyT sequence which also terminates eukaryotic pol III promoters.

Application to gene editing in *Thalassiosira pseudonana* and *Fragilariopsis cylindrus*

At the beginning of this project only one publication existed that demonstrated CRISPR-Cas in algae. Jiang et al. (2014) gave evidence for transient expression of Cas9 and gene editing in *Chlamydomonas reinhardtii*, however mutants with a functional Cas9 were not viable, leading to speculation that Cas9 is toxic in *C. reinhardtii* when expressed in-vivo. Since then CRISPR-Cas has been used to efficiently edit genes in *C. reinhardtii* via the introduction of ribonucleoproteins consisting of recombinant Cas9 and either synthetic (Shin et al. 2016) or in-vitro transcribed sgRNAs (Baek et al. 2016). Knock-out by CRISPR-cas has also been achieved in *Pheodactylum tricornutum* (Nymark et al. 2016) and *Nannochloropsis oceanica* (Wang et al. 2016) using expression based systems.

The SV40 nuclear localization signal contains the conserved sequence K-K/R-X-K/R found in classical monopartite NLS. It has been shown to direct proteins to the nucleus in both algae and higher plants (Lauersen et al. 2015; Rasala et al. 2014; Nekrasov et al. 2013), and more recently has been used to direct Cas9 in the Heterokonts *N. oceanica* (Wang et al. 2016) and *P. tricornutum* (Nymark et al. 2016). Although the Cas9 nucleases used for genome editing in *P. tricornutum* and *N. oceanica* are codon optimised for their specific species, genes with a human codon bias have been previously shown to work in several diatom species (Zaslavskaia et al. 2000; Miyagawa-Yamaguchi et al. 2011) including *T. pseudonana* (Poulsen et al. 2013; Delalat et al. 2015) and *F. cylindrus* (see chapter 2: *F. cylindrus* transformation). Additionally, it was noted that the optimised enhanced green fluorescent protein (egfp) for expression in human cells has a similar codon bias to *P. tricornutum* (Zaslavskaia et al. 2000).

The CRISPR-cas method presented here focuses on gene editing in *T. pseudonana*, although ground work has also been carried out for future editing in *F. cylindrus*. In this case CRISPR-Cas utilizes a human codon bias, *S. pyogenes*, Cas9 with a SV40 NLS driven by an endogenous fucoxanthin chlorophyll a/c binding protein promoter for high expression. Expression of chimeric sgRNAs with a 20nt target sequence are driven by an empirically determined endogenous U6 promoter and terminated by a polyT sequence.

Plasmid replication in diatoms

In addition to the construct described in Hopes et al. (2016), a construct was developed which also includes a CEN-ARS-HIS (C-A-H) sequence for maintenance and replication.

Karas et al. (2015) discovered that a C-A-H sequence, for autonomous replication in yeast also led to maintenance and low copy replication of plasmids in diatoms *P. tricornutum* and *T. pseudonana*. In yeast the autonomous replication sequence (ARS) is responsible for replication whilst the centromeric sequence (CEN) limits the number of copies and stabilises the plasmid (Stearns et al. 1990), however plasmids can be lost if there is no selective pressure (Dani & Zakian 1983). The HIS is added for selection by histidine auxotrophy in HIS deficient yeast strains. As the HIS sequence is not directly being used in diatoms, it may be possible to remove it and retain replicational functionality.

Using bacterial conjugation to transform *P. tricornutum*, Karas et al. (2015) obtained around 400-650 colonies when a C-A-H sequence was included in the plasmid compared to less than 15 colonies without. As the C-A-H sequence allows gene expression without integration, it appears that random integration into the genome is a limiting factor in transformation efficiency.

As with yeast, removing selective pressure on the episomal plasmid, in this case zeocin selection, leads to plasmid loss. About 65% of cells lost the plasmid after approximately 30 generations (Karas et al. 2015). This could be advantageous for a system such as CRISPR-Cas where the interest often lies in editing the genome, rather than the transgenes themselves. Inclusion of the C-A-H sequence has the potential to increase transformation efficiency and provide a route to remove transgenes following mutation in diatoms. Removal of Cas9 may also limit off-target mutations from long term expression.

The flexible, modular Golden-Gate cloning system, described in the accompanying paper has been used to combine the necessary elements for knock-out of the urease gene in a single construct, both with and without the C-A-H sequence described by Karas et al. (2015).

Additional methods

Construction of the urease knock-out plasmid including a CEN-ARS-HIS module

In addition to the construct described in Hopes et al. (2016) a further construct was developed which also contains a C-A-H module. pTpPuc3 (Karas et al. 2015) containing a yeast CEN6-ARSH4-HIS3, was sourced from Addgene.

Domesticating CEN6-ARSH4-HIS3

To allow Golden Gate cloning, BpiI sites were removed from ARSH4 and HIS3. The BpiI site was removed from ARSH4 using a Q5 site-directed mutagenesis (SDM) kit and accompanying protocols (NEB). Forward and reverse primers TCTGTGTAGAtGACCACACAC and CAAGATGAAACAATTCGGC were used, where the lower case letter denotes the base change. Following transformation, positive colonies were selected on LB plates with 100ug/ml ampicillin. Colonies were picked, grown in 5ml of LB media overnight at 37°C and the plasmid extracted using a Promega PureYield plasmid Miniprep kit. The resultant plasmid, pTPpuc3_ARSmut, was screened using restriction digest with BpiI (Thermo Fisher). The BpiI site was then removed from HIS3 in pTPpuc3_ARSmut to make 'domesticated pTPpuc3'. Initial attempts using the Q5 SDM kit and primers ACACCACTGAgGACTGCGGGA and GATGGTCGTCTATGTGTAAGTCAC, led to a truncated plasmid. Re-designed primers PO_4-ATCACACCACTGAgGACTGC and PO_4-GGTCGTCTATGTGTAAGTCAC were ordered with a 5' phosphate group to allow SDM to be carried out without a kit as follows. Phusion (NEB) PCR was carried out in a 20µl volume with the above primers and pTPpuc3_ARSmut as the template. Denaturation for 2 minutes was followed by 35 cycles with denaturation at 98°C for 10 seconds, annealing at 63°C for 30 seconds and extension at 72°C for 60 seconds, this was followed by a final extension at 72°C for 5 mins. After the final extension, the reaction was cooled to 37°C and 10 units of DpnI (NEB) was added. The reaction was incubated at 37°C for 3 hours. A 10µl ligation reaction was carried out with T4 DNA ligase (NEB) and 1µl of the DpnI treated PCR reaction at room temperature for 2 hours. 5µl of PCR reaction was run on a 0.8% agarose gel to check the product size before transformation. 2.5µl of the ligation reaction was transformed into 25µl of One Shot TOP10 competent *E. coli* (Thermo Fisher) according to the manual. Transformed cells were selected on 100µg/ml ampicillin LB agar plates. Colonies were screened by PCR prior to mini-prep plasmid extraction to ensure clones contained the full-length C-A-H. Go Taq Colony PCR was performed with primers aggtctcaggagCGCGAGCATCACGTGCTATAA and aggtctcaagcgGTCAAGTCCAGACTCCTGTGTAAA, designed to amplify the C-A-H region for L1 Golden-gate assembly. Two colonies which screened positive for the full-length C-A-H were picked and plasmids extracted in a mini-prep as described above. Plasmids were sent to Eurofins for sequencing with reverse primer GCCAATATATCCTGTCAAACAC, which anneals downstream of the insert.

Golden-gate cloning

C-A-H was PCR amplified from domesticated pTPpuc3 using Phusion DNA polymerase (NEB) and inserted into a L1 pICH47772 destination vector as described in Hopes et al. (2016). L1 plasmids were screened for the correct insert by digestion with BpiI. L1 modules pICH47732:FCP:NAT, pICH47742:FCP:Cas9YFP, pICH47751:U6:sgRNA_Urease 1, pICH47761:U6:sgRNA_Urease 2, pICH47772:CAH and the L5E linker pICH41800 were assembled into the L2 destination vector
pAGM4723 as described (Figure 3.1.). Constructs were screened by digestion with EcoRV. In addition to the sequencing primers described in Hopes et al. (2016), primer agcgGTCAAGTCCAGACTCCTGTGTAAA, was used to sequence pAGM4723:TpCC_Urease_CAH.

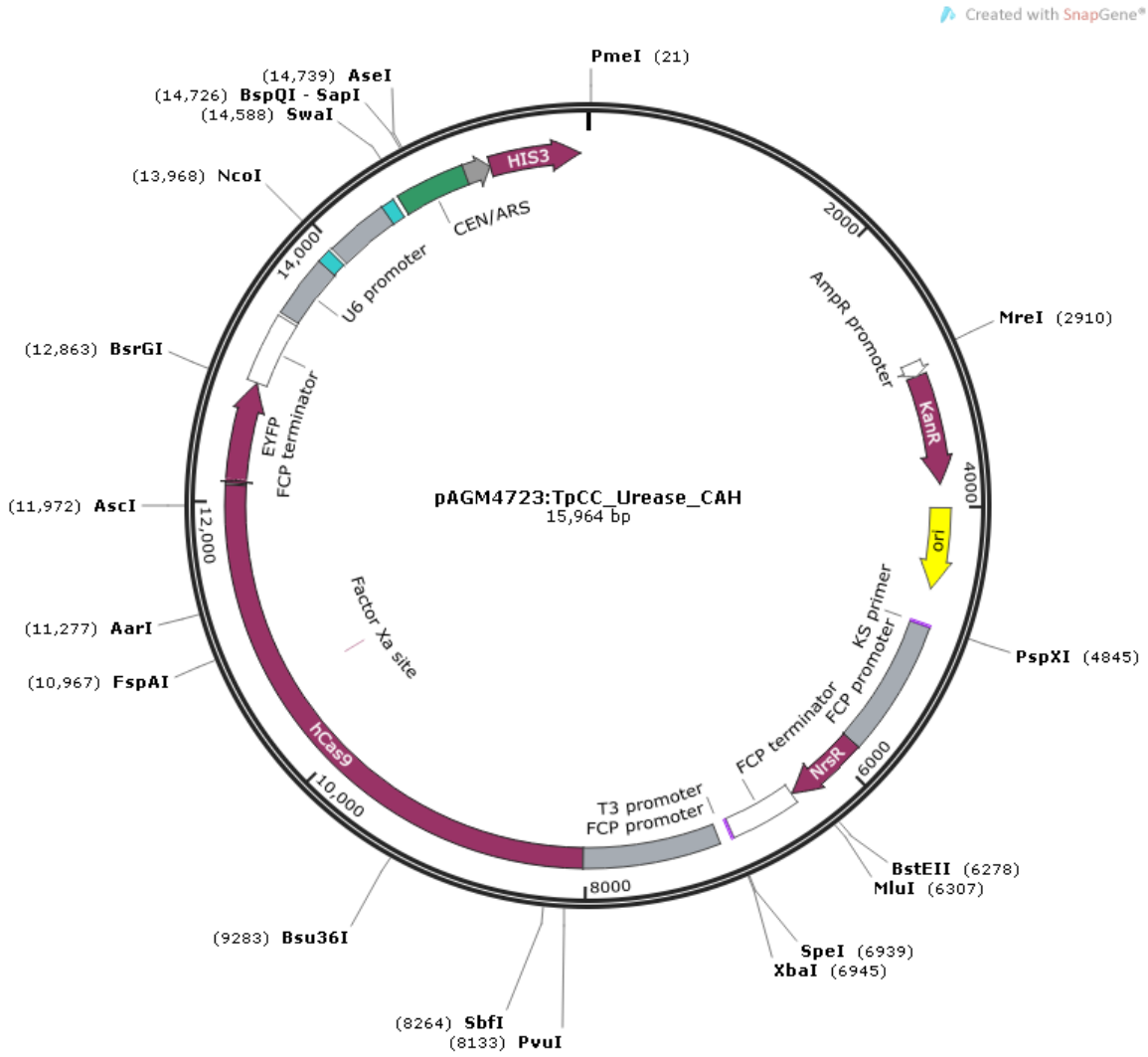


Figure 3.1. Plasmid map of pAGM4723:TpCC_Urease_CAH

Transformation, screening and phenotyping

Transformation, screening and phenotyping was carried out according to Hopes et al. (2016; See chapter 3).

Testing for the presence of self-replicating plasmids

M1, M2 and M3 primary clones were generated with the C-A-H containing construct. 20ml cultures from M1-M4 primary clones were grown to exponential phase as described (Hopes et al. 2016).

Based on the episome extraction from Karas et al. (2015), a modified alkaline lysis method using the GeneJET plasmid Miniprep kit (Thermo Fisher) was used to isolate potential plasmid DNA. Cultures were harvested by centrifugation at 4,000g for 5 mins. Cells were resuspended in 250µl of Resuspension Solution which contained 5µl of lysozyme (125µg), vortexed until homogenous and incubated at 37°C for 30 minutes. Two hundred and fifty µl of Lysis Solution containing RNase A was added and mixed by inverting the tube 6 times before incubating at room temperature for 5 minutes. Supernatant was transferred to a GeneJET spin column and centrifuged at 12,000xg for 1 minute. Columns were washed by adding 500µl of Wash Solution and centrifuging at 12,000xg for 1 minute. The wash step was repeated once more before spinning the column for 2 minutes at 12,000xg to remove residual ethanol in the Wash Solution. DNA was eluted by adding 50µl of TE to the centre of the column, incubating at room temperature for 2 minutes and spinning for 2 minutes at 12,000xg. 5µl of supernatant from each sample was transformed into NEB 5-α competent *E. coli* according to the NEB protocol. Two hundred µl of cells were plated onto LB plates with 50µg/ml kanamycin.

Screening *T. pseudonana* and *F. cylindrus* for the classical monopartite NLS signal

Several proteins (14-15) associated with the nucleus including DNA binding proteins, enzymes associated with DNA repair, transcription factors and helicases were inspected for the consensus sequence K-K/R-X-K/R, found in classical monopartite NLS, in both *T. pseudonana* and *F. cylindrus*.

Design of construct for silacidin knock-out in *T. pseudonana*

Although it has yet to be used, two constructs for silacidin knock-out in *T. pseudonana* were designed. Although I designed the constructs, sgRNAs and primers, assembly of the constructs was carried out by Marianne Jaubert at the Centre de Recherche des Cordeliers in Paris. The same constructs and Golden-Gate cloning method as described in Hopes et al. (2016) have been used with only the sgRNAs altered for targeting the silacidin gene.

The 708bp silacidin gene (chr_2: 1522471 - 1521764) contains multiple repeat sequences starting at +170bp and continuing to the end of the gene. Two sets of sgRNAs were designed; set 1 is intended to target two sites within the silacidin (but not within a repeat region to avoid multiple cut sites) at +20 and +544, to induce a 518bp deletion and frameshift, potentially disrupting remaining

repeat regions. SgRNA set 2 targets regions up and downstream of the coding region at -57 and +729 to cut out a 780bp region encompassing the entire silacidin gene.

The following plasmids developed during this chapter were sent to Paris for assembly; pICH47732:FCP:NAT, pICH47742:FCP:Cas9YFP and L0 U6 promoter, as was the plasmid for amplification of the sgRNA backbone; pICH86966_AtU6p_sgRNA_NbPDS (Nekrasov et al. 2013), L1 destination vectors pICH47751 and pICH47761 for U6:sgRNA assemblies and L2 destination vector pAGM4723 and linker pICH41780.
Forward primers for amplification of sgRNAs for L1 assembly are as follows: sgRNA 1; **aggtctc**attgtG**CGAGGACTGCAATGAAGGCG***GTTTTAGAGCTAGAAATAGCAA*, sgRNA 2; **aggtctc**attgtG**CCGTCGTCTCTCGTCCTCCG***GTTTTAGAGCTAGAAATAGCAAG*. Forward primers for set 2 are: sgRNA 1; **aggtctc**attgtG**AGGGGGAACGAGTTGCTGTG***GTTTTAGAGCTAGAAATAGCAAG* and sgRNA 2; **aggtctc**attgtG**TGATGTTTGATGCATGGGCG***GTTTTAGAGCTAGAAATAGCAAG*.
All oligos include **BsaI**, the 4nt overhang, the **target** and the forward complementary *sgRNA backbone* sequence. Additionally a 5' G has been added directly before each of the target sequences as pol III requires a G to start transcription. The reverse primer for amplification of all sgRNA products is **tggtctc**aagcgtaatgccaactttgtacaag (Urease sgRNA R).

Preliminary work for CRISPR-Cas in *F. cylindrus*

Several CRISPR-Cas constructs for *F. cylindrus* have been created by Irina Grouneva and Nigel Belshaw using the Golden-Gate *T. pseudonana* CRISPR-Cas model described in the accompanying paper and other plasmids/sequences developed during this PhD. This includes the Golden-Gate pICH47732_FCP_Shble cassette for zeocin resistance, created for SITMyb overexpression (see chapter 4: SITMyb) and FCP sequences, that were identified and tested during proof of principle *F. cylindrus* transformation (See transformation chapter), to drive Cas9. As discussed, several *F. cylindrus* nucleus associated proteins were screened for the consensus sequence associated with SV40 NLS. As with *T. pseudonan*a, a blastn search was also carried out on the *F. cylindrus* genome to determine the U6 promoter. Nigel Belshaw later empirically determined the exact end of the promoter using the same 5' RACE method used for *T. pseudonana*.

Two sgRNAs were designed to delete a region incorporating both the potential silicon transporter (SIT) and Myb domains of the SITMyb gene. Forward primers incorporating the target sequences for sgRNA 1: **aggtctc**atatt**GCTCCCTGCATATGACTCAA***GTTTTAGAGCTAGAAATAGCAAG* and sgRNA 2: **aggtctc**ttattG**AGACTACTGTGACGAGAGCG***GTTTTAGAGCTAGAAATAGCAAG* were designed including **BsaI** sites, the 4nt overhang, **target** sequence, and *sgRNA backbone* complement. Irina Grouneva built the *F. cylindrus* L1 FCP:Cas9:YFP module, constructed the L1 U6:sgRNA modules using the primers designed above and the reverse primer 'Urease sgRNA R', and put together the final L2 construct. Irina has transformed the construct into *F. cylindrus* using the method described in the *F. cylindrus* Transformation chapter. Results are currently pending.

Additional Results and Discussion

Screening for NLS signals

Potential canonical monopartite nuclear localisation signals were found in each nucleus associated protein screened for both *T. pseudonana* and *F. cylindrus*, including the sequence KKRK associated with the SV40 NLS. Although empirical testing is needed to validate potential NLS signals in DNA binding associated proteins, presence of these signals indicated that there was a chance the SV40 NLS is functional in these species. As demonstrated in the Hopes et al. (2016) paper, CRISPR-Cas is able to efficiently edit the genome of *T. pseudonana*, suggesting that Cas9 is being directed to the nucleus. CRISPR-Cas in *P. tricornutum* also utilises an SV40 NLS signal to localise the Cas9 nuclease (Nymark et al. 2016). Given that *T. pseudonana* and *P. tricornutum* are able to use the SV40 NLS, it is logical to also use the SV40 NLS in the *F. cylindrus* CRISPR system, especially given that *P. tricornutum* and *F. cylindrus* are closer in evolutionary terms, with both raphid pennate diatoms belonging to the same class (Bacillariophyceae).

Construction of plasmids for knock-out of silacidin, and SITMyb.

Constructs were successfully assembled for gene-editing of silacidin in *T. pseudonana* and SITMyb in *F. cylindrus* by three independent parties. This demonstrates that the Golden-Gate system is well-suited to assembly of CRISPR-Cas constructs, as the original *T. pseudonana* model has been easily applied to another gene in *T. pseudonana* and adapted for *F. cylindrus*.

CRISPR construct with CEN-ARS-HIS

Two replicates from the transformation with the C-A-H construct crashed, leaving only one replicate for plating following the 24 hours incubation. One replicate from the construct described in the accompanying paper also crashed. As a result, there were not enough samples to determine if an increased transformation efficiency occurred when the C-A-H sequence was included. However, the transformation efficiency of the successful C-A-H construct was 113 colonies/10^8 cells, whilst the average efficiency of the two –C-A-H replicates was 100 colonies/10^8 cells. Although this is a very small sample size there was no obvious increase in transformation efficiency compared to the ~30x increase seen for *P. tricornutum* using bacterial conjugation.

Plasmid extraction on the primary clone cultures yielded less than 2ng/µl DNA, which is below the detection limit of the nanodrop spectrophotometer, for all C-A-H construct derived cultures as well as the culture derived from the –C-A-H construct. Transformation of plasmid extractions into *E. coli* did not yield colonies. This suggests that either autonomous replicating plasmids were not present, or they were not adequately extracted.

The protocol for NEB 5-alpha *E. coli* transformation recommends at least 1pg of plasmid per 50µl of competent cells. If a single copy of each C-A-H plasmid was present in 20ml of cells at 1 x 10^6 cells per ml, and if plasmid extraction was 100% efficient, you would expect a total of 34pg to be used in each *E.coli* transformation. According to data from NEB approximately 34pg gives a

transformation efficiency of 1 x 10^9 cfu/µg of pUC19, equating to 3000 colonies per transformation. As the CRISPR-Cas plasmids are larger than puc19, you would therefore expect a lower transformation efficiency - according to NEB, a plasmid of 15.9kb has a 10x lower transformation efficiency than puc19. Therefore, approximately 300 colonies may be expected from the CRISPR construct, or 60 colonies if plating out 200µl of the transformation culture. In addition, a reduction in extraction efficiency is possible given that plasmid extraction kits are typically designed for bacteria and low gDNA extraction yields are seen for *T. pseudonana* compared to less silicified diatoms such as *F. cylindrus*.

Another possible explanation is that constructs are chemically and mechanically fragmented when introduced into the cell through microparticle bombardment (Krysiak et al. 1999; Jacobs et al. 2015). In order for plasmids to replicate, they need to remain circular. This is further supported by only 11% of clones tested screening positive for Cas9, but all clones screening positive for NAT, suggesting that fragments of the plasmid are present.

If there were cells with circular self-replicating C-A-H plasmids, un-fragmented by microparticle bombardment, it would be difficult to determine presence of these plasmids, given the low DNA yield following plasmid extraction. Methods such as qPCR, to screen for small quantities of DNA, have a chance of producing false positives, given that fragments of gDNA may be present in the final eluate, and could lead to positive results if C-A-H were integrated into the genome. Although, a general PCR of the C-A-H sequence in mutant colonies may give an insight into presence and absence of the replicating sequence, whether as a plasmid or integrated into the genome.

It's also worth considering that presence of the C-A-H sequence may alter fitness of the diatom, particularly if there are also cells with fragmented, integrated transgenes present that may have a different fitness. In yeast presence of autonomously replicating plasmids have been shown to lower viability and growth rate (Stearns et al. 1990; Falcón & Aris 2003).

Currently there is no evidence of self-replicating plasmids in these cultures, either through transformation efficiency or extraction and transformation into *E. coli*. Further work is required to investigate. Extraction methods such as the French press, which can disrupt the cell wall but leave the nucleus intact, may give better yields. A repeat transformation, with larger numbers of replicates, would be needed to see if there is an increase in efficiency with the C-A-H plasmid. Changing the transformation method to bacterial conjugation (Karas et al. 2015), to prevent fragmentation of the plasmid, may also be required in order investigate the possible advantages and disadvantages of expressing the CRISPR-Cas transgenes on an episome.

CRISPR-Cas has been established in three other algal species: *C. reinhardtii* (Jiang et al. 2014; Shin et al. 2016; Baek et al. 2016), *P. tricornutum* (Nymark et al. 2016) and *N. oceanica* (Wang et al. 2016). In each of these cases one sgRNA has been used to induce a mutation through error-prone non-homologous end joining (NHEJ). As mutations are often small insertions or deletions, screening methods such as restriction site loss (Wang et al. 2016), high resolution melt curves

(Nymark et al. 2016) and sequencing (Baek et al. 2016) are required to identify gene editing at a molecular level. Based on previous work in higher plants (Belhaj et al. 2013), two sgRNAs have been used to create a deletion in the urease gene. This allows for simple screening from cell lysate by band shift assay. This involves PCR of the target sequence and visualisation on a gel to determine if a deletion has occurred. Restriction site loss was also carried out, as was sequencing to verify the mutation. A disadvantage of restriction site loss is that the cut site of the sgRNA must sit within a restriction site. At best, this limits the location of the sgRNA, particularly as use of common 4nt cutters is limited if using digestion to enrich mutant sequences, given that any PCR product produced must contain only one copy of the restriction site. If digesting after PCR and running out bands on a gel, common cutters may lead to multiple small fragments which may be difficult to resolve. This was the case for the BccI digestion of the urease fragment. Differences in banding were observed but fragments were small leading to poor resolution. In cases where the target site needs to be very specific, it may not be possible to find a suitable sgRNA which coincides with a restriction site. This was the case when designing the construct for silacidin knock-out. Baek et al. (2016) used an effective and straightforward approach to screening by targeting genes CpFTSY and ZEP that result in colonies with visible phenotype, including a change in colour and chlorophyll a fluorescence respectively.

As with *P. tricornutum* and *N. oceanica*, CRISPR-Cas in *T. pseudonana* is achieved by in-vivo expression of transgenic Cas9 and sgRNAs. In *C. reinhardtii* the CRISPR-Cas complex is inserted as a ribonucleoprotein (RNP), formed from recombinant and either synthetic (Shin et al. 2016) or in-vitro transcribed (Baek et al. 2016) sgRNAs, due to potential toxicity of Cas9 expression (Jiang et al. 2014). Low numbers of mutants in *C. reinhardtii* following expression of a zinc finger nuclease (ZFN) and knock-out of the COP3 gene may also be due to toxicity caused by prolonged exposure of the ZFN (Sizova et al. 2013). Transcription activator-like effectors (TALE) have been used to over-express genes in *C. reinhardtii* (Gao et al. 2014) but TALE nucleases have yet to be applied to gene knock-out. It would be interesting to see if the TALEN gene editing method also leads to reduced fitness.

Mutation efficiencies from CRISPR-Cas in *T. pseudonana* fall within those seen in other algal systems. CRISPR-Cas using RNPs in *C. reinhardtii* gave very high mutation efficiencies at 0.46 – 0.56%, especially given that antibiotic selection was not used to select for mutants (Baek et al. 2016). Mutation efficiency with expression based CRISPR-Cas in *P. tricornutum* was 25-63%, whilst *N. oceanica* has a success rate of 0.1-1% (Wang et al. 2016). Considering that knock-out by HR, without initiation by DSBs, in the same species was 11-95% (Kilian et al. 2011), this is rather low. Gene-editing by meganucleases (MN) and TALENS in *P. tricornutum* was 42% and 7-56% respectively. In comparison, *T. pseudonana* mutation efficiency is around 11% in primary clones and 100% in clones containing Cas9.

For species with diploid genomes, the homozygous nature of the mutation also needs to be considered. Both *C. reinhardtii* and *N. oceanica* (Kilian et al. 2011) are haploid, as a result, as long as only one copy of the target gene exists in the genome, only one gene-editing event is required for a knock-out. If cells are diploid both alleles need to be edited, with mono-allelic mutations potentially failing to show a knock-out phenotype as seen by Weyman et al. (2015) when editing the urease gene. Following transformation it is possible for primary colonies of diploid species to be mosaic, in which nuclease induced mutations occur following cell division, leading to a mixture of cells with mutant and wild-type copies (Nymark et al. 2016; Daboussi et al. 2014). In *T. pseudonana* occurrence of bi-allelic deletions in sub-clones was 25-65%. This compares to the majority of cells showing bi-allelic mutants in *P. tricornutum* using CRISPR-Cas. In *P. tricornutum* colonies, up to 15% of cells following MN editing (Daboussi et al. 2014) and up-to 80% following TALEN editing (Daboussi et al. 2014; Nymark et al. 2016) showed mutations, although it is not specified if they are bi-allelic. Bi-allelic mutation efficiences in *T. pseudonana* are based on deletions, so do not consider potential mutants at individual sgRNA sites. When editing Solanum lycopersicum with two sgRNAs for a deletion, Brooks et al. (2014) found that although a high mutation efficiency was observed, occurrence of mutants with bi-allelic deletions was only 3%. Working with single cell autotrophs, such as diatoms, gives an advantage when dealing with mosaicism in comparison to higher plants, as bi-allelic sub-clones can be easily isolated by re-spreading on plates and picking new colonies.

The transformation method used to deliver transgenes may affect mutation efficiency. As mentioned, microparticle bombardment is known to cause chemical and mechanical fragmentation (Jacobs et al. 2015; Krysiak et al. 1999) of the plasmid. This may account for 11% of the primary colonies containing Cas9, thus decreasing the potential mutation efficiency. Micro-particle bombardment was also used to transform *P. tricornutum*, whilst *C. reinhardtii* and *N. oceanica* used electroporation. In future work, it may be worth creating a Cas9 cell line and later introducing sgRNAs to increase mutation efficiency. A similar approach has been taken in *Nicotiana benthamiana* in which a Cas9 clone was generated before introducing the sgRNA through a viral vector (Ali et al. 2015). On the opposite end of the spectrum, generation of mutants through transient Cas9 and sgRNA expression could be beneficial if transgene free cell-lines are required, either for industrial purposes where genetically modified (GMO) algae may not be desirable, or to remove potential long-term effects of Cas9 expression. As discussed, inclusion of a CEN-ARS-HIS sequence in plasmids may be a route to transient expression in diatoms. There are however, alternative methods currently being used in other species such as delivery of Cas9 mRNA (Chiu et al. 2013), delivery of RNPs (Baek et al. 2016; Shin et al. 2016), agroinfiltration in plants (Jiang et al. 2013) and use of viruses (Yin et al. 2015) and viral vectors (Ali et al. 2015; Maggio et al. 2014; Gong et al. 2017).

As well as carrying genes for nuclease activity, geminiviral vectors have been used to introduce donor sequences for homology directed repair. This system is particularly interesting as viral

sequences, located on plasmids, are able to circularise and replicate in the host cell, provided that a viral replication-initiation protein (REP) is also expressed. This can increase not only the occurrence of DSBs but also HR due to the presence of multiple donor sequences. REP, the HR donor sequence and the CRISPR-Cas genes can be expressed on separate viral vectors (Baltes et al. 2014) or as a single vector (Čermák et al. 2015). In plants, geminivirusus have been used to deliver and express the relevant genes (Yin et al. 2015). As proteins for replication are provided, this method doesn't rely on machinery from the host cell, much like CRISPR-Cas. As a result, it may be transferable from higher plants to diatoms.

Several diatom viruses have been identified which may be worth exploring along with the geminiviral vectors for transgene delivery and transient expression. To date viruses have been identified for centric diatoms belonging to the genera Rhizosolenia (Nagasaki et al. 2004), Chaetoceros (Tomaru et al. 2009; Shirai et al. 2008; Bettarel et al. 2005; Tomaru et al. 2011; Tomaru et al. 2013; Kimura & Tomaru 2015) and *Skeletonema* (Kim et al. 2015). In pennate diatoms, viruses which infect species from the genera *Thalassionema* and *Asterionellopsis* (Tomaru et al. 2012) have been identified. These include both ssRNA and circular dsDNA viruses. Viruses which infect *T. pseudonana* would first need to be identified if viral delivery were to be established. The only diatom species with both a transformation system (Ifuku et al. 2015) and known viruses is *Chetoceros gracilis*.

Since its application to eukaryotic gene editing in 2013 (Cong et al.), CRISPR-Cas has largely been used to induce mutations through error-prone NHEJ (Nymark et al. 2016), deletions (Brooks, C. et al. 2014; Zheng et al. 2014; Ordon et al. 2016) and homologous recombination (Cong et al. 2013; Baltes et al. 2014). However, CRISPR-Cas is advancing both as a gene-editing tool and for additional applications. It has been shown that large deletions can be achieved up to 245kb, easily encompassing whole genes, loci and even large fragments of chromosomes (Zhou et al. 2014; Ordon et al. 2016). Multiple genes can be targeted, with the potential to disrupt whole pathways, either through multiple sgRNA cassettes (Zheng et al. 2014) or through tracrRNA and CRISPR arrays (Cong et al. 2013). Efforts to improve HR efficiency following DSBs include exposure to multiple donor sequences (Baltes et al. 2014; Čermák et al. 2015) and disruption of NHEJ by inhibition of DNA ligase IV (Maruyama et al. 2015). The Cas9 itself has also been adapted. A nickase with an inactive RuvC domain has been developed to cut one DNA strand (Cong et al. 2013). By using this nickase in combination with two single guide RNAs which target alternative strands, a DSB can be induced. This is useful if off-target cutting is an issue, as nicks tend to repair cleanly and DSBs will only be created if the two sgRNAs are in close proximity. A further Cas9 has been adapted, termed dCas9 (deactivated Cas9), with both domains deactivated. This means that the Cas9 complex can be used for its targeting and binding properties without the nuclease activity. DCas9 can be fused to transcriptional activators or repressors to control transcription (Qi et al. 2013; Piatek et al. 2015) or to fluorescent markers to profile specific genomic loci without denaturation (CasFISH; Deng et al. 2015). It can also be used for DNA modification by attaching a

cytidine deaminase for a C to U base change (Komor et al. 2016) or by adding a DNA methyltransferase for methylation (Vojta et al. 2016).

Summary

Efficient CRISPR-Cas gene-editing in *T. pseudonana* has been established. Efficiencies compare well with other algal CRISPR systems. The band-shift screening method, following a deletion with two sgRNAs works well and gives a clear visual method for screening mutants. In addition, using two sgRNAs to induce deletions gives a higher degree of control over mutations compared to the random indels produced when using individual sgRNAs. The Golden-Gate cloning method is well suited to CRISPR-Cas applications, providing a flexible platform for construct modification and use of multiple sgRNAs. This system has since been used to create constructs for two other genes in *T. pseudonana* as well as constructs for the polar diatom *F. cylindrus*. Further work is needed in terms of developing a transient expression system for CRISPR-Cas in diatoms. Now that an efficient and flexible method is in place for gene editing in *T. pseudonana* the system can be used for targeting other genes and applications beyond the scope of basic gene editing. This will hopefully be a valuable addition to the diatom molecular toolbox.

References

Ali, Z. et al., 2015. Efficient Virus-Mediated Genome Editing in Plants Using the CRISPR/Cas9 System. *Molecular Plant*, 8(8), pp.1288–1291.

Baek, K. et al., 2016. DNA-free two-gene knockout in Chlamydomonas reinhardtii via CRISPR-Cas9 ribonucleoproteins. *Scientific reports*, 6, p.30620. Available at: http://www.nature.com/articles/srep30620\nhttp://www.ncbi.nlm.nih.gov/pubmed/27466170.

Baltes, N.J. et al., 2014. DNA Replicons for Plant Genome Engineering. *The Plant Cell*, 26(1), pp.151–163. Available at: http://www.plantcell.org/cgi/doi/10.1105/tpc.113.119792.

Belhaj, K. et al., 2013. Plant genome editing made easy: targeted mutagenesis in model and crop plants using the CRISPR/Cas system. *Plant methods*, 9(1), p.39. Available at: http://plantmethods.biomedcentral.com/articles/10.1186/1746-4811-9-39.

Bettarel, Y. et al., 2005. Isolation and preliminary characterisation of a small nuclear inclusion virus infecting the diatom Chaetoceros cf. gracilis. *Aquatic Microbial Ecology*, 40(2), pp.103–114.

Brooks, C. et al., 2014. Efficient Gene Editing in Tomato in the First Generation Using the Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-Associated9 System1. *Plant Physiol*, 166, pp.1292–1297.

Čermák, T. et al., 2015. High-frequency, precise modification of the tomato genome. *Genome biology*, 16(1), p.232. Available at: http://genomebiology.com/2015/16/1/232\nhttp://www.ncbi.nlm.nih.gov/pubmed/26541286\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4635538.

Chiu, H. et al., 2013. Transgene-free genome editing in Caenorhabditis elegans using CRISPR-Cas. *Genetics*, 195(3), pp.1167–1171.

Cong, L. et al., 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*, 339, pp.819–823.

Daboussi, F. et al., 2014. Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology. *Nature communications*, 5, p.3831. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24871200.

Dani, G.M. & Zakian, V.A., 1983. Mitotic and meiotic stability of linear plasmids in yeast. *Proc Natl Acad Sci U S A*, 80, pp.3406–3410.

Delalat, B. et al., 2015. Targeted drug delivery using genetically engineered diatom biosilica. *Nature Communications*, 6, p.8791. Available at: http://www.nature.com/doifinder/10.1038/ncomms9791.

Deng, W. et al., 2015. CASFISH: CRISPR/Cas9-mediated in situ labeling of genomic loci in fixed cells. *Proceedings of the National Academy of Sciences of the United States of America*, 112(38), pp.11870–11875. Available at: http://www.pnas.org/content/112/38/11870.abstract.html?etoc.

Falcón, A. a. & Aris, J.P., 2003. Plasmid Accumulation Reduces Life Span in Saccharomyces cerevisiae. *Journal of Biological Chemistry*, 278(43), pp.41607–41617.

Gao, H. et al., 2014. TALE activation of endogenous genes in Chlamydomonas reinhardtii. *Algal Research*, 5(1), pp.52–60. Available at: http://dx.doi.org/10.1016/j.algal.2014.05.003.

Gasiunas, G. et al., 2012. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 109(39), pp.E2579–86. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3465414&tool=pmcentrez&rendertype=abstract\nhttp://www.pnas.org/content/109/39/E2579?ijkey=dada4ad08fceeef0f38c0559436f614f04af9614&keytype2=tf_ipsecsha.

Gong, H. et al., 2017. Method for Dual Viral Vector Mediated CRISPR-Cas9 Gene Disruption in Primary Human Endothelial Cells. *Scientific Reports*, 7(February), p.42127. Available at: http://www.nature.com/articles/srep42127.

Hopes, A. et al., 2016. Editing of the urease gene by CRISPR-Cas in the diatom Thalassiosira pseudonana. *Plant Methods*, 12, pp.49–60. Available at: http://biorxiv.org/lookup/doi/10.1101/062026.

Ifuku, K. et al., 2015. A stable and efficient nuclear transformation system for the diatom Chaetoceros gracilis. *Photosynthesis Research*, 123(2), pp.203–211.

Jacobs, T.B. et al., 2015. Targeted genome modifications in soybean with CRISPR/Cas9. *BMC Biotechnol*, 15, p.16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25879861\nhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC4365529/pdf/12896_2015_Article_131.pdf.

Jiang, W. et al., 2013. Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice. *Nucleic Acids Research*, 41(20), pp.1–12.

Jiang, W. et al., 2014. Successful transient expression of Cas9 and single guide RNA genes in Chlamydomonas reinhardtii. *Eukaryotic Cell*, 13(11), pp.1465–1469.

Jinek, M. et al., 2012. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. , 337(August), pp.816–822.

Karas, B.J. et al., 2015. Designer diatom episomes delivered by bacterial conjugation. *Nature Communications*, 6, p.6925. Available at: http://www.nature.com/doifinder/10.1038/ncomms7925.

Kilian, O. et al., 2011. High-efficiency homologous recombination in the oil-producing alga Nannochloropsis sp. *Proceedings of the National Academy of Sciences*, 108(52), pp.21265–21269. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3248512&tool=pmcentrez&rendertype=abstract.

Kim, J. et al., 2015. Isolation and physiological characterization of a novel algicidal virus infecting the marine diatom Skeletonema costatum. *The Plant Pathology Journal*, 31(2), pp.186–191.

Kimura, K. & Tomaru, Y., 2015. Discovery of two novel viruses expands the diversity of single-stranded DNA and single-stranded RNA viruses infecting a cosmopolitan marine diatom. *Applied and Environmental Microbiology*, 81(3), pp.1120–1131.

Komor, A.C. et al., 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, 61(16), pp.5985–91. Available at: http://www.nature.com/doifinder/10.1038/nature17946\nhttp://www.ncbi.nlm.nih.gov/pubmed/11507039.

Krysiak, C., Mazus, B. & Buchowicz, J., 1999. Relaxation, linearization and fragmentation of supercoiled circular DNA by tungsten microprojectiles. *Transgenic Research*, 8(4), pp.303–306.

Lander, E.S., 2016. The Heroes of CRISPR. *Cell*, 164(1-2), pp.18–28.

Lauersen, K.J., Kruse, O. & Mussgnug, J.H., 2015. Targeted expression of nuclear transgenes in Chlamydomonas reinhardtii with a versatile, modular vector toolkit. *Applied Microbiology and Biotechnology*, 99(8), pp.3491–3503.

Maggio, I. et al., 2014. Adenoviral vector delivery of RNA-guided CRISPR/Cas9 nuclease complexes induces targeted mutagenesis in a diverse array of human cells. *Scientific reports*, 4, p.5105. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4037712&tool=pmcentrez&rendertype=abstract.

Mali, P. et al., 2013. RNA-Guided Human Genome Engineering via Cas9. *Science*, 339(February), pp.823–826.

Maruyama, T. et al., 2015. Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nature biotechnology*, 33(5), pp.538–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25798939\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4618510.

Miyagawa-Yamaguchi, A. et al., 2011. Stable nuclear transformation of the diatom Chaetoceros sp. *Phycological Research*, 59(2), pp.113–119.

Nagasaki, K. et al., 2004. Isolation and characterization of a novel single-stranded RNA virus infecting the bloom-forming diatom *Rhizosolenia setigera*. *Applied and environmental microbiology*, 70(2), pp.704–11. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=348932&tool=pmcentrez&rendertype=abstract.

Nekrasov, V. et al., 2013. Targeted mutagenesis in the model plant Nicotiana benthamiana using Cas9 RNA-guided endonuclease. *Nature Biotechnology*, 31(8), pp.691–693. Available at: http://dx.doi.org/10.1038/nbt.2655\npapers2://publication/doi/10.1038/nbt.2655.

Nymark, M. et al., 2016. A CRISPR/Cas9 system adapted for gene editing in marine algae. *Scientific reports*, 6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27108533.

Ordon, J. et al., 2016. Generation of chromosomal deletions in dicotyledonous plants employing a user-friendly genome editing toolkit. *The Plant journal : for cell and molecular biology*, pp.1–14. Available at: http://doi.wiley.com/10.1111/tpj.13319\nhttp://www.ncbi.nlm.nih.gov/pubmed/27579989.

Piatek, A. et al., 2015. RNA-guided transcriptional regulation in planta via synthetic dCas9-based transcription factors. *Plant Biotechnology Journal*, 13(4), pp.578–589.

Poulsen, N. et al., 2013. Pentalysine clusters mediate silica targeting of silaffins in Thalassiosira pseudonana. *Journal of Biological Chemistry*, 288(28), pp.20100–20109.

Qi, L.S. et al., 2013. Repurposing CRISPR as an RNA-Guided Platform for Sequence- Specific Control of Gene Expression. *Cell*, 152(5), pp.1173–1183.

Rasala, B. a. et al., 2014. Enhanced genetic tools for engineering multigene traits into green algae. *PLoS ONE*, 9(4).

Shin, S.-E. et al., 2016. CRISPR/Cas9-induced knockout and knock-in mutations in Chlamydomonas reinhardtii. *Scientific Reports*, 6, p.27810. Available at: http://www.nature.com/articles/srep27810.

Shirai, Y. et al., 2008. Isolation and characterization of a single-stranded RNA virus infecting the marine planktonic diatom Chaetoceros tenuissimus meunier. *Applied and Environmental Microbiology*, 74(13), pp.4022–4027.

Sizova, I. et al., 2013. Nuclear gene targeting in Chlamydomonas using engineered zinc-finger nucleases. *Plant Journal*, 73(5), pp.873–882.

Stearns, T., Ma, H. & Botstein, D., 1990. Manipulating Yeast Genome Using Plasmid Vectors. *Methods in Enzymology*, 185, pp.280–297.

Tomaru, Y. et al., 2012. First evidence for the existence of pennate diatom viruses. *The ISME Journal*, 6(7), pp.1445–1448. Available at: http://dx.doi.org/10.1038/ismej.2011.207.

Tomaru, Y. et al., 2011. Isolation and characterization of a single-stranded DNA virus infecting Chaetoceros lorenzianus Grunow. *Applied and Environmental Microbiology*, 77(15), pp.5285–5293.

Tomaru, Y. et al., 2009. Isolation and characterization of a single-stranded RNA virus infecting the bloom-forming diatom Chaetoceros socialis. *Applied and Environmental Microbiology*, 75(8), pp.2375–2381.

Tomaru, Y. et al., 2013. New single-stranded DNA virus with a unique genomic structure that infects marine diatom Chaetoceros setoensis. *Scientific reports*, 3, p.3337. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3840382&tool=pmcentrez&rendertype=abstract.

Vojta, A. et al., 2016. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Research*, 44(12), pp.5615–5628.

Wang, Q. et al., 2016. Genome editing of model oleaginous microalgae Nannochloropsis spp. by CRISPR/Cas9. *Plant Journal*, 88, pp.1071–1081.

Weyman, P.D. et al., 2015. Inactivation of Phaeodactylum tricornutum urease gene using transcription activator-like effector nuclease-based targeted mutagenesis. *Plant Biotechnology Journal*, 13(4), pp.460–470.

Yin, K. et al., 2015. A geminivirus-based guide RNA delivery system for CRISPR/Cas9 mediated plant genome editing. *Scientific Reports*, 5, p.14926. Available at: http://www.nature.com/articles/srep14926.

Zaslavskaia, L. a et al., 2000. Transformation Of the Diatom Phaeodactylum Tricornutum With a Variety of Selectable Marker and Reporter Genes. *Journal of Phycology*, 386(August 1999), pp.379–386.

Zheng, Q. et al., 2014. Precise gene deletion and replacement using the CRISPR/Cas9 system in human cells. *BioTechniques*, 57(3), pp.115–124.

Zhou, H. et al., 2014. Large chromosomal deletions and heritable small genetic changes induced by CRISPR/Cas9 in rice. *Nucleic Acids Research*, 42(17), pp.10903–10914.

Hopes *et al. Plant Methods (2016) 12:49*
DOI 10.1186/s13007-016-0148-0

Plant Methods

METHODOLOGY

Open Access

Editing of the urease gene by CRISPR-Cas in the diatom *Thalassiosira pseudonana*

Amanda Hopes[1], Vladimir Nekrasov[2], Sophien Kamoun[2] and Thomas Mock[1*]

Abstract

Background: CRISPR-Cas is a recent and powerful addition to the molecular toolbox which allows programmable genome editing. It has been used to modify genes in a wide variety of organisms, but only two alga to date. Here we present a methodology to edit the genome of *Thalassiosira pseudonana*, a model centric diatom with both ecological significance and high biotechnological potential, using CRISPR-Cas.

Results: A single construct was assembled using Golden Gate cloning. Two sgRNAs were used to introduce a precise 37 nt deletion early in the coding region of the urease gene. A high percentage of bi-allelic mutations (≤61.5%) were observed in clones with the CRISPR-Cas construct. Growth of bi-allelic mutants in urea led to a significant reduction in growth rate and cell size compared to growth in nitrate.

Conclusions: CRISPR-Cas can precisely and efficiently edit the genome of *T. pseudonana*. The use of Golden Gate cloning to assemble CRISPR-Cas constructs gives additional flexibility to the CRISPR-Cas method and facilitates modifications to target alternative genes or species.

Keywords: CRISPR-Cas, Diatom, Genome editing, Urease, Golden Gate, *Thalassiosira pseudonana*

Background

Diatoms are ecologically important microalgae with high biotechnological potential. Since their appearance about 240 million years ago [1], they have spread and diversified to occupy a wide range of niches across both marine and freshwater habitats. Diatom genomes have been shaped by secondary endosymbiosis and horizontal gene transfer resulting in genes derived from heterotrophic hosts, autotrophic endosymbionts and bacteria [2, 3]. They play a key role in carbon cycling [4], the food chain, oil deposition and account for about 20% of the world's annual primary production [5, 6]. However, they are perhaps best known for their intricate silica frustules which give diatoms a range of ecological advantages and play a key role for carbon sequestration and silica deposition.

Several aspects of diatom physiology including the silica frustule, lipid storage and photosynthesis are being applied to biotechnology. Areas of high interest include nanotechnology [7], drug delivery [8], biofuels [9], solar capture [10] and bioactive compounds [11].

Given the ecological importance of diatoms and their applications for biotechnology, it is pivotal that the necessary tools are available to study and manipulate them at a molecular level. This includes the ability to replace, tag, edit and impair genes. A recent addition to the genetic tool box, CRISPR-Cas, allows double strand breaks (DSBs) to be introduced at specific target sequences. This adapted mechanism, used by bacteria and archaea in nature as a defence system against viruses, facilitates knock-out by the introduction of mutations through repair by error prone non-homologous end joining (NHEJ) or homologous recombination (HR). This requires both a Cas9 to cut the DNA and a sgRNA to guide it to a specific sequence. Further information on the history and application of CRISPR-Cas can be found

*Correspondence: t.mock@uea.ac.uk
[1] School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK
Full list of author information is available at the end of the article

BioMed Central

© The Author(s) 2016. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

in several excellent reviews [12–14]. Zinc-finger nucleases (ZFNs), meganucleases and transcription activator-like effector nucleases (TALENs) have also been used to induce double strand breaks. TALENs and CRISPR-Cas both bring flexibility and specificity to gene editing, however CRISPR-Cas is also cheap, efficient and easily adapted to different sequences by simply changing the 20 nt guide sequence in the sgRNA.

So far, within the diverse group of algae, the diploid, pennate diatom *Phaeodactylum tricornutum* [15] and the haploid, green alga *Chlamydomonas reinhardtii* [16] have been subject to gene editing by CRISPR-Cas. NHEJ and HR have been used to repair DSBs following CRISPR-Cas or TALENs in *P. tricornutum*, introducing mutations into a nuclear coded chloroplast signal recognition particle [15], the urease gene [17] and several genes associated with lipid metabolism [18]. *Thalassiosira pseudonana* is a logical choice for CRISPR-Cas development. It is a model centric diatom with a sequenced genome (first eukaryotic marine phytoplankton to be sequenced [2]) and well-established transformation systems [19, 20]. The genus has multiple biotechnology applications [8, 21, 22], and although gene silencing has been established, a method to easily and efficiently knock-out and edit genes and the entire genome would be highly advantageous. The genus *Thalassiosira* is among the top 10 genera of diatoms in the World's Ocean in terms of ribotype (V9 of 18S) diversity and abundance [23] and the species *T. pseudonana* is a model for understanding the mechanisms behind silicification [24–26].

Golden Gate cloning can add further flexibility to CRISPR-Cas methods as demonstrated in higher plants [27]. As a modular cloning system it allows different modules, including the sgRNA, to be easily interchanged or added [28]. As a result, new constructs can be made quickly, cheaply and efficiently for new or multiple targets. This extends to any aspect of the construct, including promoters, Cas9 variants and their nuclear localisation signals (NLS). As a result, construct alterations such as replacing constitutive promoters for inducible ones, exchanging the wildtype Cas9 for a Cas9 nickase or changing the localisation signal to target other organelles can be easily carried out.

An increasing range of software tools are available for CRISPR-Cas, including programs that facilitate sgRNA target searches in a genetic locus of interest, estimate efficiencies of sgRNAs [29] and perform off-target predictions.

While off-target prediction tools tend to be species specific, there are tools that accept requests for a genome to be added to the list, or allow for a genome to be directly uploaded [30, 31]. The latter is particularly useful for less studied organisms, such as diatoms. The ability to combine several different aspects of sgRNA design can help to make an informed decision when choosing target sites for gene editing.

Our paper represents a proof of concept to demonstrate the feasibility of gene editing in the model diatom *T. pseudonana* using two sgRNAs to induce a precise deletion in the urease gene. Methods combine a flexible Golden Gate cloning approach with sgRNA design, which draws on several available online tools. This takes into account multiple factors, such as position within the gene in terms of both early protein disruption and presence in the coding region, DNA cutting efficiency and presence of restriction enzyme sites at the cut site. The latter, in combination with inducing a large deletion by targeting with two sgRNAs, allows easy screening of mutants through either the restriction enzyme site loss assay [32] or the PCR band-shift assay [33], respectively.

Methods

Strains and growth conditions

Thalassiosira pseudonana (CCMP 1335) was grown in 24 h light (100–140 µE) at 20 °C in half salinity Aquil synthetic seawater [34]. For routine growth, a 1 mM nitrate concentration was used.

5′RACE U6 promoter

To identify the U6 small nuclear RNA (snRNA) in *T. pseudonana*, an NCBI blastn search was performed on the genome against the central conserved region of the U6 sequence. Two potential guanine (G) start sites were found downstream of a TATA box in the promoter. To identify the start site of the U6 snRNA and empirically determine the end of the promoter, 5′ RACE was carried out as follows: 400 ml of culture was grown to exponential phase (1×10^6 cells ml^{-1}) and harvested. Small RNAs were extracted and enriched using a miRNeasy kit (Qiagen). 5′ template switching oligo RACE was performed according to Pinto and Lindblad [35]. For oligos used see Table 1 (Ref. Numbers 1–3). RACE products were sequenced and results aligned to the genome to determine the end of the promoter.

Plasmid construction using Golden Gate cloning

Golden Gate cloning was carried out according to Weber et al. [28] and Belhaj et al. [33]. BsaI and BpiI sites were removed in a so-called "domestication" procedure using a Q5 site-directed mutagenesis (SDM) kit (NEB). For oligos used in SDM see Table 1 (Ref. Numbers 17–20). BsaI sites and specific 4 nt overhangs for Level 1 (L1) assembly were added through PCR primers (Table 1). Plasmid DNA was extracted using a Promega mini-prep kit.

Table 1 Oligonucleotides used in this study

Name	Sequence	Ref. No.
GS U6 R	AGGTTTGCTTCTCTTCGATTATG	1
TSO	GTCGCACGGTCCATCGCAGCAGTCACAGGGGG	2
U sense	GTCGCACGGTCCATCGCAGCAGTC	3
Fcp:Nat F	tggtctc*aggag*CTCGAGGTCGACGGTATC	4
Fcp:Nat R	aggtctc*aagcg*CGCAATTAACCCTCACTAAAGG	5
FCP prom F	tggtctc*aggag*AGCTTGCGCTTTTTCCGAG	6
FCP prom R	aggtctc*acat*TTGGTATTGGTTTGGTAAATCAG	7
Cas9:YFP F	aggtctc*aATG*GACAAGAAGTACTCCATTGG	8
Cas9:YFP R	aggtctc*aaagc*TCACTTGTACAGCTCGTCCATG	9
FCP term F	aggtctc*agctt*ATACTGGATTGGTGAATCAATG	10
FCP term R	tggtctc*aagcg*GAGAACTGGAGCAGCTAC	11
U6 prom F	cggtctc*aggag*CTTCATCAAGAGAGCAACCA	12
U6 prom R	aggtctcaACAATTTCGGCAAAACGT	13
Urease sgRNA1 F	aggtctc*attgt***gtcgtaatcaagtattgccg**GTTTTAGAGCTAGAAATAGCAAG	14
Urease sgRNA2 F	aggtctc*attgt***gtttccgatctaatgtccat**GTTTTAGAGCTAGAAATAGCAAG	15
Urease sgRNA R	tggtctc*aagcg*TAATGCCAACTTTGTACAAG	16
FCP prom SDM F	TCCGCGGCAGaTCTCTGTCG	17
FCP prim SDM R	AGAAGTACCGTGTTGTTGCAGTG	18
NAT SDM F	CGACACCGTaTTCCGCGTCAC	19
NAT SDM R	GTGGTGAAGGACCCATCCAG	20
Cas9 screen F	CCGAGACAAGCAGAGTGGAAAG	21
Cas9 screen R	AGAGCCGATTGATGTCCAGTTC	22
NAT screen F	ATGACCACTCTTGACGACAC	23
NAT screen R	TTGATTCACCAATCCAGTATGC	24
Urease screen F 1	AAACAGACCACCTTCACCTC	25
Urease screen R	CTCCACCTGTACGTCTCG	26
Fcp seq F	CCATAAGTCAACGGCTCCAATC	27
NAT seq F	CTCTTGACGACACGGCTTAC	28
Cas9 seq 1 F	CATTACGGACGAGTACAAGGTG	29
Cas9 seq 2 F	TGAACACGGAGATCACCAAAG	30
Cas9 seq 3 F	CTTCCTGGACAATGAGGAGAAC	31
Cas9 seq 4 F	CAAACTGATCACACAACGGAAG	32
YFP FcpT seq F	ACTACCTGAGCTACCAGTCC	33
sg2 seq R	GTTTCCGATCTAATGTCCAT	34
sg1 seq F	TGTGTCGTAATCAAGTATTGC	35

Ref. No. 1–3: oligos used in 5′ RACE [35]. Ref. No. 4–16: primers for Golden Gate cloning, BsaI sites are underlined, 4 nt overhangs are shown in italics, and sgRNA targets are shown in bold. Upper case indicates complement to the template. Ref. No. 17–20: primers for SDM, lower case indicates base change. Ref. No. 21–26: primers for screening transformants. Ref. No. 27–35: primers for sequencing the CRISPR-Cas construct

Golden Gate reactions

Golden Gate reactions for L1 and Level 2 (L2) assembly were carried out using the method specified in Weber et al. [28]. Forty fmol of each component was included in a 20 µl reaction with 10 units of BsaI or BpiI and 10 units of T4 DNA ligase in 1 × ligation buffer. The reaction was incubated at 37 °C for 5 h, 50 °C for 5 min and 80 °C for 10 min. Five µl of the reaction was transformed into 50 µl of NEB 5-alpha chemically competent *E. coli*.

Level 0 assembly

The endogenous FCP promoter and terminator were amplified with GoTaq flexi (Promega) from domesticated pTpFCP/NAT [19] and the U6 promoter from gDNA [extracted with an Easy-DNA gDNA purification kit (Thermo Fisher)]. Both promoters are associated with high expression levels. The U6 promoter was amplified from the position −470 to −1 (the end of the promoter), cutting off a BpiI site and removing the need

for additional SDM. For oligos, see Table 1 (Ref. Numbers 6–7 and 10–13). Products were cloned into a pCR8/GW/TOPO vector (Thermo Fisher).

Domesticated human codon bias Cas9 from *Streptococcus pyogenes* with an N-terminal SV40 NLS and a C-terminal YFP tag was PCR-amplified using Phusion DNA polymerase (NEB) and L1 Cas9:YFP plasmid as a template. The PCR product was purified with a GFX PCR DNA and gel purification kit (GE Healthcare) and incubated for 20 min with Taq to add adenine overhangs before cloning directly into a pCR8/GW/TOPO vector. For oligos, see Table 1 (Ref. Numbers 8–9).

Level 1 assembly

The FCP:NAT cassette for nourseothricin resistance was PCR-amplified using Phusion polymerase and the domesticated pTpFCP/NAT as a template, purified and inserted into a L1 pICH47732 destination vector. FCP promoter, Cas9 and FCP terminator L0 modules were assembled into L1 pICH47742. For oligos, see Table 1 (Ref. Numbers 4–5).

The sgRNA scaffold was amplified from pICH86966_AtU6p_sgRNA_NbPDS [32] with sgRNA guide sequences integrated through the forward primers. Together with the L0 U6 promoter, sgRNA_Urease 1 and sgRNA_Urease 2 were assembled into L1 destination vectors pICH47751 and pICH47761, respectively. For oligos, see Table 1 (Ref. Numbers 14–16).

Level 2 assembly

L1 modules pICH47732:FCP:NAT, pICH47742:FCP:Cas9YFP, pICH47751:U6:sgRNA_Urease 1, pICH47761:U6:sgRNA_Urease 2 and the L4E linker pICH41780 were assembled into the L2 destination vector pAGM4723. Constructs were screened by digestion with EcoRV and sequenced. For oligos used in sequencing, see Table 1 (Ref. Numbers 27–35). See Fig. 1 for an overview of the Golden Gate assembly procedure and the final construct.

sgRNA design for the urease gene knockout

Two sgRNAs were designed to cut 37 nt apart early in the coding region of the urease gene (JGI ID 30193) to induce a deletion and frame-shift. Several programmes, explained below, were used to collect data and make an informed decision on sgRNA choice. Excel was used to combine, process and compare data.

Selecting CRISPR-Cas targets and estimating on-target score

Twenty bp targets with an NGG PAM were identified and scored for on-target efficiency using the Broad Institute sgRNA design programme (www.broadinstitute.org/rnai/public/analysis-tools/sgrna-design), which utilises the Doench et al. [29] on-target scoring algorithm calculated from >1800 empirically tested sgRNAs.

Determining cut positions and cross referencing to restriction recognition sites

All restriction sites and their positions within the urease gene were identified using the Emboss restriction tool (http://emboss.bioinformatics.nl/). As the Broad Institute sgRNA design programme does not give the location of CRISPR-Cas targets within a gene, this was determined using Primer map (http://www.bioinformatics.org/sms2/primer_map.html [36]). The cut site position (3 nt upstream of the start of the PAM sequence) was calculated for each sgRNA depending on sense or anti-sense strand placement. All predicted CRISPR-Cas cut sites were cross-referenced to restriction recognition sites.

Reverse complement of antisense strand CRISPR-Cas targets

The reverse complement (RC) was found for each CRISPR-Cas target using the programme: http://www.bioinformatics.org/sms2/rev_comp.html [36]. In the final spreadsheet (Additional file 1: Figure S1), if a target was located on the anti-sense strand, the RC was shown for the 'sense strand sequence' column. This allows the sgRNA to be easily searched within the original gene sequence.

Determine position of CRISPR-Cas cut sites in relation to coding region

An array was made with start and end positions for each exon/intron. Cut site positions were compared to exon/intron ranges and the relevant exon/intron returned if the data overlapped.

The final spreadsheet gives data on CRISPR-Cas target sequences and their sense sequence (if located on the antisense strand), location of target (relative to the sense strand), predicted CRISPR-Cas cut site, first nucleotide of the target, PAM sequence, location (i.e. exon, intron), strand, sgRNA score and restriction recognition sites overlapping the cut site. The table (Additional file 1: Figure S1) was sorted to prioritise sgRNAs by starting base prioritising guanine, sgRNA score, position within the gene and interaction with restriction recognition sites.

Predicting off-targets

The full 20 nt target sequences and their 3′ 12 nt seed sequences were subjected to a nucleotide BLAST search against the *T. pseudonana* genome. Resulting homologous sequences were checked for presence of an adjacent NGG PAM sequence at the 3′ end. The 8 nt sequence outside of the seed sequence was manually checked for complementarity to the target sequence. In order for a site to be considered a potential off-target the seed

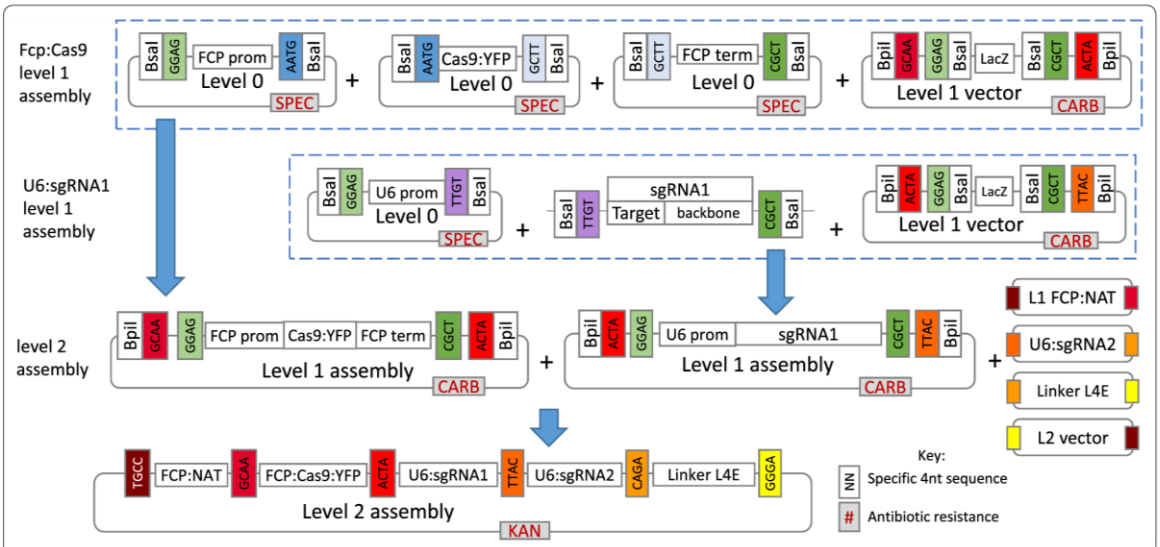


Fig. 1 Overview of level 1 (L1) and level 2 (L2) Golden Gate cloning for assembly of the CRISPR-Cas construct pAGM4723:TpCC_Urease. L0 modules are created by PCR amplifying the relevant insert, BsaI sites are added through the primers. Products are then TOPO TA cloned into pCR8/GW/TOPO vectors. Construction of the L1 and L2 modules is based on the ability of BsaI and BpiI restriction enzymes to cut outside of their restriction recognition sites leaving 4 nt overhangs specific to each module. Overhangs between adjacent modules correspond allowing multiple modules to be ligated in a particular order within the same reaction. BsaI is used to assemble L0 modules into L1 destination vectors and BpiI is used to assemble L1 modules into the final L2 destination vector. L1 assemblies of pICH47742:FCP:Cas9YFP and pICH47751:U6:sgRNA_Urease 1 are shown. L2 assembly of the final construct is shown. L1 modules containing the FCP:NAT cassette, U6:sgRNA Urease 2 cassette, the L4E linker and L2 destination vector are shown in a simplified format. Corresponding *colours* denote a shared 4 nt overhang

sequence had to match, a PAM had to be present at the 3′ end of the sequence and a maximum of three mismatches between the target and sequences from the blast search were allowed outside of the seed sequence.

Off-targets were also checked using the EuPaGDT program [31], which checks for up to 5 mismatches in the 20 nt target sequence and the CasOT program [30], which uses flexible parameters for identifying off-target sequences. Parameters were set to check for an NGG PAM, complete complementarity within the 12 nt seed sequence and up to 3 mismatches outside of the seed region.

Transformation and selection

Using the Poulsen et al. [19] method, transformations were carried out in triplicate with the CRISPR-Cas construct, pTpfcp/NAT (positive control) and water (negative control). 5×10^7 cells in exponential phase were used per shot with a rupture disc of 1350 psi and a 7 cm flight distance. Following transformation, cells were rinsed into 25 ml of media and left to recover for 24 h under standard growth conditions. Cells were counted using a Coulter counter (Beckman) and 2.5×10^7 cells from each transformation were spread onto 5, ½ salinity Aquil 0.8% agar plates (5×10^6 cells/plate) with 100 µg ml^{-1} nourseothricin. Plates were incubated under standard conditions for two weeks. The remaining sample was diluted to 1×10^6 cell ml^{-1} in media and supplemented with nourseothricin to a final concentration of 100 µg ml^{-1} for liquid selection. Liquid selection cultures were maintained under standard growth conditions with 100 µg ml^{-1} nourseothricin. Colonies were picked and transferred to 20 µl of media. Ten µl from each colony was transferred to 1 ml of selective media for further growth. The remaining sample was used in screening.

To isolate sub-clones from colonies which screened positive for mutations, 100 µl of cells at exponential phase were streaked onto ½ salinity Aquil 0.8% agar plates with 100 µg ml^{-1} nourseothricin.

Screening clones and cultures

Ten µl from each colony or culture from liquid selection, was spun down and supernatant removed. Cells were re-suspended in 20 µl of lysis buffer (10% Triton X-100, 20 mM Tris–HCl pH 8, 10 mM EDTA), kept on ice for 15 min then incubated at 95 °C for 10 min. One µl of lysate was used in Taq PCR to amplify the CRISPR-Cas targeted fragment of the urease gene. Clones were also screened for Cas9 and NAT by PCR. For PCR primers, see Table 1 (Ref. Numbers 21–26). PCR products were

run on an agarose gel to check for the lower MW band associated with a double-cut deletion in the urease gene and for the presence of Cas9 and NAT. Urease PCR products were also digested with BccI and HpaII to determine if the restriction recognition sites, which overlap the cut sites, had been mutated. Urease PCR products from all screened primary clones were sent for sequencing to look for mutations. PCR products from a selection of subclones derived from three primary clones were sent for sequencing to confirm mutations.

Growth experiments

Knockout and wild-type (WT) cultures were nitrate depleted by growing cells in nitrate free media until cell division stopped and Fv/Fm (quantum yield of photosynthesis, used as a proxy for cell stress) measured on the Phyto-PAM-ED dropped below 0.2. Cultures were then transferred in triplicate at a final concentration of 2.5×10^4 cells ml^{-1} into 25 ml of media with either 1 mM sodium nitrate or 0.5 mM urea. A media control was also carried out with WT cultures by transferring cells to fresh nitrate free media to check for any residual nitrate or nitrogen compounds that could lead to cell growth. Cell count and mean cell size were measured once a day using a Coulter counter. Fv/Fm measurements were also taken daily. Growth rates were calculated using $\mu = Ln_2 - Ln_1/T_2 - T_1$, where T is a time point corresponding to exponential growth and Ln is the natural log of cell counts ml^{-1}. Analysis of variance with Tukey's pairwise comparision was used to compare both growth rates and cell size at the end of exponential phase between samples.

Results and discussion

sgRNA design

The two CRISPR-Cas targets with the highest on target scores (0.5 and 0.79), containing a predicted cut site over a restriction site and occurring early in the coding region, were chosen. sgRNAs were designed to cut 37 nt apart at positions 138 and 175 within the urease gene. Both targets started with a G for polymerase III transcription (Fig. 2). No off-target sites were predicted for sgRNAs designed for either of the two CRISPR-Cas target sequences.

Constructing the CRISPR-Cas plasmid using the Golden Gate cloning method

A single CRISPR-Cas construct was made using Golden Gate cloning (Fig. 1). The construct included the NAT selectable marker gene for nourseothricin resistance, Cas9:YFP driven by an endogenous FCP promoter for high expression and two U6 promoter-driven sgRNAs. RNA polymerase III U6 promoters are a popular choice for expression of sgRNAs in CRISPR-Cas [15, 27, 37–39]. RACE products showed that the U6 promoter ended 23 nt after the TATA box. As a standardised, efficient, modular system, Golden Gate cloning gives a high level of flexibility to the CRISPR-Cas method and bypasses the need for co-transformation as it enables assembly of multiple expression units, such as Cas9 and sgRNAs, into a single vector backbone. Multiple sgRNA modules can be incorporated into the construct to target several genes or whole pathways. In human cells, up to 7 sgRNAs have been successfully assembled and expressed from a single construct created using the Golden Gate cloning method [39]. Golden Gate has also proved successful for building constructs for genome editing in higher plants using both TALENs [40] and CRISPR-Cas [33, 37].

In this study, only the promoters and target sequences are specific to *T. pseudonana,* which demonstrates how simple it can be to apply this method to a new species using the Golden Gate system. The *S. pyrogenes* Cas9 with a human codon bias, shown previously to work in higher plants [32, 33, 37], carries a SV40 NLS, which follows a canonical sequence found throughout eukaryotes, including *T. pseudonana.*

The long term effects from off-target mutations introduced through CRISPR-Cas are currently unknown, therefore it may be advantageous for future work to remove CRISPR-Cas constructs from mutants. Adding a yeast CEN6-ARSH4-HIS3 sequence to plasmids allows autonomous replication in diatoms and expression of genes without random integration into the genome [20]. Furthermore, removing selection leads to plasmids being discarded. By expressing CRISPR-Cas genes and selective markers on a removable episome, mutations could be introduced without integration of the plasmid. CRISPR-Cas constructs could then be expelled by removing selection. As well as considerations for long term off-target effects, this could also be advantageous for studies and applications which are sensitive to the presence of transgenes.

Selecting and screening for mutations in the urease gene

The transformation efficiency with the CRISPR-Cas construct was on average 41.5 colonies μg^{-1} plasmid (13.35–66.65 colonies μg^{-1} plasmid). Thirty-three colonies were screened by PCR and sequencing of the targeted urease gene fragment.

Four colonies showed mutations in the urease gene. All colonies screened positive for NAT but only the four colonies with mutations screened positive for Cas9, suggesting that once the Cas9 and sgRNAs are present there is a high chance of inducing mutations in the target gene. The lack of Cas9, which accounts for a third of the construct, in the majority of colonies was potentially caused

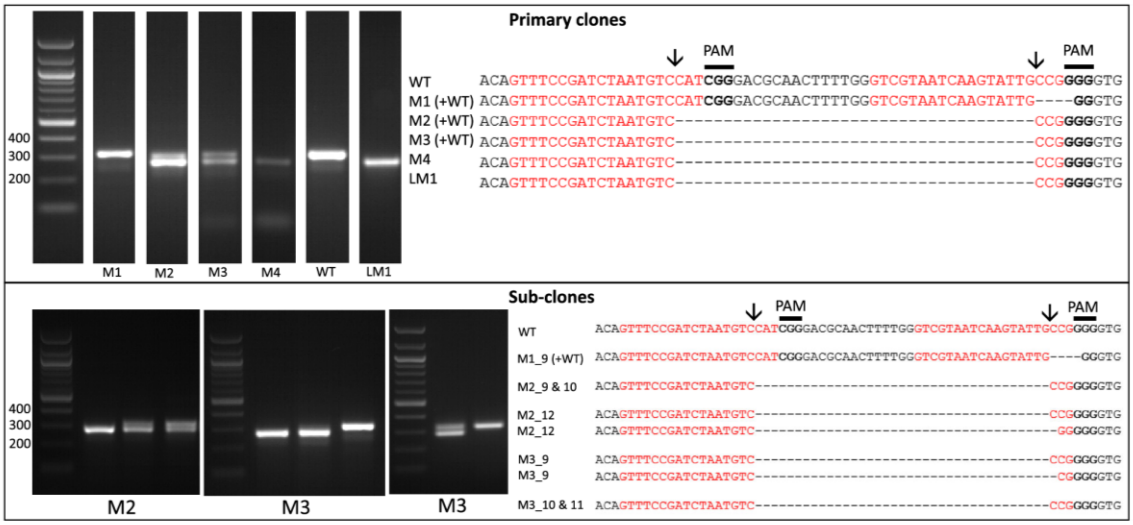


Fig. 2 Screening by PCR and sequencing. Expected sgRNA cut indicated by ↓. *Red text* shows the sgRNA target sequence and *bold text* the PAM motif. *Primary clones:* Several primary clones contain sequences showing both CRISPR-induced mutations and the wildtype (WT) sequence, as seen by the presence of two bands following PCR, these are indicated by (+WT). Mutants M1–M4 were selected on plates and LM1 underwent selection in liquid media. M1 shows a 4 nt deletion from the second sgRNA whilst mutants M2–M4 show a 37 nt deletion between the two CRISPR-Cas cut sites. *Sub-clones:* The gel shows examples of a selection of sub-clones derived from the primary clones. Sub-clones are labelled according to the primary clone and sub-clone number. With the exception of M1_9, which gives a WT sequence and 4 nt deletion as seen in the primary clone, all sub-clones chosen for sequencing are bi-allelic. Two-thirds of sequenced bi-allelic sub-clones show a single sequence with a 37 nt deletion suggesting that both alleles carry the same mutation. In sub-clones where mutations differ between alleles both sequences are shown

by shearing of the plasmid during microparticle bombardment [38] from either mechanical force or chemical breakdown [41].

Of the four primary colonies which screened positive for mutations (Fig. 2), one (M4) showed a single band with a 37 nt deletion between the two sgRNA cut sites which suggests that both copies of the urease gene contain the deletion giving a bi-allelic mutant. Two colonies (M2 and M3) produced two bands following PCR: a WT higher MW band and a lower MW band with the 37 nt deletion, confirmed by sequencing (Fig. 2). The fourth colony (M1) showed a single band associated with the WT urease, however sequencing showed two products: a WT urease and a mutant urease with a 4 nt deletion at the first sgRNA cut site. A mixture of PCR products may be due to a mono-allelic mutation, in which one allele is WT and the other displays a mutation. It can also be due to colony mosaicism where a colony contains a mixture of cells with WT and mutant alleles due to mutations occurring after transformed cells have started to divide. Both mono-allelic mutants and mosaic colonies have been observed in *P. tricornutum* [15, 18].

To determine if the colonies were mosaic or mono-allelic, cells from mutant clones producing mixed PCR products were spread onto selective plates to isolate single sub-clones. Thirty-four sub-clones from each clone were screened by PCR (a few examples are presented in Fig. 2). Two clones (M2 and M3) were mosaic with a mixture of sub-clones showing either a single band corresponding to the expected deletion (61.5 and 25%, respectively), two bands associated with the WT and expected deletion (25.5 and 28.1%, respectively) or a single band corresponding to the WT urease fragment (13 and 46.9% respectively). For each of the two clones PCR amplicons from three putative bi-allelic sub-clones were sequenced (Fig. 2). Four out of six (M2_9, M2_10, M3_10 and M3_11) showed the expected 37 nt 'clean' deletion without any additional mutations. Precise deletions, such as this, using 2 sgRNAs have previously been generated with high efficiency [37, 42], and allow a large degree of control over the mutation. Two of the sub-clones (M3_9 and M2_12) showed one allele with the expected 37 nt deletion and the other with an additional deletion at the 2nd sgRNA cut site. In addition, M2_12 showed a C → G SNP within the sgRNA1 target site. As sequencing of primary clones M2 and M3 showed the expected deletion without additional indels, this suggests that this was the more dominant product following PCR. As well as isolating bi-allelic mono-clonal cultures, sub-cloning can reveal less common mutations produced by CRISPR-Cas.

Sub-clones derived from the M1 clone showed WT and 4 nt deletion PCR amplicons as seen in the original clone, suggesting that this clone may have a mono-allelic mutation.

Using CRISPR-Cas with one sgRNA can introduce a variety of indels into a locus of interest via the error-prone NHEJ DNA repair mechanism [15]. Cas9 preferentially cuts DNA three nucleotides upstream of the PAM sequence in the seed region [43] and the NHEJ mechanism either repairs a double strand break perfectly or indels are introduced. If cut sites are not cleaved at the same time, when using two sgRNAs, mutations at each site rather than removal of the fragment in between target sites may occur [37]. In this study, however, we report a high occurrence of bi-allelic mutants with precise deletions between the CRISPR-Cas cut sites, suggesting that the Cas9/sgRNA complex is cutting efficiently and DNA ends tend to be repaired perfectly. This allows control over the introduced mutations and gives the chance to avoid introducing in-frame indels.

Restriction digest (results not shown) and sequencing (Fig. 2) demonstrated loss off the BccI site in all knock-out clones and HpaII in M2_12 and M1 as a deletion downstream of the cut site is required to remove the HpaII site. This demonstrates that restriction screening can be a valuable tool, however is this case screening for a deletion based band shift by PCR was an efficient way of identifying bi-allelic mutants especially given the limited sgRNA/restriction site interactions available for this gene.

As well as clones from plate selection, one culture from liquid selection (LM1; population of cells transferred to liquid selective media after transformation), showed a single band associated with the bi-allelic 37 nt deletion following PCR. This was confirmed by sequencing (Fig. 2). PCR screening following growth of LM1 in urea showed only the lower MW band product (results not shown), giving further evidence for a bi-allelic mutation from a population of cells. As small volumes of cells are transferred to fresh media when passaging this may have isolated bi-allelic mutants.

Growth experiments with mutants

Urease catalyses the breakdown of urea to ammonia allowing it to be used as a source of nitrogen [44]. Sub-clones from different cell-lines with 37 or 38 nt deletions were tested for knock-out of the urease gene by looking for a lack of growth when supplemented with urea as the sole nitrogen source.

Cells were nitrogen starved and then transferred to media with either nitrate or urea. Cell counts and cells size were measured daily for 7 days. Negative controls to account for any background nitrate in the media were also run in which no nitrate or urea was added for WT cultures.

Four putative bi-allelic mutants (LM1, M4, M2_10 and M3_9) were tested along with WT and the mono-allelic M1_10 over two growth curve experiments. Both LM1 from liquid selection ($p = 0.0029$) and the sub-clone M3_9 ($p = 0.0000001$) showed a significant decrease in growth rate in urea compared to nitrate (Fig. 3) as well as a significant 13–18% decrease in cell size (Fig. 4; $p = 0.0029$ and $p = 0$, respectively). The latter was also apparent with light microscopy (results not shown). Mutants in urea could be easily discerned even without cell counts, as cultures appeared much paler in colour. M4 did not show a difference in growth rate but did show a significant decrease in cell size ($p = 0.038$).The mono-allelic mutant M1_10, displayed higher growth in urea and similar growth to the WT control (Fig. 3). This correlates with results from Weyman et al. [17] which showed that despite a reduced protein concentration, a mono-allelic urease knock-out was able to grow in urea. M2_10 which screened as a bi-allelic mutant prior to growth experiments showed a smaller but still significant decrease in growth rate ($p = 0.0014$; Fig. 3) and cell size ($p = 0.0039$; Fig. 4). PCR screening of the urease gene following growth in nitrate and urea showed the expected bi-allelic mutation for LM1, M3_9 and M4, however M2_10 also showed a faint WT band in nitrate and a strong WT band in urea (Fig. 5). This suggests that M2_10 was mosaic, with cells containing a functional urease out-competing those with a mutant urease. Given that only a faint WT band was present after growth in nitrate this suggests that the majority of the cells from the sub clone contained the mutant urease, initially accounting for the majority of growth and resulting in a lower but still significant decrease in growth rate.

Knock-out of the urease gene in the diatom *P. tricornutum* prevents growth in urea [17]. Urease mutants in this study still grew in urea but with a lower growth rate and reduced cell-size, characteristics which are associated with nitrogen limitation in diatoms [45, 46] rather than nitrogen starvation. Mutant cell-lines in urea grew to the same density as the same cell-lines in nitrate, but at a lower rate (Fig. 3). As nitrogen is an essential nutrient for growth, this suggests that mutant cells in urea still have access to nitrogen, but lower growth rates and cell-size indicates that nitrogen may not be as readily available compared to cells grown with nitrate. Controls in nitrogen free media showed very little growth which suggests that growth of mutants in urea was not due to background nitrate in the culture media. It is unlikely that random integration of the CRISPR-Cas plasmid is responsible for reduced growth rate in mutants as all four individual mutant cell-lines display increased growth

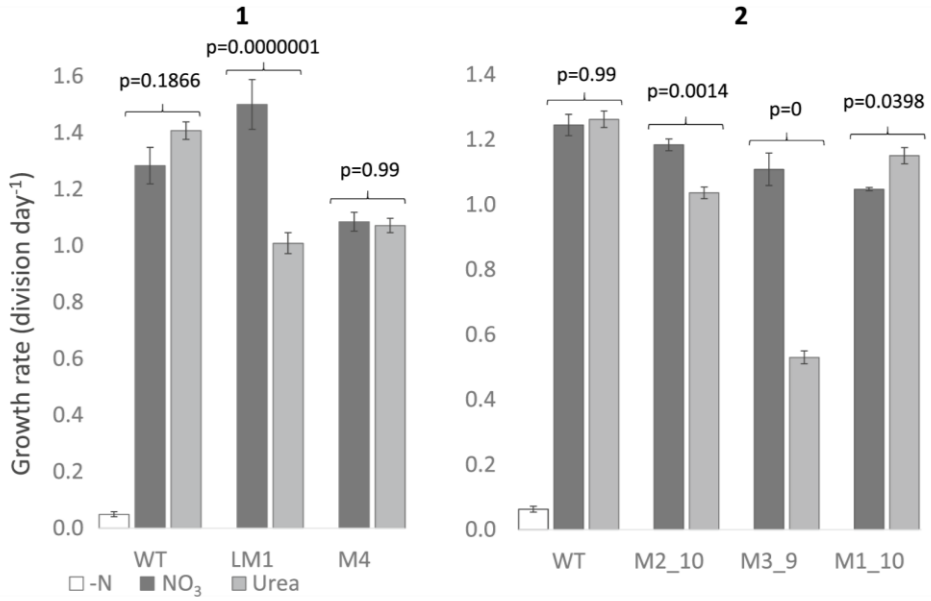


Fig. 3 Growth rate of WT and mutant urease cell lines from two separate growth experiments (**1**, **2**). The WT cell line was grown in nitrate free (*white*), nitrate (*dark grey*) and urea (*light grey*) enriched media. Mutant cell lines were grown in nitrate or urea enriched media. Growth rate (division day^{-1}) was measured in exponential phase and rates compared using analysis of variance with Tukey's pairwise comparisons

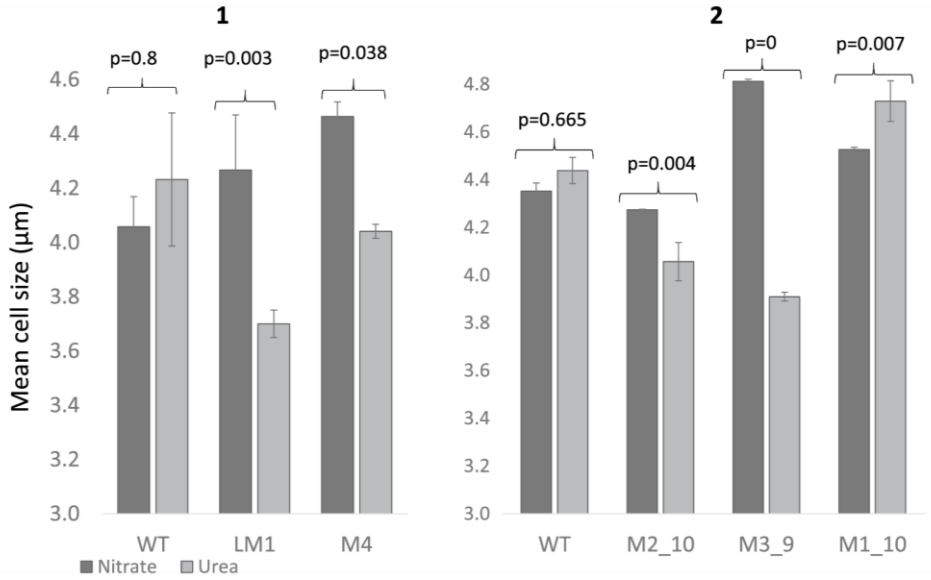


Fig. 4 Mean cell size (µm) measured at the end of exponential phase for WT and mutant cultures across two growth experiments (**1**, **2**). Cells were grown with nitrate (*dark grey*) or urea (*light grey*) as the sole nitrogen source. Cell size was compared using analysis of variance with Tukey's pairwise comparisons

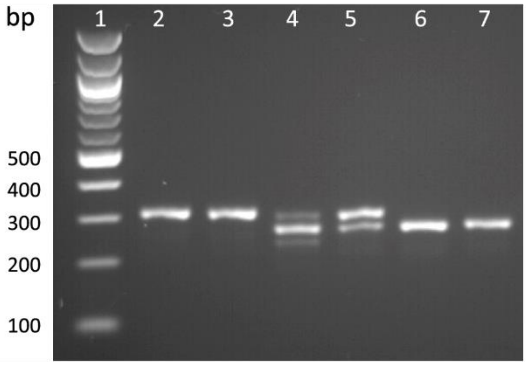


Fig. 5 PCR of the targeted urease fragment following growth of WT and mutant cell lines in nitrate or urea. NEB 100 bp ladder (*1*), WT in nitrate (*2*) and urea (*3*), M2_10 in nitrate (*4*) and urea (*5*) and M3_9 in nitrate (*6*) and urea (*7*)

rates when grown in nitrate. Therefore it seems likely that impaired growth of urease mutants in urea is due to a reduction in function of the urease gene.

There are a few possible reasons why a mutation in the urease gene appears to lead to nitrogen limitation rather than nitrogen starvation as seen in *P. tricornutum*. Cells may be able to access nitrogen from another source, separate to the breakdown of urea via urease. Some algae have an alternative pathway for breakdown of urea but this has only been found in Chlorophyceae [47] and blast searches show no evidence of urea carboxylase or allophanate hydroxylase, the enzymes involved in this pathway, in *T. pseudonana*.

The urease gene may still be active but with lower functionality. In *T. pseudonana* urease is modelled to be 807 amino acids. Urease consists of multimers of three subunits: gamma, beta and alpha, which in TP are translated as one protein. The alpha sub-unit contains the active site which catalyses the breakdown of urea to ammonia [44]. The gamma subunit has no known enzymatic function [48] but may play a role in quaternary structure and stability [44, 49].

Translations of urease sequences with both 37 and 38 nt deletions show frame shifts and early stop codons after the deletion in the gamma sub-unit, leading to major disruption of the gamma sub-unit, nonsense down-stream and short products of 24 or 44 amino acid residues (Fig. 6). Since all mono-clonal bi-allelic mutants tested for growth in urea had either two alleles with a 37 nt deletion or both a 37 and 38 nt deletion, it was predicted that the urease gene would no longer be functional. However, several mechanisms exist in eukaryotes

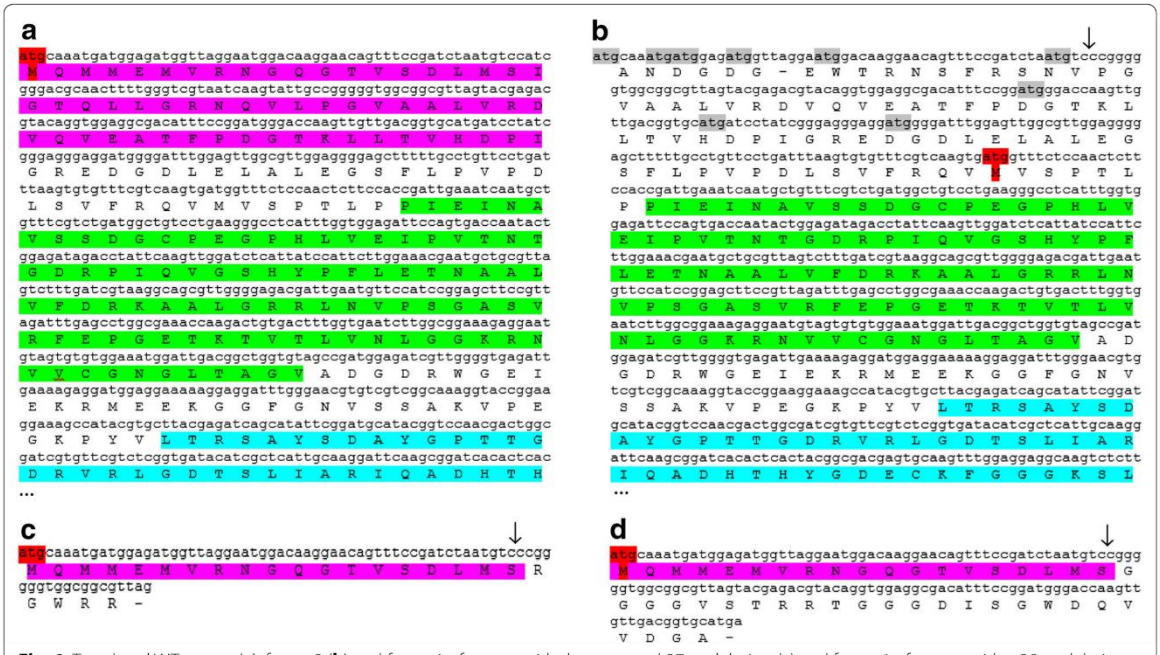


Fig. 6 Translated WT urease (**a**), frame 3 (**b**) and frame 1 of urease with the expected 37 nt deletion (**c**) and frame 1 of urease with a 38 nt deletion (**d**). Position of deletion indicated by ↓. The model WT protein contains 807aa. The figure shows the initial 260 amino acids for **a** and **b** including the start of the alpha sub-unit. Translations are identical for the unshown segments. Gamma (*pink*), Beta (*green*) and Alpha (*blue*) sub-units are highlighted in order. Expected start codon (*red*) and upstream out-of-frame start codons (*grey*) are highlighted

which can allow translation of the protein from start codons later in the coding region. These include leaky initiation, re-initiation of ribosomes and internal ribosome entry sites (IRES) [50]. IRES have been shown to become active in yeast following amino acid starvation [50]. If an in-frame translation can occur after the deletion at an IRES or via a mechanism such as re-initiation then the active site located in the alpha-subunit could still be present. The first in-frame ATG after the deletion would start translation of the protein just before the beta sub-unit, leading to an N-terminal truncated protein without the gamma sub-unit but with both the beta and alpha sub-units (Fig. 6). Earlier start codons are predicted to result in non-sense and early stop codons.

The 5′ end of the urease coding region was targeted to induce a frame shift and disrupt the protein early on, however it may be better to target the active site or entirely remove the gene. Precise deletions larger than a gene using CRISPR-Cas and two sgRNAs have been previously demonstrated [42].

Conclusions

The main aim of this research was to edit the genome of *T. pseudonana* using CRISPR-Cas. We have demonstrated that this can be achieved with both precision and efficiency. Twelve percent of initial colonies and 100% which screened positive for Cas9 showed evidence of a mutation in the urease gene, with many sub-clones showing precise bi-allelic 37 nt deletions from two sgRNA DSBs. Screening for the deletion by PCR allowed efficient identification of bi-allelic mutants and Golden Gate cloning allowed easy assembly of a plasmid for CRISPR-Cas. This included adapting the system for *T. pseudonana* by including endogenous promoters and two specific sgRNAs. Due to the flexible modular nature of the cloning system, this can be easily adapted for other genes in *T. pseudonana*. A variety of available online tools were used to design two sgRNAs that would target the early coding region of the urease gene. There is a significant difference between the phenotype of the knock-out cell lines in urea compared to nitrate. Knock-out of the urease gene is expected to have a negative impact on nitrogen acquisition from urea. This appears to be the case, however, as growth rate and cell-size was reduced rather than growth being prevented, this suggests that function of the urease may have been impaired rather than removed or an alternative source of nitrogen was available.

The CRISPR-Cas method has significant potential for future work from both an ecological and biotechnology perspective in *T. pseudonana* and can potentially be easily adapted for many other algal species.

Additional file

Additional file 1: Figure S1. Final spreadsheet for choosing sgRNAs.

Authors' contributions
AH and TM conceived the project. AH designed and conducted the laboratory and bioinformatics work. AH wrote the paper with contributions from VN, TM and SK. All authors read and approved the final manuscript.

Author details
[1] School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK. [2] The Sainsbury Laboratory, Norwich Research Park, Norwich NR4 7UH, UK.

Acknowledgements
Thanks to The Sainsbury Laboratory for supplying the Golden Gate destination vectors and linkers. We are thankful to Lewis Dunham for his contributions to the 5′ RACE for identifying the U6 promoter.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Please contact the first or corresponding author for constructs developed in this study.

Consent for publication
All authors have given consent for the manuscript entitled 'Editing of the urease gene by CRISPR-Cas in the diatom *Thalassiosira pseudonana*' to be published.

Funding
This work has been funded by a PhD studentship from the Natural Environment Research Council (NERC) awarded to Amanda Hopes. TM acknowledges partial funding from NERC (NE/K004530/1) and the School of Environmental Sciences at University of East Anglia, Norwich.

Received: 20 July 2016 Accepted: 10 November 2016
Published online: 24 November 2016

References

1. Kooistra WHCF, Medlin LK. Evolution of the diatoms (Bacillariophyta). Mol Phylogenet Evol. 1996;6:391–407.
2. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski B, 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, Kröger N, Lau WW, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakolm K, Valentin K, Vardi A, Wilkerson FP, Roksar DS. The genome of the diatom Thalassiosira pseudonana: ecology, evolution and metabolism. Science. 2004;306:79–86.
3. 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 J-P, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 2008;456:239–44.
4. Smetacek V. Diatoms and the ocean carbon cycle. Protist News. 1999;150:25–32.
5. Field CB. Primary production of the biosphere: integrating terrestrial and oceanic components. Science. 1998;281:237–40.

6. Falkowski PG, Raven JA. Aquatic photosynthesis. 2nd ed. Princeton: Princeton University Press; 2007.
7. Dolatabadi JEN, de la Guardia M. Applications of diatoms and silica nanotechnology in biosensing, drug and gene delivery, and formation of complex metal nanostructures. TrAC Trends Anal Chem. 2011;30:1538–48.
8. Delalat B, Sheppard VC, Rasi Ghaemi S, Rao S, Prestidge CA, McPhee G, Rogers M-L, Donoghue JF, Pillay V, Johns TG, Kröger N, Voelcker NH. Targeted drug delivery using genetically engineered diatom biosilica. Nat Commun. 2015;6:8791.
9. d'Ippolito G, Sardo A, Paris D, Vella FM, Adelfi MG, Botte P, Gallo C, Fontana A. Potential of lipid metabolism in marine diatoms for biofuel production. Biotechnol Biofuels. 2015;8:28.
10. Jeffryes C, Campbell J, Li H, Jiao J, Rorrer G. The potential of diatom nanobiotechnology for applications in solar cells, batteries, and electroluminescent devices. Energy Environ Sci. 2011;4:3930.
11. Kuczynska P, Jemiola-Rzeminska M, Strzalka K. Photosynthetic pigments in diatoms. Mar Drugs. 2015;13:5847–81.
12. Lander ES. The heroes of CRISPR. Cell. 2016;164:18–28.
13. Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol. 2014;32:347–55.
14. Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. Science. 2014;346:1258096.
15. Nymark M, Sharma AK, Sparstad T, Bones AM, Winge P. A CRISPR/Cas9 system adapted for gene editing in marine algae. Sci Rep. 2016;6:24951.
16. Shin S-E, Lim J-M, Koh HG, Kim EK, Kang NK, Jeon S, Kwon S, Shin W-S, Lee B, Hwangbo K, Kim J, Ye SH, Yun J-Y, Seo H, Oh H-M, Kim K-J, Kim J-S, Jeong W-J, Chang YK, Jeong B. CRISPR/Cas9-induced knockout and knock-in mutations in *Chlamydomonas reinhardtii*. Sci Rep. 2016;6:27810.
17. Weyman PD, Beeri K, Lefebvre SC, Rivera J, Mccarthy JK, Heuberger AL, Peers G, Allen AE, Dupont CL. Inactivation of *Phaeodactylum tricornutum* urease gene using transcription activator-like effector nuclease-based targeted mutagenesis. Plant Biotechnol J. 2015;13:460–70.
18. Daboussi F, Leduc S, Maréchal A, Dubois G, Guyot V, Perez-Michaut C, Amato A, Falciatore A, Juillerat A, Beurdeley M, Voytas DF, Cavarec L, Duchateau P. Genome engineering empowers the diatom *Phaeodactylum tricornutum* for biotechnology. Nat Commun. 2014;5(May):3831.
19. Poulsen N, Chesley PM, Kröger N. Molecular genetic manipulation of the diatom *Thalassiosira pseudonana* (Bacillariophyceae). J Phycol. 2006;42:1059–65.
20. Karas BJ, Diner RE, Lefebvre SC, McQuaid J, Phillips APR, Noddings CM, Brunson JK, Valas RE, Deerinck TJ, Jablanovic J, Gillard JTF, Beeri K, Ellisman MH, Glass JI, Hutchison C III, Smith HO, Venter JC, Allen AE, Dupont CL, Weyman PD. Designer diatom episomes delivered by bacterial conjugation. Nat Commun. 2015;6:6925.
21. Cook O, Hildebrand M. Enhancing LC-PUFA production in *Thalassiosira pseudonana* by overexpressing the endogenous fatty acid elongase genes. J Appl Phycol. 2015;28:897–905.
22. Doan TTY, Sivaloganathan B, Obbard JP. Screening of marine microalgae for biodiesel feedstock. Biomass Bioenergy. 2011;35:2534–44.
23. Malviya S, Scalco E, Audic S, Vincent F, Veluchamy A, Bittner L, Poulain J, Wincker P, Iudicone D, de Vargas C, Zingone A, Bowler C. Insights into global diatom distribution and diversity in the world's ocean. Proc Natl Acad Sci. 2015;348:201509523.
24. Shrestha RP, Hildebrand M. Evidence for a regulatory role of diatom silicon transporters in cellular silicon responses. Eukaryot Cell. 2015;14:29.
25. Scheffel A, Poulsen N, Shian S, Kröger N. Nanopatterned protein microrings from a diatom that direct silica morphogenesis. Proc Natl Acad Sci USA. 2011;108:3175–80.
26. Poulsen N, Scheffel A, Sheppard VC, Chesley PM, Kroger N. Pentalysine clusters mediate silica targeting of silaffins in *Thalassiosira pseudonana*. J Biol Chem. 2013;288:20100–9.
27. Xing H-L, Dong L, Wang Z-P, Zhang H-Y, Han C-Y, Liu B, Wang X-C, Chen Q-J. A CRISPR/Cas9 toolkit for multiplex genome editing in plants. BMC Plant Biol. 2014;14:327.
28. Weber E, Engler C, Gruetzner R, Werner S, Marillonnet S. A modular cloning system for standardized assembly of multigene constructs. PLoS ONE. 2011;6:e16765.
29. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ, Root DE. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol. 2014;32:1262–7.
30. Xiao A, Cheng Z, Kong L, Zhu Z, Lin S, Gao G, Zhang B. CasOT: a genome-wide Cas9/gRNA off-target searching tool. Bioinformatics. 2014;30:1180–2.
31. Tarleton R, Peng D. EuPaGDT: a web tool tailored to design CRISPR guide RNAs for eukaryotic pathogens. Microb Genom. 2015;1:1–7.
32. Nekrasov V, Staskawicz B, Weigel D, Jones JDG, Kamoun S. Targeted mutagenesis in the model plant *Nicotiana benthamiana* using Cas9 RNA-guided endonuclease. Nat Biotechnol. 2013;31:691–3.
33. Belhaj K, Chaparro-Garcia A, Kamoun S, Nekrasov V. Plant genome editing made easy: targeted mutagenesis in model and crop plants using the CRISPR/Cas system. Plant Methods. 2013;9:39.
34. Price NM, Harrison GI, Hering JG, Hudson RJ, Nirel PM, Palenik B, Morel FM. Preparation and chemistry of the artificial algal culture medium Aquil. Biol Oceanogr. 1989;6:443–61.
35. Pinto FL, Lindblad P. A guide for in-house design of template-switch-based 5′ rapid amplification of cDNA ends systems. Anal Biochem. 2010;397:227–32.
36. Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques. 2000;28:1102–4.
37. Brooks C, Nekrasov V, Lippman ZB, Van Eck J. Efficient gene editing in tomato in the first generation using the clustered regularly interspaced short palindromic repeats/CRISPR-associated9 system. Plant Physiol. 2014;166:1292–7.
38. Jacobs TB, LaFayette PR, Schmitz RJ, Parrott W. Targeted genome modifications in soybean with CRISPR/Cas9. BMC Biotechnol. 2015;15:16.
39. Sakuma T, Nishikawa A, Kume S, Chayama K, Yamamoto T. Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system. Sci Rep. 2014;4:5400.
40. Weber E, Gruetzner R, Werner S, Engler C, Marillonnet S. Assembly of designer TAL effectors by golden gate cloning. PLoS ONE. 2011;6:e19722.
41. Krysiak C, Mazus B, Buchowicz J. Relaxation, linearization and fragmentation of supercoiled circular DNA by tungsten microprojectiles. Transgenic Res. 1999;8:303–6.
42. Zheng Q, Cai X, Tan MH, Schaffert S, Arnold CP, Gong X, Chen CZ, Huang S. Precise gene deletion and replacement using the CRISPR/Cas9 system in human cells. Biotechniques. 2014;57:115–24.
43. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337:816–21.
44. Gupta S, Kathait A, Sharma V. Computational sequence analysis and structure prediction of jack bean urease. Int J. 2015;3:185–91.
45. Olson RJ, Vaulot D, Chisholm SW. Effects of environmental stresses on the cell cycle of 2 marine phytoplankton species. Plant Physiol. 1986;80:918–25.
46. Li W, Gao K, Beardall J. Interactive effects of ocean acidification and nitrogen-limitation on the diatom *Phaeodactylum tricornutum*. PLoS ONE. 2012;7:e51590.
47. Fan C, Glibert PM, Alexander J, Lomas MW. Characterization of urease activity in three marine phytoplankton species, *Aureococcus anophagefferens*, *Prorocentrum minimum*, and *Thalassiosira weissflogii*. Mar Biol. 2003;142:949–58.
48. Habel JE, Bursey EH, Rho BS, Kim CY, Segelke BW, Rupp B, Park MS, Terwilliger TC, Hung LW. Structure of Rv1848 (UreA), the Mycobacterium tuberculosis urease γ subunit. Acta Crystallogr, Sect F: Struct Biol Cryst Commun. 2010;66:781–6.
49. Jabri E, Andrew Karplus P. Structures of the *Klebsiella aerogenes* urease apoenzyme and two active-site mutants. Biochemistry. 1996;35:10616–26.
50. Hellen CUT, Sarnow P. Internal ribosome entry sites in eukaryotic mRNA molecules. Genes Dev. 2001;15:1593–612.

Chapter 4: SITMyb

Introduction

Transcription factors are essential for regulating gene expression. They bind to specific cis-acting DNA elements in promoters, activating or repressing transcription of target genes. Transcription factors with Myb domains are well characterised in eukaryotic organisms and form one of the predominant groups of transcription factors in Stramenopiles including diatoms (Buitrago-Florez et al., 2014; Rayko et al., 2010).

Biosilicification can be found in a range of eukaryotic organisms including higher plants, sponges, choanoflagellates, radiolarians and algae such as diatoms and haptophytes (Marron et al., 2016). Several proteins have been linked to uptake of silicon in this diverse range of organisms (Marron et al., 2016), however silicon transporters (SITs), first discovered in diatoms (Hildebrand et al., 1997), have only been found in a few other groups including chrysophytes (Likhoshway et al., 2006), choanoflagellates (Marron et al., 2013) and haptophytes (Durak et al., 2016).

Several genes and compounds are known to contribute towards the formation of the silica frustule in diatoms, however there is much about the process that is unknown. This extends to regulation of the components involved in silica metabolism. Although it has been shown that certain genes are up or downregulated under silicon limitation (Mock et al., 2008) or during recovery (Shrestha et al., 2012), how silicon metabolism is directly regulated in response to this is unclear.

Sequencing of the *F. cylindrus* genome (Mock et al., 2017) led to the discovery of a large gene with homology to both SIT and Myb domains. This is the first time a gene with both of these domains has been found, and it may help in understanding the processes involved in regulating the silica frustule.

Gene regulation is an essential part of cellular function and helps cells respond to changes in the intracellular and external environment. A range of different regulatory mechanisms exist in eukaryotes. TFs bind directly to specific sequences in gene promoters and either promote or supress transcription by recruiting or blocking RNA polymerase respectively. Transcription factors may also have transactivating domains to bind additional regulatory proteins such as coactivators/ corepressors, or signal sensing domains, both of which affect activity of the TF. Regulation may also be controlled through histones by post-translational modifications which can make DNA more or less assessable to other regulatory proteins and polymerase. Methylation of genes may also affect their ability to be transcribed and many eukaryotes also regulate genes post-transcriptionally by the use of small non-coding RNAs such as microRNAs and small interfering RNAs. While canonical microRNAs have yet to be detected in diatoms, small non-coding RNAs are highly expressed (Lopez-Gomollon et al., 2014; Rogato et al., 2014). Furthermore, a group of sRNAs in *P. tricornutum* are associated with genes involved in DNA methylation (Rogato et al., 2014)

Myb transcription factors normally contain 1-3 DNA-binding Myb domains, each of which forms a helix-turn-helix structure (Rayko et al., 2010). The Myb domain typically binds the consensus sequence YAACKG (Ogata et al., 2004; Rosinski and Atchley, 1998), although alternative or extended sequences such as YGRCGTTR and YAACKGHH have been observed (Deng et al., 1996). It has been shown that the YAAC sequence is key for DNA binding, whilst the second half of the motif may be linked to binding stability (Ording et al., 1994).

Myb TFs can have varied functions and in plants are associated with regulation of biosynthetic pathways, signalling and morphogenesis (Rosinski and Atchley, 1998). They have also been shown to control cell cycle, transcription factors and negative growth regulators in mammalian systems (Deng et al., 1996) and can act as both repressors and activators (Deng et al., 1996; Feller et al., 2011; Liu et al., 2015).

Regulatory networks in diatoms are still poorly understood, with only a few studies in *P. tricornutum* dedicated to characterising TFs (Huysman et al., 2013; Matthijs et al., 2017, 2016; Ohno et al., 2012). However genome sequencing (Rayko et al., 2010) and expression data has also been employed (Ashworth et al., 2013) to predict potential TFs and their networks.

In *Pheodactylum tricornutum* and *Thalassiosira psuedonana* heat shock factors account for the highest number of transcription factors, followed by Myb TFs (Rayko et al., 2010). Ashworth et al. (2013), predicted gene regulatory networks in *T. pseudonana* by analysing co-expression data and searching for known plant transcription factor binding sites (TFBS) from families including heat-shock factors (HSF), Myb, basic leucine zippers (bZIP), activating protein 2 (AP2) and E2 factors (E2F). This led to identification of potential transcription factors for various physiological and metabolic pathways, including over 1000 genes potentially regulated by Myb transcription factors. It was found that Myb TFBS are enriched in several groups of potentially co-expressed gene clusters in *T. pseudonana* and *P. tricornutum* including genes for tRNA sysnthesis, protein synthesis, amino acid metabolism and processing of proteins (Ashworth et al., 2016).

Four transcription factors have been characterised in *P. tricornutum* including three bZIP TFs (Huysman et al., 2013; Matthijs et al., 2017; Ohno et al., 2012) and a RGQ1 TF (Matthijs et al., 2016). The bZIP10 TF acts with blue light sensor auroechrome 1a to activate the diatom specific cyclin dsCYC2. The dsCYC2 protein binds to cyclin-dependent kinase, CDKA1, which is linked to control of the cell cycle at the G1-to-S phase (Huysman et al., 2013). A further bZIP TF, PtbZIP1, is associated with regulation of CO_2 acquisition. Three CO_2-cAMP responsive elements were found in the promoter of pyrenoidal β-carbonic anhydrase (PtCA1). PtbZIP11, a transcription factor with a basic zipper region, binds to these elements found 42-86 base pairs upstream of the transcription start site (Ohno et al., 2012). The third bZIP TF, bZIP14, is also linked to carbon metabolism and has been classified as a regulator of the tricarboxylic acid (TCA) cycle (Matthijs et al., 2017). The DNA binding domain for RING-Domain TF, RGQ1, is overrepresented in promoters of genes

upregulated during nitrogen starvation. Using yeast-1-hybrid, Matthijs et al. (2016) were able to demonstrate binding of RGQ1 to promoters upstream of nitrogen stress response genes.

As of yet, no TFs with a Myb domain have been characterised in diatoms, although function has been speculated based on activity in plants. Rayko et al (2010) found that Myb expression levels across growth conditions in diatoms *P. tricornutum* and *T. pseudonana* were low, although one SHAQKYF6 Myb TF with a single domain appeared to be constitutively expressed across conditions. This type of TF has been previously associated with the circadian clock in plants (Huysman et al., 2014).

Expression data in response to nutrient availability may also give insight into the TFs and co-regulators involved in nutrient aquisistion in microalgae. In *P. tricornutum* PtMyb5R, was up-regulated in response to ammonium (Rayko et al., 2010) and in *Chlorella spp*. the Myb transcription factor, ROC40, was highly upregulated under nitrogen starvation and may be linked to lipid accumulation (Goncalves et al., 2016). Differential regulation of dsCYCs are seen in response to changes in nitrate, phosphate, iron and silicon concentration (Huysman et al., 2010; Maheswari et al., 2009; Sapriel et al., 2009) and Huysman et al. (2010) suggest that dsCYCs may act as signal integrators for nutrients, particularly phosphate, during regulation of the cell cycle.

Ashworth et al. (2016) looked for groups of co-expressed genes under various environmental conditions by applying a Pearson correlation distance metric to log2 expression ratios of transcript pairs. Cis-regulatory sequence motifs were also included to gain insight into possible regulatory networks. Several genes involved in nitrogen metabolism were clustered, with many sharing a common DNA binding motif linked to the binding of HSFs in their promoters. It was also demonstrated that transcriptional regulation of several key processes may be conserved between diatom species. Pre-existing microarray data was used, including data generated during silica limitation in which clusters of genes involved in silica transport, membrane regulation and vesicle transport were observed. Furthermore a portal to access these clusters and investigate specific genes is publicly available (Ashworth et al. 2016).

Several studies have explored expression of genes, including transcription factors themselves under silicon limiting conditions to determine genes involved in silica metabolism (Du et al., 2014; Mock et al., 2008; Sapriel et al., 2009; Shrestha et al., 2012). TFs from several different families in *T. pseudonana* showed differential expression in response to silica, whilst several heat-shock factors in *P. tricornutum* were up-regulated in response to silica starvation (Rayko et al., 2010).

Certain key genes associated with silica metabolism are expressed depending on external silicon, including silicon transporters (Mock et al., 2008; Sapriel et al., 2009; Shrestha and Hildebrand, 2015) which appear to be regulated at a transcriptional and posttranscriptional level (Thamatrakoln and Hildebrand, 2007).

Formation of the silica frustule is closely linked to the diatom cell cycle (Brzezinski et al., 1990; Hildebrand et al., 2007) and an important aspect of silica metabolism is the acquisition of soluble silicic acid from the external environment. At high external concentrations, silicic acid can diffuse across the membrane (Thamatrakoln and Hildebrand, 2008), but diatoms also possess a group of proteins called silicon transporters (SITs) that can actively transport silicic acid into the cell. Silicon transporters were first discovered in *Cylindrotheca fusiformis* (Hildebrand et al., 1997). They have 10 transmembrane domains, each with an α-helical structure (Curnow et al., 2012; Thamatrakoln et al., 2006) and contain GXQ and MXD motifs which may act as silicic acid binding sites (Curnow et al., 2012; Sherbakova et al., 2005; Thamatrakoln et al., 2006). A coiled-coil domain can be found at the C-terminus, which is hypothesised to be positioned within the cell given the nature of the coiled-coil domain (Thamatrakoln et al., 2006). Transport across SITs is sodium dependent (Bhattacharyya and Volcani, 1980; Curnow et al., 2012; Hildebrand et al., 1997; Vrieling et al., 2007) and sodium binding sites can be found in the majority of these proteins (Curnow et al., 2012). Whilst some SITs appear to be constitutively expressed, several show differential expression depending on the concentration of environmental silicon present (Mock et al., 2008; Sapriel et al., 2009; Shrestha and Hildebrand, 2015) and it appears that they play different roles in the acquisition and regulation of silicon influx. There is evidence to suggest that certain SITs are up-regulated to provide adequate silica for frustule formation and cell division (Shrestha and Hildebrand, 2015). Some show high affinity/low capacity transport whilst others display low affinity/high capacity to cope with environmental changes and cellular function (Flynn and Martin-Jézéquel, 2000; Hildebrand, 2003; Thamatrakoln and Hildebrand, 2007). Not all SITs appear to be involved in external silicon transport (Sapriel et al., 2009) and alternative functions such as intracellular transport or regulation have been suggested (Shrestha and Hildebrand, 2015). Furthermore there is evidence that some SITs may be acting as silicon sensors (Shrestha and Hildebrand, 2015; Thamatrakoln and Hildebrand, 2007).

The silica frustule is formed within the silica deposition vesicle (SDV), an organelle with an acidic environment known to promote silica precipitation (Vrieling et al., 1999). The SDV is encapsulated by a membrane known as the silicalemma (Hildebrand and Lerch, 2015; Koester et al., 2016) and contains a variety of molecules and proteins which are responsible for precipitation and control of silica. These include silaffins (Kröger et al., 2001, 1999), silacidins (Richthammer et al., 2011; Wenzl et al., 2008), cingulins (Scheffel et al., 2011) and long chain polyamines (LCPA; Kröger et al., 2000).

Silaffins contain several post-translational modifications including several negatively charged lysine groups with polyamines, long chain polyamines and methyl groups, as well as cationic phosphoserines (Kröger et al., 2001). Pentalysine clusters formed of lysine rich peptides and phosphorylated serines are important for targeting to the biosilica (Poulsen et al., 2013). As zwitterions, silaffins form large aggregates, which can be even larger when LCPAs are included, which may be responsible for silica precipitation in vivo (Kroger et al., 2002; Poulsen et al., 2013).

In vitro, addition of silaffins in combination with LCPAs induces rapid precipitation, with length, concentration and source of LCPAs affecting nanoscale morphology (Kroger et al., 2002; Kröger et al., 2000). Different silaffins also show differences in activity, and can either activate or inhibit silca precipitation, depending on concentration (Poulsen and Kröger, 2004).

Cingulins are silaffin-like proteins that are associated with formation of the girdle band. They form part of an organic matrix known as microrings, which induce silica precipitation and act as a template, forming nanoscale patterning (Scheffel et al., 2011).

Silacidins are also highly phosphorylated and precipitate silica in the presence of LCPAs. They are highly acidic as they contain several acidic amino acids (Wenzl et al., 2008). Only phosphorylated silacidins rapidly and efficiency precipitate silica (Richthammer et al., 2011), highlighting the importance of negative phosphate groups which are also associated with nanoparticle size (Sumper et al., 2003). Furthermore due to their high activity and relative increase in concentration in the frustule under silicon starvation, it has been suggested that silacidins may be important for precipitation during low silicon availability (Richthammer et al., 2011).

Chitin fibres are associated with the silica frustule forming part of the organic matrix (Brunner et al., 2009; Durkin et al., 2009) which has been hypothesised to provide a template for silica deposition (Brunner et al., 2009) or act as a supporting structure (Hildebrand and Lerch, 2015). Another component of the organic matrix, P150, which is particularly associated with the girdle band contains 3 potential chitin binding sites (Davis et al., 2005), suggesting potential interaction with the chitin fibres. Carbohydrates are also associated with this matrix including mannose which is particularly prominent (Tesson and Hildebrand, 2013) and well as glycoproteins known as frustulins (Kroger et al., 1996; Kröger et al., 1994).

The cytoskeleton appears to play an important role in micro and meso-scale patterning of the frustule (Tesson and Hildebrand, 2010a, 2010b). Actin is associated with the SDV and is thought to control its positioning and shape (Pickett-Heaps et al., 1990), whilst microtubules may act to strengthen the SDV (Tesson and Hildebrand, 2010a). Both actin and tubulin are associated with formation of valves and girdle bands (Tesson and Hildebrand, 2010b). Recent work suggests that silacidins may also affect the frustule at a micro-scale, with an increase in cell size observed upon knockdown of silacidins in *T. pseudonana* (Kirkham et al., 2017).

Signalling and transport mechanisms are still largely unknown, although expression analysis highlights possible genes involved in these processes under silicon limitation or during frustule formation (Mock et al., 2008; Shrestha et al., 2012). In general, transcriptome and proteome analyses under these conditions, suggest that several genes associated with silica metabolism are regulated in response to silicon concentration or requirement (Du et al., 2014; Mock et al., 2008; Sapriel et al., 2009; Shrestha et al., 2012). These include but are not limited to genes involved in silicon transport, chitin synthesis, carbohydrate metabolism, silica precipitation and the

cytoskeleton. In addition genes involved in regulation such as transcription factors and helicases, as well genes for signal transduction and post-translational modification such as kinases can be found.

This chapter focuses on the SITMyb gene and explores its potential as a regulatory protein and its possible link to silica metabolism. Modelling of the gene aims to investigate potential domains and motifs that support a link to silica metabolism or regulation. Previously generated RNA-seq data has also been explored along with in-vitro gene modelling. The gene model is then used to design an overexpression construct in *F. cylindrus* and to explore potential SITMyb binding domains through yeast-1-hybrid, using both the modelled gene and the modelled Myb domain.

Methods

Modelling the SITMyb gene.

Two alleles of the SITMyb gene are present in the *Fragilariopsis cylindrus* genome: ID 233781 on scaffold_1:5636864-5642151 (-) and ID 250586 on scaffold_31:566466-571660 (+). Blastn and Blastp searches were performed against the NCBI database. The nucleotide sequences from both alleles were aligned using EMBL MAFFT. The protein sequence was also run through ScanProsite (de Castro et al., 2006) at high sensitivity to check for conserved domains and motifs. Structure of the Myb and SIT protein domains were modelled as well as the entire protein using Phyre2 (Kelly et al., 2015) with intensive modelling mode and SWISS-MODEL (Arnold et al., 2006). Nuclear localisation signals were predicted with cNLS mapper (Kosugi et al., 2009), NucPred (Brameier et al., 2007) and Scanprosite. Coiled-coils were predicted using Expasy COILS (Lupas et al., 1991). The protein sequence was searched for potential silicic acid binding motifs GXQ and MXD (Curnow et al., 2012; Sherbakova et al., 2005; Thamatrakoln et al., 2006).

F. cylindrus RNA-seq expression data generated by Jan Straus (Mock et al., 2017; Strauss, 2012) under the following conditions; control, high and low temperature, iron limitation, high carbon dioxide, prolonged darkness, red and blue light and silicon limitation, was visualised with the Broad Institute Integrative Genomics Viewer (IGV; Thorvaldsdóttir et al., 2013) to see expression values across the gene for both alleles compared to the genome.

The SIT domain of the SITMyb gene was aligned to several other SIT sequences found in diatoms (39), plants (7), chronoflagellates (7) and coccolithophores (1) from NCBI and UniProt using ClustalX with a gonnet matrix (Larkin et al., 2007) as demonstrated by Thamatrakoln et al. (2006) for comparison of diatom SITs. As the SIT domain from the SITMyb gene appears to align to C-terminal sequences from diatoms, a further alignment was carried out for this region. Radial phylograms were drawn using the neighbour joining method with a bootstrap of 100.

Promoter sequences were checked for the consensus Myb binding domain defined as YAACKG (Biedenkapp et al., 1988; Ogata et al., 2004) or the extended Myb consensus sequences YGRCGTTR and YAACKGHH (Deng et al., 1996).

RACE and RT-PCR of the SITMyb gene.

F. cylindrus total RNA was extracted from exponentially growing cultures using a Qiagen RNeasy kit. Rapid amplification of cDNA ends (RACE) was carried out to determine the 3' and 5' ends of the SITMyb transcript for both alleles. Initially this was performed using the Ambion RNA ligase mediated (RLM)-RACE kit. Later 5' template switching oligo RACE (TSO; Pinto and Lindblad, 2010) was used. Both 3' and 5' RACE was carried out with the RLM method following manufacturer's instructions. Mg^{2+} concentration (1mM, 2mM and 3mM) and temperature (57-69°C in 3°C increments) was optimised for the PCR steps of the RACE method.

5' RLM RACE: briefly, ten 10µg of *F. cylindrus* total RNA was treated with Calf Intestine Alkaline Phosphatase (CIP) for 1 hour at 37°C. The CIP reaction was terminated with ammonium acetate before purifying the RNA with phenol:chloroform and isopropanol and resuspending in nuclease free water. RNA was then treated with Tobacco Acid Pyrophosphatase (TAP) in a one hour reaction at 37°C before ligating the 5' RACE adapter with T4 RNA Ligase at 37°C for 1 hour. A –TAP control was also carried out by performing the ligation reaction on the CIP treated RNA. Reverse transcription was performed at 42°C for 1 hour on the ligated RNA using M-MLV Reverse transcriptase and random decamers. cDNA produced in this step was then used as a template for the outer PCR reactions. KAPA Long range DNA polymerase was used for both outer and inner PCR reactions with an annealing temperature of 60°C and 2mM $MgCl_2$, according to the manufacture's protocol. Along with the provided 5' RACE outer primer, reverse gene specific (GS) primers at locations 5'4 and 5'2 (see Table 4.1 and Figure 4.1) were used. As well as the –TAP control, a no target control with water instead of template was performed for each primer set. Nested inner PCRs were performed with 1µl of the outer PCR reaction. Along with the provided forward 5' RACE Inner primer, GS reverse primers at locations 5'1 and 5'3 were used. Both GS reverse primers were used in combination with both outer PCR reactions (using primers at 5'4 and 5'2). Products were run on a 2% agarose gel.

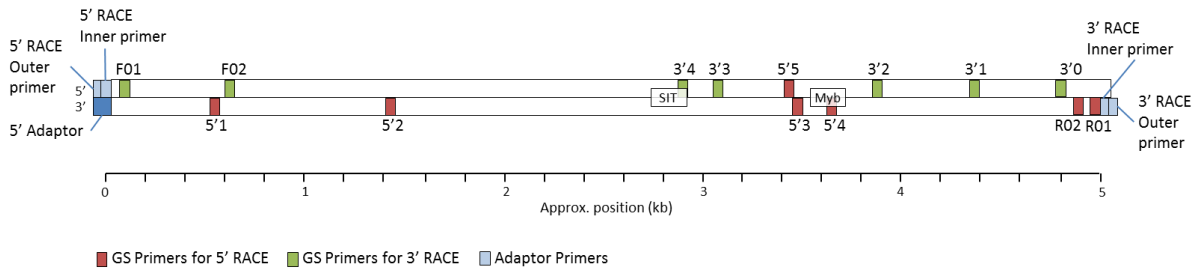


Figure 4.1. Positions of RACE primers and primers for amplification of the full coding region. Red boxes are gene specific (GS) primers for 5' RACE, green boxes are GS primers for 3' RACE and light blue boxes are primers from the RLM RACE kit, which in the case of the 5' RACE are complementary to the 5' adaptor, and for the 3' RACE correspond to the 3' adaptor. Major increments of the scale bar are at 1kb, smaller increments are at 200bp.

Primer name	Application	Forward/ Reverse	Position	Sequence	Tm (°C)	% GC
5'MybRace add_1	5' RACE	R	5'1	TTGCTGTCCGTCATCGTGGTAGT	60.3	52.2
5'MybRACE add_2	5' RACE	R	5'2	TCGCGGAACAACATGCCATTGA	59.9	50
myb 5 RACE Inner	5' RACE	R	5'3	CCTGTTGGGCTACGGTGTGTTCTT	61	54.2
myb 5 RACE Outer	5' RACE	R	5'4	GCCATTGCCTTCGCAACTTCTTC	58	45.8
myb 5 RACE GS	5' RACE	F	5'5	ACACACCGTAGCCCAACAGGATATT	60	48
Myb 3' Race A1	3' RACE	F	3'0	CCAAAGCTAAAGAGCGCGAGC	59.1	57.1
Myb 3' Race A2	3' RACE	F	3'0	CCAAAGCGAAAGAGCGCGAAC	59.6	57.1
3'MybRACE add_2	3' RACE	F	3'1	AAGAAAGCCAGGGCGTCAAAGT	59.5	50
3'MybRACE add_1	3' RACE	F	3'2	ACGGTGGCACCATTAGAGGATTGA	60.2	50
myb 3 RACE Inner	3' RACE	F	3'3	CGTCCAATCCCACGACCAACAATG	60	54.2
myb 3 RACE Outer	3' RACE	F	3'4	AGTCATATGCAGGGAGCTTATTC	54	43.5
Myb Coding+UTR_F	Full cDNA	F	F 01	TCTGCTGACAAAGGAAGTACCTGA	57.7	45.8
Myb Coding_F	Full cDNA	F	F 02	ATGGAAGCAACAACCACAGGA	57.1	47.6
Coding+UTR_R A1	Full cDNA	R	R 01	CATTGCATTTATTCATGCTATTACTCATG	53.5	31
Coding+UTR_R A2	Full cDNA	R	R 01	CATTGCATTTATTCATATTATTACTCATG	50.1	24.1
Myb Coding_R	Full cDNA	R	R 02	TCTTACACCATTACAATTTCTTCTTCCTC	55.1	34.5

Table 4.1. Primers used in RLM RACE and amplification of the full coding sequence.

3' RLM RACE: One µg of *F. cylindrus* RNA was used for reverse transcription (RT) with the 3' RACE adapter and M-MLV Reverse transcriptase at 42°C for one hour. Initially the outer PCR was performed with the provided 3' RACE Outer Primer and the GS primer at location 3'4 (Table 4.1 and Figure 4.1). This was followed by an Inner PCR with the 3' RACE Inner Primer and the GS primer at 3'4. Later PCRs involved only the outer PCR and used GS primers at location 3'4, 3'2 and 3'1. As with 5' RACE, primers were annealed at 60°C and KAPA Long range DNA polymerase was used. Products were run on a 2% agarose gel.

Transcript amplification: PCRs to amplify the full coding region, with and without UTRs were carried out. cDNA from the 3' and 5' RACE RT reactions was used as a template, as well as gDNA as a positive control. Primers at location F02 and R02 were used in amplification of the coding region. Primers F01 and R01 were used for amplification of the coding region + UTR. Two variants of R01 were designed due to SNPs present between SITMyb alleles. Amplification of the whole transcript was unsuccessful. As a result the transcript was amplified as overlapping fragments with the primer sets for both alleles shown in Table 4.2.

Allele/ Region	Forward Primer	Reverse Primer	Expected size	Tm
A1 – 5’	5’ RACE inner CCTGTTGGGCTACGGTGTGTTCTT	5'MybRace add_1 (5’1) TTGCTGTCCGTCATCGTGGTAGT	>465	60.3-61.1
A1 – 1st	Myb Coding+UTR_F (F01) TCTGCTGACAAAGGAAGTACCTGA	A1 1st CAACATGCCATTGATTAGCACTGTTG	1289	57.2-57.7
A1 – 2nd	A1 2nd F GGTCGCTTCTTAAAAGTCAACAGTG	A1 A2 2nd R GAATAAGCTCCCTGCATATGACTC	1436	55.2-56.3
A1 – 3rd	myb 3 RACE Outer (3’4) AGTCATATGCAGGGAGCTTATTC	A1 3rd TCGTCGTAGTTGTTCTCATCATT	1471	53.9-54.0
A1 – 4th	A1 4th ACCATCTGAACGTTTAGAAGATAATGAT	Myb Coding_R (R 02) TCTTACACCATTACAATTTCTTCTTCCTC	882	54.4-55.1
A1 – 3’	A1 3' RACE CCAAAGCTAAAGAGCGCGAG	3’ RACE outer	>272	55.1-56.5
A2 – 5’	5’ RACE inner CCTGTTGGGCTACGGTGTGTTCTT	A2 5' RACE CAGCTTGAGATTCAACTACTCTGCGAC	>949	59.1-61.1
A2 – 1st	Myb Coding_F(F02) ATGGAAGCAACAACCACAGGA	A2 1st AAGGGTGCACTTTCGCTTG	920	56.2-57.1
A2 – 2nd	A2 2nd F CAAGCGAAAGTGCACCCTT	A1 A2 2nd R GAATAAGCTCCCTGCATATGACTC	~1252	45.8-52.6
A2 – 3rd	myb 3 RACE Outer (3’4) AGTCATATGCAGGGAGCTTATTC	A2 3rd CGTCGTAGTTGTTCTCGTCATC	1348	54.5-55.3
A2 – 4th	A2 4th CATCTGAACGTTTAGAAGATGATGAC	Myb Coding_R (R 02) TCTTACACCATTACAATTTCTTCTTCCTC	880	53.8-55.1
A2 – 3’	A2 3' RACE CCAAAGCGAAAGAGCGCGAA	3’ RACE outer	>272	55.1-58.9

Table 4.2. Primers used to amplify overlapping fragments from cDNA of both SITMyb alleles. Bases in red indicate the presence of SNP between alleles. A1 is allele 1 (ID 233781) and A2 is allele 2 (ID 250586)

5’ TSO RACE. Template switching oligo RACE was performed according to Pinto and Lindblad (2010) using the template switching oligo (TSO: GTCGCACGGTCCATCGCAGCAGTCACAGGGGG) and U_SENSE primer (GTCGCACGGTCCATCGCAGCAGTC) described in the paper. Briefly 700ng of *F. cylindrus* total RNA was denatured with dNTPs and either the GS primer at location 5’3 or R02 (Table 4.1). A no template control was included at this stage. cDNA synthesis was carried out on denatured DNA with RevertAid H- reverse transcriptase. A no RT control was also added to check for DNA contamination. The template switching oligo was then added along with $MnCl_2$. The reverse transcriptase adds C residues to the end of the sequence in the presence of $MnCl_2$, providing a complementary sequence for the TSO, which acts as an adaptor for PCR, to anneal to. Hotstart Phire II DNA polymerase (Thermo fisher) was used to amplify fragments with the forward USense primer and either the SIT1 (ATGCACCACGGAGTATTG) or the SIT2

(TGATGTTGTCGTCGTAGTTG) GS reverse primer. An annealing temperature of 60°C with an extension time of 1 minute was used.

Creating a construct for overexpression of the SITMyb gene in *F. cylindrus*.

The SITMyb gene (233781) from +2398 to +5536 was cloned into a puc19 backbone with the *F. cylindrus* FCP promoter and terminator described in the transformation chapter. The sequence for a 6x His-tag was included before the stop codon at the end of the SITMyb sequence for potential isolation and labelling of the protein. Initially the pucFCPshble plasmid made in the transformation chapter was used as a backbone. However, this did not result in clones with the correct insert and the FCP:SITMyb cassette was assembled into puc19 instead. Following assembly into puc19, the FCP:SITMyb cassette was combined into a single construct with FCP:shble using Golden Gate cloning. Gibson assembly (GA), was used to assemble fragments in one reaction. Primers for amplification (Table 4.3) include sequences complementary to the adjacent fragment in assembly, at the 5' end, giving a 40bp overhang between adjacent fragments during GA. Phusion DNA polymerase (NEB) was used to amplify all fragments, as described in the manufacturer's instructions, including puc19 (1-2), the FCP promoter (3-4), SITMyb in three parts (5-10) and the FCP terminator (11-12). The puc19 plasmid was used as a template for vector amplification and *F. cylindrus* gDNA was used for all other fragments. The Tm of all primers can be found in Table 4.3. Following PCR, products were run on a TAE agarose gel, excised and purified using a GFX PCR DNA and Gel Band Purification Kit (GE). Products were eluted into nuclease free water. Gibson assembly was carried out as described in the transformation chapter, using the Ford (2013) GA mastermix, 50ng of vector and a 3 times molar concentration of each fragment. The pucFCP:SITMyb construct was screened by restriction digestion using SphI and BamHI, as well as PCR of the insert using primers 13 and 14 (Table 4.3). The construct was sequenced using primers 3, 5, 7, 9, 11, 15, 16, 17 and 18.

Golden-gate cloning. Details on Golden-gate cloning methods can be found in the CRISPR-Cas chapter. Site directed mutagenesis (SDM) was carried out on the pucFCP:SITMyb plasmid to remove BsaI and BpiI sites, using a Q5 site-directed mutagenesis (SDM) kit and accompanying protocols (NEB). BsaI from the FCP terminator was removed using primers 1 and 2 (Table 4.4). Primers 3-4 were used to remove a BpiI site from the coding region, with a synonymous substitution. A BpiI site within the intron was removed either by a base change (primers 5-6) or by removing the entire intron (7-8). The domesticated FCP:SITMyb cassette was amplified using primers 9 and 10, and assembled into a L1 pICH47742 backbone. The FCP:shble cassette was amplified from pucFC_FCPshble (transformation chapter), with the same primer set (9 and 10) and cloned into L1 backbone pICH47732. Domesticated CEN-ARS-HIS (CRISPR-Cas chapter) was cloned into L1 backbone pICH47752, as previously described. All three cassettes were assembled in L2 backbone pAGM4723 along with linker pICH41766, giving either pAGM_SITMybOE (with domesticated intron, Figure 4.2) or pAGM_SITMybOE_IR (with intron removed). Constructs were screened by restriction digest with XbaI and EcoRV and checked by sequencing.

Number	Primer	Primer sequence	Tm (°C)
1	Vector F	cagcacaacaacaacgactAGGCATGCAAGCTTGGC	63
2	Vector R	atttctatgccttactttgGGGTACCGAGCTCGAATTCAC	64
3	Prom F	attcgagctcggtacccCAAAGTAAGGCATAGAAATAATC	56
4	Prom R	TTGTCATTCGCCGTTTTCATtttgatatataagtttgtttttggtag	58
5	SIT1 F	aaacaaacttatatatcaaaATGAAAACGGCGAATGAC	58
6	SIT1 R	ATGCACCACGGAGTATTG	59
7	SIT2 F	TGCACAACAACAGATGTATC	58
8	SIT2 R	TGATGTTGTCGTCGTAGTTG	60
9	SIT3 F	GAAGATAATGATGAGAACAAC	54
10	SIT3 R	aataaggattaataaaatgcTTAGTGATGATGATGATGATGCACCATTACAATTTCTTCTTC	55
11	Term F	ATCATCATCATCATCACTAAgcattttattaatccttatttgatcg	58
12	Term R	gccaagcttgcatgcctAGTCGTTGTTGTTGTGCTG	59
13	Puc19 Insert F	GCTGCAAGGCGATTAAGTTG	62
14	Puc19 Insert R	GCTCGTATGTTGTGTGGAATTG	62
15	Prom_SIT1_seq	tttaccgctttcgatcttctc	60
16	SIT1_SIT2_seq	TTTGGTTGTGGTGGTAGTG	60
17	SIT2_SIT3_seq	CCGTCACATCATACACACATAG	61
18	SIT3_Term_seq	ACAAGATGCTCGTGACTATATG	60

Table 4.3. Primers for cloning the FCP:SITMyb cassette into a puc19 backbone using Gibson Assembly. 1-12) Primers for amplification prior to GA. 13-14) Primers for amplification of the insert. 15-18) Primers for sequencing junctions of the final construct. Upper case bases denote the part of the primer which anneals to the original template. Bases in lower case are from the adjacent sequence in assembly and lead to fragments with ~40bp overlaps. Tm refers to the template specific sequence of the primers.

Number	Primer name	Primer sequence	Tm (°C)	Annealing temp (°C)
1	FC_FcpT_Mut_F	ATATAGTGAGtCCCTTCCGTTGAC	64	64
2	FC_FcpT_Mut_R	TCGAATCAATGAATCGATCAAATAAGG	61	
3	SITMyb_mut code_F	ATGGTTGTCTaCTTCCGAAAG	56	59
4	SITMyb_mut code_R	CTTTGGTTCGACGTAGTTC	58	
5	SITMyb_mut_intron_F	GAGCAAGGTAAGTCTcCCAAAAC	57.6	63
6	SITMyb_mut_intron_R	CATCAGGGCGCCTATATG	58.8	
7	SITMyb_mut_intronR_F	ACCCTTTGGATTATCTTGGCAC	62.4	65
8	SITMyb_mut_intronR_R	CTTGCTCCATCAGGGCG	61.8	
9	GG_FC_vFcpP_F	a**ggtctc**aggagGCTGCAAGGCGATTAAGTTG	61.5	64
10	GG_FC_vFcpP_R	a**ggtctc**aagcgGCTCGTATGTTGTGTGGAATTG	61.5	

Table 4.4. Primers used to construct the FCP:SITMyb overexpression construct with Golden Gate cloning. Primers 1-8 were used for SDM to remove BsaI and BpiI. The lower case letter in the SDM primers denotes the base change. Primers 9-10 were used to amplify the FCP:SITMyb cassette for assembly into the pICH47742 level 1 backbone.

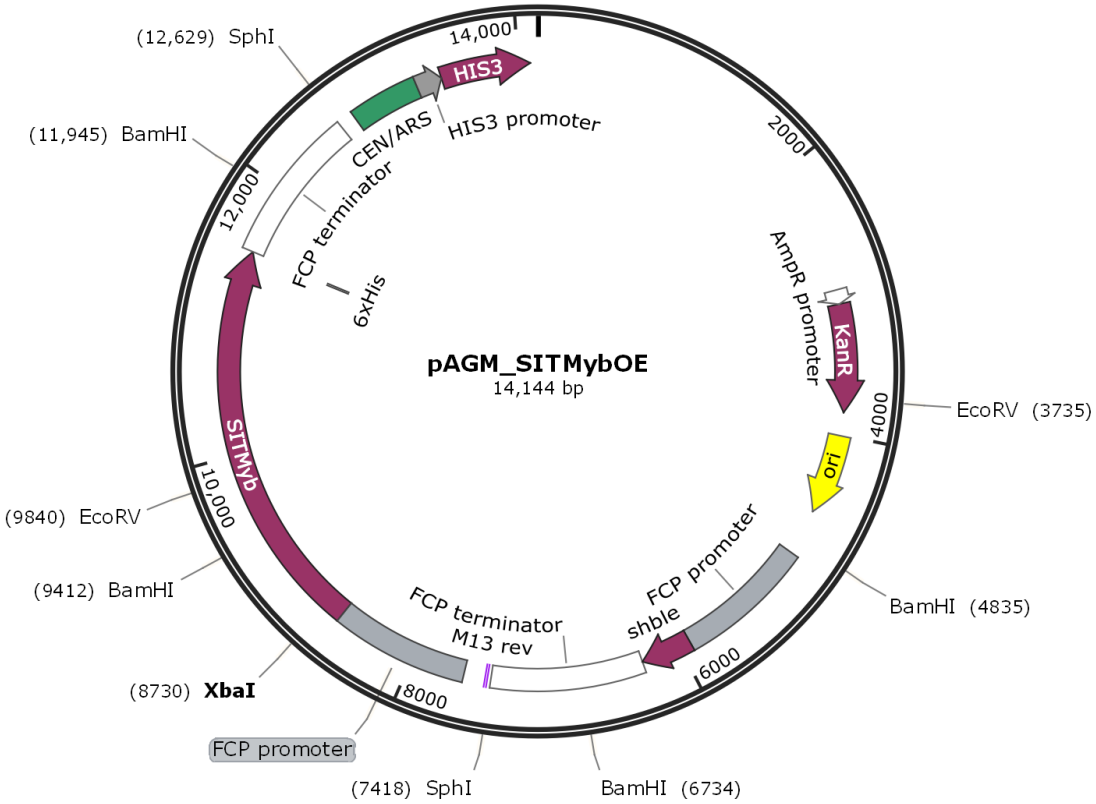


Figure 4.2. Vector map of pAGM_SITMybOE construct. Map created with SnapGene.

Transformation of SITMyb overexpression constructs into *F. cylindrus*.

Constructs pAGM_SITMybOE and pAGM_SITMybOE_IR were introduced into exponentially growing *F. cylindrus* cells through microparticle bombardment as described in the Transformation chapter. Positive and negative controls were included with pucFCFCPshble:FCPegfp and water respectively. All shots were carried out in triplicate with a 1550psi rupture disc. Selection was carried out on zeocin plates as previously described.

Screening *F. cylindrus* clones. PCR of gDNA, RT-PCR and western blots.

Colonies were picked and grown to exponential phase as described in the transformation chapter. One pAGM_SITMybOE colony (SITMyb_OE 1) and 3 pAGM_SITMybOE_IR (SITMybIR_OE_2, 4 and 6) colonies were screened for the overexpression construct. Genomic DNA was extracted from 1ml of each culture using an Easy DNA gDNA purification kit (ThermoFisher) according to protocol #3 of the product manual. PCR was carried out with 500ng of gDNA in a 20ul reaction using Phusion DNA polymerase (NEB) according to the NEB protocol. Forward primers 15 and 16 (Table 4.3) were used with reverse primer SITMyb OE_R (gcTTAGTGATGATGATGATGATGCAC), to amplify from within the promoter or early in the coding region to the His-tag to check for presence of the overexpression cassette.

RNA was extracted from exponentially growing cells using a Qiagen RNeasy kit. DNA was removed by incubating 1µg of extracted RNA with DNase I (NEB) at 37°C for 10 minutes, before stopping the reaction by addition of EDTA to a final concentration of 5mM and heat inactivation at 75°C for 10 minutes. The reaction was purified using a Qiagen RNeasy MinElute Cleanup kit and eluted into 10 µl of RNase free water. One µl of the eluate was used in a 10µl reverse transcription reaction with Superscript III (SSIII, Thermo Fisher) and oligo dT according to the SSIII protocol. Two controls were included at this stage– one with no RNA (RT control) and one with RNA but no SSIII (DNA control). The reaction was incubated at 50°C for 1 hour, before inactivating the SSIII by heating to 70°C for 15 minutes. RNA was removed by incubating with RNAse H for 20 minutes at 37°C. PCR using GoTaq DNA polymerase (Promega) with a final concentration of $MgCl_2$ at 1.25mM was carried out with 1µl of the RT reaction and primers SIT1_SIT2 seq (Table 4.3) and SITMyb OE_R SITMyb OE_R targets the overexpression transcript by annealing to the His-tag sequence at the 5' end. Products were visualised on an agarose gel.

Protein was extracted from SITMyb_OE 1, SITMybIR_OE 2 and WT cultures using lysis buffer, denaturing lysis buffer and XTractor buffer (Clonetech).

Extraction with lysis buffer: Cultures was pelleted and cells resuspended in lysis buffer (50mM Tris-HCl pH6.8, 2% SDS). One hundred µl of buffer was used for every 100ml of culture. Four µl of protease inhibitor, EDTA free (Thermo Fisher) was added (for a final 1x concentration) for every 100µl of lysis buffer. Cells were vortexed briefly to homogenize and incubated at room temperature for 30 minutes. The lysate was spun down at 13000 rpm for 30 minutes at 4°C, and the supernatant transferred to a clean Eppendorf tube. The crude protein extract was kept on ice until needed or stored at -80°C.

Extraction with denaturing lysis buffer: Cultures were treated as above, but with denaturing lysis buffer (100mM NaH_2PO_4, 10mM Tris-Cl, 8M urea, NaOH to pH 8.0).

Extraction with XTractor buffer : Protein was extracted according to the manufacturer's protocol. Briefly cells were pelleted at 3000 x g for 5 minutes at 4°C and washed with 2x PBS. Cells were resuspended in XTractor buffer, using 100µl of buffer for every 100ml of culture, by vigorously vortexing. Protease inhibitor was added before incubating for 10 minutes at room temperature. Lysate was clarified by spinning at 12000 x g for 20 minutes at 4°C, and transferred to a new Eppendorf tube. Supernatant was kept on ice until needed or stored at -80°C.

His-tag purification: Crude protein extracts were run through His-tag purification columns to enrich proteins with a His-tag. A Capturem His-tagged Purification Miniprep Kit (Clonetech) was used. Briefly, 400µl of buffer (buffer used to produce protein lysate) was used to equilibrate the spin column before loading 400µl of cleared protein lysate and spinning at 11000 x g for 1 minute at room temperature. Columns were washed with wash buffer under the same conditions and protein eluted into 100µl of Elution buffer.

Prior to His-tag purification protein concentration was measured by 260nm absorbance on a nanodrop. Following His-tag purification concentration was measured using the Bradford protein assay. Quick Start reagent (BIO-RAD) was used in a 1:1 ratio with the eluted protein according to the 'Microassay protocol' in the BIO-RAD instruction manual. Concentration was calculated against BSA standards (0-100µg/ml).

Western blots: Both crude protein extracts and His-tag purified extracts were run on NuPAGE 4-12% BIS-Tris gels using the X-cell Surelock Mini-gel electrophoresis system (Invitrogen). Protein extracts were run for 35 minutes at 200V alongside 5µl of Broad Range Color Prestained Protein Standard ladder (NEB), 1-5µl of BenchMark His-tagged Protein Standard Ladder (Invitrogen) and a positive control (His-tagged NAD dependent ligase, 80kDa) at 5, 1 and 0.25µg/ml. For crude extracts 13µl at 10mg/ml were loaded into wells. For His purified proteins 13ul at 100µg/ml were loaded. Proteins were transferred onto a nitrocellulose membrane according to the X-cell Surelock manual for 1 hour at 30V. Membranes were blocked for 1 hour in 5% BSA PBST (1 x PBS with 0.05% Tween 20) before incubating with a Cell signalling HIS antibody (rabbit) overnight at 4°C. A 1:1000 dilution was used in PBST with 5% BSA. Membranes were washed 3 x in PBST for 10 minutes each on a rocker. An anti Rabbit IgG HRP tagged secondary antibody (Promega) at a dilution of 1:2500 in PBST was added to membranes for 1 hour at room temperature, and washed 3 x in PBST for 10 minutes each. Blots were visualised by adding ECL Western blotting substrate (Pierce) for 2 minutes and chemiluminscence imaged in 30 second intervals for up to 5 minutes.

Yeast 1 hybrid

Yeast one hybrid was carried out based on the method by Yan and Burgess (2012). Plasmids pYOH1 and pYOH366 were kindly provided by Shawn Burgess. *F. cylindrus* gDNA was fragmented and cloned into pYOH366 upstream of the URA3 gene, to create pYOH366-g. The SITmyb gene and the Myb domain were each cloned into pYOH1 for expression in yeast. The gDNA library and transcription factor (TF) constructs (Figures 4.4 and 4.5 respectively) were transformed into two mating types of the same yeast strain and combined by mating. If the transcription factor is expressed and binds to a site in the *F. cylindrus* library, URA3 is expressed allowing growth in –uracil media. The pYOH366-g plasmid in the positive clone can then be amplified and sequenced to identify potential binding sites. Methods specific to YIH in *F. cylindrus* and alterations to the method are detailed below. An overview of the process as shown in the Yan and Burgess (2012) paper can be seen in Figure 4.3.

Generating pYOH366-g

Optimising gDNA digest**.** Genomic DNA was extracted from 800ml of exponentially growing *F. cylindrus* using an Easy DNA genomic extraction kit (Thermo Fisher). Digestion reactions were carried out at 37°C in Cutsmart buffer with 4nt cutter MluCI (NEB), which cuts site AATT and allows cloning of the fragments into the EcoRI site in pYOH366. Ten µg of gDNA was standardly used per 10µl reaction volume. Several different parameters were used to optimise fragment size.

MluCI at 2U and 0.01U per 10µl was trialled with incubation times at 0, 2, 5, 7.5 and 10-60 minutes in 10 minute intervals. Scaled up reactions in 50µl were incubated between 0 and 10 minutes with 10U of enzyme. Reactions were either stopped by placing on ice, or by adding 10 µl of 50 mM EDTA (pH 8.0). Different methods of purification were also tested. Reactions were either run on a gel and bands between 200-1000bp excised and purified using a Qiagen QIAquick gel extraction kit, cleaned-up with the same kit but without running on a gel, or purified using Isopropanol precipitation. Isopropanol precipitation was carried out by adding 1/5th volume of 3M sodium acetate followed by 1 volume of 100% isopropanol. This was incubated at room temperature for 1 hour before centrifuging at 14,000 RPM for 20 minutes then washing with 70% ethanol. Following a further centrifugation step, the ethanol was removed and the pellet air dried before resuspending in nuclease free water.

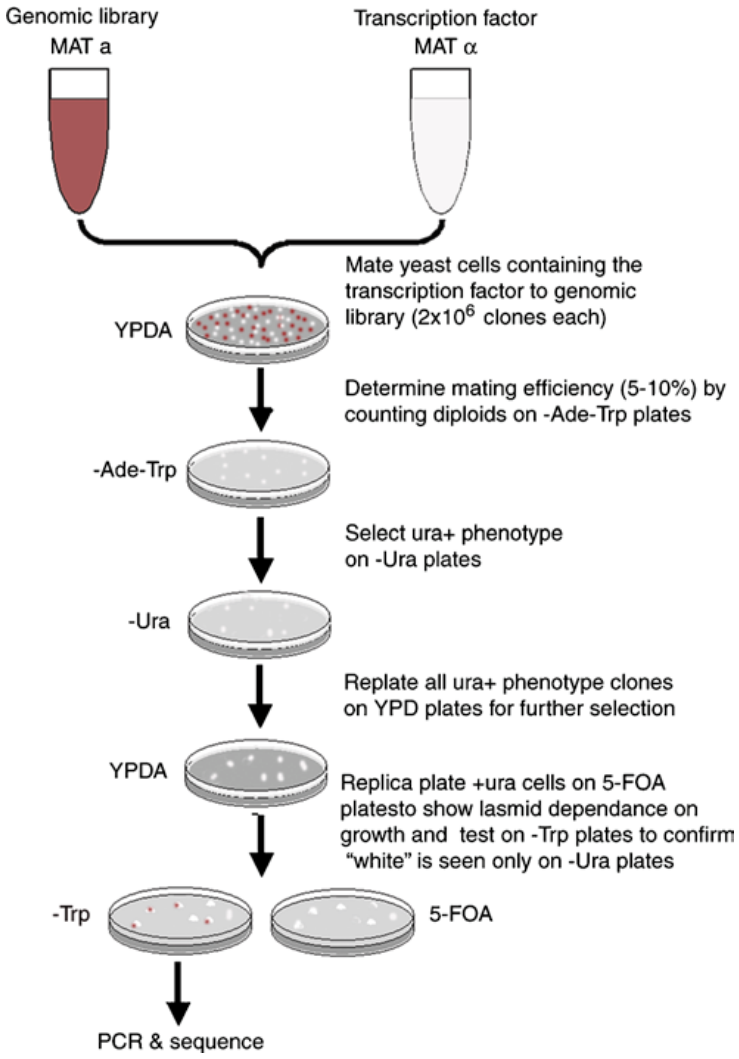


Figure 4.3. Overview of the steps for yeast-1-hybrid (Yan and Burgess, 2012).

Optimising ligation of gDNA fragments into pYOH366. Fifty µg of pYOH366 was digested in a 300µl reaction with EcoRI-HF (NEB) at 37°C overnight. EcoRI was denatured at 65°C for 20 minutes then treated with Alkaline Phosphatase, Calf Intestinal (CIP). Six µl of CIP was added to the reaction and incubated for 30 minutes at 37°C before a further 6µl was added and the reaction incubated for an additional 30 minutes. The digested vector was run on an agarose gel, excised and purified using columns from the Qiagen QIAquick gel extraction kit, using one column for every 10µg. Ligation reactions were carried out with the digested gDNA and vector using molar ratios of vector:insert at 1:1, 1:3 and 1:20 in 10µl reactions with 50ng of vector. All ligation reactions were incubated overnight at 16°C. Five µl of each ligation reaction was transformed into NEB 5-alpha *E.coli* as described. Three different ligases were tested: Promega T4 DNA ligase (1U/ 10µL), NEB T4 DNA ligase (400U/ 10µL) and Thermofisher T4 DNA ligase (0.5 and 2.5U/ 10µL). Different units of ligase activity are used between suppliers, so unit concentrations were based on the manufacturer's protocols. Reactions were trialled with 100ng of vector, and scaled up to volumes of 200µl. Ligation reactions were cleaned up with Qiagen columns and eluted DNA at 50, 100 and 200ng was transformed into *E.coli*. Positive controls with undigested pYOH366 and negative controls with digested and phosphorylated vector were included.

Final gDNA digestion and ligation protocol for pYOH366-g. Fifty µg of genomic DNA was digested with 10U of MluCI (NEB) in a reaction volume of 50µl for 5 minutes. The reaction was stopped with 10µl of 50mM EDTA (pH 8.0). Digested DNA was purified straight from the reaction using 5 Qiagen columns. One µg of each reaction was run on a gel to check sizing - the size range of fragments was between 70 to approximately 2000bp, with the strongest signal around 250 to 300bp. Fifty µg of pYOH366 was digested with EcoRI-HF, treated with CIP and purified as described in the above paragraph. Ligation of the gDNA fragments into pYOH366 was carried out in molar ratio of 1:3 vector:insert. Five µg of vector was used in a 500µl reaction overnight at 16°C. Two reactions were performed – one with the gDNA fragments at an average of 250bp and one with fragments at 300bp. Each reaction was cleaned with a Qiagen column and eluted into 30µl of nuclease free water. Eighteen transformations into NEB 5-alpha *E.coli* were carried at with 200ng of product per transformation. Following recovery in 1ml of SOC, transformations for each reaction were pooled. For each, 3 x 25µl was spread onto 3 LB-agar plates with 100µg/ml ampicillin. The remaining transformation culture was spread onto plates at 200µl per plate. Plates were left to incubate overnight at 37°C. Colonies were counted from the 25µl plates and used to calculate the total number of colonies for each reaction. Based on the average size of the fragments, number of colonies was used to calculate the genome coverage based on a genome size of 80.5 Mbp using the equation:

$$coverage = \frac{fragment\ size\ (bp)\ x\ 2\ x\ number\ of\ clones}{genome\ size\ (bp)}$$

Colonies from all plates were scrapped off, pooled and a plasmid maxiprep carried out according to Yan and Burgess (2012). In addition size of inserted fragments was checked by Phusion PCR using forward primer CAGGAGCTGGTCAAGTTCAG (Tm = 61.8) and reverse primer TTTGTCGGCGGCTATTTCTC (Tm= 62.2) as well as digest of pYOH366-g with EcoRI.

Generating pYOH1-TF

The SITMyb gene including a 3' His-tag, previously described for overexpression in *F. cylindrus* (from +2107 to +4872 of the JGI transcript) and the Myb domain (+3517 to +3732) were cloned into pYOH1, downstream of, and in-frame with the HA-tag (Figure 4.4). Inserts were amplified from pAGM_SITMybOE_IR using Phusion DNA polymerase (NEB) as previously described. Primers for SITMyb incorporated XmaI and XhoI sites at the 3' and 5' ends respectively, whilst primers for the Myb domain incorporated EcoRI and XhoI sites (Table 4.5). One µg of each PCR product was double digested with their respective restriction enzymes for 4 hours at 37°C in 1x cutsmart buffer in a 50µl volume (NEB). At the same time 10µg of pYOH1 vector was double digested with either XmaI/XhoI or EcoRI/XhoI in 50µl. Following digestion of the vector, 6µl of Antarctic phosphatase reaction buffer and 2.5µl of Antarctic phosphatase were added to the reaction and further incubated for 30 minutes at 37°C. The phosphatase was inactivated by heating at 80°C for 2 minutes. All vector and PCR reactions were run on a gel, excised and purified. The purified PCR products and vectors were ligated in 20µl reactions with T4 DNA ligase (NEB) using a 1:3 vector:insert molar ratio and 100ng of vector. The reaction was incubated overnight at 16°C before transforming into NEB 5-alpha competent *E.coli* (NEB) according to manufacturer's instructions. Clones were screened by colony PCR as previously described using either the Myb or SITMyb primer sets from Table 4.5. Plasmids from clones were also screened by restriction digest with XmaI and XhoI for pYOH1-SITMyb and EcoRI and XhoI for pYOH1-Myb. pYOH1-SITMyb was sequenced using primers 5, 6 (Table 4.5) and 16 (Table 4.3). pYOH1-Myb was sequenced with primer 5 (Table 4.5).

Number	Primer	Sequence	Tm (°C)
1	R_YIH_SITMyb F	at***cccggg***ATGAAAACGGCGAATGACAA	60.6
2	R_YIH_SITMYb R	atat***ctcgag***TTAGTGATGATGATGATGATGCAC	60.8
3	R_YIH_Myb F	atatt***gaattc***gaGGAAAATGGACGCCCGAAGA	64.2
4	R_YIH_MYb R	atat***ctcgag***TTAGTAGTCTTGTTTATACTTTTTACATGTC	59.2
5	pYOH1-TF seq F	AACTATCTATTCGATGATGAAGATACC	60.3
6	pYOH1-TF seq R	TGCACGATGCACAGTTGAAG	62.9

Table 4.5. Primers for generating and screening the pYOH1-TF constructs. Primers 1-2 amplify the SITMyb sequence for pYOH1-SITMyb, primers 3-4 amplify the Myb domain for pYOH1-Myb and primers 5-6 were used to screen the constructs.

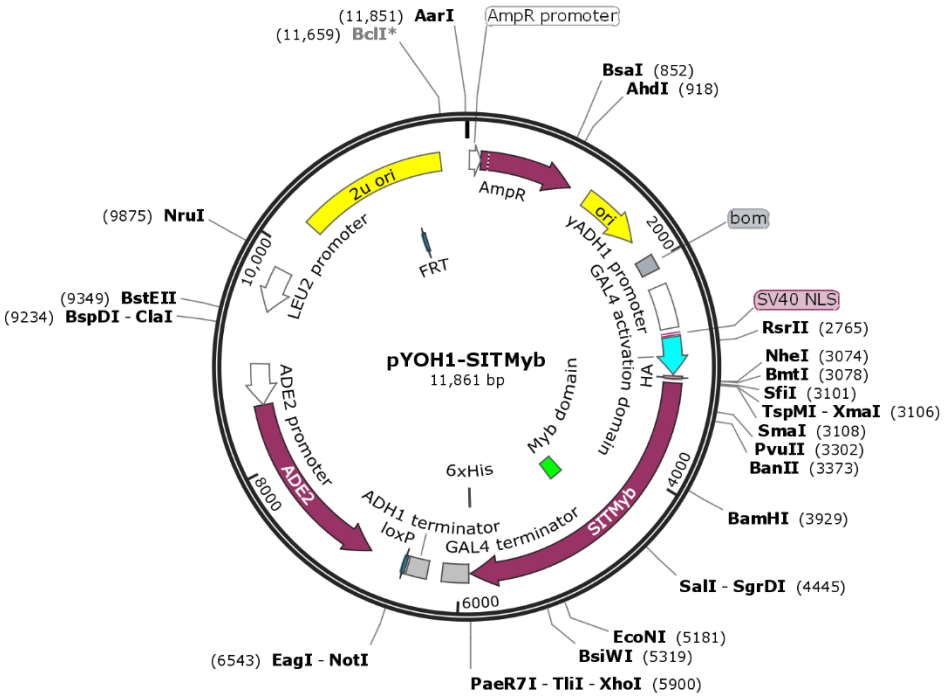


Figure 4.4. Vector map of pYOH1-SITMyb. Created with SnapGene.

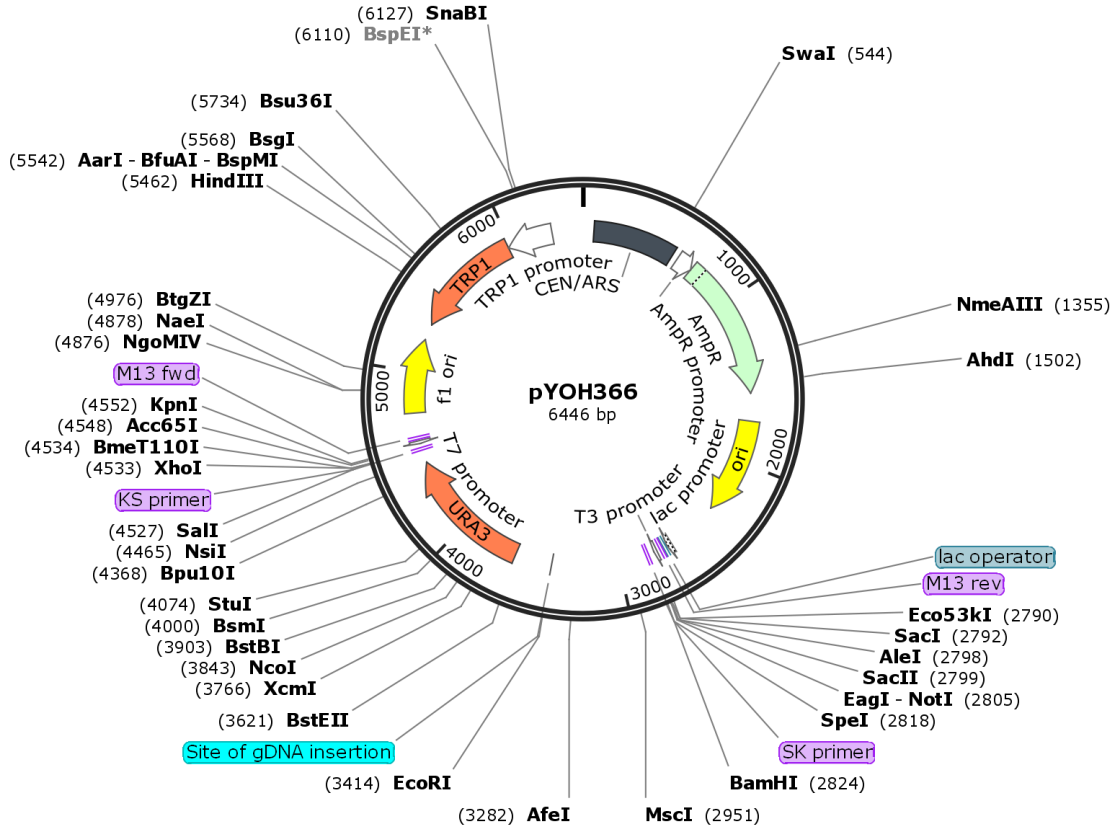


Figure 4.5. Vector map of pYOH366. Created with Snapgene.

Yeast strains and growth conditions

Yeast strains W303 (MATa and MATα) were purchased from GE Dharmacon and cultured according to the Yan and Burgess (2012) protocol.

Materials for YPDA media were sourced from Sigma-Aldrich as were indiviudal amino acids. Yeast nitrogen base, Synthetic dropout media (SD -Ade-Trp-Ura and SD - Ade-His-Trp-Ura) and 5-Fluoro Orotic Acid were sourced from Formedium.

Transforming the gDNA library and the pYOH1-TF constructs into yeast.

Transformation of pYOH366-g into the MATa strain was carried out according to Agatep et al. (1998), protocol 1. Briefly, cells were made competent by treatment with LiAc and transformed using heat shock and 1µg of pYOH366-g per transformation. 19 transformations were carried out. Transformations were then processed according to Yan and Burgess (2012). Transformations were pooled and 5 100µl aliquots at a dilution of 1-100 were spread onto SD -Trp plates. These were used to calculate transformation efficiency. The remaining cells were divided between 10 large SD -Trp plates with 5 FOA, grown and harvested according to the protocol.

pYOH1-SITMyb and pYOH1-Myb were transformed into W303 MATα according to the protocol with the exception of pTpPuc3 being used to deliver the HIS3 cassette instead of pRS313-HIS3. This produced MAT α-pYOH1 SITMyb and MAT α-pYOH1 Myb. Briefly yeast cells in log phase were made competent by treatment with LiAc, and 0.1µg of each pYOH1-TF construct and pTpPuc3 were co-transformed into cells by heat-shock. Positive clones were selected on SD-His-Ade plates.

Mating of the MATa and MATα yeast strains and screening of the transcription factor binding sites was carried out according to the protocol. Due to a poor mating efficiency for Myb strains, only SITMyb was carried forward for screening. Colony PCR was performed on clones which were white on -Ura plates, sectioning on -Trp plates and pink or showing no growth on 5-FOA plates. Several other colonies which showed the correct colour on two plates but not the third were also screened. Colony PCR was carried out by touching a toothpick to a colony and resuspending in 20µl of 0.02M NaOH and heating at 99°C for 10 minutes. One µl of the treated sample was used in PCR reaction with Go-Taq polymerase according to the manufacturer's protocol with the addition of betaine to a final concentration of 1M. Forward primer pYOH366_F (CAGGAGCTGGCTAAGTTCAG) and reverse primer pYOH366_R (TTTGTCGGCGGCTATTTCTC) were used to amplify the insert with an annealing temperature of 53°C and an extension time of 3 minutes. Forty-eight colonies in total were screened. Following PCR, products were sequenced using P3 described in the protocol.

Checking for expression of the transcription factor/ transcription factor domain in yeast.

Proteins were extracted using a mixture of methods from the Clonetech xTractor Buffer manual and the Clonetech Yeast protocols handbook. Six MAT α-pYOH1 SITMyb and 2 MAT α-pYOH1 Myb cultures were grown to log phase overnight and 12ml harvested by centrifugation at 700 x g

for 5 min at 4°C. Cells were washed with water and re-pelleted. Cells were transferred to a 1.5-ml screw-cap microcentrifuge tube and 400µl of xTractor buffer was added. 320µl of glass beads were then added and the sample vortexed in a Mini-Beadbeater (BioSpec) for 1 minute. Debris and beads were pelleted by centrifugation at 14000 RPM for 5 minutes at 4°C and the supernatant transferred to a new tube. Proteins were stored at -80°C.

His-tag purification: Crude protein extracts were run through His-tag purification columns to enrich proteins with a His-tag. A Capturem His-tagged Purification Miniprep Kit (Clonetech) was used. Briefly, 400µl of Xtractor buffer was used to equilibrate the spin column before loading 400µl of cleared protein lysate and spinning at 11000 x g for 1 minute at room temperature. Columns were washed with wash buffer under the same conditions and protein eluted into 100µl of Elution buffer.

Western blots: Both crude protein extracts and His-tag purified extracts were run on NuPAGE 4-12% BIS-Tris gels using the X-cell Surelock Mini-gel electrophoresis system (Invitrogen). Protein extracts were run for 35 minutes at 200V alongside 5µl of Broad Range, Color Prestained Protein Standard ladder (NEB). For westerns using a His-tag antibody, 1-5µl of BenchMark His-tagged Protein Standard Ladder was also included. For crude extracts 13µl at 2mg/ml were loaded into wells. For His purified proteins 13ul at 40-60µg/ml were loaded. Proteins were transferred onto a nitrocellulose membrane according to the X-cell Surelock manual for 1 hour at 30V. Membranes were blocked for 1 hour in 5% BSA PBST (1 x PBS with 0.05% Tween 20). Crude lysate from both pYOH1-SITMyb and pYOH1-Myb cultures were incubated with HA-Tag (C29F4) Rabbit mAb (Cell Signalling), overnight at 4°C. A 1:1000 dilution was used in PBST with 5% BSA. Crude lysate and His-tag purified proteins from pYOH1-SITMyb cultures were incubated with a HIS-antibody as described earlier for SITMYb overexpression in *F. cylindrus*. Membranes were washed 3 x in PBST for 10 minutes each on a rocker. An anti Rabbit IgG HRP tagged secondary antibody (Promega) at a dilution of 1:2500 in PBST was added to membranes for 1 hour at room temperature, and washed 3 x in PBST for 10 minutes each. Blots were visualised by adding ECL Western blotting substrate (Pierce) for 2 minutes and chemiluminscence imaged in 30 second intervals for up to 5 minutes.

Results and Discussion

Sequencing of the *F. cylindrus* genome revealed a large gene with both a SIT and a Myb domain, previously unseen in other sequenced diatoms. Myb transcription factors are some of the most prevalent in the Stramenopiles (Rayko et al., 2010) whilst silicon transporters play an important role in silicon acquisition (Hildebrand et al., 1997; Thamatrakoln and Hildebrand, 2008). Furthermore SITs are the first proteins in diatoms shown to bind silica (Hildebrand et al., 1997) and may have a role in sensing silicic acid and regulation of cell cycle progression (Shrestha and Hildebrand, 2015). If SITMyb is a transcription factor that is able to either bind silicic acid or regulatory protein involved in silicon sensing, presence or absence of the substrate may alter its

activity. Given the domains present in the SITMyb gene, it is possible that it regulates genes involved in silica metabolism.

In order to investigate this hypothesis, several different methods have been performed. Initially the gene model was investigated, through both in-silico and in-vitro methods. This was followed by cloning the gene into *F. cylindrus* under the control of a highly expressing FCP promoter to observe differences in phenotype upon overexpression. Finally, inverse yeast-1-hybrid was carried out to determine potential binding sites. The above methods need further work for more conclusive results, however, preliminary data and development of methods should be useful for further investigation of this gene, and potentially other *F. cylindrus* transcription factors.

Modelling the SITMyb gene

Two alleles of the SITMyb gene are present (ID 233781 and 250586) in the *F. cylindrus* genome. The later has an incomplete sequence with a section missing in the middle of the gene, just after the start of the first intron. As a result, this chapter largely concentrates on allele 233781 (Figure 4.6). Alignment of alleles (Appendix Figure 1) showed high conservation after the 1st intron (99% identity from 2832-5288bp, based on 233781), with an increase in divergence at the beginning of the gene (87% identity from 1 to 1881 bp). SIT and Myb domains are found in the highly conserved region and show 100% identity between alleles.

Blastn searches gave hits with low coverage of the gene at 5%. Along with the Myb domain, which showed homology to Myb domains in higher plants, several hits from 3 short regions were found. Region 720-821 bp showed homology (78% identity) to a hypothetical protein from *Salpingoeca rosetta*, a silicifying choanoflagellate, as well as homology to a hypothetical protein from *Plasmodium sp*. The sequence from 2279-2329 had a 90% identity to a putative Na+/H+ exchanger in *Eimeria acervulina*, a parasitic apicomplexan. This may have implications for the function of the SIT domain, as silicon transport in diatoms is sodium dependent (Curnow et al., 2012; Hildebrand et al., 1997). Apicomplexans such as Eimeria or Plasmodium are part of the alveolates. Algae have complex evolutionary origins involving multiple endosymbiotic events. One theory for diatom evolution is that several groups of algae including diatoms, cryptophytes, coccolithophores and dinoflagellates are part of larger group called the chromalveolates, with an ancestral alveolate host involved in a secondary endosymbiotic event which later diverged into these main groups (Cavalier-Smith, 1999). It is therefore possible that part of the SITMyb gene may have origins from the heterotrophic host.

Finally, the region from 2484-2567 shows homology (84-86%) to several different genes from *Eimeria sp*, including a putative ATP synthase, a ZIP zinc transporter and a DEAD/DEAH box helicase. This sequence also shows an alignment (79%) to a putative CCAAT transcription factor in *Plasmodium falciparum*. The presence of a zinc transporter domain is interesting as it has been suggested that silicic acid may be bound via a zinc atom (Grachev et al., 2005; Sherbakova et al., 2005), whilst homology to DNA binding proteins such as helicases and transcription factors may

be linked to transcriptional activity of the SITMyb gene. MXD motifs, where X = L or I, found in several silicon transporters, have been proposed to bind zinc which in turn may bind silicic acid (Grachev et al., 2005, 2008; Sherbakova et al., 2005), however no MXD motifs are present in the 233781 gene model (Figure 4.6) or in the sequenced part of the 250586 allele. GXQ motifs (X= Q,G,R or M) have also been associated with binding of silicic acid in silicon transporters (Thamatrakoln et al., 2006) but based on ClutalX alignment, are not typically found in the domain associated with the SITMyb gene. However, a GXQ motif can be seen downstream of the Myb domain (Figure 4.6) and a GKQ motif found within the Myb domain may also be worth further investigation (Paul Curnow – personal communication). Another interesting motif in the SITMyb gene is the stretch of proline residues found at the beginning of the gene (Figure 4.6). Stretches of amino acid repeats can be associated with transcription factors (Rado-Trilla et al., 2015) and Gerber et al. (1994) found that fusing runs of proline or glutamine to the binding domain of a GAL4 factor led to activation of transcription. For proline, in cell transfection assays, 10 residues gave the highest activity. A slightly shorter repeat of seven residues can be seen in the SITMyb gene.

MPRVAKNFDKYLLTKEVPDSPNGTGITKKNIEHIRVQWDYELEDTTRVIYHRFRNQTEYEKYHNNKVSQGVRKRNGLYETGQKLRK
KKKKEDDDENMTTDGTTTTTTTKIGSTVAAVAVAVAEKPSSTSSLLSLKKKASALASAAVGTTTMTDSNDNAGVSTYEDGIDAGLP
STTSYHCQPIVGTRLVITPSNIPPPPSPPLQaKEHDTMEATTTGSPPPPPPPLGKTIIQTTHKDGNDDEVVVVEEPAKTASPSALSTST
TTAATTAATTATATSLIIDLTIDDDPDNDVVGGSLVPARAVESKGKPKMKRAKRQMLPLNRRKMQDEEADNNNDVVGGSRASSK
LLRMEMIMKLPTAEELKAESIVINDVNDHDILMGSRKCNKHPGNKVYRDIVRKYQPLLEKETIGDRAIVVSMVIDHIHDQIGGRFLK
VNSANQWHVVPRLDVITKITKALVELGNPAIFRLALPKSVPTSTMLESIGESIGQSIVGQDNKRPKRKCTLPKVIEEQVDDEPMEKKF
TKTTTMRERSSRIEDEKITKIDDDDAESNKNDDNNDDHHHLRMVATVVEEEEENGFSSIPTTAAAAVMTNANSTADADSIIPTSW
KDWVSRVNNLVESSPSCMDYSPASERYKILLDYIPMEDIVSQIRYKHMMYQIFISDRPGDRECIHAFPNQISFNSIEKKYNKQISNGK
SVKKKRSQLQMKTANDKYRSKKKRKTDASDAFRTTVGSIQEEQAARATTITTGIALEEKNNEYLELFTAAAAAADTTTATATATATK
EVACEEIKRALSNAGIAAVAKQNDTNASVITQTQKEYTRLPPHRTSAARNNNNNDNNNNNNDNNSTKHQQQEQQKSMTASTR
GQDEETNATPVSNQYPIAIAATVDTTSQLKPQTKPLSHMQGAYSSRREKVMANIKELRSQISQATFDEEKIAFEQAFKLEIESLGRLN
KDEMKSKLLFEGDKIDVIEEAELVNGSNASNPTTNNVNLLYPPYNQFGMEMMANDFGCGGSANIGVIGGISHGFGASGNLGAPY
SASFYAQQQMYQYSVVHPHQIVHQHPYFQHHHHKHQAIAVTNTDRPDDPFIMMDPAIPDGSSDKNKREQNENENRGGCTAVT
ADIATHSLPLSEEHTVAQQDIAQDDSIISDAVDGKNTTGSDGNADANFNANTKPTATTKGKWTPEEHEEVAKAMAKYGPRVSGK
QISIEFVKGRTPLQLNSYINRKKSELLATCKKYKQDYCDESEDDDDGGTIRGLKFHQTRSCDRDGDDKKTTNTDVEKNVQYRNAERE
LRRTKDGCLLPKGGKEKYLKDDGTYRRPDGARPFGLSWHKIRGLWVPSERLEDNDENNYDDNINKSGYTNYSSGDTIAKATSSNCE
QSYDRSALPRGLKTHIRDPVGGCYWTPLGSRRKLTAKEASRKSKKRKPGRQSAGAKTKEKETRALAYVTPLEILQGKKPTPSNLFLSL
HSVIPEAAMNENFKDDDDDDDDDEGYESWTSGSWCLLQAQRDASASAVAESEAKKSAPDEEEQQVRKCKSIAAEAKAKERERIG
ATCTGNQKDESANCGDDSSQDSSKRRRLSICEQSRIINPVQKINVQRKKKKSFLKAKQDARDYMLAKYGQGNEEEEIVMV

Figure 4.6. SITMyb gene model of 233781. The silicon transporter domain (SIT) is highlighted in green and the Myb domain in blue. A domain with homology to Myb DNA-bind 6 is underlined. Potential nuclear locailsation signals are shown highlighted in yellow and GXQ motifs are shown in red. Predicted coiled-coil motifs are double underlined.The proline repeat is shown in purple.

To access and act on the genome, transcription factors need to be transported to the nucleus. Several mono and bi-partite nuclear localisation signals (NLS) were found in the SITMyb gene suggesting that it may be localised to this organelle.

Blastp searches revealed hits to SIT and Myb/SANT domains (Figure 4.6). Hits for Myb placed the domain at 45-66 amino acid residues. Myb transcription factors can contain, 3, 2 or 1 Myb domains of around 50 amino acids. Plants often have transcription factors with a single Myb domain and diatoms *P. tricornutum* and *T. pseudonana* show genes with 1-3 Myb domains (Rayko et al., 2010). Many of the hits against the Myb domain were from hypothetical proteins and transcription factors (TFs). The latter included Myb TFs from cryptophytes, chlorophytes and several higher plants.

The SIT domain showed homology to diatom silicon transporters from *F. cylindrus*, *Cylindrotheca fusiformis* and *Nitschia alba*. The protein sequence of the SIT domain was aligned to several diatom silicon transporters. This showed alignment towards the C-terminus of the diatoms SITs after the 10th transmembrane domain (Figure 4.7).

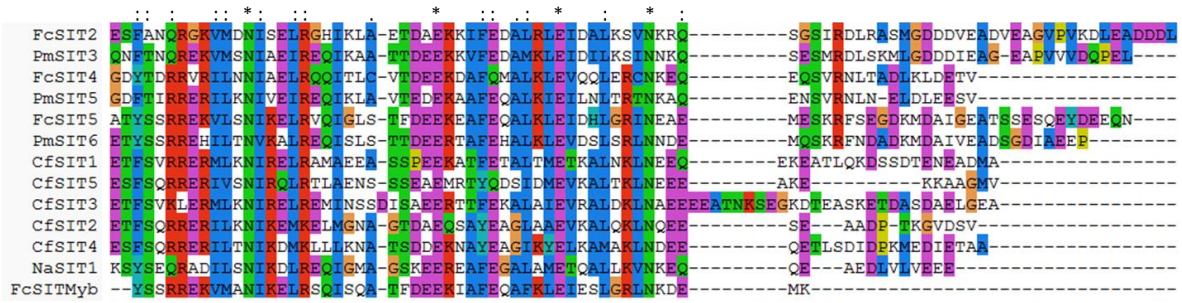


Figure 4.7. ClustalX alignment of the *F. cylindrus* SITMyb SIT domain and C-terminal regions of closely aligned diatom SITs. Default ClustalX colours used (http://www.jalview.org/help/html/colourSchemes/clustal.html). * indicates a single conserved residue, : indicates a fully conserved 'strong group (gonnet PAM250 matrix score>0.5)' and . a fully conserved weak group (score<0.5).

The phylogenetic tree produced for this region (Figure 4.8), shows that the SITMyb domain clusters with SIT5 from *F. cylindrus* and SIT6 from *Pseudonitzchia multiseries*. In addition, this region contains a coiled-coil domain, also found in *Thalasiosira pseudonana* and *Phaodactylum tricornutum* SITs (Shrestha et al., 2012).

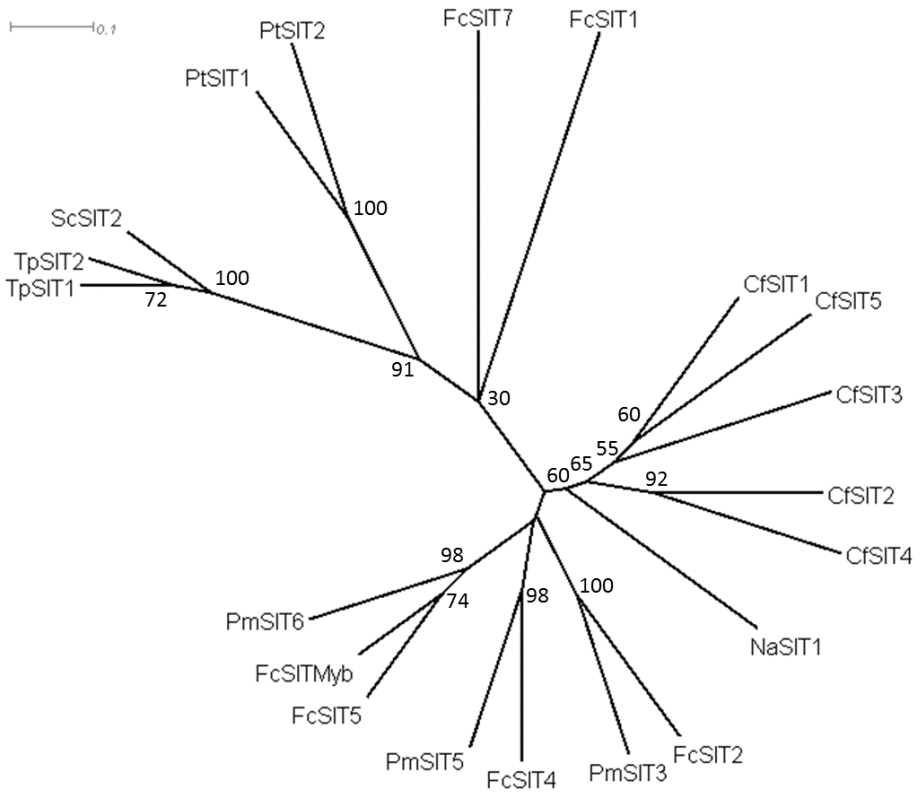


Figure 4.8. Neighbour joining tree of C-terminal SIT regions. Cf; *Cylindrotheca fusiformis*, Fc; *Fragilariopsis cylindrus*, Na; *Nitzchia alba*, Pt; *Pheodactylum tricornutum*, Sc; *Skeletonema costatum*, Tp; *Thalasiosira pseudonana*.

A total of 6 coiled-coil motifs can be found in the SITMyb gene (Figure 4.6). Coiled-coil domains which are typically intracellular (Thamatrakoln et al., 2006) are associated with transcription factors (Mason and Arndt, 2004) and protein-protein interactions (Mier et al., 2017), both of which can be linked to transcriptional regulation. In *Arabidopsis thalinania* and *Chlamydomonas reinhardtii* there is a Myb TF with both a single Myb domain and a coiled-coil. This TF is involved in phosphate regulation and is hypothesised to be part of a larger regulatory network, with signal-transduction occurring upstream, and activity determined either by post-translational modification or presence of a co-regulator which responds to phosphate starvation (Rubio et al., 2001). Currently only silicon transporters have been shown to interact with silicic acid (Hildebrand et al., 1997; Shrestha and Hildebrand, 2015) and Shrestha and Hildebrand (2015) point out that some transporters are known to act as sensors. Given the presence of a silicic acid binding motif and the SIT domain there is possibility that SITMyb may be directly interacting with silicic acid. Alternatively, the coiled-coiled domain in the SIT region may form protein-protein interactions with intermediates, possibly involved in sensing silicon.

The SITMyb protein was modelled with both Phyre 2 and Swiss Model. Both gave strong models for the helix-turn-helix of the Myb domain. The majority of hits (top 19/20) modelled with Phyre 2, aligned to Myb domains of transcription factors and DNA/RNA binding domains. For the Myb

domain highlighted in blue (Figure 4.6) 89% of residues were modelled with 96% confidence. For the underlined Myb domain 75% of residues were modelled with a confidence of 90% (Figure 4.9).

As with Phyre2, Swiss model showed hits which corresponded to Myb proteins, C-Myb and DNA binding with a high coverage (>90%). Similar models to Phyre 2 were created. The majority of the remaining protein was unmodelled due to low identity. Modelling of the SIT domain independently with Phyre 2 led to 31% of residues modelled with a 39.5% confidence. Only three hits over 26% confidence were returned. These were linked to hydrolase inhibition or protein binding. The highest identity hit for Swiss model was for a DNA binding protein at 30.77%.

These results suggest that there's a high chance the SITMyb gene contains a Myb domain, and is therefore likely linked to transcription. The majority of the protein could not be modelled to a high confidence, which is unsurprising given that genome, transcriptome and proteome analysis of diatoms often show a high number of unknown genes (Armbrust et al., 2004; Frigeri et al., 2006; Mock et al., 2017, 2008; Shrestha et al., 2012), and models rely on structures of previously determined proteins.

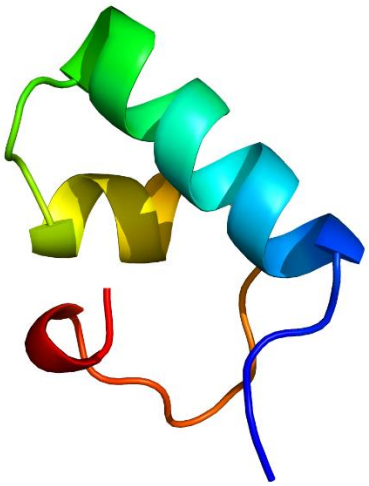


Figure 4.9. Phyre 2 protein model of the Myb domain: GKWTPEEHEEVAKAMAKYGP RVSGKQISIEFVKGRTPLQLNSYIN. Blue represents the N terminus and red the C terminus. Modelling shows the domain follows a helix-turn-helix structure with homology to Myb. 89% of residues are modelled with a 96% confidence.

Regulation of SITMyb alleles

Only two conditions in each allele showed strong up or down regulation against the control, although these conditions are different for each allele; cold and high CO2 for 233781, and Iron limited and prolonged darkness for 250586, suggesting that SITMyb alleles are differentially regulated (Figure 4.6). Both alleles show low expression under control conditions (233781 – rank 12366 and 250586 – rank 19147). Low expression of Myb genes can also be seen under a variety of different conditions for *P. tricornutum* (Rayko et al., 2010).

Allele	Fe vs Ctrl	Dark vs Ctrl	Cold vs Ctrl	CO2 vs Ctrl	Blue vs Ctrl	Si vs Ctrl	Red vs Ctrl	Heat vs Ctrl
233781	0.41	0.24	-1.53	-2.18	-0.26	-0.78	0.78	-0.60
250586	1.02	2.87	0.43	-0.13	-0.10	-0.60	0.97	-0.70

Table 4.6. LogFC values from *F. cylindrus* RNA seq data produced by Strauss (2012) under different growth conditions for SITMyb alleles 233781 and 250586. LogFC values above 1 or below -1 are shown in red. All LogFC values above 0.77 or below -0.77 have p values below 0.01.

RACE and amplification of the transcript

Rapid amplification of cDNA ends (RACE) was performed for both the 3' and 5' end of each SITMyb allele. RNA-ligase mediated (RLM) RACE was used to assess the 3' end and both RLM-RACE and template switching oligo (TSO) RACE were used for the 5' end.

Transcript ends appear to be variable depending on the RACE experiment. Start and end position vary depending on the experiment replicate, primer set and method. Products from each RACE PCR have multiple bands and often show smears (Figure 4.10, lanes 5 and 6; Figure 4.11). Additionally, non-specific products can be seen, as evidenced by products produced by PCR with a reverse primer that sits downstream of the RT primer (Figure 4.11, lane 3). The JGI gene model puts the size of the transcript for allele 233781 at 4875bp and allele 250586 at 5004bp. There is a gap in genome coverage towards the 5' end for 250586, however, so the latter may not be accurate. 5' RACE products put the start of transcription between -219 to +2910, whilst 3' RACE products put the end of the transcript between +3279 to +5054. If RNA or cDNA is damaged or truncated before the adaptors have been added this can lead to shorter products which tend to be favoured during PCR. Amplification of the full transcript as a single PCR product either from cDNA or gDNA was unsuccessful, despite successfully testing all primers by amplifying shorter fragments. This may be due to the length of the SITMyb gene or secondary structure. Taking into account the length of the RT product and outer PCR the longest products covered 3569bp from the 5' end and 3359 from the 3' end of the JGI model. Internal, overlapping, fragments of the SITMyb gene amplified from cDNA produced during 5' and 3' RLM RACE show clear bands at the correct size. Some multiple banding can be seen, however, the strongest band corresponds to the expected size. Internal fragments cover positions +1296 to +4023 for allele 233781 and +1479 to 4168 for allele 250586. The full transcript appears to be covered, with the banding of some RACE products extending into the 5' and 3' UTRs. However several products finish within the gene model and results are inconsistent. This may be due to non-specificity, truncated products, splice variants or differences in the final transcript and JGI gene model.

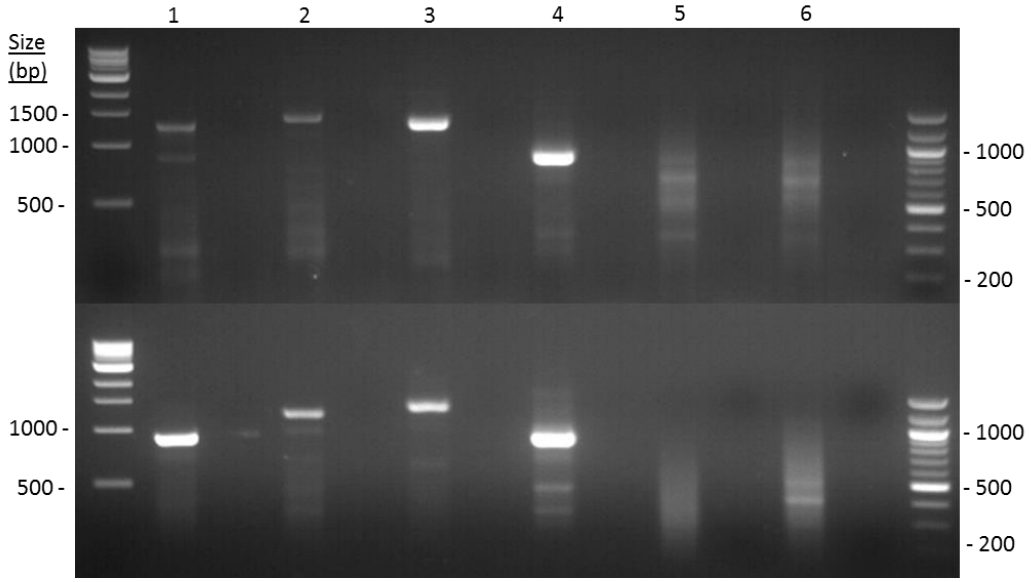


Figure 4.10. PCR products from RLM-RACE and internal fragments of the SITMyb gene amplified from cDNA. Products from allele 1 (233781) are shown in the upper panel and products from allele 2 (250586) in the lower panel. Lanes adjacent to numbered lanes are the equivalent negative no target control. Lanes 1-4) Internal fragments 1-4, 5) 3' RACE, 6) 5' RACE. See Table 4.2 for more details on fragments.

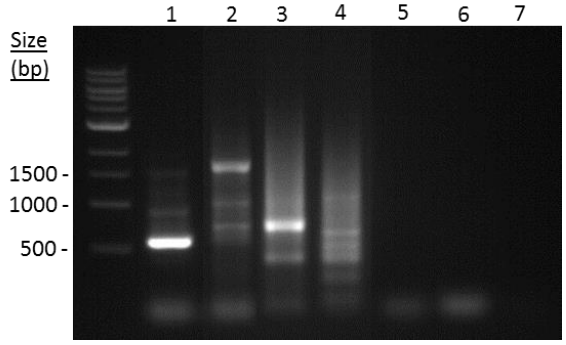


Figure 4.11. PCR products from TSO RACE. 1) RT with 5'3 and PCR with SIT1, 2) RT with R02 and PCR with SIT1, 3) RT with 5'3 and PCR with SIT2, 4) RT with R02 and PCR with SIT2.

To further examine the transcript in comparison to the JGI gene model, RNA-seq coverage was visualised using IGV (Figure 4.12). The 3' end is highly covered in both alleles across all conditions. As the sequence progresses to the 5' end coverage is reduced. This may be due to unequal expression across the gene, possibly from a splice variant towards the 3' end. Alternatively this may be an artefact from sequencing. Coverage bias towards the 3' end is often seen when cDNA is generated from the polyA tail (Nagalakshmi et al., 2008), however, in this case first strand synthesis was performed on total RNA using random hexamers (Mock et al., 2017). It has been

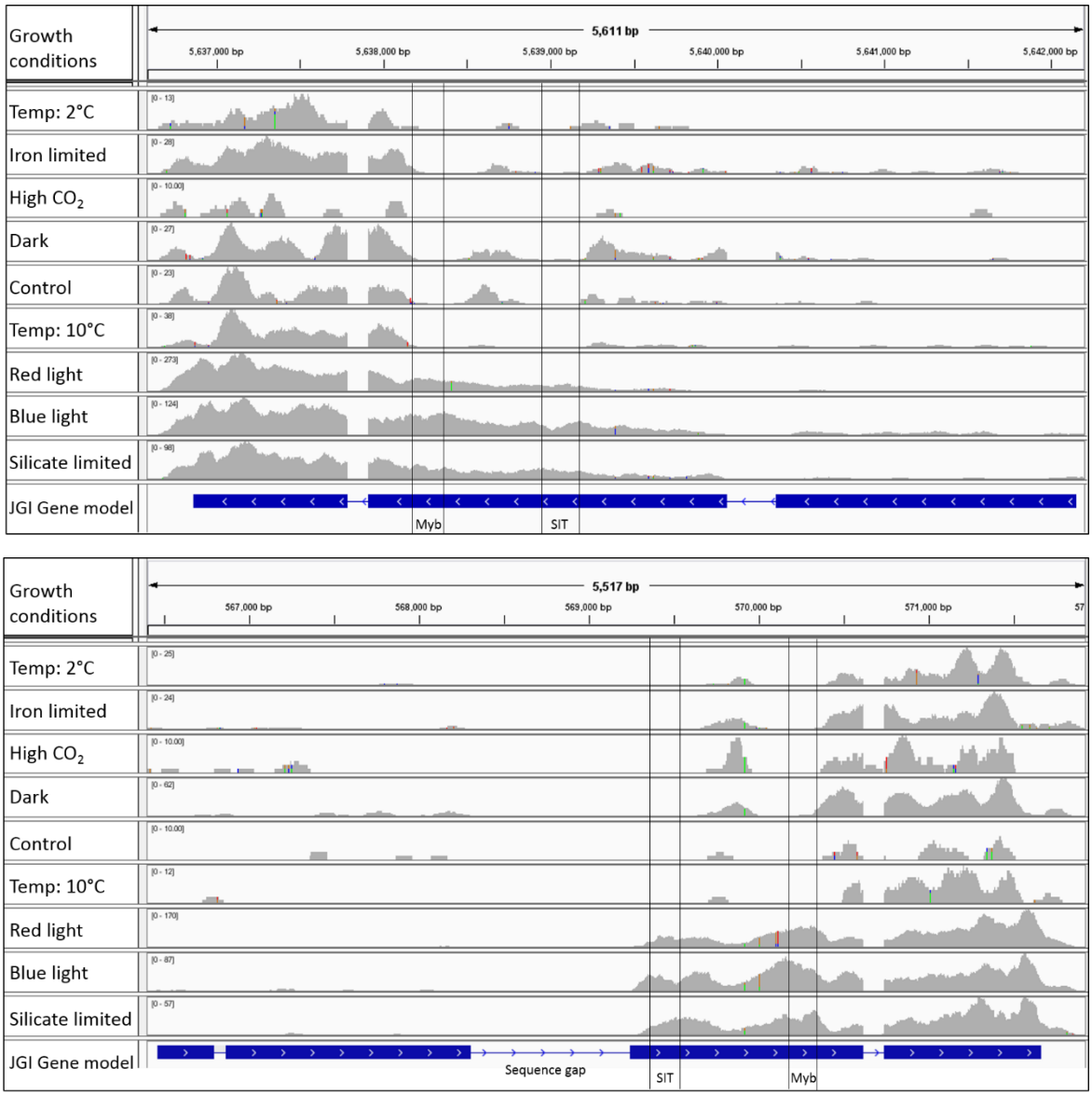


Figure 4.12. *F. cylindrus* RNA-seq data produced by Jan Strauss under multiple conditions visualised in IGV. The upper panel is allele 233781 and the lower panel is allele 250586. Due to its position on the negative strand 233781 is reversed. JGI gene model track: exons are shown by the blue bar and introns by the blue line. Arrows indicate direction of the gene. The gap in the genome sequence for allele 250586 is indicated. Red light, blue light and silicate limited data were produced separately to the other conditions.

found that higher GC content may increase coverage when using random hexamers (Dohm et al., 2008; Zheng et al., 2011), which may help to explain the differences in coverage across the gene, given that GC content in the sequence after the Myb domain is 5% higher than before it.

RNA-seq data (Strauss, 2012) and sequencing of internal fragments confirm the sequence compared to the JGI transcript, including splicing of the introns. However due to variability and quality of the 5' RACE results, as well as low RNA-seq coverage towards the 5' end, the start of the gene is poorly characterised.

Building a SITMyb overexpression construct and transformation into *F.cylindrus*.

Because RACE data is fairly inconclusive, the 5' end is poorly characterised and there were difficulties amplifying the full 5kb gene either from DNA or cDNA, it was decided to clone a shorter sequence than the predicted JGI model. The gene was amplified from an in-frame ATG of allele 233781 after the first intron, from a point which displayed clear expression in the RNA seq data (from +2398 to the end of the predicted gene +5536). This sequence includes the highly conserved region between alleles and both the SIT and Myb domains. SITMyb was amplified from gDNA rather than cDNA due to uncertainty with the transcripts following RACE. Also as an endogenous gene it is expected to be correctly spliced.

Even with a truncated gene, PCR was fairly inefficient when amplified in one segment. As a result the sequence was amplified in three parts and assembled with the FCP promoter and terminator into puc19 using Gibson assembly. The pucFCP:SITMyb construct was then used as the template for site directed mutagenesis to remove BsaI and BpiI sites, before cloning the cassette into a level 1 Golden-gate vector. A BpiI site within the intron was removed both by inducing a point mutation and by removing the entire intron. Domesticated SITmyb cassettes were then combined with FCP:shble into level 2 backbones to create the overexpression constructs. A CEN-ARS-HIS module was also included, however no follow through work, as yet, has been carried out in relation to this sequence. Both the variants with and without the intron were cloned, with the final overexpression (OE) constructs designated pAGM_SITMybOE and pAGM_SITMybOE_IR, respectively. Constructs were transformed into *F. cylindrus* as described in the transformation chapter.

Screening F. cylindrus clones with SITMyb OE construct

Of the 9 colonies that were picked, 5 of them were successfully transferred from plates and grown in liquid selective media. Four of these were screened; 1 from SITMybOE and 3 from SITMybOE_IR. The fifth appeared on plates after initial screening. Screening was carried out by PCR to check presence in the genomic DNA, by reverse transcription PCR to check presence of the transcript and by western blotting to check for the protein. PCRs on genomic DNA and cDNA were carried out with a reverse primer targeted to the chimeric His-tag in order to enrich for the overexpressed gene rather than the WT SITMyb.

Figure 4.13 shows results from PCR of the genomic DNA using a forward primer at the end of the promoter to amplify the full OE SITMyb (a) and a shorter fragment from within the coding region (b). Both colony 1 from SITMybOE and colony 2 from SITMybOE_IR show the correct band in both PCR reactions. Additional bands can also be seen across samples, however some of the lower bands also occur in the WT, suggesting that the primers may also bind non-specifically to a sequence in the WT genomic DNA. One other colony, SITMybOE_IR 6, also shows the correct band for the shorter PCR, however a smaller product than expected is seen for the full length SITMyb, suggesting that part of the earlier sequence may be missing.

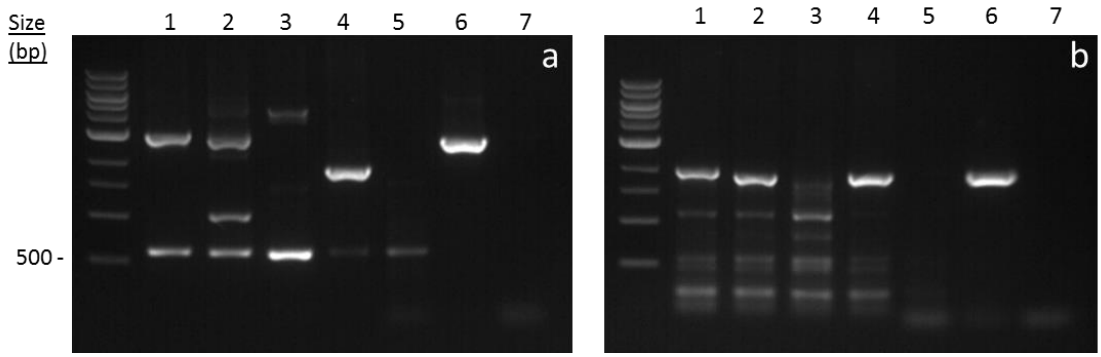


Figure 4.13. Screening for the overexpression cassette in *F. cylindrus* clones via PCR of gDNA. a) Used forward primer prom_SIT1 F, starting amplification within the promoter. b) Used forward primer SIT1_SIT2 F, amplifying the product from within the coding region. Both PCR reactions used reverse primer SITMyb OE R which anneals to the His-tag, specific to the OE cassette, at the 5' end of the coding region. Lanes: 1) SITMyb_OE 1, 2) SITMybIR_OE_2, 3) SITMybIR_OE 4, 4) SITMybIR_OE 6, 5) Wildtype, 6) Positive control (pAGM_SITMybOE_IR), 7) Negative control.

Reverse-transcription (RT) PCR of a 2000bp fragment was then carried out to check presence of the transcript in SITMybOE 1 and SITMybOE_IR2 (Figure 4.14). Both colonies show a band at the correct size as well as two smaller fainter bands. The two lower bands can also be seen in the WT (not shown) suggesting that the primers may bind non-specifically to the WT gDNA. RT, DNA and no target controls were clean (not shown), suggesting that the product was amplified from RNA. Sequencing of the higher bands was concurrent with the expected transcript.

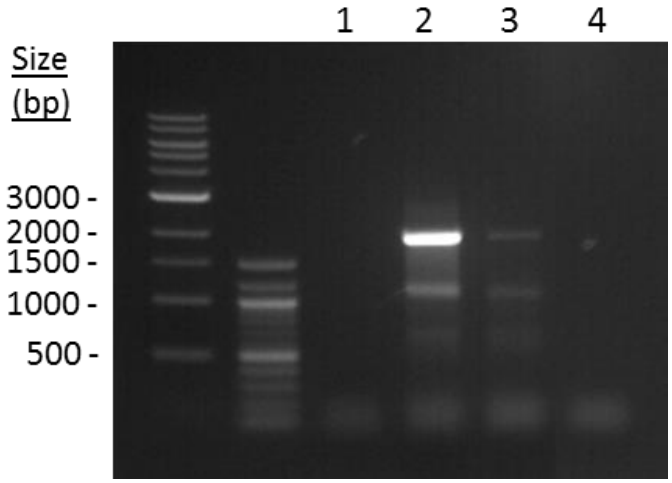


Figure 4.14. RT-PCR of overexpressed SITMyb. a) PCR of cDNA from overexpression cell lines. The forward primer lies within the SITMyb coding region and the reverse primer anneals to the His-tag sequence at the 5' end for amplification of transcript produced from the overexpression construct. 1) Negative RT control. 2) SITMyb_OE 1. 3) SITMybIR_OE_2. 4) Negative no target control.

Along with the genomic DNA PCR, this suggests that the overexpression cassette has been successfully introduced and expressed in the overexpression strains. However, no protein at the correct size, could be seen in the westerns blots either from crude lysate or from His-tagged purified protein (Figure 4.15). Both the ladder and the positive controls were clear suggesting that labelling was successful. In addition a very faint band just above 50kDa could be seen for all His-tag purified samples processed with Xtractor buffer, including the WT, which suggests that a protein is present in the WT which interacts with the His-tag antibody. Protein concentration following His-tag purification was similar between OE cell lines and the WT suggesting that either low concentrations of His-tagged proteins were present in the OE cell lines or carryover of non-tagged proteins occurred.

At this point, given that there was no evidence of overexpression of the SITmyb protein, efforts were concentrated on the yeast-1-Hybrid method. It is worth mentioning however, that as seen later in this chapter, expression of the same SITMyb was achieved in yeast. The SITMyb expressed in yeast, had a HA-tag at the N-terminus of the gene in addition to the His-tag at the C-terminus. No protein could be seen when probing blots with the His antibody, however a clear correctly sized protein was revealed when probing with a HA antibody (Figure 4.18).

It appears that the His-tag in the SITMyb gene may not be functional or accessible for probing. If denaturing conditions are required to expose the His-tag, this is unlikely to be seen when using the denaturing buffer as His-tag purification columns did not appear to be compatible with the buffer given the low yield returned compared to Xtractor buffer. Denaturing conditions are applied during westerns, however if the His-tag is not accessible under standard conditions, then purification may be removing the target protein along with other non-tagged proteins. No His-tagged proteins from samples could be seen in the blot with crude lysate which indicates that either no protein is present, the His-tag is not functional or protein is present in low concentrations and needs to be enriched.

Given the lack of signal when blotting with a His antibody, it's not possible at this point to say whether or not the SITMyb protein is overexpressed in the *F. cylindrus* OE cell-lines. Initially the His-tag approach was chosen as SITMyb is a large gene and there were concerns about maintaining functionality and handling a large gene with a fused protein. In addition His-tagged proteins have previously been expressed and purified from transformed diatoms (Apt et al., 2002; Joshi-Deo et al., 2010)

In-vivo His targeted labelling approaches such as Ni^{2+}-nitrilotriacetate (Ni-NTA) probes (Lai et al., 2015) were also considered for visualising the overexpressed SITMyb, however given the current results it may be worth creating a SITMyb:egfp fusion. If there are issues with size and functionality, it may be possible to include a cleavage domain at the fusion junction (Wang et al., 2015), to separate the protein post-translation. It may also be worth using a HA-tag at the N-terminus, as seen in the yeast lines, to determine presence or absence of protein before further phenotyping.

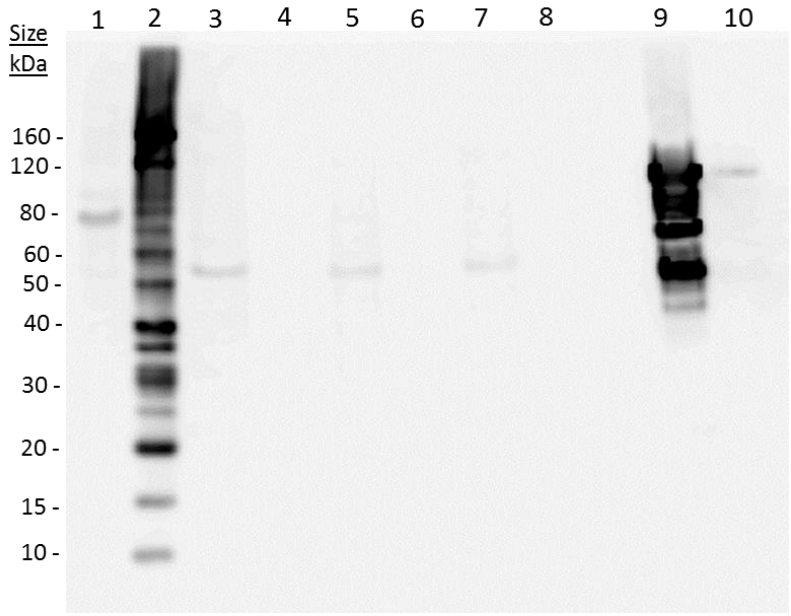


Figure 4.15. Western blots of His-tag purified proteins from SITMyb overexpression and WT cell lines. Lane 1) Colour protein standard, 2) Benchmark His-tagged standard, 3-4) WT, 5-6) SITMybOE 1, 7-8) SITMybOE_IR 2, 9) 5ug positive control, 10) 0.25ug positive control. Odd lanes between 3-8 were extracted with Xtractor buffer and even lanes with denaturing buffer.

Yeast-1-hybrid

According to JGI, two hundred and fifty eight genes with a Myb domain or Myb-like domain can be found in the *F. cylindrus* genome. This compares to 114 in *T. pseudonana* and 60 in *P. tricornutum*. The problem with predicting genes controlled by Myb transcription factors by searching transcription factor binding sites (TFBS), is that the binding site is expected to occur frequently due to the degenerate bases. It is pointed out by Berge et al. (2001), that in yeast a potential Myb binding sequence could occur at random, on average every 1024 bp. *F. cylindrus* has a similar GC content to *Saccharomyces cerevisiae* so it would be reasonable to expect a similar random occurrence of this motif. As a result a more empirical method is required to elucidate potential binding sites.

Yeast-1-hybrid was chosen to determine potential TFBS of SITMyb and the Myb domain. The protocol used was closely based on the one described by Yan and Burgess (2012). The method involves creating yeast clones containing a gDNA library with fragments cloned upstream of a URA3 gene and mating them with clones containing the target transcription factor fused to a GAL4 activation domain. If the transcription factor:GAL4 fusion binds to a cloned fragment in the gDNA library, then the URA3 gene can be expressed, allowing growth on media deficient in uracil. The library vector in positive clones can then be sequenced to determine the binding site and target gene. Library clones also contain the Trp gene for selection with tryptophan deficient media and

the TF clones contain the ADE2 gene for selection on adenine deficient media and pink/white colony selection.

Optimising digest of F. cylindrus gDNA

F. cylindrus gDNA was digested with MluCI, a restriction enzyme with a 4nt recognition site which can cut frequently. Fragment size was optimised by changing enzyme concentrations, reaction volume and reaction time. Different methods for stopping and purifying digested DNA were also trialled. The optimised protocol can be found within the methods. Figure 4.16 shows *F. cylindrus* gDNA fragmented with 1U of enzyme/10µg with different reaction times. Figure 4.17 also shows different reaction times but in combination with different purification methods. Fragment size was measured by running products on a gel and determining the size at which signal was strongest.

Two gDNA digests were used for the final library preparations. One with a modal fragment size, in terms of signal at 250bp and one at 300bp. Fragment sizes ranged from 70bp to 2000bp following purification. The protocol calls for fragments to be 200-800bp in length, however extracting fragments from agarose gels to reduce the size range, resulted in large decreases in yield and low concentrations following elution.

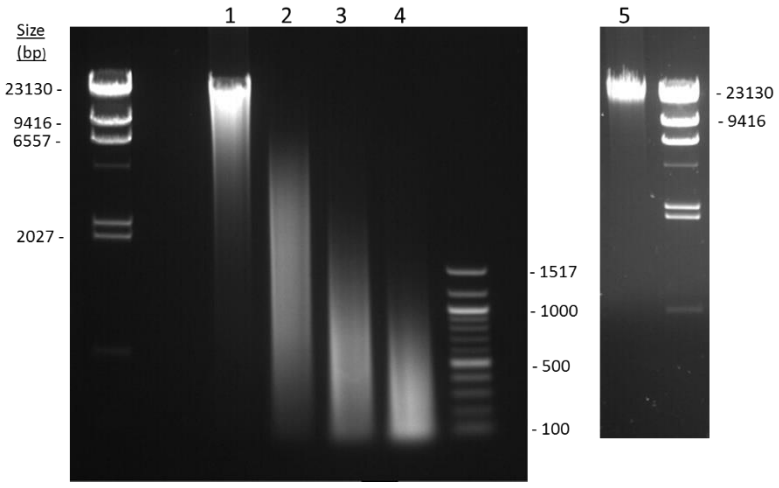


Figure 4.16. Optimising digest of *F. cylindrus* gDNA for the Y1H gDNA library. Digest with 1U MluCI/10µg of DNA for O (1), 2 (2), 5 (3) and 10 (4) minutes, followed by storage on ice. Lane 5 shows untreated gDNA.

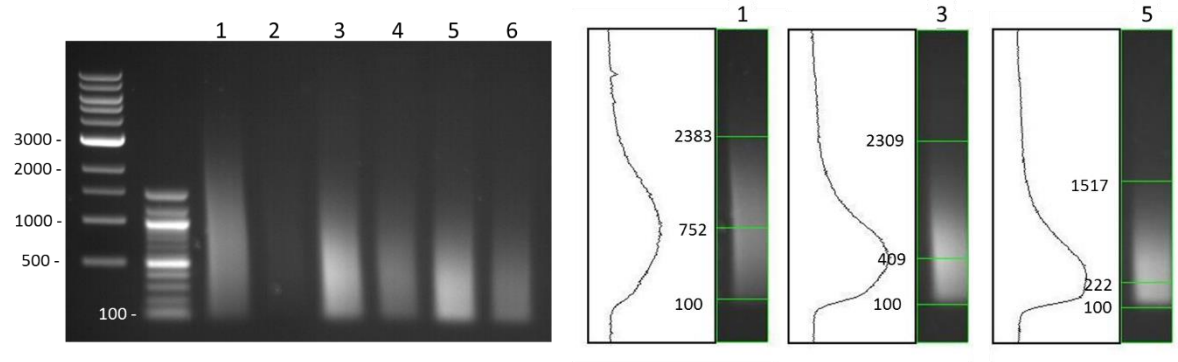


Figure 4.17. Optimising digest of *F. cylindrus* gDNA for the Y1H gDNA library. Lanes 1, 3 and 5 were purified with a Qiagen QIAquick, lanes 2, 4 and 6 were purified by isopropanol precipitation following digestion. Lanes 1-2 were digested with 1U MluCI/10µg of DNA for 3 minutes, lanes 3-4 for 5 minutes and lanes lanes 5-6 for 7 minutes.

Furthermore, transcription factor binding sites in yeast tend to bind at around 50-400 bp upstream of the transcription start site (Lin et al., 2010) and one study involving characterisation of a leucine zipper in *P. tricornutum* showed binding sites 42-86bp upstream of the transcription start site (Ohno et al., 2012). Studies on transcription factors and binding sites in diatoms are underrepresented (Matthijs et al., 2016), making it difficult to gauge the fragment length needed. For this reason, it was decided to digest gDNA and purify straight from the reaction without gel mediated size exclusion. As the majority of fragments appear to be under 1000bp and larger fragments tend to clone less efficiently, the chance of larger fragments being incorporated was considered to be less of an issue compared to loss of yield from gel extraction. Digests with the majority of fragments sizes around 300bp were chosen so that the binding site would not be too far from the start of transcription, but also so that the fragment was large enough to give a reasonable library coverage. Qiagen columns purify fragments from 70bp. As a result very small fragments were removed which may have otherwise been preferentially cloned due to their size. Optimised cloning of the gDNA fragments into pYOH366 gave an *E. coli* library size of 9 x 10^5 clones. Transformation into yeast resulted in 5.8 x 10^6 colonies. Following sequencing of YIH clones after mating and screening, the average insert size was found to be 275bp, which corresponds well with the original gDNA digests which gave an average size of 250 and 300bp. An average fragment size of 275 equates to a 6.15 times genome coverage in the bacterial library, which is well within the 5-10 times range suggested by Yan and Burgess (2012).

Generating pYOH1-TF

The same SITMyb sequence as used in the SITMyb_IR overexpression contruct for *F. cylindrus*, was cloned into pYOH1 in frame with the GAL activation domain and HA-tag. The His-tag at the C- terminus was also included. The individual Myb domain was also cloned into pYOH1. The domain was based on data from blastn, blastp and protein modelling seen earlier in the chapter.

Transformation into yeast resulted in plenty of white colonies, indicating the presence of ADE2 and therefore the TF.

Although several white colonies appeared after transformation of pYOH1-TF into MATα yeast strains, transferring these colonies to SD-His-Ade media to grow overnight consistently led to slightly pink cultures. A HIS3 cassette is co-transformed at the same time to also allow selection on SD-His-Ura plates following mating. This should lead to loss of the pYOH1-TF plasmid (which contains ADE2) if the TF is not required for growth on -Ura plates, thereby highlighting false positives in which the URA3 gene is functional without the TF by a white to pink colour change.

The change in colour of white pre-mated pYOH1-TF colonies to pink after picking from initial SD-Ade-His plates suggests that cells are losing the ADE2 gene and therefore the transcription factor, despite the absence of adenine in the media.

Western blots were carried out on crude and His-tag purified protein lysates from these cultures and probed with HA-tag and His-tag antibodies to check presence of the TF. Blots which targeted the HA-tag (Figure 4.18) showed a clear band around the expected size of the GAL4AD:SITMyb protein in each of the SITMyb samples (lanes 5-10), indicating that the TF is being expressed. Additional banding can be seen, however this is also present in the WT, suggesting that several of the endogenous yeast proteins are probed by the HA-tag antibody. Samples for overexpression of the Myb domain (lanes 3-4) also show a clear band at the correct size, however a faint band at this size is also present in the WT. In addition a slightly larger protein with a strong band around 32 kDa, exclusive to the Myb samples can be seen. Bands which correspond to the WT in these samples are present in similar quantities to the WT, whereas the expected band at 24.7kDa is stronger in the Myb samples, indicating that the Myb protein may be expressed. Proteins don't always run to the correct size on gels, and post-translational modification can also affect size and migration, meaning that the higher, exclusive band around 32 kDa may also be the Myb protein. No result was seen when probing the His-tag suggesting that it may be non-functional or inaccessible.

It was decided to carry on with the YIH method using the palest cultures, as Western blots suggested the transcription factor or Myb domain were present in at least a population of the cultures and changing the pYOH1 vector at this stage was not feasible due to time. In some cases, expression of the TF can be toxic to yeast (Zhu et al., 2016), however since white pYOH1-TF colonies grew well on SD –Ade-His plates, and subsequent cultures showed expression of the protein, this doesn't seem likely in this case.

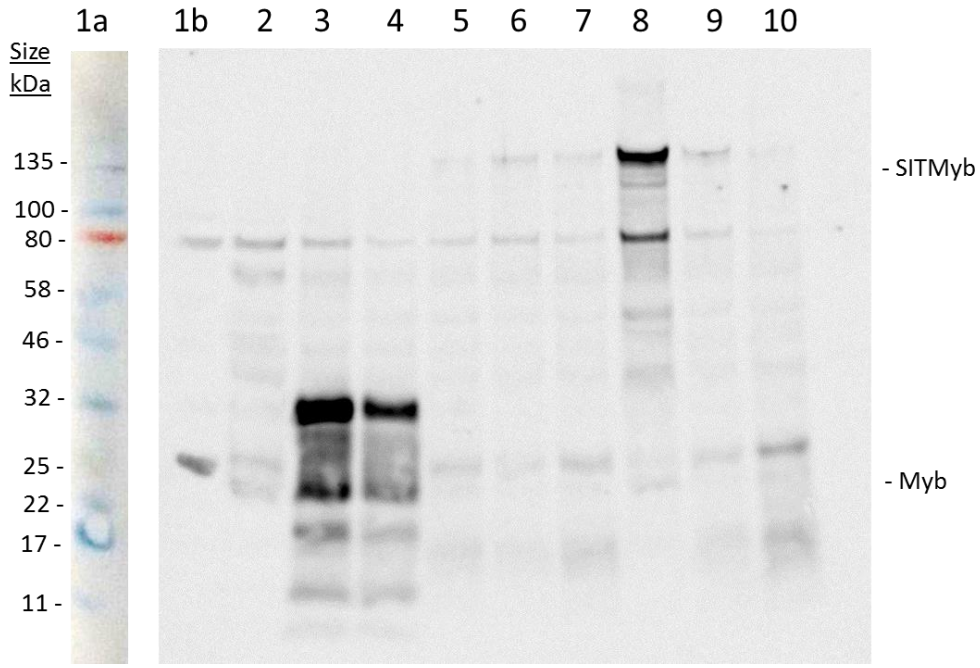


Figure 4.18. HA-tag western blots with crude protein lysate from SITMyb and Myb yeast overexpression cell-lines for yeast-1-hybrid. 1) Broad Range, Color Prestained Protein Standard ladder (NEB), a- brightfield, b- chemiluminescence, 2) WT 3-4) W303 pYOH1-Myb cell lines, 5-10) W303 pYOH1-SITMyb cell lines. The expected size for SITMyb and Myb, with the GAL4 Activation domain and HA-tag is 120.5 and 24.7 kDa respectively.

Mating of yeast cell lines and screening

Following mating, cells were initially selected on SD-His-Ura plates. Mating efficiency which was calculated from dilutions on SD-His-Trp plates, was around 10% for SITMyb cell lines, however it was very low for Myb, with only a few colonies present, suggesting that mating may not have been successful for the latter. White colonies from Myb SD-His-Ura plates also failed to grow when later re-plated out onto SD-Ura media. Mating with strains containing pYOH1-Myb needs to be repeated. Just under 200 white SITMyb colonies from the SD-His-Ura plates were transferred to YPDA and replica plated on screening plates (Figure 4.19). After transferring to YPDA, around 75% of clones reverted to a pink colour, suggesting that the ADE2 gene and therefore the SITMyb gene was lost after transfer to YPDA. As discussed a lack of selective pressure can lead to loss of plasmids, especially when they don't contain a CEN sequence, as is the case with pYOH1-TF (Dani and Zakian, 1983), which may explain the large shift from white to pink colonies at this stage. Despite this, the majority of colonies grew on -Ura following replica plating from the YPDA plate. True positives, in which URA3 is expressed following binding of the TF to a fragment in the gDNA library, should be white on –Ura, Pink on +5-FOA and should section on –Trp plates. Presence of +5-FOA is toxic in yeast with a functional URA3 gene and as there is no selective pressure for the pYOH1-TF either on +5-FOA or -Trp plates, ADE2 should be gradually lost, leading to only pink cells on +5-FOA and colonies with a mixture of white and pink cells on - Trp.

A mixture of different phenotypes occurred across the plates. Table 4.7 describes the different variations, frequency and possible explanation for each, along with methods to either verify true positives or solutions for false positives. Whilst, white, pink and no-growth phenotypes were seen on YPDA, SD -Ura and +5-FOA, colonies on SD -Trp plates showed either no growth or appeared as pink or yellow, rather than sectioning.

As mentioned, absence of adenine when growing colonies with pYOH1-TF plasmids doesn't appear to prevent loss of the plasmid. It has been shown that plasmids in yeast are rapidly lost without selective pressure (Dani and Zakian, 1983). This can be reduced but not prevented by the presence of a centromeric (CEN) sequence (Dani and Zakian, 1983; Stearns et al., 1990). Dani and Zakian (1983) found that plasmids without a CEN sequence were lost at a rate of 21% per generation without selection, compared to a rate of 3% when a CEN sequence was included. If lack of adenine is not providing adequate selective pressure in the pYOH1-TF cell lines then high rates of loss and multiple incubations seen by the samples at this point may have led to loss of the TF in a large portion of the colonies. If this is the case then pink colonies may have no ADE2 gene and yellow colonies may be due to reversal of the ADE2 mutation. Furthermore pink colonies, which are expected to have lost the ADE2 gene and therefore the transcription factor, should not be able to grow on –Ura plates. This suggests possible binding of endogenous yeast TFs to gDNA fragments, leaky expression of URA3 from the pYOH366-g plasmid, or URA3 expression independent of pYOH366-g, possibly by integration into the genome or reversal of the URA3 mutation found in the original W303 cell lines.

The presence of white colonies or no growth on 5-FOA certainly suggests that for some colonies expression of the URA3 gene is not linked to presence of the transcription factor. In these cases white colonies may be due to the presence of the ADE2 gene, possibly through reversion of the mutant or integration, rather than from the pYOH1-TF. Colonies with no-growth on 5-FOA suggest a URA3 revertant rather than URA3 expression from the pYOH366-g plasmid (either from binding of yeast TFs or SITMyb), which should give pink or sectioning colonies under this condition given that 5-FOA selects against either pYOH1-TF or pYOH366-g.

Only 6.5% of colonies were white under –Ura, pink under +5-FOA and either pink or yellow on –Trp. These are colonies most likely to be true positives. Thirty six genomic DNA inserts from clones with the white -Ura phenotype were amplified (Figure 4.20) and sequenced. Of these, 11 showed inserts comprised of multiple fragments, 25 clones contained a single fragment; 8 showed 2 concatenated fragments and 3 clones contained 3 fragments. The average insert size was 275bp, whilst the average fragment size was 203bp. Twenty of the fragments fell within a coding region and 7 within an intergenic region. Six of the latter were upstream of a gene and may be potential candidates for binding. Two fragments showed homology to the chloroplast genome, two occurred

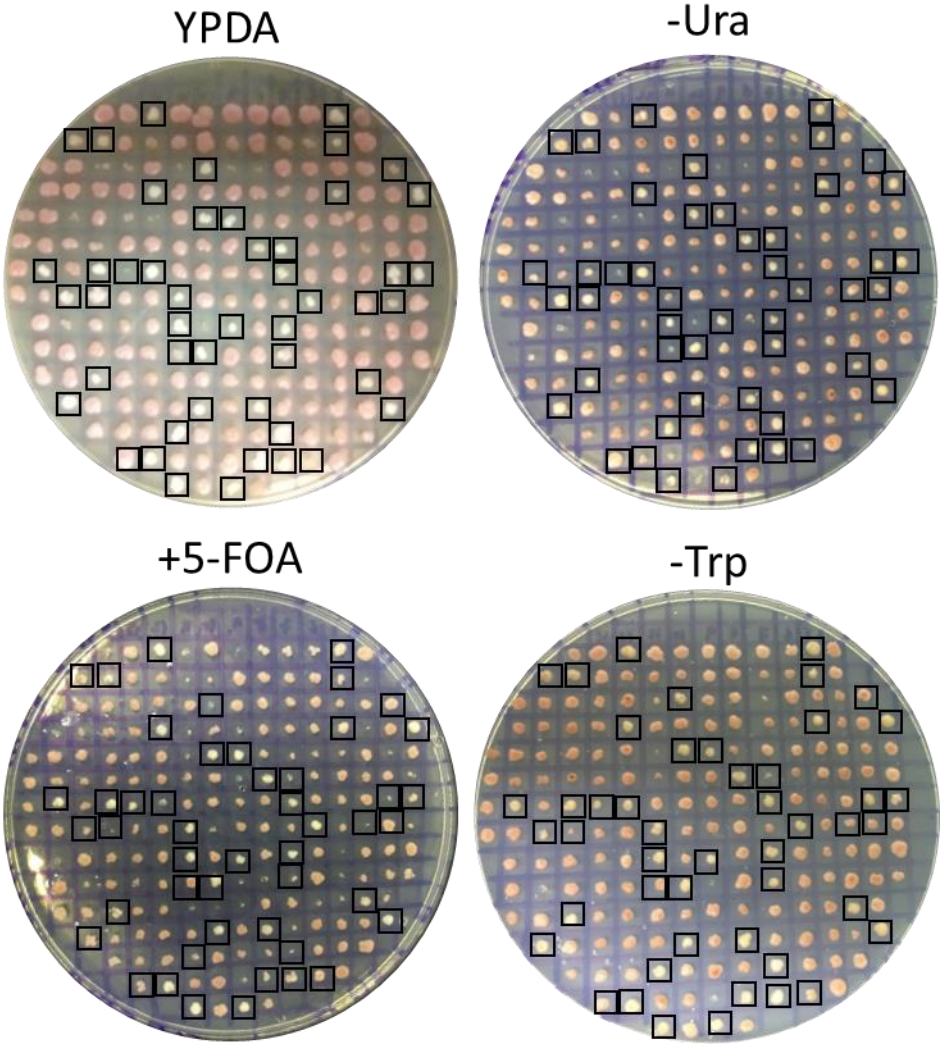


Figure 4.19. Ura+ yeast one hybrid colonies. W303a pYOH1-SITmyb and W303α pYOH366-g were mated and colonies with the Ura+ phenotype selected on –His-Ura plates. Colonies were picked, re-plated onto the YPDA plate and allowed to grow before replica plating onto SD –uracil (-Ura), +5-FOA and SD –tryptophan (-Trp) plates. Colonies 1-48 (top to bottom, left to right) selected for PCR have been highlighted.

in an intronic region and one fragment from a highly repetitive element was present. Several could not be identified or showed poor sequence quality. Twelve of the inserts included a consensus Myb binding site, designated as YAACKG (Biedenkapp et al., 1988; Ogata et al., 2004). Of the inserts with a potential Myb binding site, 4 were located in a potential promoter region upstream of a gene. Three of these were unique and sat upstream of genes 23174, 234174 (duplicated) and 182302. The first two genes are predicted proteins with homology to hypothetical proteins in *P. tricornutum* and the third is a serine peptidase: peptidase S9 propyl oligopeptidase. This may be interesting as both silaffins and silacidins are rich in serine residues (Kroger et al., 2002; Wenzl et al., 2008).

However, the colony for this gene, as well as the other non-duplicated colony gave a phenotype of white/white/yellow on SD –Ura/+5FOA/ -Trp plates, suggesting that they may be false positives. The duplicate colony gave a white/pink/yellow phenotype, so may be a true positive but this needs further validation.

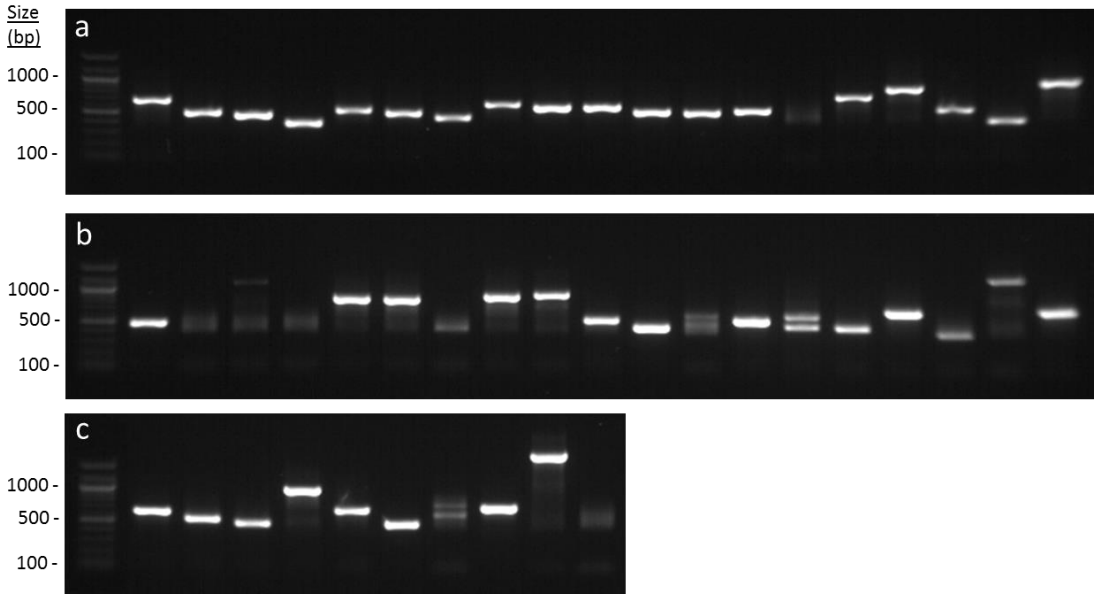


Figure 4.20. Colony PCR of *F. cylindrus* gDNA inserts in pYOH366-g, following screening of potential SITMyb binding sites in yeast.

Considering the large number of fragments found within a coding region and the low occurrence of white/pink colonies on -Ura/+5FOA plates it seems likely that the majority of colonies are false positives. It may be that, due to the occurrence or revertants, random integration of selective markers, leaky URA3 expression or binding of endogenous yeast TFs, false positives are expected to occur. Lack of potential true positives, however makes the false positives more apparent. At this stage, it's difficult to say if SITMyb is able to bind as there are clearly some problems with execution of the current method. The most obvious issues to address in the method at this stage, are the loss of the TF prior to mating which will reduce the chances of seeing true positives and the high number of pink colonies on –Ura plates, which suggests that the URA3 gene is expressed independently of the TF. In addition, screening of colonies for the pYOH1-TF and pYOH366-g using a method such as colony PCR, may help to determine false positives.

False positives due to binding of endogenous yeast TFs should be removed during selection on 5-FOA prior to mating. If clones with a compatible binding site for endogenous TFs are passing this screening then it may be necessary to increase concentration of 5-FOA or repeat the screening more than once. This method uses 1mg/ml 5-FOA, though other methods can be found which use higher concentrations (Saghbini et al., 2001). In this method, the gDNA fragment is inserted into a

SPO13 promoter upstream of the URA3 gene. As genes under the control of this promoter are repressed, it should prevent expression of URA3 unless a gDNA fragment is inserted which can bind a transcription factor. This method to reduce leaky URA3 has proved successful in other systems (Vidal et al., 1996; Yanai, 2013), so is less likely to be the cause of false positives, compared to binding of endogenous yeast TFs, which is one of the main constraints in YIH systems (Zhu et al., 2016).

As there is a problem with maintaining the pYOH1-TF vectors in W303 MATα under adenine selection, it may be worth trying an additional or different selection marker. It's common for other yeast-1-hybrid and yeast-2-hybrid methods to use markers such as LEU2, HIS3 or URA3 to select for the binding protein (Hosoda et al., 2015; James, 2001; Liu et al., 1993; Taniguchi-Yanai et al., 2010; Vidal et al., 1996; Yanai, 2013). Vectors compatible with this method such as pGADT7-GW (Lu et al., 2010), are already available for cloning of the transcription factor. Inclusion of a centromeric region may help to stabilize the pYOH1-TF plasmid (Dani and Zakian, 1983; Stearns et al., 1990), however this would also change the copy number from high to 1-2 copies which may have an effect on expression of the protein.

Condition	%	Possible cause of phenotype.	Verification and Solutions
PPP	50.7%	Pink colonies indicate that pYOH1 containing ADE2 and the TF is not present. Growth on -Ura and +5-FOA suggests that URA3 is expressed but URA3 expression can be lost. This points to expression from pYOH366-g, suggesting either binding of an endogenous TF or leaky URA3 expression.	The presence of pYOH366-g can be tested by PCR as can absence of the TF. 5-FOA concentration can be increased to help reduce binding of yeast TFs to the pYOH366-g gDNA inserts. Loss of pYOH1-TF under ADE2 selection needs to be addressed.
ZPNP	15.7%	Pink colonies indicate that pYOH1 containing ADE2 and the TF is not present. Growth on –Ura but not on +5-FOA suggests URA3 expression from the genome rather than a plasmid. This points towards reversion of the URA3 mutant in the W303 yeast strain or integration.	The presence of pYOH366-g can be tested by PCR as can absence of the TF. PCR and sequencing of the URA3 gene can be carried out to verify the variant. Inclusion of a CEN sequence reduces copy number to 1-2 copies, this may help to reduce the chances of integration if the URA3 from pYOH366-g is being integrated.
WWY	10.7%	White colonies indicate that a functional ADE2 gene or a white mutant (Ugolini and Bruschi, 1996) is present. Functional ADE2 may be present either from pYOH1-TF, reversion of ADE2-1 or integration.	The presence of pYOH366-g can be tested by PCR as can absence of the TF. PCR and sequencing of the ADE2 gene can be carried out to verify the variant.

		Growth on –Ura but presence of white colonies on 5-FOA suggests that URA3 expression is not linked to binding of SITMyb and is likely expressed from pYOH366-g. This is supported by the majority of white clones containing pYOH366-g. As colonies do not sector on -Ura, the white phenotype is unlikely to be from pYOH1-TF.	Inclusion of a CEN sequence may help to reduce the chances of integration.
NPP	7.6%	No growth on -Ura- suggests absence of the URA3 gene. Pink colonies indicate that pYOH1 containing ADE2 and the TF is not present. Growth was seen on original SD –His-Ura selection, suggesting an initial false positive or loss of pYOH1-TF before transfer to screening plates	Reduce cell numbers plated on SD –His-Ura plates to prevent growth on dead cells. Include a CEN sequence to help stabilise pYOH1-TF during non-selective YPDA incubations to reduce loss of plasmid.
WPY	5%	White colonies indicate that a functional ADE2 gene or a white mutant is present. Pink colonies on 5-FOA suggest the white phenotype is linked to the pYOH-TF. Colonies with phenotype may be true positives, however only 2 colonies from this group showed an insert from a possible promoter region, suggesting that the TF is present but may not be specifically responsible for URA3 expression.	The presence of pYOH366-g and pYOH1-TF can be tested by PCR.
WNY	2.5%	White colonies indicate that a functional ADE2 gene or a white mutant is present. Growth on -Ura but not on +5-FOA suggests URA3 expression from the genome rather than a plasmid. This points towards reversion of the URA3 mutant in the W303 yeast strain or integration. As colonies do not sector on -Ura, the white phenotype is unlikely to be from pYOH1-TF. Sequencing suggested pYOH366-g was absent from most clones which indicates growth on SD -Trp may be from a revertant.	The presence of pYOH366-g and pYOH1-TF can be tested by PCR. PCR and sequencing of the ADE2, URA3 and TRP2 genes can be carried out to verify the variants. Inclusion of a CEN sequence may help to reduce the chances of integration.

WPP	1.5%	White colonies indicate that a functional ADE2 gene or a white mutant is present. Pink colonies on 5-FOA suggest the white phenotype is linked to the pYOH-TF. Colonies with this phenotype may be true positives, however in this case, sequenced colonies showed inserts from coding regions rather than promoters.	The presence of pYOH366-g and pYOH1-TF can be tested by PCR.
WNP	1%	White colonies indicate that a functional ADE2 gene or a white mutant is present. Growth on -Ura but not on +5-FOA suggests URA3 expression from the genome rather than a plasmid. This points towards reversion of the URA3 mutant in the W303 yeast strain or integration. As colonies do not sector on -Ura, the white phenotype is unlikely to be from pYOH1-TF, however pink colonies on -Trp contrast this suggesting ADE2 gene can be lost, likely originating from pYOH1-TF.	The presence of pYOH366-g and pYOH1-TF can be tested by PCR. PCR and sequencing of the ADE2 and URA3 genes can be carried out to verify the variants.
PNN	2%	Pink colonies indicate that pYOH1 containing ADE2 and the TF is not present. No growth on 5-FOA suggests that functional URA3 expression is originates from the genome – this is supported by lack of growth on -Trp which suggests absence of pYOH366-g. URA3 is likely to be a revertant.	Presence of URA3 can be checked with PCR and sequencing. PCR can be used to determine the presence of pYOH366.
PPN	1%	Pink colonies indicate that pYOH1 containing ADE2 and the TF is not present. Growth on 5-FOA suggests that URA3 is expressed from pYOH366-g, however lack of growth in -Trp suggests the same plasmid is absent.	The presence of pYOH366-g and pYOH1-TF can be tested by PCR.

Table 4.7. Phenotypes of mated YIH SITMyb clones on screening plates. Possible reasons for the phenotype are shown along with possible solutions to troubleshoot false positives. Codes in the left-hand column refer to phenotypes on plates as follows: W; White colonies, P; Pink colonies, N; No growth.

In summary of the yeast-1-Hybrid section, the purpose of this method was to try and elucidate potential binding sites of the SITMyb gene, if it binds to DNA. Generation of the pYOH366-g library by insertion of *F. cylindrus* gDNA fragments into the SPO13 promoter upstream of a URA3 in pYOH366-g appears to have been successful. Average insert size was 275bp, equating to a 6.15 x genome coverage.

The overall method, however, leads to phenotypes or gDNA inserts which suggest an abundance of false positives. Only 1% (2 replica clones) of colonies give a phenotype and insert which could potentially be a true positive, with further validation needed.

Two main issues stand out with the method. First, the TF carrying plasmid, pYOH1-TF, appears to be gradually rejected from yeast cultures following transformation, despite the presence of ADE2 which should put a positive selection pressure on this plasmid. Western blots show proteins at the correct size in cultures transformed with both the SITMyb gene and the individual Myb domain. However, cultures transformed with both are pale pink indicating loss of pYOH1. This is also seen later following mating, when transferring clones to YPDA. This may explain the low yield of potential true positives.

Second, there are a lot of false positives, which are made more evident by the lack of true positives. The most likely explanation, given the phenotypes observed and known issues with the yeast-1-hybrid method (Zhu et al., 2016), is that negative selection on 5-FOA is not removing all clones with gDNA inserts capable of binding endogenous yeast transcription factors.

Other factors may also lead to false positives such as occurrence of revertants or integration of selective markers into the genome. Hishida et al. (2002) found that reversion frequencies of TRP1-1 and ADE2-1 on YPDA were 2 and 3.6 in 10^7 cells respectively and user manuals for ADE2-1 containing mutant yeast strains stress the need to grow cells with additional adenine to prevent frequent reversion.

Outlook for yeast-1-hybrid

The question at this point is, how can the method be improved and what further work is needed?

Further investigation can be carried out to try and determine the cause of the false positives, including colony PCRs to directly test for presence of the pYOH1-TF, pYOH366-g and different variants of the selective markers.

Higher concentrations of 5-FOA need to be trialled when negatively selecting W303 MATa pYOH366-g, possibly with multiple screenings to reduce carryover of clones with sites capable of binding endogenous yeast transcription factors.

One of the most important issues to address is loss of the plasmid pYOH1-TF, containing SITMyb or Myb. Changing the selective marker or adding additional selective markers to increase selective pressure may help. It may also be beneficial to include a CEN sequence to stabilize the plasmid and

reduce the rate of loss as well as double checking the selective media to ensure it doesn't contain adenine.

Only the SITMyb strain was successfully mated and further work would need to be carried out with both SITMyb and Myb. It would also be interesting to use the full SITMyb gene and both alleles described earlier in the chapter. If mating remains inefficient for particular strains, then other methods such as the one shown by Yanai (2013), which first transforms the library into yeast and then transforms the TF plasmid into the library strain, may be used.

It should be pointed out that the SITMyb protein may not bind DNA under the current system, or it may not bind DNA at all. Not all proteins are folded or correctly modified in yeast which can affect activity (Gasser et al., 2008; Tang et al., 2016), and there's still a level of uncertainty regarding the general gene model of SITMyb which may require further work before an active protein can be produced. In addition, if SITMyb does bind to a consensus Myb DNA binding site, then it may be possible for certain endogenous Myb yeast TFs to bind at the same site, removing clones from the experimental pool through negative 5-FOA selection and reducing occurrence of positive clones. It's difficult to say how likely this is, given that a variety of Myb TFs exist with different binding sites (Feller et al., 2011; Williams and Grotewold, E., 1997) and that structural differences in Myb TFs between organisms can affect binding specificity (Williams and Grotewold, E., 1997). Furthermore, binding activity of endogenous TFs may be influenced by additional regulatory mechanisms (Pireyre and Burow, 2015).

The SANT domain has a close homology to Myb domains and is found in transcriptional co-regulators (Aasland et al., 1996) and proteins that bind to chromatin (Iyer et al., 2008). SANT proteins control regulation through protein-protein interactions by binding histone tails or other transcriptional regulators (Aasland et al., 1996; Iyer et al., 2008). A lack of results from YIH may be due to SITMyb forming protein-protein interactions rather than DNA-protein interactions. If this is the case then Yeast-2-Hybrid may be a more appropriate method to determine its target. It may also be worth pursuing this method to determine if the SIT domain is involved in protein-protein interactions.

Concluding remarks and future work

The function of the SITMyb gene is yet to be determined, but several of the domains and motifs present, support its potential as a regulatory protein that may be linked to silicon metabolism. This includes a single Myb domain with a modelled helix-turn-helix structure that is associated with DNA binding and belongs to a family of prevalent transcription factors in stramenopiles (Rayko et al., 2010). Myb-like SANT domains are also linked to regulation through protein interactions. Furthermore, the presence of the SIT domain suggests a role linked to silicon transporters. The SIT domain present in SITMyb corresponds to the C-terminus of several diatom SITs, rather than any of the transmembrane domains which are predicted to be involved in silicic acid binding via a GXQ motif (Thamatrakoln et al., 2006). However, a GXQ motif is present downstream of the SIT

domain in SITMyb. The SIT domain also contains a coiled-coil motif which is associated with protein binding (Mier et al., 2017). As not all SITs appear to have a transport role (Sapriel et al., 2009) and some are proposed to function as a sensor (Shrestha and Hildebrand, 2015), there is a possibility that the SIT domain may be binding a sensory or regulatory protein. Additional motifs are also present that are positively linked to a role in regulation such as the presence of putative NLS sequences for localisation to the nucleus, a stretch of proline residues found in other transcription factors (Gerber et al., 1994) and several coiled-coil motifs.

In terms of the transcript, modelling of this gene needs further work as RACE experiments and previously generated RNA-seq data do not give a clear indication of the transcription start site. Overexpression of the SITMyb gene just after the predicted 1st intron, led to transcripts in overexpression cell lines but no evidence of protein. This appears to be due to a problem with labelling via the His-tag considering that later expression in yeast appeared to be successful when probing with the HA-tag, but not the His. This means that the SITMyb protein may be present but this cannot be confirmed in the current overexpression cell-lines without development of a SITMyb antibody. Additional overexpression constructs are needed with the full predicted JGI protein as well as the current model. Functional tags need to be built in, such as HA-tags or egfp to confirm expression of the protein before phenotyping. HA-tags would allow the protein to be purified and egfp could help to confirm whether or not this is a nuclear protein, as expected from a transcription factor. If evidence of the overexpressed protein can be found, then RNA-sequencing may help to determine if an overabundance of this protein affects regulation of other genes. This could also be applied to knock-out cell-lines which are currently being developed in the Mock lab using CRISPR-Cas and sgRNAs designed in this chapter. If this is achieved, then phenotyping including determining the levels of precipitated silica with PDMPO or examining frustule morphology at different scales will need to be carried out for both overexpression and knock-out cell-lines.

Yeast-1-hybrid was carried out to try and determine possible binding sites of the SITMyb gene and Myb domain. The gDNA library represents the ‘prey’ and the transcription factor the ‘bait’. This method is often carried out in reverse to determine the transcription factor for a specific binding domain. A gDNA library linked to a URA3 selective marker, was produced with a 6x coverage and proteins at the correct size were observed for the TF SITMyb/Myb:GAL4 AD overexpression lines. However, plasmid loss was observed for the TF lines and ultimately very few colonies were observed that could be true positives. This is likely to be due to binding of endogenous yeast transcription factors to fragments in the library. In addition, loss of the TF plasmid will reduce the chance of seeing true positives. This can potentially be addressed by increasing screening of the gDNA library and adding additional selective markers and a CEN-ARS sequence to the TF plasmid to encourage maintenance and retention. It may be that SITMyb does not bind to DNA. If this is the case, then it may be worth carrying out Yeast-2-hybrid to determine if SITMyb binds to other proteins. This would also be a useful tool to examine the SIT domain.

With the expression of SITMyb and Myb in yeast, further work to elucidate the function of this gene and the Myb domain can be carried out. However, it would also be interesting to see if overexpressed protein can be purified from the *F. cylindrus* overexpression lines once the construct has been altered, as folding and post-translational modification is more likely to be correct if produced in an endogenous host. This is especially relevant for *F. cylindrus* as a polar organism given that temperature can affect solubility and protein folding (San-Miguel et al., 2013; Vasina and Baneyx, 1996). Mobility shift assays can be carried out, with targets identified from Y1H or genes linked to the silica frustule to explore DNA binding. Silicic acid binding can also be examined using methods such as protein crystallography.

Although no conclusions can be drawn at this stage, this work provides preliminary data and method development required to try and elucidate the function of this gene. It provides leads for changes needed to develop functional methods such as overexpression in *F. cylindrus* and yeast-1-hybrid and provides tools such as a gDNA library for elucidation of transcription factor binding sites. In addition, modelling of this gene gives an intriguing glimpse into its potential as a silicon-linked regulatory protein, which may help to shed light on the regulatory networks involved in silica metabolism.

References

Aasland, R., Stewart, A.F., Gibson, T., 1996. The SANT domain: a putative DNA-binding domain in the SWI-SNF and ADA complexes, the transcriptional co-repressor N-Cor and TFIIB. rends Biochem. Sci. 21, 87–88.

Agatep, R., Kirkpatrick, R.D., Parchaliuk, D.L., Woods, R. a., Gietz, R.D., 1998. Transformation of Saccharomyces cerevisiae by the lithium acetate/single-stranded carrier DNA/polyethylene glycol protocol. Tech. Tips Online 3, 133–137. doi:10.1016/S1366-2120(08)70121-1

Apt, K.E., Zaslavkaia, L., Lippmeier, J.C., Lang, M., Kilian, O., Wetherbee, R., Grossman, A.R., Kroth, P.G., 2002. In vivo characterization of diatom multipartite plastid targeting signals. J. Cell Sci. 115, 4061–4069. doi:10.1242/jcs.00092

Armbrust, E. V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., Brzezinski, M.A., Chaal, B.K., Chiovitti, A., Davis, A.K., Demarest, M.S., Detter, J.C., Glavina, T., Goodstein, D, D., 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science (80-.). 306, 79–86.

Arnold, K., Bordoli, L., Kopp, J., Schwede, T., 2006. The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. Bioinformatics 22, 195–201. doi:10.1093/bioinformatics/bti770

Ashworth, J., Coesel, S., Lee, A., Armbrust, E.V., Orellana, M. V, Baliga, N.S., 2013. Genome-wide diel growth state transitions in the diatom Thalassiosira pseudonana. Proc. Natl. Acad. Sci. U. S. A. 110, 7518–23. doi:10.1073/pnas.1300962110

Ashworth, J., Turkarslan, S., Harris, M., Orellana, M. V., Baliga, N.S., 2016. Pan-transcriptomic analysis identifies coordinated and orthologous functional modules in the diatoms Thalassiosira pseudonana and Phaeodactylum tricornutum. Mar. Genomics 26, 21–28. doi:10.1016/j.margen.2015.10.011

Berge, T., Bergholtz, S.L., Andersson, K.B., Gabrielsen, O.S., 2001. A novel yeast system for in vivo selection of recognition sequences: defining an optimal c-Myb-responsive element. Nucleic Acids Res. 29, E99.

Bhattacharyya, P., Volcani, B.E., 1980. Sodium-dependent silicate transport in the apochlorotic marine diatom Nitzschia alba. Proc. Natl. Acad. Sci. U. S. A. 77, 6386–90. doi:10.1073/pnas.77.11.6386

Biedenkapp, H., Borgmeyer, U., Sippel, a E., Klempnauer, K.H., 1988. Viral myb oncogene encodes a sequence-specific DNA-binding activity. Nature. doi:10.1038/335835a0

Brameier, M., Krings, A., MacCallum, R.M., 2007. NucPred - Predicting nuclear localization of proteins. Bioinformatics 23, 1159–1160. doi:10.1093/bioinformatics/btm066

Brunner, E., Richthammer, P., Ehrlich, H., Paasch, S., Simon, P., Ueberlein, S., Van Pée, K.H., 2009. Chitin-based organic networks: An integral part of cell wall biosilica in the diatom thalassiosira pseudonana. Angew. Chemie - Int. Ed. 48, 9724–9727. doi:10.1002/anie.200905028

Brzezinski, M., Olson, R., Chisholm, S., 1990. Silicon availability and cell-cycle progression in marine diatoms. Mar. Ecol. Prog. Ser. 67, 83–96. doi:10.3354/meps067083

Buitrago-Florez, F.J., Restrepo, S., Riano-Pachon, D.M., 2014. Identification of transcription factor genes and their correlation with the high diversity of stramenopiles. PLoS One 9, 1–8. doi:10.1371/journal.pone.0111841

Cavalier-Smith, T., 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J. Eukaryot. Microbiol. 46, 347–366. doi:10.1111/j.1550-7408.1999.tb04614.x

Curnow, P., Senior, L., Knight, M.J., Thamatrakoln, K., Hildebrand, M., Booth, P.J., 2012. Expression, purification, and reconstitution of a diatom silicon transporter. Biochemistry 51, 3776–3785. doi:10.1021/bi3000484

Dani, G.M., Zakian, V.A., 1983. Mitotic and meiotic stability of linear plasmids in yeast. Proc Natl Acad Sci U S A 80, 3406–3410.

Davis, A.K., Hildebrand, M., Palenik, B., 2005. A stress-induced protein associated with the girdle band region of the diatom Thalassiosira pseudonana (Bacillariophyta). J. Phycol. 41, 577–589. doi:10.1111/j.1529-8817.2005.00076.x

De Castro, E., Sigrist, C.J. a, Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., Hulo, N., 2006. ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 34, 362–365. doi:10.1093/nar/gkl124

Deng, Q.L., Ishii, S., Sarai, A., 1996. Binding site analysis of c-Myb: Screening of potential binding sites by using the mutation matrix derived from systematic binding affinity measurements. Nucleic Acids Res. 24, 766–774. doi:10.1093/nar/24.4.766

Dohm, J.C., Lottaz, C., Borodina, T., Himmelbauer, H., 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36. doi:10.1093/nar/gkn425

Du, C., Liang, J.R., Chen, D.D., Xu, B., Zhuo, W.H., Gao, Y.H., Chen, C.P., Bowler, C., Zhang, W., 2014. ITRAQ-based proteomic analysis of the metabolism mechanism associated with silicon response in the marine diatom thalassiosira pseudonana. J. Proteome Res. 13, 720–734. doi:10.1021/pr400803w

Durak, G.M., Taylor, A.R., Walker, C.E., Probert, I., de Vargas, C., Audic, S., Schroeder, D., Brownlee, C., Wheeler, G.L., 2016. A role for diatom-like silicon transporters in calcifying coccolithophores. Nat. Commun. 7, 10543. doi:10.1038/ncomms10543

Durkin, C. a., Mock, T., Armbrust, E.V., 2009. Chitin in diatoms and its association with the cell wall. Eukaryot. Cell 8, 1038–1050. doi:10.1128/EC.00079-09

Feller, A., MacHemer, K., Braun, E.L., Grotewold, E., 2011. Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. Plant J. 66, 94–116. doi:10.1111/j.1365-313X.2010.04459.x

Flynn, K.J., Martin-Jézéquel, V., 2000. Modelling Si–N-limited growth of diatoms. J. Plankton Res. 22, 447–472. doi:10.1093/plankt/22.3.447

Ford, E., 2013. 2X Gibson Assembly Master Mix [WWW Document]. URL https://ethanomics.files.wordpress.com/2013/06/2x-gibson-assembly-master-mix.pdf

Frigeri, L.G., Radabaugh, T.R., Haynes, P. a., Hildebrand, M., 2006. Identification of Proteins from a Cell Wall Fraction of the Diatom Thalassiosira pseudonana Insights into Silica Structure Formation. Mol. Cell. Proteomics 5, 182–193. doi:10.1074/mcp.M500174-MCP200

Gasser, B., Saloheimo, M., Rinas, U., Dragosits, M., Rodríguez-Carmona, E., Baumann, K., Giuliani, M., Parrilli, E., Branduardi, P., Lang, C., Porro, D., Ferrer, P., Tutino, M., Mattanovich, D., Villaverde, A., 2008. Protein folding and conformational stress in microbial cells producing recombinant proteins: a host comparative overview. Microb. Cell Fact. 7, 11. doi:10.1186/1475-2859-7-11

Gerber, H.P., Seipel, K., Georgiev, O., Höfferer, M., Hug, M., Rusconi, S., Schaffner, W., 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263, 808–11. doi:10.1126/science.8303297

Goncalves, E.C., Koh, J., Zhu, N., Yoo, M.J., Chen, S., Matsuo, T., Johnson, J. V., Rathinasabapathi, B., 2016. Nitrogen starvation-induced accumulation of triacylglycerol in the green algae: Evidence for a role for ROC40, a transcription factor involved in circadian rhythm. Plant J. 85, 743–757. doi:10.1111/tpj.13144

Grachev, M. a., Annenkov, V. V., Likhoshway, Y. V., 2008. Silicon nanotechnologies of pigmented heterokonts. BioEssays 30, 328–337. doi:10.1002/bies.20731

Grachev, M., Sherbakova, T., Masyukova, Y., Likhoshway, Y., 2005. A potential zinc binding motif in silicic acid transport proteins of diatoms. Diatom Res. 20, 409–411.

Hildebrand, M., 2003. Biological processing of nanostructured silica in diatoms. Prog. Org. Coatings 47, 256–266. doi:10.1016/S0300-9440(03)00142-5

Hildebrand, M., Frigeri, L.G., Davis, A.K., 2007. Synchronized growth of Thalassiosira pseudonana (Bacillariophyceae) provides novel insights into cell-wall synthesis processes in relation to the cell cycle. J. Phycol. 43, 730–740. doi:10.1111/j.1529-8817.2007.00361.x

Hildebrand, M., Lerch, S.J.L., 2015. Diatom silica biomineralization: Parallel development of approaches and understanding. Semin. Cell Dev. Biol. 46, 27–35. doi:10.1016/j.semcdb.2015.06.007

Hildebrand, M., Volcani, B.E., Gassmann, W., Schroeder, J.I., 1997. A gene family of silicon transporters. Nature 385, 689.

Hishida, T., Ohno, T., Iwasaki, H., Shinagawa, H., 2002. Saccharomyces cerevisiae MGS1 is essential in strains deficient in the RAD6-dependent DNA damage tolerance pathway. EMBO J. 21, 2019–2029. doi:10.1093/emboj/21.8.2019

Hosoda, K., Kanno, Y., Sato, M., Inajima, J., Inouye, Y., 2015. Identification of CAR / RXR α Heterodimer Binding Sites in the Human Genome by a Modified Yeast One-Hybrid Assay 83–97.

Huysman, M.J., Martens, C., Vandepoele, K., Gillard, J., Rayko, E., Heijde, M., Bowler, C., Inzé, D., 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 Biol. 11, R17. doi:10.1186/gb-2010-11-2-r17

Huysman, M.J.J., Fortunato, A.E., Matthijs, M., Costa, B.S., Vanderhaeghen, R., Van den Daele, H., Sachse, M., Inzé, D., Bowler, C., Kroth, P.G., 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). Plant Cell 25, 1–15. doi:10.1105/tpc.112.106377

Huysman, M.J.J., Vyverman, W., De Veylder, L., 2014. Molecular regulation of the diatom cell cycle. J. Exp. Bot. 65, 2573–2584. doi:10.1093/jxb/ert387

Iyer, L.M., Anantharaman, V., Wolf, M.Y., Aravind, L., 2008. Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. Int. J. Parasitol. 38, 1–31. doi:10.1016/j.ijpara.2007.07.018

James, P., 2001. Yeast Two-Hybrid Vectors and Strains. Methods Mol. Biol. 177, 41–84.

Joshi-Deo, J., Schmidt, M., Gruber, A., Weisheit, W., Mittag, M., Kroth, P.G., Büchel, C., 2010. Characterization of a trimeric light-harvesting complex in the diatom Phaeodactylum tricornutum built of FcpA and FcpE proteins. J. Exp. Bot. 61, 3079–3087. doi:10.1093/jxb/erq136

Kelly, L. a., Mezulis, S., Yates, C., Wass, M., Sternberg, M., 2015. The Phyre2 web portal for protein modelling, prediction, and analysis. Nat. Protoc. 10, 845–858. doi:10.1038/nprot.2015-053

Kirkham, A., Richthammer, P., Schmidt, K., Wustmann, M., Maeda, Y., Hedrich, R., Brunner, E., Tanaka, T., van Pée, K., A, F., Mock, T., 2017. A role for the cell-wall protein silacidin in cell size of the diatom Thalassiosira pseudonana. ISME. doi:10.1038/ismej.2017.100

Koester, J., Brownlee, C., Taylor, A.R., 2016. Algal Calcification and Silicification. eLS 1–10. doi:10.1002/9780470015902.a0000313.pub2

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, 10171–10176. doi:10.1073/pnas.0900604106

Kroger, N., Bergsdorf, C., Sumper, M., 1996. Frustulins: Domain Conservation in a Protein Family Associated with Diatom Cell Walls. Eur. J. Biochem. 239, 259–264. doi:10.1111/j.1432-1033.1996.0259u.x

Kröger, N., Bergsdorf, C., Sumper, M., 1994. A new calcium binding glycoprotein family constitutes a major diatom cell wall component. EMBO J. 13, 4676–4683.

Kröger, N., Deutzmann, R., Bergsdorf, C., Sumper, M., 2000. Species-specific polyamines from diatoms control silica morphology. Proc. Natl. Acad. Sci. U. S. A. 97, 14133–14138. doi:10.1073/pnas.260496497

Kröger, N., Deutzmann, R., Sumper, M., 2001. Silica-precipitating peptides from diatoms: The chemical structure of silaffin-1A from Cylindrotheca fusiformis. J. Biol. Chem. 276, 26066–26070. doi:10.1074/jbc.M102093200

Kröger, N., Deutzmann, R., Sumper, M., 1999. Polycationic Peptides from Diatom Biosilica That Direct Silica Nanosphere Formation. Science (80-.). 286, 1129–1132. doi:10.1126/science.286.5442.1129

Kroger, N., Lorenz, S., Brunner, E., Sumper, M., 2002. Self-Assembly of Highly Phosphorylated Silaffins and Their Function in Biosilica Morphogenesis. Science (80-.). 298, 584–586. doi:10.1126/science.1076221

Lai, Y.-T., Chang, Y.-Y., Hu, L., Yang, Y., Chao, A., Du, Z.-Y., Tanner, J. a., Chye, M.-L., Qian, C., Ng, K.-M., Li, H., Sun, H., 2015. Rapid labeling of intracellular His-tagged proteins in living cells. Proc. Natl. Acad. Sci. 112, 201419598. doi:10.1073/pnas.1419598112

Larkin, M. a., Blackshields, G., Brown, N.P., Chenna, R., Mcgettigan, P. a., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, a., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948. doi:10.1093/bioinformatics/btm404

Likhoshway, Y. V., Masyukova, Y. a., Sherbakova, T. a., Petrova, D.P., Grachev, M. a., 2006. Detection of the gene responsible for silicic acid transport in chrysophycean algae. Dokl. Biol. Sci. 408, 256–260. doi:10.1134/S001249660603015X

Lin, Z., Wu, W.-S., Liang, H., Woo, Y., Li, W.-H., 2010. The spatial distribution of cis regulatory elements in yeast promoters and its implications for transcriptional regulation. BMC Genomics 11, 581. doi:10.1186/1471-2164-11-581

Liu, J., Osbourn, A., Ma, P., 2015. MYB transcription factors as regulators of phenylpropanoid metabolism in plants. Mol. Plant 8, 689–708. doi:10.1016/j.molp.2015.03.012

Liu, J., Wilson, T.E., Milbrandt, J., Johnston, M., 1993. Identifying DNA-Binding Sites and Analyzing DNA-Binding Domains Using a Yeast Selection System. METHODS A Companion to Methods Enzymol. 5, 125–137.

Lopez-Gomollon, S.M.B.T.R.S.M.F.M.I.M.V.M.T.D. and T.M., Beckers, M., Rathjen, T., Moxon, S., Maumus, F., Mohorianu, I., Moulton, V., Dalmay, T., Mock, T., 2014. Global discovery and characterization of small non-coding RNAs in marine microalgae. BMC Genomics 2014, 697. doi:10.1186/1471-2164-15-697

Lu, Q., Tang, X., Tian, G., Wang, F., Liu, K., Nguyen, V., Kohalmi, S.E., Keller, W. a., Tsang, E.W.T., Harada, J.J., Rothstein, S.J., Cui, Y., 2010. Arabidopsis homolog of the yeast TREX-2 mRNA export complex: Components and anchoring nucleoporin. Plant J. 61, 259–270. doi:10.1111/j.1365-313X.2009.04048.x

Lupas, A., Van Dyke, M., Stock, J., 1991. Predicting coiled coils from protein sequences. Science (80-.). 252, 1162–1164. doi:10.1126/science.252.5009.1162

Maheswari, U., Mock, T., Armbrust, E.V., Bowler, C., 2009. Update of the Diatom EST Database: A new tool for digital transcriptomics. Nucleic Acids Res. 37. doi:10.1093/nar/gkn905

Marron, A.O., Alston, M.J., Heavens, D., Akam, M., Caccamo, M., Holland, P.W.H., Walker, G., 2013. A Family of Diatom-Like Silicon Transporters in the Siliceous Loricate Choanoflagellates. P. Roy. Soc. B-Biol. Sci. 280, 20122543. doi:10.1098/rspb.2012.2543

Marron, A.O., Ratcliffe, S., Wheeler, G.L., Goldstein, R.E., King, N., Not, F., de Vargas, C., Richter, D.J., 2016. The Evolution of Silicon Transport in Eukaryotes. Mol. Biol. Evol. 33, msw209. doi:10.1093/molbev/msw209

Mason, J.M., Arndt, K.M., 2004. Coiled coil domains: Stability, specificity, and biological implications. ChemBioChem 5, 170–176. doi:10.1002/cbic.200300781

Matthijs, M., Fabris, M., Broos, S., Vyverman, W., Goossens, A., 2016. Profiling of the Early Nitrogen Stress Response in the Diatom *Phaeodactylum tricornutum* Reveals a Novel Family of RING-Domain Transcription Factors. Plant Physiol. 170, 489–498. doi:10.1104/pp.15.01300

Matthijs, M., Fabris, M., Obata, T., Foubert, I., Franco-Zorrilla, J., Solano, R., Fernie, A., Vyverman, W., Goossens, A., 2017. The transcription factor bZIP14 regulates the TCA cycle in the diatom Phaeodactylum tricornutum. EMBO J. 36, 1559–1576.

Mier, P., Alanis-Lobato, G., Andrade-Navarro, M. a., 2017. Protein-protein interactions can be predicted using coiled coil co-evolution patterns. J. Theor. Biol. 412, 198–203. doi:10.1016/j.jtbi.2016.11.001

Mock, T., Otillar, R.P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., Ward, B.J., Allen, A.E., Dupont, C.L., Frickenhaus, S., Maumus, F., Veluchamy, A., Wu, T., Barry, K.W., Falciatore, A., Ferrante, M.I., Fortunato, A.E., Glöckner, G., Gruber, A., Hipkin, R., Janech, M.G., Kroth, P.G., Leese, F., Lindquist, E. a., Lyon, B.R., Martin, J., Mayer, C., Parker, M., Quesneville, H., Raymond, J. a., Uhlig, C., Valas, R.E., Valentin, K.U., Worden, A.Z., Armbrust, E.V., Clark, M.D., Bowler, C., Green, B.R., Moulton, V., van Oosterhout, C., Grigoriev, I. V., 2017. Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature 541, 536–540. doi:10.1038/nature20803

Mock, T., Samanta, M.P., Iverson, V., Berthiaume, C., Robison, M., Holtermann, K., Durkin, C. a., BonDurant, S.S., Richmond, K., Rodesch, M., Kallas, T., Huttlin, E.L., Cerrina, F., Sussman, M.R., Armbrust, E.V., 2008. Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. Proc. Natl. Acad. Sci. 105, 1579–1584. doi:10.1073/pnas.0707946105

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M., 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science (80-.). 320, 1344–1349. doi:10.1126/science.1158441.The

Ogata, K., Tahirov, T.H., Ishii, S., 2004. The c-Myb DNA binding domain, in: Frampton, J. (Ed.), Myb Transcription Factors: Their Role in Growth, Differentiation and Disease. Springer Netherlands, pp. 223–238. doi:10.1007/978-1-4020-2869-4_11

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 Physiol. 158, 499–513. doi:10.1104/pp.111.190249

Ording, E., Kvavik, W., Bostad, A., Gabrielsen, O.S., 1994. Two functionally distinct half sites in the DNA-recognition sequence of the Myb oncoprotein. Eur. J. Biochem. 222, 113–120. doi:10.1111/j.1432-1033.1994.tb18848.x

Pickett-Heaps, J., Schmid, A., Edgar, L., 1990. The cell biology of diatom valve formation. Prog. Phycol. Res. 7, 1–68.

Pinto, F.L., Lindblad, P., 2010. A guide for in-house design of template-switch-based 5??? rapid amplification of cDNA ends systems. Anal. Biochem. 397, 227–232. doi:10.1016/j.ab.2009.10.022

Pireyre, M., Burow, M., 2015. Regulation of MYB and bHLH transcription factors: A glance at the protein level. Mol. Plant 8, 378–388. doi:10.1016/j.molp.2014.11.022

Poulsen, N., Kröger, N., 2004. Silica morphogenesis by alternative processing of silaffins in the diatom Thalassiosira pseudonana. J. Biol. Chem. 279, 42993–42999. doi:10.1074/jbc.M407734200

Poulsen, N., Scheffel, A., Sheppard, V.C., Chesley, P.M., Kroger, N., 2013. Pentalysine clusters mediate silica targeting of silaffins in Thalassiosira pseudonana. J. Biol. Chem. 288, 20100–20109. doi:10.1074/jbc.M113.469379

Rado-Trilla, N., Arato, K., Pegueroles, C., Raya, A., De La Luna, S., Mar Alba, M., 2015. Key role of amino acid repeat expansions in the functional diversification of duplicated transcription factors. Mol. Biol. Evol. 32, 2263–2272. doi:10.1093/molbev/msv103

Rayko, E., Maumus, F., Maheswari, U., Jabbari, K., Bowler, C., 2010. Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. New Phytol. 188, 52–66. doi:10.1111/j.1469-8137.2010.03371.x

Richthammer, P., Börmel, M., Brunner, E., Van Pée, K.H., 2011. Biomineralization in Diatoms: The Role of Silacidins. ChemBioChem 12, 1362–1366. doi:10.1002/cbic.201000775

Rogato, A., Richard, H., Sarazin, A., Voss, B., Cheminant Navarro, S., Champeimont, R., Navarro, L., Carbone, A., Hess, W.R., Falciatore, A., 2014. The diversity of small non-coding RNAs in the diatom Phaeodactylum tricornutum. BMC Genomics 15, 698. doi:10.1186/1471-2164-15-698

Rosinski, J. a, Atchley, W.R., 1998. Molecular evolution of the Myb family of transcription factors: evidence for polyphletic origin. J. Mol. Evol. 46, 74–83.

Rubio, V., Linhares, F., Solano, R., Mart'in, A.C., Iglesias, J., Leyva, A., Paz-Ares, J., 2001. A conserved MYB transcription factor involved in phosphate starvation signaling both in vascular plants and in unicellular algae. Genes Dev. 15, 2122–2133. doi:10.1101/gad.204401.availability

Saghbini, M., Hoekstra, D., Gautsch, J., 2001. Media Formulations for Various Two-Hybrid Systems Michael Saghbini, Denise Hoekstra, and Jim Gautsch. Methods Mol. Biol. 177, 15–40.

San-Miguel, T., Perez-Bermudez, P., Gavidia, I., 2013. Production of soluble eukaryotic recombinant proteins in E. coli is favoured in early log-phase cultures induced at low temperature. Springerplus 2, 89. doi:10.1186/2193-1801-2-89

Sapriel, G., Quinet, M., Heijde, M., Jourdren, L., Tanty, V., Luo, G., Le Crom, S., Lopez, P.J., 2009. Genome-wide transcriptome analyses of silicon metabolism in Phaeodactylum tricornutum reveal the multilevel regulation of silicic acid transporters. PLoS One 4. doi:10.1371/journal.pone.0007458

Scheffel, A., Poulsen, N., Shian, S., Kröger, N., 2011. Nanopatterned protein microrings from a diatom that direct silica morphogenesis. Proc. Natl. Acad. Sci. U. S. A. 108, 3175–3180. doi:10.1073/pnas.1012842108

Sherbakova, T. a, Masiukova, I. a, Safonova, T. a, Petrova, D.P., Vereshchagin, a L., Minaeva, T. V, Adel'shin, R. V, Triboĭ, T.I., Stonik, I. V, Aĭzdaĭcher, N. a, Kozlov, M. V, Likhoshvaĭ, E. V, Grachev, M. a, 2005. Conservative motif CMLD in silicic acid transport proteins of diatom algae. M Mol Biol 39, 303–16.

Shrestha, R., Tesson, B., Norden-Krichmar, T., Federowicz, S., Hildebrand, M., Allen, A.E., 2012. Whole transcriptome analysis of the silicon response of the diatom Thalassiosira pseudonana. BMC Genomics 13, 499. doi:10.1186/1471-2164-13-499

Shrestha, R.P., Hildebrand, M., 2015. Evidence for a regulatory role of diatom silicon transporters in cellular silicon responses. Eukaryot. Cell 14, 29. doi:10.1128/EC.00209-14

Stearns, T., Ma, H., Botstein, D., 1990. Manipulating Yeast Genome Using Plasmid Vectors. Methods Enzymol. 185, 280–297.

Strauss, J., 2012. A genomic analysis using RNA - Seq to investigate the adaptation of the psychrophilic diatom Fragilariopsis cylindrus to the polar environment. University of East Anglia.

Sumper, M., Lorenz, S., Brunner, E., 2003. Biomimetic Control of Size in the Polyamine-Directed Formation of Silica Nanospheres. Angew. Chemie - Int. Ed. 42, 5192–5195. doi:10.1002/anie.200352212

Tang, H., Wang, S., Wang, J., Song, M., Xu, M., Zhang, M., Shen, Y., Hou, J., Bao, X., 2016. N-hypermannose glycosylation disruption enhances recombinant protein production by regulating secretory pathway and cell wall integrity in Saccharomyces cerevisiae. Sci. Rep. 6, 25654. doi:10.1038/srep25654

Taniguchi-Yanai, K., Koike, Y., Hasegawa, T., Furuta, Y., Serizawa, M., Ohshima, N., Kato, N., Yanai, K., 2010. Identification and characterization of glucocorticoid receptor-binding sites in the human genome. J. Recept. Signal Transduct. Res. 30, 88–105. doi:10.3109/10799891003614816

Tesson, B., Hildebrand, M., 2013. Characterization and Localization of Insoluble Organic Matrices Associated with Diatom Cell Walls: Insight into Their Roles during Cell Wall Formation. PLoS One 8. doi:10.1371/journal.pone.0061675

Tesson, B., Hildebrand, M., 2010a. Extensive and intimate association of the cytoskeleton with forming silica in diatoms: Control over patterning on the meso- and micro-scale. PLoS One 5. doi:10.1371/journal.pone.0014300

Tesson, B., Hildebrand, M., 2010b. Dynamics of silica cell wall morphogenesis in the diatom Cyclotella cryptica: Substructure formation and the role of microfilaments. J. Struct. Biol. 169, 62–74. doi:10.1016/j.jsb.2009.08.013

Thamatrakoln, K., Alverson, A.J., Hildebrand, M., 2006. Comparative sequence analysis of diatom silicon transporters: Toward a mechanistic model of silicon transport. J. Phycol. 42, 822–834. doi:10.1111/j.1529-8817.2006.00233.x

Thamatrakoln, K., Hildebrand, M., 2008. Silicon Uptake in Diatoms Revisited: A Model for Saturable and Nonsaturable Uptake Kinetics and the Role of Silicon Transporters. Plant Physiol. 146, 1397–1407. doi:10.1104/pp.107.107094

Thamatrakoln, K., Hildebrand, M., 2007. Analysis of Thalassiosira pseudonana silicon transporters indicates distinct regulatory levels and transport activity through the cell cycle. Eukaryot. Cell 6, 271–279. doi:10.1128/EC.00235-06

Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192. doi:10.1093/bib/bbs017

Ugolini, S., Bruschi, C. V., 1996. The red/white colony color assay in the yeast Saccharomyces cerevisiae: Epistatic growth advantage of white ade8-18, ade2 cells over red ade2 cells. Curr. Genet. 30, 485–492. doi:10.1007/s002940050160

Vasina, J. a, Baneyx, F., 1996. Recombinant protein expression at low temperatures under the transcriptional control of the major Escherichia coli cold shock promoter cspA. Appl. Environ. Microbiol. 62, 1444–1447.

Vidal, M., Brachmann, R.K., Fattaey, a, Harlow, E., Boeke, J.D., 1996. Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions. Proc. Natl. Acad. Sci. U. S. A. 93, 10315–10320. doi:10.1073/pnas.93.19.10315

Vrieling, E.G., Gieskes, W.W.C., Beelen, T.P.M., 1999. SILICON DEPOSITION IN DIATOMS: CONTROL BY THE pH INSIDE THE SILICON DEPOSITION VESICLE. J. Phycol. 35, 548–559. doi:10.1046/j.1529-8817.1999.3530548.x

Vrieling, E.G., Sun, Q., Tian, M., Kooyman, P.J., Gieskes, W.W.C., van Santen, R. a, Sommerdijk, N. a J.M., 2007. Salinity-dependent diatom biosilicification implies an important role of external ionic strength. Proc. Natl. Acad. Sci. U. S. A. 104, 10441–10446. doi:10.1073/pnas.0608980104

Wang, Y., Wang, F., Wang, R., Zhao, P., Xia, Q., 2015. 2A self-cleaving peptide-based multi-gene expression system in the silkworm Bombyx mori. Sci. Rep. 5, 16273. doi:10.1038/srep16273

Wenzl, S., Hett, R., Richthammer, P., Sumper, M., 2008. Silacidins: Highly acidic phosphopeptides from diatom shells assist in silica precipitation in vitro. Angew. Chemie - Int. Ed. 47, 1729–1732. doi:10.1002/anie.200704994

Williams and Grotewold, E., C.E., 1997. Differences between plant and animal Myb domains are fundamental for DNA-binding, and chimeric Myb domains have novel DNA-binding specificities. . J. Biol. Chem. 272, 563–571.

Yan, J., Burgess, S.M., 2012. Using a Yeast Inverse One-Hybrid System to Identify Functional Binding Sites of Transcription Factors. Methods Mol. Biol. 786, 275–290. doi:10.1007/978-1-61779-292-2_17

Yanai, K., 2013. A Modified Yeast One-Hybrid System for Genome-Wide Identification of Transcription Factor Binding Sites. Methods Mol. Biol. 977, 125–136. doi:10.1007/978-1-62703-239-1_1

Zheng, W., Chung, L.M., Zhao, H., 2011. Bias detection and correction in RNA-Sequencing data. BMC Bioinformatics 12, 290. doi:10.1186/1471-2105-12-290

Zhu, Z.X., Yu, Z.M., Taylor, J.L., Wu, Y.H., Ni, J., 2016. The application of yeast hybrid systems in protein interaction analysis. Mol Biol 50, 751–759. doi:10.7868/S0026898416050189

Chapter 5: Summary

Transformation chapter

Key findings

- Stable nuclear transformations of only 12 species have been published. All of these species are temperate, the majority are marine, and all can be grouped into 6 orders across 2 classes. In short, the current range of transformable diatom species gives a poor representation of their geographical and biological diversity.
- *F. cylindrus* is an ecologically important, psychrophilic, pennate diatom found at both poles and in seasonally cold marine waters.
- In this chapter a transformation system for this model species was developed. This is the first transformation system for a polar microalga and may be the first for any psychrophilic eukaryote.
- The fucoxanthin chlorophyll a/c binding protein (FCP; ID 267576) was found to be the highest expressed gene in *F. cylindrus* across a range of growth conditions, based on previously produced RNA-seq data (Mock et al., 2017).
- A construct was developed using Gibson assembly to express 2 separate cassettes: the shble gene for zeocin resistance and egfp with a human codon bias, both under the control of the endogenous FCP promoter and terminated by the FCP terminator.
- Microparticle bombardment was used to introduce the construct into cells.
- The microparticle transformation system was adapted from *Pheodactylum tricornutum* (Kroth, 2007) for use with a polar species. Key changes involved filtering cells and shooting on a filter (on top of agar) rather than drying at room temperature and ensuring that all steps of the protocol involving cells, were carried out on ice or at 4°C.
- The pressure under which the 0.7µm tungsten particles, coated in construct, was introduced into the cells was optimised. A pressure of 1550psi gave the highest transformation efficiency at 30 colonies/10^8 cells.
- Sensitivity of *F. cylindrus* to zeocin was tested. It was found that zeocin prevents growth from at least 50µg/ml. Full salinity or half salinity media does not appear to affect zeocin sensitivity in this species. Transformation with FCP:shble successfully provides resistance against 100µg/ml zeocin, both on plates and in liquid cultures.
- Colonies took longer to appear on plates compared to temperate species with 3-5 weeks required for transformants to become apparent.
- Presence of both shble and egfp was observed from gDNA of transformed cultures. 60% of colonies with shble also contained egfp which supports previous findings from other diatom transformation systems.

- Fluorescence from egfp could be seen in all cultures with the egfp gene. Flow cytometry results showed a clear increase in fluorescence in the green channel and epifluorescence microscopy showed a highly visible signal localised to the cytosol.
- The shble and egfp genes appear to be stably transformed, and are still present in mutant cell lines 2 years after transformation.

Concluding remarks

Development of a stable transformation system for the first psychrophilic microalga has been achieved. A functional promoter, terminator, selective marker and reporter gene have been successfully identified and tested, along with delivery of transgenes via microparticle bombardment. As a key ecological species in polar regions and a model organism for eukaryotic psychrophiles, this is an important tool for reverse genetics to understand key biological functions such as adaptations to an extreme polar environment. It may also be a useful tool for biotechnology given that preferential solubility, folding, yield and stability of recombinant proteins can be observed under cold conditions (San-Miguel et al., 2013; Vasina and Baneyx, 1996), especially when producing proteins of a eukaryotic origin, as not all post-translational modifications can be achieved when using a prokaryotic system (Demain and Vaishnav, 2009).

Future work

- Although transformation efficiency sits within the range expected for diatom transformation using microparticle bombardment, some studies see a much higher number of transformant colonies using this method (Buhmann et al., 2014; Poulsen et al., 2006). In terms of parameters, only pressure has been optimised and it may be worth trying alternative settings, particle sizes and recovery conditions to boost efficiency.
- Alternative methods could also increase efficiency such as electroporation or bacterial conjugation (Karas et al., 2015; Miyahara et al., 2013). For the latter a suitable bacterial host would be required to transform and transfer the conjugative plasmid. In diatoms *E. coli* is used – this is also the case when transforming psychrophilic bacteria (Miyake et al., 2007) but incubation is carried out at lower temperatures for a longer period of time. However, the temperature used (18°C) is still 10°C higher than the maximum used for culturing *F. cylindrus*. Either trials at lower temperatures would be needed or alternative psychrophilic bacteria would need to be found.
- In the above methods whole, circular plasmids can be transformed which would allow exploration with self-replicating episomes, potentially bypassing the need for integration, increasing the number of transformants with both cassettes and increasing efficiency (Karas et al., 2015).
- A larger toolbox in terms of promoters, selective markers and reporter genes would be useful. Inducible promoters as well as promoters with different expression levels would allow a greater control over transgenes.

- Identification of targeting signals would be valuable for localisation to the various organelles, cell membrane and cell wall.

CRISPR-Cas chapter

Key findings

- CRISPR-Cas is a recent and powerful tool which allows programmable genome editing by inducing double strand breaks, leading to subsequent mutations following repair by non-homologous end-joining.
- Two constructs were made using the flexible, modular Golden-gate cloning approach to target two sites in the urease gene in *Thalassiosira pseudonana*, to induce a deletion.
- Modules for each component including promoters, genes, terminators, sgRNAs and CEN-ARS-HIS, were created and assembled into cassettes. The cassettes were then combined together in a single construct. Both of the final constructs contained a NAT cassette, Cas9 cassette and two sgRNA cassettes. One construct also contained a CEN-ARS-HIS sequence for replication in diatoms.
- The U6 promoter was used to express each sgRNA. The *T. pseudonana* genome was blasted to find the U6 promoter and 5' template switching oligo RACE was performed to elucidate its exact end.
- Twenty nt targets for sgRNAs were designed to cut 37nt apart and over a restriction site, for screening by the band shift assay and restriction site loss assay, respectively. All targets began with a G residue for expression by RNA polymerase III, recruited by the U6 promoter.
- Following transformation, presence of Cas9 was seen in 4/33 colonies. Each of these 4 colonies displayed mutations. Band shift assays gave clear confirmation of deletions for 3 of the colonies. Sequencing highlighted a 4nt deletion at one of the sgRNAs in the 4th colony.
- Colonies with deletions were mosaic, with cells displaying a mixture of the WT and edited urease gene. Sub-cloning was required to isolate colonies with the edited urease in both alleles (bi-allelic). A high percentage of bi-allelic colonies up to 61.5% was observed. Urease in two thirds of sub-clones with bi-allelic deletions was repaired by 'clean' non-homologous end-joining, resulting in only the 37nt deletion and no additional indels. One third of sub-clones showed additional 1nt deletions and substitutions at the cut site.
- Urease is required for growth in urea as the sole nitrogen source. Growth experiments on mutant and WT cell lines were carried out to observe growth in urea and nitrate. All cultures grew well in nitrate, with no clear difference observed between 'knock-out' and WT cell-lines. Bi-allelic mutant cell-lines did grow in urea, but at a significantly slower rate. Cells were also much smaller, which is an indicator of nitrogen limitation. The deletion occurs early on in the urease gene and is expected to create a frame-shift,

knocking-out the protein. Based on growth results, the urease protein may be expressed from an alternative translation start site. From the position of the deletion and potential in-frame ATGs, this is likely to include the alpha sub-unit, which contains the active site, along with the beta sub-unit, but is expected to truncate the gamma sub-unit, potentially reducing the activity of the urease gene and accounting for sub-optimal growth.

- The CEN-ARS-HIS sequence, does not appear to result in autonomous replication of the CRISPR-Cas plasmid. This is likely due to fragmentation of the plasmid during microparticle bombardment.
- A construct to knock out the silacidin gene in *T. pseudonana* was also designed along with sgRNAs for editing of the SITMyb gene in *F. cylindrus*.

Concluding remarks

CRISPR-Cas has been used in *T. pseudonana* to successfully edit the urease gene. The Golden-gate system works well, with modules now publicly available and being used within our group and by other groups for gene editing of alternative genes and diatom species. One of the main limitations with the current system is the need to integrate a large Cas9 cassette, resulting in only 12% of clones containing this gene. However, once the Cas9 is present, CRISPR-cas with the current sgRNAs appears to be very efficient, leading to 100% of primary clones with a mutation. Sub-cloning is important to isolate colonies with bi-allelic deletions, however these occur very frequently and screening for deletions via the band-shift assay is quick and effective. As editing of the urease gene led to a nitrogen-limited phenotype, rather than a nitrogen starved phenotype, future target sites needs to be carefully considered. It may be more effective to target/remove the active site, or even the whole gene given that large deletions have been achieved in other systems (Ordon et al., 2016; Zheng et al., 2014).

Future work

- If expression of Cas9 is not dependent on integration into the genome, then it may be possible to increase occurrence of mutations in primary clones. One method to do this would be to introduce the necessary CRISPR-Cas components on a self-replicating episome. This has the additional benefit of being able to expel the episome (Karas et al., 2015) after mutations have been induced, by removing selection, which would be beneficial if off-target activity from long term Cas9 exposure occurs. This would require a method such as electroporation or bacterial conjugation to introduce whole, circular plasmids.
- Alternatively a Cas9 transgenic cell-line could be produced so that only the small sgRNA cassette would need to be introduced. This would require either an effective secondary antibiotic for selection or a means to introduce the Cas9 without a secondary selective marker. This could potentially be achieved by introducing one set of CRISPR sequences, i.e. Cas9 cassette, U6:sgRNA and selective marker on an episome and inducing a double

strand break. A plasmid could be co-transformed with a separate donor Cas9 cassette for homologous recombination (HR) at the cut site, introducing the cassette into the genome. Selection could then be removed, leading to expulsion of the episome and leaving just the Cas9 integrated by HR.

- HR using CRISPR-Cas is currently being carried out in the Mock lab with promising results so far. It may be possible to use geminiviral vectors (Yin et al., 2015) to introduce donor sequences for repair by HR. Vectors are introduced in a linear state which then circularise and replicate, increasing the chance of a donor sequence being used for repair.
- With only two examples of CRISPR-Cas working in diatoms, including this one, there is currently no information on off-target mutations. It would be interesting to look at predicted off-target sites at various intervals following transformation to see if, and how quickly CRISPR-Cas induces mutations. This may also be dependent on the specific sgRNA sequence used.
- To remove the need to sub-clone, it could be worth optimising the incubation time following transformation, prior to plating on selective media. The incubation time would need to be long enough to allow the mutation to 'settle' but short enough to prevent too many replica clones.
- Very large deletions have been achieved using CRISPR-Cas in other systems (Feng et al., 2013; Ordon et al., 2016). It would be interesting to see how large a section can efficiency be removed from diatoms. This would be useful information, particularly if a gene model or active site is uncertain, in which case the entire gene could potentially be removed.
- As a modular system, Golden-gate cloning lends itself to targeting multiple genes at once (Sakuma et al., 2014). This could potentially allow partial or whole pathways to be targeted. Whilst multiple U6:sgRNA cassettes can be assembled into a single construct, it may be more efficient to try a CRISPR array based approach (Cong et al., 2013), in which the tracrRNA is separately expressed to the pre-crRNA array which contains the guide target sequences interspersed by direct CRISPR repeats. Both are expressed by a U6 promoter and are processed by and annealed using common eukaryotic enzymes.
- This could be combined with a slightly different Golden-gate approach to the current system. Golden-gate vectors exist that allow assembly into the level two backbone, with a module for future insertions. The modules developed in this chapter are currently being used to assemble the FCP:NAT, FCP:Cas9 cassettes into a level two vector along with one of these modules. The U6:sgRNAs are then added later (Bermejo Martinez, unpublished). It could be a good idea to take this further and assemble the selective marker, Cas9 cassette, U6:tracrRNA and another U6 promoter directly upstream of a 'insert module' into a level 2 vector. Pre-crRNA arrays could then be introduced at a later date under the control of the single U6 promoter. This would allow cloning of a single, much shorter fragment which may increase cloning efficiency, compared to assembling multiple

U6:sgRNA cassettes. It would also allow anyone wanting to knock-out a gene or a set of genes in this species, to simply clone a synthesised pre-crRNA array directly into a single vector. It would of course be necessary to check the array based method is functional first.

- Now that a system has been established, several adapted versions of CRISPR-Cas can be pursued. These include using Cas9 nickases to reduce off-target (Cong 2013), using dCas9 as a binding protein to activate or repress transcription (Qi 2013, Piatek et al 2015), inducing DNA modifications such as methylation (Vojta et al. 2016) or fluorescently tagging specific sequences (Deng 2015).

SITMyb chapter

Key findings – In silico modelling

- Sequencing of the *F. cylindrus* genome (Mock et al., 2017) led to the discovery of a novel gene predicted to have both a Myb domain and a domain with homology to silicon transporters.
- Two alleles of this gene are present, with the second half of the gene, including the SIT and Myb domains, showing a high degree of conservation.
- The majority of the gene shows no homology to existing proteins, however besides the SIT and Myb domains, 3 short regions show homology to other proteins. This includes a hypothetical protein from a silicifying choanoflagellate and two regions with similarities to alveolate proteins potentially linked to either transcription or silicon transport.
- Blast searches and modelling of the Myb domain strongly support its helix-turn-helix structure and homology to DNA binding proteins. Myb domains are also closely related to SANT domains which regulate expression through protein-protein interactions with histones or regulatory proteins (Aasland et al., 1996; Iyer et al., 2008). The SITMyb protein also contains coiled-coil motifs which are often involved in protein-protein interactions or transcription (Mason and Arndt, 2004; Mier et al., 2017). Modelling therefore supports the potential of the SITMyb gene as a regulatory protein.
- The SIT domain aligns with the C-terminal sequence of several diatom SITs after the 10th transmembrane domain and contains a coiled-coil motif, which is expected to be located intracellularly. The closest alignment is to a SIT from *F. cylindrus*.
- GXQ motifs, linked to silicic acid binding in SITs (Thamatrakoln et al., 2006) are present, in regions of no known homology, at the C-terminal end of the SITMyb gene.
- A homopolymer of proline residues is found towards the start of the gene, which has been previously linked to transcription activation (Gerber et al., 1994)
- Several nuclear localisation signals which are required by TFs for transport to the nucleus, are present.

Concluding remarks– In silico modelling

There are several very interesting sequences present in this gene with key domains/motifs such as SIT, Myb, coiled-coils and NLS signals conserved between both alleles. Structure of this gene points towards binding activity linked to regulation whilst the presence of the SIT domain suggests that it may be performing an additional function. Literature surrounding gene regulation in diatoms is far from comprehensive, however expression analyses show that several genes respond to changes in nutrient concentration. This includes response to silica starvation, with many genes linked to silica metabolism and frustule formation being differentially expressed (Mock et al., 2008; Shrestha et al., 2012). This indicates that a regulatory mechanism is involved. Transcription factors are almost certainly involved given that they are a core component of gene regulation in eukaryotes and signalling pathways are also likely to be involved. SITs are currently the only proteins in diatoms known to bind silicic acid (Curnow et al., 2012; Hildebrand et al., 1997) and there is evidence that some SITs may be acting as sensors for frustule formation/cell cycle progression (Shrestha and Hildebrand, 2015). The occurrence of a protein with a Myb domain linked to regulation of transcription through DNA/protein binding and a SIT domain which could be involved in silicon/protein binding is very exciting and adds weight for further exploration of this gene as a silicon-linked regulator. It is possible that further examination of this gene could lead to identification of a larger regulatory network.

Key findings – In vitro modelling

- 5’ and 3’ RACE experiments did not give a conclusive transcript start or end, with varied products ranging from the predicted 5’ UTR through to the 3’ UTR.
- RACE results and amplification of overlapping fragments from cDNA support presence of the full transcript, however amplification of the full gene, either from gDNA or cDNA was not possible.
- RT-PCR, sequencing and previously generated RNA-seq data supported the intron-exon model from JGI. The 3’ end is well modelled due to excellent RNA-seq coverage, however the 5’ end is poorly covered and along with poor and inconsistent 5’ RACE results, an exact transcription start site remains elusive.

Concluding remarks– In vitro modelling

Due to uncertainty in the start site, poor coverage at the 5’ end and difficulties with amplifying the full gene, a slightly shorter sequence compared to the original JGI model was chosen for overexpression in *F. cylindrus* and yeast. An in-frame ATG after the first intron was chosen to start the sequence, which resulted in the highly conserved region between alleles being selected including the SIT and Myb domains.

Key findings – Overexpression of SITMyb in *F. cylindrus*

- A construct was produced for overexpression of SITMyb, based on in-vitro modelling, under the control of a promoter from the highly expressed gene fucoxanthin chlorophyll a/c binding protein. A 6x His-tag was included at the C-terminus for purification and labelling.
- Two clones were isolated with the SITMyb cassette.
- RT-PCR demonstrated that the gene was transcribed.
- However, no protein was observed for SITMyb during western blots using the His-tag.
- This may be due to problems with the His-tag, as later expression in yeast gave a protein at the correct size when using a HA-tag. Using the His-tag in the yeast system also gave no results.

Concluding remarks – Overexpression of SITMyb in *F. cylindrus*

The SITMyb cassette was successfully introduced and expressed as a transcript in *F. cylindrus*. However no protein was observed. This may be due to a non-functional or in accessible His-tag rather than a lack of protein.

Key findings – Yeast-1-Hybrid

- A Yeast-1-hybrid (Y1H) *F.cylindrus* gDNA library with a 6x coverage has been constructed for elucidating the binding sites of TFs in this species.
- Two Y1H constructs to overexpress SITMyb and the Myb domain in yeast have been produced. Both constructs produce a protein of the expected size when transformed into yeast suggesting that SITMyb and Myb are expressed.
- Loss of plasmids, containing the TF, in overexpression lines is observed, possibly due to problems with selection. This has implications on the efficiency of the final method and numbers of true positives.
- Only clones with the SITMyb TF were successfully mated with the gDNA library. Mating efficiency of Myb domain-containing clones was poor and produced no colonies following transfer to –uracil plates.
- The majority of colonies appear to be false positives with just 1 colony out of 200 with the right circumstances to be true positive. The most likely candidate for false positives is binding of endogenous yeast transcription factors to domains in the gDNA library, although other factors may also produce false negative results.

Concluding remarks – Yeast-1-Hybrid

Progress has been made in this method, but it is not yet fully functional for *F. cylindrus*. Production of the gDNA library with a good coverage is a useful tool, not just for elucidating potential binding sites of SITMyb, but also for other *F. cylindrus* transcription factors. Although the SITMyb gene appears to be expressed, there are issues with loss of the overexpression construct which will reduce efficiency and may be responsible for the low number of potential true positives observed. Further work is needed to optimise this method and tackle the high number of false positives

compared to potential true positives. This includes re-examination of the Myb domain using the optimised method. It is also possible that a longer version of the SITMyb gene is required for binding. It may be that SITMyb binds protein rather than DNA, given that Myb and SANT domains have similar homology. If this is the case then Yeast-2-hybrid (Y2H) would be a more appropriate method. Y2H could also be used to determine if the SIT domain binds proteins, which would require a further overexpression construct with this domain.

Future work

- Further empirical testing is needed to determine the function of the SITMyb gene and its domains.
- DNA binding needs to be assessed for the current gene model, the Myb domain and the full JGI predicted gene model. In order to do this, the Yeast-1-hybrid system needs to be optimised. The most pertinent features of this method to be addressed, are increasing selective pressure on transformants to maintain the overexpression construct and reducing false positives, likely caused by binding of endogenous yeast TF, by more vigorous screening. It may also be worth directly looking at binding of allele 250586. This would need the gap in the JGI sequence to be amplified and sequenced.
- If Yeast-1-hybrid cannot be successfully applied, then alternatives such as ChIP seq are available.
- Now that proteins for SITMyb and Myb appear to be expressed in yeast, it should be possible to purify the recombinant proteins via the HA-tag and test for DNA and silicon binding. Although, purification from an *F. cylindrus* overexpression line, grown at lower temperatures, may yield a more accurately folded protein, given the psychrophilic nature of *F. cylindrus*.
- Mobility shift assays could help to determine DNA binding of proteins. Genes known to be involved in silica metabolism could be tested as could genes identified by yeast-1-hybrid. This could also help to validate results from the later. Shift assays could also be performed in the presence of silicic acid to see if this affects binding activity.
- Silicon binding could be determined using protein Crystallography.
- The overexpression construct for *F. cylindrus* needs to be reconsidered. Although the transcript appeared to be expressed, no protein was observed. As with Y1H constructs, plasmids with both the current gene model and longer JGI predicted model need to be created. An alternative to the C-terminal His-tag is needed to purify and probe any overexpressed protein. A C-terminal HA-tag is an option, given that it works well for yeast overexpression lines. It may also be worth adding egfp as a fusion, for a more direct indication of overexpression. Worries about solubility/activity of a larger fusion protein could be addressed by adding a cleavage domain at the fusion junction (Wang et al., 2015). An egfp-tagged SITMyb (without a cleavage domain) could also be used to determine localisation of the SITMyb, as a TF would be expected in the nucleus.

- If evidence of the overexpressed SITMyb gene can be found, then cell-lines can be phenotyped to examine potential changes in the frustule. This could include examining morphology at different scales, as well as observing levels of precipitated silica.
- It would also be possible to carryout RNA-seq to examine any changes in expression levels, associated with a higher concentration of SITMyb.
- Yeast-2-hybrid could help to determine any potential protein-protein interactions. This could be carried out for the full SITMyb gene as well as both the Myb and SIT domains which have features potentially associated with protein binding.
- Finally CRISPR-Cas of SITMyb is ongoing in the lab and if mutations can be produced, cell-lines will need to be phenotyped and RNA-seq carried out.

References

Aasland, R., Stewart, A.F., Gibson, T., 1996. The SANT domain: a putative DNA-binding domain in the SWI-SNF and ADA complexes, the transcriptional co-repressor N-Cor and TFIIB. rends Biochem. Sci. 21, 87–88.

Buhmann, M.T., Poulsen, N., Klemm, J., Kennedy, M.R., Sherrill, C.D., Kröger, N., 2014. A tyrosine-rich cell surface protein in the diatom Amphora coffeaeformis identified through transcriptome analysis and genetic transformation. PLoS One 9. doi:10.1371/journal.pone.0110369

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., Zhang, F., 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science (80-.). 339, 819–823. doi:10.1126/science.1231143

Curnow, P., Senior, L., Knight, M.J., Thamatrakoln, K., Hildebrand, M., Booth, P.J., 2012. Expression, purification, and reconstitution of a diatom silicon transporter. Biochemistry 51, 3776–3785. doi:10.1021/bi3000484

Demain, A.L., Vaishnav, P., 2009. Production of recombinant proteins by microbes and higher organisms. Biotechnol. Adv. 27, 297–306. doi:10.1016/j.biotechadv.2009.01.008

Feng, Z., Zhang, B., Ding, W., Liu, X., Yang, D.-L., Wei, P., Cao, F., Zhu, S., Zhang, F., Mao, Y., Zhu, J.-K., 2013. Efficient genome editing in plants using a CRISPR/Cas system. Cell Res. 23, 1229–1232. doi:10.1038/cr.2013.114

Gerber, H.P., Seipel, K., Georgiev, O., Höfferer, M., Hug, M., Rusconi, S., Schaffner, W., 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263, 808–11. doi:10.1126/science.8303297

Hildebrand, M., Volcani, B.E., Gassmann, W., Schroeder, J.I., 1997. A gene family of silicon transporters. Nature 385, 689.

Iyer, L.M., Anantharaman, V., Wolf, M.Y., Aravind, L., 2008. Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. Int. J. Parasitol. 38, 1–31. doi:10.1016/j.ijpara.2007.07.018

Karas, B.J., Diner, R.E., Lefebvre, S.C., McQuaid, J., Phillips, A.P.R., Noddings, C.M., Brunson, J.K., Valas, R.E., Deerinck, T.J., Jablanovic, J., Gillard, J.T.F., Beeri, K., Ellisman, M.H., Glass, J.I., Hutchison III, C. a., Smith, H.O., Venter, J.C., Allen, A.E., Dupont, C.L., Weyman, P.D., 2015. Designer diatom episomes delivered by bacterial conjugation. Nat. Commun. 6, 6925. doi:10.1038/ncomms7925

Kroth, P.G., 2007. Genetic Transformation A Tool to Study Protein Targeting in Diatoms. Methods Mol. Biol. 390, 257–267.

Mason, J.M., Arndt, K.M., 2004. Coiled coil domains: Stability, specificity, and biological implications. ChemBioChem 5, 170–176. doi:10.1002/cbic.200300781

Mier, P., Alanis-Lobato, G., Andrade-Navarro, M. a., 2017. Protein-protein interactions can be predicted using coiled coil co-evolution patterns. J. Theor. Biol. 412, 198–203. doi:10.1016/j.jtbi.2016.11.001

Miyahara, M., Aoi, M., Inoue-Kashino, N., Kashino, Y., Ifuku, K., 2013. Highly Efficient Transformation of the Diatom Phaeodactylum tricornutum by Multi-Pulse Electroporation. Biosci. Biotechnol. Biochem. 77, 874–876. doi:10.1271/bbb.120936

Miyake, R., Kawamoto, J., Wei, Y.L., Kitagawa, M., Kato, I., Kurihara, T., Esaki, N., 2007. Construction of a low-temperature protein expression system using a cold-adapted bacterium, Shewanella sp. strain Ac10, as the host. Appl. Environ. Microbiol. 73, 4849–4856. doi:10.1128/AEM.00824-07

Mock, T., Otillar, R.P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., Ward, B.J., Allen, A.E., Dupont, C.L., Frickenhaus, S., Maumus, F., Veluchamy, A., Wu, T., Barry, K.W., Falciatore, A., Ferrante, M.I., Fortunato, A.E., Glöckner, G., Gruber, A., Hipkin, R., Janech, M.G., Kroth, P.G., Leese, F., Lindquist, E. a., Lyon, B.R., Martin, J., Mayer, C., Parker, M., Quesneville, H., Raymond, J. a., Uhlig, C., Valas, R.E., Valentin, K.U., Worden, A.Z., Armbrust, E.V., Clark, M.D., Bowler, C., Green, B.R., Moulton, V., van Oosterhout, C., Grigoriev, I. V., 2017. Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature 541, 536–540. doi:10.1038/nature20803

Mock, T., Samanta, M.P., Iverson, V., Berthiaume, C., Robison, M., Holtermann, K., Durkin, C. a., BonDurant, S.S., Richmond, K., Rodesch, M., Kallas, T., Huttlin, E.L., Cerrina, F., Sussman, M.R., Armbrust, E.V., 2008. Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. Proc. Natl. Acad. Sci. 105, 1579–1584. doi:10.1073/pnas.0707946105

Ordon, J., Gantner, J., Kemna, J., Schwalgun, L., Reschke, M., Streubel, J., Boch, J., Stuttmann, J., 2016. Generation of chromosomal deletions in dicotyledonous plants employing a user-friendly genome editing toolkit. Plant J. 1–14. doi:10.1111/tpj.13319

Piatek, A., Ali, Z., Baazim, H., Li, L., Abulfaraj, A., Al-Shareef, S., Aouida, M. and Mahfouz, M.M., 2015. RNA-guided transcriptional regulation in planta via synthetic dCas9-based transcription factors. Plant biotechnology journal. 13(4), pp.578-589. doi: 10.1111/pbi.12284

Poulsen, N., Chesley, P.M., Kröger, N., 2006. Molecular genetic manipulation of the diatom Thalassiosira pseudonana (Bacillariophyceae). J. Phycol. 42, 1059–1065. doi:10.1111/j.1529-8817.2006.00269.x

Sakuma, T., Nishikawa, A., Kume, S., Chayama, K., Yamamoto, T., 2014. Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system. Sci. Rep. 4, 5400. doi:10.1038/srep05400

San-Miguel, T., Perez-Bermudez, P., Gavidia, I., 2013. Production of soluble eukaryotic recombinant proteins in E. coli is favoured in early log-phase cultures induced at low temperature. Springerplus 2, 89. doi:10.1186/2193-1801-2-89

Shrestha, R., Tesson, B., Norden-Krichmar, T., Federowicz, S., Hildebrand, M., Allen, A.E., 2012. Whole transcriptome analysis of the silicon response of the diatom Thalassiosira pseudonana. BMC Genomics 13, 499. doi:10.1186/1471-2164-13-499

Shrestha, R.P., Hildebrand, M., 2015. Evidence for a regulatory role of diatom silicon transporters in cellular silicon responses. Eukaryot. Cell 14, 29. doi:10.1128/EC.00209-14

Thamatrakoln, K., Alverson, A.J., Hildebrand, M., 2006. Comparative sequence analysis of diatom silicon transporters: Toward a mechanistic model of silicon transport. J. Phycol. 42, 822–834. doi:10.1111/j.1529-8817.2006.00233.x

Vasina, J. a, Baneyx, F., 1996. Recombinant protein expression at low temperatures under the transcriptional control of the major Escherichia coli cold shock promoter cspA. Appl. Environ. Microbiol. 62, 1444–1447.

Vojta, A., Dobrinic, P., Tadic, V., Bockor, L., Korac, P., Julg, B., Klasic, M., Zoldos, V., 2016. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. Nucleic Acids Res. 44, 5615–5628. doi:10.1093/nar/gkw159

Wang, Y., Wang, F., Wang, R., Zhao, P., Xia, Q., 2015. 2A self-cleaving peptide-based multi-gene expression system in the silkworm Bombyx mori. Sci. Rep. 5, 16273. doi:10.1038/srep16273

Yin, K., Han, T., Liu, G., Chen, T., Wang, Y., Yu, A.Y.L., Liu, Y., 2015. A geminivirus-based guide RNA delivery system for CRISPR/Cas9 mediated plant genome editing. Sci. Rep. 5, 14926. doi:10.1038/srep14926

Zheng, Q., Cai, X., Tan, M.H., Schaffert, S., Arnold, C.P., Gong, X., Chen, C.Z., Huang, S., 2014. Precise gene deletion and replacement using the CRISPR/Cas9 system in human cells. Biotechniques 57, 115–124. doi:10.2144/000114196

List of Abbreviations

Below is a list of abbreviations used throughout this thesis with their meanings.

Abbreviation	Meaning
Ade	Adenine
ARS	Autonomous replicating sequence
BC	Bacterial conjugation
bZIP	Basic leucine zippers
C-A-H	CEN-ARS-HIS
cDNA	Complementary DNA
CEN	Centromeric sequence
CIP	Calf intestine alkaline phosphatase
CRISPR	Clustered regularly interspersed short palindromic repeats
crRNA	CRISPR RNA
dCas9	Deactivated Cas9
DSB	Double strand break
dsCYC	Diatom specific cyclin
EF2	Elongation factor 2
Egfp	Enhanced green fluorescent protein
EST	Expressed sequence tag
Eyfp	Enhanced yellow fluorescent protein
FCP	Fucoxanthin chlorophyll a/c binding protein
GA	Gibson assembly
gDNA	Genomic DNA
GS	Gene specific
GUS	β-glucuronidase
H4	Histone 4
His	Histidine
HSF	Heat shock factors
IGV	Integrated Genomics Viewer
JGI	Joint genome institute
LCPA	Long chain polyamines
Luc	Luciferase
MPB	Microparticle bombardment
MPE	Multi-pulse electroporation
NAT	Nourseothricin N-acetyl transferase
NEB	New England Biolabs
NLS	Nuclear localisation signal
OE	Overexpression
oriT	Origin of transfer
PCR	Polymerase chain reaction
pol III	RNA polymerase III
qPCR	Quantitative PCR
RACE	Rapid amplification of cDNA ends
RLM	RNA-ligase mediated
RT	Reverse transcription

SDM	Site directed mutagenesis
SDV	Silica deposition vesicle
sgRNA	Single guide RNA
SIT	Silicon transporter
snRNA	Small non-coding RNA
sRNA	Small RNA
SSIII	Superscript III
TF	Transcription factor
TFBS	Transcription factor binding site
tracrRNA	Trans-activating crRNA
Trp	Tryptophan
TSO	Template switching oligo
TSS	Transcription start site
Ura	Uracil
WT	Wildtype
Y1H	Yeast-1-hybrid

Appendix

Appendix figure 1. Alignment of SITMyb alleles 233781 and 250586. Text highlighted in green indicates the SIT domain, blue indicates the Myb domain and the ATG highlighted in red shows the start site of SITMyb sequence for overexpression.

```
233781          ATGCCTAGAGTTGCTAAGAACTTTGATAAGTATCTGCTGACAAAGGAAGTACCTGATTCA
250586          ATGCCGAAAGTTGCGAAGAACTTTGATCAGTATATGCTGACAAAGGAAGTACCTGATTCA
                ***** *.****** ************ ***** **************************

233781          CCAAACGGTACAGGAATTACCAAGAAGAATATAGAACATATCCGAGTACAATGGGATTAT
250586          CCAAACGGTACAGGAATTACCAAGAAGAATATCGAACATATCCGAGTACAATGGGATTAT
                ******************************** ***************************

233781          GAATTAGAAGATACAACAAGAGTAATCTATCATCGATTTAGAAATCAAACAGAGTATGAA
250586          GCATTAGAGGATACAACAAAAGTAATCTATCATAGATTTAGAAATCAAACAGAGTATGAA
                * ******.**********.************* **************************

233781          AAATATCATAATAATAAGGTCAGTCAAGGTGTTCGAAAAAGAAATGGATTATATGAGACT
250586          AAATATCATAATAATAAGGTCAGTCAAGGTGTTCGAAGAAGAAATGGATTATATGAGATT
                ************************************.********************.*

233781          GGACAAAAATTAAGAAAAAAGAAGAAAAAAGAGGATGATGACGAAAACATGACGACGGAT
250586          GGACAAAAATTAAGAAAA---AAGAAAAAAGAGGATGATAATGAAAACAGGACTACGGAT
                ******************   ******************.*.******* *** ******

233781          GGTACTACTACTACTACGACTAAAATCGGATCGACGGTGGCGGCGGT------GGCT
250586          GG---TACTACTACTACGACTAAAAACGGATCGGTGGTGGCAGTGGCTGTGGAGGCG
                **   ********************** *******..******.*.**.      ***

233781          GTGGCTGTGGCGGAAAAGCCTTCATCAACATCTTCATTATTGTCATTGAAGAAAAAGGCA
250586          GAGGCGGAGGCGGAGAAGCCTTCATCAACATCTTC---ATCGTCATCGAAGAAAATGGCA
                * *** * ******.********************   **.*****.******** ****

233781          TCGGCATTGGCATCAGCGGCGGTTGGTAC------TACCACGATGAC---GGACAGCAAT
250586          TCGGCATCGGCATCAGCGGCGGTTGGTACTACTAGTACCACGATGACTAGTGACAACAAT
                *******.**********************      ************    ****.****

233781          GATAA------TGCGGGTGTATCAA---CATACGAGGATGGTATTGATGCTGGTCTGCCG
250586          GATAATGATACTGTGGGTGTATCAACATCATACGATGATGGTATTGATGCTGGTTCGCTG
                *****      **.***********   ******* ******************..**.*

233781          TCCACTACCTCCTACCATTGTCAACCAATTGTTGGGACAAGACTTGTTATTACTCCCTCT
250586          TCCACTACCACTTCCCATTGTCGACCAATTGTTGGGACAAGAGTTGGTATTACTGCCTCT
                ********* *.* ********.****************** *** ******* *****

233781          AATATTCCTCCTCCTCCTTCTCCTCCTCTTCAGAAAGAACATGATACCATGGAAGCAACA
250586          AATCTTCCTCCTCCTTC------------TCAAAAAGAACATGATAACATGGAAGCAACA
                *** ***********.*            ***.************* *************

233781          ACCACAGGATCTCCTCCTCCTCCTCCTCCTCTAGGGAAGACTATCATTCAGACCACA
250586          ACCACAGGAT---CTCCTCCTCCTCTTCCTCCTCTAAGGAAGACTATCATTCAGACCACA
                **********   ***********.**********.**********************

233781          CATAAGGATGGTAATGATGATGAAGTAGTTGTAGTAGAAGAACCGGCAAAAACAGCATCC
250586          CATAAGGATGGTAATGATGATGAAGTAGTTGTAGTAGAAGAACTGGCAAAAACAGCATTC
                ******************************************.**************.*

233781          CCATCAGCTTTATCAACATCAACAACAACAGCAGCAA------------CAACAGCAGCA
250586          CCATCAGCTTTATCAGCATCAACAACATCAACAGCAATATCAACAACAGCAGCAACAACA
                ***************.*********** **.******           **.**.**.**

233781          ACAACAGCAACAGCAACGTCTCTAATAATCGATCTAACAATTGATGATGATCCCGATAAC
250586          TCAACATCAACATCAACGTCTCTAATAATCGATCTAACAATTGATGATGATCCCGATAAC
                 ***** ***** ********************************************

233781          GATGTGGTCGGAGGAAGTCTTGTCCCTGCGCGGGCAGTTGAATC----------------
250586          GATGTGGTCGGAGGAAGTCTTGGTCGCA---GAGTAGTTGAATCTCAAGCTGAAGCTGCA
                ********************* .* ..   *.*.*********
```

```
233781          ------------------------------------AAAAGGTAAACCAAAAATGAAAAGA
250586          TCTGGCGCGGTGGATGTGAATGTGGCGTCTAGAAATGGAGGTAAACCAAAAATGAAAAGA
                                                   ..**********************

233781          GCGAAGAGACAAATGTTACCATTGAATCGTCGAAAGATGCAAGATGAAGAGGCGGATAAT
250586          GCTAAGAGACAAATGTTACCATTGAATCGTCGAAAGATGCTCGGAGGA------------
                ** ************************************  *. *.*

233781          AATAACGATGTGGTCG-----GAGGAAGTCGTGCCTCATCAAAATTATTAAGAATGGAGA
250586          ---AATCTTGTGGTTGCGACCAACAAACCCAATCCTTCGCCTTCTTCTCAATCATCAAAA
                   **.  ******.*     .* .** .*.  ***.  *    ** *.**  ** .*.*

233781          TGATAATGAAATTACCAACCGCCGAGGAATTAAAAGCAGAAAGCATTGTGATTAATGATG
250586          T--TAATAACAATGCCAACCGCCGAGGAATTAAAAGCAGAAAGCATTGTGATTGATGATG
                *  ****.* * *.****************************************.******

233781          TGAATGATCATGATATTCTAATGGGATCACGGAAATGTAACAAACATCCTGGGAATAAAG
250586          TGAATGATCATGATATTCTAATGGGATCAAAGAAATGTAACAAACATCCTGGGAATAAAG
                **************************** .****************************

233781          TGTATAGGGATATTGTGCGAAAATATCAACCACTATTGGAGAAGGAGACTATAGGTGATA
250586          TGTATAGGGATATTGTACGAAAATATCAACCACAATT------GGAGACTATAGGTGATA
                ****************.*************** ***      *****************

233781          GAGCGATAGTTGTTAGTATGGTTATTGATCATATACATGATCAAATTGGTGGTCGCTTCT
250586          GAGCAATAGTTGCTAGTATGGTTATTGATCATATTCATAATCAAATTGGTGGTCGATTCT
                ****.*******.********************** ***.**************** ****

233781          TAAAAGTCAACAGTGCTAATCAATGGCATGTTGTTCCGCGATTGGATGTCATAACCAAGA
250586          TAAAAGTTGATAGTACTAATCAATGGCATGTTGTTCCGCGATTGGATGTCATAACCAAGA
                *******..*.***.*********************************************

233781          TTACAAAAGCATTAGTTGAATTGGGAAATCCAGCTATTTTTCGTCTCGCCCTCCCAAAAT
250586          TTACAAAAGCATTAGTTGAATTGGGAAATCCATCTATTCTTCGTCTCGCCCTCCCAAAAT
                ******************************** ******.*********************

233781          CTGTTCCTACTAGTACCATGTTGGAGTCGATTGGGGAGTCGATTGGACAATCAATTGTTG
250586          CTGTTCCTACTAGTACCATGTTGGAGTCGATTGGGGAGTCGATTGGACAATCAATTGTTA
                ************************************************************.

233781          GGCAAGATAACAAAAGACCTAAGCGAAAATGTACCCTCCCAAAAGTGATAGAGGAACAAG
250586          TCAAAGATAACAAAAGACCCAAGCGAAAGTGCACCCTTCCAAAAGTGATAGAGGAACAAG
                   ****************.********.**.******.**********************

233781          TCGATGATGAACCAATGGAAAAGAAATTTACTAAAACGACGACGATGCGGGAAAGAAGTA
250586          TCGATGATGAACCAATGGAAATGAAATTTACTAAAACGACGACGATGCGGGGAAGAAGTA
                ********************* ******************************.********

233781          GTAGAATCGAGGATGAGAAGATTACAAAAATAGATGATGACGACGCTGAATCCAATAAGA
250586          GCAGAATCGAGGATGAGAAGATTACAAAAATAGATGATGACGACGCTGAATCCAATAAGA
                *.************************************************************

233781          ACGACGACAACAATGACGACCATCATCATTTAAGGATGGTGGCAACTGTGGTGGAGGAGG
250586          ACGACGACATCAATGACGACCATCATCATTTAAGGATGGTGGCAACTGTGGTGGTGGAGG
                ********* ******************************************** *****

233781          AGGAGGAGAATGGGTTTTCTTCTATTCCCACCACTGCCGCCGCCGCCGTGATGACGAATG
250586          AGGAGGAGAATGGGTTTTCTTCTATTCCCATCACTGCCGCCGCCGCCGTGATGACGAATG
                ****************************.*****************************

233781          CAAATTCTACGGCGGACGCTGATTCCATAATGTAAGTAAAAATCAAAACGAATTCAAAAT
250586          CAAATGCTACGGCGGACGCTGATTCCATAATGTAAGTAAAAATCAAAACGAATTCAAAAT
                ***** ****************************************************

233781          CGAATTCAAAATCGAATTCAAAATCGAATTCAAATTCAAATTTAAATTCAAATTCAGATT
250586          CGAATTCAAAATCGAATTCAAAATCGAATTCAAAATCGAATCAAA--TCAAAT-CAAAT-
                ********************************* **.***. **  ****** **.**
```

```
233781          CAAATATCCCGCTGTACTGCATTGGGGGTATTCGTATTCGTATTAAAAATCGTTCCTATT
250586          ------------------------------------------------------------

233781          TTTTTAAAAATTGATGCATATTCATATCTTTTTCCTTTCTTTTAATAATTTTTATCCTTA
250586          ------------------------------------------------------------

233781          TTTTTTTGTGCATTCATTAATACTTCGGCCGGCTAAATATATAAATTTATATGATGATAA
250586          ------------------------------------------------------------

233781          TATTAATAATGACGGAATATAGTCCTACATCATGGAAGGATTGGGTATCAAGAGTAAATA
250586          ------------------------------------------------------------

233781          ACTTAGTTGAATCATCACCATCATGTATGGATTACTCCCCTGCTTCTGAAAGATATAAGA
250586          ------------------------------------------------------------

233781          TTCTCTTAGATTACATTCCGATGGAAGATATTGTATCACAAATACGTTACAAACATATGA
250586          ------------------------------------------------------------

233781          TGTATCAGATTTTTATATCAGATCGTCCTGGTGATAGAGAATGTATTCATGCATTTCCGA
250586          ------------------------------------------------------------

233781          ATCAAATAAGTTTTAATTCAATAGAAAAAAAATACAACAAACAGATAAGTAATGGTAAAT
250586          ------------------------------------------------------------

233781          CTGTTAAGAAAAAAAGAAGTCAATTGCAAATGAAAACGGCGAATGACAAGTATCGTTCGA
250586          ------------------------------------------------------------

233781          AAAAAAAAAGGAAAACAGATGCATCTGATGCATTCAGGACGACGGTCGGTAGTATTCAAG
250586          ------------------------------------------------------------

233781          AAGAACAAGCTGCAAGGGCCACCACCATCACAACTGGTATTGCTCTAGAGGAAAAGAATA
250586          ------------------------------------------------------------

233781          ACGAATATTTGGAATTATTTACTGCTGCTGCTGCTGCAGCTGATACAACAACAGCAACAG
250586          ------------------------------------------------------------

233781          CAACAGCAACAGCAACAAAGGAAGTGGCGTGTGAAGAAATTAAAAGGGCTCTTTCGAATG
250586          ------------------------------------------------------------

233781          CAGGTATCGCTGCTGTTGCGAAGCAGAATGATACAAATGCTTCTGTGATTACACAAACCC
250586          ------------------------------------------------------------

233781          AGAAAGAATACACGCGTCTACCTCCTCACCGTACTTCCGCTGCTCGTAATAATAATAATA
250586          ------------------------------------------------------------

233781          ACGATAATAACAACAATAATAATGATAACAACAGTACCAAGCATCAGCAGCAGGAACAGC
250586          -------------------------------------------CATCAGCAGCAGGAACAGC
                                                           ******************

233781          AAAAATCAATGACAGCATCGACAAGGGGGCAGGACGAAGAAACGAATGCAACCCCGGTGT
250586          AAAAATCAATGACAGCATCGACAAGGGGGCAGGACGAAGAAACGAATGCAACCCCGGTGT
                ************************************************************

233781          CAAATCAATATCCGATAGCAATCGCCGCCACCGTCGATACTACTAGTCAATTGAAACCCC
250586          CAAATCAATATCCGATAGCAATCGCCGCCACCGTCGATACTACTAGTCAATTGAAACCCC
                ************************************************************
```

```
233781          AAACCAAACCATTGAGTCATATGCAGGGAGCTTATTCCAGTCGTCGTGAAAAGGTCATGG
250586          AAACCAAACCATTGAGTCATATGCAGGGAGCTTATTCCAGTCGTCGTGAAAAGGTCATGG
                ************************************************************

233781          CAAACATTAAAGAACTTCGTTCACAAATTAGTCAAGCAACATTCGATGAAGAAAAGATAG
250586          CAAACATTAAAGAACTTCGTTCACAAATTAGTCAAGCAACATTCGATGAAGAAAAGATAG
                ************************************************************

233781          CTTTTGAACAAGCTTTTAAATTAGAAATTGAATCATTAGGACGTTTAAATAAAGATGAAA
250586          CTTTTGAACAAGCTTTTAAATTAGAAATTGAATCATTAGGACGTTTAAATAAAGATGAAA
                ************************************************************

233781          TGAAATCGAAACTTCTCTTCGAAGGCGATAAGATAGATGTTATTGAAGAAGCTGAATTAG
250586          TGAAATCGAAACTTCTCTTCGAAGGCGATAAGATAGATGTTATTGAAGAAGCTGAATTAG
                ************************************************************

233781          TGAACGGATCCAATGCGTCCAATCCCACGACCAACAATGTAAATTTATTGTACCCCCCAT
250586          TGAACGGATCCAATGCGTCCAATCCCACGACCAACAATGTAAATTTATTGTACCCCCCAT
                ************************************************************

233781          ATAATCAATTTGGTATGGAGATGATGGCGAACGATTTTGGTTGTGGTGGTAGTGCAAATA
250586          ATAATCAATTTGGTATGGAGATGATGGCGAACGATTTTGGTTGTGGTGGTAGTGCAAATA
                ************************************************************

233781          TTGGCGTAATTGGCGGGATTTCCCATGGATTTGGAGCTAGTGGGAACTTAGGCGCCCCTT
250586          TTGGCGTAATTGGCGGGATTTCCCATGGATTTGGAGCTAGTGGGAACTTAGGCGCCCCTT
                ************************************************************

233781          ATTCGGCATCGTTTTATGCACAACAACAGATGTATCAATACTCCGTGGTGCATCCACATC
250586          ATTCGGCATCGTTTTATGCACAACAACAGATGTATCAATACCCCGTGGTGCATCCACATC
                ****************************************.*******************

233781          AAATTGTACATCAGCATCCGTATTTTCAACACCATCATCATAAGCATCAGGCGATTGCAG
250586          AAATTGTACATCAGCATCCGTATTTTCAACACCATCATCATAAGCATCAGGCGATTGCAG
                ************************************************************

233781          TGACGAACACTGATCGACCCGACGATCCATTTATAATGATGGACCCAGCTATACCCGACG
250586          TGACGAACACTGATCAACCCGACGATCCATTTATAATGATGGACCCAGCTATACCCGACG
                ***************.********************************************

233781          GTAGTAGCGACAAGAACAAAAGAGAACAGAATGAAAATGAAAATCGAGGAGGATGTACTG
250586          GTAGTAGCGACAAGAACAAAAGAGAACAGAATGAAAATGAAAATCGAGGAGGATGTACTG
                ************************************************************

233781          CTGTTACTGCTGATATCGCTACCCATTCTCTTCCCCTCAGTGAAGAACACACCGTAGCCC
250586          CTGTTACTGCTGATATCGCTACCCATTCTCTTCCCCTCAGTGAAGAACACACCGTAGCCC
                ************************************************************

233781          AACAGGATATTGCTCAAGATGACAGCATCATTAGCGATGCCGTCGACGGCAAAAATACCA
250586          AACAGGATATTGCTCAAGATGACAGCATCATTAGCGATGCCGTCGACGGCAAAAATACCA
                ************************************************************

233781          CCGGGTCTGATGGTAATGCCGATGCCAATTTCAATGCCAATACTAAACCTACTGCAACAA
250586          CCGGGTCTGATGGTAATGCCGATGCCAATTTCAATGCCAATACTAAACCTACTGCAACAA
                ************************************************************

233781          CAAAAGGAAAATGGACGCCCGAAGAGCATGAAGAAGTTGCGAAGGCAATGGCCAAATACG
250586          CAAAAGGAAAATGGACGCCCGAAGAGCATGAAGAAGTTGCGAAGGCAATGGCCAAATACG
                ************************************************************

233781          GACCTAGAGTAAGTGGGAAACAAATTTCAATTGAATTTGTTAAGGGTCGGACCCCCCTAC
250586          GACCTAGAGTAAGTGGGAAACAAATTTCAATTGAATTTGTTAAGGGTCGGACCCCCCTAC
                ************************************************************

233781          AACTCAATAGCTATATAAATCGCAAAAAAAAGTGAGTTATTAGCGACATGTAAAAAGTATA
250586          AACTCAATAGCTATATAAATCGCAAAAAAAAGTGAGTTATTAGCGACATGTAAAAAGTATA
                ************************************************************

233781          AACAAGACTACTGTGACGAGAGCGAGGATGACGACGACGGTGGCACCATTAGAGGATTGA
250586          AACAAGACTACTGTGACGAGAGCGAGGATGACGACGACGGTGGCACCATTAGAGGATTGA
                ************************************************************
```

```
233781          AGTTTCACCAAACAAGATCTTGTGATCGTGATGGGGATGACAAAAAAACAACTAATACCG
250586          AGATTCACCAAACAAGATCTTGTGATCGTGATGGGGATGACAAAAAAACAACTAATACCG
                ** ************************************************************

233781          ACGTCGAGAAGAATGTGCAGTACCGAAATGCGGAACGAGAACTACGTCGAACCAAAGATG
250586          ACGTCGAGAAGAATGTGCAGTACCGAAATGCGGAACGAGAACTACGTCGAACCAAAGATG
                ************************************************************

233781          GTTGTCTTCTTCCGAAAGGCGGAAAAGAAAAGTATTTGAAGGATGATGGAACATATAGGC
250586          GTTATCTTCTTCCGAAAGGCGGAAAAGAAAAGTATTTGAAGGATGATGGAACATATAGGC
                ***.********************************************************

233781          GCCCTGATGGAGCAAGGTAAGTCTTCCAAAACTACTATATACATGACTTCCATATTTTCT
250586          GCCCTAATGGAGCAACGTAAGTCTTCCAAAACTACTATATACATGACTTCCATATTTTCT
                *****.********* ******************************************

233781          TTATCTGAAATTACTCATCCTTTCTAACTTTCTAAATTTTGAATTAAATCAAACCACCGT
250586          TTATCTAAAATTACTCATCCTTTCTAACTTTCTAAATTTTGAATTAAATCAAACCACCGT
                ******.*****************************************************

233781          CACATCATACACACATAGACCCTTTGGATTATCTTGGCACAAAATTCGAGGTTTATGGGT
250586          CACATCATACACACATAGACCCTTTGGATTATCTTGGCACAAAATTCGAGGTTTATGGGT
                ************************************************************

233781          ACCATCTGAACGTTTAGAAGATAATGATGAGAACAACTACGACGACAACATCAACAAATC
250586          ACCATCTGAACGTTTAGAAGATGATGACGAGAACAACTACGACGACAACATCAACAAATC
                **********************.****.********************************

233781          GGGCTACACGAACTATAGTAGCGGCGATACCATCGCAAAAGCGACATCTTCCAATTGCGA
250586          GGGCTACACGAACTATAGTAGCGGCGATACCATCGCAAAAGCGACATCTTTCAATTGCGA
                ************************************************.*********

233781          GCAATCATATGATAGAAGTGCTTTACCAAGAGGCCTGAAAACGCATATCCGTGACCCTGT
250586          GCAATCATATGATAGAAGTGCTTTACCAAGAGGCCTGAAAACGCATATCCGTGACCCTGT
                ************************************************************

233781          CGGAGGTTGCTACTGGACCCCTTTAGGTTCAAGAAGGAAACTAACTGCAAAAGAGGCTTC
250586          CGGAGGTTGCTACTGGACCCCTTTAGGTTCAAGAAGGAAACTAACTGCAAAAGAGGCTCC
                ************************************************************.*

233781          ACGCAAGTCCAAGAAAAGAAAGCCAGGGCGTCAAAGTGCAGGAGCCAAAACGAAGGAGAA
250586          ACGCAAGTCCAAGAAAAGAAAGCCAGGGCGTCAAAGTGCAGTAGCCAAAACGAAGGAGAA
                ****************************************** ******************

233781          GGAGACGAGAGCGTTGGCGTACGTAACACCTCTCGAAATACTTCAGGGGAAAAAACCAAC
250586          GGAGACGAGAGCGTTGGCGTACGTAACACCTCTCGAAATACTTCAGGGGAAAAAACCAAC
                ************************************************************

233781          TCCGTCTAATCTTTTCCTTTCTTTGCACAGTGTCATTCCTGAAGCGGCGATGAATGAAAA
250586          TCCGTCTAATCTTTTCCTTTCTTTGCACAGTGTCATTGCTGAAGCGGCGATGAATGAAAA
                ************************************ ***********************

233781          CTTTAAGGACGACGATGATGATGATGACGATGACGAAGGATACGAATCTTGGACGAGTGG
250586          CTTCAAGGACGACGATGATGATGATGACGATGACGAAGGATACGAATCTTGGACGAGTGG
                ***.********************************************************

233781          CTCGTGGTGTTTACTTCAAGCTCAAAGGGATGCTTCGGCATCGGCAGTAGCAGAATCAGA
250586          CTCGTGGTGTTTACTTCAAGCTCAAAGGGATGCTTCGGCATCGGCAGTAGCAGAATCAGA
                ************************************************************

233781          AGCGAAAAAGTCTGCCCCCGACGAGGAAGAACAACAAGTACGCAAATGCAAATCTATTGC
250586          AGCGAAAAAGTCTGCCCCCGACGAGGAAGAACAACAAGTACGCAAATGCAAATCTATTGC
                ************************************************************

233781          AGCCGAGGCCAAAGCTAAAGAGCGCGAGCGTATTGGAGCTACCTGTACCGGAAATCAGAA
250586          AGCCGAGGCCAAAGCGAAAGAGCGCGAACGTATTGGAGCTACCTGTACCGGAAATCAGAA
                *************** ***********.**********************************

233781          AGACGAGTCCGCGAATTGTGGGGATGATAGTAGCCAAGATAGTAGCAAACGCCGACGGCT
250586          AGACGAGTCCGCGAATTGTGGAGATGATAGTAGCCAAGATAGTAGCAAACGCCGACGGCT
                **********************.*************************************
```

```
233781          GAGCATATGTGAGCAATCTCGTATCATTAATCCTGTTCAGAAAATTAATGTTCAGAGAAA
250586          GAGCATATGTGAGCAATCTCGTATCATTAATCCTGTTCAGAAAATTAATGTTCAGAGAAA
                ************************************************************

233781          AAAGAAGAAAAGTTTCTTGAAAGCAAAACAAGATGCTCGTGACTATATGTTGGCCAAGTA
250586          AAAGAAGAAAAGTTTCTTGAAAGCAAAACAAGATGCTCGTGACTATATGTTGGCCAAGTA
                ************************************************************

233781          CGGTCAAGGGAATGAGGAAGAAGAAATTGTAATGGTGTAA
250586          CGGTCAAGGGAATGAGGAAGAAGAAATTGTAATGGTGTAA
                ****************************************
```